ON THE PROBABILITY THEORY OF LINKAGE IN MENDELIAN HEREDITY

By Hilda Geiringer

Bryn Mawr College

1. Introduction. If for a certain generation the distribution of genotypes is known and a certain law of heredity is assumed, the distribution of genotypes in the next generation can be computed. Suppose there are N different genotypes in the nth generation in the proportions $x_1^{(n)}, \cdots, x_N^{(n)}$ where $\sum_{i=1}^{N} x_i^{(n)} = 1$ and denote by $p_{\kappa\lambda}^{i}$ the probability that an offspring of two parents of types κ and λ be of type i where $\sum_{i=1}^{N} p_{\kappa\lambda}^{i} = 1$ for all κ and λ, and $p_{\kappa\lambda}^{i} = p_{\lambda\kappa}^{i}$. Assuming panmixia, identical distributions $x_i^{(n)}$ for males and females, etc., we can derive $x_i^{(n+1)}$ from $x_i^{(n)}$ by means of the formula

$$x_i^{(n+1)} = \sum_{\kappa,\lambda=1}^{N} p_{\kappa\lambda}^{i} x_\kappa^{(n)} x_\lambda^{(n)} \qquad (i = 1, 2, \cdots N). \tag{1}$$

Thus if the distribution $x_i^{(0)}$ is given for an initial generation we can deduce successively the $x_i^{(1)}, x_i^{(2)}, \cdots$ for subsequent generations. Besides, one may wish to express the $x_i^{(n)}$, for any n, explicitly in terms of the initial distribution $x_i^{(0)}$, i.e. to "solve" the system (1). A further problem consists in determining the limit-distribution of the genotypes $\lim_{n\to\infty} x_i^{(n)}$ $(i = 1, \cdots, N)$.

Mendel's heredity theory is based on some ingenious assumptions which are known as Mendel's first and second law. They enable us to define the possible genotypes and to establish the recurrence formula (1); they will be explained and formulated in sections 2 and 3. It is well known that in Mendel's theory it makes an essential difference whether one or more "Mendelian characters" are considered. In the first case Mendel's first law only is used; there are with respect to this character but $N = 3$ different types and the recurrence formula (1) can be derived without difficulty. As early as 1908 G. H. Hardy [5] established the simple but most remarkable result that under random breeding a state of equilibrium is reached in the first filial generation, i.e. $x_i^{(1)} \neq x_i^{(0)}$ (in general) but $x_i^{(n)} = x_i^{(1)}$ $(n = 2, 3, \cdots,)$.[1]

In the case of $m \geqq 2$ Mendelian characters Mendel assumed *independent assortment* of these characters (Mendel's second law). However, within four years after the dramatic rediscovery of Mendel's fundamental paper [10], observations were reported that did not show the results expected for two independent characters. T. H. Morgan [11] and collaborators in basic contributions, con-

[1] See also [12] where the stability of the particular ratio 1:2:1 is recognized.

cluded that a certain *linkage* of genes was to be assumed.[2] Taking that as the starting point, the main purpose of this paper is to establish the basic recurrence formula for the general case of linkage, to solve the corresponding system of difference equations, and to determine the limit distribution of genotypes. Throughout the paper "multiple alleles" are considered instead of making the frequent restriction to two alleles. This generalization is, however, an obvious one (see section 1).

In order to deal with the general problem a *linkage distribution* (l.d.), is introduced. This concept, which seems to be basic to the whole problem, refers to the probability theory of arbitrarily linked events [3]. The *crossover probabilities*, (c.p.), defined by Morgan and Haldane are, notwithstanding their high importance, not sufficient for our purpose. (They turn out to be certain marginal distributions of the l.d.) If, however, $m = 2$ and $N = 10$ (for two alleles), a case studied by W. Weinberg [16] H. S. Jennings [7] and R. B. Robbins [14], the c.p. is equivalent to the l.d. But for the general case the l.d. is needed and the desired results must be derived by other methods than explicit computation, which is feasible if m equals one or two. The original problem of independent assortment appears, of course, as a particular case of the general linkage. This problem was completely solved in 1923 by H. Tietze [15] in a very interesting but rather involved paper. The proof of the limit theorem given in the following pages for the general case is far simpler and shorter than the treatment of the particular case in the older paper and is therefore a simpler proof of Tietze's theorem.

