A Study Through a Markov Chain of a Population Undergoing Certain Mating Systems in the Presence of Linkage*

by

Prem S. Puri

Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series No. 130
December 1967

*To appear in a volume of contributed papers being published by the Indian Society of Agricultural Statistics for presentating to Dr. V. G. Panse on his birthday in recognition to his services to the Society and his contributions to the field of statistics.

A Study Through a Markov Chain of a Population Undergoing Certain Mating Systems in the Presence of Linkage

by

Prem S. Puri

Purdue University, Lafayette, Indiana

1. Introduction. How the distribution of various genotypes and its properties vary from one generation to another in a population undergoing a given system of mating has been and perhaps still is the subject of extensive study among several research workers. A great deal concerning this can be found in literature (see Kempthorne [4]). There are left however a great number of cases which are either still unsolved or are such that their properties have not been fully explored. One such case forms the subject of this paper. One of the key assumptions made in most of the cases studied thus far is that of independent segregation of the factors involved. The only case known in literature where the two factors are assumed to be linked, is the one where the population is subjected to random mating. In the present paper an attempt is made to eliminate the assumption of independent segregation among factors.

To begin with, we shall study the case of two linked factors for a diploid population undergoing selfing startubg with an arbitrary initial genotypic distribution. later this is generalised to a case of great interest to plant breeders namely that of mixed self-fertilization and random mating. Here it is assumed that the initial population is in equilibrium with respect to random mating system. The case with an arbitrary initial genotypic distribution is somewhat involved and will be reported elsewhere in order not to overload a single paper (see Puri [7]). The populations undergoing such a mixed mating system have been studied by several research workers in the past and more recently by Ghai ([1], [2]), all under the assumption of independent segregation of the factors involved.

The model considered here is still restrictive in the sense that it assumes an absence of selection and mutation and assumes also an equal mortality and fertility over all genotypes. The model incorporating selection presents interesting problems and is left for a later study. Reader is referred to an interesting recent paper on selection by Li [5].

2.0 A Markov Chain related to model where population is undergoing continued selfing. Consider the case of a diploid population with two linked factors in mind each with two alleles and let A-a and B-b be the corresponding gene-pairs with p $(q = 1-p)$ denoting the usual recombination value or the proportion of cross-over gametes. Following the standard convention we restrict p to take values between 0 and $\frac{1}{2}$. The case with $p = 0$ is that of complete linkage and p is $\frac{1}{2}$ when the factors segregate independently. The four possible gametes are AB, Ab, aB and ab, and the ten possible genotypes are

$$(1) \quad \begin{cases} 1.\ AB/AB \quad 2.\ AB/Ab \quad 3.\ Ab/Ab \quad 4.\ AB/aB \quad 5.\ AB/ab \\ 6.\ Ab/aB \quad 7.\ Ab/ab \quad 8.\ aB/aB \quad 9.\ aB/ab \quad 10.\ ab/ab \end{cases} .$$

Let their initial genotypic frequency vector be given by

$$(2) \qquad \underset{\sim}{f}^{(0)} = (f_{22}^{(0)}, f_{21}^{(0)}, f_{20}^{(0)}, f_{12}^{(0)}, f_{11c}^{(0)}, f_{11r}^{(0)}, f_{10}^{(0)}, f_{02}^{(0)}, f_{12}^{(0)}, f_{01}^{(0)}, f_{00}^{(0)})'$$

with vector $\underset{\sim}{f}^{(n)}$ in an analogous manner corresponding to the nth generation, where the components of these vectors add up to one and where the numbers in the subscripts stand for the numbers of genes A or B present in the corresponding genotypes. Also the letters c and r stand for the coupling and the recessive phases of the double hetrozygotes numbered 5 and 6 respectively in (1). The prime in (2) denotes the transpose of a matrix.

We consider first the case where the above population is undergoing self-fertilization indefinitely. The immediate problem is to find the genotypic distribution vector $\underset{\sim}{f}^{(n)}$ of the population at the nth stage of continued selfing. To this end, we visualise a discrete time Markov chain $\{X_n; n = 0,1,2,\ldots\}$ with the state space S consisting of the ten possible genotypes 1,2,...,10 as listed in (2). The chain is considered to be in state i at nth step if a randomly chosen genotype out of the population (assumed to be infinite in size) at the nth generation turns out to be of type i with $i = 1,2,\ldots,10$. The initial probability distribution is governed by the vector $\underset{\sim}{f}^{(0)}$. The Markov property of the chain is obvious, since the genotypic distribution at (n+1)st step (generation) depends only on the distribution at the nth step. The only thing left in order to specify the Markov chain completely is the one-step transition probability matrix (P_{ij}) which is stationary in the present case. This is obtained by considering for each genotype individually the genotypic composition it produces after it is once selfed. For instance, given $X_n = 5$ (i.e. AB/ab), the vector of probabilities $P_{5j} = P(X_{n+1} = j | X_n = 5)$; $j=1,2,\ldots,10$, is given by

$$(q^2/2,\ pq/2,\ p^2/4,\ pq/2,\ q^2/2,\ p^2/2,\ pq/2,\ p^2/4,\ pq/2,\ q^2/4),$$

where $q = 1-p$, and so on. The matrix $\underset{\sim}{P} = (P_{ij})$ obtained in this manner is given by

$$(3) \quad \underset{\sim}{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/4 & 0 & 0 \\ q^2/4 & pq/2 & p^2/2 & pq/2 & q^2/2 & p^2/2 & pq/2 & p^2/4 & pq/2 & q^2/4 \\ p^2/4 & pq/2 & q^2/2 & pq/2 & p^2/2 & q^2/2 & pq/2 & q^2/4 & pq/2 & p^2/4 \\ 0 & 0 & 1/4 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Thus following the usual Markov chain argument the genotypic distribution in the nth generation is given by

$$(4) \qquad \underset{\sim}{f}^{(n)} = (\underset{\sim}{P}')^n \cdot \underset{\sim}{f}^{(0)} .$$

