THE GENETICAL STRUCTURE

OF

NORTHEASTERN BRAZIL

A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF THE

UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN GENETICS

JANUARY 1966

By

Norikazu Yasuda

Thesis Committee:

Newton E. Morton, Chairman
Chin S. Chung
Estel H. Cobb
Ming-Pi Mi
Lucian M. Sprague

ACKNOWLEDGMENT

I wish to express my sincere thanks to Dr. Newton E. Morton, Professor of Genetics, for his interest, supervision and encouragement throughout these investigations, and for his helpful advice and stimulating discussion in the preparation of this dissertation. Parts of the manuscript were read by Dr. G. Malecot, to whom I am grateful for his advice.

I am indebted to Dr. M. P. Mi for teaching me the technique of computer programming and for his discussion during its progress.

I also wish to thank especially Dr. C. W. Cotterman for aid in designing and equipping the Brazilian project, Dr. H. Cagno for providing space and assistance at the Hospedaria de Imigrantes, São Paulo, Brazil, and Dr. C. Pavan and Dr. O. Frota-Pessoa for facilitating the work in Brazil. Thanks are also due to Dr. W. Nance and Dr. O. Smithies who carried out the hemoglobin and serum protein, Dr. A. Steinberg the Gm and Inv determinations, and Dr. R. E. Rosenfield the first thousand f, V, Lu and Js tests and supplementary tests for other systems.

I would like to thank Dr. E. Azevedo, Dr. A. Freire-Maia and Mr. H. Krieger for their cooperation in the Brazilian study.

I am grateful to Antonia Alipio, Urusana Braga, Maria Dias, Ana Kokovin, Lourdes Moreira, Vera Nehemy, and Raimunda Santos for help in collecting the data, to Daisy Almeida, Dertia Freire-Maia, Gail Mackey, Barbara Plummer, and Joan Sellner for technical assistance in blood typing, and above all to the families whose cooperation made this study possible.

I owe much also to the Computing Center of the University of Hawaii, where the data were analyzed with IBM 7040 computer. I also wish to acknowledge a grant from the U. S. National Institutes of Health, which supported the entire work.

My gratitude goes to Dr. Y. Hiraizumi and Dr. M. Kimura who offered me to follow a career in human population genetics.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

From the standpoint of human population genetics, one of the most pertinent problems is to assess frequencies of individual phenotypes and mating types in terms of gene frequencies and the inbreeding coefficient. An assumption of random mating is often made in order to estimate gene frequencies and then genotype and mating type frequencies based on Hardy-Weinberg binomial law. However, there might be several genetic barriers to prevent a random combination of genes.

The first part of the present thesis is devoted to a theoretical investigation on genetic barriers which has led, as an extension of Wahlund's principle, to derivation of mating type frequencies as a function of gene frequencies and the inbreeding coefficient, provided that the inbreeding coefficient is not greater than the smallest gene frequencies (which is true in almost all human polymorphic systems). In this connection, the effect of inbreeding on segregation analysis is also examined.

As a parameter describing genetic barriers, the most meaningful interpretation of the inbreeding coefficient is as a coefficient of correlation between uniting gametes in Wright's sense. The positive correlation which measures effects of genetic barriers on combination of genes consists of ascertained and unascertained consanguinity. The negative correlation which may be observed in a small population is also included in the unascertained inbreeding coefficient. The total inbreeding coefficient is due to contributions from both close and remote consanguinity.

In the second part, a method of maximum likelihood scoring is developed to estimate simultaneously population gene frequencies and the

inbreeding coefficient from individual frequencies and from mating type frequencies by use of an electronic computer. It has been found that the two allelic system with complete dominance and the ABO blood group system do not give any information about the inbreeding coefficient when individual data are used.

The present theory has been tested with 1068 families from north-eastern Brazil by employing sixteen polymorphic systems. Pedigree analysis and bioassay revealed that the remote inbreeding coefficient is as great as the close inbreeding coefficient that is ascertainable from pedigree analysis. Therefore the elimination rate for rare recessive genes is greater than had been estimated previously. Racial endogamy contributes only 22 per cent of the total inbreeding coefficient for polymorphisms, and less for rare genes (monomorphisms).

To describe human population structure, the marital distance, defined as distance between birth places of mates, is the most pertinent measure. An exponential relation between the inbreeding coefficient and the marital distance has been predicted by Malecot. This is a good approximation in remote consanguinity, but deviation apparently due to preferential mating of relatives is found for close consanguinity. A tendency for the inbreeding coefficient to decrease by generation cannot be detected through pedigree study due to incomplete ascertainment, but is found by the new methods of bioassay and study of migration functions.

1. Introduction

Population genetics may be divided into two major fields, either study of populations in which gene frequencies are assumed to be constant, or populations in which gene frequencies vary in time. The main problem in the first branch is to describe phenotype (or genotype) frequencies in terms of gene frequencies and factors which depend on the structure of population. From the standpoint of human population genetics, the pertinent problem is concerned with all those factors which determine mating preference.

The research in this line is still infant due either to lack of reasonable theories or presence of several difficulties in field work. For example, the Hardy-Weinberg binomial law (Hardy, 1908; Weinberg, 1908) may be applied under the assumption of random mating. In human populations, however, the assumption of random mating may be unrealistic, since there exist several factors which cause departure from the Hardy-Weinberg law. Apart from selection, mutation and random genetic drift, there is isolation due to geographical, racial, religious, social, economic, professional, and other barriers. The effect of this isolation on the relation between gene frequencies and genotype frequencies is evident in a population with local differentiation of gene frequencies. Furthermore, isolation may result in assortative mating and consanguineous marriage. Migration in subdivided populations gives rise to racial mixture, clines, and discontinuous gene distributions. Much research has been made on the proportion of consanguineous marriages in man, but little has been studied in relation to isolation.

It was Wahlund (1928) who worked out some mathematical consequences of a subdivided population to explain why the breakdown of isolates decreases homozygosity. The following year, Dahlberg (1929) formulated the idea of "isolate size" by pointing out an important demographic concept that the number of consanguineous marriages was closely related to family size and population number. These results were, however, not applied to actual populations.

In the field of quantitative genetics, Wright (1921) introduced the inbreeding coefficient, F, as a measure of nonrandomness between uniting gametes. Later, Malecot (1948) established that the inbreeding coefficient could be understood from the standpoint of probability theory. The most significant application of the inbreeding coefficient in population genetics has been the expression of genotype frequencies as a function of gene frequencies and the inbreeding coefficient.

It is not surprising therefore that Wright (1943, 1956, 1951) and Malecot (1948, 1950, 1959) have made an attempt to adapt their model to describe human population structure. Wright extended his theory to include isolation by distance, retaining the notion of isolate size. He derived the mean inbreeding coefficient in terms of systematic pressure on gene frequency and neighborhood size, defined for a normal migration in two dimensions as the effective number in a circle of radius twice the standard deviation of offspring in one direction. Malecot, on the other hand, derived the inbreeding coefficient as a function of marital distance, defined as the distance between birth places of mates, from a stochastic equation with an empirical migration function. He suggested that the rate of decreasing in inbreeding coefficient by distance was independent of neighborhood size.

Both models have been applied to several human populations and have encountered difficulty in estimating isolate or neighborhood size and distance between mates born within the same parish, village, or other demographic unit.

Combining theory with practice, it has been realized that one of the main problems in studying population structure in man is to ascertain the inbreeding coefficient of a given population.

Recently, I have developed a model from Wahlund's principle which permits us to evaluate mating type frequencies in nonrandomly mating populations and to estimate the total inbreeding coefficient, including the contribution due to remote consanguinity. Also, I have generalized the meaning of the inbreeding coefficient for describing human population structure. In this thesis, a systematic description of population structure in man will be given in the second chapter, some statistical methods will be discussed in the third, and then the theory will be applied to a population from northeastern Brazil (Morton, 1964).

2. Theoretical Studies

2.1. Randomly mating populations

A diplophase generation begins when gametes from a gene pool are combined pairwise into zygotes according to some rule; these zygotes experience migration, mutation, and differential mortality and fertility; and the generation terminates with the haplophase gene pool of the next generation. By panmixia or random mating we mean that uniting gametes are drawn independently from the gene pool, without restrictions due to finite population size, inbreeding, or assortative mating, and are enumerated before differential selection has acted. Accordingly, genotype frequency and mating type frequency can be calculated by the Hardy-Weinberg binomial law.

Confusions in definition of random mating should be pointed out here, since only an infinite population allows an inbreeding coefficient of zero.

Thus the binomial calculation of expected genotype frequencies and mating type frequencies from a gene array can be applied approximately in a finite population, but this is not a Hardy-Weinberg population unless the population size is indefinitely large. The term "homogamy" has been proposed (Malecot, 1948) for random mating by zygotes, but in this thesis, we will consider only random combination of genes as random mating.

2.2. The inbreeding coefficient

Wright's inbreeding coefficient is the most meaningful quantity in population genetics to describe departure from randomness between uniting gametes. Consider a single locus with two alleles and their frequencies p and q (=1-p) in a given population. The probabilities that two gametes

should unite are p^2, $2pq$ and q^2 under random mating. When a positive correlation between uniting gametes is observed, these frequencies change in such a way that homozygosity increases and heterozygosity decreases, with no change in gene frequency. It is easily shown that the change in genotype frequencies is pqF, where F is the correlation coefficient between uniting gametes, also called as Wright's inbreeding coefficient (Table 2.1). Having need of theoretically clear bases for studying systems of mating in experimental populations, Malecot (1948) has given another important interpretation of Wright's inbreeding coefficient from the viewpoint of probability theory. Suppose that a given population is partitioned in two parts; in one fraction mating is random and in the other combination of gametes is restricted to those which are identical by descent. If f is taken as the latter proportion, that is also the probability that two genes are identical by descent, then the increment of each homozygote is given by pqf (Table 2.1). Thus $F=f$, and the two definitions are identical.

When we consider several such populations together in which gene frequencies are the same but f may vary among populations, the frequencies of three genotypes in the total population will be $\Sigma w_i p(p+qf_i)$, $\Sigma w_i 2pq(1-f_i)$ and $\Sigma w_i q(q+pf_i)$, where w_i is the relative size of the i-th population whose fraction f_i is the probability that two genes are identical by descent and summation takes over all populations. Those proportions may be written in terms of the mean inbreeding coefficient α $(=\Sigma w_i f_i)$, that has been introduced a priori by Bernstein (1930), (Table 2.1). Bernstein's coefficient of inbreeding may also be interpreted in a population whose inbreeding coefficient f is heterogeneous.

It is obvious, therefore, that the inbreeding coefficient can be understood as a measure of non-randomness that also describes zygote frequencies, the correlation between uniting gametes, and the probability that two genes are identical by descent. Besides these, it can measure degree of differentiation in subdivided populations, and mating type frequencies can be described by it.

Wright (1943) has proposed the inbreeding coefficient for populations with hierarchic structure. Defining the panmictic index as the complement of the inbreeding coefficient, the panmictic index of individuals relative to the total is equal to the product of the hierarchical panmictic indices. This is merely an approach to defining the inbreeding coefficient between populations in his study. Malecot (1950) has given a recurrence relation of the inbreeding coefficient among populations. (Detail of this will be explained with discussion of distance since his works are less familiar so far.) This line of development of the coefficient of inbreeding, taking an arbitrary reference population will be helpful when we consider a subdivided population.

In addition to the above general definition of the inbreeding coefficient, it is worthwhile to consider the number of coefficients for an autosomal locus. Since the definition of the inbreeding coefficient is concerned with only one gene, if the number of alleles is n, then the number of inbreeding coefficients at the locus is also n. Consequently, there are n-1 independent inbreeding coefficients at a locus. The inbreeding coefficient of the locus, f, is then,

$$f = \Sigma f_i p_i$$

where f_i and p_i are the inbreeding coefficient of the i-th allele and its

frequency, respectively. In a polymorphic system f_i may be nearly f for all i. Discussion on this point will be given in chapter three. With definition of the inbreeding coefficient of an allele, it is of interest how the heterozygote frequency shall be written in terms of gene frequencies and the inbreeding coefficient of alleles. Suppose that three genotype frequencies, say P_{ii}, $2P_{ij}$, and P_{jj} $(i \neq j)$ are expressed with respect to two alleles frequencies p_i and p_j (it is not necessary that $p_i+p_j = 1$), their inbreeding coefficient f_i and f_j, and a decreasing proportion of heterozygote, f_{ij}

$$P_{ii} = p_i^2 + p_i(1-p_i)f_i$$

$$2P_{ij} = 2p_ip_j(1-f_{ij})$$

and

$$P_{jj} = p_j^2 + p_j(1-p_j)f_j.$$

Since the sum of three genotypes should be expressed in form, $R(1-F) + IF$, where $R = (p_i+p_j)^2$, $I = (p_i+p_j)$ and $F = (p_if_i+p_jf_j)/(p_i+p_j)$, then we obtain

$$f_{ij} = (f_i+f_j)/2.$$

Namely, the fraction in which heterozygosity is depressed can be calculated from a simple average of the inbreeding coefficients of the two alleles.

2.3. Wahlund's principle

2.3.1. Discrete model

Suppose that a population is divided into many endogamous panmictic smaller populations ("isolates") restricted by geographical, racial, religious, social, or 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 frequency p_i and q_i in the i-th isolate, respectively, then the mean frequency of gene A, p, and its variance, σ^2, in the total population are

$$p = \Sigma p_i w_i$$

and

$$\sigma^2 = \Sigma(p_i - p)^2 w_i = \Sigma p_i^2 w_i - p^2,$$

respectively, where summation is over all isolates. Since the frequencies of AA, Aa and aa genotypes in the total population are given by $\Sigma p_i^2 w_i$, $2\Sigma p_i q_i w_i$ and $\Sigma q_i^2 w_i$, the subdivision results in increasing homozygosity by an amount equal to the gene frequency variance, σ^2 (Table 2.1). Wahlund (1928) discovered this result and discussed it in the cases of dominance and co-dominance.

Comparison of heterozygous frequencies with Wright's or Malecot's result leads to Wright's formula

$$\sigma^2 = p(1-p)F \qquad \text{(Wright, 1949).}$$

All of the above arguments hold for an arbitrary number of alleles, for each of which an inbreeding coefficient can be defined as in the last section, and this leads to an interesting formula:

$$F = \Sigma \frac{\sigma_{p_i}^2}{1-p_i}.$$

It should be borne in mind that an "artificial" subdivision of a population does not always result in increasing homozygosity. There would be no change observed whenever a given gene frequency was exactly the same for all isolates, i.e. gene frequency variance was zero. This suggests an association between isolate size and probability density of gene frequency which will be discussed in the succeeding section.

2.3.2. The breakdown of isolates

Since the Wahlund's principle has mainly been employed to explain why the breakdown of isolates decreases homozygosity, there has been no mathematical treatment of how the proportion of homozygosity decreases by removing a single barrier. To visualize the situation, the following discussion may be helpful.

Suppose that a population consists of three isolates 1, 2 and 3 whose relative sizes are w_1, w_2 and w_3 ($\Sigma w_i=1$), and in which frequencies of gene A are p_1, p_2 and p_3, respectively. In this population, the gene frequency p_{III} and its variance σ^2_{III} are

$$p_{III} = p_1w_1 + p_2w_2 + p_3w_3$$

and

$$\sigma^2_{III} = p_1^2w_1 + p_2^2w_2 + p_3^2w_3 - p^2_{III},$$

respectively. Suppose that the barrier between the isolate 2 and 3 is removed, creating a new isolate in which mating ultimately goes on at random, (perhaps after a few generations in which a gene cline persists). The relative size W of this new panmictic isolate and its gene frequency P are

$$W = w_2 + w_3,$$

and

$$P = (w_2p_2 + w_3p_3)/W,$$

respectively.

The gene frequency, p_{II} and its variance σ^2_{II} of the total population become, therefore,

$$p_{II} = p_1w_1 + PW$$

$$= p_1w_1 + p_2w_2 + p_3w_3 = p_{III},$$

and $$\sigma_{II}^2 = p_1^2 w_1 + P^2 W - p_{II}^2$$
$$= \sigma_{III}^2 - [p_2^2 w_2 + p_3^2 w_3 - (p_2 w_2 + p_3 w_3)^2/(w_2 + w_3)]$$

or

$$\sigma_{II}^2 = \sigma_{III}^2 - \frac{w_2 w_3}{(w_2 + w_3)} (p_2 - p_3)^2.$$

Apart from mutation, selection and random genetic drift, the population gene frequency does not change, whereas the gene frequency variance decreases in an amount that depends on the relative sizes and difference of gene frequencies of isolates whose barriers are removed. As a corollary, the change in the inbreeding coefficient is given by using Wright's formula and $p_{II} = p_{III} = p$,

$$F_{II} = F_{III} - F_B,$$

where F_B is a contribution due to the breakdown of isolates and, in our terminology, $F_B = \left[\frac{w_2 w_3}{w_2 + w_3}\right] \frac{(p_2 - p_3)^2}{p(1-p)}$. More general treatment is given in Appendix 1.

This elaboration of Wahlund's results may be applied to human populations. For instance, the barrier that is removed may be racial endogamy, and the effect of this on the inbreeding coefficient is immediately apparent. (See application to Brazilian data below.) On the other hand, when new barriers are created under a certain circumstance, it is clear from the above discussion that the inbreeding coefficient increases with the amount F_B.

2.4. Extension of Wahlund's principle

Wahlund's principle described the basic effects of partitioning a population, but there still remain several aspects which are valuable for studying systems of mating in non-experimental populations.

2.4.1. Continuous model

Although it has been assumed that the barriers are discrete, an actual barrier is usually continuous, or we may not know what type of barrier it is. One of the approaches to bridge the gap is, then, extension of Wahlund's model to continuous or mixed barriers. Since the result from Wahlund's discussion is described in terms of mean and variance of population gene frequencies, sums can be replaced by integrals. In this continuous model each individual gene has a "probability density" to contribute to population gene, genotype, or mating type frequencies. Therefore, gene frequency and its variance in the population can be expressed by Lebesgue-Stieltjes integral (Cramer, 1946),

$$p = \int p_w dW$$

and

$$\sigma^2 = \int p_w^2 dW - p^2,$$

where sums are taken for the discrete model and integrals for the continuous case. In the mixed case the barriers may be separated into discrete and continuous type. Thus, Wahlund's principle covers any type of heterogeneous population. For instance, the continuous model where a population is divided by distance has been studied by Holgate (1964). Furthermore, in case of subdivision by time or generation the probability density, w, may correspond to the solution of the Fokker-Planck equation in population genetics (Wright, 1946). The situation in man is so complicated by factors such as time, space, population size, and human

behavior that it may be difficult to find, even approximately, the appropriate probability density function. It should be emphasized here, however, that Wahlund's principle holds even for an unknown density function, and this extension replaces the concept of "isolate size" by a "probability density of gene frequencies". A genetical interpretation of the probability density could be as a tendency of genes to combine that would be affected by several genetic barriers.

2.4.2. Moments of a subdivided population

Since gene and genotype frequencies of a population are given by the first and the second moments with respect to possible isolates in the population, it seems worthwhile to consider the biological meanings of the higher moments. The first moment gives the gene frequencies, and the second moment the genotype frequencies. The third and the fourth moments gives the mating type frequencies at a sex-linked and an autosomal locus, respectively, since three and four genes are concerned with each gene combination. More generally, whenever we consider a set of genes, the order of the moment corresponds to the number of gene involved. These higher moments appear in studies of linkage, illegitimacy, polyploidy, heritability and so on, but we shall restrict attention to the fourth and lower moments that correspond to mating type frequency for study of population structure in man, though the results are completely general.

Let us consider a locus with two alleles A and a, and their frequency p and q(=1-p) in a subdivided population with the inbreeding coefficient α. Suppose that the difference between gene frequency of an isolate p_w and of 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 integrals are in the sense of Lebesgue-Stieltjes. For the first and second moments, the following relations hold precisely:

$$m_1 = 0,$$

and $$m_2 = p(1-p)\alpha.$$

For the population moment M_a,

$$\begin{aligned} M_a &= \int p_w^a dW = \int (p+\Delta p_w)^a dW \\ &= \int \sum_{r=0}^{a} \binom{a}{r} p^{a-r} (\Delta p_w)^r dW \\ &= \sum_{r=0}^{a} \binom{a}{r} p^{a-r} m_r, \end{aligned}$$

or

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

In the above expression, if cubic and higher power of Δp_w are negligible, it follows that

$$M_a = p^a + \frac{a(a-1)}{2} p^{a-1}(1-p)\alpha \qquad (a = 0,1,2,\ldots).$$

for example,

$$M_1 = p$$

$$M_2 = p^2 + p(1-p)\alpha$$

$$M_3 \doteq p^3 + 3p^2(1-p)\alpha$$

and $$M_4 \doteq p^4 + 6p^3(1-p)\alpha.$$

Exact expression of the moments by gene and the inbreeding coefficient may be made when a distribution function of isolates is known. For instance, one or two parameter probability functions such as binomial, Poisson, normal, exponential, gamma and beta distributions have been

applied to this case, the beta probability being especially interesting because it corresponds to a steady state distribution of gene frequency (Wright, 1931) (Appendix 2). All these cases indicate that the moment of population can be expressed as a polynomial of α, with the quadratic and higher order powers negligible when α is not greater than p or 1-p.

However, it is extremely difficult to determine the distribution of isolate size in human populations as stated previously. It is, therefore, necessary to approach this problem without assuming a distribution. In the above general argument we assumed that higher than cubic moment of Δp_w are negligible. This limiting form is valid provided that all gene frequencies exceed the inbreeding coefficient, as is certainly the case for the polymorphisms to which this model will be applied. Extensive study with known distribution forms has suggested that whenever isolate size distributions are symmetrical, then $m_3 = 0$ and $m_4 = 0(\alpha^2)$. Even when asymmetric functions such as gamma are assumed, the limiting form holds with sufficient accuracy if the smallest gene frequency is greater than the inbreeding coefficient, which does not exceed two per cent in human populations (Wright, 1950).

The population moments as a function of gene frequencies and the inbreeding coefficient can be obtained in case of more than two alleles at a given locus. Since algebraic argument will be given in Appendix 3, the results are simply reproduced here.

For the tri-allelic locus with frequencies p, q and r (p+q+r=1), the moment $M_{a,b}$ in limiting form is

$$M_{a,b} = \int p_w^a q_w^b dW$$

$$= p^a q^b + \left[\frac{a(a-1)}{2} p^{a-1}(1-p)q^b + \frac{b(b-1)}{2} p^a q^{b-1}(1-q) - abp^a q^b\right]\alpha$$

$$(a,b = 0,1,2,\ldots).$$

For more than three alleles,

$$M_{a,b,c,d} = \int p_w^a q_w^b r_w^c s_w^d dW$$

$$= p^a q^b r^c s^d + p^{a-1} q^{b-1} r^{c-1} s^{d-1}\left[\binom{a}{2}(1-p)qrs + \binom{b}{2}p(1-q)rs + \right.$$

$$\left.\binom{c}{2}pq(1-r)s + \binom{d}{2}pqr(1-s) - (ab+ac+ad+bc+bd+cd)pqrs\right]\alpha$$

$$(a,b,c,d = 0,1,2\ldots),$$

where $\binom{x}{2} = x(x-1)/2$ and it is not necessary that $p+q+r+s = 1$.

The most general formula to be obtained for the moments of population is as follows:

$$M_{a_1,\ldots,a_n} = \int p_{1w}^{a_1} \ldots p_{nw}^{a_n} dW$$

$$= \prod_{i=1}^{n} p_i^{a_i} + \left(\prod_{i=1}^{n} p_i^{a_i - 1}\right)\left[\sum_{i=1}^{n}\binom{a_i}{2}(1-p_i)\left(\prod_{j\neq i} p_i\right) - \right.$$

$$\left.\left(\prod_{i=1}^{n} p_i\right)\left(\sum_{i>j} a_i a_j\right)\right]\alpha \qquad (a_i = 0,1,2,\ldots \text{for all } i).$$

2.5. Mating type frequencies

As mentioned in the preceding section, mating type frequencies of a given genetic system can be obtained from the moments of population. It is thus straightforward to evaluate the frequencies in the case of two alleles at an autosomal and a sex-linked locus. To illustrate, let us take the intercross $\underline{Aa}$ x $\underline{Aa}$ and its relative frequency fr. In an

isolate the proportion of this type is $4p_w^2(1-p_w)^2dW$ so that

$$\begin{aligned} fr &= \int 4p_w^2(1-p_w)^2 dW \\ &= 4M_2 - 8M_3 + 4M_4 \\ &= 4p^2(1-p)^2 + 4p(1-p)[1-6p(1-p)]\alpha \\ &= 4p^2q^2 + 4pq(1-6pq)\alpha, \end{aligned}$$

where $p+q = 1$ and M denotes the population moments defined in section 2.1. Mating type frequencies and the proportions of their possible children in the limiting form are shown in Table 2.5.1. for the autosomal locus and in Table 2.5.2. for the sex-linked locus. In the latter case, we assumed that gene frequencies are the same in both sexes. Justification of the moment method to describe mating type frequencies is immediate when frequencies of possible offspring are evaluated as $p^2+pq\alpha$, $2pq(1-\alpha)$ and $q^2+pq\alpha$ for genotypes AA, Aa and aa, respectively, at the autosomal locus without dominance as well as in the other cases. Mating type frequencies when a distribution is assumed are also given in Table 2.5.3. for the autosomal locus and in Table 2.5.4. for the sex-linked locus. (Incidentally, both normal and rectangular distributions give exactly the same frequencies for sex-linked mating types.) These results permit us to apply them to a higher level of inbreeding population, justifying the distribution assumption. When α approaches unity, incross frequencies go to corresponding gene frequencies as happens also for genotype frequencies with beta and binomial distributions. For the other distributions examined, a convergency of incross frequencies to the gene frequencies failed when $\alpha \to 1$. As $\alpha \to 1$, the distribution condenses into 2 poles, at 0,1. This cannot be represented by one of these distributions. Since

our purpose is to describe human population structure, we are not going to discuss further in this line.

It may be worthwhile to note here that the correlation coefficient, m, between mates is equal to $2\alpha/(1+\alpha)$ at an autosomal locus without dominance, which was first given by Wright (1921) and was discussed by Li (1955). This indicates that if there is no genetic correlation between mates, then the inbreeding coefficient is zero and mating is random (and vice versa).

Dominance does not create any difficulty to obtain phenotypic mating type frequencies since it requires simple additions of terms of genotypic mating type frequencies whose phenotypes are same (Table 2.5.1. and 2.5.2.).

The main effect of inbreeding on the frequencies of zygotes has already been mentioned as a decrease of heterozygosity. In Figure 2.5.1 we demonstrated the amount of decreasing on zygote frequencies due to inbreeding in case of two alleles. The abscissa denotes the frequencies of gene A and the ordinate stands for the coefficient of α, or inbred component of zygote frequency, I. The effects of inbreeding are most enhanced when both genes are in equal frequency.

This presentation will extend to the mating type frequencies. Only a two-allelic locus will be discussed here, since the essential features of the inbreeding effects can be observed in it. In the following discussion, p and q stand for frequencies of gene A and a, respectively, where gene A may be dominant over gene a. Mating type frequencies can be written in the form, $R+I\alpha$ in the neighborhood of $\alpha = 0$, where R is the mating type frequency in randomly mating populations and I is the inbred

component. In order to see the inbreeding effects we will examine I as a function of p.

Autosome, without dominance: Figure 2.5.2 gives us the general features of the inbreeding effects on the six different mating types. Both incrosses (AA x AA and aa x aa) are always increasing. Interestingly, both backcrosses (AA x Aa and aa x Aa) decrease when the gene frequency is small, and are compensatory to all other types of mating if $p < .212$ or $q < .212$. It is clear that the inbreeding effect is more striking in backcrosses than incrosses. As an extreme case, when $p = .18$, the I-value for backcross aa x Aa reaches a minimum, -.93. If we take $\alpha = .006$, the decreasing frequency due to inbreeding is $.93 \times .006 = .0056$ which is $.0056/.019 = .029$ or about three per cent of the mating type frequency calculated from the Hardy-Weinberg law. Thus an assumption of random mating for the estimation of gene frequencies in man might be justified as a first approximation.

Autosome with complete dominance: In Figure 2.5.3, A- denotes the dominant phenotype. The effect of inbreeding is greatest when the frequency of the dominant gene is nearly .25. The cross of both dominant phenotypes (A- x A-) compensates the other two matings if $p > .577$.

Sex-linked without dominance: General tendency of effects of inbreeding is similar to the autosome without dominance. The effects are rather weaker at the sex-linked locus than the autosomal locus. Figure 2.5.4 indicates the change of I by gene A frequency.

Sex-linked with complete dominance: The effects of inbreeding balance

each other when mating type frequencies in population are classified by male (or hemizygote).

In summary, the frequency of incrosses is always enhanced by inbreeding, as homozygotes are. The effect on the other types of crosses depends on the gene frequency. Roughly, the effects of inbreeding or subdivision of population on mating type frequencies are enhanced when gene frequency is nearly .25 (or .75 in co-dominant locus) instead of .5. These predictions can be immediately tested with mating type frequencies with such a locus as MN, Ss, PTC, Secretor, and so forth.

When there are more than two alleles at a locus, the number of possible mating types becomes very large. For instance, with three alleles at an autosomal locus, the possible numbers of zygotes and mating types are six and twenty-one, respectively, and with ten alleles the corresponding values become 55 and 1545. If a, g and m stand for the numbers of alleles, genotypes and mating types, respectively, then $m=g(g+1)/2=a(a+1)(a^2+a+1)/8$ for autosome and $m=ag=a^2(a+1)/2$ for sex-linked locus. In this large number of mating types, however, there are only seven basic types of crosses at the autosomal locus and four at the sex-linked locus. They are tentatively called "incross" (AA x AA), "backcross" (AA x AB), "intercross" (AB x AB), "outcross" (AA x BB), "3-ways-intercross" (AB x AC), "3-ways-outcross" (AA x BC) and "4-ways-intercross" (AB x CD) for the autosomal locus, and "incross" (AA x A), "outcross" (AA x B), "backcross" (AB x A) and "intercross" (AB x C) at the sex-linked locus, where A, B, C and D denote different alleles. Any dominance relation between alleles will diminish the number of mating types. These mating type frequencies are also derived from the population moments (Table 2.5.5. and 2.5.6.).

2.6. Effect of subdivision on Snyder's ratio (S) and the proportion of mating pairs which cannot segregate one phenotype (h)

Studies on modes of inheritance in man have received attention from geneticists as well as physicians ever since biochemical individuality was initially reported by Garrod (1902). One classical method that is still of importance, is pedigree analysis or study of familial distribution of a character. Population geneticists are no longer behind in this feature. Snyder (1932) gave a test of the hypothesis of autosomal recessive inheritance in non-tasting of phenylthiocarbamide (PTC). According to Snyder, the expected frequencies of recessive children from the mating T- x T-, T- x tt and tt x tt are S^2 : S : 1, respectively, where t is non-taster gene with frequency q, T is for taster gene and S is the conditional frequency of t gene among T- persons, or Snyder's ratio (Morton, 1965). S is expressed in terms of q; $S = q/(1+q)$. Insignificant deviations of observed frequencies of non-taster children from the ratio were taken for a decision that PTC tasting was dominant over non-tasting. The underlying assumptions in this method are unit inheritance, recessivity, random mating, complete penetrance, no extramarital children or classification errors, and no selection. Among these simplifications, random mating has a direct effect on frequencies of mating types and of their offspring. In the populations where mating is not at random but inbreeding is not high (say, less than 2 per cent), we may examine the effect of inbreeding on Snyder's ratio. From Table 2.5.1., we obtain two conditional probabilities S and R of recessive offspring from matings, dominant x recessive and dominant x dominant,

$$S = \frac{q^2 + 3q(1-2q)\alpha}{q(1+q)+(1-6q^2)\alpha} \doteq \frac{q}{(1+q)} + \frac{2-3q}{(1+q)^2}\alpha$$

and

$$R = \frac{q^2(1-q) + q(1-6q+6q^2)\alpha}{(1-q)(1+q)^2-2q(1-3q^2)\alpha} \doteq \left(\frac{q}{1+q}\right)^2 + \frac{q(1-3q-6q^2)}{(1+q)^4}\alpha,$$

respectively.

When mating is at random, $\alpha = 0$ so that $R = S^2$, where S is Snyder's ratio defined in a randomly mating population. Expansions in series form with respect to power of α are permitted whenever $q > \alpha < 1-q$, which is the essential condition for expressing population moments as a linear function of α. In Figure 2.6.1, the constants of the linear term of α (C_1 and C_2) are plotted against the recessive gene frequency q. The effect of subdivision or inbreeding is more pronounced on S than R when the recessive gene is rare, and little effect on R is observed for any frequency of recessive gene. Thus, Snyder's ratio is still useful for polymorphic systems without serious error from the assumption of random mating. (Especially, if the recessive gene frequency lies about between .3 and .7.)

In this method, however, the more serious error due to heterogeneity in segregation frequencies among families has been painted out in the usual chi-square test (Morton, 1965). No study will be made here on this topic except for one of parameters, h, the proportion of parents who cannot segregate (Morton, 1959, 1962) which was introduced to avoid the possible statistical errors in Snyder's ratio. This probability is directly related to inbreeding while segregation frequencies do not depend on system of mating. Therefore it may be worthwhile to examine the effect

of inbreeding or of subdivision of population on h-values.

We usually encounter three different types of h for an autosomal locus: (1) the proportion of homozygotes among the dominant phenotypes, designated by h_1; (2) the proportion of non-segregating couples among outcrosses with respect to the dominant and recessive phenotypes, h_2; and (3) the proportion of non-segregating couples among incrosses with respect to the dominant phenotype, h_3. Two more h-values are defined at the sex-linked locus (h_4 and h_5 for outcross and incross, respectively), relative to the probability that heterozygous females will give birth to both type of males regardless of phenotypes of their mate. For reference, let P(.) be the probability that an event . occurs. In this terminology, the five h-values are

$$h_1 = \frac{P(AA)}{P(A-)}, \qquad h_2 = \frac{P(AA \times aa)}{P(A- \times aa)}, \qquad \text{and}$$

$$h_3 = 1 - \frac{P(Aa \times Aa)}{P(A- \times A-)} \qquad \text{for autosomal locus,}$$

and

$$h_4 = \frac{P(AA \times a)}{P(A- \times a)} \qquad \text{and} \qquad h_5 = \frac{P(AA \times A)}{P(A- \times A)}$$

for sex-linked locus, where A and a are alleles with frequencies p and q (p+q=1), respectively, and A is dominant over a. Thus we obtain from Table 2.5.1. and 2.5.2. (Whenever $q > \alpha < 1-q = p$, the expression in linear forms is permissible except for h_4 where the additional restriction $|\alpha| < 2/3$ must be posed for mathematical reasons.)

$$h_1 = \frac{p^2+pq\alpha}{p^2+2pq+pq(1-2\alpha)} \doteq \frac{p}{1+q} + \frac{2q}{(1+q)^2}\alpha,$$

$$h_2 = \frac{2p^2q^2+2pq(1-6pq)\alpha}{2pq^2(1+q)+2pq(1-6q^2)\alpha} \doteq \frac{p}{1+q} + \frac{2(3q-2)}{q(1+q)^2}\alpha,$$

$$h_3 = \frac{p^3(1+3q)-6p^2q(1-3q)\alpha}{p^2(1+q)^2-2pq(1-3q^2)\alpha} \doteq \frac{p(1+3q)}{(1+q)^2} + \frac{4q(6q^2+3q-1)}{(1+q)^4}\alpha,$$

$$h_4 = \frac{p^2q+pq(1-3p)\alpha}{pq(1+q)-3pq^2\alpha} \doteq \frac{p}{1+q} - \frac{2(1-q)}{(1+q)^2}\alpha$$

and

$$h_5 = \frac{p^3+3p^2q\alpha}{p^2(1+q)+pq(2-3p)\alpha} \doteq \frac{p}{1+q} + \frac{4q}{(1+q)^2}\alpha.$$

In randomly mating population where $\alpha = 0$, all h converges to the same value except h_3. As we see in Figure 2.6.2, h_1 for the zygote is little affected by inbreeding. When the recessive gene frequency is low, the effect appears on h_2 and h_4 for outcrosses, while if the recessive gene frequency is high h_3 for incross will be affected. With the inbreeding coefficient $\alpha = .006$, for instance, if $q = .1$, h_1 changes .001/.818 = .001 or .1 per cent increasing relative to random mating, h_2 decreases about 2 per cent, h_3 decreases .1 per cent, h_4 decreases about 9 per cent and h_5 is enhanced about 2 per cent. While if $q = .9$, h_1 increases about 6 per cent, h_2 increases 4 per cent, h_3 increases about 11 per cent, and h_4 and h_5 have practically no effect. Thus it seems reasonable to adopt the assumption of random mating in segregation analysis, but when the recessive gene frequency is high, the effect of inbreeding on h_3 with respect to intercross segregation may not be negligible as well as on h_4 for outcross at the sex-linked locus when the recessive gene frequency is low.

2.7. Distance approach to ascertain remote inbreeding coefficient

So far discussion on population structure was concerned with gene frequencies and the inbreeding coefficient together. Since systems of mating themselves do not alter frequencies of genes but of genotypes (without selection on zygotes and no random genetic drift), it is natural that studies of a structured population has been made with the inbreeding coefficient only, especially by Wright and Malecot. Wright (1943) introduced the concept of distance as the relation between the effective population number and the generation: the effective number of the neighborhood size is directly proportional to the generation number in the two dimensional model and to the square root of generation number in the one dimensional model. In his theory, the migration between parents and offspring is the most important function and normal migration was assumed. Although his results are very suggestive, many studies have shown the migration is leptokurtic, with mode near zero (Bateman, 1950; Cavalli, 1958; Skellam, 1951). It is difficult to estimate distance for mates born within the same parish, village, or other demographic unit, except by arbitrarily assigning them the value 0, as if the unit were a geometric point. Despite this, there is a tendency for migration to be leptokurtic.

The choice of migration function depended only on experience and mathematical convenience. However, the inbreeding coefficient as a function of distance at a stationary state is more intricate. Since leptokurtic relations between the inbreeding coefficient and distance between two individuals have been predicted by Malecot (1950) and the results shall be applied to the Brazilian population, a probabilistic approach to

obtain the inbreeding coefficient shall be briefly discussed here. Before going on to the inbreeding coefficient, distance will be defined as a measurable geographical length between birth places of mates instead of parent-offspring, because, (1) genetic implications of the inbreeding coefficient become more clear; for instance, the estimated inbreeding coefficient from mating type frequencies gives the inbreeding coefficient of offspring, and (2) estimation of distance is much easier in field work, in other words, only a single measure is necessary for a couple while many parent-offspring distances might be possible for a family and could not be independent to each other. In the following discussion, distance will be understood to be between birthplaces of mates, though this is not a necessary restriction.

Consider p isolates whose sizes are $N_1, N_2, \ldots, N_p$. Let l_{ki} be the probability that a person who was born at the k-th isolate and reproduced at the i-th isolate. This l_{ki} is called the coefficient of migration. The coefficient of consanguinity of two individuals at the generation n, denoted by $f_{ij}(n)$, who are taken at random, one from the i-th isolate and the other from the j-th isolate, is given by

$$f_{ij}(n) = \sum_k (1-u)^4 l_{ki} l_{kj} \frac{1}{N_k} \frac{1+f_{kk}(n-2)}{2} + \sum_k (1-u)^2 l_{ki} l_{kj} \left(1-\frac{1}{N_k}\right) f_{kk}(n-1) + \sum_{k \neq h} (1-u)^2 l_{ki} l_{hj} f_{kh}(n-1) \qquad (2.7.1.)$$

where u is mutation rate from the gene in question to its alleles and the subscripts i and j could be the same. In (2.7.1.), the first summation is the probability that two homologous genes came from the same individual in the previous generation, the second is the probability that two genes were from different individuals who lived in the same isolate, and the third is the probability that two genes were from two different isolates. If the population size N_k remains constant throughout generations, the coefficient of consanguinity becomes the inbreeding coefficient and at a stationary state ($f_{ij}(n) \to f_{ij}$ as $n \to \infty$)

$$(2.7.2) \qquad f_{ij} = \sum_{k,h}(1-2u)l_{ki}\left(f_{kh} + \frac{1-f_{kk}}{2N_k}\delta_{kh}\right)l_{hj}$$

where $\delta_{kh}=0$ if $k \neq h$ and $\delta_{kh}=1$ if $k=h$, and the higher powers of u are ignored.

The general solution of this system of linear equations has been given by Malecot (1950), using matrix algebra. Since we are only interested in solutions which can be applied on data, some simplifications are necessary. The most acceptable model as a first approach is such a homogeneous population structure that: (1) all isolates are the same size; $N_k=N$, (2) the coefficient of migration is invariant throughout generations, and (3) the inbreeding coefficient within an isolate is constant for all isolates; $f_{ii} = f_0$. Under the conditions, the following situations are useful:

Symmetrical migration in one dimension: Suppose that infinitely large number of isolates are on a line, with equal distance between each other (Figure 2.7.1). The coefficient of migration ($l_{ij}=l_{ji}$ in this

case) is only dependent on the absolute value $|j-i|$ which is the difference of indices of two corresponding isolates. From (2.7.2), after some calculations,

$$(2.7.3.) \qquad 2 + \frac{4N}{1-f_0} F(\lambda) = \frac{1}{1-(1-u)G(\lambda)} + \frac{1}{1+(1-u)G(\lambda)}$$

where $F(\lambda)$ and $G(\lambda)$ are moment generating functions of f_{ij} and l_{ij}, respectively. The mutation rate u is small so that the largest solution in absolute value of the equation $G(\lambda)=1/(1-u)\doteq 1+u$, λ_1, represents to a good approximation the decreasing rate of the inbreeding coefficient with distance (See Feller, 1957; pp. 257-259). If each isolate receives immigrants only from two neighbors (with proportion m/2) in each generation, then $G(\lambda) = 1-m + \frac{m}{2}(\lambda + \lambda^{-1})$ which is equal to $1+u$. Thus $\lambda_1 = 1 + \frac{u}{m} - \sqrt{\frac{2u}{m} + (\frac{u}{m})^2}$ so that, for small $|i-j|$, $4Nf_{ij}(1-f_0) \doteq c/\lambda_1^{|i-j|+1}$ where $c = \lambda_1/\sqrt{u(u+2m)}$. Therefore

$$(2.7.4.) \qquad f_{ij} = \frac{1-f_0}{4Nu(u+2m)} \left[1 + \frac{u}{m} - \sqrt{(\frac{u}{m})^2 + \frac{2u}{m}}\right]^{|i-j|}$$

For near zero distance, $f_{ii}=f_0$, we obtain $f_0 = 1/[1+4N\sqrt{u(u+2m)}]$. This result can be compared with Wright's result $f_0 = 1/[1+4N(u+m)]$ which corresponds to immigration from an infinitely large population. If d denotes the distance between birthplaces of mates, we have approximately

$$(2.7.4a.) \qquad f(d) = f(0)\exp(-\sqrt{\frac{2u}{m}}\, d) \qquad (f_0 = f(0))$$

which can be fitted with data by the least squares method.