After a brief consideration of the classical case $m = 1$ (section 2), the problem of m arbitrarily linked characters is discussed in section 3 with a particular view to a clear statement of the biological and mathematical assumptions. The l.d., its relation to the c.p.. and some basic properties of both are considered in section 4. Then, after a very concise consideration of the case $m = 2$ (section 5), the basic recurrence formula is established in section 6 from which we deduce in section 7 two general limit theorems. The main point is that the limit distribution of genotypes is "uncorrelated" and equals the product of certain marginal distributions of first order deduced from the distribution for the first filial generation. As a kind of an appendix section 8 contains the solution of the system of equations furnished by the general recurrence formula.

In the second part of the paper an attempt is made to contribute to the *linear theory* or theory of the linear order of the genes, from the point of view of probability theory. Accordingly, the linear theory consists in certain assumptions on the l.d., or on an equivalent distribution which will be called *crossover distribution*, (c.d.), and which is more appropriate for this purpose. (Sections 9

[2] "To T. H. Morgan and his associates and students is due the credit for opening up this new field of genetic research; and the small vinegar fly Drosophyla Melanogaster upon which most of their work has been based, has now assumed as great an importance in genetics as the famous peas studied by Mendel." (Sinnot and Dunn, *Principles of Genetics*, New York 1939.)

and 10.) In this connection in section 10 a probability definition of the "distance" d_{ij} of two genes is proposed which, far from being contradictory to Morgan's ideas on the subject seems to formulate them mathematically; (the distance d_{ij} between two genes i and j is defined as the mathematical expectation of the number of crossovers between i and j). This distance is of course additive as it ought to be in the framework of the linear theory.

A problem frequently discussed is whether the crossover probabilities are independent of each other (this independence is not identical with Mendel's free assortment). Observations (see [4a]) did not seem to substantiate this as a general assumption. Then it was concluded that there exists a so-called *interference* which prevents, i.e. diminishes the probability of crossovers too "near" to each other. (See also [13].) It seems to the author that observations on interference should be interpreted in terms of appropriate assumptions regarding the l.d. or the c.d. Again the remark holds that the c.p.'s are not sufficient for describing the situation. Hence in section 11 an attempt is made to understand "interference" by means of the c.d., accepting however the linear theory. It is well known that the explicit presentation of consistent dependent distributions is not trivial (see e.g. [2]). Not many different types of "contagious" distributions are known. In section 11 two such schemes are proposed which, though simple enough, seem to correspond to the general idea of interference. They contain as particular cases the case of independent and the case of disjoint crossovers.

2. One Mendelian character. Hardy's theorem. It will be helpful to start with the simple and well known case of one character introducing the basic concepts in a way appropriate for generalization.

Mendel recognized that the distribution of certain hereditary attributes in organisms is similar to the distribution of attributes in a probability distribution. With respect to a *Mendelian character* each individual is characterized by *two* elements called *genes* which represent two possible alternatives. The color of the flower of peas is such a character, the alternatives being red and white. With respect to this character each plant belongs to one of the three types: red-red, red-white, white-white.[3] These are three different *genotypes*.—In this paper genotypes only will be considered. The difference between genotypes and phenotypes and the related concepts of dominant and recessive qualities will not be dealt with. This is an example of a *two-valued* Mendelian character, i.e. a character for which only two possibilities exist or, using a more technical term,

[3] It will be assumed throughout that the individuals considered are "diploid" That means in the terminology of the preceding example that the only possible types are RR, RW, and WW; or, using A and a: AA, Aa, and aa. Modern research has however revealed that situations may arise where "tetraploids", "hexaploids", etc. briefly "polyploids" prevail, i.e. types like $A^x a^y$ (with $x + y = 2p$). In this case the reproduction cell segregates $A^{x_1} a^{y_1}$ (with $x_1 + y_1 = p$). Stability is no longer reached in the first filial generation. See [4b].

with two alleles. The case of two alleles is most frequently considered in the biological literature where the two possibilities correspond mostly to a dominant and a recessive quality. There is, however, no difficulty in considering from the very beginning the general case of *multiple alleles* where the character under consideration is assumed to be r-valued, i.e. susceptible to r different manifestations (e.g. $r = 5$ possible colors of a plant). These r possible values may be distinguished by the r arguments, $1, 2, \cdots r$.[4]

In the consideration of only one Mendelian character *Mendel's first law* only is used which may be stated as follows:

(a). With respect to one r-valued Mendelian character each individual belongs to one of the $r(r+1)/2$ possible types, each type being determined by a pair of elements (genes) x and y $\begin{pmatrix} x = 1, \cdots, r \\ y = 1, \cdots, r \end{pmatrix}$.