The problem now is to find the nth power of $\underset{\sim}{P}$ needed in (4). For this, on solving the characteristic equation $|\underset{\sim}{P} - \lambda \underset{\sim}{I}| = 0$ of matrix $\underset{\sim}{P}$ for λ, we observe that the ten eigen values of $\underset{\sim}{P}$ are given by

$$(5) \qquad [\, 1, 1, 1, 1, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \frac{1-2pq}{2}, \frac{1-2p}{2} \,] .$$

Fortunately for the present case the matrix $\underset{\sim}{P}$ lends itself to a spectral representation. Following the standard approach (see Karlin [3]) for finding the spectral representation of a matrix by obtaining the left and right eigen vectors for various eigen values, we find that

(6) $$\underset{\sim}{P} = \underset{\sim}{M}\,\underset{\sim}{D}\,\underset{\sim}{L}$$

where

(7) $$D = \mathrm{dg}\,(1,\ 1,\ 1,\ 1,\ \tfrac{1}{2},\ \tfrac{1}{2},\ \tfrac{1}{2},\ \tfrac{1}{2},\ a_2,\ a_3) \quad ,$$

(8) $$\underset{\sim}{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ a_1/2 & pa_1 & pa_1 & a_1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/\sqrt{2} & 1/\sqrt{2} \\ pa_1 & a_1/2 & a_1/2 & pa_1 & 1/2 & 1/2 & 1/2 & 1/2 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

(9) $$\underset{\sim}{L} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1/2 & 1 & -1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1 & 0 & 0 & 0 & -1/2 & 0 & 0 \\ 0 & 0 & -1/2 & 0 & 0 & 0 & 1 & 0 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1/2 & 1 & -1/2 \\ 1/2\sqrt{2} & -1/\sqrt{2} & 1/2\sqrt{2} & -1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} & -1/\sqrt{2} & 1/2\sqrt{2} & -1/\sqrt{2} & 1/2\sqrt{2} \\ -a_1a_3/\sqrt{2} & 0 & a_1a_3/\sqrt{2} & 0 & 1/\sqrt{2} & -1/\sqrt{2} & 0 & a_1a_3/\sqrt{2} & 0 & -a_1a_3/\sqrt{2} \end{pmatrix}$$

with $\underset{\sim}{M}\,\underset{\sim}{L} = \underset{\sim}{I}$ and

$$(10) \qquad a_1 = \frac{1}{(1+2p)} \;;\quad a_2 = \frac{1-2pq}{2} \;;\quad a_3 = \frac{1-2p}{2} \quad .$$

Here $\underset{\sim}{I}$ is a 10 x 10 identity matrix. Also we have used the convention of writing an $n \times n$ diagonal matrix $\underset{\sim}{A} = (a_{ij})$ with $a_{ij} = \delta_{ij}\, a_{ii}$ by $\underset{\sim}{A} = dg\,(a_{11}, a_{22}, \ldots, a_{nn})$. Using (6) and the fact that $\underset{\sim}{M}\,\underset{\sim}{L} = \underset{\sim}{I}$, we finally have

$$(11) \qquad \underset{\sim}{P}^n = \underset{\sim}{M}\,\underset{\sim}{D}^n\,\underset{\sim}{L} \ ,$$

yielding for $n = 0, 1, 2, \ldots,$

$$(12) \qquad \underset{\sim}{f}^{(n)} = \underset{\sim}{L}'\,\underset{\sim}{D}^n\,\underset{\sim}{M}'\,\underset{\sim}{f}^{(0)} \ ,$$

where

$$(13) \qquad \underset{\sim}{D}^n = dg(1,\ 1,\ 1,\ 1,\ \frac{1}{2^n},\ \frac{1}{2^n},\ \frac{1}{2^n},\ \frac{1}{2^n},\ a_2^n,\ a_3^n) \quad .$$

Writing (12) descriptively for later use, we have the expressions for $f_{ij}^{(n)}$ for various i,j values, given in table 1.

It is clear from (3) that the states 1, 3, 8 and 10 are absorption state while the remaining six states are transient. Thus it follows from the theory of finite Markov chains that with probability one the ultimate absorption takes place to one of the four absorption states. The transient states here correspond to those of hetrozygotes while the absorption states to those of homozygotes.

Table 1. Elements of $\underset{\sim}{f}^{(n)}$

$$f_{22}^{(n)} = f_{22}^{(0)}+\tfrac{1}{2}(f_{12}^{(0)}+f_{21}^{(0)})[1-(\tfrac{1}{2})^n]+\tfrac{1}{2}f_{11c}^{(0)}[a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n-a_1a_3^{n+1}] + \tfrac{1}{2}f_{11r}^{(0)}[1-a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n+a_1a_3^{n+1}]$$

$$f_{21}^{(n)} = (\tfrac{1}{2})^n f_{21}^{(0)} + \tfrac{1}{2}(f_{11c}^{(0)}+f_{11r}^{(0)})[(\tfrac{1}{2})^n-a_2^n]$$

$$f_{20}^{(n)} = f_{20}^{(0)}+\tfrac{1}{2}(f_{21}^{(0)}+f_{10}^{(0)})[1-(\tfrac{1}{2})^n] + \tfrac{1}{2}f_{11c}^{(0)}[1-a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n+a_1a_3^{n+1}] + \tfrac{1}{2}f_{11r}^{(0)}[a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n-a_1a_3^{n+1}]$$