If each isolate receives a proportion m of immigrants who come from not only the neighbors but also from some other groups, then λ_1 becomes approximately $1-\sqrt{2u/\sigma^2}$ where σ^2 is the migration variance. Furthermore, if a weak selection acts on the gene in question, the parameter for mutation rate, u, may be replaced by u+s(=U), where s is the selection coefficient. As a conclusion, in the symmetrical homogeneous migration in one dimension, the inbreeding coefficient decreases exponentially with rate constant $\sqrt{2a}$, where a is the ratio of systematic pressures to migration rate.

Symmetrical migration in two dimensions: Suppose that a population consists of an unlimited square net of isolates, each of which can be represented by two indices (Figure 2.7.2). It can be verified that (2.7.3.) holds in two dimensions, taking λ as a vector with two nuisance elements. The coefficients of migration and of inbreeding are also specified by two indices, p and q, as l(p,q) and f(p,q), respectively. As a special case, if each isolate exchanges only individuals with its four neighbors each generation, or mathematically,

$$l(1,0) = l(-1,0) = m/2,$$
$$l(0,1) = l(0,-1) = m'/2,$$
$$l(0,0) = 1-m-m', \qquad (M=m+m')$$
$$l(p,q) = 0 \text{ for otherwise,}$$

then a leptokurtic distribution of two dimensional variables, p and q, for the inbreeding coefficient with distance has been given by Malecot (1950) and Kimura and Weiss (1964). Fixing p=0, Malecot obtained that the marginal distribution approaches the exponential function in a one dimensional model when distance q is large, whereas Kimura et al. calculated

the distribution of $d = \sqrt{p^2 + q^2}$ which becomes asymptotically $\exp(-\sqrt{4U/M}\, d)/\sqrt{d}$ for large distance.

When we consider a continuous model, (2.7.2.) becomes

$$f(q,r) = \iint (1-2u) f(p,m) l(p,q) l(m,r) dS_p dS_m$$

(2.7.5.)

$$+ \int (1-2u) l(p,q) dS_p \frac{l(p,r)dS_p}{d(p)dS_p} \cdot \frac{1-f(p,p)}{2}$$

where $d(p)$ is density at neighborhood of point p, $l(p,q)dS_p$ is the probability that an individual was born in a unit area at the neighborhood of point p and reproduced at the neighborhood of point q, and $f(q,r)$ is the coefficient of consanguinity between two individuals, one taken randomly from q and the other from r. The size of isolate (N_k) is replaced by $d(p)dS_p$.

In the homogeneous model if density is constant ($d(p)=d$) throughout the population and the migration coefficient depends only on the distance between two points, (2.7.3.) holds in both dimensions. Particularly, when migration is normal in two dimensions, as Wright assumed, $f(p,q)$ becomes

$$f(p,q) = \frac{1-f_0}{8\pi\sigma_x\sigma_y d} \sum_{p=1}^{\infty} \frac{(1-u)^{2p}}{p} \exp\left\{-\left[\frac{(x_p-x_q)^2}{\sigma_x^2} + \frac{(y_p-y_q)^2}{\sigma_y^2}\right]/4p\right\}$$

which gives the inbreeding coefficient $f_0=f(p,p)$ in the neighborhood of p (putting $x_p=x_q$ and $y_p=y_q$) as $f_0=1/[1+8\pi\sigma_x\sigma_y d(-1/\ln 2u)]$ and, the asymptotic form of $f(p,q)$ for large distance becomes

$$f(p,q) \propto \exp(-\sqrt{2u}\,\frac{d}{\sigma})/\sqrt{d}$$

where $(d/\sigma)^2 = ((x_p-x_q)/\sigma)^2+((y_p-y_q)/\sigma_y)^2$ (Malecot, 1959). This result agrees with Kimura et al. who studied the correlation coefficient of gene frequencies between two isolates in one, two and three dimensional homogeneous stepping stone models without assuming any migration function. Their result may be summarized in a function:

$$f(d) \propto d^{-\frac{n-3}{2}-1} \exp\left(\sqrt{\frac{2u}{m}}\, d\right)$$

for large distance, where n is the number of dimension (n = 1, 2, and 3), and m is the average migration rate per coordinate ($m = \Sigma m_i/n$). Apparently, this agreement has not been recognized since Malecot himself stated in his discussion on the decrease of relationship with distance (discussion in Kimura (1955)): "So the coefficient of inbreeding f_0 is much influenced by the number of dimensions; on the contrary, the decrease with distance of the coefficient of relationship or of correlation is approximately the same, $e^{-\sqrt{\frac{2u}{\sigma^2}}\, x}$ in all cases;...".

Before closing theoretical discussion on the relationship between distance and the inbreeding coefficient, the significance of f_0 should be considered. The inbreeding coefficient f_0 itself is within an isolate, including contributions from self-fertilization, brother-sister

mating and so forth. Since no selfing occurs in man[1] and brother-sister marriages are prohibited by law, the expectation of f_0 decreases. The inbreeding function with distance becomes flat near zero if we remove contributions from close consanguinity (Morton and Yasuda, 1962). Therefore, there will be some error in estimating the systematic pressure and the migration rate from the observed f_0 and the estimated decreasing rate of inbreeding with distance in human populations, although Lamotte (1951) has succeeded to estimate both parameters in natural populations of Cepaea nemoralis, by this distance approach. Furthermore, the dependency of inbreeding function with distance on dimension is troublesome when interpreting data, since we do not know "dimension" with respect to human migration, though it has been suggested that it varies from 1 to 2 (cited in Kimura et al., 1964). As an extreme case, a population of organisms living along a river, coastal line or mountain ridge may be described by the one dimensional model and the two dimensional model may cover a population on a plane.

In summary leptokurtic relationships between the inbreeding coefficient and distance between birthplaces of mates has been predicted

---

[1] When we consider the distance between birthplaces of mates, d, d=0 might be observed in such a case that after a boy was born in place P, his family moved to a different place Q; and another family moved into the place P and gave birth to a girl. If the boy and girl become a couple, then d=0. Although we expect such couples to be rare, no information is available at present.

theoretically, exponential in one dimension and more rapidly decreasing than exponential in two dimensions. However, the decreasing constant is invariant regardless of dimension of models.

2.8. Discussion and problems in ascertaining the inbreeding coefficient

The most common way to ascertain the inbreeding coefficient in human population is to classify marriages into known degrees of inbreeding and to take their average weighted by the relative frequencies of observed numbers. This method is called pedigree analysis, requires a complete knowledge of pedigrees, and assumes the nominal coefficient of consanguinity is equal to the inbreeding coefficient. Tracing generation paths and making loop(s) through common ancestor(s) the inbreeding coefficient for a particular marriage would be estimated by Wright's formula:

$$F = \Sigma \left(\frac{1}{2}\right)^{m+f} \frac{1+F'}{2}, \qquad \text{Wright (1921)}$$

where f and m are number of generations from father and mother respectively to the common ancestor whose inbreeding coefficient is F' (usually, we assumed that F'=0) and summation is taken over all possible loops. For instance, F=0 for non-consanguineous marriages, F=1/8 for uncle-niece, double first cousin,..., F=1/16 for first cousin, F=1/32 for second cousin, and so on. Some devices have been made for unusually complex pedigrees (Wright and Mcphee, 1925; Kudo, 1962). The average inbreeding coefficient is obtained by $\alpha = \Sigma c_i F_i$, where c_i is the corresponding proportion of marriage. However, this does not cover unrecognized remote consanguinity. For instance, under favorable circumstances

the ascertainment of consanguinity can extend several generation into the past. In some area, records of Roman Catholic marriage dispensations go back hundreds of years (Moroni, 1962). Under these conditions ascertained consanguinity is likely to account for a large fraction of the total inbreeding coefficient. Formally we may represent the situation as

(2.8.1.) $$\alpha_t = \alpha + \alpha_r,$$

where α_t, α and α_r denote the inbreeding coefficient due to total consanguinity, ascertained consanguinity and undetected, remote consanguinity, respectively. Unfortunately, as we go backward in time the proportion of ancestors who were migrants increases, so that ascertainment of consanguinity will always be incomplete even for populations with extensive marriage records.

Although we may hope that α_r/α is small, doubt arises even in the most favorable cases. For example, birth records in the Alpine village of Bosco-Gurin permit reconstruction of pedigrees for ten generations (Moor-Jankowski and Huser, 1957). There was little migration into the village. We might expect that all important consanguinity had been ascertained. But in fact, history shows that the villagers migrated into the area in the thirteenth and fourteenth centuries from the Valais. It is likely that inbreeding during the ages before the birth records began had effects on gene frequencies which are still appreciable and contribute to the α_r of Switzerland.

Since no system of records, however complete, can ascertain the total inbreeding coefficient, we must look for other ways. There are two approaches to pursue the remote inbreeding coefficient: use of a

biological indicator (bioassay) and of migration functions. In both methods, the remote inbreeding coefficient is calculated as the difference of the total inbreeding and the close inbreeding ascertained by pedigree analysis. In this connection, we define remote consanguinity as relationship more distant than first cousins once removed ($F < 1/32$).

The inbreeding coefficient can be estimated from phenotype and mating type frequencies. Differential selection, illegitimacy, and misclassification are the main sources to disturb an accurate estimate of the inbreeding coefficient, and, generally speaking, they affect phenotype frequencies more than mating type frequencies (see 2.5.). Differential selection, especially against homozygotes, might tend to give smaller, or even a negative estimate of the inbreeding coefficient. Illegitimacy or misclassification has in a statistical sense the same effects on the biological indicator as selection does. And genes whose frequencies are relatively small are excluded from the probability models for mating types, and should be pooled with more common alleles to meet the restriction $p > F$.

Sanghvi (1955) and Schull (1965) pointed out the insensitiveness of phenotype frequencies to estimate the inbreeding coefficient. Regardless of these difficulties which will be shown mathematically in the next chapter, there is no such trouble in estimating the inbreeding coefficient from mating type data. The statistical properties of the biological indicator will be given in the succeeding section.

Use of a migration function, $m(x)$, defined as the probability among all marriages that the marital distance is x, requires determination of a migration function and evaluation of the genetic correlation coefficient, $f(x)$, of children whose parents had a marital distance x. If these

two functions with distance are found, the total mean inbreeding coefficient is calculated by

$$\alpha = \int_0^\infty f(x)m(x)dx.$$

Since nearly all the information comes from rare homozygotes, whether there is dominance or not and whether they are ascertained prospectively, as a random sample of the population, or retrospectively as probands for a rare homozygous condition, the conditional probability that the parents of a rare homozygote had marital distance x leads to powerful and informative results (Morton and Yasuda, 1962). As mentioned in 2.7., human migration does not follow a normal distribution expected for dispersion of genes by a diffusion process (Cavalli, 1958). This is not surprising because many of the barriers which generate isolates in human population. Thus at present a choice of migration function is not completely specified except (1) the function is leptokurtic, (2) the proportion of near zero distance should be finite: $m(0) < +\infty$, and (3) the function has better be a mathematically and statistically simple form. Under these conditions, the suggestive functions are exponential, square root exponential (Cavalli, 1958), log-normal, beta, double exponential and so forth. A gamma function that includes an exponential distribution as a special case (n=1) has been fitted to a northern Italian population (Cavalli, 1962). The fit is good but the estimate of the distribution parameter, n, are always less than one so that m(0) tends to be infinite. This is unrealistic. It is expected, however, that no distribution would fit well because of a practical difficulty to estimate near zero distance frequency.

The genetic correlation with distance, f(x), is more intricate. This can be derived if a migration function is known (Malecot, 1948; Kimura, 1963), but it seems that assumption of migration function is not necessary (Malecot, 1950; Kimura et al., 1964). A difficulty in practice here is the fact that f(x) depends on the dimension of human migration. Fortunately f(x) can be determined empirically as a genetic correlation,

$$\frac{\Sigma(p_a-p)(p_{x+a}-p)}{\Sigma(p_a-p)^2} \qquad \text{(Malecot, 1955)}$$

where the summation taking over locations a where gene frequency is p_a.

The separation of the total inbreeding coefficient into contributions due to ascertained and remote consanguinity involves an important concept of population structure. Wahlund's principle tells us that if random mating is assumed within isolates, the inbreeding coefficient due to barriers is always positive since the coefficient is defined with respect to gene frequency variance. The more barriers there are, a higher value of the inbreeding coefficient is expected. However, all barriers would not be ascertained in practice. If F_i designates the ascertained inbreeding coefficient by the i-th degree procedure, for instance, the first degree may be due to ascertainment of close consanguinity less distant than second cousin, the second degree up to known consanguinity and so forth, then the total inbreeding coefficient α_t can be obtained from

$$\alpha_t = \Sigma F_i .$$

However, the assumption of random mating within isolates may not be

justified in particular situations. For example, suppose an isolate consists of two types of homozygotes, AA and aa, and mating occurs only between different genotypes. Obviously, the inbreeding coefficient for the isolate is not zero but minus one in the sense of a negative correlation between uniting gametes. This leads to

$$\alpha_t = \Sigma F_i + r, \qquad (2.8.2.)$$

where r is the correlation coefficient due to non-random mating in isolates and the following relation holds:

$$-1 \leq r \leq 0 \leq \alpha_t \leq \Sigma F_i \leq 1.$$

In other words, all positive correlations of uniting gametes are considered due to genetic barriers which might have been generated by random genetic drift, geographical, sociological and other factors. In practice, however, the ascertainment of ΣF_i is dependent on technique so that

$$\alpha_t = \underset{A}{\Sigma} F_i + ({}_{T-A}\Sigma F_i + r),$$

where $\underset{A}{\Sigma}$ and ${}_{T-A}\Sigma$ mean summations of ascertained inbreeding coefficient with respect to the degree of procedure and of unascertained positive correlation between uniting gametes, respectively. This is equivalent to (2.8.1.) if we put $\alpha = \underset{A}{\Sigma} F_i$ for the ascertained inbreeding and $\alpha_r = ({}_{T-A}\Sigma F_i + r)$ for the remote consanguinity. α_r can be negative if the negative correlation in isolates is high.

An alternative model has been proposed by Wright (1943) for consideration of breeds of cattle. If a population has hierarchic structure, the total inbreeding coefficient, F_{IT}, is related to the inbreeding

coefficient within subpopulation, F_{IS}, and due to subdivision, F_{ST}, in the following manner:

(2.8.3.) $$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$$

which can be extended into any degree of hierarchic structure, or $1 - F_{IT} = \prod (1-F_{S_i})$. This relation can be deduced from the moment theory as a special case. Suppose that a population consists of isolates whose size and a gene frequency are w_{ij} ($\Sigma\Sigma_{ij} w_{ij}=1$) and p_{ij}, respectively, and within which mating is at random. Then the total frequency of a homozygote in the population is $\Sigma\Sigma_{ij} p_{ij}^2 w_{ij} = p^2 + p(1-p)F_{IT}$ where $p = \Sigma\Sigma_{ij} p_{ij} w_{ij}$. On the other hand, when we consider barriers with respect to i, within the i-th aggregate of isolates, the homozygote frequency then is $\Sigma_j p_{ij}^2 w_{ij} = [p_i^2 + p_i(1-p_i)F_i]w_i$, where $p_i = \Sigma_j p_{ij} w_{ij}/w_i$, $w_i = \Sigma_j w_{ij}$ and F_i is the inbreeding coefficient of the i-th aggregate. The total homozygote frequency is therefore $\Sigma_i [p_i^2 + p_i(1-p_i)F_i]w_i = \Sigma p_i^2 w_i + \Sigma_i p_i(1-p_i)F_i w_i$, where $\Sigma p_i^2 w_i = p^2 + p(1-p)F_{ST}$, so that $p^2 + p(1-p)F_{IT} = p^2 + p(1-p)F_{ST} + \Sigma_i p_i(1-p_i)F_i w_i$ or

(2.8.4.) $$F_{IT} = F_{ST} + \frac{\Sigma F_i p_i (1-p_i) w_i}{p(1-p)} .$$

If the inbreeding coefficients for all aggregates are same: $F_i = F_{IS}$ for all i, we obtain $F_{IT} = F_{ST} + F_{IS}(1-F_{ST})$ which is equivalent to (2.8.3.). The result can be proved with respect to heterozygote frequency.

It is obvious that an hierarchic pattern of barriers is specified in (2.8.3.), whereas no such scheme is made in (2.8.2.). A genetic

barrier is hard to recognize in human population, while it is rather easy to set up such a pattern in experimental populations like cattle. F_{ST} and F_{IS} should be always positive with respect to genetic barriers and from a probablistic point of view, but Wright (1951) stated that F_{IS} could be negative. This is true only when mating is not random within basic units of population or isolates. In this situation F_{IS} corresponds to r, and whenever the F-value becomes negative, the independency between system of mating and gene frequency breaks since any homozygote frequency cannot be less than zero. Therefore r must be near zero in human populations. This implies that the gene has potentially an equal probability to unite with the neighbors in the sense of probability density. On the other hand, the identification of isolates is almost impossible in man without knowledges of "original composition of population". At present, we do not have any method to estimate r but r=0 for human population. Further research is desirable.

The hierarchic description is a good approximation of population structure. Taking the logarithm of a general form of (2.8.3.), and expanding in series, we obtain approximately

$$F_{IT} = \Sigma F_{S_i} + F_{IS},$$

where F_{IS} is zero if mating in isolates is at random, otherwise

$$-1 \leq F_{IS} < 0.$$

2.9. Summary

A new theory for describing human population structure has been proposed by replacing the concept of isolate size in Dahlberg's sense or neighborhood size in Wright's sense by an idea of probability density for genes or a tendency that a gene shall combine with the neighbors in order to form genotype, mating type, and other gene combinations. These genetic quantities can be described in terms of moments of population whose order corresponds to a number of genes combined. The main results when the inbreeding coefficient is not greater than the smallest gene frequency are that: (1) mating type frequencies are given as a function of gene frequencies and the inbreeding coefficient at autosomal and sex-linked loci; (2) the effect of inbreeding or subdivision of population on mating type frequencies and on segregation analysis has been examined and no serious effects is found; and (3) the relation between the inbreeding coefficient and marital distance describes genetic isolation in populations is leptokurtic, including an exponential function.

A method to estimate the total, ascertained and remote inbreeding coefficient has been derived. The statistical procedure will be given in the next chapter. Components of three kinds of the inbreeding coefficient were considered. Two components in describing system of mating in terms of correlation coefficient between uniting gametes should be distinguished: positive and negative correlations. All of the positive correlations may be described with genetic barriers, while the negative correlation is observed if the basic unit of population, or isolate, cannot be considered as randomly mating group. The ascertained inbreeding coefficient consists of positive correlations and the remote inbreeding

may include both correlations. Comparison with Wright's hierarchic structure of population is also made.

3. Statistical Procedures

3.1. Introduction

This chapter devotes itself to develop statistical methods for utilization of theories which have been discussed in the preceding chapter. Since Fisher's maximum likelihood method (Fisher, 1922 and later) will be employed frequently, it is worthwhile to look at one of its developments, the scoring method that is powerful when likelihood equations are too complicated to obtain analytic solutions for parameters. There are good summaries on this method (Rao, 1952; Morton, 1959; Bailey, 1961). To visualize, the description below is for a single parameter, but its generality is not lost when the number of parameters is arbitrary.

Suppose that L denotes a likelihood with a single parameter θ and we define the score of θ as the first derivative of lnL with respect to θ; i.e., $u_\theta = \partial \ln L/\partial\theta$. The amount of information for θ is $k_\theta = E(u_\theta^2) = -E(\partial^2 \ln L/\partial\theta^2)$, where E is an operational notation to take expectation. With independent samples the scores and the amount of informations are additive so that the total score U_θ and the information K_θ are obtained as

$$U_\theta = \Sigma u_\theta \text{ and } K_\theta = \Sigma k_\theta,$$

where summation is over all independent sampling units. Applying Taylor's series expansion to the likelihood with a tentative value θ_0, the improved estimate θ_1 will be

$$\theta_1 = \theta_0 + U_{\theta_0}/K_{\theta_0}$$

and its variance $\sigma^2 = 1/K_{\theta_1}$. The discrepancy between θ_0 and θ_1 are tested by $\chi^2 = (\theta_1 - \theta_0)^2 K_\theta$ with one degree of freedom. If χ^2 indicates

a significant difference, then we may repeat the above process until no significance appears. The heterogeneity test between units is carried out by $\chi^2 = \Sigma u^2/k - U^2/K$ with degree of freedom being number of independent units minus one. Nature of convergency in iteration will be discussed in Appendix 4.

3.2. The ascertainment of inbreeding coefficient by pedigree study

In the general population, the inbreeding coefficient f_i has frequencies $w_i (\Sigma w_i = 1)$, mean $\alpha = \Sigma f_i w_i$ and variance $\sigma^2 = \Sigma f_i^2 w_i - \alpha^2$. The unbiased estimates are

$$\alpha = \Sigma f_i n_i / n,$$

and

$$s^2 = [\Sigma f_i^2 n_i - n\alpha^2]/(n-1),$$

where n and n_i are the total number of individuals studied and the number of individuals whose inbreeding coefficient is f_i, respectively. Thus the variance of α is obtained by

$$\mathrm{Var}(\alpha) = \frac{s^2}{n}.$$

The method is heavily dependent upon information about pedigrees.

3.3. Bioassay of the inbreeding coefficient and gene frequencies

Since the inbreeding coefficient and gene frequencies are fundamental quantities to describe human population structure, a statistically powerful and biologically meaningful method is required for estimating both parameters. At hand, two types of models and data are available for this purpose: individual phenotype frequencies and mating type frequencies.

In the following we shall use the inbreeding coefficient α to denote either α_t or α_r, according as known consanguineous marriages are omitted or not. In both cases, the first task is to test a null hypothesis that $\alpha=0$, or mating is panmictic. Under the assumption, gene frequencies are estimated by the maximum likelihood method so that an estimate of the inbreeding coefficient in the neighborhood of zero can be obtained by iteration process with respect to α. To visualize the method, mathematical descriptions for some simple cases are presented in what follows.

3.3.1. Individual phenotype frequencies

Case 1. Two alleles without dominance: Let a,b,c be the observed numbers of genotypes AA, Aa and aa in a random sample from the general population, and p be the frequency of A. Assuming that the individuals in the sample are unrelated, that mating is panmictic but for an inbred component α, and that genotypes are enumerated before differential selection has acted, we have

$$P(\underline{AA}) = p^2 + p(1-p)\alpha$$
$$2P(\underline{Aa}) = 2p(1-p)(1-\alpha)$$
$$P(\underline{aa}) = (1-p)^2 + p(1-p)\alpha$$

The log likelihood is

$$L = (a+b)\ln p + (b+c)\ln(1-p) + a\ln(p+\alpha-\alpha p) + b\ln(1-\alpha) + c\ln(1-p+\alpha p)$$

and the maximum likelihood scores are

$$U_p = (a+b)\left[\frac{1}{p}\right] + a\left[\frac{1-\alpha}{p+\alpha-\alpha p}\right] - (b+c)\left[\frac{1}{1-p}\right] - b\left[\frac{1-\alpha}{1-p+\alpha p}\right]$$

$$U_\alpha = a\left[\frac{1-\alpha}{p+\alpha-\alpha p}\right] - b\left[\frac{1}{1-\alpha}\right] + c\left[\frac{p}{1-p+\alpha p}\right]$$

The solutions are

$$p = (2a+b)/2(a+b+c) \qquad \alpha = (4ac-b^2)/(2a+b)(2c+b).$$

The variances of the scores are

$$K_{pp} = a\left[\frac{1}{p} + \frac{1-\alpha}{p+\alpha-\alpha p}\right]^2 + b\left[\frac{1-2p}{p(1-p)}\right]^2 + c\left[\frac{1}{1-p} + \frac{1-\alpha}{1-p+\alpha p}\right]^2$$

$$K_{p\alpha} = a\left[\frac{1}{p} + \frac{1-\alpha}{p+\alpha-\alpha p}\right]\left[\frac{1-p}{p+\alpha-\alpha p}\right] - b\left[\frac{1-2p}{p(1-p)(1-\alpha)}\right] - c\left[\frac{1}{1-p} + \frac{1-\alpha}{1-p+\alpha p}\right] \cdot$$

$$\left[\frac{p}{1-p+\alpha p}\right]$$

$$K_{\alpha\alpha} = a\left[\frac{1-p}{p+\alpha-\alpha p}\right]^2 + b\left[\frac{1}{1-\alpha}\right]^2 + c\left[\frac{p}{1-p+\alpha p}\right]^2$$

To test the null hypothesis that $\alpha=0$ we may evaluate the scores and their variances when $\alpha=0$. Then $K_{p\alpha}=0$, and

$$U_\alpha = (a+b+c)(4ac-b^2)/(2a+c)(2c+b)$$

$$K_{\alpha\alpha} = a+b+c$$

Note that $U_\alpha/K_{\alpha\alpha}$ is exactly the maximum likelihood estimate of α. The variance of this estimate in the neighborhood of the null hypothesis is $1/K_{\alpha\alpha}$.

When there are k alleles, the amount of information about α under the null hypothesis that $\alpha=0$ is given by $(k-1)N$, where N is the total number of individuals (Appendix 5).

Case 2. Two alleles with complete dominance: Dominance creates the difficulty that, with only two phenotypes, it is impossible to estimate p and α simultaneously from phenotype frequencies alone. However, if there is other information about either parameter, for example from segregation analysis (Morton, 1959) or marital distance, the information from this can be combined to yield both estimates.

Suppose that two phenotypes are aa, with frequency $p(p+\alpha-\alpha p)$, and A- with frequency $(1-p)(1+p-\alpha p)$. Let the observed numbers be a and b, respectively. Then the log likelihood is

$$L = a\ln p + b\ln(1-p) + a\ln(p+\alpha-\alpha p) + b\ln(1+p-\alpha p)$$

and the scores for p and α are

$$U_p = a\left[\frac{1}{p} + \frac{1-\alpha}{p+\alpha-\alpha p}\right] - b\left[\frac{1}{1-p} - \frac{1-\alpha}{1+p-\alpha p}\right]$$

$$U_\alpha = a\left[\frac{1-p}{p+\alpha-\alpha p}\right] - b\left[\frac{p}{1+p-\alpha p}\right].$$

The variances of the scores are

$$K_{pp} = a\left[\frac{1}{p} + \frac{1-\alpha}{p+\alpha-\alpha p}\right]^2 + b\left[\frac{1}{1-p} - \frac{1-\alpha}{1+p-\alpha p}\right]^2$$

$$K_{p\alpha} = a\left[\frac{1}{p} + \frac{1-\alpha}{p+\alpha-\alpha p}\right]\left[\frac{1-p}{p+\alpha-\alpha p}\right] + b\left[\frac{1}{1-p} - \frac{1-\alpha}{1+p-\alpha p}\right]\left[\frac{p}{1+p-\alpha p}\right]$$

$$K_{\alpha\alpha} = a\left[\frac{1-p}{p+\alpha-\alpha p}\right]^2 + b\left[\frac{p}{1+p-\alpha p}\right]^2$$

When $\alpha=0$, we obtain

$$U_p = a\left[\frac{2}{p}\right] - b\left[\frac{2p}{1-p^2}\right], \qquad U_\alpha = a\left[\frac{1-p}{p}\right] - b\left[\frac{p}{1+p}\right]$$

$$K_{pp} = 4(a+b)/(1-p^2)$$

$$K_{\alpha\alpha} = (a+b)(1-p)/(1+p)$$

$$K_{p\alpha} = 2(a+b)/(1+p).$$

Note that the K-matrix is singular ($K_{pp}K_{\alpha\alpha} = K_{p\alpha}^2$), as expected, and U-scores are not linearly independent ($U_\alpha = \frac{1-p}{2}U_p$), indicating that p and α cannot be estimated simultaneously from this material alone.

Case 3. ABO and MNSsU blood group systems: More complications due to dominance are discovered at the ABO locus and effects of linkage or segregant factor pairs in the same system are also found in MNSsU system.

With anti-A and -B, we are able to classify human population into four phenotypic groups: O, A, B and AB. Bernstein (1925, 1930) established that three genes A, B, and O at a single locus were responsible for the phenotypes and gave conventional formulae to evaluate gene frequencies which were biased estimates from individual sample of random population. Later, Stevens (1938) proved that Bernstein's formulae did not exactly satisfy the maximum likelihood equations and $p+q+r \neq 1$, where p,q,r were frequencies of gene A, B, O, respectively.

In populations with inbred proportion α, the phenotype frequencies are given by

$$P(O) = r^2 + r(1-r)\alpha$$

$$P(A) = p^2 + 2pr + [p(1-p) - 2pr]\alpha$$

$$P(B) = q^2 + 2qr + [q(1-q) - 2qr]\alpha$$

$$P(AB) = 2pq - 2pq\alpha$$

The log likelihood is

$$L = O\cdot\ln[r^2+r(1-r)\alpha] + A\cdot\ln[p^2+2pr + \{p(1-p)-2pr\}\alpha] + B\cdot\ln[q^2+2qr + \{q(1-q)-2qr\}\alpha] + AB\cdot\ln(2pq-2pq\alpha)$$

and the maximum likelihood scores under the null hypothesis that $\alpha=0$ are

$$U_p = O\left[-\frac{2}{r}\right] + A\left[\frac{2r}{p^2+2pr}\right] + B\left[\frac{-2}{q+2r}\right] + AB\left[\frac{1}{p}\right]$$

$$U_q = O\left[-\frac{2}{r}\right] + A\left[\frac{-2}{p+2r}\right] + B\left[\frac{2r}{q^2+2qr}\right] + AB\left[\frac{1}{q}\right]$$

$$U_\alpha = O\left[\frac{1-r}{r}\right] + A\left[\frac{q-r}{p+2r}\right] + B\left[\frac{p-r}{q+2r}\right] + AB[-1].$$

It is easily verified that

$$U_\alpha = -\frac{p}{2}U_p - \frac{q}{2}U_q,$$

indicating that no information about α is yielded from this material alone. This is surprising result from viewpoints of statistics and genetics. Statistically, we could expect to estimate simultaneously three independent parameters (p, q and α) with three degrees of freedom (since we have four phenotypes). It has been observed that simultaneous estimates of gene frequencies and the inbreeding coefficient in the ABO system are very unstable (Schull, 1965). We have just proved that no estimate for

the inbreeding coefficient can be obtained from the ABO locus by this method, Schull's estimates being based entirely on rounding error in estimating a singular matrix. This is not improved by subtyping of A into A_1 and A_2 with anti-A_1 and -A sera. U-scores in this case have a relation

$$U_\alpha = -\frac{p_1}{2} U_{p_1} - \frac{p_2}{2} U_{p_2} - \frac{q}{2} U_q,$$

where p_1, p_2 and U_{p_1}, U_{p_2} are frequencies of A_1, A_2 genes and their U-scores, respectively (Appendix 6). Although we now have 5 degrees of freedom (six phenotypes), yet we can neither estimate simultaneously four independent parameters (p_1, p_2, q and α), nor test the null hypothesis that $\alpha=0$.

From the standpoint of genetics, no reliable information about the inbreeding coefficient would be expected even if one and only one O gene exists in the ABO locus,* whereas the complete absence of O genes generates a codominant system with A and B genes and, as we know, that system gives information about α as much as the total number of observed samples (case 1.). If either A or B gene is absent or both genes are pooled, it becomes the system with two alleles with complete dominance and no information about α is expected (case 2.).

This discovery at the ABO locus is very discouraging for studies in human population structure because hundreds of thousands of observations

---

* This situation may be observed in the Duffy system with anti-Fy^a and anti-Fy^b where Fy gene is rare in Caucasians but is common in Negroes.

on ABO blood groups has been reported from around the world (ex. Mourant et al., 1958). However, if we can specify numerically the location of population studied, for example, with latitude and longitude, then the empirical correlation method with distance may be applied to the ABO data.

An interesting relationship in MNSsU system between the ABO-type dominance and segregating factor pairs in the same system should be mentioned here. At the MNSsU blood group system, the factors M and N forms a codominant system while the factor S, s and * (= U+u that is observed in a phenotype S(-)s(-)U(?) (Morton et al., 1965)) have the ABO type dominance relation. A question is what effect of the S-series would be observed on the amount of information about the inbreeding coefficient, compared to the information from the MN series alone. In the analysis of the northeastern Brazilian population by maximum likelihood scoring method, detail of which will be found in Appendix 7, the amount of information about α with the MN factors alone is I_{MN} = 2128 units, whereas it becomes I_{MNSsU} = 4536 units with all factors, or I_{MNSsU}/I_{MN} = 2.13. It is therefore concluded that no improvement in the information about the inbreeding coefficient is expected by subtyping at ABO-like systems, while segregating factor pairs in the same system increase the information, compared to a codominant system.

3.3.2. Mating type frequencies

Mating type frequencies give more reliable estimates of the inbreeding coefficient and no difficulty due to dominance is observed. Generally speaking, mating type frequencies yield more information about the inbreeding coefficient while the estimated gene frequencies and

their variances from mating type are exactly same as from individual phenotype frequencies.

Case 1. Autosomal locus with two codominant alleles: Suppose that the observed numbers of mating types AA x AA, AA x Aa, Aa x Aa, AA x aa, Aa x aa and aa x aa are n_1, n_2, n_3, n_4, n_5, and n_6, respectively, where A and a are alleles with frequencies 1-q and q at a given locus. Assuming that the couples in the sample are at random, and that the couples are enumerated before differential selection has acted, the mating type frequencies in population whose inbred component is α are given by:

$$P_1 = (1-q)^4 + 6q(1-q)^3\alpha$$
$$P_2 = 4q(1-q)^3 + 12q(1-q)^2(2q-1)\alpha$$
$$P_3 = 4q^2(1-q)^2 + 4q(1-q)[1-6q(1-q)]\alpha$$
$$P_4 = 2q^2(1-q)^2 + 2q(1-q)[1-6q(1-q)]\alpha$$
$$P_5 = 4q^3(1-q) + 12q^2(1-q)(1-2q)\alpha$$
$$P_6 = q^4 + 6q^3(1-q)\alpha.$$

The log likelihood is

$$L = \sum_{i=1}^{6} n_i \ln P_i$$

and the maximum likelihood scores are

$$U_q = \sum_{i=1}^{6} \frac{n_i}{P_i}\left(\frac{\partial P_i}{\partial q}\right) \qquad \text{and} \qquad U_\alpha = \sum_{i=1}^{6} \frac{n_i}{P_i}\left(\frac{\partial P_i}{\partial \alpha}\right)$$

The variances of scores are

$$K_{qq} = \sum_{i=1}^{6} \frac{n_i}{P_i}\left(\frac{\partial P_i}{\partial q}\right)^2 \quad K_{q\alpha} = \sum_{i=1}^{6} \frac{n_i}{P_i}\left(\frac{\partial P_i}{\partial q}\right)\left(\frac{\partial P_i}{\partial \alpha}\right) \quad K_{\alpha\alpha} = \sum_{i=1}^{6} \frac{n_i}{P_i}\left(\frac{\partial P_i}{\partial \alpha}\right)^2 .$$

Under the null hypothesis that $\alpha=0$, we may evaluate the scores and their variances when $\alpha=0$. Then

$$U_q = \frac{1}{q(1-q)}\left[(2c+b)-2(a+b+c)q\right] = 0$$

$$U_\alpha = \frac{3(a+b+c)}{(2a+b)(2c+b)}\left[(4ac-b^2) + \frac{4}{3}(a+b+c)(n_3-2n_4)\right]$$

and

$$K_{qq} = \frac{2(a+b+c)}{q(1-q)}, \quad K_{q\alpha} = 0, \quad K_{\alpha\alpha} = 3(a+b+c),$$

where a $(=2n_1+n_2+n_4)$, b $(=n_2+2n_3+n_5)$ and c $(=n_4+n_5+2n_6)$ correspond to the observed numbers of genotype AA, Aa, and aa, respectively. The results are remarkable: no improvement is obtained in information about gene frequencies, whereas the amount of information about the inbreeding coefficient is three times as great from mating type frequencies as from individual phenotypic data, indicating that mating types are yielding much information on population structure.

Case 2. Autosomal locus with two alleles with complete dominance: Let n_1, n_2, n_3 be the observed numbers of mating types A- x A-, A- x aa, aa x aa in a random sample from population and q be the frequency of the recessive gene a. Assuming that the couples in the sample are unrelated except for an inbred component α, and that the couples are enumerated before differential selection has acted, the mating type frequencies are

given by:

$$P_1 = (1-q^2)^2 - 2q(1-q)(1-3q^2)\alpha$$
$$P_2 = 2q^2(1-q^2) + 2q(1-q)(1-6q^2)\alpha$$
$$P_3 = q^4 + 6q^3(1-q)\alpha$$

The log likelihood is

$$L = \sum_{i=1}^{3} n_i \ln P_i$$

and the maximum likelihood scores are

$$U_q = \sum_{i=1}^{3} n_i u_{qi}, \qquad U_\alpha = \sum_{i=1}^{3} n_i u_{\alpha i},$$

where $u_{qi} = \frac{1}{P_i}\left(\frac{\partial P_i}{\partial q}\right)$ and $u_{\alpha i} = \frac{1}{P_i}\left(\frac{\partial P_i}{\partial \alpha}\right)$. The variances of the scores are

$$K_{qq} = \sum_{i=1}^{3} n_i u_{qi}^2$$

$$K_{q\alpha} = \sum_{i=1}^{3} n_i u_{qi} u_{\alpha i}$$

$$K_{\alpha\alpha} = \sum_{i=1}^{3} n_i u_{\alpha i}^2.$$

To test the null hypothesis that $\alpha=0$, we may evaluate the scores and their variances when $\alpha=0$. Then

$$U_q = \frac{2}{q(1-q^2)}[b-(a+b)q^2] = 0 \qquad U_\alpha = \frac{8b}{(1+q)^2}$$

and the information matrix is given by

$$K = \begin{bmatrix} K_{qq} & K_{q\alpha} \\ K_{q\alpha} & K_{\alpha\alpha} \end{bmatrix}$$

where

$$K_{qq} = \frac{4(a+b)}{1-q^2}, \quad K_{q\alpha} = \frac{2(a+b)}{1+q}, \quad K_{\alpha\alpha} = \frac{(a+b)(1+7q^2)}{(1+q)^2}.$$

($a=2n_1+n_2$, $b=n_2+2n_3$ are the observed number of A- and aa individuals, respectively.)

The covariance matrix is then

$$K^{-1} = \begin{bmatrix} K^{qq} & K^{q\alpha} \\ K^{q\alpha} & K^{\alpha\alpha} \end{bmatrix}$$

where

$$K^{qq} = \frac{(1+7q^2)(1-q^2)}{32(a+b)q^2}, \quad K^{q\alpha} = -\frac{(1+q)^2(1-q)}{16(a+b)q^2}, \quad K^{\alpha\alpha} = \frac{(1+q)^2}{8(a+b)q^2}.$$

The amount of information about α will be

$$K_\alpha = \frac{1}{K^{\alpha\alpha}} = 8(a+b)\left(\frac{q}{1+q}\right)^2$$

Again, no improvement is observed in estimating gene frequency. However, it is possible with mating type frequencies to test the hypothesis that $\alpha=0$ and then to estimate the inbreeding coefficient. It is of interest that the amount of information about α is proportional to the expected frequency of recessive children from dominant x dominant type matings in randomly mating population.

Case 3. Sex-linked locus: Since we have little interest in the inbreeding coefficient α at the sex-linked locus, no detailed discussion will be given here, but results are listed in the Table 3.1.

Verification of formulae is immediate.

3.3.3. Factor union algebra

With increasing numbers of alleles, it becomes more complicated to obtain analytic form of maximum likelihood scores and their variances. For estimation of gene frequency only, assuming random mating, the method of gene counting (Ceppelini et al., 1955; Smith, 1957) may be useful provided that individual phenotypic frequency data are available. For our purpose to estimate simultaneously gene frequencies and the inbreeding coefficient, however, the counting method is not satisfactory. Before going to develop a new method that is the most economical with electronic computer, we will introduce a concept of factor union algebra for grouping of genotypes whose phenotype is the same.

Let us take the A_1A_2BO blood group system for explanation of factor union algebra. There are four main alleles $\underline{A}_1$, $\underline{A}_2$, $\underline{B}$ and $\underline{O}$ at this locus and $\underline{A}_1$ is dominant to $\underline{A}_2$ and $\underline{O}$; $\underline{A}_2$ to $\underline{O}$; $\underline{B}$ to $\underline{O}$; but $\underline{B}$ is co-dominant with $\underline{A}_1$ and $\underline{A}_2$. These alleles are detected by reactions with corresponding sera: $\underline{A}_1$ by anti-A_1 and -A; $\underline{A}_2$ by anti-A; $\underline{B}$ by anti-B; and $\underline{O}$ by no agglutination with either sera. If we assign $\underline{1}$ for positive reaction and $\underline{O}$ for negative (this number will be called a factor), then alleles at the locus are characterized by an array of factors which is named as gene vector:

$$\underline{A}_1 = (1, 1, 0)$$
$$\underline{A}_2 = (0, 1, 0)$$
$$\underline{B} = (0, 0, 1)$$
$$\underline{O} = (0, 0, 0),$$

Where the order of factors is conventional and is anti-A_1, -A, and -B in this case.

Since man is diploid so that two genes are responsible for phenotype of individual, combination of gene vectors should be performed with logical union of factors or factor union algebra (Cotterman, 1965):

$$0 + 0 = 0$$

$$0 + 1 = 1$$

$$1 + 0 = 1$$

$$1 + 1 = 1.$$

The additional operator, "+", is actually union, one of the binary operations in Boolean algebra (Birkhoff and MacLane, 1965). The genotype vectors are then generated as

$$\begin{aligned} A_1A_1 &= (1, 1, 0) + (1, 1, 0) \\ &= (1+1, 1+1, 0+0) \\ &= (1, 1, 0) \\ A_1A_2 &= (1, 1, 0), \\ &\cdots\cdots\cdots\cdots \\ OO &= (0, 0, 0). \end{aligned}$$

Six of ten genotype vectors have different phenotype vectors:

phenotype vector	genotype
(1, 1, 0)	A_1A_1, A_1A_2, A_1O
(0, 1, 0)	A_2A_2, A_2O
(0, 0, 1)	BB, BO
(1, 1, 1)	A_1B
(0, 1, 1)	A_2B
(0, 0, 0)	OO.

It is therefore clear that the probability and score of genotypes can be summed by phenotype through factor union algebra. Moor-Jankowski et al. (1964) suggested that a binary system for phenotypes might prove useful for coding. Introduction of factor union algebra allows not only coding of phenotypes but also of genes and genotypes, and besides these, it characterizes a genetic system through logical unions of factors. For instance, if a particular genotype shows only one phenotype (such as genotype A_1O being A_1-phenotype) which excludes genetic systems with incomplete penetrance, then such a system has been called "regular phenotype system" (Cotterman, 1953). Whenever we apply the factor union algebra to genetic systems, it will be noted that the phenotypic vector of homozygotes is always equal to the corresponding gene vector. This imposes another restriction to the regular phenotype system: that is, none of homozygotes may show phenotype of the other homozygotes. Thus the factor union system is a special case of the regular phenotype system. Since it has been suggested that any genetic system can be described by Boolian algebra (Morton, 1965b) involving three binary operations: union, intersection and complementation, which correspond to dominance, recessivity, and taking the complementary alleles, respectively, it is then no wonder that the factor union system is a subset of the regular system. However, this subset covers almost all regular phenotype systems arising in genetics.