(b). In the formation of a new individual each parent transmits one of its two genes to the new individual, the other gene coming from the other parent.

(c). The probability for the transmission of either gene is the same and thus equals $\frac{1}{2}$.

We wish to deduce the distribution of genotypes in the $(n+1)$st generation from the distribution of genotypes in the nth generation *under the assumption of complete panmixia* (random breeding). Moreover, assume that the given initial distributions of genotypes as well as the laws of heredity are the *same for males and females*.[5] In computing successively the new distribution from the preceding one we shall always assume that the distribution of individuals participating in the process of procreation is the same as their distribution when born.

Let us denote a genotype by $(x; y)$, $(x = 1, \cdots, r; y = 1, \cdots, r)$. To fix the ideas we shall assume through this paper that the gene x before the semicolon was transmitted by the mother, and the y after the semicolon by the father of the individual. In some cases which will be considered later this distinction will be relevant. Denote by $w^{(n)}(x; y)$ the probability of the type $(x; y)$ in the nth generation. Since the laws of heredity are the same for males and females we have $w^{(n)}(x; y) = w^{(n)}(y; x)$ and thus have for each generation a symmetric distribution of genotypes with r^m probabilities whose sum is one. There is, however, according to principle (a) no difference between the types $(x; y)$ and $(y; x)$ and therefore it is preferable to group together these types, thus introducing for $x = 1, \cdots, r; y = 1, \cdots, r$:

$$
\begin{aligned}
v^{(n)}(x; x) &= w^{(n)}(x; x) \\
v^{(n)}(x; y) &= w^{(n)}(x; y) + w^{(n)}(y; x) \text{ where } x < y.
\end{aligned} \tag{2}
$$

[4] "It is simplest to deal with mere pairs of alternative conditions (alleles) but a theory remains seriously inadequate unless capable of extension to multiple alleles." ([17] p. 224).

[5] It is obvious that we may admit without any change of result different distributions for males and females in the initial generation, as long as random mating takes place afterwards.

Consequently there are $r(r+1)/2$ such probabilities:

(2′) $v^{(n)}(1;1), v^{(n)}(1;2), \cdots v^{(n)}(1;r), v^{(n)}(2;2), \cdots v^{(n)}(2;r), \cdots v^{(n)}(r;r)$

where

$$\sum_{x \leqq y} v^{(n)}(x;y) = 1 \qquad (n = 0, 1, 2, \cdots). \tag{3}$$

Now define $p^{(n)}(x)$ as *the probability that in the nth generation a male (or a female) individual transmits the gene* x. Obviously we have:

$$p^{(n)}(x) = \tfrac{1}{2} v^{(n)}(1;x) + \tfrac{1}{2} v^{(n)}(2;x) + \cdots + v^{(n)}(x;x) + \tfrac{1}{2} v^{(n)}(x;x+1) + \cdots + \tfrac{1}{2} v^{(n)}(x;r) \tag{4}$$

and

$$\sum_{x=1}^{r} p^{(n)}(x) = 1 \tag{4′}$$

In fact, the gene x will be transmitted, if an individual possesses this gene and also transmits it. The individuals of type $(y;x)$ (or $(x;y)$) all possess the gene x and transmit it with probability $\frac{1}{2}$ if $y \neq x$ and with probability 1 if $y = x$. Besides, the probability of the type $(x;y)$ in the $(n+1)$st generation is obviously $p^{(n)}(x)p^{(n)}(y)$:

$$w^{(n+1)}(x;y) = p^{(n)}(x)p^{(n)}(y) = w^{(n+1)}(y;x) \tag{5}$$

or in terms of the $v^{(n)}(x;y)$

$$\begin{aligned} v^{(n+1)}(x;x) &= [p^{(n)}(x)]^2 \\ v^{(n+1)}(x;y) &= 2p^{(n)}(x)p^{(n)}(y) \qquad (x \leqq y). \end{aligned} \tag{5′}$$