$$f_{12}^{(n)} = (\tfrac{1}{2})^n f_{12}^{(0)} + \tfrac{1}{2}(f_{11c}^{(0)}+f_{11r}^{(0)})[(\tfrac{1}{2})^n - a_2^n]$$

$$f_{11c}^{(n)} = \tfrac{1}{2}f_{11c}^{(0)}[a_2^n+a_3^n] + \tfrac{1}{2}f_{11r}^{(0)}[a_2^n-a_3^n]$$

$$f_{11r}^{(n)} = \tfrac{1}{2}f_{11c}^{(0)}[a_2^n-a_3^n] + \tfrac{1}{2}f_{11r}^{(0)}[a_2^n+a_3^n]$$

$$f_{10}^{(n)} = (\tfrac{1}{2})^n f_{10}^{(0)} + \tfrac{1}{2}(f_{11c}^{(0)} + f_{11r}^{(0)})[(\tfrac{1}{2})^n-a_2^n]$$

$$f_{02}^{(n)} = f_{02}^{(0)}+\tfrac{1}{2}(f_{01}^{(0)}+f_{12}^{(0)})[1-(\tfrac{1}{2})^n]+\tfrac{1}{2}f_{11c}^{(0)}[1-a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n+a_1a_3^{n+1}] + \tfrac{1}{2}f_{11r}^{(0)}[a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n-a_1a_3^{n+1}]$$

$$f_{01}^{(n)} = (\tfrac{1}{2})^n f_{01}^{(0)} + \tfrac{1}{2}(f_{11c}^{(0)}+f_{11r}^{(0)})[(\tfrac{1}{2})^n-a_2^n]$$

$$f_{00}^{(n)} = f_{00}^{(0)}+\tfrac{1}{2}(f_{10}^{(0)}+f_{01}^{(0)})[1-(\tfrac{1}{2})^n] + \tfrac{1}{2}f_{11c}^{(0)}[a_1-(\tfrac{1}{2})^n+\tfrac{1}{2}a_2^n-a_1a_3^{n+1}] + \tfrac{1}{2}f_{11r}^{(0)}[1-a_1-(\tfrac{1}{2})^n + \tfrac{1}{2}a_2^n + a_1a_3^{n+1}] \quad .$$

Thus as expected, the hetrozygosity disappears eventually with probability one, with the propositions of various homozygotes in the ultimate population given by letting $n \to \infty$ in $f_{ij}^{(n)}$, yielding

$$(14) \quad \begin{cases} f_{22}^{(\infty)} = f_{22}^{(0)} + \frac{1}{2}(f_{21}^{(0)} + f_{12}^{(0)}) + \frac{1}{2}[a_1 f_{11c}^{(0)} + (1-a_1)\, f_{11r}^{(0)}] \\ f_{20}^{(\infty)} = f_{20}^{(0)} + \frac{1}{2}(f_{21}^{(0)} + f_{10}^{(0)}) + \frac{1}{2}[(1-a_1)\, f_{11c}^{(0)} + a_1\, f_{11r}^{(0)}] \\ f_{02}^{(\infty)} = f_{02}^{(0)} + \frac{1}{2}(f_{12}^{(0)} + f_{01}^{(0)}) + \frac{1}{2}[(1-a_1)\, f_{11c}^{(0)} + a_1\, f_{11r}^{(0)}] \\ f_{00}^{(\infty)} = f_{00}^{(0)} + \frac{1}{2}(f_{10}^{(0)} + f_{01}^{(0)}) + \frac{1}{2}[a_1 f_{11c}^{(0)} + (1-a_1)\, f_{11r}^{(0)}] \end{cases} ,$$

where $f_{ij}^{(\infty)} = 0$ for transient states. Furthermore the population approaches to homozygosity (fixation) as $n \to \infty$ at least as fast as $(\frac{1}{2})^n$ to zero.

2.1 **Distribution of Time to Homozygosity.** Having found that after continuous selfing the population approaches with probability one to homozygosity, it is natural to ask as to how much time does it take before it attains homozygosity with respect to one or the other factor or both. To this end, let T_A and T_B denote the times when the population reaches for the first time to homozygous state with respect to gene pairs A-a and B-b respectively, so that

$$(15) \quad \Pr(T_A \leq m,\ T_B \leq n) = \Pr(X_m \in \{1, 2, 3, 8, 9, 10\} \text{ and } X_n \in \{1, 3, 4, 7, 8, 10\}) .$$

Let $m \leq n$. Using the fact that the states 1, 3, 8 and 10 are the absorption states and are homozygous with respect to both factors, we have

$$(16)\qquad \Pr(T_A \leq m\ ,\ T_B \leq n) = f_{22}^{(m)} + f_{20}^{(m)} + f_{02}^{(m)} + f_{00}^{(m)} + f_{21}^{(m)} \Pr(X_n \ \epsilon\{1,3\}\,|X_m = 2) + f_{01}^{(m)} \Pr(X_n \ \epsilon\{8,10\}\,|X_m = 9)\ .$$

From the stationarity property of the Markov Chain it follows that

$$(17)\qquad \begin{cases} \Pr(X_n \epsilon\{1,\ 3\}\,|X_m = 2) = (f_{22}^{(n-m)} + f_{20}^{(n-m)}\,|f_{21}^{(0)} = 1) \\ \Pr(X_n \epsilon\{8,10\}\,|X_m = 9) = (f_{02}^{(n-m)} + f_{00}^{(n-m)}\,|f_{01}^{(0)} = 1)\ . \end{cases}$$