To illustrate this, the haptoglobin system, one of serum protein polymorphisms, will serve as a good example. This haemoglobin-binding protein of serum was shown to vary in different individuals by starch gel electrophoresis (Smithies, 1955), and a simple genetic hypothesis

involving two autosomal alleles Hp^1 and Hp^2 has been proposed to account for the inheritance of the three haptoglobin types. Assigning 1 for the presence of band(s) at a specific position on starch and 0 for the absence, gene vectors are then

$$Hp^1 = (1, 0)$$
$$Hp^2 = (0, 1).$$

Therefore genotype or phenotype vectors are:

genotype	phenotype	binary
Hp^1/Hp^1	Hp1-1	(1, 0)
Hp^1/Hp^2	Hp2-1	(1, 1)
Hp^2/Hp^2	Hp2-2	(0, 1).

Further studies, subjecting purified haptoglobin to reductive cleavage and starch gel electrophoresis revealed isoalleles 1F and 1S from Hp^1 (Connell et al., 1962). Arranging factors in order from the fastest moving band to the slowest, gene vectors now become:

$$Hp^{1F} = (1, 0, 0)$$
$$Hp^{1S} = (0, 1, 0)$$
$$Hp^2 = (0, 0, 1)$$

and six genotype or phenotype vectors are

genotype	phenotype	binary
Hp^{1F}/Hp^{1F}	1F	(1, 0, 0)
Hp^{1F}/Hp^{1S}	1F-1S	(1, 1, 0)
Hp^{1F}/Hp^2	2-1F	(1, 0, 1)
Hp^{1S}/Hp^{1S}	1S	(0, 1, 0)
Hp^{1S}/Hp^2	2-1S	(0, 1, 1)
Hp^2/Hp^2	2	(0, 0, 1).

In this case even if one more factor for the Hp^1 is added, the number of phenotypes remains as above, indicating that binary expression of gene is not unique though the order of factors is determined.

Variants of Hp^2 are also studied (Giblett, 1959). One of them, the allele Hp^{2m} that is responsible for the phenotype Hp2-1(Mod) showing lighter and fewer slow moving bands than the common Hp2-1 phenotype is especially of interest since the gene Hp^{2m} can be identified only in heterozygous condition with Hp^1 or Hp^{1F} or Hp^{1S}. The homozygote Hp^{2m}/Hp^{2m} cannot be distinguished from the other homozygotes which are generated by several variants of Hp^2 (for example, Hp^2/Hp^2). In our terminology, a genetic system with three alleles Hp^1, Hp^2 and Hp^{2m} is a regular phenotypic but not factor union system. By the way, this is the only example of a regular system that is not a factor union system so far observed among human polymorphisms. Regardless of the above limitation, the factor union algebra extremely simplifies grouping operations of genotypes by phenotype, particularly when an electronic computer is available.

One more aspect of binary expression of genes should be mentioned before closing this section. It is possible to figure out the total number of phenotypes in a factor union system. For instance, at the A_1A_2BO blood group system, the factors against anti-A_1 and anti-A generate three phenotypes, while the factor against anti-B produces two phenotypes so that as a whole $2 \times 3 = 6$; six phenotypes are expected, as it should be. The same logic does not always hold for the other cases, but a principle is, first, to find "independent factors" in vector which may form sub-vectors and, second, to figure out all possible phenotypes

from each sub-vector and multiply them. Otherwise, it is necessary to evaluate all possible genotype vectors and to pick out different types of vectors as possible phenotype vectors which can be performed easily by use of computer. Examples of factor union systems which has been used for studies of northeastern Brazilian population are given in Table 3.2.

3.3.4. Generalized maximum likelihood scoring method for estimating gene frequencies and the total inbreeding coefficient

The method that will be discussed here is primarily intended to use an electronic computer to estimate systematically gene frequencies and the inbreeding coefficient. However, the procedure itself is general so that there is no difficulty to follow it with desk calculator and a sheet of paper. At hand, two kinds of data are available; mating type frequencies and individual phenotype frequencies data. Since the second kind of data can be prepared from the first kind, it is possible to analyze same data by different procedures in this case. Both methods may be incorporated into a single program for computer, but discussion in the following will be separated. The program for the estimation of gene frequencies and the inbreeding coefficient from individual frequency data is called G-TYPE while MATYPE is for mating type frequency data. Maximum likelihood methods are of course employed. It has already observed that contribution from genetic model to the maximum likelihood method is in U-scores. In other words, once we have obtained U-scores, K-scores then follow and the iteration process for improving the tentative values goes automatically. Therefore the evaluation of U-scores is the most pertinent problem in both programs.

Program G-TYPE: At the beginning, tentative values of gene frequencies and the inbreeding coefficient must be found from any conventional methods or previous studies. For the inbreeding coefficient, assumption of random mating suggests to take $\alpha=0$ as a tentative value.

Let n_φ be the observed number of individuals of phenotype φ, determined by a locus $\underline{A}$ with alleles $\underline{A}_1$, $\underline{A}_2$, ..., $\underline{A}_k$ in frequencies p_1, p_2, ..., p_k in a population with inbred component α. Then the expected frequency of φ is

$$P(\varphi) = \sum_g P_g,$$

where P_g is frequency of a zygote. It should be called to attention that contribution to U-score from the phenotype φ is a ratio of a first derivative of phenotype frequency to phenotype frequency. Thus

$$U_\alpha = \sum_\varphi \frac{n_\varphi}{P(\varphi)} \left(\frac{\partial P(\varphi)}{\partial \alpha}\right) = \sum_\varphi n_\varphi u_{\alpha\varphi}$$

and

$$U_{p_1} = \sum_\varphi \frac{n_\varphi}{P(\varphi)} \left(\frac{\partial P(\varphi)}{\partial p_1}\right) = \sum_\varphi n_\varphi u_{p_1\varphi} \qquad (1 = 1, \ldots, k)$$

For a particular genotype, we have

$$P_g = p_i^2 + p_i(1-p_i)\alpha \qquad \text{for homozygote}$$

$$= 2p_ip_j(1-\alpha) \qquad \text{for heterozygote}$$

and

$$\left[\frac{P(\varphi)}{\partial \alpha}\right]_g = p_i(1-p_i) \qquad \text{for homozygote } (i=1, \ldots, k)$$

$$= -2p_ip_j \qquad \text{for heterozygote } (i \neq j)$$

$$\left[\frac{P(\varphi)}{\partial p_1}\right]_g = 2p_i + (1-2p_i)\alpha \qquad \text{for homozygote } (1=i)$$

$$= 0 \qquad \text{for homozygote } (1\neq i)$$

$$= 2p_j(1-\alpha) \qquad \text{for heterozygote } (1=i)$$

$$= 2p_i(1-\alpha) \qquad \text{for heterozygote } (1=j)$$

$$= 0 \qquad \text{otherwise}$$

so that $P_\varphi u_{\alpha\varphi} = \sum_g \left[\frac{\partial P(\varphi)}{\partial \alpha}\right]_g$ and $P_\varphi u_{p_1\varphi} = \sum_g \left[\frac{\partial P(\varphi)}{\partial p_1}\right]_g$. Since we have genotype vector generated from gene vectors in binary code, grouping of the same binary vectors makes it possible to perform $\sum_g$ operations. In other words, genotypes whose vectors are same can be summed to obtain both probabilities and their scores. This process is easily performed in computer. To illustrate the method, Gm factor of human gamma globulin serves as a good example. When we take five factors a, b, x, c, and d (which will be discussed in the next chapter), five genes are shown in binary system as:

gene	factors a	b	x	c	d
Gm^a	= (1,	0,	0,	0,	0)
Gm^{ab}	= (1,	1,	0,	0,	0)
Gm^{ax}	= (1,	0,	1,	0,	0)
Gm^{abc}	= (1,	1,	0,	1,	0)
$Gm^{b(1,2)}$	= (0,	1,	0,	0,	1)

whose tentative frequencies are p_1, p_2, p_3, p_4 and p_5, respectively, and $\alpha=0$. (In the following the name of the locus Gm is omitted.). The possible genotypes, their vector which are generated by factor union algebra and their derivative with respect to α and one of five derivatives with

respect to gene frequency, say of abc, p_4, are as follows:

genotype	binary	frequency	derivatives with respect to: α: $P_g U_{\alpha g}$	p_4: $P_g U_{p_4 g}$
a/a	10000	p_1^2	$p_1(1-p_1)$	0
ax/ab	11100	$2p_2p_3$	$-2p_2p_3$	0
ax/abc	11110	$2p_3p_4$	$-2p_3p_4$	$2p_3$
ax/b(1,2)	11101	$2p_3p_5$	$-2p_3p_5$	0
ax/ax	10100	p_3^2	$p_3(1-p_3)$	0
ax/a	10100	$2p_1p_3$	$-2p_1p_3$	0
ab/ab	11000	p_2^2	$p_2(1-p_2)$	0
ab/a	11000	$2p_1p_2$	$-2p_1p_2$	0
abc/abc	11010	p_4^2	$p_4(1-p_4)$	$2p_4$
abc/ab	11010	$2p_2p_4$	$-2p_2p_4$	$2p_2$
abc/a	11010	$2p_1p_4$	$-2p_1p_4$	$2p_1$
abc/b(1,2)	11011	$2p_4p_5$	$-2p_4p_5$	$2p_5$
ab/b(1,2)	11001	$2p_2p_5$	$-2p_2p_5$	0
a/b(1,2)	11001	$2p_1p_5$	$-2p_1p_5$	0
b(1,2)/b(1,2)	01001	p_5^2	$p_5(1-p_5)$	0

Thus ten phenotypes with frequencies and U-scores, by grouping the same genotype vectors, are:

phenotype	binary	frequency (P)	$P \cdot u_{\alpha\varphi}$	$P \cdot u_{p_4\varphi}$
a	10000	p_1^2	$p_1(1-p_1)$	0
abx	11100	$2p_2p_3$	$-2p_2p_3$	0
abcx	11110	$2p_3p_4$	$-2p_3p_4$	$2p_3$
axbd	11101	$2p_3p_5$	$-2p_3p_5$	0
ax	10100	$p_3^2+2p_1p_3$	$p_3(1-p_3)-2p_1p_3$	0
ab	11000	$p_2^2+2p_1p_2$	$p_2(1-p_2)-2p_1p_2$	0
abc	11010	$p_4^2+2p_1p_4+2p_2p_4$	$p_4(1-p_4)-2p_1p_4-2p_2p_4$	$2p_1+2p_2+2p_4$
abcd	11011	$2p_4p_5$	$-2p_4p_5$	$2p_5$
abd	11001	$2p_1p_5+2p_2p_5$	$-2p_1p_5-2p_2p_5$	0
bd	01001	p_5^2	$p_5(1-p_5)$	0

It should be that $\sum_\varphi P(\varphi) = 1$, $\sum_\varphi(\frac{\partial P(\varphi)}{\partial \alpha}) = \Sigma P(\varphi)u_{\alpha\varphi} = 0$ and $\sum_\varphi P(\varphi)u_{p_4\varphi} = 2$, and which may be used for check of scores. U-scores are therefore calculated by

$$U_\alpha = \sum_\varphi n_\varphi u_{\alpha\varphi}$$

$$U^*_{p_4} = \sum_\varphi n_\varphi u_{p_4\varphi}.$$

All the other U-score for gene frequencies are calculated in the same manner. Imposing the restriction that $p_5 = 1-p_1-p_2-p_3-p_4$, the independent U-scores with respect to gene frequency become

$$U_{p_i} = U^*_{p_i} - U^*_{p_5} \qquad (i = 1,2,3,4)$$

Then the variances are given by

$$K_{\alpha\alpha} = \sum_{\varphi} n_{\varphi} u^2_{\alpha\varphi}$$

$$K_{\alpha p_i} = \sum_{\varphi} n_{\varphi} u_{\alpha\varphi} (u_{p_i\varphi} - u_{p_5\varphi})$$

$$K_{p_i p_j} = \sum_{\varphi} n_{\varphi} (u_{p_i\varphi} - u_{p_5\varphi})(u_{p_j\varphi} - u_{p_5\varphi}) \qquad (i,j = 1,2,3,4)$$

Thus improved estimates of parameters are obtained by

$$\hat{p}^* = p + UK^{-1},$$

where all quantities are in matrix notation.

Further improvement may be performed from beginning with value $\hat{p}^*$. Standard errors of estimates are obtained from the square root of corresponding diagonal elements of K^{-1} matrix. The estimate of the dependent parameter, in this case p_5, is calculated from the relation $p_5 = 1-p_1-p_2-p_3-p_4$ and its variance is the sum of all elements with respect to gene frequency in K^{-1} matrix. Detail of program G-TYPE and its instruction for user is presented in Appendix 7.

Program MATYPE: General principle is the same as program G-TYPE.

Suppose that n_{φ} is the observed number of couples of phenotypic mating type φ, determined by a locus G with alleles A, B, C, D, ... in frequencies p_A, p_B, p_C, p_D, ... in a population with inbred component α. Then the expected frequency of φ and its U-scores are

$$P(\varphi) = \sum_{M} P_M$$

$$U_{\alpha} = \sum_{\varphi} \frac{n_{\varphi}}{P(\varphi)} \left[\sum_{M} \frac{\partial M}{\partial \alpha} \right]$$

$$U_{p_G} = \sum_{\varphi} \frac{n_{\varphi}}{P(\varphi)} \left[\sum_{M} \frac{\partial M}{\partial p_G} \right] \qquad (G = A,B,C,D,\ldots)$$

where P_M, $\frac{\partial M}{\partial \alpha}$ and $\frac{\partial M}{\partial p_G}$ are genotypic mating type frequency and its first derivative with respect to α and p_G, respectively. These are given in the Table 2.5.5. for seven basic genotypic mating types, where $\sum_M$ is taken as summation over genotypic mating types whose phenotypic mating type is same. It should be explained about the operation $\sum_M$ in this case that it is somewhat different from the G-TYPE method which required only a single comparison between two genotypic vectors for grouping. The A_1A_2BO blood group system is again taken as an example to illustrate the procedure. Ten genotypes and their vectors are generated through the factor union algebra from four alleles with their binary vectors:

genotypes	binary
$\underline{A_1A_1}$	110
$\underline{A_1A_2}$	110
$\underline{A_1B}$	111
$\underline{A_1O}$	110
$\underline{A_2A_2}$	010
$\underline{A_2B}$	011
$\underline{A_2O}$	010
$\underline{BB}$	001
$\underline{BO}$	001
$\underline{OO}$	000

Then genotypic mating types, their binary, frequencies, and derivatives are, for instance, as follows:

mating type	binary	frequency (P)	derivative of P Pu_α	Pu_{p_1}	
$\underline{A_1A_1} \times \underline{A_1A_1}$	110,110	p_1^4	$6p_1^3(1-p_1)$	$4p_1^3$	...
$\underline{A_1A_1} \times \underline{A_1A_2}$	110,110	$4p_1^3p_2$	$12p_1^2p_2(1-2p_1)$	$12p_1^2p_2$	...
$\underline{A_1A_1} \times \underline{A_1B}$	110,111	$4p_1^3p_3$	$12p_1^2p_3(1-2p_1)$	$12p_1^2p_3$	...
.....	.....	.....	.....	.....	...
$\underline{A_1A_2} \times \underline{A_10}$	110,110	$8p_1^2p_2p_4$	$8p_1p_2p_4$	$16p_1p_2p_4$	...
.....	.....	.....	.....	.....	...
$\underline{A_1B} \times \underline{A_10}$	111,110	$8p_1^2p_3p_4$	$8p_1p_3p_4$	$16p_1p_5p_4$	...
.....	.....	.....	.....	.....	...
$\underline{00} \times \underline{00}$	000,000	p_4^4	$6p_4^3(1-p_4)$	0	...

Where $\Sigma P=1$, $\Sigma Pu_\alpha = 0$ and $\Sigma Pu_{p_i} = 4$ which may serve for checking of scores. Grouping is performed if two mating type vectors consist of the same genotypic vectors; for example, both mating types $\underline{A_1A_1} \times \underline{A_1A_1}$ and $\underline{A_1A_2} \times \underline{A_10}$ have binary (110,110) so that they are grouped in probability and its derivative. However, this does not cover all of cases. For instance, mating $\underline{A_1A_1} \times \underline{A_1B}$ and $\underline{A_1B} \times \underline{A_10}$ have binary code (110,111) and (111,110), respectively. Yet we have to have grouped both mating types since we are not concerned with sex. This is simply carried out by a change of order in genotype vectors in one of mating type binary codes.

Thus we obtain the probabilities and their derivatives for each of twenty-one phenotypic mating type. Calculation of K-scores and the iteration process are now straightforward. Instructions for program MATYPE and information about programming logic are discussed in Appendix 8.

3.4. Fitting of migration function

In order to fit any function for describing human migration, we shall use the probability-integral transformation as a means of transforming any known continuous distribution to the rectangular distribution of interval (0, 1) (Kendall and Stuart, 1961). The moment method to obtain mean marital distance and its variance will be avoided because of high frequency of zero distance class that is unescapable in practice and of bias due to a few case with large distance. If we have assumed a simple migration function m(x) with distance x, then the variable

$$P(x) = \int_0^x m(t)dt$$

is distributed on (0, 1). Thus if we have a set of n observations x_i and transform them to a new set P_i by the probability-integral transformation and use a function of the P_i to test the departure of the P_i from rectangularity, the distribution of the test statistics will be distribution-free, not merely asymptotically but for any number of observation. In other words, if the distribution data and probabilities are given by:

class interval	observed number	probability
$0 - x_1$	n_0	P_0
$x_1 - x_2$	n_1	P_1
.....	.	.
$x_t - \infty$	n_t	P_t

then the likelihood of observations becomes

$$L = \frac{n!}{n_0!n_1! \ldots n_t!} P_0^{n_0} P_1^{n_1} \ldots P_t^{n_t}$$

where $\Sigma n_i = n$ and

$$P_i = \int_{x_i}^{x_{i+1}} m(t)dt = P(x_{i+1}) - P(x_i).$$

It is obvious therefore that not only test of migration function can be made, but the distribution parameters are also estimated by the maximum likelihood scoring method and we may calculate from them the mean marital distance, its variance, the inbreeding coefficient by marital distance, and so forth.

The functions we have examined for human populations are listed in the Table 3.3., including six single parameter probability functions, two with two parameters and one of three parameters. These functions have met the conditions that we have discussed in 2.8. Results of fitting will be given in the next chapter. The estimates by this method are somewhat different from the ordinary maximum likelihood (ML) estimates which might be obtained directly from migration function and without probability-integral transformation. In the latter method, we have to assume the observation of individual distance with great accuracy. It is practically impossible to measure distance "precisely". However, the multinomial ML estimates which are obtained from probability-integral transformation might give us sufficient informations about migration pattern without serious error due to a few biased observation. The multinomial ML estimates are, however, dependent upon the number of class intervals and the observed frequencies in each class. There are recommendations for χ^2 test (Kendall et al. cited) for determination of the

number of classes and choice of class interval. For our purpose, we have chosen intervals exponentially, considering a leptokurtotic migration pattern.

To visualize the method, a log normal function will serve. The function has a single parameter and its form is

$$m(x) = \frac{a}{\sqrt{\pi}\,x} e^{-a^2(\ln x)^2}$$

where a may be called an attraction parameter. By the probability integral transformation,

$$P(x) = \int_0^x \frac{a}{\sqrt{\pi}\,t} e^{-(a\ln t)^2} dt \qquad [\text{Note that } P(0) = 0]$$

$$= \frac{1}{2}\Phi(\sqrt{2}a\ln x) \qquad [\text{where } \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-y}^{y} e^{-\frac{t^2}{2}} dt]$$

so that

$$P_i = \frac{1}{2}[\Phi(\sqrt{2}a\ln x_{i+1}) - \Phi(\sqrt{2}a\ln x_i)]$$

whose derivative with respect a is

$$P_i u_{ai} = \frac{1}{\sqrt{\pi}} [(\ln x_{i+1}) e^{-(a\ln x_{i+1})^2} - (\ln x_i) e^{-(a\ln x_i)^2}]$$

Then the U-score and the variance are

$$U_a = \sum_i n_i u_{ai} \qquad K_{aa} = \sum_i n_i u_{ai}^2$$

The improved estimates is obtained from:

$$a_1 = a + U_a/K_{aa}$$

and χ^2-test for parameter is performed by

$$\chi^2 = U_a^2/K_{aa}$$

with one degree of freedom. Goodness of fit may be made by $\chi^2 = \Sigma\frac{(O-e)^2}{e}$ with k-2 degree of freedom, where k is the number of class intervals.

Thus the mean marital distance is

$$\bar{x} = \int_0^\infty tm(t)dt$$

$$= \frac{a}{\sqrt{\pi}} \int_0^\infty e^{-(alnt)^2} dt = e^{\frac{1}{4a^2}}$$

and its variance is obtained from

$$Var(\bar{x}) = \bar{x}^2/16a^6K_{aa}.$$

Given an inbreeding function with distance f(x), the inbreeding coefficient of population will be

$$\alpha = \int_0^\infty f(x)m(x)dx \doteq \sum_{x_i} f(x_i)m(x_i)\Delta x_i.$$

The discrete approximation may be used when the integral is complicated and Δx_i is class interval. x_i may be taken as an average distance weighted by the observed numbers in the interval. In case of exponential inbreeding and lognormal migration function, we have

$$\alpha \doteq \frac{af_0}{\sqrt{\pi}} \Sigma \left[e^{-bx_i-(alnx_i)^2}\right] \frac{\Delta x_i}{x_i}$$

The process of convergence to obtain the final estimate is usually very slow, taken on the average about twenty iterations with an initial value estimated by moment method. Consequently, the more parameters are involved in migration function, the more iterations are expected, or even failure in convergence. Experience with this method has told us that the migration function examined would not carry more than three parameters in spite of a biological interest in constructing functions such as a linear combination of three normal distributions (Cavalli-Sforza et al., 1965). The study of northeastern Brazilian population will answer to the above problems in the next chapter.

3.5. Summary

The maximum likelihood scoring method is applied to estimate gene frequencies and the inbreeding coefficient, to test an assumption of random mating, and to fit migration function. From individual data, it is found that the estimation of inbreeding coefficient and the test of random mating are statistically impossible at the ABO blood group system and two allelic loci with complete dominance. Introducing factor union algebra which defines dominance relations between alleles in binary code, a highly convenient method with using an electronic computer is devised. An efficient method of fitting migration functions is also discussed.

4. Practical surveys

4.1. Introduction and material

4.1.1. General feature of population

The data to be analyzed here have been collected primarily under the supervision of Dr. Newton E. Morton at the Hospedaria de Imigrantes in São Paulo, Brazil during the year beginning June 1962 (Morton, 1964), for determining the effects of various genetic factors on mortality and morbidity in a rigorous environment. However, the study was also made for an investigation of isolation by distance. The 1068 migrant families from northeastern Brazil passed through the Hospedaria de Imigrantes with government aid to the interior of the states of Sao Paulo and Parana. The sample was taken from a government registry according to pre-established criteria of rural origin, wife under 50 years of age, presence of both husband and wife and long cohabitation time. The population was characterized by mixture of three major races: 11, 30 and 59 percent of Indian, Negro and Caucasian, respectively (Krieger et al., 1965). Each sampled family submitted to an interview and medical examination, during which blood and saliva specimens were taken. For the sake of studying population structure, an interview was also made with each parent separately for information about birthplaces, marital distances and consanguinity in both parental and grandparental generations. Discrepancies were checked by confronting the two parents and discussing the point at issue. The marital distances were also checked on the map (IBGE, 1958) and coded in km. Because of low level of literacy (Krieger et al., 1965) and of usage of a large unit of scale, légua (1 légua = 6 km.), on distance, the estimated distances were only

an approximation. The population density of birthplaces were calculated from government official report (IBGE, 1961) by dividing rural population from where individual was drawn by area of the município, or county.

While the husband and wife were being interviewed, the children were assigned numbers and proceeded through the data collection center, where a nurse determined phenylthiocarbamide sensibility (PTC) and anthropometrics, and each subject contributed about 12 cc of blood and a saliva sample. Duplicate laboratories using different antisera typed A_1, A, B, P, C, D, E, M, N, S, s, and K blood groups. Fy^a and c were tested in duplicate with the same antisera. Le and ABH secretion and the Le^a blood group were tested only in one laboratory, and Di, Js^a, f, V, k, e, Lu^a and Le^b blood groups were tested in part of the sample. Red cells were collected in EDTA, stored at 4^oC, and typed blind the next day in two duplicate laboratories. Discrepancies between laboratories, parental exclusions, and serological curiosities were retested using the same and different antisera. Additional tests were performed on problem families by Dr. R. E. Rosenfield in New York using glycerolized cells shipped in dry ice from Brazil. A glycerolized red cell sample (Crawford et al., 1954) was sent to R. E. Rosenfield for supplementary tests, and another to W. Nance and O. Smithies for tests of serum proteins such as haptoglobin and transferrin, and to A. Steinberg for Gm and Inv. All this material was kept in dry ice until receipt in the United States where it was stored in different laboratories at temperatures ranging from -20^oC to -70^oC.

4.1.2. Description of genetic systems employed

Bioassay of the inbreeding coefficient uses genetic systems

which are little affected by selection and technical errors, and are polymorphic with complete penetrance. Sixteen polymorphic systems were employed for the bioassay: ABH secretion, Lewis, Lutheran, PTC, P, Duffy, Inv, Diego, Haptoglobin, Hemoglobin, Transferrin, Kell, ABO, MNSs, Gm, and Rh. Formal genetics of these systems is discussed in brief, according to the Brazilian survey.

Secretor system: Inhibition test of saliva used anti-A, anti-B of human sera and anti-H of saline extracts of the seed of Ulex europaeus. A_2, B, and O red cells were used as indicators to provide classification of individuals into secretor and non-secretor. The ability to secrete the A, B, or H antigen in the saliva is inherited as a Mendelian dominant character. Two alleles, Se and se, are known at the locus. The system is also known as the first example in man of autosomal linkage and autosomal crossing-over with the Lutheran blood group system (the recombination value is estimated as about fifteen percent).

Lewis system: According to inhibition test of saliva with anti-Le^a, individual phenotype was determined as either positive or negative. Thus, this system is treated as two alleles, Le and le, with complete dominance in bioassay. No attempt will be made to explain a current theory of Grubb and of Ceppellini (Grubb, 1951; Ceppellini, 1955) on association between the secretor system and the Lewis system, but the theory has been used for confirmation of phenotype of Lewis and secretor systems in saliva from that on red cells.

Lutheran system: At least two alleles, Lu^a and Lu^b, have been described at this locus. We had tested, however, about two hundred couples with

anti-Lu^a alone in saline suspension of red cells. Therefore, the system is considered as a locus with two alleles, Lu^a and Lu, with complete dominance.

Phenylthiocarbamide sensitivity (PTC): There are two alleles, T and t, with complete dominance of the ability to taste PTC. A continuous anti-mode distribution of sensory threshold by different dilutions of PTC (Azevedo et al., 1965) makes it difficult to classify doubtful cases into either positive or not. In the Brazilian study, subjects were classified as non-tasters if they could not discriminate the solution 5, which contains 81.25 mg PTC/litter. This criterion was derived from pedigree studies of doubtful cases involved. Children age eight years or less have been excluded because of multiple sorting errors and of being significantly higher phenotypic frequency of non-taster than among older peoples.

P system: Three alleles, P_1, P_2 and p, in this system have a similar relation as A_1, A_2 and O, in the ABO locus. Since we have used only anti-P_1 to detect the cold agglutinogen on red cells, the system becomes two alleles, P_1 and P_2+p, with complete dominance.

Duffy system: Three alleles, Fy^a, Fy^b and Fy, with ABO type dominance (Fy^a corresponds to A, Fy^b to B and Fy to O) have been described with anti-Fy^a and anti-Fy^b. The high frequency of Fy(a-b-) individuals in Negro but rare in Caucasian have offered an anthropological interest and have led to study of dosage effect of Fy^a antigen on red cells (Race et al., 1953). Again only anti-Fy^a was available in the Brazilian study so

that the system was considered as two alleles, $\underline{Fy^a}$ and $\underline{Fy^b}$ + $\underline{Fy}$, with complete dominance.

Inv system: Three different genetic factors, Inv(a), Inv(b) and Inv(1) have been described in β_2M globulins (19S γ-globulins), β_{2A}-globulins and Bence-Jones proteins by inhibition tests. Dr. A. G. Steinberg typed Inv(a) and Inv(b) factors on Brazilian material. However, only the results of Inv(a) factor was used in bioassay analysis because of unreliable reactions with Inv(b) reagents (Steinberg, 1964). Thus the locus consists of two alleles, $\underline{Inv^a}$ and $\underline{Inv}$ (=$\underline{Inv^b}$ + $\underline{Inv^1}$), with complete dominance.

Diego system: The $\underline{Di^a}$ antigen on red cells is essentially a Mongolian character so that we had decided to investigate Brazilian material. About two hundred couples and eighty-seven children whose one of parents was positive were submitted in typing. The system is considered as two alleles, $\underline{Di^a}$ and $\underline{Di}$, with complete dominance.

Haptoglobin system: Some of aspects on the locus have been discussed in 3.3.3. This, one of the serum protein polymorphisms, is treated as three alleles, $\underline{Hp^{1F}}$, $\underline{Hp^{1S}}$ and $\underline{Hp^2}$, without dominance (codominance). Incidentally, a variant of Hp^1, Carlsberg, and ahaptoglobulinemia subjects are ignored from study.

Hemoglobin system: Hemoglobin, the oxygen carrying protein molecule, consists of four polypeptide chains and a heme group attached to each chain. The polypeptide chains are usually classified into two pairs of identical chains, called α-chain and β-chain, so that its structure can

be written $\alpha_2\beta_2$. While synthesis of each chain is controlled by genes such as thalassaemia genes and regulators, the structure of amino acid sequence in each chain is also genetically controlled. Besides the normal hemoglobin, Hb-A, the genetic variants at the sixth position from the N-terminal in β-chain, Hb-S (glutamine → valine) and Hb-C (glutamine → lysin) are relatively common (clinically known as sickle cell anemia major and minor, according to whether all β chains are affected or only one half) and, therefore, they are used in bioassay. The other variants observed in the field work such as Hb-F and Hb-A_2 were not used for bioassay because of rare frequencies. The locus is thus considered as three alleles, $\underline{Hb^A}$, $\underline{Hb^S}$ and $\underline{Hb^C}$, without dominance. All parents are typed, whereas only children whose both parents are not $\underline{Hb^A/Hb^A}$ were submitted to test.

Transferrin system: One of β-globulins is called transferrin or siderophilin which transports plasma iron to bone marrow and tissue storage areas. Starch gel electrophoresis shows three major variants B, C and D, in order of faster moving bands, which are genetically controlled and are codominant to each other. It has been suggested that their different mobility rates may be due to alternatives in the number of sialic acid residues.

Kell system: The recent discovery that the Sutter antigens Js^a and Js^b on red cells are localized in the Kell system (Stroup et al., 1964; Morton et al., 1965) gives three alleles, K, k and k^s, at this locus. The other alleles, k^o and k^p were not distinguished from k in Brazilian material, since we used anti-K2 (Cellano) only on Kell-positive cells and

did not use anti-K3 (Penney). Thus the locus consists of three alleles, K, k and k^s, without dominance between K and k or k^s but with complete dominance k^s to k. In Brazil, all individuals were tested with anti-K1 (Kell), but only the last fifth of the sample, beginning at family 855, with anti-K6 (Sutter).

ABO system: Using the antisera, anti-A_1, anti-A and anti-B, all of which are commonly found in sera of human beings who do not have the corresponding antigen, six phenotypes can be distinguished. Based on Bernstein's multiple allele hypothesis, Thomsen et al. (1930) put forward the four allele, A_1, A_2, B and O, theory of inheritance. The dominance relation between alleles are so-called "ABO type dominance" which was explained in 3.3.3.

MNSsU system: The segregating factor pairs in this locus, the MN-series and the Ss-series, are one of the most interesting blood group systems. The discovery of the MN-series was dependent upon making anti-M, injected human red cells into rabbits (heteroimmunization). This series alone consists of two major alleles, M and N, without dominance if anti-M and anti-N are used. In the Ss-series, however, the factors S and s are not always complementary, especially in Negro populations. It has been interpreted as the ABO type dominance between three alleles, S, s and s^u, using anti-S and anti-s. The s^u gene was proposed to explain a phenotype S(-)s(-). The discovery of a third antibody, anti-U, has rendered some complications. It has been observed so far that all S(+) and s(+) red cells are U(+) (Wiener et al., 1953), but not all S(-)s(-) phenotypes are U(-) (Francis et al., cited by Race and Sanger, 1962, p. 91).

Ignoring the reaction with anti-U on red cells, Morton et al. (1965) have suggested a notation * for S(-)s(-) reactions, instead of u or U.

Combination of two series results in six alleles MS, Ms, M*, NS, Ns, and N* at the locus.

Gm system: The Gm factors are determined by an intricate series of alleles. Normal human sera may or may not inhibit agglutinating activities of rheumatoid sera (agglutinator) which agglutinate red cells coated with incomplete anti-D. The ability of inhibition is genetically controlled, and combination of anti-D and agglutinator determines the gamma globulin factors such as Gm(a), Gm(b), Gm(c), Gm(x) and Gm(b2). The series of antigens on the 7S-globulin (γ_2-globulin) molecules of man occurs with different frequencies in different populations. For instance, Gm(c) occurs only in Negroes who in turn do not have Gm(x); and only the Caucasoids show variation in the frequencies of Gm(a). Among results in Brazilian material, Gm(a), Gm(b1), Gm(c), Gm(x) and Gm(b2) factors are employed for our purpose. Since the agglutinator "Davis" was used to type Gm(b2) (Steinberg et al., 1965), we simply designated this factor by "d" instead of b2 in binary code (see 3.3.3.). Thus, the system consists of five alleles, Gm^{a}, $Gm^{ab(1)}$, Gm^{abc}, Gm^{ax} and $Gm^{b(1,2)}$, which generate ten phenotypes.

Rh system: During the Brazilian study, anti-D, anti-C, anti-c and anti-E were used in the routine work and E+ samples were typed with anti-e. Only 354 families were tested with the sera, anti-f and anti-V, in addition to the above antisera. In this locus, eight alleles, cde, cdE, Cde, CdE, cDe, cDE, CDe, and CDE are currently accepted. When the

relative frequencies of alleles such as CDE, cdE and CdE are rare, we have ignored them. The last 354 families were used to analyze the 11 allele system: cde, cde^s, Cde, Cde^s, cdE, CdE, cDe, cDe^s, CDe, cDE and CDE, where anti-f and anti-V react with gene complex ce and ce^s, respectively. It has assumed that the allele, Cde^s, in Negroes reacts with anti-c (Rosenfield, 1964).

4.2. Population gene frequencies

Population gene frequencies in northeastern Brazil have been estimated by G-TYPE under the assumption of panmixia. It is confirmed that MATYPE gives exactly the same results in sixteen polymorphic systems, which was theoretically predicted for the special case of a two allelic locus (3.3.2.). Table 4.2.1 shows gene frequencies of the total population in the parental generation. This table also includes gene frequencies where couples are divided by their marital distance. On the whole, differentiation in gene frequencies by marital distance is hardly observed in this population. When the couples with close consanguinity ($F \geq 1/32$) are removed from the total population, the gene frequencies are estimated (Table 4.2.2.) to be practically the same as for the total population.

For the bioassay of the inbreeding coefficient, the calculation of gene frequencies was also made for the subpopulations defined by distance times the square root of density, in order to simulate a population of uniform density (Table 4.2.3. and 4.2.4.). Bioassay was also made for unknown consanguinity (Table 4.2.5.) and for children (Table 4.2.5.) to see breakdown of isolates by generation.

4.3. The inbreeding coefficient

4.3.1. Pedigree study

In order to ascertain the inbreeding coefficient, the couples are divided by the inbreeding coefficient (F) of their child, where F can be expressed in term of power of 1/2: $(1/2)^c$. Since it has been necessary to have a code of consanguinity for data collection, the

negative of the logarithm of inbreeding coefficient to the base 2, $c = -\log_2 F$, has been taken as the code (Moroni, 1962) which may cover an interval: $[c-0.5, c+0.5)$ where [and) mean "including the border value" and "not including the border value", respectively. This can then be transformed into F (Table 4.3.1.). This interval classification of consanguineous marriages is useful when a large body of data is available.

Table 4.3.2. summarizes distributions of couples by the inbreeding coefficient and the marital distance in northeastern Brazil, and the mean inbreeding coefficient and its standard error were calculated by formulae given in 3.2. The total inbreeding coefficient is .0059 ± .0006 in children and .0036 ± .0004 in parents. The lower level of the inbreeding coefficient in parents than in children might be due to incomplete ascertainment of consanguinity in the parental generation. These values may be comparable with α=.0050 which is obtained from Catholic marriages in parental rural populations in northeastern Brazil (Freire-Maia, 1957), which is also heterogeneous in time. Thus it seems reasonable that our migrant families as representative of northeastern Brazil with respect to the ascertained inbreeding coefficient.

It is also seen in Table 4.3.2. that close consanguinity up to the second cousin marriage accounts for .0051/.0059 = .86 or 86 percent of the total inbreeding coefficient ascertained by pedigree analysis.

4.3.2. The total inbreeding coefficient by bioassay

Two methods, G-TYPE with individual phenotype frequencies and MATYPE with mating type frequencies, are applied to the total and the remote populations. Since we did not find any significant difference in gene frequencies between total and remote populations (Table 4.2.1. and

4.2.2.), the frequencies in the total population will be used.

In the G-TYPE method, only eight systems could be submitted to analysis since a singularity at two allelic locus with complete dominance was well-established in advance. The estimated inbreeding coefficient in the total and in the remote population are .0170 ± .0086 and .0132 ± .0089, respectively. The difference between them is .0038, which may be interpreted as due to close consanguinity in agreement with the value .0036 obtained as the ascertained coefficient of inbreeding of parents in the pedigree study. The heterogeneity of α among eight systems is highly significant in both populations ($\chi^2_7 = 0(10^3)$). However, we have met a peculiar property of the information about α at the ABO locus (Table 4.3.4.), which has in turn been explained mathematically as mentioned in 3.3. The other three systems which show a significant deviation from the hypothesis that $\alpha=0$ also give little information.

Taking away the four systems with very small amounts of information (the ABO, hemoglobin, transferrin and Kell loci), the inbreeding coefficients become .0246 ± .0086 in the total population and .0208 ± .0089 in the remote population. The difference is .0038, in agreement with the previous result. Heterogeneity of α among the four remaining systems is insignificant ($\chi^2_3 = 7.43$ for the total and $\chi^2_3 = 6.07$ for the remote populations).

Although the insensitiveness of G-TYPE method has been discussed, multiple allelic systems without dominance, or even with little dominance, seem to give good information about α. The number of iteration for α with all eight systems was six in our material.

The MATYPE method gives the coefficient of consanguinity of parents or the inbreeding coefficient of children. All sixteen polymorphic systems now contribute to information about α. The estimated inbreeding coefficients are .0133 ± .0035 in the total and .0082 ± .0034 in the remote population, the difference of .0051 corresponding to the close inbreeding coefficient of children in pedigree analysis. The heterogeneity test on α among systems is again highly significant ($\chi^2_{15}=0(10^2)$). When we remove the systems with very small amounts of information about α, Secretor, Lutheran, Diego and Kell, the inbreeding coefficients become .0160 ± .0035 in the total and .0106 ± .0035 in the remote populations and the heterogeneity tests become nonsignificant (χ^2_{11} = 11.25 and χ^2_{11} = 11.12, respectively). Again, the difference, .0054, agrees well with the estimate from pedigree study.

A word should be said about the four systems which have been removed from the estimation of α in the above procedure. Three systems, Lutheran, Diego and Kell, have an allele whose frequency is nearly the same order as the estimated inbreeding coefficient, so there is a possible violation of the restriction that the smallest gene frequency be greater than the inbreeding coefficient. However, the amount of information provided by these systems is too small to justify this speculation.

The superiority of the MATYPE method, comparing with G-TYPE, is observed at the hemoglobin, Transferrin and ABO loci in the greatly increased amount of information about α.

The inbreeding coefficient of children may also be estimated from individual frequencies of children by the G-TYPE method (Table 4.3.6.), which gives .0121 for the total and .0073 for the remote populations.

Although no standard error may be assigned because children are not independent samples, the estimates agree well with .0133 ± .0033 and .0082 ± .0034 respectively which were estimated by MATYPE. This means there is no evidence of heterozygote advantage for these polymorphisms.

The inbreeding coefficient for consanguineous marriages of unknown degree is estimated from mating type frequencies as .0086 ± .0152 (Table 4.3.7.) which corresponds to the degree between second cousin and second cousin once removed. In Table 4.3.8., the total inbreeding coefficients estimated from the available systems are summarized for the sake of comparison. The inbreeding coefficient decreases by generation, indicating that the breakdown of isolates is occurring in northeastern Brazil. This is also supported by the study of migration functions since mean marital distance increases by generation (see 4.4.).

This study demonstrates that the MATYPE method gives more efficient and stable estimate about α than G-TYPE. And convergence with sixteen systems required only six iterations starting from $\alpha=0$. In the following, therefore, we will employ MATYPE for further bioassay analysis of population structure.

4.3.3. Components of the inbreeding coefficient

In northeastern Brazil, the total inbreeding coefficient consists of contributions from: (1) close consanguinity ($F \geq 1/32$) ascertained from pedigree analysis, (2) remote consanguinity ($F < 1/32$) ascertained from pedigree analysis, (3) unascertained consanguinity within a racial group, and (4) racial endogamy. It has already been shown in the pedigree study that the first two components in parents are .0051

and .0009 respectively. There would be a higher value for the ascertained inbreeding coefficient due to remote consanguinity if pedigrees of consanguineous marriages were intensively traced. Since the bioassay analysis indicates .0133 in the total inbreeding coefficient, the ascertained inbreeding coefficient from pedigree analysis is only forty-four percent of it. This points to an error in assessment of inbreeding effects on the basis of the inbreeding coefficient estimated from pedigree study, for the elimination rate for rare recessive genes is greater than had been estimated previously.

Since the population consists of three main racial groups, Negro, Indian, and Caucasian, racial endogamy contributes to the total inbreeding coefficient. Based on correlations for three racial groups, Krieger et al. (1965) estimated the mean endogamy coefficient of the same population to be .030 and the equivalent inbreeding coefficient for fixed loci in the ancestral populations was estimated to be .095 ± .011 (Chung et al., 1965), so that the inbreeding coefficient due to racial endogamy is .030 x .095 = .0029. This is twenty-two percent of the total inbreeding coefficient. The regression of the estimated inbreeding coefficient from the ancestral population on p(1-p) within loci is .68 ± .19, which is highly significant ($F_{1,12} = 12.54$). There is no significant regression on gene frequency p. Thus as the mean gene frequency approaches .5 from either direction, the divergence among populations increases. This can only mean that the polymorphisms are more subject to local selection than are rare genes which may be almost uniformly deleterious, and therefore the contribution of racial endogamy must be less for monomorphisms. And the difference between the endogamy coefficients in

ancestral and in present populations, .095 - .003 = .092, indicates how much breakdown of racial isolates have occurred in the present population after their migration.