Hence by (4) and (5′), $v^{(n+1)}$ has been expressed in terms of $v^{(n)}$ and the recurrence-problem is solved. The distribution $w^{(n+1)}(x;y)(n \geqq 0)$ shows "independence," and is therefore known to be stable. In fact, computing in the same way $p^{(n+1)}(x)$ we get

$$\begin{aligned} p^{(n+1)}(x) &= \tfrac{1}{2}\cdot 2p^{(n)}(1)p^{(n)}(x) + \cdots p^{(n)}(x)p^{(n)}(x) + \tfrac{1}{2}\cdot 2p^{(n)}(x)p^{(n)}(x+1) + \cdots + \tfrac{1}{2}\cdot 2p^{(n)}(x)p^{(n)}(r) \\ &= p^{(n)}(x)\cdot\sum_{\rho=1}^{r} p^{(n)}(\rho) = p^{(n)}(x) \end{aligned}$$

or

$$p^{(n+1)}(x) = p^{(n)}(x) \qquad (n = 0, 1, 2, \cdots), (x = 1, 2, \cdots r). \tag{6}$$

This last formula contains G. H. Hardy's famous result [5] *that* $p^{(n)}(x)$ *is the same for all* n:

$$p^{(n)}(x) = p^{(0)}(x) \qquad (n = 1, 2, \cdots,) \tag{7}$$

and because of (5′):

$$(7') \qquad v^{(n)}(x; y) = v^{(1)}(x; y) \qquad (x \leqq y, n = 2, 3, \cdots).$$

In case of only one Mendelian property the distribution of genotypes reaches a stationary state in the first filial generation.

3. Basic assumptions in case of m Mendelian characters. A new situation presents itself if there is more than one character. In case of m characters a genotype is described by $2m$ numbers $(x_1, \cdots, x_m; y_1, \cdots, y_m)$ or briefly $(x; y)$ (e.g. for $m = 5$, $r = 9$:(1,2,3,4,6; 2,7,3,5,9). There are primarily $N = r^{2m}$ possible types because on each of the $2m$ places any of the r numbers can be written. Now, if the types $(x; y)$ and $(y; x)$ are considered as identical genotypes, the number of different genotypes reduces to $N_1 = \frac{r^m}{2}(r^m + 1)$ (e.g. for $r = 2$, $m = 1$:$N_1 = 3$; for $r = m = 2$: $N_1 = 10$). It is essential for the understanding of linkage that in counting this way two types like (1,3; 5.7) and (1,7; 5,3) or (1,1;2,2) and (1,2;2,1) are considered as different although in both cases the individual possesses with respect to the first character the gene pair 1,5 and with respect to the second the pair 3,7. If no difference is assumed between two such types the number of different genotypes reduces to $N_2 = \left(\frac{r(r+1)}{2}\right)^m$. (E.g. for $r = 2$:$N = 4^m$, $N_1 = \frac{1}{2}\cdot 2^m(2^m + 1)$, $N_2 = 3^m$; hence for $m = r = 2$:$N = 16$, $N_1 = 10$, $N_2 = 9$ or for $r = 2$, $m = 3$:$N = 64$, $N_1 = 36$, $N_2 = 27$). Which method of counting is the correct one?

The answer is that there are but N_2 different genotypes if *Mendel's second law*, the *law of independent assortment*, is accepted. Then and only then there is no difference between types like (1,3:5,7) and (1,7:5,3). Under the assumption of general linkage however, these types must be distinguished, not as individuals, but with respect to their heredity properties, i.e. considered as parents of a new generation. Under this assumption there are in general N_1 different types. This will be discussed presently in more detail.

Let us first consider Mendel's original theory as contained in his *first and second law*. Analogous to (a), (b) and (c) in §2 we now formulate as follows:

(a′) With respect to m characters the genotype of an individual is characterized by m pairs of numbers. Two individuals are of the same type if to each of the m characters corresponds the same pair. Hence there are $N_2 = \left(\frac{r(r+1)}{2}\right)^m$ genotypes.

(b′) In the formation of a new individual a parent of type $(x_1, \cdots, x_m; y_1, \cdots, y_m)$ transmits to the offspring, corresponding to each of the m characters, one of the two genes which he (or she) possesses with respect to this character.

(c′) The probability of transmitting any of these 2^m combinations is the same and therefore equal to $1/2^m$.

Consider e.g. the individual (1,2,3:1,4,7); the pair 1,1 corresponds to the first character the pair 2,4 to the second and 3,7 to the third. Under the assumptions of Mendel's original theory this individual is of the same type with (1,4,3; 1,2,7) and (1,2,7; 1,4,3), and of course with (1,4,7; 1,2,3), etc. As $m = 3$, it may transmit eight combinations which in the preceding example reduce to four, because the individual is homozygous in the first character. These four combinations are 1,2,3 or 1,4,3 or 1,2,7 or 1,4,7 each with probability $2 \times \frac{1}{8} = \frac{1}{4}$.