Now on using the expressions for $f_{ij}^{(n)}$ in (17) from table 1, we obtain

$$(18)\qquad \Pr(T_A \leq m,\ T_B \leq n) = f_{22}^{(m)} + f_{20}^{(m)} + f_{02}^{(m)} + f_{00}^{(m)} + f_{21}^{(m)} + f_{01}^{(m)} - (\tfrac{1}{2})^{n-m}(f_{21}^{(m)} + f_{01}^{(m)})\ .$$

Similarly for $m \geq n$, we have

$$(19)\qquad \Pr(T_A \leq m\ ,\ T_B \leq n) = f_{22}^{(n)} + f_{20}^{(n)} + f_{02}^{(n)} + f_{00}^{(n)} + f_{12}^{(n)} + f_{10}^{(n)} - (\tfrac{1}{2})^{m-n}(f_{12}^{(n)} + f_{10}^{(n)})\ .$$

Using (18) and (19) and the expressions of $f_{ij}^{(n)}$ of table 1, one easily obtains after some algebra the joint distribution of T_A and T_B as given below.

$$(20)\qquad \begin{cases} \Pr(T_A = T_B = 0) = f_{22}^{(0)} + f_{20}^{(0)} + f_{02}^{(0)} + f_{00}^{(0)}\ , \\ \Pr(T_A = 0, T_B = n) = (\tfrac{1}{2})^n(f_{21}^{(0)} + f_{01}^{(0)})\ ;\ \text{for}\ \ n \geq 1\ , \\ \Pr(T_A = m, T_B = 0) = (\tfrac{1}{2})^m(f_{12}^{(0)} + f_{10}^{(0)})\ ;\ \text{for}\ \ m \geq 1\ , \\ \Pr(T_A = T_B = n) = (a_2)^n(f_{11c}^{(0)} + f_{11r}^{(0)});\ \text{for}\ n \geq 1\ , \\ \Pr(T_A = m, T_B = n) = (\tfrac{1}{2})^{n-m}pq(a_2)^{m-1}(f_{11c}^{(0)} + f_{11r}^{(0)});\ \text{for}\ 1 \leq m < n\ , \\ \Pr(T_A = m, T_B = n) = (\tfrac{1}{2})^{m-n}pq(a_2)^{n-1}(f_{11c}^{(0)} + f_{11r}^{(0)});\ \text{for}\ 1 \leq n < m\ . \end{cases}$$

In the rest of the paper, we shall freely use the convention of replacing a subscript by dot to indicate that a summation has been carried out over that subscript; for instance

$$f^{(n)}_{\cdot 1} = f^{(n)}_{01} + f^{(n)}_{11c} + f^{(n)}_{11r} + f^{(n)}_{21} \quad ,$$

$$f^{(n)}_{11\cdot} = f^{(n)}_{11r} + f^{(n)}_{11c} \ ; \quad \text{and so on.}$$

Now using (20) it is easy to establish that

$$(21) \quad \begin{cases} E(T_A) = 2f^{(0)}_{1\cdot} \ ; \ E(T_B) = 2f^{(0)}_{\cdot 1} \\ Var(T_A) = 2f^{(0)}_{1\cdot}(3-2f^{(0)}_{1\cdot}) \ ; \ Var(T_B) = 2f^{(0)}_{\cdot 1}(3-2f^{(0)}_{\cdot 1}) \\ Cov(T_A,T_B) = [\ \dfrac{6f^{(0)}_{11\cdot}}{1+2pq} - 2f^{(0)}_{1\cdot}\ f^{(0)}_{\cdot 1}] \\ \rho_{T_A T_B} = \dfrac{\{3f^{(0)}_{11\cdot}/(1+2pq)\} - 2f^{(0)}_{1\cdot}\ f^{(0)}_{\cdot 1}}{[f^{(0)}_{\cdot 1} f^{(0)}_{1\cdot}(3-2f^{(0)}_{1\cdot})(3-2f^{(0)}_{\cdot 1})]^{\frac{1}{2}}} \end{cases}$$

As expected the correlation coefficient ρ between T_A and T_B depends both on the initial distribution $\underset{\sim}{f}^{(0)}$ and on how closely the two factors are linked. The closer are they linked - that is smaller is the value of p - the greater is the correlation between them. In particular if $f^{(0)}_{11\cdot} = 1$, then

$$(22) \qquad \rho_{T_A T_B} = (1-4pq)/(1+2pq) \quad ,$$

which is non-negative and is equal to 1 or 0 according as $p = 0$ or $\frac{1}{2}$.

Consider now another random variable T denoting the time when for the first time the population attains homozygosity with respect to both the factors. One can obtain the distribution of T by using the joint distribution of T_A and T_B given by (20), but more easily by noticing that

$$(23) \qquad \Pr(T \leq n) = f_{22}^{(n)} + f_{20}^{(n)} + f_{02}^{(n)} + f_{00}^{(n)} \ ; \ n = 0,\ 1,\ 2,\ \ldots,$$

so that for $n \geq 1$,

$$(24) \qquad \begin{aligned} \Pr(T = n) &= \Pr(T \leq n) - \Pr(T \leq n - 1) \\ &= (1,0,1,0,0,0,0,1,0,1)(\underset{\sim}{f}^{(n)} - \underset{\sim}{f}^{(n-1)}) \\ &= (1,0,1,0,0,0,0,1,0,1)\ \underset{\sim}{L}'(\underset{\sim}{D}^{n} - \underset{\sim}{D}^{n-1})\ \underset{\sim}{M}'\ \underset{\sim}{f}^{(0)} \quad . \end{aligned}$$