The last contribution due to unascertained consanguinity is as important as the ascertained inbreeding coefficient. Table 4.3.9 summarizes the above discussion.

4.3.4. Relationship between the inbreeding coefficient and marital distance

When couples were grouped by marital distance, the inbreeding coefficient might be given as a function of distance and dimension of migration. For a large distance, if migrants move in one dimension, the function is approximately reduced to a simple exponential form: $f=ae^{-bx}$, where f is the inbreeding coefficient at marital distance x, a is the inbreeding coefficient at x=0, and b is a constant measuring decrease of inbreeding with distance. If two dimensional migration on a plane was assumed, the relation would be $f=ae^{-bx}/\sqrt{x}$ for a large distance. A test of this theory was performed with sixteen polymorphisms in northeastern Brazil, estimating the inbreeding coefficient for three distance groups: 0-3 km., 3-27 km. and 27-∞ km., by pedigree study and by bioassay with mating type frequencies (MATYPE).

Two bioassay methods were employed in order to see effects of variations in gene frequencies with distance on the inbreeding coefficient. In bioassay A, gene frequencies are taken from the estimates of the total population (Table 4.2.1.), hence no differentiation in gene frequencies with distance is assumed. On the other hand, the estimated gene frequencies for each distance group are used in bioassay B. In Table 4.3.10, the close inbreeding coefficients for bioassay A and B are taken from

the difference between the total and remote inbreeding coefficient estimated by the corresponding methods. Since the differences in the two methods are within sampling error, the estimates from bioassay A shall be taken in the following discussions. The smaller inbreeding coefficients in 0-3 km. group than those in 3-27 km. population by bioassay might suggest possible avoidance of close consanguineous matings in the shorter marital distance which would lead to a negative correlation between uniting gametes, or they might simply be due to sampling error. A possible selection is ruled out in this case since no differentiation in gene frequencies with distance has been observed. No such reduction of inbreeding for small distances was seen in pedigree study.

When one-dimensional theory was applied to the data by the least square method, taking a distance weighted by the number of couples as a representative quantity in each population and weighting by the information on the inbreeding coefficient, the inbreeding coefficient at x=0 by bioassay A were .0212 ± .0058, .0147 ± .0062 and .0064 ± .0056 in total, remote and close population respectively (note that .0212 ≒ .0147 + .0064) and an exponential relation fits in total and close populations (Figure 4.3.1 - 3.). Deviation from the exponential in the remote population is due to the inbreeding coefficient of shorter marital distance where negative values of F are observed. This would mean that the ascertainment of consanguineous marriages was nearly complete when the marital distance was near zero. On the other hand, the pedigree study showed the opposite result: only the remote population fits well with the exponential hypothesis and the estimated inbreeding coefficient at x=0 were .0085 ± .0011, .0018 ± .0002 and .0064 ± .0010 in total,

remote and close population respectively (again, note that .0085 ≐ .0018 + .0064).

These conclusions are not substantially altered where marital distance is multiplied by the square root of population density so as to simulate a population of uniform density. The units of subdivision of population are 0-29, 30-179, and 180-∞ which give about three hundred couples in the first two subgroups. The goodness of fit in bioassay to an exponential function is improved by this treatment, and there is no significant deviation from the hypothesis in total, close and remote consanguinity (Table 4.3.11.). The inbreeding coefficients at zero distance are substantially the same in the total population by bioassay and pedigree study and in the remote and close groups in pedigree study, suggesting little effects of a heterogeneity in population density on the inbreeding coefficient of couples whose marital distance is nearly zero. Different estimates in the remote population by bioassay might come from different grouping intervals in the distance and density-corrected analyses.

Based on the estimates of a and b, tentative values of average systematic pressure, U, and average migration pressure, M, on sixteen polymorphisms are made by transformations (δ is population density):

$$\begin{cases} a = \dfrac{1}{1+4M\delta\sqrt{2U}} \\ \\ b = \dfrac{\sqrt{2U}}{M} \end{cases}$$

Solving for M and U, we obtain

$$\begin{cases} M = \frac{1}{2}\sqrt{\frac{1-a}{ab\delta}} \\ U = \frac{b(1-a)}{8a\delta} \end{cases}$$

The variances of M and U are also calculated by

$$\begin{cases} \sigma_M^2 = M^2\left[\frac{\sigma_a^2}{4a^2(1-a)^2} + \frac{\sigma_{ab}}{2a(1-a)b} + \frac{\sigma_b^2}{4b^2}\right] \\ \sigma_U^2 = U^2\left[\frac{\sigma_a^2}{a^2(1-a)^2} - \frac{2\sigma_{ab}}{ba(1-a)} + \frac{\sigma_b^2}{b^2}\right] \end{cases}$$

The estimated mean population density was 20 persons per square kilometer in the rural population of northeastern Brazil. Table 4.3.12. gives results from six methods where an exponential inbreeding relation with distance was assumed. Since M is a standard deviation of migration distance, it might be understood as a mean marital distance when an exponential migration was assumed. The most meaningful estimates among them are by bioassay A on the total population. The estimates of M, 12.31 and 105.57, are comparable with 80 and 244 which are obtained from study of migration function (Table 4.5.2. and 4.5.4.). Heterogeneity of density among populations seems to reduce the genetic effect of migration.

The systematic pressures estimated are due to mutation and weak selection as a linear pressure on gene frequencies. The two estimates of U, .0011 ± .0000 and .0068 ± .0055, should be comparable since mutation rate is independent of population density.

The estimate of M is highly dependent upon the parameters a and b in the inbreeding function while U depends mostly on a. Inaccuracy of estimation is indicated by the large standard errors, but strong selective forces on these polymorphisms seem to be unlikely.

The discussion should be extended to two dimensional theory. Since the approximate inbreeding function itself involves a difficulty in the neighborhood of zero distance, the convergence process is poor. In the distance study, only three cases; pedigree study in remote population and bioassay A and B in close populations gave convergent estimates of a and b, whereas the study of distance multiplied by square root of population density resulted in convergency of three analyses for remote population and of bioassay A for close population. From the study with two dimensional model, the following are suggested: (1) The one-dimensional migration may hold for remote consanguineous marriages ($F < 1/32$), and (2) significant differences in b-values between both dimension models may be due to incorrect use of approximations, that is, there must be more exact forms which take account of avoidance of too close inbreeding, such as selfing and brother-sister mating, especially in the two dimensional model. At any rate, we may take the one dimensional model as a first approximation to describe the relation between the inbreeding coefficient and distance, although a study of migration function has suggested that the migration in northeastern Brazil may have dimension 1.7 to 1.9, or nearly two dimensional migration (Table 4.5.1-4.).

As conclusions, an exponential relation between the inbreeding coefficient and the marital distance may hold in the total and remote inbreeding coefficients by bioassay and in the remote inbreeding

coefficient by pedigree study. Deviation from the exponential relation for the inbreeding coefficients in pedigree study of the total and close populations may indicate preferential consanguineous marriages at large distances.

4.3.5. The inbreeding coefficient for alleles

So far the inbreeding coefficient has been discussed as if only one parameter exists per locus. As discussed in chapter 2, the concept of assigning an inbreeding coefficient to each allele might be helpful in understanding population structure, since Wahlund's principle could be applied to the gene in question and heterogeneity in the inbreeding coefficient among alleles might indicate a consequence of random genetic drift, mutation or selection on a particular gene.

A possible statistical method to estimate the inbreeding coefficient for alleles is to reduce a genetic system with multiple alleles to the case of two alleles, the one in question and the others pooled, and then to apply the MATYPE method. The results of analysis for eight systems are shown in Table 4.3.13. The inbreeding coefficients for a locus are calculated by $\alpha = \Sigma \alpha_i p_i$, where p_i and α_i are the frequency of the ith allele (Table 4.2.1.) and its inbreeding coefficient, respectively. In parentheses, the values which had been calculated directly (see Table 4.3.5.) are given for comparison. Both quantities are generally the same order, except in the Kell, MNSsU and Rh systems. Discrepancies in these systems might arise from the method itself because no such couples as $k^s \times k^s$, K x K in the Kell system, N* x N* and M* x M* in the MNSs system, and incrosses with respect to phenotypes r', r'' and r^y, are observed, giving a negative inbreeding coefficient for these alleles or

even a singular matrix. The singularity in the Rh locus required that D(-) alleles be pooled as one allele. Apparently, the lack of some incrosses of homozygotes presents no problem in a codominant system such as Haptoglobin, Hemoglobin and Transferrin. Apart from the negative estimates of the inbreeding coefficient, the probability to be identical by descent seems to vary among alleles even in polymorphic systems. The likeliness may be also suggested by the endogamy coefficients estimated in the ancestral population of northeastern Brazil (Table 4.3.14) where the coefficients may represent a diversity of allele frequencies between three ethnic groups. The causes of variation among the inbreeding coefficients for allele might reflect random genetic drift, mutation or selection, but no conclusion can be drawn at this stage.

4.4. Mating type frequencies and the related probabilities

Mating type: Mating type frequencies have offered a method of joint estimation of gene frequencies and the total inbreeding coefficient which are basic quantities to describe population structure. Thus the expected mating type frequencies as a function of gene frequencies and the inbreeding coefficient must fit well with the observed frequencies if deviations due to selection or misclassification are small. Theoretical discussion of two-allelic loci was made in section 2.5. The observations in northeastern Brazil provided autosomal cases for testing. Five codominant systems, Ss and MN blood group factors, Haptoglobin, Hemoglobin and Transferrin serum variants, and eight dominant systems, Diego, Lutheran, Inv, Duffy, Lewis, ABH secretor, P and PTC polymorphisms serve in this case. Six observed numbers of mating types and their expected numbers under the hypotheses that H_0: $\alpha=0$ and H_1: $\alpha=.0133$ are given in

Table 4.4.1 for five codominant loci in order to test whether the present theory is satisfactory or not but the improvement in goodness of fit is not striking.

Non-segregating probability h: Although it has been demonstrated theoretically that effects of inbreeding on segregation analysis are nearly negligible, the tendency of deviation in the observed non-segregating probability from the expected under the hypothesis that H_0: $\alpha=0$ should be tested with data. In a study of selection acting on sixteen polymorphisms in Brazilian material, Morton et al. (1966) have found that the discrepancies in h_2 and h_3 attributed to selection are infrequent. Since the observed deviations are ranging from about .3 to 10 percent except of h_3 in A_2 factor and since they are appeared to be within variations due to subdivision of population or inbreeding, it will be worthwhile to compare them with the expected deviations in the population with the inbreeding coefficient α = .0133. Correlation coefficient of the observed deviation, $e = U_h/K_{hh}$, and the expected, $\Delta h = c\alpha$, or a linear term of α in the expression of h as a function of gene frequencies and the inbreeding coefficient, are calculated (Table 4.4.3.). The high positive correlation means that the direction of discrepancies in the observed h can be attributed to an effect of inbreeding. The simple correlation coefficients are calculated weighting by the amount of information K_{hh}. In our Brazilian material from that seventeen factors are examined, the correlation coefficients for incompatible backcross, compatible backcross and incross with respect to a dominant phenotype are r = .35, .33 and .46, respectively. Although they are not significant from the null hypothesis that r = 0 ($P > .10$), the figures suggest on

effect of inbreeding.

The result again encourages us to justify the moment method to describe human population structure.

4.5. Migration function

Studies on the distribution of marital distance have contributed to human biology in two senses; to define the migration pattern in man and to obtain the relation between the inbreeding coefficient and the distance. Several probability density functions have been proposed to fit data which showed unescapably leptokurtotic patterns. In figure 4.5.1., relations among the proposed functions are summarized. In the following, each distribution shall be examined one by one in accordance with its relation to the others and goodness of fit with northeastern Brazilian population where a measured marital distance and the distance multiplied square root of population density are studied in grandparental and parental generations. Populations are subdivided into eight classes whose representative distance was calculated as a mean weighted by the number of couples in each group. In general, $m(x;a)$, for example, stands for a probability density function at distance x with an attraction parameter a, and the mean marital distance and its standard error are denoted by $\bar{x}$ and $\sigma_{\bar{x}}$, respectively. From these, the total inbreeding coefficient is tentatively calculated by the exponential relation between the inbreeding coefficient and the distance: $\alpha = f_0 k$, where

$$k = \int_0^\infty e^{-bx} m(x;a)dx,$$

$f_0 = .0212$ and $b = .0038$ in the distance study and $f_0 = .0199$ and $b = .0011$ for the distance multiplied by the square root of population density

so as to simulate a uniform density population. Whenever the integral is complicated, a discrete approximation is made such that $k \doteq \Sigma e^{-bx_i} n_i/n$, where x_i, n_i and n designate the representative value of distance in the i-th class, the expected number of couples in the i-th class and the total number of couples sampled. As far as the Brazilian population is concerned, errors due to the discrete approximation are of order 10^{-4} with respect to the inbreeding coefficient α, which may be tolerable in the present analysis. The inbreeding coefficient for the empirical observation is thus obtained by the discrete approximation, and the mean marital distance is estimated by the ordinary moment method.

Normal distribution: $m(x;a) = 2ae^{-(ax)^2}/\sqrt{\pi}$, $\bar{x} = .5642/a$ and $\sigma_{\bar{x}} = \bar{x}/\varepsilon\sqrt{K_{aa}}$. Wright's studies on isolation by distance have suggested a normal migration in man, which turned out to be unrealistic in practice. The function is also ruled out for the distribution of marital distance in man (Cavalli, 1958). As a reference hypothesis, we examined normal migration and found it not to fit (Table 4.5.1-4.). The technique employed here have also suggested that the normal function may not be applied to a marital distance so large that the expected proportion of couples is nearly zero. A grouping of the last large distance class was therefore made. Recently, Cavalli et al. (1965) have suggested a sum of normal probabilities $m(x; a_i) = \Sigma w_i 2a_i e^{-(a_i x)^2}/\sqrt{\pi}$ ($\Sigma w_i = 1$) in order to save Wright's theory to cover possible leptokurtosis. However, this meets two difficulties: first, experience tells us that functions with more than two parameters often do not converge and second, the linear combination of probabilities is only an empirical description and would not be a plausible model of migration.

Exponential distribution: $m(x;a) = ae^{-ax}$, $\bar{x} = 1/a$, $\sigma_{\bar{x}} = \bar{x}/a\sqrt{K_{aa}}$ and $k = a/(a+b)$. A suggested empirical function which is leptokurtic and has a mode near zero is an exponential distribution (Sutter and Tran-Ngoc-Toan, 1957). Because of its simplicity, the function has been used frequently in theory and in practice. Brazilian populations do not fit with it at all (Table 4.5.1-4.). This tempts us to generalize into two directions. One is to use a gamma distribution, which includes the exponential distribution as a special case (Cavalli, 1962). The gamma distribution therefore has two parameters, an attraction parameter and a dimension parameter n: $m(x;a,n) = a^n x^{n-1} e^{-ax}/n!$ An unrealistic point in this migration function is that it has a leptokurtic form only when $0 < n < 1$, which leads $m(x;a,n)$ to be infinite at $x=0$. Therefore, the function was not used.

Double exponential function: $m(x;a,b,p) = (1-p)ae^{-ax} + pbe^{-bx}$, $\bar{x} = (1-p)/a+p/b$. Another extension to save an exponential migration is to make a sum of exponential probabilities. As stated in the normal hypothesis, this approach is only for an empirical description of migration. Chi-squares for goodness of fit are considerably improved compared with previous analysis. Searches for trial values for parameters are, however, tedious, especially with p. The only suggestion which was helpful is to find estimates from graph. And if we take the point values of estimates, the couples consist of short migrant group (about 10 km. or 25 km. $\sqrt{\text{density}}$) and large migrant group (about 100 km. or 250 km. $\sqrt{\text{density}}$), where the short range proportion is 53 percent in grandparents and 46 percent in parents. The mean marital distance increased by generation and the estimated inbreeding coefficient decreased in the amount.

.001 or .1 percent by generation, where a stationary state is assumed with respect to the inbreeding function.

Square root exponential distribution: $m(x;a) = a^2 e^{-a\sqrt{x}}/2$, $\bar{x} = 6/a^2$ and $\sigma_{\bar{x}} = 2\bar{x}/a\sqrt{K_{aa}}$. This function was introduced by Cavalli (1958) to describe migration in the northern Italian population. Although the function does not fit well near zero distance, among functions with a single parameter it fits better than the normal and exponential functions. The present method of maximum likelihood turned out to converge extraordinarily slowly, usually taking more than sixty iteration starting from a tentative value calculated by the moment method. Occasionally, the goodness of fit with an intermediate estimate of parameter a was better than that with the final converged estimate. The function fits fairly well the northeastern Brazilian population. A generalization of the square root exponential function is to have a modified gamma distribution: $m(x;a,n) = a^n e^{-ax^{1/n}}/n!$ The function includes a normal distribution ($n = 1/2$) and an exponential distribution ($n=1$) as a special case. By transformation, $y=x^{1/n}$, it reduced to a gamma distribution. The attempt to fit to actual data remains a possibility.

Lognormal distribution: $m(x;a) = ae^{-(a\ln x)^2}/\sqrt{\pi}\, x$, $\bar{x} = \exp(1/4a^2)$ and $\sigma_{\bar{x}} = \bar{x}/4a^3\sqrt{K_{aa}}$. By a simple transform, $y = \ln x$, the function reduces to a normal probability. This implies that if human dispersion can be described by a lognormal function, then all of theories, which are based on normal migration, in isolation by distance will be saved. The fit of a lognormal distribution is better if nearly half of the sample falls in 0-1 unit class. Since northeastern Brazilian populations did not meet this criterion, it is no wonder that chi-squares for

goodness of fit have rather large values. Populations such as northern Italy (Cavalli, 1962) and northern Japan (Hiraizumi, 1965) where nearly half of couples have marital distance less than one km. class of marital distance may be well described by the lognormal distribution. This will be discussed elsewhere.

A Bessel distribution: $m(x;a) = a^2 x K_0(ax)$, $\bar{x} = 1.5708/a$ and $\sigma_{\bar{x}} = \bar{x}/e\sqrt{K_{aa}}$. This function has been suggested by Kimura (1963) from a purely theoretical point of view. Mathematical properties of $K_0(x)$ may be found in Lebedev (1965) but $K_0(x)$ is more leptokurtotic than the exponential and $K_0(x) \to \infty$, $xK_0(x) \to 0$ as $x \to 0$. The goodness of fit is worse than other single parameter distributions in Brazilian material (Table 4.5.1-4.).

Skellam distribution: $m(x;a) = 2ax/(1+x^2)^{a+1}$, $\bar{x} = \infty$ if $a < 1$. When individual mobility, or attraction parameter in our terminology, follows a gamma distribution, a normal migration is replaced by the Skellam distribution (Skellam, 1951). There are no finite moments unless a is greater than one. The fit of this function is as good as the square root exponential distribution, and better fits are observed with shorter distance (Table 4.5.1-4.). The meaning of parameter a is not clear in this case since no moments are available with such small values of a as .16, .12, .09 and .07 in four studies of our material. By a transformation, $y=x^2$, the distribution reduces to a beta probability.

A generalized Skellam distribution: $m(x;a,b) = 2abx/(1+ax^2)^{b+1}$, $\bar{x} = \infty$ if $b < .5$. An extension of Skellam distribution was made to separate the attraction and distribution parameters. However, this

generalization gave little improvement and the difficulty in estimating parameters is formidable. Convergence could not be obtained in the grandparent population. All b-values are about .2, indicating that no moments can be derived from the distribution parameters. A transformation, $y=x^2$, again introduces a beta distribution.

A beta distribution: $m(x;a,b) = ab/(1+ax)^{b+1}$, $\bar{x} = \infty$ if $b < 1$. Cavalli (1962) has suggested this type of function in the "gravitational" component of migration in man where an exploration range will be proportional to the inverse of some powers of distance. The parameters a and b may be called the attraction factor and dimensional index, where a is a scale parameter and b is dependent upon the range of exploration. This is the best function among those tested (Figure 4.5.2). [redacted] attraction factor is smaller in parental generation than in grandparental generation (.0916 ± .0089 → .0549 ± .0063 in the distance analysis and .0381 ± .0034 → .0236 ± .0028 in the distance multiplied by square root of population density) while the dimension index did not alter appreciably by generation (.8971 ± .0554 → .8765 ± .0627 in distance analysis and .7114 ± .0371 → .6667 ± .0458 in study of distance times square root of density). This result suggests that breakdown of isolates is occurring in northeastern Brazil, the migrants exploring for mates with dimension $b+1 \doteq 1.8$, or essentially two dimensional migration (Table 4.5.1-4.).

Since migration in man presumably has a dimension between one and two, the b-value would not exceed one so that no moments of distribution can be described in terms of a and b.

We thought from the beginning that the mean marital distance would

be explained by a decrease of inbreeding coefficient by generation, and that the parameters of a distribution which fitted well with data would give an unbiassed estimate of this decrease. The mean marital distance varies considerably for different types of migration function, but the estimated mean inbreeding coefficients do not vary much, which supports theoretical evidence that the inbreeding coefficient may be little related to the form of the migration function. The difference of .001 in the estimated inbreeding coefficients by generation is also almost constant. It is also observed that the mean inbreeding coefficient of population is roughly about eighty percent of that at zero distance for a variety of migration functions.

As a summary, a beta probability function was found best to describe distribution of marital distance in northeastern Brazil. Breakdown of isolates and dimension of human migration may be studied from the basis of distribution parameters, but not from mean marital distance. A double exponential probability was also fitted fairly well. Several other distribution functions which have been proposed were also examined.

4.6. Discussion

The bioassay method to ascertain the total inbreeding coefficient has shown several advantages in studies of population structure in northeastern Brazil. First of all, it requires only one generation data. Difficulties in tracing human pedigrees are common, so that the inbreeding coefficient is often underestimated even in an intensive pedigree survey. In bioassay, on the other hand, if data are collected by couple as a sampling unit, the investigator may estimate both the total

coefficient of consanguinity based on mating type frequencies and the total inbreeding coefficient of parents themselves treated as if they were collected as a random sample from the population. It is not necessary to study children, but they could serve as a confirmation of the bioassay method by comparing the inbreeding coefficient of children (G-TYPE) with the coefficient of consanguinity of parents (MATYPE). In addition, a comparison of two generations permits study of the breakdown of isolates which seems to be occurring in northeastern Brazil.

The total inbreeding coefficient thus obtained is due to all genetic barriers, and includes effects of consanguineous marriages, random genetic drift, mutation and selection. Effect of random genetic drift is theoretically known to increase homozygote frequencies but its effect is unknown in the present study. However, it might be small since our sample was from a tri-racial mixed population whose endogamy coefficient was estimated to account for only 22 percent of the total inbreeding coefficient. Mutation may be omitted from discussion at the present stage because of no mutation being observed in sixteen polymorphic systems.

Selection would be one of the sensitive factors, but technical error or misclassification often cannot be distinguished from it in statistical analysis. As is well-known, if selection has acted against homozygotes (heterosis), then the inbreeding coefficient estimated by bioassay tends to be small or even to be negative. This was almost ruled out in sixteen polymorphic systems from northeastern Brazil when positive values of the coefficient of consanguinity in parent and the inbreeding coefficient in children were obtained. A possibility of heterozygote disadvantage (negative heterosis) may enhance the inbreeding coefficient by

bioassay but no such evidences were found (Morton et al., 1966). Apparently, the effect of selection on the inbreeding coefficient is very small.

In connection with a negative inbreeding coefficient, the concept of "isolate size" should be concretely stated at this moment. In a finite population with size N, a correlation coefficient of two samples without replacement is always negative: $r = -1/(N-1)$ (See, for example, Wilks, 1962. p 217). When a finite population consisting of 2N gametes or haploids is in question, then the correlation coefficient of uniting gametes becomes $-1/(2N-1)$. The derivation of the correlation coefficient permits selfing and incest because of the assumption that no genetic barriers exist in the population. Genetic barriers in a population always give a non-negative correlation of uniting gametes (f_B) which may depend upon the population size. Therefore, the total inbreeding coefficient observed as a net must be

$$\alpha_T = -\frac{1}{2N-1} + f_B$$

which would be zero when $N = (1+f_B)/2f_B$ and also when $N \to \infty$ and $f_B = 0$.

The number $N_0 = 2N$ is called the isolate size in the sense of the probability density theory. In northeastern Brazil, $f_B \geq .0133$ has been observed (4.3.2.) as an amount of inbreeding and which gives $N_0 \leq 38$, assuming that f_B is independent from population size, and all genetic barriers are taken care in f_B. Subpopulation of size greater than this are expected to have positive inbreeding coefficients. This critical value of 38 is much smaller than Dahlberg's theory which is unsatisfactory not so much because of its implausibility, as because it leads to

no quantitative prediction of the relation between gene frequencies and genotype frequencies. The value N_0 means only that when the number of individuals in population is equal to it, the net inbreeding coefficient becomes zero. If the population size is smaller than N_0, then a negative correlation between uniting gametes is expected.

The alternative possibility for a zero inbreeding coefficient is a population in so-called "Hardy-Weinberg equilibrium" where the population size is infinite and neither mutation, selection, nor migration occurs. In an infinite population with genetic barriers, however, the correlation coefficient would be positive. An interesting problem is thus: what will be the inbreeding coefficient in practice when the population size becomes large? Three possibilities may be suggested: (1) the inbreeding coefficient increases with the population size since genetic barriers increase, (2) the inbreeding coefficient will be constant after a given population size and (3) the inbreeding coefficient will approach to zero when the size becomes infinite. Although the second possibility is most plausible, it remains for further research.

Wright (1943) has estimated the long-term inbreeding coefficient in man from his theory of isolation by distance as less than .02, which corresponds to 200 couples without mutation, selection, or long range migration. Although his approach assumed normal migration of parent-offspring distance, a remarkable consistency with our results which gave .017 in parent and .0133 in children revealed his great insight. Had he developed his theory of isolation by distance with another migration function such as an exponential or a beta distribution, and had he reached the same predicted estimate, this would be one of evidences that

migration function seemed not to be related to the total inbreeding coefficient of population. There are theoretical predictions on this by Malecot (1950) and Kimura et al. (1964), and our Brazilian material where almost all inbreeding coefficients estimated from 9 different distributions of marital distance gave very similar values (Table 4.5.). This is one of the most important aspects in human population genetics since the marital distance serves not only for study of migration and thus to evaluate the total inbreeding coefficient, but also for study of inbreeding by distance. We have tried to test both one and two dimensional models for the inbreeding function with distance but approximation seemed to be crude, especially in two dimensional function. Further researches are desirable.

In summary, the theory developed in chapter 2 has been applied successfully to a population from northeastern Brazil.

5. Summary

A new theory to describe human population structure was developed, based on Wahlund's principle. The possible consequences from the theory were tested by using a method of maximum likelihood scoring with 1068 migrant families from northeastern Brazil.

The following facts were predicted in theory and emerged from the analysis:

1. Mating type frequencies are given as a function of gene frequencies and the inbreeding coefficient, provided that the smallest gene frequencies is not less than the inbreeding coefficient. As far as autosomal loci with two alleles are concerned, this gave a better fit to the Brazilian population than the assumption of Hardy-Weinberg mating type

frequencies.

2. The effect of inbreeding, or subdivision of population, on segregation analysis was examined theoretically and with 17 human polymorphic factors. No serious effects were found of assuming Hardy-Weinberg equilibrium in segregation analysis when the inbreeding coefficient is smaller than the smallest gene frequency.

3. The relation between the inbreeding coefficient and marital distance describing genetic isolation in populations was leptokurtic and approximately exponential for the total and remote inbreeding coefficients and the remote coefficient ascertained by pedigree study. The deviation from exponential in the ascertained inbreeding coefficient was apparently due to preferential consanguineous mating at large distances.

4. Four alternative methods to estimate the inbreeding coefficient were applied: pedigree study which showed .0059, bioassays from individual phenotype frequencies and mating type frequencies, .0170 and .0133, respectively, and use of migration function, .018.

5. The breakdown of isolates was measured in term of the inbreeding coefficient $f_B = .095$, taking ancestral ethnic populations as a reference.

6. A method to estimate the inbreeding coefficient for alleles was devised.

7. The ABO blood group system does not give any information about the inbreeding coefficient by bioassay method with individual data.

The following results were from the analysis.

8. The pedigree study of the inbreeding coefficient resulted in higher estimate in offspring due to incomplete ascertainment in the previous generation.

9. Comparison of the inbreeding coefficients by pedigree study and by bioassay with sixteen polymorphic systems indicated that the remote inbreeding coefficient was as great as the close inbreeding coefficient.

10. No differentiation of gene frequencies with respect to marital distance was observed.

11. Racial endogamy contributed only 22 percent of the total inbreeding coefficient for polymorphisms, and less for rare alleles or monomorphisms.

12. A tendency for the inbreeding coefficient to decrease by generation could not be detected through pedigree analysis, but is found by the new methods of bioassay and distribution of marital distance.

Further researches are needed in the following aspects.

13. Theoretical works on the inbreeding function with marital distance, taking care to exclude self-fertilization and incest.

14. An empirical correlation method with distance instead of bioassay for describing inbreeding functions.

15. Study in other human populations than Brazil should be conducted for bioassay of the inbreeding coefficient. On this occasion, an accurate record of location (for example, the longitude and the latitude, of birthplace of individual) must be made in order to extract more information on human population structure.

16. Some mathematical models might be checked with other species in the laboratory. A homogeneous symmetrical migration population structure, for instance, could be tested with Drosophila melanogaster.

17. There are still several unexplored possibilities to examine the present theory. Mating type frequencies at sex-linked loci are waiting for the applications and the moment theory itself may be applied to other

aspects such as linkage, illegitimacy, polyploidy, heritability and so on, which appear in genetic aspects of human biology.

Appendix 1. General discussion on the breakdown of isolates

The conclusion in text is not altered when we consider more than three isolates. Although several models may be developed, we shall discuss only two of them: (i) the breakdown of isolates in a part of the total population, and (ii) hierarchic model of removing of barriers.

(i). Suppose that a population consisted of n isolates and k of n ($k \leq n$) isolates were grouped into a new panmictic isolate by removing of barriers so that the population consists of n-(k-1) isolates.

At the first phase, the population is characterized by:

$$\sum_{i=1}^{n} w_i = 1,$$

$$p_N = \sum_{i=1}^{n} p_i w_i,$$

and

$$\sigma_N^2 = \sum_{i=1}^{n} p_i^2 w_i - p_N^2,$$

where w_i and p_i are the relative size and gene frequency of the i-th isolate, respectively, and also p_N and σ_N^2 are a gene frequency and its variance in the total population, respectively. Let us take the first k isolates being grouped. The present population is now specified by

$$W + \sum_{i=K+1}^{n} w_i = 1,$$

$$p_{N-K} = PW + \sum_{i=K+1}^{n} p_i w_i,$$

and

$$\sigma_{N-K}^2 = P^2 W + \sum_{i=K+1}^{n} p_i^2 w_i - p_{N-K}^2,$$

where $W = \sum_{i=1}^{K} w_i$, $P = \sum_{i=1}^{K} p_i w_i / W$ and p_{N-K} and σ_{N-K}^2 stand for gene frequency and its variance in the present population, respectively.

Comparison of two phases results in

$$p_N = p_{N-K}(\equiv p),$$

and
$$\sigma^2_{N-K} = P^2W + \sum_{i=K+1}^{n} p_i^2 w_i - p^2_{N-K}$$

$$= \sigma^2_N - \left[\left(\sum_{i=1}^{K} w_i\right)\left(\sum_{i=1}^{K} p_i^2 w_i\right) - \left(\sum_{i=1}^{K} p_i w_i\right)^2 \right] \Big/ \sum_{i=1}^{K} w_i$$

or

$$\sigma^2_{N-K} = \sigma^2_N - \sum_{i<j}^{K} w_i w_j (p_i - p_j)^2 \Big/ \sum_{i=1}^{K} w_i .$$

The relation in the inbreeding coefficient will be

$$F_{N-K} = F_N - F_B,$$

where $F_B = \sum_{i<j}^{K} \frac{(w_i w_j)}{W} \frac{(p_i - p_j)^2}{p(1-p)}$

(ii). Although the breakdown of isolates has occurred in several parts of a population, some of barriers still remain so that the population consists of a number of new isolates. Taking the same notations of (i) but X to be the number of new isolates, we can easily verify the relation in the inbreeding coefficient between two phases of the population: $F_{N-\ldots-X} = F_N - F_B$, where $F_{N-\ldots-X}$ stands for the inbreeding coefficient at the second phase and

$$F_B = \Sigma \sum_{i<j} \left(\frac{w_i w_j}{W}\right)(p_i - p_j)^2 / p(1-p).$$

The first summation is taken for new isolates.

Example. The comparison of the endogamy coefficient in three racial ancestral populations with the inbreeding coefficient from a tri-racial mixture population. Suppose that three racial groups are Indian, Negro and Caucasian whose relative sizes and gene A frequencies are I, N and C, and p_i, p_n and p_c, respectively. Let F_M and F_E be the inbreeding coefficient of the tri-racial mixture population and the endogamy coefficient in the ancestral populations. We obtain

$$F_M = F_E - F_B,$$

where
$$F_B = \frac{IC(p_i-p_c)^2 + CN(p_c-p_n)^2 + NI(p_n-p_i)^2}{p(1-p)}$$

since $p = Ip_i + Np_n + Cp_c$ and $I + N + C = 1$.

Consideration of several genes will provide more information for F_B.

(In the above discussion, we assumed no mutation, selection, and accident by sampling which may result in changing of the mean gene frequency of population. Since it is always difficult to estimate gene frequencies in ancestral populations, F_B may be, therefore, taken as the first approximation. Examining the mean gene frequency of population at different stages, we may justify this method if difference is not significant from zero.)

Appendix 2. Moments of a subdivided population given a distribution of isolate size

A general idea to obtain the moments, knowing a distribution function of isolates, would be demonstrated by beta probability in a locus with two alleles since this case represented a steady state distribution of gene frequency under Wright's island model, where the population consisted of isolates of equal size and each isolate exchanged constantly individuals into the neighbors. Only the third and fourth moments will be given in other distributions for the sake of comparison. For the higher moments, a method of moment generating function will be helpful.

(i). Beta distribution: Suppose that a density function is given by

$$dW = \frac{(a+b-1)!}{(a-1)!(b-1)!} p_w^{a-1} q_w^{b-1} dp_w \qquad (p_w + q_w = 1),$$

where a and b are distribution parameters and p_w is the gene frequency in the neighborhood of the point w. The moment of population will be

$$M_k = \int p_w^k dW$$

$$= \frac{(a+b-1)!}{(a-1)!(b-1)!} \int_0^1 p_w^{a+k-1}(1-p_w)^{b-1} dp_w$$

$$= \frac{(a+b-1)!(a+k-1)!}{(a-1)!(a+b+k-1)!} \qquad (k = 0, 1, 2, \ldots)$$

which gives

$$M_1 = a/(a+b) \text{ and } M_2 = M_1(a+1)/(a+b+1).$$

Since the first and second moments correspond to the population gene frequency p and an homozygous frequency $p^2 + pq\alpha$, where α is the inbreeding coefficient, the parameters may be written in terms of gene frequencies and the inbreeding coefficient, or

$$a/(a+b) = p$$

$$(a+1)/(a+b+1) = p + q\alpha$$

so that $a = p(1-\alpha)/\alpha$ and $b = q(1-\alpha)/\alpha$. We obtain, therefore,

$$M_3 = \frac{1}{(1+\alpha)}\,[p^3 + p^2(1+2q)\alpha + pq(1+q)\alpha^2]$$

and

$$M_4 = \frac{1}{(1+\alpha)(1+2\alpha)}\,[p^4 + 3p^3(1+q)\alpha + p^2(2+6q+3q^2)\alpha^2 + pq(1+q)(2+q)\alpha^3],$$

or, if α is small (say, less than two per cent),

$$M_3 = p^3 + 3p^2q\alpha - pq(1+3q)\alpha^2 + 2q(1-2q)\alpha^3 + \ldots$$

and

$$M_4 = p^4 + 6p^3q\alpha - p^2q(8-19q)\alpha^2 + 2pq(7-27q+23q^2)\alpha^3 + \ldots$$

In the island model, $a=4Nmp$ and $b=4Nmq$, where N is the effective size of isolates and m is the migration rate, so that $\alpha = 1/(1+4Nm)$. (Wright, 1931).

The following results are straightforward. (The form of distribution function may be found in Mood and Graybill, 1963).

(ii). Binomial distribution:

$$M_3 = p^3 + 3p^2q\alpha - pq(1-2q)\alpha^2$$

$$M_4 = p^4 + 6p^3q\alpha - p^2q(4-11q)\alpha^2 + pq(1-6q+6q^2)\alpha^3.$$

(iii). Poisson distribution:

$$M_3 = p^3 + 3p^2q\alpha + pq^2\alpha^2$$
$$M_4 = p^4 + 6p^3q\alpha + 7p^2q^2\alpha^2 + pq^3\alpha^3.$$

(iv). Rectagular distribution:

$$M_3 = p^3 + 3p^2q\alpha$$
$$M_4 = p^4 + 6p^3q\alpha + (9/5)p^2q^2\alpha^2.$$

(v). Normal distribution:

$$M_3 = p^3 + 3p^2q\alpha$$
$$M_4 = p^4 + 6p^3q\alpha + 3p^2q^2\alpha^2.$$

(vi). Gamma distribution: $(p \geq q)$

$$M_3 = p^3 + 3p^2q\alpha + 2pq^2\alpha^2$$
$$M_4 = p^4 + 6p^3q\alpha + 11p^2q^2\alpha^2 + 6pq^3\alpha^3.$$

Thus the square and higher powers of α may be ignored when $|\alpha|$ is not greater than the smallest gene frequency.

Appendix 3. Derivation of a general formula for the moment of population

When the number of alleles increases beyond two, the covariance moments become of importance, which are given in the Wahlund's principle as frequencies of heterozygotes.

Let p and q be gene frequencies of population and p_w and q_w be of an isolate (it is not required that p+q=1 and $p_w+q_w=1$). Denoting differences in gene frequencies between the isolate and the population by Δp_w and Δq_w, their covariance moment is given by

$$m_{ij} = \int(\Delta p_w)^i(\Delta q_w)^j dW = E(\Delta p_w)^i(\Delta q_w)^j,$$

where E is an operational symbol denoting expectation. For example,

$$m_{10} = m_{01} = 0,$$
$$m_{20} = p(1-p)\alpha,\ m_{11} = -pq\alpha,\ \text{and}\ m_{02} = q(1-q)\alpha,$$

where α is the inbreeding coefficient. The moment of population is now

$$\begin{aligned} M_{a,b} &= E(p_w^a q_w^b) \\ &= E(p+\Delta p_w)^a(q+\Delta q_w)^b \\ &= \sum_{r,s}\binom{a}{r}\binom{b}{s}p^{a-r}q^{b-s}m_{rs} \\ &= p^a q^b + ap^{a-1}q^b m_{10} + bp^a q^{b-1}m_{01} + \binom{a}{2}p^{a-2}q^b m_{20} + \\ &\quad abp^{a-1}q^{b-1}m_{11} + \binom{b}{2}p^a q^{b-2}m_{02} + \cdots\cdots \end{aligned}$$

or, ignoring the higher terms of m_{rs} ($r+s>2$, $r\geq 0$ and $s\geq 0$),

$$M_{a,b} = p^a q^b + \left[\frac{a(a-1)}{2}p^{a-1}(1-p)q^b + \frac{b(b-1)}{2}p^a q^{b-1}(1-q) - abp^a q^b\right]\alpha.$$

Justification to ignore the higher moments, m_{rs}, is also seen in a series of calculations of moments assuming distribution functions.

The sufficient condition is again that the smallest gene frequency is greater than the inbreeding coefficient.

Generalization is now straightforward. The moment of population is

$$M_{a_1,\ldots,a_n} = E\left[\prod_{i=1}^{n}(p_i + \Delta p_{iw})^{a_i}\right],$$

where p_i is the frequency of i-th allele. By expansion of binomial product terms and by replacing m_{rs} in terms of gene frequencies and the inbreeding coefficient, we obtain

$$M_{a_1,\ldots,a_n} = \prod_{i=1}^{n} p_i^{a_i} + \prod_{i=1}^{n} p_i^{a_i-1}\left[\sum_{i=1}^{n}\binom{a_i}{2}(1-p_i)\prod_{i>j}^{n} p_j^{*} - \left(\prod_{i=1}^{n} p_i\right)\sum_{i<j} a_i a_j\right]\alpha.$$

The assumptions for deriving the general formula are the same as the previous arguments.

Appendix 4. The application of Newton-Raphson method to solve maximum likelihood equation (so called maximum likelihood scoring method) and its convergency

Let L be a log-likelihood with a parameter θ, and take its first derivative with respect to θ as $u(\theta) = \partial L/\partial\theta$. Furthermore let θ_0 denote an approximate value of the true one, and e be the correction which must be applied to θ_0 to give the exact value of the solution, so that

$$\theta = \theta_0 + e.$$

The maximum likelihood equation $u(\theta) = 0$ then becomes

$$u(\theta_0 + e) = 0.$$

Expanding this in the series form by Taylor's theorem, we have

$$u(\theta_0+e) = u(\theta_0) + e\,u'(\theta_0) + \frac{e^2}{2}u''(\theta_0+\delta e) \qquad (0 \leq \delta \leq 1)$$

Hence

$$u(\theta_0) - e\,k(\theta_0) - \frac{e^2}{2}k'(\theta_0+\delta e) = 0 \qquad (1)$$

where $k(\theta_0) = -(\partial u/\partial\theta)_{\theta=\theta_0}$.

Now if e is relatively small, we may neglect the term containing e^2 and get the simple relation

$$u(\theta_0) - e_1 k(\theta_0) = 0$$

from which

$$e_1 = u(\theta_0)/k(\theta_0) \qquad (2).$$

The improved value of the root is then

$$\theta_1 = \theta_0 + e_1 = \theta_0 + \frac{u(\theta_0)}{k(\theta_0)} \qquad (3).$$

The succeeding value of the root is

$$\theta_2 = \theta_1 + e_2 = \theta_1 + \frac{u(\theta_1)}{k(\theta_1)}$$

$$\theta_3 = \theta_2 + u(\theta_2)/k(\theta_2)$$

.....

$$\theta_n = \theta_{n-1} + u(\theta_{n-1})/k(\theta_{n-1}).$$

Equation (3) is the fundamental formula in the maximum likelihood scoring method.