The distribution of genotypes in successive generations under the assumption of Mendel's second law has been investigated by H. Tietze [15] who also considers the limiting distribution as $n \to \infty$. His results will appear as a particular case of our general considerations.

In order to discuss the basic facts which lead to the idea of linkage let us for the moment consider the case $m = 2$. Soon after the rediscovery of Mendel's work Bateson and Punett reported observations which did not give the expected numerical results. To understand the type of such an observation assume that a homozygous male of type (1,1; 1,1), [or any other homozygous type, e.g. (2,3; 2,3)] is mated to a homozygous female of type (2,2; 2,2) [or to any homozygous type different from the first e.g. (4,5; 4,5)]. Obviously, in this case there is only one possible kind of offspring namely (2,2; 1,1), [or (4,5; 2,3)]. But if now one of these daughters is mated to a homozygous male of the original type (1,1; 1,1), there are four kinds of possible offspring, namely (2,2; 1,1), (2,1; 1,1), (1,2; 1,1), and (1,1; 1,1), corresponding to the four combinations of genes transmitted by the heterozygous (dihybrid) daughter [or (4,5; 2,3), (4,3; 2,3), (2,5; 2,3), and (2,3; 2,3)]; and according to the idea of free assortment each of these four combinations should appear with the same relative frequency: $\frac{1}{4}$. But it was observed that the combined frequency of the two types (1,1; 1,1) and (2,2; 1,1) was larger than that of the types (2,1; 1,1) and (1,2; 1,1). "The characters that went in together have come out together in a much higher percentage than expected from Mendel's second law, viz. the law of independent assortment" [11]. Morgan, in his theory of the gene called this "tendency" *linkage*. The idea is that the two genes 2,2 and 1,1 which have been together in the maternal individual tend to stay together and that nature has to make an effort to produce a so-called *crossing-over*, i.e. a separation of the genes "that came in together,"—such that a female of type (2,2; 1,1) may transmit the group 1,2 or the group 2,1. In other words, *the idea of linkage implies an influence of the grandparents.*

According to observation the percentage of crossing over varies from 0 to 50 per cent, i.e. from *complete linkage* to *free assortment*. It will appear however that in principle crossover-values greater than 50 per cent cannot be excluded. It was also observed that the percentage of individuals of type (1,1; 1,1) equals very nearly that of individuals of type (2,2; 1,1), as we would expect. In the same way the percentages of types (2,1; 1,1) and (1,2; 1,1) are nearly equal, their sum yielding the *crossover-ratio*. Hence the four probabilities correspond-

ing to the formation of the four types (1,1; 1,1), (2,2; 1,1), (2,1; 1,1), and (1,2; 1,1) are assumed to be $(1 - c)/2$, $(1 - c)/2$, $c/2$, and $c/2$. It is important to notice that these are at the same time the probabilities that the female of type (2,2; 1,1) (which was mated to the homozygous (1,1; 1,1), transmits the groups 1,1 or 2,2 or 2,1 or 1,2 respectively.

In the general case of m characters there are $\binom{m}{2} = \frac{m(m-1)}{2}$ crossover probabilities. In this case Morgan assumes *linkage-groups*, each group consisting of m_i elements with $\sum_i m_i = m$, such that "there is linkage between the elements of each group but that the members belonging to different linkage groups assort independently, in accordance with Mendel's second law." This idea will be reconsidered in Section 4.

If we now wish to solve our first basic problem, i.e. to derive the distribution of genotypes in any later generation from an initial distribution of genotypes, then the concept of crossover probabilities does not suffice. The complex possibilities which arise if Mendel's second law is no longer accepted as universally valid cannot be adequately described in terms of crossover probabilities. Or, more exactly: It will be seen that if $m \geqq 4$ the crossover probabilities are no longer sufficient, whereas for $m = 2$ and $m = 3$ this concept is general enough. For the complete description of the hereditary mechanism in the general case a so-called *linkage distribution*, l.d., is needed which involves 2^m probabilities with sum equal to one. Let us define this distribution.