Here the last step follows from (12). Simplifying (24), we have

$$(25) \qquad \begin{cases} \Pr(T = 0) = f_{22}^{(0)} + f_{20}^{(0)} + f_{02}^{(0)} + f_{00}^{(0)} \\ \Pr(T = n) = (\tfrac{1}{2})^{n}[f_{21}^{(0)} + f_{12}^{(0)} + f_{10}^{(0)} + f_{01}^{(0)}] + f_{11\cdot}^{(0)}[(\tfrac{1}{2})^{n-1} - (\frac{1+2pq}{2})(a_2)^{n-1}] \ ; \end{cases}$$

$$n \geq 1 \quad ,$$

which yields

$$(26) \begin{cases} E(T) = 2(f_{21}^{(0)} + f_{12}^{(0)} + f_{10}^{(0)} + f_{01}^{(0)}) + \frac{2(1+4pq)}{(1+2pq)}\, f_{11\cdot}^{(0)} \\ \mathrm{Var}(T) = 6(f_{21}^{(0)} + f_{12}^{(0)} + f_{10}^{(0)} + f_{01}^{(0)}) + 2f_{11\cdot}^{(0)}[\ \frac{3+26pq+24p^2q^2}{(1+2pq)^2}] - [E(T)]^2 \ . \end{cases}$$

It is clear from (26) that the expected time it takes to attain complete homozygosity, increases with p (Note that p takes values between 0 and $\frac{1}{2}$) with minimum value when $p = 0$, the complete linkage case and maximum when $p = \frac{1}{2}$, the case where the factors segregate independently. For the special case with $f_{11\cdot}^{(0)} = 1$, we have

$$(27) \quad E(T) = \frac{2(1+4pq)}{(1+2pq)} \; ; \quad Var(T) = \frac{2(1+10pq-8p^2q^2)}{(1+2pq)^2} \quad .$$

From (27) one finds that behavior of $Var(T)$ differs from that of $E(T)$ in that it does not always monotonically increase with p . In fact $Var(T)$ is equal to 2 at $p = 0$, from where it increases with p until it reaches its maximum value $11/4$ at $p = \frac{1}{2}(1-\sqrt{1/3}) \approx \cdot 211$ after which it decreases to $8/3$ at $p = \frac{1}{2}$.

2.2 Loss of Hetrozygosity. This section deals with the study of loss of hetrozygosity for which there appears to be no standard single measure available in literature, particularly when more than one locii are involved. Two of the measures however appear to be more in common use, namely F_n and F_n^* as defined below for the nth generation.

Following Ghai [2], let $H_0^{(n)}$, $H_1^{(n)}$ and $H_2^{(n)}$ denote the proportion in the nth generation of homozygotes, single hetrozygotes and the double hetrozygotes respectively, so that $H_0^{(n)} + H_1^{(n)} + H_2^{(n)} = 1$. Let $H^{(n)} = H_1^{(n)} + H_2^{(n)}$ denote the proportion of all hetrozygotes in the nth generation. F_n is then defined to be the loss in hetrozygosity (single and double hetrozygotes combined) at nth generation relative to that in the initial population and is given by

$$(28) \qquad F_n = 1 - \frac{H^{(n)}}{H^{(0)}} \quad .$$

Again if Z_n denote the number of hetrozygous locii in a genotype randomly drawn from the population at the nth generation, we have

$$(29) \qquad P(Z_n = k) = H_k^{(n)} \; ; \; k = 0, 1, 2 .$$

F_n^* is now defined to be the loss in the mean number of hetrozygous locii that is $E(Z_n)$ relative to $E(Z_0)$, and is given by

$$(30) \qquad F_n^* = 1 - \frac{E(Z_n)}{E(Z_0)} = 1 - \frac{H_1^{(n)} + 2H_2^{(n)}}{H_1^{(0)} + 2H_2^{(0)}} \quad .$$

Another quantity which is of some interest is $Var(Z_n)$ given by

$$(31) \qquad Var(Z_n) = (H_1^{(n)} + 4H_2^{(n)}) - (H_1^{(n)} + 2H_2^{(n)})^2 \quad .$$

Notice that the two measures F_n and F_n^* coincide for the case where only one locus is under consideration. For our present case using the results of table 1, we have

$$(32) \qquad F_n = \frac{(f_{1.}^{(0)} + f_{.1}^{(0)})[1-(\frac{1}{2})^n] - f_{11.}^{(0)} \, [1-(\frac{1-2pq}{2})^n]}{(f_{1.}^{(0)} + f_{.1}^{(0)} - f_{11.}^{(0)})} \quad ,$$

$$(33) \qquad F_n^* = 1 - (\tfrac{1}{2})^n \quad ,$$

and

$$(34)\qquad \begin{cases} E(Z_n) = (\tfrac{1}{2})^n f_{11\cdot}^{(0)} \\ \\ Var(Z_n) = (\tfrac{1}{2})^n [f_{1\cdot}^{(0)} + f_{\cdot 1}^{(0)} - (\tfrac{1}{2})^n f_{11\cdot}^{(0)}]^2 + 2(\frac{1-2pq}{2})^n f_{11\cdot}^{(0)} \end{cases} .$$

It is interesting to note that whereas the measure F_n depends both on the linkage fraction p and the initial distribution $\underset{\sim}{f}^{(0)}$, the measure F_n^* on the other hand is independent of both of these. As expected, both these measures tend to unity as $n \to \infty$. Again, $E(Z_n)$ is independent of p while $Var(Z_n)$ is not. $Var(Z_n)$ increases with the decrease in p. These observations concerning Z_n are not new and can be found in a remark made by Kempthorne (page 80, [4]).