It is evident from this formula that the larger the k-score, the smaller is the correction which must be applied to get the true value of the estimate. This means that when the graph of likelihood equation is nearly vertical where it crosses the θ-axis the correct value of the root can be found with great rapidity and very little labor. If, on the other hand, the numerical value of the k-score should be small in the neighborhood of the root, the values of e given by (2) would be large and the computation of the root by this method would be a slow process or might even fail altogether. This method should never be used when the graph of $u(\theta)$ is nearly horizontal where it crosses the θ-axis. In such cases, regula falsi interpolation might be useful (Barrai et al., 1965). The process will evidently fail if $k(\theta)=0$ at the neighborhood of the root and such an example in bioassay has been seen at the ABO blood group system (see text 3.3.1.).

In the above process we neglected the term involving e^2 and got an approximate value e_1 from the equation $u(\theta_0)=e_1 k(\theta_0)$. Subtracting it from (1), we obtain

$$(e-e_1)k(\theta_0) + \frac{e^2}{2} k'(\theta_0 + \delta e) = 0$$

or

$$e-e_1 = -e^2 \frac{k'(\theta_0+\delta e)}{2k(\theta_0)} \qquad (= E_1) \qquad (4)$$

Now since e is the true value of the required correction, and e_1 is its approximate value, it is plain that E_1 is the error in e_1. Let M be the maximum value of $k'(\theta_0)$ at the neighborhood of θ_0+e_1, then

$$e - e_1 = - \frac{e^2 M}{2k(\theta_0)}$$

or $Me^2 + 2k(\theta_0)e - 2k(\theta_0)e_1 = 0$. Solving it,

$$e = (-k(\theta_0) + \sqrt{[k(\theta_0)]^2 + 2Mk(\theta_0)\, e_1}\,)/M$$

so that

$$E_1 \leq \left| \frac{Me^2}{2k(\theta_0)} \right|$$

This is error in θ_1. In general the error in θ_n is therefore

$$E_n \leq \left| \frac{Me_n^2}{2k(\theta_{n-1})} \right|.$$

If $|M/2k(\theta_n)| \leq 1$, as usual in cases where the Newton-Raphson method can be applied, we then have

$$E_n \leq e_n^2.$$

This result is most important, for in finding the correction from (2), the division of $u(\theta_0)$ by $k(\theta_0)$ need to be carried out to only one more significant figure than number of zero between the decimal point and first significant figure.

Appendix 5. Maximum likelihood estimation of gene frequencies and the inbreeding coefficient from individual frequency data: <u>In case of k-alleles without dominance</u>

Let p_i ($\sum_{i=1}^{k} p_i = 1$) be the frequency of allele A_i and n_{ij} ($=n_{ji}$) be the observed number of an ordered genotype A_iA_j ($=A_jA_i$) whose frequency is $p_ip_j + p_i(\delta_{ij} - p_j)\alpha$, where $\delta_{ij}=1$ for $i=j$ and $\delta_{ij}=0$ for $i\neq j$. The log likelihood is

$$L = \sum\sum_{ij} n_{ij} \ln[p_ip_j + p_i(\delta_{ij}-p_j)\alpha]$$

and the scores are

$$U^*_{p_i} = \frac{2p_i + (1-2p_i)\alpha}{p_i^2 + p_i(1-p_i)\alpha} n_{ii} + \frac{1}{p_i} \sum_{j\neq i} (n_{ij}+n_{ji}) \qquad (i=1, \ldots, k)$$

$$U^*_{\alpha} = \sum_{i=1}^{k} \frac{(1-p_i)}{p_i+(1-p_i)\alpha} n_{ii} - \frac{1}{1-\alpha} \sum_{j\neq i} (n_{ij}+n_{ji})$$

Imposing the restriction that $p_k = 1 - \sum_{i=1}^{k-1} p_i$, the maximum likelihood scores are

$$U_{p_i} = U^*_{p_i} - U^*_{p_k} \qquad U_{\alpha} = U^*_{\alpha}.$$

The variances of U-scores are

$$K_{p_ip_j} = K^*_{p_ip_j} - K^*_{p_ip_k} - K^*_{p_jp_k} + K^*_{p_kp_k} \qquad (i,j = 1, \ldots, k-1)$$

$$K_{p_i\alpha} = K^*_{p_i\alpha} - K^*_{p_k\alpha}$$

$$K_{\alpha\alpha} = K^*_{\alpha\alpha},$$

where

$$K^*_{p_i p_i} = \left[\frac{2p_i+(1-2p_i)\alpha}{p_i^2+p_i(1-p_i)\alpha}\right]^2 n_{ii} + \frac{1}{p_i^2}\sum_{j\neq i}(n_{ij}+n_{ji})$$

$$K^*_{p_i p_j} = \frac{1}{p_i p_j}(n_{ij}+n_{ji}) \qquad (i\neq j)$$

$$K^*_{p_i \alpha} = \left[\frac{1-p_i}{p_i+(1-p_i)\alpha}\right]\left[\frac{2p_i+(1-2p_i)\alpha}{p_i^2+p_i(1-p_i)\alpha}\right] n_{ii} - \frac{1}{p_i(1-\alpha)}\sum_{j\neq i}(n_{ij}+n_{ji})$$

$$K^*_{\alpha\alpha} = \sum_{i=1}^{k}\left[\frac{1-p_i}{p_i+(1-p_i)\alpha}\right]^2 n_{ii} + \left[\frac{1}{1-\alpha}\right]^2 \sum_{j\neq i}(n_{ij}+n_{ji})$$

Under the null hypothesis that $\alpha=0$, the estimates of gene frequencies are obtained from the likelihood equations $U_{p_i}=0$ for i=1, ..., k-1: since

$$\frac{2n_{ii} + \sum_{i\neq j}(n_{ij}+n_{ji})}{p_i} = \frac{2n_{kk} + \sum_{j\neq k}(n_{kj}+n_{jk})}{p_k}$$

for i=1, ..., k-1, there should exist, therefore, a constant, C, such that

$$2n_{ii} + \sum_{j\neq i} n_{ij} = p_i C \qquad \text{for all } i.$$

Adding for i,

$$\sum_{i=1}^{k}[2n_{ii} + \sum_{j\neq i} n_{ij}] = C\sum_{i=1}^{k} p_i$$

so that $C = 2[\sum_{i=1}^{k} n_{ii} + \sum_{i>j} n_{ij}](n_{ij}=n_{ji})$, and which is equal to

$2\Sigma\Sigma n_{ij}=2N$ or the total number of genes. Therefore

$$p_i = (2n_{ii} + \sum_{j\neq i} n_{ij})/2N.$$

The variances of scores thus become $K_{p_ip_i}=2N\left[\frac{1}{p_i} + \frac{1}{p_k}\right]$, $K_{p_ip_j} = \frac{2N}{p_k}$,

$K_{p_i\alpha} = 0$ and $K_{\alpha\alpha} = (k-1)N$ since

$$K^*_{p_i\alpha} = \left[\frac{1-p_i}{p_i^2}\right]Np_i^2 - \frac{1}{p_i}\sum_{j\neq i} Np_ip_j$$

$$= N[(1-p_i) - \sum_{j\neq i} p_j]$$

$$= N(1- \sum_{j=1}^{k} p_j) = 0,$$

and

$$K_{\alpha\alpha} = \sum_{i=1}^{k} (\frac{1-p_i}{p_i})^2 Np_i^2 + \sum_{j\neq i} p_ip_jN$$

$$= N\left[\sum_{i=1}^{k} (1-p_i)^2 + \sum_{j\neq i} p_ip_j\right]$$

$$= N\left[\Sigma(1-2p_i+p_i^2) + \Sigma p_ip_j\right]$$

$$= N\left[k-2 + \sum_{i=1}^{k} p_i^2 + \sum_{j\neq i} p_ip_j\right]$$

$$= N\left[k-2 + (\sum_{i=1}^{k} p_i)^2\right] = N(k-1).$$

The improved estimate of α is calculated from

$$\alpha_1 = \alpha_0 + U_{\alpha\alpha}/K_{\alpha\alpha} \qquad (\alpha_0 = 0)$$

The iteration is then carried out until the estimate converges. It has been observed that the final estimate is always obtained after three or four iterations.

Appendix 6. Instability of estimating the inbreeding coefficient at the A_1A_2BO blood group system

Let p_1, p_2, q and r be frequencies of gene A_1, A_2, B and O, respectively. Six phenotypes are observed so that their frequencies in population with inbred proportion α are given by

$$\begin{aligned}
P(O) &= r^2 + r(1-r)\alpha \\
P(A_1) &= p_1^2 + 2p_1p_2 + 2p_1r + p_1(1-p_1-2p_2-2r)\alpha \\
P(A_2) &= p_2^2 + 2p_2r + p_2(1-p_2-2r)\alpha \\
P(B) &= q^2 + 2qr + q(1-q-2r)\alpha \\
P(A_1B) &= 2p_1q - 2p_1q\alpha \\
P(A_2B) &= 2p_2q - 2p_2q\alpha .
\end{aligned}$$

The log likelihood is

$$\begin{aligned}
L = {} & O\cdot\ln[r^2+r(1-r)\alpha] + A_1\cdot\ln[p_1^2+2p_1p_2+2p_1r+p_1(1-p_1-2p_2-2r)\alpha] + \\
& A_2\cdot\ln[p_2^2+2p_2r+p_2(1-p_2-2r)\alpha] + B\cdot\ln[q^2+2qr+q(1-q-2r)\alpha] + \\
& A_1B\cdot\ln[2p_1q-2p_1q\alpha] + A_2B\cdot\ln[2p_2q-2p_2q\alpha]
\end{aligned}$$

and the maximum likelihood scores under the null hypothesis that $\alpha=0$ are, in vector forms,

$$U_{p_1} = \left(-\frac{2}{r}, \frac{2p_2+2r}{p_1^2+2p_1p_2+2p_1r}, \frac{-2}{p_2+2r}, \frac{-2}{q+2r}, \frac{1}{p_1}, 0\right)n^1$$

$$U_{p_2} = \left(-\frac{2}{r}, 0, \frac{2r}{p_2^2+2p_2r}, \frac{-2}{q+2r}, 0, \frac{1}{p_2}\right)n^1$$

$$U_q = \left(-\frac{2}{r}, \frac{-2}{p_1+2p_2+2r}, \frac{-2}{p_2+2r}, \frac{2r}{q^2+2qr}, \frac{1}{q}, \frac{1}{q}\right)n^1$$

$$U_\alpha = \left(\frac{1-r}{r}, \frac{1-p_1-2p_2-2r}{p_1+2p_2+2r}, \frac{1-p_2-2r}{p_2+2r}, \frac{1-q-2r}{q+2r}, -1, -1\right)n^1,$$

where n^1 is a transpose vector of observation n such that;

$$n = (0, A_1, A_2, B, A_1B, A_2B).$$

It is easily verified that

$$U_\alpha = -\frac{p_1}{2}U_{p_1} - \frac{p_2}{2}U_{p_2} - \frac{q}{2}U_q .$$

Appendix 7.

PROGRAM G-TYPE

I. Instruction for user

G-TYPE is written in FORTRAN IV language for the IBM 7040 computer, and is designated to estimate gene frequencies and the inbreeding coefficient from phenotype data on a given genetic system by the maximum likelihood scoring method. It may handle any genetic system in which gene-genotype relation can be expressed by factor union algebra and consists of up to 24 alleles, 10 factors and 300 phenotypes.

CONTROL CARDS

Card 5 — This type of card is used for control of the null hypothesis that $\alpha=0$.

Col. 1-5	55555	
6	blank	
7	1	only one iteration
	0	otherwise
8	blank	
9	1	no estimation of gene frequency
	0	otherwise

Card 0 — This type of card is used for data description.

Col. 1-5	00000
6-10	blanks
11-70	description of data

Card I

Col. 1-5	11111
6-10	blanks
11-12	total number of alleles
13-14	total number of factors
15-17	total number of phenotypes

Card II — It is used for characterization of alleles, one card for each allele. Total number of cards must be equal to the number given on Card I, column 11-12. (The First allele is treated as dependent variable in the process.)

Col.	1-5	22222
	6-7	order of allele
	8-10	blanks
	11-15	common name of allele
	16-25	allele in binary code
	26-30	trial value of gene frequency
	31	iteration index

0 or blank: iteration is desired

1: in the first allele, the biological indicator (α) is set to zero (i.e., random mating).

Card III — This type of card(s) is used as input data, one card for each phenotype. Total number of cards must be equal to the number given on Card I, column 15-17; that is always less than 300.

Col.	1-5	33333
	6-8	order of phenotype, if necessary
	9-10	blanks
	11-15	common name of phenotype
	16-25	phenotype in binary code
	26-35	observed phenotype frequency (in observed number)

Card IV — This card, a trailer, is used only at the end of a job. For multiple-runs, card III should be followed immediately by a new set of controls.

Col. 1-80 9

Output from the program includes the following:

1. Trial value of gene frequency (given)
2. Phenotype frequency
   a) observed number (given)
   b) expected
   c) corresponding χ^2 value for goodness of fit and L-ratio.
3. Number of iteration cycles performed
4. Final χ^2-value for maximum likelihood estimation of parameters
5. Log-determinant of information matrix (base 10)
6. Information matrix (or K-matrix)
7. Covariance matrix (or inverse matrix)
8. Maximum likelihood estimate (M.L.E.)(error in 10^{-4}), standard deviation and U-scores for each parameter under iteration, the amount of information and the chi-square for parameter.

These quantities except M.L.E. are evaluated after convergence and the maximum number of iteration is 99.

Estimated maximum time is about 1 minute for a system.

II. A worked example

An example is to test the null hypothesis that $\alpha=0$, having maximum likelihood estimates of gene frequencies, at the MNsSU blood group system (In the following, b indicates a blank.).

III. Input control cards

column 1 in a card

```
55555blbbbO
OOOOObbbbbMNS$U SYSTEM, BRAZILIAN SEROTYPE(TOTAL)
11111     0604012
22222blbbbMSbbbOllObbbbbbO19701
22222 2   M$    0101      03451
22222 3   M*    0100      00091
22222 4   NS    1010      00762
22222 5   N$    1001      03485
22222 6   N*    1000      00241          <- column 35
33333bbl  M*    0100bbbbbbbbbbbbbbbbl
33333  2  MS    0110               93
33333  3  MS$   0111              298
33333  4  M$    0101              256
33333  5  MN*   1100                0
33333  6  MNS   1110               68
33333  7  MNS$  1111              410
33333  8  MN$   1101              572
33333  9  N*    1000                4
33333 10  NS    1010               22
33333 11  NS$   1011              123
33333 12  N$    1001              281
```

IV. Notes for preparation of input control cards

1. $ stands for s since computer does not distinguish between capital and small letters.

2. Any error in control cards will be printed out in the following message: 11111, 22222 or 33333

   ERROR IN CONTROL CARD ↓, EXECUTION DELETED.

   And the program goes to the next job if multiple-runs are made. Otherwise, it will stop. However, error in binary codes is

not always detected by machine so that a careful check is important with output, especially the observed number.

3. When convergence fails, either a message;

SINGULAR MATRIX

is printed out and the program skips to the next job, or all output is printed out with the iteration number 99.

4. Order of genes and of phenotypes in the control cards, 22222 and 33333, is arbitrary. There is no dependence between cards II and III.

5. In the above example, if column 7 in the first card, 55555, is punched in 0 or blank instead of 1 then ordinary iterations will be continued until improved estimates are obtained.

V. Output

MNS$U SYSTEM, BRAZILIAN SEROTYPE(TOTAL)

MS = .1970
M$ = .3451
M* = .0091
NS = .0762
N$ = .3485
N* = .0241

NO.	PHENOTYPE	FREQ.	OBS.	EXP.	CHI-2	L-RATIO
1	MS	0.04239	93.	90.22	0.09	5.65
2	MS$	0.13597	298.	289.34	0.26	17.57
3	MNS	0.04091	68.	87.05	4.17	-33.58
4	MNS$	0.18990	410.	404.11	0.09	11.86
5	M$	0.12537	256.	266.80	0.44	-21.15
6	MN$	0.26351	572.	560.75	0.23	22.72
7	N*	0.00008	1.	0.18	3.85	3.47
8	MN*	0.00044	0.	0.93	0.93	0.00
9	NS	0.00948	22.	20.17	0.17	3.82
10	NS$	0.05311	123.	113.02	0.88	20.81
11	N$	0.13825	281.	294.20	0.59	-25.79
12	N*	0.00058	4.	1.24	6.18	9.40
13	TOTAL	1.00000	2128.	2128.00	17.86	14.78

ITERATION NO. = 2 CHI-SQUARE = 6.25

LOG(10) DETERMINANT = .026303599E 02

K-MATRIX

...

INVERSE MATRIX

...

	ESTIMATE	ST.DEV.	U-SCORE	INFORMATION
ALPHA	.03712	.01485	169.8659	4536.4402
MS	.19700	.00698		
M$	.34510	.01021	1.3815	
M*	.00910	.00803	4.4823	
NS	.07620	.00545	-0.1118	
N$	.34850	.00975	0.7303	
N*	.02410	.00750	0.2939	

VI. Notes for output

1. L-ratio shows whether the observed is larger or smaller than the expected. If it is positive, the observed is larger than the expected and be it negative, then the observed is smaller. The sum of L-ratios converges to the chi-square value that is obtainable from a goodness of fit test. In the program, L=0, when no observation is in a class.
2. The chi-square for goodness of fit is 17.86 in the example, and 6.25 for parameters.
3. In the example, gene MS was considered as a dependent variable but one of the other might be taken.
4. In case of two allele with complete dominance, the value, 9.00000, will appear as an estimate of ALPHA, inbreeding coeffi-

cient. No such provision was made for the ABO-type dominance.

5. When instruction for no iteration was punched on card, as in the example, ITERATION NO. output is always set to 2. Otherwise, it indicates correct number of iteration processes.

Appendix 8.

PROGRAM MATYPE

I. Instruction for user

MATYPE is a FORTRAN IV program, designed to estimate gene frequencies and the inbreeding coefficient from mating type data on a given genetic system by the maximum likelihood scoring method. It can handle any genetic system in which gene-genotype relations may be expressed by factor union algebra and consists of up to 15 alleles, 10 factors and 36 phenotypes (corresponding to 666 phenotypic-mating types).

CONTROL CARDS

Card 5	Col. 1-5	55555
	6-7	1 is no iteration 0 or blank, otherwise
	11-21	Initial value of α
Card 0	This type of card is used for data description.	
	Col. 1-5	00000
	6-10	blanks
	11-70	description of data
Card I	Col. 1-5	11111
	6-10	blanks
	11-13	Total number of phenotypic mating types observed (plus in-cross)
	14-15	Total number of alleles
	16-17	Total number of factors
	18-20	Total number of phenotypes
Card II	It is used for characterization of alleles, one	

card for each allele. Total number of cards must be equal to the number given on Card I, column 14-15. (The first allele is treated as dependent variable in the process).

Col. 1-5 22222

6-7 Order of allele

8-10 blanks

11-15 Common name of allele

16-25 Allele in binary code

26-30 Trial value of gene frequency

31 Iteration index

0 or blank: iteration is desired

1: in the first allele, the biological indicator is set to zero (i.e. random mating).

Card III

This type of card(s) is used as input data, one card for each pheno-mating type. Total number of cards must be equal to the number given on Card I, column 11-13; that is always less than or equal to 666.

Col. 1-5 33333

6-8 Order of pheno-mating type

9-10 blanks

11-15 If both mates are of the same phenotype (in-cross), give a name for the phenotype, otherwise blanks

16-25 Phenotype of one of parents in binary code

26-35 Phenotype of the other in binary code

36-45 Observed mating frequency (in observed number)

Card IV This card, a trailer, is used only at the end of a job. For multiple-runs, card III should be followed immediately by a new set of controls.

Col. 1-80 9

Output from the program includes the following:

1. Trial value of gene frequency (given)
2. Pheno-mating type frequency
   a. observed number (given)
   b. expected
   c. corresponding χ^2 value for goodness of fit
3. Number of iteration cycles performed
4. Final χ^2-value for maximum likelihood estimation of parameters
5. Log-determinant of information matrix
6. Information matrix (or K-matrix)
7. Covariance matrix (or inverse matrix)
8. Maximum likelihood estimate (M.L.E.)(error in 10^{-4}) standard deviation and U-scores for each parameter under iteration.

These quantities except M.L.E. are evaluated after convergence and the maximum number of iteration is 20.

Estimated maximum time is about 1 to 2 minutes for a system.

II. Note

The features of this program are the same as the program G-TYPE except the preparation of control card III. Information on all in-crosses should be punched on cards, regardless of the observed number of couples. This is necessary for printing out "common name" of phenotype and for identification of the total number of homozygotes or the total number of genes. For the other crosses, control cards may be omitted if the observed number is zero. Deletion of these cards should be considered in the columns 11-13 in card I.

LITERATURE CITED

Azevedo, E., H. Krieger, M. P. Mi and N. E. Morton, 1965. PTC taste sensitivity and endemic goiter in Brazil. Am. J. Hum. Genet. 17: 87-90.

Bailey, N. T. J., 1961. Introduction to the Mathematical Theory of Genetic Linkage. The Clarendon Press, Oxford.

Barrai, I., M. P. Mi, N. E. Morton and N. Yasuda, 1965. Estimation of prevalence under incomplete selection. Am. J. Hum. Genet. 17: 221-236.

Bateman, A. J., 1950. Is gene dispersion normal? Heredity 4: 353-363.

Bernstein, F., 1925. Zusammenfassende Betrachtungen über die erblichen Blutstruktururen des Menschen. Zeit. Abstgs. u. Vererbgsl. 37: 237-270.

Bernstein, F., 1930. Fortgesezte Untersuchungen aus der Theorie der Blutgruppen. Zeit. Abstgs. u. Vererbgsl. 56: 233-273.

Birkhoff, G. and S. MacLane, 1965. A Survey of Modern Algebra. (3rd ed.) The Macmillan Co., New York.

Cavalli, L. L., 1958. Some data on the genetic structure of human populations. Proc. X Inter. Cong. Genet. 1: 389-407.

Cavalli, L. L., 1962. The distribution of migration distance: Models, and applications to genetics. In Les Deplacements humains, Entretien de Monaco en Sciences Humaines. J. Sutter, Ed. I. Hachette. pp. 139-166.

Cavalli, L. L., I. Barrai and A. W. F. Edwards, 1964. Analysis of human evolution under random genetic drift. Cold Sp. Harbor Symp. Quant. Biol. 29: 9-20.

Ceppelini, R., 1955. On the genetics of secretor and Lewis characters: a family study. Proc. V. Inter. Cong. Blood Trans., Paris. pp. 207-211.

Ceppellini, R., M. Siniscalco and C. A. B. Smith, 1955. The estimation of gene frequencies in a random mating population. Ann. Hum. Genet., London. 20: 97-115.

Chung, C. S., N. E. Morton and N. Yasuda, 1965. Genetics of interracial crosses. Ann. N. Y. Acad. Sci. (in press)

Connel, G. E., G. H. Dixon and O. Smithies, 1962. Subdivision of the three common haptoglobin types based on hidden differences. Nature 183: 505-506.

Cotterman, C. W., 1953. Regular two-allele and three-allele phenotypes systems. Part I. Am. J. Hum. Genet. 5: 193-235.

Cotterman, C. W., 1965. Personal communication.

Cramer, H., 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton.

Crawford, H., M. Cutbuth and P. L. Mollison, 1954. Preservation of red cells for blood grouping tests. Vox Sang. 4:149-154.

Dahlberg, G., 1929. Inbreeding in man. Genetics 14: 421-454.

Feller, W., 1961. An Introduction to Probability Theory and Its Applications. Vol. 1. (2nd ed.) John Wiley, New York.

Fisher, R. A., 1922. On the mathematical foundations of theoretical statistics. Phi. Trans. Roy. Soc. Lond. Ser. A. 222: 309-368.

Freire-Maia, N., 1957. Inbreeding in Brazil. Am. J. Hum. Genet. 9: 284-297.

Garrod, A. E., 1902. The incidence of Alkaptonuria: A study in chemical individuality. Lancet 1902-2: 1616-1620.

Giblett, E. R., 1959. Haptoglobin types in American Negroes. Nature 183: 192-193.

Grubb, R., 1951. Observations on the human group system Lewis. Acta path. microbiol. scand., 28: 61-81.

Hardy, G. H., 1908. Mendelian proportions in a mixed population. Science 28: 49-50.

Hiraizumi, Y., 1965. Unpublished data.

Holgate, P., 1964. Genotype frequencies in a section of a cline. Heredity 19: 507-509.

I. B. G. E., 1958. "Mapa fisico dos Estados Unidos do Brasil". Conselho Nacional de geografia, Divisao de Cartografia Rio de Janeiro, GB., Brasil.

I. B. G. E., 1961. Anuario Estatistico do Brasil. Conselho Nacional de Estatistica Rio de Janeiro, GB., Brasil.

Kendall, M. A. and A. Stuart, 1961. The Advanced Theory of Statistics. Vol. 2. Hafner Pub. Co., New York.

Kimura, M., 1955. Stochastic processes and distribution of gene frequencies under natural selection. Cold Sp. Harbor Symp. Quant. Biol., 20: 33-53.

Kimura, M., 1963. Personal communication.

Kimura, M. & G. H. Weiss, 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49: 561-576.

Krieger, H., N. E. Morton, M. P. Mi, E. Azevedo, A. Freire-Maia and N. Yasuda, 1965. Racial admixture in northeastern Brazil. Ann. Hum. Genet., London. 29: 113-125.

Kudo, A., 1962. A method for calculating the inbreeding coefficient. Am. J. Hum. Genet. 14: 426-432.

Lamotte, M., 1951. Recherches sur la structure genetique des populations Naturelles de Cepaea Nemoralis (L.). Bull. Biol. France et Belgique, Supp. 35.

Levedev, N. N., 1965. Special Functions and Their Applications. (Translated by R. A. Silverman). Prentice-Hall, Inc., New Jersey.

Li, C. C. and D. G. Horvitz, 1953. Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. 5: 107-117.

Li, C. C., 1955. Population Genetics. The University of Chicago Press, Chicago.

Malecot, G., 1948. Les Mathematique de l'Heredite. Masson, Paris.

Malecot, G., 1950. Quelques Schemas Probabilistes sur La Variabilite des Populations Naturelles. Ann. de L'Universite de Lyon, 13. Section A. In Sciences Mathematiques et Astronomie (Paris, Masson et C). pp. 37-60.

Malecot, G. 1955. The decrease of relationship with distance. In discussion of Kimura, 1955. Cold Sp. Harbor Symp. Quant. Biol., 20: 52-53.

Malecot, G., 1959. Les Modeles Stochastiques en Genetique de Population. Publications de L'Institut de Statistique de L'Universite de Paris. 8: 173-210.

Mood, A. M. and F. A. Graybill, 1963. Introduction to the Theory of

Statistics. (2nd ed.) McGraw-Hill Co., Inc., New York.

Moor-Jankowski, J. K. and H. J. Huser, 1957. Sero-anthropological investigations in the Walser and Romansh isolates in the Swiss Alps and their methodological aspects. Acta Genet. Stat. Med. 6: 527-531.

Moor-Jankowski, J. and A. S. Wiener and C. M. Rogers, 1964. Blood groups of chimpanzees: demonstrated with isoimmune serums. Science 145: 1441-1443.

Moroni, A., 1962. Sources, reliability, and usefulness of consanguinity data with special reference to Catholic records. U.N./W.H.O. Seminar on use of vital and health statistics for genetics and radiation studies.

Morton, N. E., 1955. Non-randomness in consanguineous marriage. Ann. Hum. Genet. London. 20: 116-124.

Morton, N. E., 1959. Genetic tests under incomplete ascertainment. Am. J. Hum. Genet. 11: 1-16.

Morton, N. E., 1962. Segregation and linkage. In Methodology in Human Genetics, Ed., W. J. Burdette, 17-52. Holden-Day, San Francisco.

Morton, N. E. and N. Yasuda, 1962. The genetical structure of human population. In Les Deplacements humains, Entretien de Monaco en Sciences Humaines. J. Sutter, Ed. I. Hachette. pp. 186-203.

Morton, N. E., 1964. Genetic studies of northeastern Brazil. Cold Sp. Harbor Symp. Quant. Biol. 29: 69-79.

Morton, N. E., 1965a. Models and evidence in human population genetics. In Genetics Today (ed. S. J. Geerts). Vol. 3. Proc. XI Inter. Cong. Genet., The Hague, Netherlands. September, 1963. pp. 935-951. Pergamon Press.

Morton, N. E., 1965b. Parental exclusions in northeastern Brazil. In preparation.

Morton, N. E., H. Krieger, A. G. Steinberg and R. E. Rosenfield, 1965. Localization of Sutter in the Kell system. Vox Sang. 10: 608-613.

Morton, N. E., M. P. Mi and N. Yasuda, 1965. The S^u problem in northeastern Brazil. Vox Sang. (in press)

Morton, N. E., H. Krieger, and M. P. Mi, 1966. Natural selection on polymorphisms in northeastern Brazil. Am. J. Hum. Genet. (in press)

Mourant, A. E., A. C. Kopec and K. Domaniewska-Sobczak, 1958. The ABO Blood Groups: Comprehensive Tables and Maps of World Distribution. Charles C. Thomas, Springfield, Illinois.

Race, R. R., Sanger, Ruth and D. Lehane, 1953. Quantitative aspects of the blood group antigen Fy^a. Ann. Eug., London. 17: 255-266.

Race, R. R. and R. Sanger, 1962. Blood Groups in Man. (4th ed.) Blackwell, Oxford.

Rao, C. R., 1952. Advanced Statistical Methods in Biometric Research. John Wiley, New York.

Rosenfield, R. E., 1964. Personal communication.

Sanghvi, L. D., 1955. Inbreeding, genes and phenotypes. Am. Nat. 89: 247-248.

Schull, W. J., 1965. Estimation of Genetic parameters in population studies. In Genetics and the Epidemiology of Chronic Diseases (U. S. Dept. of Health, Education, and Welfare); pp. 45-60.

Skellam, J. G., 1951. Gene dispersion in heterogeneous population. Heredity 5: 433-435.

Smithies, O., 1955. Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. Biochem. J. 61: 629-641.

Smith, C. A. B., 1957. Counting methods in genetical statistics. Ann. Hum. Genet., London. 21: 254-276.

Snyder, L. H., 1932. The inheritance of taste deficiency in man. Ohio J. Sci. 32: 436-440.

Steinberg, A., 1964. Personal communication.

Stevens, W. L., 1938. Estimation of blood group gene frequencies. Ann. Eug., London. 8: 362-375.

Stroup, Majory, Mija MacIlroy, R. Walker and Jane Aydelotte, 1964. The probable relationship of the Sutter and Kell Blood group systems. In preparation.

Sutter, J. and Tran-Ngoc-Toan, 1957. The problem of structure of isolates and of their evolution among human population. Cold Sp. Harbor Symp. Quant. Biol. 22: 379-383.

Thomsen, O., V. Friedenreich and E. Worsaae, 1930. Uber die Moglichkeit der Existenz zweier neuer Blutgrupen; auch ein Beitrag zur Beleuchtung sogennanter Untergrupen. Acta path. microbiol. scand. 7: 157-190.

Wahlund, S., 1928. Zusammensetzung von Populationen und Korrelationserscheinungen von Standpunkt der Vererbungslehre aus betrachtet. Hereditas 2: 65-106.

Weinberg, W., 1908. Uber den Nachweis der Vererbung beim Menschen. Jahr. d Ver. Vater. Natur. Wuttem., Sturrgart. 64: 368-382.

Wiener, A. S., Unger, L. J. and E. B. Gordon, 1953. Fetal hemolytic transfusion reaction caused by sensitization to a new blood factor U. J. Am. Med. Ass. 153: 1444-1446.

Wilks, S. S., 1962. Mathematical Statistics. John Wiley, New York.

Wright, S., 1921. System of mating. Genetics 6: 111-178.

Wright, S. and H. C. McPhee, 1925. An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. J. Agr. Res. 31: 377-383.

Wright, S., 1931. Evolution in Mendelian populations. Genetics. 16: 97-159.

Wright, S., 1943. Isolation by distance. Genetics 28: 114-138.

Wright, S., 1946a. The differential equation of the distribution of gene frequencies. Proc. Nat. Acad. Sci. 31: 382-389.

Wright, S., 1946b. Isolation by distance under diverse system of mating. Genetics 31: 39-59.

Wright, S., 1950. Discussion on population genetics and radiation. J. Cell. Comp. Phys. Suppl. 1, 35: 187-205.

Wright, S., 1951. The genetical structure of population. Ann. Eug., London. 15: 323-354.

Table 2.1.

Expression of genotype frequencies by different concepts.

Genotype	Hardy-Weinberg (1908)	Wright (1921)	Wahlund (1928)	Bernstein (1930)	Malecot (1948)
AA	p^2	p^2+pqF	$p^2+\sigma^2$	$p(p+\alpha q)$	$p^2(1-f)+pqf$
Aa	$2pq$	$2pq-2pqF$	$2pq-2\sigma^2$	$2pq(1-\alpha)$	$2pq(1-f)$
aa	q^2	q^2+pqF	$q^2+\sigma^2$	$q(q+\alpha p)$	$q^2(1-f)+pqf$
Total	$(p+q)^2=1$	1	1	1	1

Where p and q are frequency of gene A and a, respectively, F is Wright's inbreeding coefficient, i.e. correlation coefficient of uniting gametes, f probability that two genes are identical by descent, α the mean inbreeding coefficient, and σ^2 gene frequency variance. Note that $F = f = \alpha = \sigma^2/pq$.

Table 2.5.1.

Frequencies of mating types and their offspring.
(Two alleles at an autosomal locus).

(i) No dominance

Mating type	Frequency of mating type	Frequency of offspring		
		AA	Aa	aa
AA x AA	$p^4+6p^3q\alpha$	$p^4+6p^3q\alpha$	--	--
AA x Aa	$4p^3+12p^2q(1-2p)\alpha$	$2p^3q+6p^2q(1-2p)\alpha$	$2p^3q+6p^2q(1-2p)\alpha$	--
Aa x Aa	$4p^2q^2+4pq(1-6pq)\alpha$	$p^2q^2+pq(1-6pq)\alpha$	$2p^2q^2+2pq(1-6pq)\alpha$	$p^2q^2+pq(1-6pq)\alpha$
AA x aa	$2p^2q^2+2pq(1-6pq)\alpha$	--	$2p^2q^2+2pq(1-6pq)\alpha$	--
Aa x aa	$4pq^3+12pq^2(1-2q)\alpha$	--	$2pq^3+6pq^2(1-2q)\alpha$	$2pq^3+6pq^2(1-2q)\alpha$
aa x aa	$q^4+6pq^3\alpha$	--	--	$q^4+6pq^3\alpha$
Total	1	$p^2+pq\alpha$	$2pq-2pq\alpha$	$q^2+pq\alpha$

Table 2.5.1. -- Continued

(ii) Complete dominance

Mating type	Frequency of mating type	Frequency of offspring	
		A-	aa
A- x A-	$p^2(1+q)^2-2pq(1-3q^2)\alpha$	$p^2(1+2q)-3pq(1-2q)\alpha$	$p^2q^2+pq(1-6pq)\alpha$
A- x aa	$2pq^2(1+q)+2pq(1-6q^2)\alpha$	$2pq^2+2pq(1-3q)\alpha$	$2pq^3+6pq^2(1-2q)\alpha$
aa x aa	$q^4 + 6pq^3\alpha$	--	$q^4+6pq^3\alpha$
Total	1	$1-q^2-pq\alpha$	$q^2+pq\alpha$

Where A and a are alleles with frequency p and q (p+q=1), respectively and α is the inbreeding coefficient. It is assumed that $p,q > \alpha$.

Table 2.5.2.

Frequency of mating types and their offspring.

(Two alleles at a sex-linked locus).

(i) No dominance

Mating type	Frequency of mating type	Frequency of female offspring			Frequency of male offspring	
		AA	Aa	aa	A	a
AA x A	$p^3+3p^2q\alpha$	$p^3+3p^2q\alpha$	--	--	$p^3+3p^2q\alpha$	--
Aa x A	$2p^2q+2pq(1-3p)\alpha$	$p^2q+pq(1-3p)\alpha$	$p^2q+pq(1-3p)\alpha$	--	$p^2q+pq(1-3p)\alpha$	$p^2q+pq(1-3p)\alpha$
AA x a	$p^2q+pq(1-3p)\alpha$	--	$p^2q+pq(1-3p)\alpha$	--	$p^2q+pq(1-3p)\alpha$	--
Aa x a	$2pq^2+2pq(1-3q)\alpha$	--	$pq^2+pq(1-3q)\alpha$	$pq^2+pq(1-3q)\alpha$	$pq^2+pq(1-3q)\alpha$	$pq^2+pq(1-3q)\alpha$
aa x A	$pq^2+pq(1-3q)\alpha$	--	$pq^2+pq(1-3q)\alpha$	--	--	$pq^2+pq(1-3q)\alpha$
aa x a	$q^3+3pq^2\alpha$	--	--	$q^3+3pq^2\alpha$	--	$q^3+3pq^2\alpha$
Total	1	$p^2+pq\alpha$	$2pq-2pq\alpha$	$q^2+pq\alpha$	p	q

Where A,a and p,q(p+q=1) are alleles and their corresponding frequency, respectively.

Table 2.5.2. -- Continued

(ii) Complete dominance

Mating type	Frequency of mating type	Frequency of female offspring		Frequency of male offspring	
		A-	aa	A	a
A- x A	$p^2(1+q)+pq(2-3p)\alpha$	$p^2(1+q)+pq(2-3q)\alpha$	--	$p^2+pq\alpha$	$p^2q+pq(1-3p)\alpha$
A- x a	$pq(1+q)-3pq^2\alpha$	$pq-pq\alpha$	$pq^2+pq(1-3q)\alpha$	$pq-pq\alpha$	$pq^2+pq(1-3q)\alpha$
aa x A	$pq^2+pq(1-3q)\alpha$	$pq^2+pq(1-3q)\alpha$	--	--	$pq^2+pq(1-3q)\alpha$
aa x a	$q^3+3pq^2\alpha$	--	$q^3+3pq^2\alpha$	--	$q^3+3pq^2\alpha$
Total	1	$1-q^2-pq\alpha$	$q^2+pq\alpha$	p	q

In case of complete dominance, it is assumed that A is dominant over a. α is the inbreeding coefficient.

It is also assumed that $p,q > \alpha$.

Table 2.5.3

Mating type frequency at a two allelic autosomal locus without dominance when a distribution is assumed. (p and q(p+q=1) are frequency of alleles A and a respectively, and α is the inbreeding coefficient).

Beta distribution

(i) Exact expression

Mating type	(α=0)	Frequency ($0 < \alpha < 1$)	(α=1)
AAxAA	p^4	$c[\ p^4+3p^2(1+q)\alpha + p^2(2+6q+3q^2)\alpha^2 + pq(1+q)(2+q)\alpha^3]$	p
AAxAa	$4p^3q$	$c[\ 4p^3q+12p^2q^2\alpha - 4pq(1-3q^2)\alpha^2 - 4pq^2(1+q)\alpha^3]$	0
AAxaa	$2p^2q^2$	$c[2p^2q^2+2pq(1-3pq)\alpha - 2pq(1-3pq)\alpha^2 - 2p^2q^2\alpha^3]$	0
AaxAa	$4p^2q^2$	$c[4p^2q^2+4pq(1-3pq)\alpha - 4pq(1-3pq)\alpha^2 - 4p^2q^2\alpha^3]$	0
Aaxaa	$4pq^3$	$c[\ 4pq^3+12p^2q^2\alpha - 4pq(1-3p^2)\alpha^2 - 4p^2q(1+p)\alpha^3]$	0
aaxaa	q^4	$c[\ q^4+3q^3(1+p)\alpha + q^2(2+6p+3p^2)\alpha^2 + pq(1+p)(2+p)\alpha^3]$	q
Total	1	1	1

Where $c = 1/(1+\alpha)(1+2\alpha)$.

(ii) Asymptotic expression in the neighborhood of α=0.

Mating type	Frequency
AAxAA	$p^4+6p^3q\alpha - p^2q(8-19q)\alpha^2 + 2pq(7-27q+23q^2)\alpha^3 + \ldots$
AAxAa	$4p^3q+12p^2q(q-p)\alpha + 4pq(2-15p+19p^2)\alpha^2 - 8pq(6-25q+23q^2)\alpha^3 + \ldots$
AAxaa	$2p^2q^2+2pq(1-6pq)\alpha - 2pq(4-19pq)\alpha^2 + 4pq(5-pq)\alpha^3 + \ldots$
AaxAa	$4p^2q^2+4pq(1-6pq)\alpha - 4pq(4-19pq)\alpha^2 + 8pq(5-pq)\alpha^3 + \ldots$
Aaxaa	$4pq^3+12pq^2(p-q)\alpha + 4pq(2-15q+19q^2)\alpha^2 - 8pq(6-25p+23p^2)\alpha^3 + \ldots$
aaxaa	$q^4+6pq^3\alpha - pq^2(8-19p)\alpha^2 + 2pq(7-27p+23p^2)\alpha^3 + \ldots$
Total	1

Table 2.5.3. - Continued

Binomial distribution

Mating type	(α=0)	Frequency ($0 < \alpha < 1$)		(α=1)
AAxAA	p^4	$p^4+6p^3q\alpha + p^2q(7-11p)\alpha^2 + pq(1-6pq)\alpha^3$		p
AAxAa	$4p^3q$	$4p^3q+12p^2q(1-2p)\alpha + 4p^2q(3-11q)\alpha^2 - 4pq(1-6pq)\alpha^3$		0
AAxaa	$2p^2q^2$	$2p^2q^2+2pq(1-6pq)\alpha - 2pq(1-11pq)\alpha^2 + 2pq(1-6pq)\alpha^3$		0
AaxAa	$4p^2q^2$	$4p^2q^2+4pq(1-6pq)\alpha - 4pq(1-11pq)\alpha^2 + 4pq(1-6pq)\alpha^3$		0
Aaxaa	$4pq^3$	$4pq^2+12pq^2(1-2q)\alpha + 4pq^2(3-11p)\alpha^2 - 4pq(1-6pq)\alpha^3$		0
aaxaa	q^4	$q^4+6pq^3\alpha + pq^2(7-11q)\alpha^2 + pq(1-6pq)\alpha^3$		q
Total	1	1		1

Poisson distribution (p>q)

Mating type	Frequency				
AAxAA	$p^4+6p^3q\alpha$	+	$7p^2q^2\alpha^2$	+	$pq^3\alpha^3$
AAxAa	$4p^3q+12p^2q(1-2p)\alpha$	+	$4pq^2(1-7p)\alpha^2$	-	$4pq^3\alpha^3$
AAxaa	$2p^2q^2+2pq(1-6pq)\alpha$	-	$2pq^2(2-7p)\alpha^2$	+	$2pq^3\alpha^3$
AaxAa	$4p^2q^2+4pq(1-6pq)\alpha$	-	$4pq^2(2-7p)\alpha^2$	+	$4pq^3\alpha^3$
Aaxaa	$4pq^3+12pq^2(1-2q)\alpha$	+	$4pq^2(3-7p)\alpha^2$	-	$4pq^3\alpha^3$
aaxaa	$q^4+6pq^3\alpha$	-	$pq^2(4-7p)\alpha^2$	+	$pq^3\alpha^3$
Total	1				

Table 2.5.3. - Continued

Rectangular distribution

Mating type	Frequency				
AAxAA	p^4	+	$6p^3q\alpha$	+	$1.8p^2q^2\alpha^2$
AAxAa	$4p^3q$	+	$12p^2q^2(1-2p)\alpha$	-	$7.2p^2q^2\alpha^2$
AAxaa	$2p^2q^2$	+	$2pq(1-6pq)\alpha$	+	$3.6p^2q^2\alpha^2$
AaxAa	$4p^2q^2$	+	$4pq(1-6pq)\alpha$	+	$7.2p^2q^2\alpha^2$
Aaxaa	$4pq^3$	+	$12p^2q^2(1-2q)\alpha$	-	$7.2p^2q^2\alpha^2$
aaxaa	q^4	+	$6pq^3\alpha$	+	$1.8p^2q^2\alpha^2$
Total	1				

Normal distribution

Mating type	Frequency				
AAxAA	p^4	+	$6p^3q\alpha$	+	$3p^2q^2\alpha^2$
AAxAa	$4p^3q$	+	$12p^2q(1-2p)\alpha$	-	$12p^2q^2\alpha^2$
AAxaa	$2p^2q^2$	+	$2pq(1-6pq)\alpha$	+	$6p^2q^2\alpha^2$
AaxAa	$4p^2q^2$	+	$4pq(1-6pq)\alpha$	+	$12p^2q^2\alpha^2$
Aaxaa	$4pq^3$	+	$12pq^2(1-2q)\alpha$	-	$12p^2q^2\alpha^2$
aaxaa	q^4	+	$6pq^3\alpha$	+	$3p^2q^2\alpha^2$
Total	1				

Table 2.5.3. - Continued

Gamma distribution (q>p)

Mating type	Frequency							
AAxAA	p^4	+	$6p^3q\alpha$	-	$p^2q(8-11q)\alpha^2$	+	$6p^3q\alpha^3$	
AAxAa	$4p^3q$	+	$12p^2q(q-p)\alpha$	+	$4p^2q(6-11q)\alpha^2$	-	$24p^3q\alpha^3$	
AAxaa	$2p^2q^2$	+	$2pq(1-6pq)\alpha$	-	$2p^2q(4-11q)\alpha^2$	+	$12p^3q\alpha^3$	
AaxAa	$4p^2q^2$	+	$4pq(1-6pq)\alpha$	-	$4p^2q(4-11q)\alpha^2$	+	$24p^3q\alpha^3$	
Aaxaa	$4pq^3$	+	$12pq^2(p-q)\alpha$	+	$4p^2q(2-11q)\alpha^2$	-	$24p^3q\alpha^3$	
aaxaa	q^4	+	$6pq^3\alpha$	+	$11p^2q^2\alpha^2$	+	$6p^3q\alpha^3$	
Total	1							

Table 2.5.4.