Remark. Chai [2] has used the measure F_n in order to study the loss in hetrozygosity for the case of k independently segregating factors where the population is subjected successively to mixed random mating and selfing. In this case, he has observed that F_n for k independent locii cannot be expressed in terms of the value of F_n obtained for a single locus except for the case where the population is completely selfed in successive generations. To this we may now add keeping (32) in mind that this observation still holds in the presence of linkage even if the population is completely selfed. Here in order to obtain F_n for a single locus it is understood that we ignore the other locus completely. Contrary to the behavior of F_n, the measure F_n^* which turns out to be

$$(35)\qquad F_n^* = \frac{v}{2-v}\,[1-(\tfrac{v}{2})^n]$$

for the case considered by Ghai [2], is independent of the number of independently segregating factors involved. (Here v is the proportion of the population self-fertilized at each generation and the remaining (1-v) kept under random mating).

As such the above observation made for F_n no longer holds if we instead use the measure F_n^* . In this sense then F_n^* may be preferred to F_n , for it does not depend upon the number of independent or linked factors but only on the mode of the mating system used.

2.3 Some Genetic Properties of the Population Undergoing Selfing. Let us consider two quantitative genetic characters, one governed by the gene pair A-a and the other by B-b , both linked. In Mather's notation [6], table 2 gives the various genotypic frequencies in the nth generation along with their genotypic values.

Table 2. Genotypic values and Genotypic frequencies.

	BB	Bb	bb	Total
AA	$f_{22}^{(n)}$ (d_a, d_b)	$f_{21}^{(n)}$ (d_a, h_b)	$f_{20}^{(n)}$ $(d_a, -d_b)$	$f_{2\cdot}^{(n)}$
Aa	$f_{12}^{(n)}$ (h_a, d_b)	$f_{11\cdot}^{(n)} = f_{11c}^{(n)} + f_{11r}^{(n)}$ (h_a, h_b)	$f_{10}^{(n)}$ $(h_a, -d_b)$	$f_{1\cdot}^{(n)}$
aa	$f_{02}^{(n)}$ $(-d_a, d_b)$	$f_{01}^{(n)}$ $(-d_a, h_b)$	$f_{00}^{(n)}$ $(-d_a, -d_b)$	$f_{0\cdot}^{(n)}$
Total	$f_{\cdot 2}^{(n)}$	$f_{\cdot 1}^{(n)}$	$f_{\cdot 0}^{(n)}$	1

Let $M_a^{(n)}$ and $M_b^{(n)}$ be the genotypic means, $\sigma_{aa}^{(n)}$ and $\sigma_{bb}^{(n)}$ the genotypic variances of the two characters respectively. Also, let $\sigma_{ab}^{(n)}$ be their covariance and $\rho_{ab}^{(n)}$ their correlation coefficient. The expressions for these quantities can be easily derived using results of tables 1 and are given as

$$M_a^{(n)} = d_a(2q_A - 1) + \frac{h_a}{2^n} f_{1\cdot}^{(0)} \tag{36}$$

$$M_b^{(n)} = d_b(2q_B - 1) + \frac{h_b}{2^n} f_{\cdot 1}^{(0)} \tag{37}$$

$$\sigma_{aa}^{(n)} = d_a^2[4q_A(1-q_A) - \frac{1}{2^n} f_{1\cdot}^{(0)}] + h_a^2 \frac{f_{1\cdot}^{(0)}}{2^n} (1 - \frac{f_{1\cdot}^{(0)}}{2^n}) - 2d_a h_a \frac{f_{1\cdot}^{(0)}}{2^n} (2q_A - 1) \tag{38}$$

$$\sigma_{bb}^{(n)} = d_b^2[4q_B(1-q_B) - \frac{1}{2^n} f_{\cdot 1}^{(0)}] + h_b^2 \frac{f_{\cdot 1}^{(0)}}{2^n} (1 - \frac{f_{\cdot 1}^{(0)}}{2^n}) - 2d_b h_b \frac{f_{\cdot 1}^{(0)}}{2^n} (2q_B - 1) \tag{39}$$

$$\sigma_{ab}^{(n)} = \sigma_L^{(n)} + \sigma_P^{(n)} \tag{40}$$

$$\rho_{ab}^{(n)} = \frac{\sigma_L^{(n)}}{\sqrt{\sigma_{aa}^{(n)} \ \sigma_{bb}^{(n)}}} + \frac{\sigma_P^{(n)}}{\sqrt{\sigma_{aa}^{(n)} \ \sigma_{bb}^{(n)}}} \tag{41}$$

where

$$\sigma_L^{(n)} = d_a d_b (f_{11c}^{(0)} - f_{11r}^{(0)})(\frac{1-2p}{1+2p})[1 - (\frac{1-2p}{2})^n] + f_{11\cdot}^{(0)}(\frac{h_a h_b}{2^n})[(1-2pq)^n - (\tfrac{1}{2})^n] , \tag{42}$$

$$(43) \quad \sigma_P^{(n)} = d_a d_b (f_{22}^{(0)} + f_{00}^{(0)} - f_{20}^{(0)} - f_{02}^{(0)} - 4q_A q_B + 2q_A + 2q_B - 1)$$

$$+ \frac{h_a h_b}{2^{2n}} (f_{11\cdot}^{(0)} - f_{1\cdot}^{(0)} f_{\cdot 1}^{(0)}) + \frac{d_a h_b}{2^n} [(f_{21}^{(0)} - f_{01}^{(0)}) - f_{\cdot 1}^{(0)}(2q_A - 1)]$$

$$+ \frac{d_a d_b}{2^n} [(f_{12}^{(0)} - f_{10}^{(0)}) - f_{1\cdot}^{(0)}(2q_B - 1)] ,$$

and q_A and q_B are the gene frequencies of genes A and B respectively given by

$$(44) \quad q_A = f_{2\cdot}^{(n)} + \tfrac{1}{2} f_{1\cdot}^{(n)} = f_{2\cdot}^{(0)} + \tfrac{1}{2} f_{1\cdot}^{(0)} \; ; \; q_B = f_{\cdot 2}^{(n)} + \tfrac{1}{2} f_{\cdot 1}^{(n)} = f_{\cdot 2}^{(0)} + \tfrac{1}{2} f_{\cdot 1}^{(0)} ,$$

keeping in mind that the gene frequencies remain unchanged from generation to generation. It is interesting to note that the first component of the genetic correlation coefficient (41) is entirely due to the presence of linkage and is zero if $p = \frac{1}{2}$. On the other hand the second component of (41) is based only on the initial genotypic distribution $\underset{\sim}{f}^{(0)}$ and is independent of p. As expected, the first component disappears again if $f_{11\cdot}^{(0)} = 0$.

One can easily find the limiting expressions for the above quantities as $n \to \infty$. Also one may study the behaviour of these quantities for several special cases such as the one with complete dominance or with no dominance etc. Finally, if instead the same character is governed by both the factors A-a and B-b and if the affects are additive, one can derive the expressions for the genotypic mean and variance in a similar manner. However, we shall not touch these various possibilities here any further. Instead in the next section, we proceed to consider the important case of mixed random mating and selfing.

3 *Population undergoing mixed random mating and self-fertilization.* In this section we consider the case where each generation is produced by subjecting a proportion v of the previous generation to self-fertilization and the remaining portion $u = 1 - v$ to random mating, starting with an initial population with the distribution vector $\underset{\sim}{f}^{(0)}$ of (2). Let $r_{11}^{(n)}$, $r_{10}^{(n)}$, $r_{01}^{(n)}$ and $r_{00}^{(n)}$ denote the proportions of gametes AB, Ab, aB and ab respectively produced by the nth generation, so that

$$(45) \quad \begin{cases} r_{11}^{(n)} = f_{22}^{(n)} + \frac{1}{2}(f_{21}^{(n)} + f_{12}^{(n)} + qf_{11c}^{(n)} + pf_{11r}^{(n)} \\ r_{10}^{(n)} = f_{20}^{(n)} + \frac{1}{2}(f_{21}^{(n)} + f_{10}^{(n)} + pf_{11c}^{(n)} + qf_{11r}^{(n)}) \\ r_{01}^{(n)} = f_{02}^{(n)} + \frac{1}{2}(f_{12}^{(n)} + f_{01}^{(n)} + pf_{11c}^{(n)} + qf_{11r}^{(n)}) \\ r_{00}^{(n)} = f_{00}^{(n)} + \frac{1}{2}(f_{10}^{(n)} + f_{01}^{(n)} + qf_{11c}^{(n)} + pf_{11r}^{(n)}) \quad . \end{cases}$$

(45) can be rewritten in the matrix form as

$$(46) \qquad \underset{\sim}{r}^{(n)} = \underset{\sim}{B}\,\underset{\sim}{f}^{(n)} \quad ,$$

where

$$(47) \qquad \underset{\sim}{r}^{(n)} = (r_{11}^{(n)}, r_{10}^{(n)}, r_{01}^{(n)}, r_{00}^{(n)})'$$

and

$$(48) \quad \underset{\sim}{B} = \begin{pmatrix} 1 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{q}{2} & \frac{p}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 1 & 0 & \frac{p}{2} & \frac{q}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{p}{2} & \frac{q}{2} & 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{q}{2} & \frac{p}{2} & \frac{1}{2} & 0 & \frac{1}{2} & 1 \end{pmatrix} \quad .$$

For the later developments, we need the following two lemmas.

Lemma 1. *A population with genotypic distribution vector* $\underset{\sim}{f}^{(0)}$ *is in equilibrium with respect to random mating if and only if*

$$(49)\quad \underset{\sim}{f}^{(0)} = ([r_{11}^{(0)}]^2,\ 2r_{11}^{(0)}r_{10}^{(0)},\ [r_{10}^{(0)}]^2,\ 2r_{11}^{(0)}r_{01}^{(0)},\ 2r_{11}^{(0)}r_{00}^{(0)},\ 2r_{10}^{(0)}r_{01}^{(0)},\ 2r_{10}^{(0)}r_{00}^{(0)},\ [r_{01}^{(0)}]^2,\ 2r_{01}^{(0)}r_{00}^{(0)},\ [r_{00}^{(0)}]^2)'$$

and

$$(50)\qquad r_{11}^{(0)}r_{00}^{(0)} = r_{10}^{(0)}\, r_{01}^{(0)} \quad ,$$

where $r_{ij}^{(0)}$'s *are as defined in* (45) *for* $n = 0$.

The proof of this lemma can be found in Kempthorne (pages 38-41, [4]).

Lemma 2. *For any distribution vector* $\underset{\sim}{f}^{(0)}$ *satisfying the conditions of lemma 1* ,

$$(51)\qquad \underset{\sim}{B}(\underset{\sim}{P}')^n\, \underset{\sim}{f}^{(0)} = \underset{\sim}{r}^{(0)} \ ; \quad n = 0,\ 1,\ 2,\ \ldots \ .$$

The proof follows by direct computation and from the facts that $\underset{\sim}{P}^n = \underset{\sim}{M}\,\underset{\sim}{D}^n\,\underset{\sim}{L}$ and that $r_{11}^{(0)}\, r_{00}^{(0)} = r_{10}^{(0)}\, r_{01}^{(0)}$.