Mating type frequency at a two allelic sex-linked locus without dominance when a distribution is assumed. (p and q(p+q=1) are frequency of alleles A and a respectively, and α is the inbreeding coefficient).

Beta distribution

(i) Exact expression

Mating type	(α=0)	Frequency ($0 < \alpha < 1$)	(α=1)
AAxA	p^3	$c[p^3 + p^2(1+2q)\alpha + pq(1+q)\alpha^2]$	p
AaxA	$2p^2q$	$c[2p^2q + 2pq(q-p)\alpha - 2pq^2\alpha^2]$	0
AAxa	p^2q	$c[p^2q + pq(q-p)\alpha - pq^2\alpha^2]$	0
Aaxa	$2pq^2$	$c[2pq^2 + 2pq(p-q)\alpha - 2p^2q\alpha^2]$	0
aaxA	pq^2	$c[pq^2 + pq(p-q)\alpha - p^2q\alpha^2]$	0
aaxa	q^3	$c[q^3 + q^2(1+2p)\alpha + pq(1+p)\alpha^2]$	q
Total	1	1	1

Where $c=1/(1+\alpha)$

(ii) Asymptotic expression in the neighborhood of α=0.

Mating type	Frequency				
AAxA	p^3	+	$3p^2q\alpha$	+	$2pq(q-p)\alpha^2\beta$
AaxA	$2p^2q$	+	$2pq(1-3p)\alpha$	-	$4pq(q-p)\alpha^2\beta$
AAxa	p^2q	+	$pq(1-3p)\alpha$	-	$2pq(q-p)\alpha^2\beta$
Aaxa	$2pq^2$	+	$2pq(1-3q)\alpha$	-	$4pq(p-q)\alpha^2\beta$
aaxA	pq^2	+	$pq(1-3q)\alpha$	-	$2pq(p-q)\alpha^2\beta$
aaxa	q^3	+	$3pq^2\alpha$	+	$2pq(p-q)\alpha^2\beta$
Total	1				

Where $\beta = \sum_{k=0}^{\infty}(-1)^k\alpha^k$

Table 2.5.4. - Continued

Binomial distribution

Mating type	(α=0)	Frequency (0 < α < 1)					(α=1)
AAxA	p^3	p^3	+	$3p^2q\alpha$	+	$pq(1-2p)\alpha^2$	p
AaxA	$2p^2q$	$2p^2q$	+	$2pq(1-3p)\alpha$	-	$2pq(1-2p)\alpha^2$	0
AAxa	p^2q	p^2q	+	$pq(1-3p)\alpha$	-	$pq(1-2p)\alpha^2$	0
Aaxa	$2pq^2$	$2pq^2$	+	$2pq(1-3q)\alpha$	-	$2pq(1-2q)\alpha^2$	0
aaxA	pq^2	pq^2	+	$pq(1-3q)\alpha$	-	$pq(1-2q)\alpha^2$	0
aaxa	q^3	q^3	+	$3pq^2\alpha$	+	$pq(1-2q)\alpha^2$	q
Total	1	1					1

Poisson distribution

Mating type	Frequency				
AAxA	p^3	+	$3p^2q\alpha$	+	$pq^2\alpha^2$
AaxA	$2p^2q$	+	$2pq(1-3p)\alpha$	-	$2pq^2\alpha^2$
AAxa	p^2q	+	$pq(1-3p)\alpha$	-	$pq^2\alpha^2$
Aaxa	$2pq^2$	+	$2pq(1-3q)\alpha$	+	$2pq^2\alpha^2$
aaxA	pq^2	+	$pq(1-3q)\alpha$	+	$pq^2\alpha^2$
aaxa	q^3	+	$3p^2q\alpha$	-	$pq^2\alpha^2$
Total	1				

Table 2.5.4. - Continued

Gamma distribution (p>q)

Mating type	Frequency				
AAxA	p^3	+	$3p^2q\alpha$	+	$2pq^2\alpha^2$
AaxA	$2p^2q$	+	$2pq(1-3p)\alpha$	-	$4pq^2\alpha^2$
AAxa	p^2q	+	$pq(1-3p)\alpha$	-	$2pq^2\alpha^2$
Aaxa	$2pq^2$	+	$2pq(1-3q)\alpha$	+	$4pq^2\alpha^2$
aaxA	pq^2	+	$pq(1-3q)\alpha$	+	$2pq^2\alpha^2$
aaxa	q^3	+	$3pq^2\alpha$	-	$2pq^2\alpha^2$
Total	1				

Table 2.5.5.

Frequency of basic mating type (Autosome) and their derivative.

Mating type		Frequency (M)
incross	$\underline{AA} \times \underline{AA}$	$p_A^4 + 6p_A^3(1-p_A)\alpha$
backcross	$\underline{AA} \times \underline{AB}$	$4p_A^3p_B + 12p_A^2p_B(1-2p_A)\alpha$
outcross	$\underline{AA} \times \underline{BB}$	$2p_A^2p_B^2 + 2p_Ap_B(p_A+p_B-6p_Ap_B)\alpha$
intercross	$\underline{AB} \times \underline{AB}$	$4p_A^2p_B^2 + 4p_Ap_B(p_A+p_B-6p_Ap_B)\alpha$
3-ways outcross	$\underline{AA} \times \underline{BC}$	$4p_A^2p_Bp_C + 4p_Ap_Bp_C(1-6p_A)\alpha$
3-ways intercross	$\underline{AB} \times \underline{AC}$	$8p_A^2p_Bp_C + 8p_Ap_Bp_C(1-6p_A)\alpha$
4-ways intercross	$\underline{AB} \times \underline{CD}$	$8p_Ap_Bp_Cp_D(1-6\alpha)$

Table 2.5.5. - Continued

$\frac{\partial M}{\partial \alpha} = Mu_{\alpha}$	$\frac{\partial M}{\partial p_A} = Mu_{p_A}$	$\frac{\partial M}{\partial p_B} = Mu_{p_B}$	$\frac{\partial M}{\partial p_C} = Mu_{p_C}$	$\frac{\partial M}{\partial p_D} = Mu_{p_D}$
$6p_A^3(1-p_A)$	$4p_A^3+6p_A^2(3-4p_A)\alpha$	0	0	0
$12p_A^2p_B(1-2p_A)$	$12p_A^2p_B+24p_Ap_B(1-3p_A)\alpha$	$4p_A^3+12p_A^2(1-2p_A)\alpha$	0	0
$2p_Ap_B(p_A+p_B-6p_Ap_B)$	$4p_Ap_B^2+2p_B(2p_A+p_B-12p_Ap_B)\alpha$	$4p_A^2p_B+2p_A(p_A+2p_B-12p_Ap_B)\alpha$	0	0
$4p_Ap_B(p_A+p_B-6p_Ap_B)$	$8p_Ap_B^2+4p_B(2p_A+p_B-12p_Ap_B)\alpha$	$8p_A^2p_B+4p_A(p_A+2p_B-12p_Ap_B)\alpha$	0	0
$4p_Ap_Bp_C(1-6p_A)$	$8p_Ap_Bp_C+4p_Bp_C(1-12p_A)\alpha$	$4p_A^2p_C+4p_Ap_C(1-6p_A)\alpha$	$4p_A^2p_B+4p_Ap_B(1-6p_A)\alpha$	0
$8p_Ap_Bp_C(1-6p_A)$	$16p_Ap_Bp_C+8p_Bp_C(1-12p_A)\alpha$	$8p_A^2p_C+8p_Ap_C(1-6p_A)\alpha$	$8p_A^2p_B+8p_Ap_B(1-6p_A)\alpha$	0
$-48p_Ap_Bp_Cp_D$	$8p_Bp_Cp_D(1-6\alpha)$	$8p_Ap_Cp_D(1-6\alpha)$	$8p_Ap_Bp_D(1-6\alpha)$	$8p_Ap_Bp_C(1-6\alpha)$

A, B, ... and p_A, p_B, ... are alleles and the corresponding frequencies, respectively ($p_A+p_B+\ldots = 1$). α is the inbreeding coefficient. It is assumed that p_A, p_B, ... $> \alpha$.

Table 2.5.6.

Frequency of basic mating types (Sex-linked) and their derivative.

Mating type		Frequency (M)	$\frac{\partial M}{\partial \alpha}$	$\frac{\partial M}{\partial p_A}$	$\frac{\partial M}{\partial p_B}$	$\frac{\partial M}{\partial p_C}$
incross	$\underline{AA} \times \underline{A}$	$p_A^3+3p_A^2(1-p_A)\alpha$	$3p_A^2(1-p_A)$	$3p_A^2+3p_A(2-3p_A)\alpha$	0	0
outcross	$\underline{AA} \times \underline{B}$	$p_A^2p_B+p_Ap_B(1-3p_A)\alpha$	$p_Ap_B(1-3p_A)$	$2p_Ap_B+p_B(1-6p_A)\alpha$	$p_A^2+p_A(1-3p_A)\alpha$	0
backcross	$\underline{AB} \times \underline{A}$	$2p_A^2p_B+2p_Ap_B(1-3p_A)\alpha$	$2p_Ap_B(1-3p_A)$	$4p_Ap_B+2p_B(1-6p_A)\alpha$	$2p_A^2+2p_A(1-3p_A)\alpha$	0
intercross	$\underline{AB} \times \underline{C}$	$2p_Ap_Bp_C(1-3\alpha)$	$-6p_Ap_Bp_C$	$2p_Bp_C(1-3\alpha)$	$2p_Ap_C(1-3\alpha)$	$2p_Ap_B(1-3\alpha)$

$\underline{A}$, $\underline{B}$, $\underline{C}$ and p_A, p_B, p_C are alleles and the corresponding frequencies, respectively ($p_A+p_B \ldots = 1$). α is the inbreeding coefficient. It is assumed that p_A, p_B, $\ldots > \alpha$.

Table 3.1.

Maximum likelihood scores and variances at a sex-linked locus with two alleles under the hypothesis that $\alpha=0$.

Model and phenotype data	Score
No dominance	
$\underline{AA}\ f_1\ \underline{A}\ m_1$	$u_q = [f_2+2f_3+m_2-(M+2F)q]/q(1-q)$
$\underline{Aa}\ f_2\ \underline{a}\ m_2$	$u_\alpha = \dfrac{[(M+F)(4f_1f_3-f_2^2)+(f_1m_2^2-m_1m_2f_2+f_3m_1^2)]}{(m_1+f_2+2f_1)(m_2+f_2+2f_3)}$
$\underline{aa}\ f_3\ \underline{\quad}$ F M	$K_{qq} = (M+2F)/q(1-q),\quad K_{q\alpha} = 0,\quad K_{\alpha\alpha} = F$
Complete dominance	
$\underline{A\text{-}}\ f_1\ \underline{A}\ m_1$	$u_q = (m_2+2f_2)/q-[(2f_1+m_1)q+m_1]/(1-q^2)$
$\underline{aa}\ f_2\ \underline{a}\ m_2$	$u_\alpha = f_1-f_2+[f_2-Fq]/q(1-q)$
F M	$K_{qq} = 4F/(1-q^2)+M/q(1-q),\quad K_{q\alpha} = 2F/(1+q),$ $K_{\alpha\alpha} = F(1-q)/(1+q)$
No dominance	
$\underline{AA} \times \underline{A}\ \ n_1$	$u_q = [n_{23}+2n_{45}+3n_6-3Nq]/q(1-q)$
$\underline{Aa} \times \underline{A}\ \ n_2$	$u_\alpha = \dfrac{3N[(3n_1+n_{23})(n_{45}+3n_6)-(n_{23}+n_{45})^2]}{(3n_1+2n_{23}+n_{45})(n_{23}+2n_{45}+3n_6)}$
$\underline{AA} \times \underline{a}\ \ n_3$	
$\underline{Aa} \times \underline{a}\ \ n_4$	$K_{qq} = 3N/q(1-q),\quad K_{q\alpha} = 0,\quad K_{\alpha\alpha} = 3N$
$\underline{aa} \times \underline{A}\ \ n_5$	
$\underline{aa} \times \underline{a}\ \ \underline{n_6}$ N	$\begin{bmatrix} n_{23} = n_2+n_3 \\ n_{45} = n_4+n_5 \end{bmatrix}$

Table 3.1. - Continued

Model and phenotype data		Score
Complete dominance		
$\underline{A}\text{-} \times \underline{A}\ \underline{A}$	n_1	$u_q = [(2n_2+n_3+3n_4)-(n_1+n_3)q-3Nq^2]/q(1-q^2)$
$\underline{aa} \times \underline{A}\ \underline{A}$	n_2	$u_\alpha = [3(n_1+n_2)-(n_1+3n_2)q]/(1-q^2)+(n_3+3n_4)/q-3N$
$\underline{A}\text{-} \times \underline{a}\ \underline{a}$	n_3	$K_{qq} = N(1+5q)/q(1-q^2)$, $K_{q\alpha} = 2N/(1+q)$,
$\underline{aa} \times \underline{a}\ \underline{a}$	n_4	$K_{\alpha\alpha} = N(1+3q)/(1+q)$
	N	

$\underline{A}$, $\underline{a}$ and p, q are alleles and the corresponding frequency, respectively. α is the inbreeding coefficient. f_i, m_i, and n_i denote the observed number.

Table 3.2.

Brazilian serotypes with reference of their binary code.

Genetic system	Gene		Phenotype		Possible genotype
Secretor	Se	1	Se	1	Se/Se, Se/se
	se	0	se	0	se/se
Lewis	Le	1	Le	1	Le/Le, Le/le
	le	0	le	0	le/le
Lutheran	Lu^a	1	a+	1	Lu^a/Lu^a, Lu^a/Lu
	Lu	0	a-	0	Lu/Lu
Gm	a	10000	a	10000	a/a
	ab	11000	abx	11100	ax/ab
	ax	10100	abcx	11110	ax/abc
	abc	11010	axbd	11101	$ax/b^{(1,2)}$
	$b^{(1,2)}$	01001	ax	10100	ax/ax, ax/a
			ab	11000	ab/ab, ab/a
			abc	11010	abc/abc, abc/ab, abc/a
			abcd	11011	$abc/b^{(1,2)}$
			abd	11001	$ab/b^{(1,2)}$, $a/b^{(1,2)}$
			bd	01001	$b^{(1,2)}/b^{(1,2)}$
Inv	Inv^a	1	a+	1	Inv^a/Inv^a, Inv^a/Inv
	Inv	0	a-	0	Inv/Inv
PTC	T	1	T+	1	T/T, T/t
	t	0	T-	0	t/t
P	P_1	1	P+	1	P_1/P_1, P_1/P
	P	0	P-	0	P/P

Table 3.2. - Continued

Genetic system	Gene		Phenotype		Possible genotype
ABO	A_1	110	A_1	110	A_1/A_1, A_1/A_2, A_1/O
	A_2	010	A_1B	111	A_1/B
	B	001	A_2	010	A_2/A_2, A_2/O
	O	000	A_2B	011	A_2/B
			B	001	B/B, B/O
			O	000	O/O
Kell	K	100	K	100	K/K
	k	010	Kk	110	K/k
	k^s	011	Kk^s	111	K/k^s
			k	010	k/k
			k^s	011	k/k^s, k^s/k^s
Duffy	Fy^a	1	a+	1	Fy^a/Fy^a, Fy^a/Fy
	Fy	0	a-	0	Fy/Fy
Diego	Di^a	1	a+	1	Di^a/Di^a, Di^a/Di
	Di	0	a-	0	Di/Di
MNSsU	MS	1010	M	1000	M*/M*
	Ms	1001	MS	1010	$\hat{MS}/\hat{MS}$, MS/M*
	M*	1000	MSs	1011	$\hat{MS}$/Ms
	$\hat{NS}$	0110	Ms	1001	Ms/Ms, Ms/M*
	Ns	0101	MN	1100	$\hat{M*}/\hat{N*}$
	N*	0100	MNS	1110	$\hat{MS}/\hat{NS}$, MS/N*, M*/NS
			MNSs	1111	MS/Ns, Ms/$\hat{NS}$
			MNs	1101	Ms/Ns, Ms/N*, M*/Ns

Table 3.2. - Continued

Genetic system	Gene		Phenotype		Possible genotype
			N	0100	N*/N*
			NS	0110	$\hat{NS}/\hat{NS}$, NS/N*
			NSs	0111	NS/Ns
			Ns	0101	Ns/Ns, Ns/N*
Rh	cde(r)	00011	r	00011	r/r
	Cde(r')	00101	R_1r	10111	R_1/r, R_1/R_0, R_0/r',
	cdE(r'')	01010			$(R_1/r'^n, R_0/r'^n)$
	CdE(r^y)	01100	R_1	10101	R_1/R_1, R_1/r'
	cDe(R_0)	10011	R_0	10011	R_0/R_0, R_0/r
	CDe(R_1)	10101	R_2r	11011	R_2/r, R_2/R_0, R_0/r''
	cDE(R_2)	11010	R_2	11010	R_2/R_2, R_2/r''
	CDE(R^z)	11100	R_1R_2	11111	R_1/R_2, R_1/r'', R_0/R^z, R_0/r^y,
	[$Cde^s(r'^n)$	00111]			R_2/r', R^z/r, $(R^z/r'^n, R_2/r'^n)$
			r'	00101	r'/r'
			rr'	00111	r/r', $(r'^n/r'^n, r'/r'^n,$
					$r/r'^n)$
			R_1R^z	11101	R_1/R^z, R_1/r^y, R^z/r'
			$r'r^y$	01101	r'/r^y
			r'r''	01011	r/r''
			R^z	11100	R^z/R^z, R^z/r^y
			R_2R^z	11110	R_2/R^z, R_2/r^y, R^z/r''
			r^y	01100	r^y/r^y
			r^yr''	01110	r^y/r''
			r''	01010	r''/r''

Table 3.2. - Continued

Genetic system	Gene		Phenotype		Possible genotype
Tf	B	100	B	100	B/B
	C	010	BC	110	B/C
	D	001	BD	101	B/D
			C	010	C/C
			CD	011	C/D
			D	001	D/D
Hemoglobin					
	A	100	AA	100	A/A
	S	010	AS	110	A/S
	C	001	AC	101	A/C
			SS	010	S/S
			SC	011	S/C
			CC	001	C/C
Haptoglobin					
	1F	100	1F	100	1F/1F
	1S	010	1F-1S	110	1F/1S
	2	001	1F-2	101	1F/2
			1S	010	1S/1S
			1S-2	011	1S/2
			2	001	2/2

Table 3.3.

Possible migration function, its probability and derivative.

Exponential	$m(x;a) = ae^{-ax}$
	$P = e^{-a\alpha} - e^{-a\beta}$
	$U_a P = \beta e^{-a\beta} - \alpha e^{-a\alpha}$
Normal	$m(x;a) = 2ae^{-(ax)^2}/\sqrt{\pi}$
	$P = \Phi(\sqrt{2}a\beta) - \Phi(\sqrt{2}a\alpha)$
	$U_a P = 2[\beta e^{-(a\beta)^2} - \alpha e^{-(a\alpha)^2}]/\sqrt{\pi}$
	$\left[\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-t^2/2} dt\right]$
Lognormal	$m(x;a) = ae^{-(a\ln x)^2}/x\sqrt{\pi}$
	$P = [\Phi(\sqrt{2}a\ln\beta) - \Phi(\sqrt{2}a\ln\alpha)]/2$
	$U_a P = [\ln\beta e^{-(a\ln\beta)^2} - \ln\alpha e^{-(a\ln\alpha)^2}]/\sqrt{\pi}$
	$[P(0) = 0]$
Square root Exponential	$m(x;a) = a^2 e^{-a\sqrt{x}}/2$
	$P = (1+a\sqrt{\alpha})e^{-a\sqrt{\alpha}} - (1+a\sqrt{\beta})e^{-a\sqrt{\beta}}$
	$U_a P = a[\beta e^{-a\sqrt{\beta}} - \alpha e^{-a\sqrt{\alpha}}]$

Table 3.3. - Continued

Skellam type

$$m(x;a) = 2ax/(1+x^2)^{a+1}$$

$$P = 1/(1+\alpha^2)^a - 1/(1+\beta^2)^a$$

$$U_a P = \ln(1+\beta^2)/(1+\beta^2)^a - \ln(1+\alpha^2)/(1+\alpha^2)^a$$

$$\left[\lim_{x\to\infty} \frac{\ln(1+x^2)}{(1+x^2)^a} = 0\right]$$

Bessel

$$m(x;a) = a^2 x K_0(ax)$$

$$P = a[\alpha K_1(a\alpha) - \beta K_1(a\beta)]$$

$$U_a P = a[\beta^2 K_0(a\beta) - \alpha^2 K_0(a\alpha)]$$

$$\left[\lim_{x\to 0} xK_n(x) = n \quad (n=0,1)\right]$$

Beta type

$$m(x;a,b) = ab/(1+ax)^{b+1}$$

$$P = 1/(1+a\alpha)^b - 1/(1+a\beta)^b$$

$$U_a P = b[\beta/(1+a\beta)^{b+1} - \alpha/(1+a\alpha)^{b+1}]$$

$$U_b P = \ln(1+a\beta)/(1+a\beta)^b - \ln(1+a\alpha)/(1+a\alpha)^b$$

Generalized Skellam type

$$m(x;a,b) = 2abx/(1+ax^2)^{b+1}$$

$$P = 1/(1+a\alpha^2)^b - 1/(1+a\beta^2)^b$$

$$U_a P = b[\beta^2/(1+a\beta^2)^{b+1} - \alpha^2/(1+a\alpha^2)^{b+1}]$$

$$U_b P = \ln(1+a\beta^2)/(1+a\beta^2)^b - \ln(1+a\alpha^2)/(1+a\alpha^2)^b$$

Table 3.3. - Continued

Double Exponential	$m(x;a,b,p) = (1-p)ae^{-ax} + pbe^{-bx}$
	$P = (1-p)(e^{-a\alpha} - e^{-a\beta}) + p(e^{-b\alpha} - e^{-b\beta})$
	$U_p P = e^{-b\alpha} - e^{-b\beta} - (e^{-a\alpha} - e^{-a\beta})$
	$U_a P = (1-p)(\beta e^{-a\beta} - \alpha e^{-a\alpha})$
	$U_b P = p(\beta e^{-b\beta} - \alpha e^{-b\alpha})$

Where $m(x;...)$ is migration function with distance x, and a, b and p are parameters. P is defined as $P = \int_{\alpha}^{\beta} m(x;...)dx$ $(\beta > \alpha)$ and U is score. See text.

Table 4.2.1.

Gene frequency at the sixteen polymorphic systems in northeastern Brazil.

(Total population in parent)

(p and σ stand for gene frequency and its standard error, respectively)

Marital distance (km.) →			[0-∞)	[0-3]	(3-27]	(27-∞)
Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
Secretor	Se	1	.5537 ± .0097	.5474 ± .0220	.5619 ± .0177	.5552 ± .0141
	se	0	.4463 ± .0097	.4526 ± .0220	.4381 ± .0177	.4448 ± .0141
Lewis	Le	1	.5317 ± .0096	.5409 ± .0220	.5573 ± .0176	.5161 ± .0138
	le	0	.4683 ± .0096	.4591 ± .0220	.4427 ± .0176	.4839 ± .0138
Lutheran	Lu^a	1	.0321 ± .0061	.0451 ± .0180	.0294 ± .0103	.0311 ± .0085
	Lu	0	.9679 ± .0061	.9549 ± .0180	.9706 ± .0103	.9689 ± .0085
PTC	T	1	.6250 ± .0101	.6337 ± .0230	.6336 ± .0183	.6193 ± .0145
	t	0	.3750 ± .0101	.3663 ± .0230	.3664 ± .0183	.3807 ± .0145
P	P_1	1	.6270 ± .0101	.6143 ± .0228	.6177 ± .0181	.6355 ± .0146
	P_2+p	0	.3730 ± .0101	.3857 ± .0228	.3823 ± .0181	.3645 ± .0146
Duffy	Fy^a	1	.2833 ± .0076	.2929 ± .0175	.2973 ± .0140	.3184 ± .0115
	Fy^-	0	.7167 ± .0076	.7071 ± .0175	.7027 ± .0140	.6816 ± .0115
Inv	Inv^a	1	.2099 ± .0067	.2137 ± .0153	.2009 ± .0119	.2131 ± .0098
	Inv^-	0	.7901 ± .0067	.7863 ± .0153	.7991 ± .0119	.7869 ± .0098

Table 4.2.1. - Continued

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
Diego	Di^a	1	.0219 ± .0051	.0158 ± .0111	.0257 ± .0096	.0223 ± .0074
	Di	0	.9781 ± .0051	.9842 ± .0111	.9743 ± .0096	.9777 ± .0074
Haptoglobin	1F	100	.2159 ± .0064	.2072 ± .0152	.2222 ± .0114	.2500 ± .0126
	1S	010	.2554 ± .0068	.2513 ± .0155	.2416 ± .0124	.2633 ± .0139
	2	001	.5287 ± .0079	.5414 ± .0187	.5362 ± .0142	.4867 ± .0154
Hemoglobin	A	100	.9731 ± .0025	.9813 ± .0048	.9731 ± .0046	.9689 ± .0040
	S	010	.0199 ± .0022	.0149 ± .0043	.0206 ± .0040	.0221 ± .0033
	C	001	.0070 ± .0013	.0037 ± .0022	.0063 ± .0022	.0090 ± .0021
Transferrin	B	100	.0031 ± .0009	.0100 ± .0035*	.0063 ± .0022	.0015 ± .0009
	D	001	.0132 ± .0018		.0118 ± .0031	.0155 ± .0028
	C	010	.9837 ± .0020	.9900 ± .0035	.9819 ± .0038	.9830 ± .0029
Kell	K	100	.0257 ± .0055	.0286 ± .0202	.0145 ± .0073	.0354 ± .0092
	k_s	010	.9447 ± .0077	.9424 ± .0286	.9523 ± .0131	.9383 ± .0116
	k^s	011	.0296 ± .0059	.0290 ± .0202	.0332 ± .0109	.0263 ± .0078
ABO	A_1	110	.1566 ± .0058	.1753 ± .0136	.1616 ± .0107	.1477 ± .0082
	A_2	010	.0526 ± .0038	.0411 ± .0077	.0601 ± .0073	.0529 ± .0055
	B	001	.0808 ± .0043	.0670 ± .0089	.0828 ± .0078	.0857 ± .0064
	O	000	.7100 ± .0074	.7166 ± .0169	.6955 ± .0137	.7137 ± .0104

Table 4.2.1. - Continued.

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
MNSsU	MS	1010	.1970 ± .0068	.1914 ± .0162	.2154 ± .0123	.1855 ± .0100
	Ms	1001	.3451 ± .0086	.3363 ± .0250	.3331 ± .0143	.3563 ± .0124
	M*	1000	.0091 ± .0042	.0199 2 .0240	--- **	.0119 ± .0055
	NS	0110	.0762 ± .0053	.0578 ± .0108	.0812 ± .0086	.0895 ± .0082
	Ns	0101	.3485 ± .0089	.3660 ± .0210	.3703 ± .0141	.3366 ± .0125
	N*	0100	.0241 ± .0050	.0286 ± .0119	--- **	.0202 ± .0063
Gm	a	10000	.2233 ± .0094	.2144 ± .0208	.2062 ± .0169	.2426 ± .0138
	ab	11000	.2203 ± .0092	.2245 ± .0210	.2227 ± .0166	.2123 ± .0133
	ax	10100	.0762 ± .0042	.0746 ± .0095	.0828 ± .0079	.0749 ± .0006
	abc	11010	.0643 ± .0039	.0723 ± .0094	.0569 ± .0066	.0649 ± .0057
	$b^{(1,2)}$	01001	.4159 ± .0075	.4142 ± .0169	.4314 ± .0135	.4053 ± .0109
Rh	cde(r)	00011	.2870 ± .0100	.2964 ± .0228	.3071 ± .0182	.2679 ± .0145
	Cde(r')	00101	.0175 ± .0037	.0256 ± .0097	.0120 ± .0054	.0157 ± .0053
	cdE(r'')	01010	.0025 ± .0014	--- **	--- **	--- **
	CdE(r^y)	01100	.0007 ± .0008	--- **	--- **	--- **
	cDe(R_0)	10011	.2273 ± .0097	.2109 ± .0216	.2230 ± .0174	.2361 ± .0142
	CDe(R_1)	10101	.3276 ± .0078	.3041 ± .0181	.3289 ± .0137	.3444 ± .0115
	cDE(R_2)	11010	.1349 ± .0054	.1630 ± .0127	.1290 ± .0094	.1359 ± .0076
	CDE(R^z)	11100	.0025 ± .0014	--- **	--- **	--- **

Table 4.2.1. - Continued

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
Rh	cde(r)	00011	.2852 ± .0102	---	---	---
	Cde(r')	01001	.0095 ± .0113	---	---	---
	cdE(r'')	00110	.0025 ± .0014	---	---	---
	CdE(r^y)	01100	.0007 ± .0008	---	---	---
	cDe(R_0)	10011	.2261 ± .0098	---	---	---
	CDe(R_1)	11001	.3300 ± .0086	---	---	---
	cDE(R_2)	10110	.1349 ± .0054	---	---	---
	CDE(R^z)	11100	.0026 ± .0015	---	---	---
	$cde^s(r'^n)$	01011	.0085 ± .0108	---	---	---

* Genes, B and D, are pooled because of low frequency.

** Since difficulty in iterations is observed, the gene is dropped from the analysis.

Table 4.2.2.

Gene frequency at sixteen polymorphic systems in northeastern Brazil.

(Remote population in parent)

(p and σ stand for gene frequency and its standard error, respectively)

Marital distance (km.) →			[0,∞)	[0,3]	(3,27]	(27,∞)
Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
Secretor	Se	1	.5531 ± .0101	.5436 ± .0243	.5588 ± .0184	.5572 ± .0144
	se	0	.4469 ± .0101	.4564 ± .0243	.4412 ± .0184	.4428 ± .0144
Lewis	Le	1	.5264 ± .0100	.5325 ± .0242	.5539 ± .0183	.5116 ± .0141
	le	0	.4736 ± .0100	.4675 ± .0242	.4461 ± .0183	.4884 ± .0141
Lutheran	Lu^a	1	.0313 ± .0062	.0386 ± .0169	.0286 ± .0107	.0321 ± .0088
	Lu	0	.9687 ± .0062	.9614 ± .0169	.9714 ± .0107	.9679 ± .0088
PTC	T	1	.6256 ± .0105	.5490 ± .0291	.6429 ± .0191	.6185 ± .0149
	t	0	.3744 ± .0105	.4510 ± .0291	.3571 ± .0191	.3815 ± .0149
P	P_1	1	.6286 ± .0105	.6066 ± .0251	.6230 ± .0189	.6370 ± .0150
	P_2+p	0	.3714 ± .0105	.3934 ± .0251	.3770 ± .0189	.3630 ± .0150
Duffy	Fy^a	1	.2810 ± .0079	.2908 ± .0192	.2953 ± .0145	.2663 ± .0109
	Fy^-	0	.7190 ± .0079	.7092 ± .0192	.7047 ± .0145	.7337 ± .0109
Inv	Inv^a	1	.2054 ± .0070	.4097 ± .0222	.2031 ± .0124	.2143 ± .0101
	Inv^-	0	.7946 ± .0070	.5903 ± .0222	.7969 ± .0124	.7857 ± .0101

Table 4.2.2. - Continued

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
Diego	Di^a	1	.0205 ± .0051	.0163 ± .0114	.0204 ± .0090	.0230 ± .0076
	Di	0	.9795 ± .0051	.9837 ± .0114	.9796 ± .0090	.9770 ± .0076
Haptoglobin	1F	100	.2168 ± .0068	.1964 ± .0166	.2264 ± .0120	.2143 ± .0096
	1S	010	.2536 ± .0071	.2598 ± .0171	.2382 ± .0127	.2625 ± .0103
	2	001	.5296 ± .0082	.5438 ± .0207	.5355 ± .0147	.5232 ± .0117
Hemoglobin	A	100	.9730 ± .0027	.9788 ± .0057	.9707 ± .0050	.9714 ± .0039
	S	010	.0197 ± .0023	.0167 ± .0050	.0224 ± .0044	.0196 ± .0032
	C	001	.0073 ± .0014	.0045 ± .0026	.0069 ± .0024	.0090 ± .0022
Transferrin	B	100	.0031 ± .0009	.0088 ± .0044*	.0060 ± .0023	.0016 ± .0009
	D	001	.0139 ± .0019		.0128 ± .0033	.0158 ± .0029
	C	010	.9830 ± .0021	.9912 ± .0044	.9812 ± .0040	.9826 ± .0030
Kell	K	100	.0258 ± .0057	.0151 ± .0114	.0202 ± .0090	.0340 ± .0091
	k	010	.9442 ± .0079	.9542 ± .0160	.9470 ± .0146	.9390 ± .0117
	k^s	011	.0300 ± .0060	.0307 ± .0161	.0328 ± .0114	.0270 ± .0081
ABO	A_1	110	.1564 ± .0061	.1764 ± .0155	.1600 ± .0111	.1502 ± .0085
	A_2	010	.0514 ± .0039	.0295 ± .0073	.0632 ± .0078	.0515 ± .0055
	B	001	.0837 ± .0045	.0678 ± .0099	.0857 ± .0083	.0889 ± .0067
	O	000	.7085 ± .0077	.7263 ± .0182	.6911 ± .0143	.7094 ± .0108

Table 4.2.2. - Continued

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ	p ± σ
MNSsU	MS	1010	.1958 ± .0072	.1951 ± .0179	.2127 ± .0127	.1839 ± .0103
	Ms	1001	.3428 ± .0091	.3149 ± .0282	.3334 ± .0147	.3549 ± .0128
	M*	1000	.0109 ± .0048	.0347 ± .0281	--- **	.0136 ± .0062
	NS	0110	.0775 ± .0055	.0535 ± .0124	.0813 ± .0092	.0916 ± .0084
	Ns	0101	.3548 ± .0092	.3989 ± .0243	.3726 ± .0147	.3421 ± .0127
	N*	0100	.0182 ± .0053	.0029 ± .0200	--- **	.0139 ± .0062
Gm	a	10000	.2196 ± .0098	.2040 ± .0232	.1976 ± .0176	.2427 ± .0141
	ab	11000	.2249 ± .0097	.2318 ± .0238	.2322 ± .0176	.2137 ± .0136
	ax	10100	.0748 ± .0044	.0664 ± .0099	.0820 ± .0082	.0757 ± .0062
	abc	11010	.0636 ± .0040	.0668 ± .0099	.0579 ± .0069	.0652 ± .0058
	$b^{(1,2)}$	01001	.4171 ± .0078	.4310 ± .0189	.4303 ± .0141	.4027 ± .0112
Rh	cde	00011	.2862 ± .0105	.2972 ± .0251	.2985 ± .0192	.2707 ± .0149
	Cde	00101	.0182 ± .0039	.0270 ± .0111	.0132 ± .0059	.0157 ± .0053
	cdE	01010	.0027 ± .0016	--- **	--- **	.0037 ± .0026
	CdE	01100	.0007 ± .0009	--- **	--- **	.0034 ± .0017
	cDe	10011	.2284 ± .0101	.2043 ± .0235	.2283 ± .0184	.2364 ± .0146
	CDe	10101	.3255 ± .0082	.3053 ± .0200	.3300 ± .0192	.3358 ± .0118
	cDE	11010	.1359 ± .0057	.1662 ± .0141	.1300 ± .0098	.1342 ± .0081
	CDE	11100	.0024 ± .0015	--- **	--- **	.0001 ± .0025

* Genes, B and D, are pooled because of low frequency.

** The gene is omitted from the iteration process because of mathematical difficulty.

Table 4.2.3.

Gene frequency at sixteen polymorphic systems in northeastern Brazil.

(Total population in parent)

(p and σ stand for gene frequency and its standard error, respectively)

Distance x $\sqrt{\text{density}}$ →			[0-30)	[30-180)	[180-∞)
Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ
Secretor	Se	1	.5543 ± .0258	.5514 ± .0175	.5583 ± .0161
	se	0	.4457 ± .0258	.4486 ± .0176	.4417 ± .0161
Lewis	Le	1	.5376 ± .0254	.5474 ± .0174	.5185 ± .0158
	le	0	.4624 ± .0254	.4526 ± .0174	.4815 ± .0158
Lutheran	Lu^a	1	.0531 ± .0211	.0197 ± .0087	.0347 ± .0098
	Lu	0	.9469 ± .0211	.9803 ± .0087	.9653 ± .0098
PTC	T	1	.5864 ± .0261	.6530 ± .0183	.6174 ± .0166
	t	0	.4136 ± .0261	.3470 ± .0183	.3826 ± .0166
P	P_1	1	.6039 ± .0263	.6364 ± .0182	.6317 ± .0167
	P_2+p	0	.3961 ± .0263	.3636 ± .0182	.3683 ± .0167
Duffy	Fy^a	1	.2929 ± .0202	.2812 ± .0136	.2766 ± .0124
	Fy^-	0	.7071 ± .0202	.7188 ± .0136	.7234 ± .0124
Inv	Inv^a	1	.2063 ± .0176	.2018 ± .0119	.2198 ± .0114
	Inv	0	.7937 ± .0176	.7982 ± .0119	.7802 ± .0114
Diego	Di^a	1	.0087 ± .0086	.0241 ± .0097	.0268 ± .0088
	Di	0	.9913 ± .0086	.9759 ± .0097	.9732 ± .0088
Haptoglobin	1F	100	.1993 ± .0163	.2341 ± .0117	.2089 ± .0107
	1S	010	.2692 ± .0184	.2366 ± .0122	.2734 ± .0116
	2	001	.5315 ± .0214	.5294 ± .0140	.5177 ± .0132
Hemoglobin	A	10	.9747 ± .0065	.9727 ± .0046	.9691 ± .0045
	S+C	01	.0253 ± .0065	.0273 ± .0046	.0309 ± .0045
Transferrin	B+D	10	.0101 ± .0036	.0217 ± .0041	.0145 ± .0031
	C	01	.9899 ± .0036	.9783 ± .0041	.9855 ± .0031

Table 4.2.3. - Continued

Genetic system	Gene Name	Gene Binary	p ± σ	p ± σ	p ± σ
Kell	K	100	.0431 ± .0197	.0078 ± .0055	.0398 ± .0107
	k	010	.9219 ± .0231	.9725 ± .0103	.9285 ± .0136
	k^s	011	.0350 ± .0176	.0197 ± .0087	.0317 ± .0094
ABO	A_1	110	.1738 ± .0161	.1544 ± .0104	.1452 ± .0093
	A_2	010	.0598 ± .0108	.0601 ± .0072	.0508 ± .0061
	B	001	.0696 ± .0105	.0838 ± .0078	.0850 ± .0073
	O	000	.6967 ± .0203	.7017 ± .0131	.7190 ± .0121
MNSs	MS	1010	.1962 ± .0168	.2144 ± .0124	.1862 ± .0112
	Ms	1001	.3366 ± .0215	.3487 ± .0139	.3644 ± .0134
	NS	0110	.0858 ± .0136	.0819 ± .0088	.0933 ± .0090
	Ns	0101	.3814 ± .0225	.3550 ± .0137	.3561 ± .0133
Gm	a	10000	.2521 ± .0246	.1890 ± .0176	.2512 ± .0158
	ab	11000	.2074 ± .0228	.2390 ± .0175	.2057 ± .0151
	ax	10100	.1075 ± .0131	.0732 ± .0074	.0770 ± .0070
	abc	11010	.0674 ± .0105	.0599 ± .0067	.0641 ± .0064
	$b^{(1,2)}$	01001	.3656 ± .0186	.4389 ± .0135	.4020 ± .0127
Rh	cde	00011	.3088 ± .0252	.2796 ± .0186	.2735 ± .0165
	CDe	11001	.3352 ± .0196	.3404 ± .0140	.3381 ± .0132
	cDe	10011	.2141 ± .0238	.2288 ± .0180	.2395 ± .0162
	cDE	10110	.1307 ± .0137	.1361 ± .0095	.1330 ± .0087
	Cde	01001	.0112 ± .0079	.0151 ± .0062	.0159 ± .0060

Table 4.2.4.

Gene frequency at sixteen polymorphic systems in northeastern Brazil.

(Remote population in parent)

(p and σ stand for gene frequency and its standard error, respectively)

Distance x $\sqrt{\text{density}}$ →			[0-30)	[30-180)	[180-∞)
Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ
Secretor	Se	1	.5615 ± .0279	.5506 ± .0179	.5580 ± .0166
	se	0	.4385 ± .0279	.4494 ± .0179	.4420 ± .0166
Lewis	Le	1	.5255 ± .0272	.5155 ± .0248	.5138 ± .0162
	le	0	.4745 ± .0272	.4845 ± .0248	.4862 ± .0162
Lutheran	Lu^a	1	.0493 ± .0215	.0211 ± .0093	.0359 ± .0102
	Lu	0	.9507 ± .0215	.9789 ± .0093	.9641 ± .0102
PTC	T	1	.5949 ± .0282	.6598 ± .0189	.6141 ± .0171
	t	0	.4051 ± .0282	.3402 ± .0189	.3859 ± .0171
P	P_1	1	.5605 ± .0276	.6353 ± .0186	.6365 ± .0172
	P_2+p	0	.4395 ± .0276	.3647 ± .0186	.3635 ± .0172
Duffy	Fy^a	1	.2770 ± .0213	.2828 ± .0140	.2758 ± .0127
	Fy^-	0	.7230 ± .0213	.7172 ± .0140	.7242 ± .0127
Inv	Inv^a	1	.1986 ± .0186	.2004 ± .0122	.2125 ± .0115
	Inv^-	0	.8014 ± .0186	.7996 ± .0122	.7875 ± .0115
Diego	Di^a	1	.0000	.0214 ± .0095	.0278 ± .0092
	Di	0	1.0000	.9786 ± .0095	.9722 ± .0092
Haptoglobin	1F	100	.1920 ± .0175	.2412 ± .0121	.2097 ± .0112
	1S	010	.2720 ± .0196	.2332 ± .0124	.2694 ± .0119
	2	001	.5360 ± .0226	.5256 ± .0143	.5209 ± .0136
Hemoglobin	A	10	.9705 ± .0076	.9729 ± .0047	.9716 ± .0044
	S+C	01	.0295 ± .0076	.0271 ± .0047	.0284 ± .0044
Transferrin	B+D	10	.0136 ± .0051	.0230 ± .0043	.0152 ± .0033
	C	01	.9864 ± .0051	.9770 ± .0043	.9848 ± .0033

Table 4.2.4. - Continued

Genetic system	Gene Name	Binary	p ± σ	p ± σ	p ± σ
Kell	K	100	.0481 ± .0219	.0083 ± .0059	.0382 ± .0106
	k	010	.9126 ± .0258	.9749 ± .0102	.9289 ± .0138
	k^s	011	.0391 ± .0196	.0168 ± .0083	.0329 ± .0098
ABO	A_1	110	.1791 ± .0176	.1512 ± .0106	.1490 ± .0097
	A_2	010	.0551 ± .0112	.0622 ± .0075	.0496 ± .0063
	B	001	.0729 ± .0115	.0868 ± .0082	.0885 ± .0076
	0	000	.6929 ± .0219	.6998 ± .0134	.7129 ± .0126
MNSs	MS	1010	.1870 ± .0175	.2145 ± .0128	.1836 ± .0115
	Ms	1001	.3300 ± .0225	.3488 ± .0143	.3663 ± .0138
	NS	0110	.0885 ± .0153	.0844 ± .0091	.0951 ± .0093
	Ns	0101	.3945 ± .0245	.3523 ± .0139	.3550 ± .0137
Gm	a	10000	.2447 ± .0264	.1872 ± .0179	.2504 ± .0161
	ab	11000	.2211 ± .0253	.2390 ± .0178	.2072 ± .0155
	ax	10100	.1018 ± .0137	.0727 ± .0076	.0784 ± .0073
	abc	11010	.0672 ± .0112	.0605 ± .0070	.0640 ± .0066
	$b^{(1,2)}$	01001	.3652 ± .0203	.4406 ± .0138	.4000 ± .0131
Rh	cde	00011	.2995 ± .0273	.2795 ± .0192	.2987 ± .0194
	CDe	11001	.3485 ± .0217	.3346 ± .0144	.3420 ± .0151
	cDe	10011	.2119 ± .0258	.2302 ± .0186	.1963 ± .0185
	cDE	10110	.1269 ± .0148	.1401 ± .0098	.1408 ± .0093
	Cde	01001	.0132 ± .0093	.0156 ± .0064	.0222 ± .0086

Table 4.2.5.

Gene frequency at sixteen polymorphic systems in northeastern Brazil

(p and σ stand for gene frequency and its standard error, respectively)

Genetic system	Gene Name	Gene Binary	Parents whose degree of consanguinity is unknown p ± σ	Children Total p	Children Remote p
Secretor	Se	1	.5296 ± .0400	.5456	.5438
	se	0	.4704 ± .0400	.4544	.4562
Lewis	Le	1	.5565 ± .0406	.5216	.5175
	le	0	.4435 ± .0406	.4784	.4825
Lutheran	Lu^a	1	.0253 ± .0250	.0319	.0329
	Lu	0	.9747 ± .0250	.9681	.9671
PTC	T	1	.5565 ± .0406	.6060	.6060
	t	0	.4435 ± .0406	.3940	.3940
P	P_1	1	.6613 ± .0426	.6118	.6185
	P_2+p	0	.3387 ± .0426	.3882	.3815
Duffy	Fy^a	1	.3535 ± .0345	.2819	.3034
	Fy^-	0	.6465 ± .0345	.7181	.6966
Inv	Inv^a	1	.1702 ± .0253	.2153	.2112
	Inv^-	0	.8298 ± .0253	.7847	.7888
Diego	Di^a	1	--- **	.3135*	.3118*
	Di	0	--- **	.6865*	.6882*
Haptoglobin	1F	100	.2193 ± .0278	.2133	.2164
	1S	010	.2456 ± .0266	.2546	.2492
	2	001	.5351 ± .0324	.5321	.5344
Hemoglobin	A	100	.9788 ± .0095	.7558***	.7485***
	S	010	.0212 ± .0095	.1802***	.1835***
	C	001		.0640***	.0680***
Transferrin	B	100	--- **	.0026	.0026
	D	001	.9877 ± .0071	.0145	.0151
	C	010	.0123 ± .0071	.9829	.9823

Table 4.2.5. - Continued

Genetic system	Gene Name	Gene Binary	Parents whose degree of consanguinity is unknown	Children Total	Children Remote
			p ± σ	p	p
Kell	K	100	.0410 ± .0130	.0273	.0283
	k_s	010	.9590 ± .0130	.9440	.9423
	k^s	011		.0287	.0293
ABO	A_1	110	.1790 ± .0261	.1541	.1523
	A_2	010	.0561 ± .0164	.0560	.0552
	B^2	001	.0587 ± .0155	.0820	.0862
	0	000	.7062 ± .0298	.7079	.7063
MNSs	MS	1010	.2235 ± .0260	.2012	.1999
	Ms	1001	.3339 ± .0301	.3359	.3347
	M*	1000	--- **	.0186	.0192
	NS	0110	.0470 ± .0169	.0749	.0747
	Ns	0101	.3956 ± .0328	.3558	.3570
	N*	0100	--- **	.0136	.0145
Gm	a	10000	.2154 ± .0481	.2363	.2322
	ab	11000	.1598 ± .0394	.2099	.2116
	ax	10100	.0985 ± .0208	.0739	.0728
	abc	11010	.0349 ± .0123	.0575	.0586
	$b^{(1,2)}$	01001	.4914 ± .0336	.4224	.4248
Rh	cde	00011	.3439 ± .0445	.2862	.2848
	Cde	00101	.0440 ± .0201	.0143	.0149
	cdE	01010	--- **	.0023	.0025
	CdE	01100	--- **	.0012	.0013
	cDe	10011	.1888 ± .0398	.2330	.2348
	CDe	10101	.3044 ± .0355	.3192	.3163
	cDE	11010	.1189 ± .0198	.1424	.1448
	CDE	11100	--- **	.0014	.0006

* Conditional probability that at least one of parents is Diego positive.

** Ignored because of rare frequency.

*** Conditional probability that both parents are not genotype AA.

Table 4.3.1.

Code of inbreeding.

Code	Inbreeding coefficient	
c	F	[a,b)
0	0	no known consanguinity
3	1/8	[.1768 - .0884)
4	1/16	[.0884 - .0442)
5	1/32	[.0442 - .0221)
6	1/64	[.0221 - .0110)
7	1/128	[.0110 - .0055)
8	1/256	[.0055 -0.]
9	F > 0	but degree unknown

Where $c = -\log_2 F$

$a = -\text{antilog}_2 (c-1/2)$

and

$b = -\text{antilog}_2 (c+1/2)$

Table 4.3.2.

Distribution of couples by the coefficient of consanguinity and the marital distance.

(Parental generation)

F = coefficient of consanguinity, α = estimated inbreeding coefficient, σ = standard error of α.

(i) Total Population

Distance (km.)	F → 0	1/8	1/16	1/32	1/64	1/128	1/256	Degree unknown (F > 0)	Total		$\alpha \pm \sigma$	
[0-.5)	20	1	5		2			3	31	206	.0157 ± .0055	.0131 ± .0018
[.5-3.5)	113	2	25	4	10	4	2	15	175		.0127 ± .0019	
[3.5-9.5)	98	1	13	3	5	1	2	14	137	326	.0088 ± .0018	.0061 ± .0010
[9.5-27.5)	155	2	4	3	5	2	1	17	189		.0042 ± .0012	
[27.5-81.5)	217	1	9	3	3	1	1	9	244		.0037 ± .0009	
[81.5-243.5)	161		7	3		1		3	175	510	.0031 ± .0010	.0031 ± .0006
[243.5-729.5)	62		2	1					65		.0024 ± .0014	
[729.5-∞)	26								26		.0000	
unknown	26								26			
Total	878	7	65	17	25	9	6	61	1068		.0059 ± .0006	
Grand parental generation	1462	5	19	13	14	0	1	129	1703		.0036 ± .0004	

Table 4.3.2. - Continued

(ii) Close Population ($F \geq 1/32$)

Distance (km.)	1/8	1/16	1/32	0	Total	$\alpha \pm \sigma$	
[0-.5)	1	5		25	31	.0141 ± .0056	.0115 ± .0018
[.5-3.5)	2	25	4	144	175	.0111 ± .0019	
[3.5-9.5)	1	13	3	120	137	.0075 ± .0018	.0050 ± .0010
[9.5-27.5)	2	4	3	180	189	.0031 ± .0012	
[27.5-81.5)	1	9	3	231	244	.0032 ± .0009	
[81.5-243.5)		7	3	165	175	.0030 ± .0010	.0030 ± .0006
[243.5-729.5)		2	1	62	65	.0024 ± .0014	
[729.5- ∞)				26	26	.0000	
unknown				26	26		
Total	7	65	17	979	1068	.0051 ± .0006	

(iii) Remote Population ($F < 1/32$)

Distance (km.)	$\alpha \pm \sigma$	
[0-.5)	.0020 ± .0009	.0020 ± .0003
[.5-3.5)	.0020 ± .0003	
[3.5-9.5)	.0015 ± .0003	.0012 ± .0002
[9.5-27.5)	.0011 ± .0002	
[27.5-81.5)	.0005 ± .0001	
[81.5-243.5)	.0002 ± .0001	.0003 ± .0001
[243.5-729.5)	.0000	
[729.5- ∞)	.0000	
Total	.0009 ± .0001	

Table 4.3.3.

Distribution of couples by the coefficient of consanguinity and distance x $\sqrt{\text{density}}$.

(Parental generation)

(i) Total Population

Distance x $\sqrt{\text{density}}$	F →	0	1/8	1/16	1/32	1/64	1/128	1/256	Degree unknown (F > 0)	Total	$\alpha \pm \sigma$
[0.-.5)		20	1	5		2			3	31	.0157 ± .0055
[.5-3.5)		24		8	1	2	1	1		37	.0155 ± .0042
[3.5-9.5)		62	2	11	2	4	1	1	11	94	.0121 ± .0027
[9.5-27.5)		101		16	4	7	3	2	13	146	.0098 ± .0016
[27.5-81.5)		148	3	6	3	6	2	1	20	189	.0057 ± .0014
[81.5-243.5)		184		9	1	3	1	1	10	209	.0034 ± .0009
[243.5-729.5)		182	1	6	3		1		4	197	.0032 ± .0010
[729.5- ∞)		124		4	2					130	.0024 ± .0010
unknown		33			1	1				35	.0013 ± .0010
Total		878	7	65	17	25	9	6	61	1068	.0059 ± .0006

Table 4.3.3. - Continued

(ii) Close Population ($F \geq 1/32$)							(iii) Remote Population ($F < 1/32$)
Distance x $\sqrt{\text{density}}$ $F \rightarrow$	1/8	1/16	1/32	0	Total	$\alpha \pm \sigma$	$\alpha \pm \sigma$
[0-.5)	1	5		25	31	.0141 ± .0056	.0020 ± .0009
[.5-3.5)		8	1	28	37	.0144 ± .0043	.0015 ± .0008
[3.5-9.5)	2	11	2	79	94	.0106 ± .0027	.0018 ± .0004
[9.5-27.5)		16	4	126	146	.0077 ± .0017	.0017 ± .0004
[27.5-81.5)	3	6	3	177	189	.0045 ± .0014	.0013 ± .0003
[81.5-243.5)		9	1	199	209	.0028 ± .0009	.0006 ± .0002
[243.5-729.5)	1	6	3	187	197	.0030 ± .0010	.0002 ± .0001
[729.5- ∞)		4	2	124	130	.0024 ± .0010	.0000
unknown			1	34	35	.0009 ± .0003	.0005 ± .0005
Total	7	65	17	979	1068	.0051 ± .0006	.0009 ± .0001

Table 4.3.4.

Bioassay of the inbreeding coefficient from individual parental phenotype frequencies (G-TYPE).

Genetic system	(i) Total Population: Inbreeding coefficient α	Score U_α	Information K_α	χ^2	(ii) Remote Population: Inbreeding coefficient α	Score U_α	Information K_α	χ^2
Haptoglobin	.0421	100.0718	3971.8275	2.50	.0385	93.0729	3666.8614	2.32
Hemoglobin	-35.5618	-59.4975	1.6463	2150.24**	-35.5632	-54.3034	1.5086	1954.68**
Transferrin	-59.3274	-35.7096	.5932	2149.56**	-57.0635	-33.8968	.5878	1954.75**
Kell	-14.3742	-11.0863	.7505	163.76**	-13.9683	-10.4429	.7292	149.55**
A_1A_2BO	-.5180	-6.6584	.0108	4087.50**	-4.3739	-4.6985	.0062	3532.55**
MNSsU	.0262	-51.8931	3611.5645	.75^^	.0236	-29.9755	3423.1910	.26^^
Gm	.0480	52.3557	2187.0116	1.26	.0414	44.7351	2008.2060	1.00
Rh	-.0079	-105.9314	3893.2823	2.92	-.0112	-94.3918	3613.4341	2.49
With eight systems:								
	α = .0170 ±.0086	-118.3488	13666.6867	8558.49** --	α = .0132 ±.0089	-89.9009	12714.5243	7597.60** --
Without four systems being significant with 1 percent level:								
	α = .0246 ±.0086	-5.3970	13663.6859	7.43	α = .0208 ±.0089	13.4407	12711.6925	6.07

** 1 percent level significant

^^

Table 4.3.5.

Bioassay of the inbreeding coefficient from mating type frequencies.

(MATYPE)

(i) Total Population

Genetic system	α	U_α	K_α	χ^2
Secretor	-.0930	-122.3843	1085.8265	13.79**
Lewis	.0267	10.9124	1574.1185	.08
Lutheran	-7.2940	-30.5751	4.0075	233.27**
PTC	.0222	-2.9862	1067.7441	.01
P	.0258	.4378	1072.2795	.00
Duffy	.0527	111.6582	2941.3587	4.24*
Inv	.0251	35.6936	3270.4829	.39
Diego	-10.8141	-20.1046	1.7786	227.26**
Haptoglobin	.0205	85.3637	11843.341	.62
Hemoglobin	.0088	-25.9538	5621.2518	.12
Transferrin	.0237	83.2232	8143.9629	.85
Kell	-3.1883	-63.6790	19.0722	212.61**
ABO	.0085	-48.2697	8976.2461	.26
MNSsU	.0124	-76.6036	12041.649	.49
Gm	.0209	90.9479	13688.338	.60
Rh	-.0011	-174.5448	11749.369	2.59
	α = .0133 ± .0035	-146.8643	83100.8263	697.18**
Removed four systems (see text).				
	.0160 ± .0035	89.8787	81990.1415	10.25

* 5 percent level significant

** 1 percent level significant

Table 4.3.5. - Continued

(ii) Remote Population

Genetic system	α	U_α	K_α	χ^2
Secretor	-.0850	-103.8535	1070.1724	10.08**
Lewis	.0237	17.9954	1542.5702	.21..
Lutheran	-7.5827	-27.4778	3.5230	214.32**
PTC	.0150	-.1847	1046.3972	.00..
P	-.0160	-28.4140	909.2341	.89
Duffy	.0405	85.3153	2716.1083	2.68
Inv	.0183	28.9092	3014.0368	.28
Diego	-11.7126	-17.3511	1.4360	209.65**
Haptoglobin	.0181	112.9659	11476.496	1.11..
Hemoglobin	-.0014	-55.4074	5677.6502	.54
Transferrin	.0190	138.4814	13258.250	1.45
Kell	-3.1781	-58.9556	17.9991	193.11**
ABO	.0057	-23.4637	8311.9101	.07..
MNSsU	.0080	-46.2461	11482.968	.19
Gm	.0146	72.4727	12786.132	.41
Rh	-.0089	-189.2872	10901.084	3.29
	α = .0082 ± .0034	-94.5012	84215.9674	638.28** ..

Removed four systems (see text).

.0106 ± .0035	113.1368	83122.8369	11.12

** 1 percent level significant

..

Table 4.3.6.

Bioassay of the inbreeding coefficient in children (G-TYPE).

(i) Total Population					(ii) Remote Population			
Genetic system	α	U_α	K_α	χ^2	α	U_α	K_α	χ^2
Haptoglobin	.0435	265.8770	8494.8494	8.32**	.0400	254.7923	7796.9463	8.33**
Transferrin	.0175	29.7645	5946.5908	.15^^	.0141	74.8080	11615.3595	.48^^
Kell	-15.7171	-27.3352	1.7075	437.61**	-14.8998	-24.8450	1.6547	373.04**
ABO	5.3686	-10.4590	.0122	-- ^^	4.0736	-6.0841	.0043	-- ^^
MNSsU	.0148	-48.7718	10059.1870	.24	.0096	-25.5608	9751.3988	.07
Gm	.0078	-42.4787	4718.7649	.38	.0081	-9.5117	4310.3675	.02
Rh	-.0197	-293.5307	8736.9353	9.86**	-.0354	-341.9613	7799.5369	14.99**
	α = .0121	-126.9339	37958.0471	^^	α = .0073	-78.3626	41275.2680	^^

Note the instability at the ABO locus.

The hemoglobin system is excluded from analysis because only children whose both parents are not AA were typed.

* 5 percent level significant

** 1 percent level significant

^^

Table 4.3.7.

Bioassay of the inbreeding coefficient for consanguineous marriages whose degree is unknown.

(MATYPE)

Genetic system	α	U_α	K_α	χ^2
Secretor	.0006	-1.0749	91.0790	.01
Lewis	-.0376	-3.7578	74.5953	.19
Lutheran	-9.4295	-1.0944	.1123	10.67**
PTC	-.0376	-3.7578	74.5953	.19
P	-7.7110	-8.1236	1.0236	64.47**
Duffy	-.0731	-12.1221	146.0468	1.01
Inv	-.0413	-9.5677	188.2109	.49
Haptoglobin	.0458	39.3717	1060.5684	1.46
Hemoglobin	-7.5015	-8.0435	1.0461	61.85**
Transferrin	-13.0526	-4.7883	.3580	64.03**
Kell	-3.8025	-16.3874	4.1999	63.94**
ABO	.0643	48.6854	881.9926	2.69
MNSs	.0167	5.7452	767.4288	.04
Gm	.0267	10.1208	483.4427	.18
Rh	-.0754	-39.4076	465.7609	3.33
	α = .0086 ± .0152	-4.2020	4340.4606	274.55**

** 1 percent level significant

Table 4.3.8.

THE TOTAL INBREEDING COEFFICIENT IN NORTH-EASTERN BRAZIL

METHOD \ GENERATION	PARENT	CHILDREN	
PEDIGREE	.0036±.0004	.0059±.0006	
BIOASSAY (GENOTYPE)	.0170±.0086	.0121	UNKNOWN CONSANGUINITY
BIOASSAY (MATING TYPE)	———	.0133±.0035	.0086±.0152

Table 4.3.9.

Components of the total inbreeding coefficient of children in northeastern Brazil.

Source	Contribution to α	%
Close consanguinity $F \geq 1/32$ (From Pedigree Study)	.0051	38
Remote consanguinity $F < 1/32$ (From Pedigree Study)	.0009	6
Racial endogamy	.0029	22
Other	.0045	34
Total (From Bioassay)	.0133	100

Table 4.3.10.

Inbreeding coefficient with marital distance.

[I. $f = ae^{-bx}$, II. $f = ae^{-bx}/\sqrt{x}$]

(i) Total Population				(ii) Remote Population			(iii) Close Population		
Distance (km.)	Inbreeding coefficient Bioassay (A)	Bioassay (B)	Pedigree Study+	Bioassay (A)	Bioassay (B)	Pedigree Study+	Bioassay (A)	Bioassay (B)	Pedigree Study+
.0	--	--	.0157	--	--	.0020	--	--	.0141
1.7	.0150	.0103	--	-.0028	-.0113	--	.0178	.0216	--
2.0	--	--	.0127	--	--	.0020	--	--	.0111
6.5	--	--	.0088	--	--	.0015	--	--	.0075
12.1	.0241	.0199	--	.0244	.0198	--	-.0003	.0001	--
16.1	--	--	.0042	--	--	.0011	--	--	.0031
49.9	--	--	.0037	--	--	.0005	--	--	.0032
134.2	--	--	.0032	--	--	.0002	--	--	.0030
167.2	.0110	.0082	--	.0043	.0044	--	.0067	.0038	--
408.5	--	--	.0024	--	--	--	--	--	.0024
I.									
a	.0212 ± .0058	.0169 ± .0056	.0085 ± .0011	.0147 ± .0062	.0076 ± .0056	.0018 ± .0002	.0064 ± .0056	.0086 ± .0057	.0064 ± .0010
b	.0038 ± .0033	.0041 ± .0042	.0130 ± .0040	.0062 ± .0066	.0023 ± .0077	.0233 ± .0051	-.0000 ± .0072	.0064 ± .0114	.0085 ± .0036
χ^2 for goodness of fit	.80 (df=1)	.91 (df=1)	19.41** (df=5)	5.90* (df=1)	8.55** (df=1)	2.20 (df=4)	2.66 (df=1)	3.85* (df=1)	17.73** (df=5)

Table 4.3.10. - Continued

II.									
a	--	--	--	--	--	.0034 ± .0004	.0201 ± .0138	.0584 ± .5604	--
b	--	--	--	--	--	.0032 ± .0031	.0143 ± .1203	.4171 ± 5.5374	--
χ^2	--	--	--	--	--	3.52	2.38	.57	--

* 5 percent level significant

** 1 percent level significant

\+ Distance zero is omitted in study II.

Table 4.3.11.

Inbreeding coefficient with marital distance x $\sqrt{\text{density}}$.

[I. $f = ae^{-bx}$, II. $f = ae^{-bx}/\sqrt{x}$]

(i) Total Population				(ii) Remote Population			(iii) Close Population		
Distance x $\sqrt{\text{density}}$	Inbreeding coefficient Bioassay (A)	Bioassay (B)	Pedigree Study+	Bioassay (A)	Bioassay (B)	Pedigree Study+	Bioassay (A)	Bioassay (B)	Pedigree Study+
.0	--	--	.0157	--	--	.0020	--	--	.0141
2.2	--	--	.0155	--	--	.0015	--	--	.0144
6.5	--	--	.0121	--	--	.0018	--	--	.0106
10.8	.0201	.0206	--	.0178	.0169	--	.0023	.0038	--
17.1	--	--	.0092	--	--	.0017	--	--	.0077
50.1	--	--	.0057	--	--	.0013	--	--	.0045
86.5	.0179	.0144	--	.0178	.0135	--	.0001	.0009	--
153.6	--	--	.0034	--	--	.0006	--	--	.0028
422.9	--	--	.0032	--	--	.0002	--	--	.0030
897.5	.0074	.0038	--	-.0010	-.0038	--	.0064	.0076	--
1939.4	--	--	.0024	--	--	--	--	--	.0024
I.									
a	.0199 ± .0061	.0187 ± .0066	.0102 ± .0013	.0215 ± .0094	.0193 ± .0119	.0018 ± .0002	.0014 ± .0115	.0029 ± .0113	.0051 ± .0007
b	.0011 ± .0010	.0019 ± .0021	.0062 ± .0017	.0034 ± .0054	.0050 ± .0096	.0062 ± .0012	.0059 ± .1371	.0060 ± .0652	.0067 ± .0004
χ^2	.00 (df=1)	.12 (df=1)	16.83** (df=6)	.26 (df=1)	.51 (df=1)	.85 (df=5)	1.20 (df=1)	1.70 (df=1)	19.68** (df=6)

Table 4.3.11. - Continued

II.									
a	--	--	--	.0812 ± .0308	.0713 ± .0338	.0051 ± .0007	.0072 ± .0397	--	--
b	--	--	--	.0004 ± .0034	.0014 ± .0080	.0003 ± .0009	.0052 ± .1799	--	--
χ^2	--	--	--	2.77	1.84	15.08**	1.19	--	--

** 1 percent level significant

+ Zero class is omitted in study II.

Table 4.3.12.

Estimation of systematic and migration pressures from inbreeding function.

One Dimensional Model

	Migration pressure	Systematic pressure
Distance study (σ = 20)		
Total bioassay A	12.31 ± 4.01	.0011 ± .0000
Remote pedigree	17.40 ± 2.79	.0824 ± .0140
Distance x $\sqrt{\text{density}}$ (σ = 1)		
Total bioassay A	105.57 ± 58.18	.0068 ± .0055
Remote pedigree	150.62 ± 22.19	.4313 ± .0689
Remote bioassay A	57.80 ± 55.42	.0194 ± .0251
Close bioassay A	174.69 ± 2638.30	.5280 ±9.3326

Table 4.3.13.

Inbreeding coefficient for alleles.

(MATYPE method)

System	Gene	α	U_α	K_α	χ^2
Haptoglobin	Hp_{1S}^{1F}	.0314	222.6620	7102.4668	6.98**
	Hp_2	.0094	62.2098	6620.1251	.58
	Hp	.0122	72.9228	5973.7173	.89
		.0080	(.0088)		
Hemoglobin	A	.0081	56.5935	7029.8098	.46
	S	.0039	26.5355	6783.7092	.10
	C	.0055	100.7670	18195.993	.56
		.0156	(.0205)		
Transferrin	C	.0119	146.0154	12286.561	1.74
	B	.0040	302.8376	76261.795	1.20
	D	.0103	147.2000	14248.070	1.52
		.0118	(.0237)		
Kell	k	.0268	4.7334	176.4688	.13
	k^s	-8.1823	-26.1477	3.1964	213.90**
	K	-6.3182	-33.8704	5.3608	214.00**
		-.3793	(-3.1883)		
ABO	A_1	.0204	75.0759	3674.7279	1.53
	A_2	.0008	3.2667	4122.3860	.00
	B	.0051	20.6541	4046.0060	.11
	O	.0079	23.1313	2935.2961	.18
		.0093	(.0085)		
MNSs	Ms	.0251	52.6489	2096.2716	1.32
	MS	-.0010	-3.1271	3217.1407	.00
	M*	-1151.6108	-1.1153	.0008	1550.82**
	NS	-.0030	- 10.5522	3536.3054	.03
	Ns	-.0038	-8.9054	2341.1535	.03
	N*	-266.7327	-4.0305	.0150	1086.26**
		-16.9049	(.0124)		
Gm	a	.0522	91.0640	1743.7084	4.76*
	ax	-.0085	-30.3458	3563.4585	.26
	ab	.0406	84.2559	2074.1880	3.42
	abc	.0157	70.6034	4511.5305	1.10
	b	.0311	73.0562	2349.6991	2.27
		.0268	(.0209)		

Table 4.3.13. - Continued

System	Gene	α	U_α	K_α	χ^2
Rh	r				
	r'				
	r''	-.0224	-58.3296	2608.2877	1.30
	r'''				
	R_0	.0011	4.1194	3860.8571	.00
	R_1	.0038	10.2876	2696.4591	.04
	R_2	.0128	47.9146	3757.4264	.61
	R^z	-151.6496	-7.0290	.0462	1068.92**
		-.3828 (-.0011)			^^

* 5 percent level significant

** 1 percent level significant

^^

Table 4.3.14.

Estimated tri-racial gene frequencies (Krieger et al., 1965), with the equivalent inbreeding coefficient f [$= \sigma^2/p(1-p)$].
(The proportions of Negro, Indian and Caucasian are .301, .114 and .585, respectively).

System	Gene	Negro	Indian	Caucasian	mean p	f	$Q=\Sigma pf$
Secretor	Se	.538	.500	.570	.552	.0023	.0023
	se	.462	.500	.430	.448	.0023	
Lewis	Le	.319	.545	.660	.544	.0932	.0932
	le	.681	.455	.340	.456	.0932	
Lutheran	Lu_a^-	.964	1.000	.964	.968	.0042	.0042
	Lu^a	.036	.000	.036	.032	.0042	
PTC	t	.207	.207	.506	.382	.0920	.0920
	T	.793	.793	.494	.618	.0920	
P	P_2+p	.246	.570	.458	.407	.0512	.0512
	P_1	.754	.430	.542	.593	.0512	
Duffy	Fy_a^-	1.000	.318	.603	.691	.2301	.2301
	Fy^a	.000	.682	.397	.309	.2301	
Inv	Inv_a^-	.684	.693	.900	.811	.0723	.0723
	Inv^a	.316	.307	.100	.189	.0723	
Diego	Di_a^-	1.000	.830	1.000	.981	.1550	.1550
	Di^a	.000	.170	.000	.019	.1550	
Haptoglobin	Hp_2^1	.624	.731	.384	.496	.0743	.1048
	Hp_{2m}	.235	.251	.616	.459	.1387	
	Hp	.141	.018	.000	.045	.0946	
Hemoglobin	A	.910	1.000	1.000	.973	.0644	.0637
	S	.066	.000	.000	.020	.0468	
	C	.024	.000	.000	.007	.0167	
Transferrin	C	.939	.996	.994	.977	.0294	.0295
	B	.000	.002	.006	.004	.0020	
	D	.061	.002	.000	.019	.0422	

Table 4.3.14. - Continued

System	Gene	Negro	Indian	Caucasian	mean p	f	$\alpha=\Sigma pf$
Kell	k_s	.851	1.000	.952	.927	.0400	.0421
	k^s	.144	.000	.000	.043	.1046	
	K	.005	.000	.048	.030	.0167	
ABO	A_1	.105	.000	.236	.170	.0505	.0538
	A_2	.052	.000	.068	.055	.0085	
	B^2	.150	.000	.066	.084	.0230	
	O	.693	1.000	.630	.691	.0612	
MNSsU	MS	.118	.236	.270	.220	.0269	.0148
	Ms	.358	.462	.310	.342	.0103	
	M*	.052	.000	.000	.016	.0368	
	NS	.058	.107	.076	.074	.0030	
	Ns	.336	.195	.344	.325	.0099	
	N*	.078	.000	.000	.023	.0556	
Gm	a	.000	.775	.204	.207	.3028	.4469
	ax	.000	.123	.092	.068	.0329	
	ab	.786	.102	.000	.249	.6708	
	abc	.214	.000	.000	.065	.1592	
	$b^{(1,2)}$	.000	.000	.704	.411	.4969	
Rh	r	.113	.000	.404	.270	.1330	.1440
	r^v	.075	.000	.000	.023	.0533	
	r'	.000	.000	.008	.005	.0033	
	r'^v	.026	.000	.000	.008	.0181	
	r''	.006	.000	.005	.005	.0007	
	r^y	.000	.000	.001	.001	.0004	
	R_0	.475	.016	.053	.176	.2661	
	R_0^2	.150	.000	.000	.045	.1092	
	R_1	.082	.533	.415	.327	.1245	
	R_2	.073	.439	.110	.136	.1026	
	R^z	.000	.012	.004	.004	.0033	
G-6-P-D	S	.609	.930	1.000	.874	.2801	.2620
	F	.221	.000	.000	.067	.1646	
	F^d	.170	.067	.000	.059	.1036	

Table 4.4.1.

Mating type frequency at two allelic loci without dominance.

System	Mating type	Observed	Expected α=.0133	Expected α=0
Ss (S=.28)	SS x SS	10	8.03	6.69
	SS x Ss	65	72.34	68.14
	SS x ss	98	85.60	86.79
	Ss x Ss	165	171.21	173.59
	Ss x ss	433	431.39	442.19
	ss x ss	288	290.43	281.60
	Total	1059	χ^2_4 = 3.28	χ^2_4 = 3.99
Haptoglobin (Hp^1=.48)	1-1 x 1-1	51	57.36	52.73
	1-1 x 1-2	248	233.76	232.81
	1-1 x 2-2	123	125.09	128.50
	1-2 x 1-2	242	250.19	256.99
	1-2 x 2-2	286	282.64	283.69
	2-2 x 2-2	83	83.96	78.29
	Total	1033	χ^2_4 = 1.93	χ^2_4 = 2.47
MN (M=.55)	MM x MM	98	104.57	98.21
	MM x MN	326	317.56	319.88
	MM x NN	126	123.42	130.23
	MN x MN	258	257.06	260.45
	MN x NN	208	213.96	212.06
	NN x NN	48	47.42	43.17
	Total	1064	χ^2_4 = .65	χ^2_4 = .90
Hemoglobin (A=.97) (S*=S+C)	AA x AA	935	935.49	933.43
	AA x AS*	100	99.21	103.21
	AA x S*S*	0	2.03	1.43
	AS* x AS*	6	4.08	2.85
	AS* x S*S*	0	.19	.08
	S*S* x S*S*	0	.00	.00
	Total	1041	χ^2_4 = 3.13	χ^2_4 = 5.99

Table 4.4.1. - Continued

System	Mating type	Observed	Expected α=.0133	Expected α=0
Transferrin (C=.98) (B*=B+D)	CC x CC	980	978.87	977.58
	CC x CB*	60	62.25	64.79
	CC x B*B*	0	.94	.54
	CB* x CB*	4	1.88	1.07
	CB* x B*B*	0	.06	.02
	B*B* x B*B*	0	.00	.00
	Total	1044	$\chi^2_4 = 3.47$	$\chi^2_4 = 8.92$

Table 4.4.2.

Mating type frequency at two allelic loci with complete dominance.

System	Mating type	Observed	Expected α=.0133	Expected α=0
Diego (Di^a=.02)	a x a	0	.61	.39
	a x -	18	16.68	17.22
	- x -	190	190.71	190.39
	Total	208	χ_1^2 = .72	χ_1^2 = .43
Lutheran (Lu^a=.03)	a x a	0	1.17	.85
	a x -	27	24.51	25.29
	- x -	187	188.31	187.85
	Total	214	χ_1^2 = 1.44	χ_1^2 = .97
Inv (Inv^a=.21)	a x a	158	151.12	147.09
	a x -	467	476.20	488.81
	- x -	417	414.67	406.10
	Total	1042	χ_1^2 = .50	χ_1^2 = 2.07
Duffy (Fy^a=.28)	a x a	275	254.77	251.70
	a x -	485	519.64	531.60
	- x -	304	289.59	280.70
	Total	1064	χ_1^2 = 4.63*	χ_1^2 = 8.18**
Lewis (Le=.53)	Le x Le	652	643.65	646.00
	Le x le	351	360.75	363.01
	le x le	57	55.60	51.00
	Total	1060	χ_1^2 = .41	χ_1^2 = 1.16
Secretor (Se=.55)	Se x Se	668	678.26	681.11
	Se x se	365	337.44	338.76
	se x se	29	46.31	42.12
	Total	1062	χ_1^2 = 8.87**	χ_1^2 = 6.37*
PTC (T=.62)	T x T	788	781.22	785.03
	T x t	251	257.96	256.95
	t x t	24	23.82	21.03
	Total	1063	χ_1^2 = .25	χ_1^2 = .57

Table 4.4.2. - Continued

System	Mating type	Observed	Expected α=.0133	Expected α=0
P (P_1=.63)	$P_1 \times P_1$	792	784.67	788.59
	$P_1 \times P_2$	248	255.97	254.83
	$P_2 \times P_2$	24	23.36	20.59
	Total	1064	χ_1^2 = .33	χ_1^2 = .76

* 1 percent level significant

** 5 percent level significant

Table 4.4.3

Effects of inbreeding (α=.0133) on h.

Factor	Gene frequency p	h_2 (α=0)	Δh	Incompatible backcross $e = \frac{U_h}{K_{hh}}$	K_{hh}	Compatible backcross e	K_{hh}	h_3 (α=0)	Δh	Incross with respect to dominance e	K_{hh}
Se	.5537	.3828	-.0084	-.0700	566	-.0427	571	.6191	.0083	.0031	1555
Le	.5317	.3621	-.0073	-.0283	554	-.0559	603	.5931	.0092	.0134	1416
Gm^x	.0762	.0396	.0055	-.0227	1473	-.0021	1295	.0776	.0247	.0135	68
Gm^c	.0643	.0332	.0057	.0254	1163	-.0465	1529	.0653	.0250	-.0925	40
Inv^a	.2099	.1173	.0031	.0097	1346	-.0033	1299	.2208	.0209	-.0110	317
T	.6250	.4545	-.0123	-.0931	245	.0504	264	.7024	.0054	-.0332	1225
A_1	.1566	.0850	.0042	-.0007	1685	.0236	1311	.1628	.0225	.0662	213
A_2	.0624	.0322	.0058	-.0283	714	.0451	994	.0634	.0251	.5000	16
B	.0808	.0421	.0055	-.0243	1454	-.0093	1677	.0824	.0246	-.0627	75
D	.6923	.5294	-.0168	-.0054	701	-.0596	272	.7785	.0028	.0113	2971
P	.6270	.4567	-.0124	-.0660	379	-.0454	1026	.7048	.0053	-.0109	2162
Fy^a	.2833	.1650	.0014	.0374	1118	-.0163	1111	.3028	.0186	.1156	546
Lu^a	.0321	.0163	.0062	.0150	160	-.0584	226	--	--	--	--
Di^a	.0219	.0111	.0064	-.0397	156	-.0221	190	--	--	--	--

Table 4.4.3. - Continued

Factor	Gene frequency p	h_2 (α=0)	Δh	Incompatible backcross $e = \frac{U_h}{K_{nh}}$	K_{nh}	Compatible backcross e	K_{nh}	h_3 (α=0)	Δh	Incross with respect to dominance e	K_{nh}
f	.5168	.3484	-.0067	-.0412	199	-.0257	183	.5754	.0098	.0276	456
v	.0840	.0438	.0054	-.0678	478	-.0719	416	.0857	.0245	.0000	20
k^s	.0296	.0150	.0062	.0498	261	-.0547	223	--	--	--	--
				$\bar{\Delta}h$ = -.0114 ± .0326 $\bar{e}$ = .0013 ± .0068 r = .35		$\bar{\Delta}h$ = -.0158 ± .0323 $\bar{e}$ = .0013 ± .0066 r = .33				$\bar{\Delta}h$ = .0080 ± .0381 $\bar{e}$ = .0076 ± .0054 r = .46	

Δh = the expected deviation due to inbreeding (α= .0133)
e = the observed deviation
K_{nh} = weight of observation
r = correlation coefficient between Δh and e

Table 4.5.1.

Distribution of couples with marital distance in grandparent.

Distance (km.)	Observed x	Observed	Expected I	Expected II	Expected III	Expected IV	Expected V	Expected VI	Expected VII	Expected VIII	Expected IX
.0 - .5	0	74	17.76	4.10	747.18	23.98	59.70	.28	67.07	--	68.60
.5 - 3.5	2.04	385	102.72	24.79	297.80	114.83	516.78	8.64	308.98	--	325.62
3.5 - 9.5	6.45	272	187.04	50.56	159.85	172.04	299.12	38.32	355.83	--	363.97
9.5 - 27.5	16.13	348	438.63	158.18	155.34	336.66	237.55	189.29	420.34	--	324.92
27.5 - 81.5	49.93	338	648.38	492.81	128.41	481.07	173.07	622.33	300.18	--	258.29
81.5 - 243.5	134.24	200	298.14	874.01	92.65	409.19	123.10	741.67	149.57	--	283.80
243.5 - 729.5	408.45	74	10.33	98.56	59.07	150.42	86.92	41.09	62.28	--	77.02
729.5 - ∞	850.08	12	.00		62.69	14.82	206.75	61.39	38.75	--	.78
Total		1703	1703.00	1703.00	1703.00	1703.00	1703.00	1703.00	1703.00	--	1703.00

Table 4.5.1. - Continued

		Observed	Expected							
			I	II	III	IV	V	VI	VII	VIII	IX
χ^2 for goodness of fit		--	372018	8195	1461	987	481	37797	94	--	248
parameters estimated	a	--	.0210 ±.0003	.0085 ±.0000	.2983 ±.0028	.2517 ±.0007	.1599 ±.0025	.0156 ±.0001	.0916 ±.0089	--	.1460 ±.0177
	b	--	--	--	--	--	--	--	.8971 ±.0554	--	.0095 ±.0003
	p	--	--	--	--	--	--	--	--	--	.4596 ±.0177
mean distance	$\bar{x}$	53.17	47.7±.6	66.2±.3	16.5±.4	94.7±.6	--	100.9±.4	--	--	52.2
inbreeding coefficient	α	.0185*	.0187	.0152	.0192*	.0168*	.0169*	.0151*	.0186*	--	.0181

x = a point weighted by observed number of couples

I = exponential, II = normal, III = lognormal, IV = square root exponential, V = Skellam, VI = Bessel, VII = beta, VIII = generalized Skellam, and IX = double exponential

* Discrete approximation (see text)

Table 4.5.2.

Distribution of couples with marital distance in parent.