From hereon we assume that our initial population is in equilibrium with respect to random mating so that the distribution vector $\underset{\sim}{f}^{(0)}$ satisfies the conditions of lemma 1. From (4), (46) and (51), it follows that if this population is subjected successively to complete selfing, then $\underset{\sim}{r}^{(n)} = \underset{\sim}{r}^{(0)}$ for all n . In our case, this being true separately for both the portions of the population under selfing and under random mating at each generation, we conclude that the gametic frequency vector $\underset{\sim}{r}^{(n)}$ remains unchanged over all generations even when the population is subjected simultaneously to both types of mating systems in the manner specified above. Furthermore, from this it follows that under mixed self-fertilization and random mating, the distribution vector of the nth generation is given by

$$(52)\qquad \underset{\sim}{f}^{(n)} = u\,\underset{\sim}{f}^{(0)} + v\,\underset{\sim}{P}'\,\underset{\sim}{f}^{(n-1)} \quad ,$$

where $\underset{\sim}{f}^{(0)}$ is as given in (49). Interating (52) over n and using (11) we obtain

$$(53)\qquad \underset{\sim}{f}^{(n)} = u\,\underset{\sim}{L}'[\underset{\sim}{I} + v\,\underset{\sim}{D} + v^2\underset{\sim}{D}^2 + \ldots + v^{n-1}\underset{\sim}{D}^{n-1}]\underset{\sim}{M}'\underset{\sim}{f}^{(0)} + v^n\underset{\sim}{L}'\underset{\sim}{D}^n\underset{\sim}{M}'\underset{\sim}{f}^{(0)} \quad .$$

This simplifies further to

$$(54)\qquad \underset{\sim}{f}^{(n)} = u\,\underset{\sim}{L}'\,\underset{\sim}{G}_n\,\underset{\sim}{M}'\,\underset{\sim}{f}^{(0)} + v^n\,\underset{\sim}{L}'\,\underset{\sim}{D}^n\,\underset{\sim}{M}'\,\underset{\sim}{f}^{(0)} \quad ,$$

where

$$(55)\qquad \underset{\sim}{G}_n = dg(\frac{1-v^n}{u}, \frac{1-v^n}{u}, \frac{1-v^n}{u}, \frac{1-v^n}{u}, \frac{1-(\frac{v}{2})^n}{(\frac{1+u}{2})}, \frac{1-(\frac{v}{2})^n}{(\frac{1+u}{2})}, \frac{1-(\frac{v}{2})^n}{(\frac{1+u}{2})} \quad ,$$

$$\frac{1-(\frac{v}{2})^n}{(\frac{1+u}{2})}, \frac{1-(v\,\frac{1-2pq}{2})^n}{1-v\,(\frac{1-2pq}{2})}, \frac{1-(v\,\frac{1-2p}{2})^n}{1-v(\frac{1-2p}{2})}).$$

One can immediately find the various elements of $\underset{\sim}{f}^{(n)}$ similar to those in table 1 from (54). Again on letting $n \to \infty$ in (54) we have the limiting genotypic distribution vector given by

$$(56) \qquad \underset{\sim}{f}^{(\infty)} = u\, \underset{\sim}{L}'\, \underset{\sim}{G}_{\infty}\, \underset{\sim}{M}'\, \underset{\sim}{f}^{(0)} \quad ,$$

where

$$(57) \quad \underset{\sim}{G}_{\infty} = dg(\frac{1}{u}, \frac{1}{u}, \frac{1}{u}, \frac{1}{u}, \frac{2}{1+u}, \frac{2}{1+u}, \frac{2}{1+u}, \frac{2}{1+u}, \frac{2}{2-v(1-2pq)}, \frac{2}{2-v(1-2p)}) \; .$$

$$= (\underset{\sim}{I} - v\, \underset{\sim}{D})^{-1} \; .$$

Using (54) one can as in section 2.3 study the behavior of various genetic properties of the distribution at nth generation, with changes in u, p and other basic elements. As expected, the expressions of these properties are bit lengthy and we will not tackle them here. We close this section with the final remark that in the presence of random mating the problem of ultimate attainment of homozygosity as discussed in section 2.1 does not arise here.

4 Concluding Remarks. In section 2.2, the author has avoided labeling F_n as the coefficient of inbreeding, simply because he is unaware of any extension of the concept of coefficient of inbreeding to the case of linked factors. In the event there is no such extension available, this appears to be an interesting problem for further investigation. Again the methods used in section 2 based on the study of a Markov chain are similar to what Kempthorne [4] calls as the "Generation matrix method". It is however more revealing to study these problems in the light of theory of Markov chains where ever possible, as some of the already available results of the fairly well developed theory of Markov chains can be applied without any extra cost. The results of section 3 have been generalised to the case where the initial population has an arbitrary distribution and will be communicated elsewhere (see Puri [7]). Also, the above model can be made more realistic by

incorporating factors such as selection etc.. Unfortunately however, this makes the algebra somewhat more involved. Finally, I hope that the inadequacies of this paper will not disguise my respects, admiration and affection for Dr. V. G. Panse, to whom it is dedicated.

References

[1] Ghai, G. L. (1964). The genotypic composition and variability in plant populations under mixed self-fertilization and random mating, J. Ind. Soc. Agr. Stat. 16, pp. 93-125.

[2] Ghai, G. L. (1966). Loss of hetrozygosity in populations under mixed random mating and selfing, J. Ind. Soc. Agr. Stat., 18, pp. 71-81.

[3] Karlin, S. (1966). A first course in Stochastic Processes, Academic Press, New York.

[4] Kempthorne, O. (1957). An Introduction to Genetic Statistics, John Wiley and Sons, Inc., London.

[5] Li, C. C. (1967). Genetic equilibrium under selection, Biometrics, 23, pp. 397-484.

[6] Mather, K. (1949). Biometrical Genetics, Methuen, London.

[7] Puri, Prem S. (1967). A model with mixed self-fertilization and random mating in the presence of linkage. Mimeo Series No. 135, Department of Statistics, Purdue University, Lafayette, Indiana.