Distance (km.)	Observed	Expected							
		I	II	III	IV	V	VI	VII	VIII	IX
.0 - .5	31	6.49	2.16	464.26	8.52	28.58	.17	24.44	5.79	27.25
.5 - 3.5	175	38.11	13.07	161.90	43.13	258.34	5.19	124.33	167.63	137.77
3.5 - 9.5	137	72.07	26.59	87.78	69.43	161.41	23.06	171.97	254.58	179.68
9.5 - 27.5	189	186.44	82.87	88.20	150.89	137.63	114.19	256.12	230.65	208.43
27.5 - 81.5	244	362.68	260.85	77.23	258.13	108.14	377.76	230.34	151.08	164.38
81.5 - 243.5	175	326.55	545.35	60.42	296.39	83.07	456.68	134.02	92.41	213.19
243.5 - 729.5	65	49.54	111.11	42.74	176.92	63.37	33.51	60.62	55.77	106.66
729.5 - ∞	26	.11		59.48	38.59	201.45	31.45	40.16	84.08	4.63
Total	1042	1042.00	1042.00	1042.00	1042.00	1042.00	1042.00	1042.00	1042.00	1042.00

Table 4.5.2. - Continued

	Observed	Expected							
		I	II	III	IV	V	VI	VII	VIII	IX
χ^2 for goodness of fit	--	6626	3242	1156	663	475	12030	66	344	188
parameters estimated a	--	.0125 ±.0002	.0074 ±.0001	.2663 ±.0029	.1890 ±.0011	.1246 ±.0027	.0154 ±.0001	.0549 ±.0063	.0974 ±.0122	.1068 ±.0088
b	--	--	--	--	--	--	--	.8765 ±.0627	.2319 ±.0134	.0066 ±.0003
p	--	--	--	--	--	--	--	--	--	.5325 ±.0245
mean distance $\bar{x}$	86.90	80.0±1.4	76.8±.5	34.0	167.9±2.1	--	102.0±.6	--	--	85.2
inbreeding coefficient α	.0175*	.0163	.0142	.0187*	.0153*	.0153*	.0151*	.0175*	.0173*	.0167

I = exponential, II = normal, III = lognormal, IV = square root exponential, V = Skellam, VI = Bessel, VII = beta, VIII = generalized Skellam, and IX = double exponential

* Discrete approximation (see text)

Table 4.5.3.

Distribution of couples with marital distance weighted by square root of density in grandparent.

Distance x $\sqrt{\text{density}}$	x	Observed	Expected								
			I	II	III	IV	V	VI	VII	VIII	IX
.0- .5	0	74	5.26	2.69	764.48	7.70	33.69	.26	22.56	2.51	20.68
.5- 3.5	2.22	85	31.20	16.21	229.56	40.76	317.81	7.98	121.56	100.48	115.44
3.5- 9.5	6.51	204	60.68	32.85	125.28	69.45	213.91	35.62	189.56	295.26	192.07
9.5- 27.5	17.09	313	169.00	101.69	129.51	164.08	195.49	177.72	341.73	418.87	362.20
27.5- 81.5	50.07	353	406.83	322.19	119.21	324.30	165.35	598.40	396.08	318.82	345.50
81.5-243.5	153.56	300	646.41	827.17	100.17	474.95	136.90	753.55	296.56	206.26	250.22
243.5-729.5	422.89	235	352.69	387.19	77.69	422.65	112.62	76.00	167.25	129.93	296.81
729.5-∞	1939.35	120	17.94		144.10	186.11	514.24	40.47	154.70	217.87	107.06
Total		1690	1690.00	1690.00	1690.00	1690.00	1690.00	1690.00	1690.00	1690.00	1690.00

Table 4.5.3. - Continued

		Observed	Expected							
			I	II	III	IV	V	VI	VII	VIII	IX
χ^2 for goodness of fit		--	2336	3856	2202	1184	1123	23443	169	2262	179
parameters estimated	a	--	.0062 ±.0001	.0056 ±.0001	.2341 ±.0016	.1395 ±.0011	.0902 ±.0019	.0149 ±.0001	.0381 ±.0034	.0280 ±.0026	.0438 ±.0025
	b	--	--	--	--	--	--	--	.7114 ±.0371	.2132 ±.0084	.0027 ±.0001
	p	--	--	--	--	--	--	--	--	--	.4647 ±.0183
mean distance	$\bar{x}$	238.3	160.5 ± 2.6	100.2 ± .9	95.8 ± 3.0	308.3 ± 4.8	--	105.1 ± .5	--	--	182.
inbreeding coefficient	α	.0167*	.0169	.0175	.171*	.0150*	.0137*	.0174*	.0167*	.0164*	.0170

x = a point weighted by observed number of couples

I = exponential, II = normal, III = lognormal, IV = square root exponential, V = Skellam, VI = Bessel, VII = beta, VIII = generalized Skellam, and IX = double exponential

* Discrete approximation

Table 4.5.4.

Distribution of couples with marital distance weighted by square root of density in parent.

Distance x $\sqrt{\text{density}}$	Observed	Expected								
		I	II	III	IV	V	VI	VII	VIII	IX
.0 - .5	31	2.11	1.36	473.08	3.37	17.09	.16	8.04	.71	9.42
.5 - 3.5	37	12.58	8.19	123.68	18.17	164.35	4.87	45.17	30.97	52.76
3.5 - 9.5	94	24.69	16.56	67.69	31.76	114.38	21.75	77.06	119.89	88.55
9.5 - 27.5	146	70.55	51.02	71.33	77.91	107.85	108.53	162.54	234.56	171.41
27.5 - 81.5	189	182.99	161.48	68.02	164.37	94.30	365.57	234.83	210.14	182.37
81.5 - 243.5	209	358.65	453.75	60.16	268.91	80.75	460.77	215.97	146.41	175.60
243.5 - 729.5	197	329.21	340.66	49.87	287.58	68.73	48.01	140.15	97.81	240.21
729.5 - ∞	130	52.22		119.16	180.94	385.56	23.35	149.25	192.50	112.66
Total	1033	1033.00	1033.00	1033.00	1033.00	1033.00	1033.00	1033.00	1033.00	1033.00

Table 4.5.4. - Continued

		Observed	Expected							
			I	II	III	IV	V	VI	VII	VIII	IX
χ^2 for goodness of fit		--	950	1425	1581	488	835	7631	107	1477	76
parameters estimated	a	--	.0041 ±.0001	.0047 ±.0001	.2073 ±.0017	.1174 ±.0016	.0748 ±.0023	.0149 ±.0001	.0236 ±.0028	.0148 ±.0018	.0422 ±.0038
	b	--	--	--	--	--	--	--	.6667 ±.0458	.1873 ±.0094	.0023 ±.0001
	p	--	--	--	--	--	--	--	--	--	.6045 ±.0232
mean distance	$\bar{x}$	356.0	244.5 ± 6.0	121.1 ± 1.8	571.4 ± 26.1	435.5 ± 11.5	--	102.2 ± .5	--	--	267
inbreeding coefficient	α	.0154*	.0157	.0170	.0172*	.0138*	.0125*	.0174*	.0154*	.0152*	.0158

I = exponential, II = normal, III = lognormal, IV = square root exponential., V = Skellam, VI = Bessel, VII = beta, VIII = generalized Skellam, and IX = double exponential

* Discrete approximation (see text).

Figure 2.5.1.

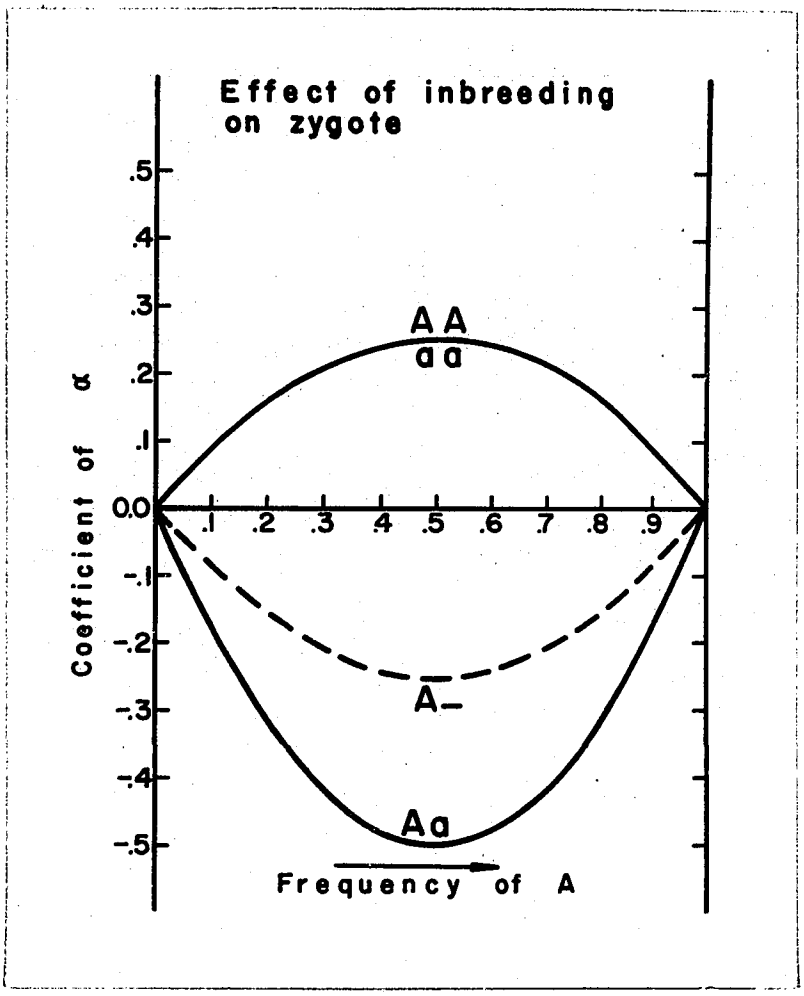

Effect of inbreeding on zygote
Coefficient of α
.5
.4
.3
.2
.1
0.0
-.1
-.2
-.3
-.4
-.5
.1 .2 .3 .4 .5 .6 .7 .8 .9
AA
aa
A–
Aa
Frequency of A

Figure 2.5.2.

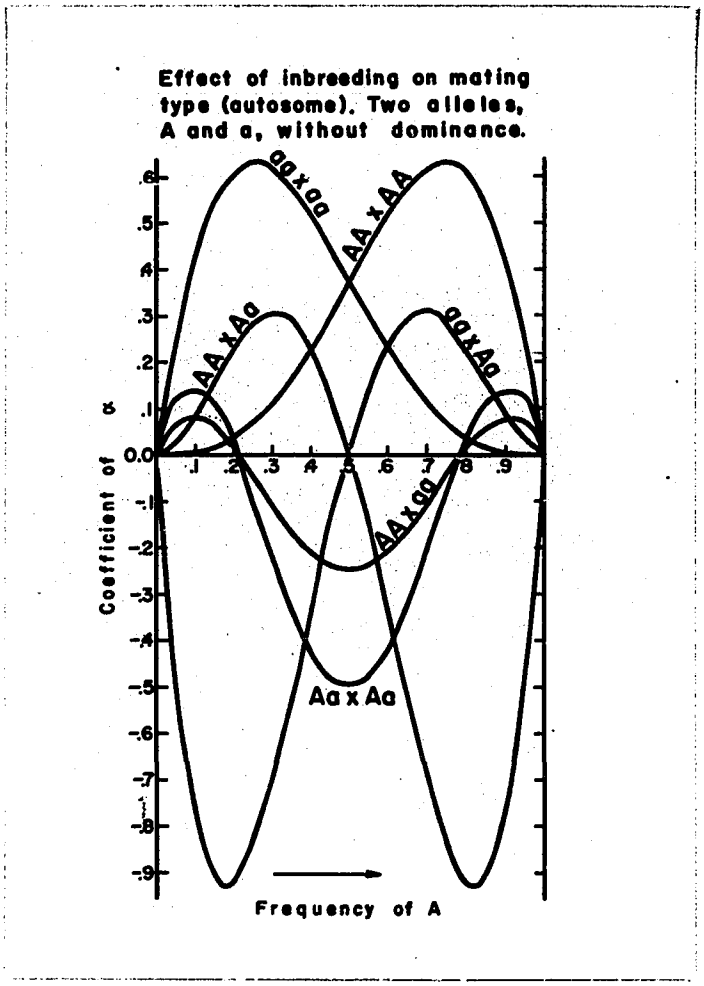

Effect of inbreeding on mating type (autosome). Two alleles, A and a, without dominance.
aa x aa
AA x AA
AA x Aa
aa x Aa
AA x aa
Aa x Aa
Coefficient of α
Frequency of A

Figure 2.5.3.

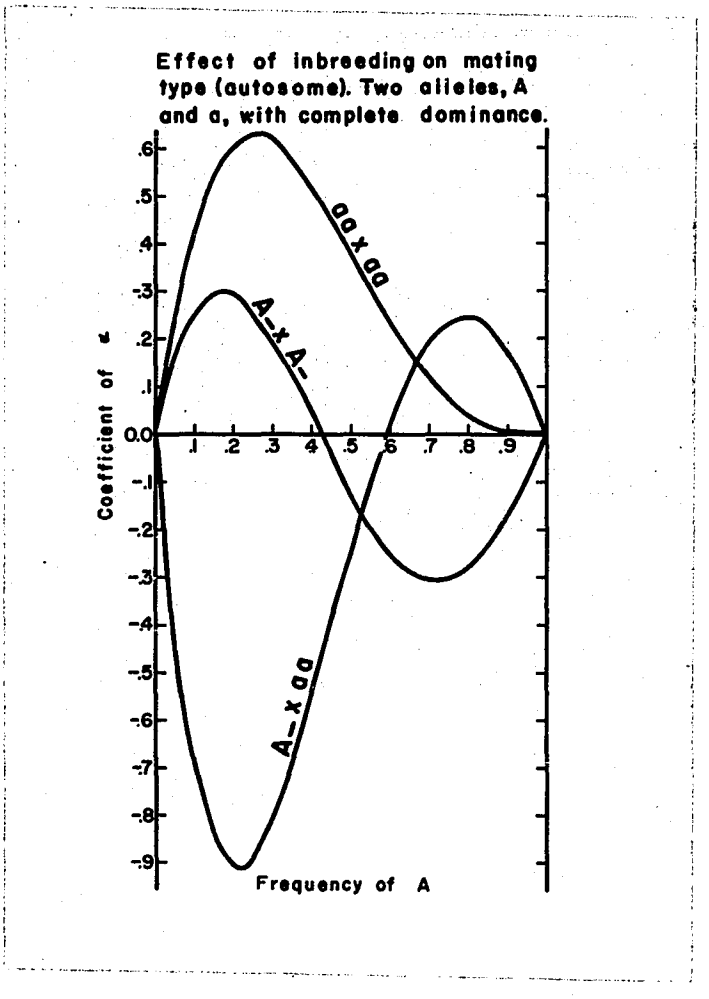

Effect of inbreeding on mating type (autosome). Two alleles, A and a, with complete dominance.
Coefficient of F
.6
.5
.4
.3
.2
.1
0.0
-.1
-.2
-.3
-.4
-.5
-.6
-.7
-.8
-.9
.1 .2 .3 .4 .5 .6 .7 .8 .9
aa x aa
A- x A-
A- x aa
Frequency of A

Figure 2.5.4.

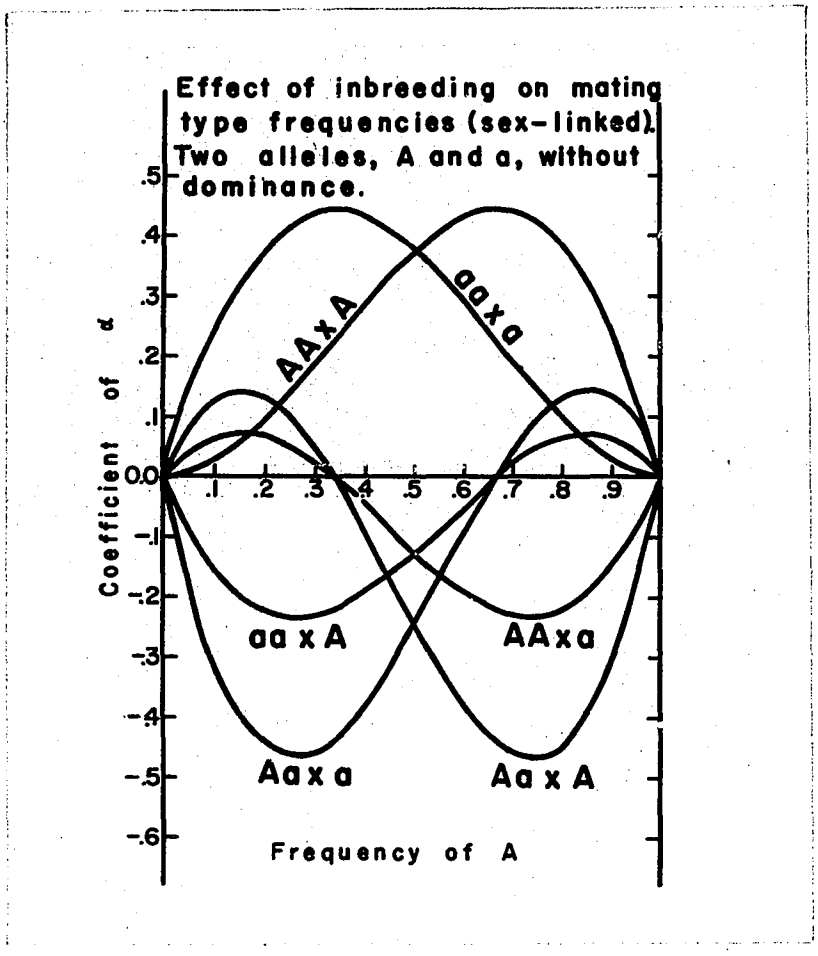

Effect of inbreeding on mating type frequencies (sex-linked). Two alleles, A and a, without dominance.
Coefficient of α
.5
.4
.3
.2
.1
0.0
-.1
-.2
-.3
-.4
-.5
-.6
.1
.2
.3
.4
.5
.6
.7
.8
.9
AAxA
aaxa
aa x A
AAxa
Aaxa
Aa x A
Frequency of A

Figure 2.5.5.

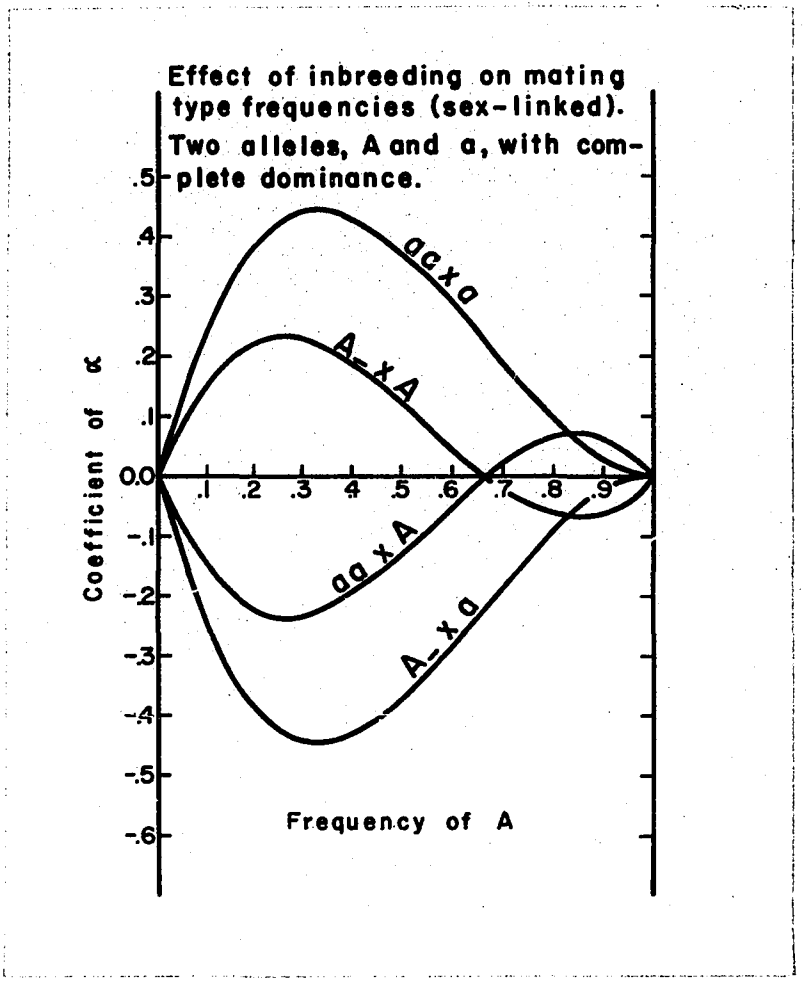

Effect of inbreeding on mating type frequencies (sex-linked).
Two alleles, A and a, with complete dominance.
Coefficient of α
.5
.4
.3
.2
.1
0.0
-.1
-.2
-.3
-.4
-.5
-.6
.1
.2
.3
.4
.5
.6
.7
.8
.9
aa x a
A- x A
aa x A
A- x a
Frequency of A

Figure 2.6.1.

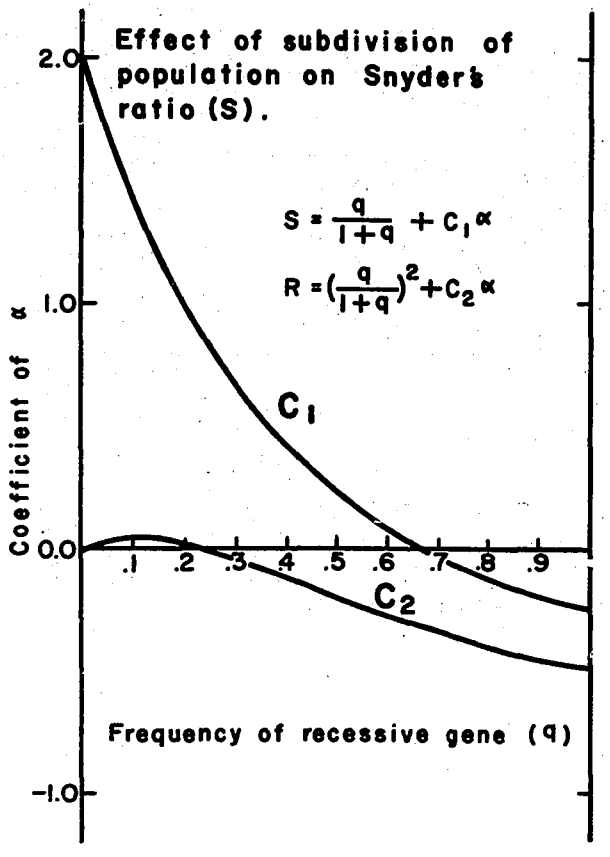

Effect of subdivision of population on Snyder's ratio (S).
$S = \frac{q}{1+q} + C_1\alpha$
$R = (\frac{q}{1+q})^2 + C_2\alpha$
Coefficient of α
2.0
1.0
0.0
-1.0
.1
.2
.3
.4
.5
.6
.7
.8
.9
C1
C2
Frequency of recessive gene (q)

Figure 2.6.2.

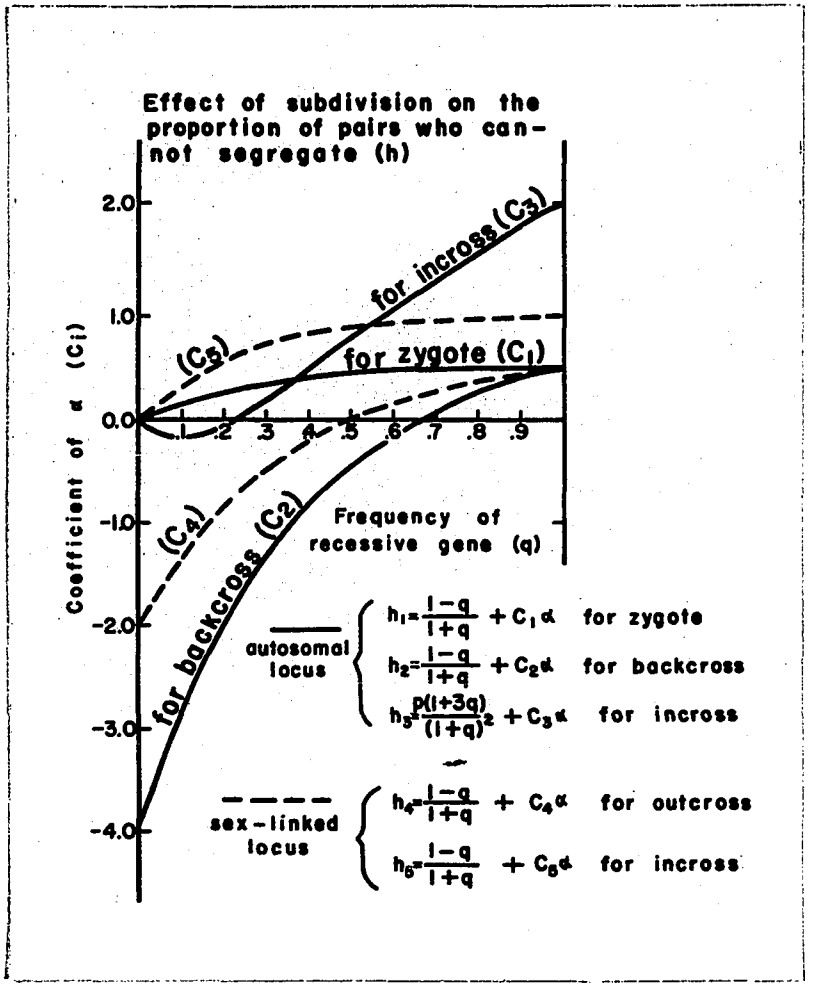
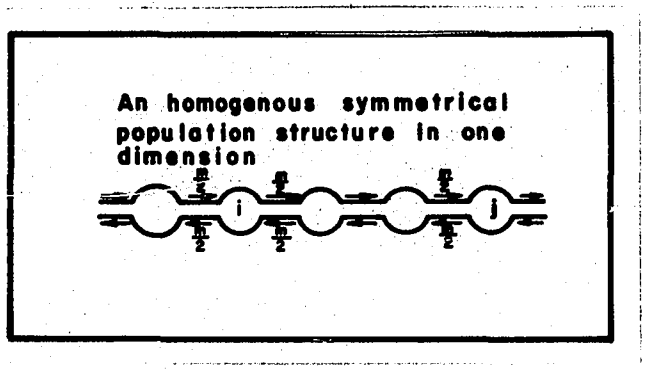
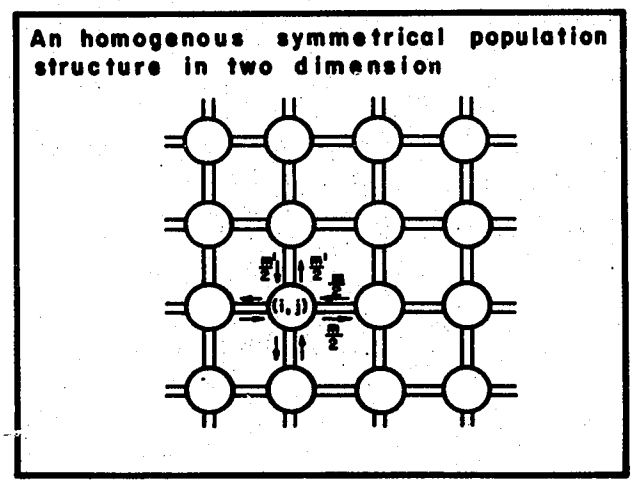

Effect of subdivision on the proportion of pairs who cannot segregate (h)
Coefficient of α (Ci)
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
.1 .2 .3 .4 .5 .6 .7 .8 .9
Frequency of recessive gene (q)
for incross (C3)
(C5)
for zygote (C1)
(C4)
for backcross (C2)
autosomal locus
h1=(1−q)/(1+q) + C1α for zygote
h2=(1−q)/(1+q) + C2α for backcross
h3=p(1+3q)/(1+q)² + C3α for incross
sex-linked locus
h4=(1−q)/(1+q) + C4α for outcross
h5=(1−q)/(1+q) + C5α for incross

Figure 2.7.1.



The arrows denote flow of migrants and $\frac{m}{2}$ is the relative proportion of migration to an isolate by each generation. The indices are arbitrarily assigned for isolates. All distances between two isolates are the same.

Figure 2.7.2.



The arrows denote flow of migrants and $\frac{m}{2}$ and $\frac{m'}{2}$ are the relative proportion of migration to an isolate by each generation. The indices are arbitrarily assigned for isolates. All distances between two isolates are the same.

Figure 4.3.1.

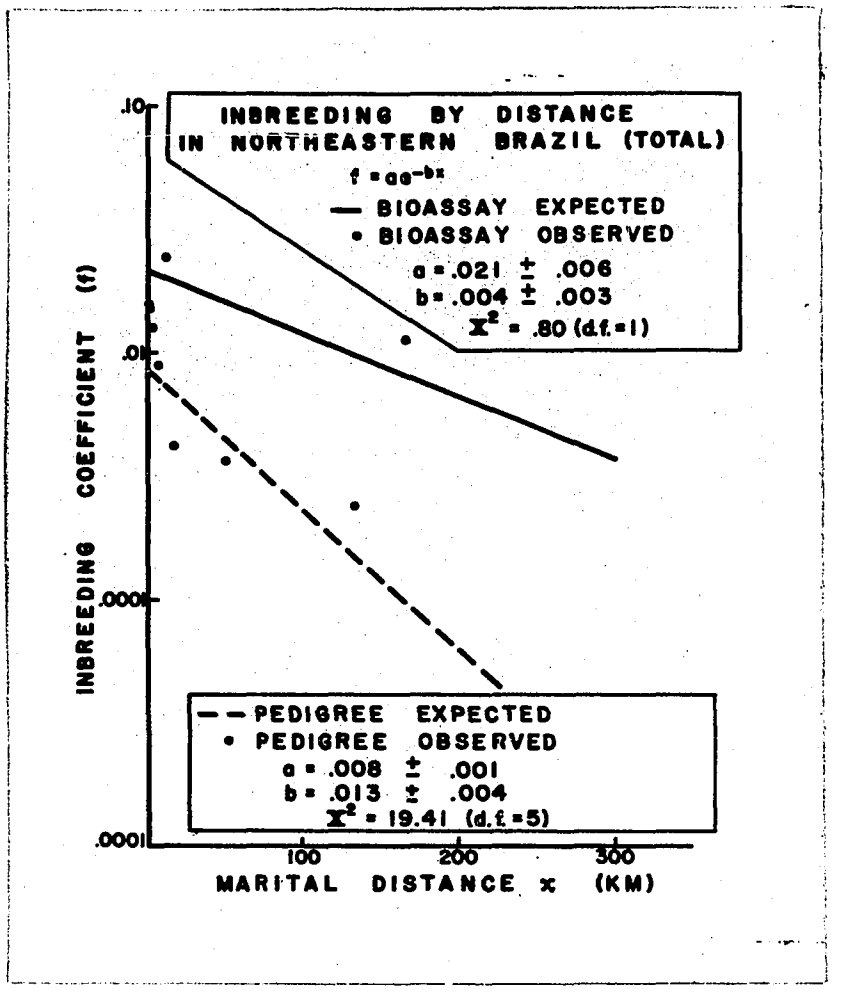

INBREEDING BY DISTANCE
IN NORTHEASTERN BRAZIL (TOTAL)
f = ae^-bx
— BIOASSAY EXPECTED
• BIOASSAY OBSERVED
a = .021 ± .006
b = .004 ± .003
X² = .80 (d.f.=1)
— — PEDIGREE EXPECTED
• PEDIGREE OBSERVED
a = .008 ± .001
b = .013 ± .004
X² = 19.41 (d.f.=5)
INBREEDING COEFFICIENT (f)
.10
.01
.001
.0001
100
200
300
MARITAL DISTANCE x (KM)

Figure 4.3.2.

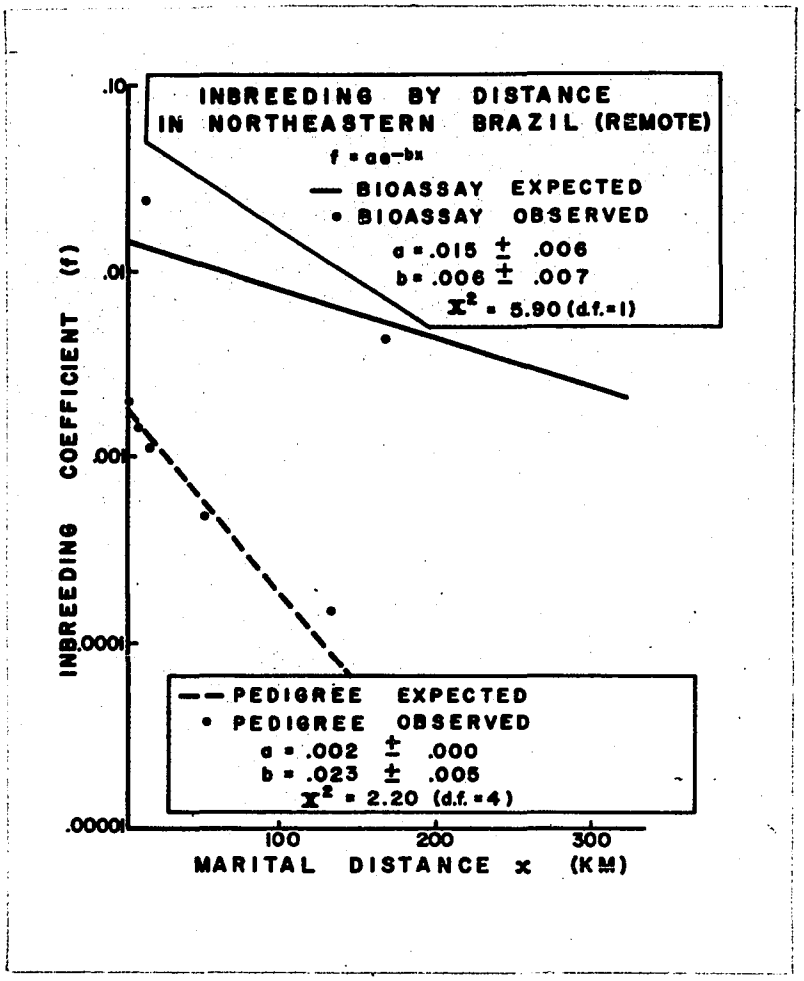

INBREEDING BY DISTANCE
IN NORTHEASTERN BRAZIL (REMOTE)
f = ae^-bx
— BIOASSAY EXPECTED
• BIOASSAY OBSERVED
a = .015 ± .006
b = .006 ± .007
χ² = 5.90 (d.f.=1)
— — PEDIGREE EXPECTED
• PEDIGREE OBSERVED
a = .002 ± .000
b = .023 ± .005
χ² = 2.20 (d.f.=4)
INBREEDING COEFFICIENT (f)
.10
.01
.001
.0001
.00001
100
200
300
MARITAL DISTANCE x (KM)

Figure 4.3.3.

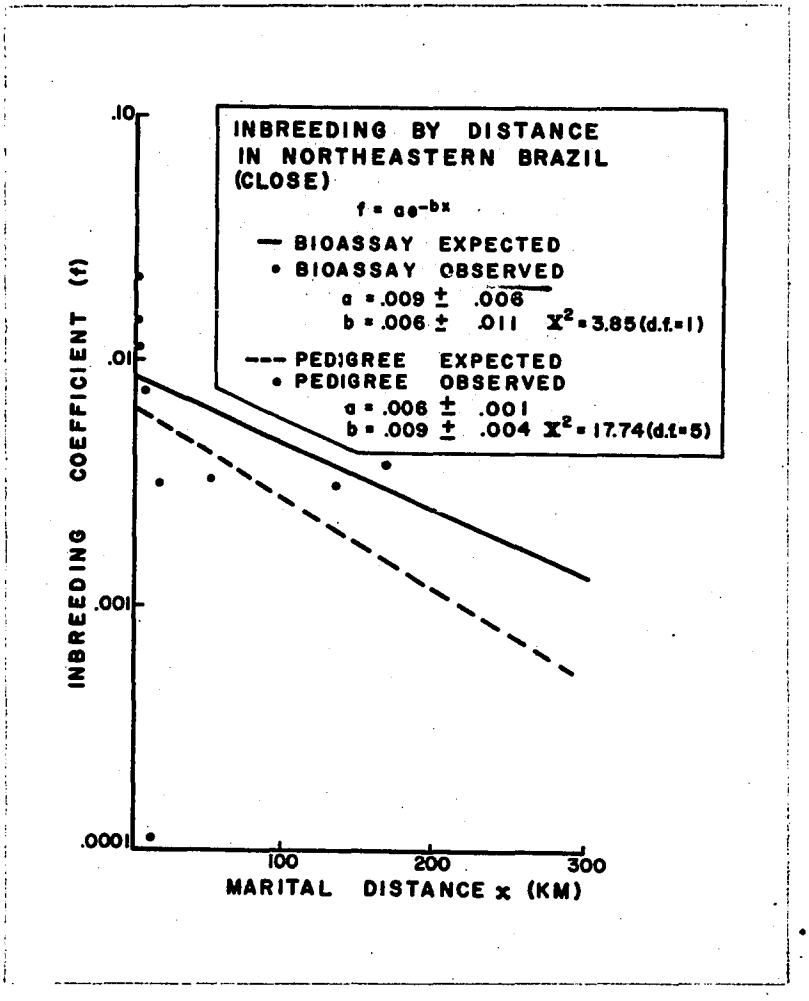
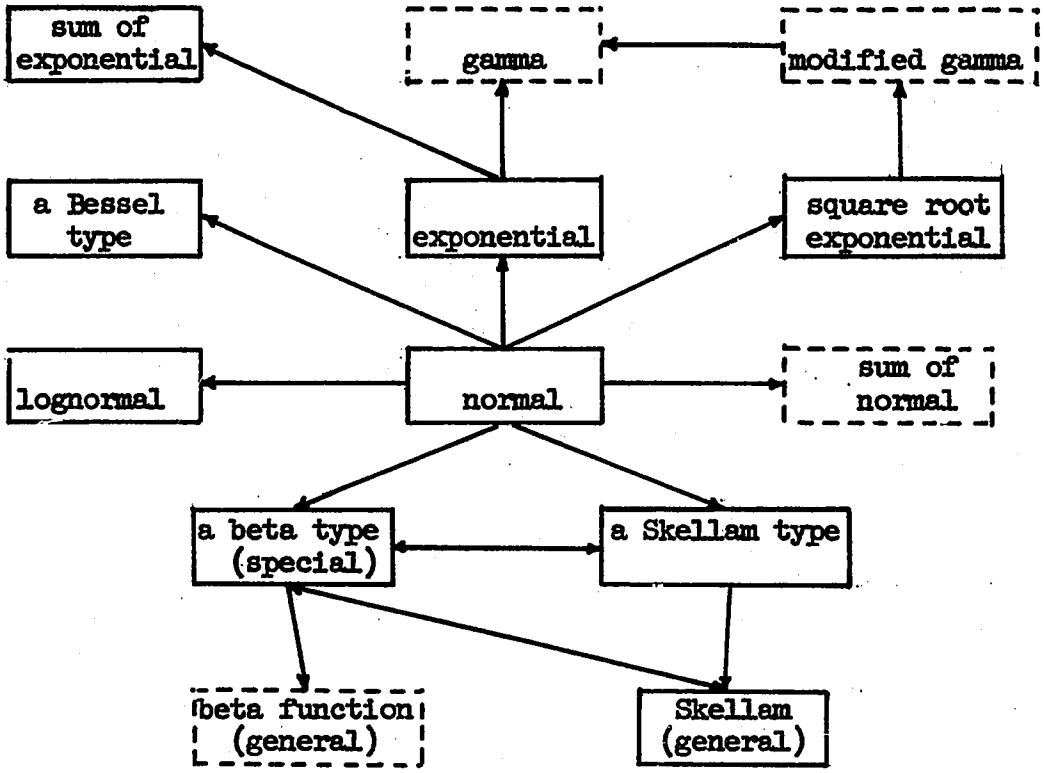

INBREEDING BY DISTANCE
IN NORTHEASTERN BRAZIL
(CLOSE)
f = ae^-bx
— BIOASSAY EXPECTED
• BIOASSAY OBSERVED
a = .009 ± .006
b = .006 ± .011 X² = 3.85(d.f.=1)
--- PEDIGREE EXPECTED
• PEDIGREE OBSERVED
a = .006 ± .001
b = .009 ± .004 X² = 17.74(d.f.=5)
INBREEDING COEFFICIENT (f)
.10
.01
.001
.0001
100
200
300
MARITAL DISTANCE x (KM)

Figure 4.5.1.

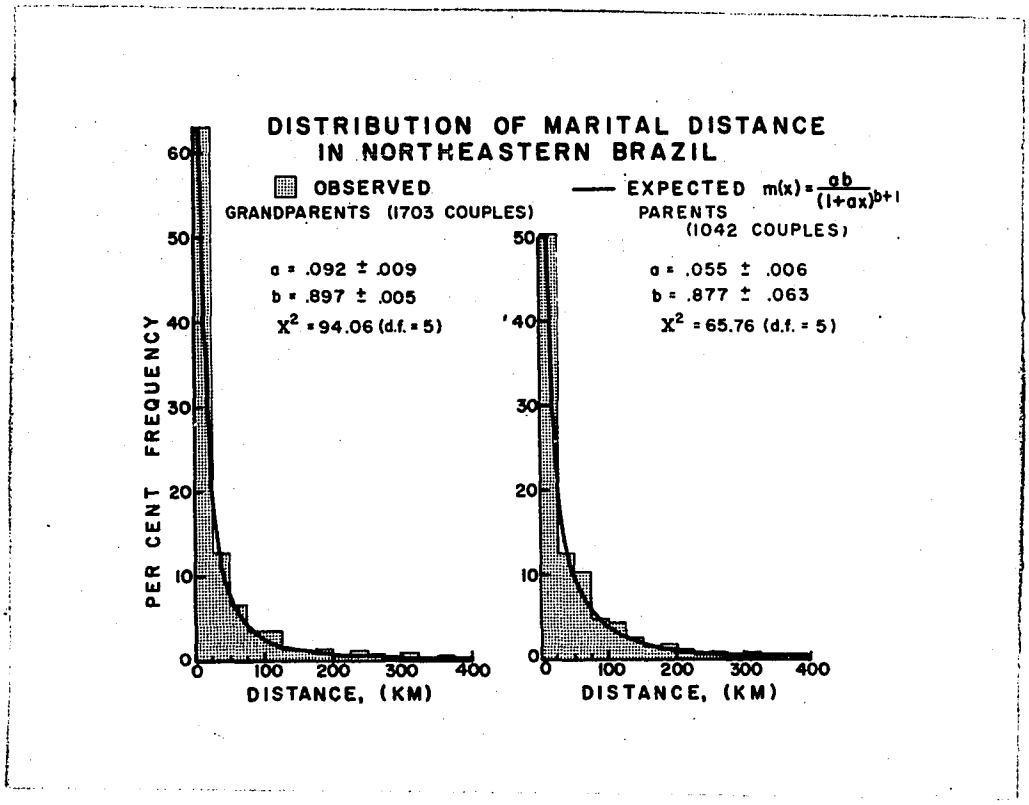
Migration functions and their relation.



Studied in text.

Not studied in text.

Figure 4.5.2.


DISTRIBUTION OF MARITAL DISTANCE
IN NORTHEASTERN BRAZIL
OBSERVED
GRANDPARENTS (1703 COUPLES)
a = .092 ± .009
b = .897 ± .005
X² = 94.06 (d.f. = 5)
EXPECTED m(x) = ab/(1+ax)^b+1
PARENTS
(1042 COUPLES)
a = .055 ± .006
b = .877 ± .063
X² = 65.76 (d.f. = 5)
PER CENT FREQUENCY
DISTANCE, (KM)
DISTANCE, (KM)