Consider an individual of type $(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv (x; y)$, where the x are the maternal genes, the genes contributed by the mother of the individual, and the y the paternal genes. Denote by S the set of the m numbers $1,2, \cdots, m$, by A any subset of S, and by A' the complementary subset $A' = S - A$. *Denote by $l(A)$ the probability that the individual $(x; y)$ transmits the paternal genes belonging to A and the maternal genes belonging to A'.* There are $1 + m + \binom{m}{2} + \cdots + 1 = 2^m$ such subsets A and accordingly 2^m probabilities $l(A)$ where

$$\sum_{(A)} l(A) = 1. \tag{8}$$

In accordance with the previously reported observations and with our assumption of equal conditions for both sexes one must assume that

$$l(A) = l(A'). \tag{8'}$$

The conditions (8) and (8') reduce the number of freely disposable values of the l-distribution to $(2^{m-1} - 1)$. The l-distribution is a socalled *m-dimensional* or *m-variate* alternative which could also and occasionally will be denoted by $l(\epsilon_1, \epsilon_2, \cdots, \epsilon_m)$ where $\epsilon_i = 0$ or 1. Thus e.g. $l(1,1,1,0,0,1)$ is the probability that the genes $y_1, y_2, y_3, x_4, x_5, y_6$, are the genes contained in the germ cell of the individual $(x; y)$. Here the set A consists of the numbers 1,2,3,6, and A' of 4,5.

Analogous to the statements (a), (b), (c) of §2 and (a'), (b'), (c') of the present section we may now formulate the *principles of Mendel's theory of heredity under the assumption of a possible linkage of the genes:*

(a'') With respect to m Mendelian characters an individual is characterized by two sets of numbers each consisting of m numbers, viz. $x_1, \cdots, x_m$ and $y_1, \cdots, y_m$, where $\begin{matrix} x_i \\ y_i \end{matrix} = 1,2, \cdots, r$. If the type of an individual is designated by $(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv (x; y)$ where x and y denote the maternal and paternal contributions respectively, then $(x; y) = (y; x)$. Hence there are $N_1 = \frac{1}{2}r^m(r^m + 1)$ types of individuals.

(b'') $\equiv$ (b') In the formation of a new individual each parent transmits to the offspring one set of m genes.

(c'') For each parent, the 2^m probabilities of transmitting any one of these 2^m possible sets are given by a linkage distribution $l(A)$ where A is a subset of the set S consisting of the m numbers 1, 2, $\cdots$, m, and $l(A)$ is the probability that the transmitted set consists of the paternal genes belonging to A and of the maternal genes belonging to $A' = S - A$, and $l(A) = l(A')$.

4. Some properties of the linkage distribution and of the crossover probabilities. In the following we shall need marginal distributions, that is partial sums, of the probabilities within a distribution. In a usual notation:

$$
\begin{aligned}
&l_1(x_1) = \sum_{x_2}\sum_{x_3}\cdots\sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots\ldots\ldots\ldots m \text{ distributions} \\
&l_{12}(x_1, x_2) = \sum_{x_3}\cdots\sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots\ldots\ldots \binom{m}{2} \text{ distributions} \\
&l_{123\cdots m-1}(x_1, x_2, \cdots, x_{m-1}) = \sum_{x_m} l(x_1, x_2, \cdots, x_m), \ldots\ldots\ldots m \text{ distributions} \\
&l_{12\ldots,m}(x_1, x_2, \cdots, x_m) = l(x_1, x_2, \cdots, x_m) \ldots\ldots \text{the original distribution.}
\end{aligned} \tag{9}
$$

These are general formulae for any discontinuous distribution. But if the distribution happens to be an alternative, as the l.d., where x_i takes only two values, any marginal distribution can be completely characterized by two subsets A and A_1 of S where $A \supset A_1$. *Denote by $l_A(A_1)$ the sum of all possible linkage probabilities which contain all points of A_1 and no point of $A - A_1$.* If, e.g. $m = 8$ and A consists of 1,3,5,6 and A_1 of 1,3,6 then $l_A(A_1) = l_{1356}(1,1,0,1) = \sum_{x_2,x_4,x_7,x_8} l(1, x_2, 1, x_4, 1, 0, x_7, x_8)$. According to the previous notation we have as usual

$$l_S(A_1) = l(A_1), \text{ or } l_{1,2,\ldots,m}(x_1, \cdots, x_m) = l(x_1, \cdots, x_m) \tag{10}$$

and

$$l_0(O) = 1, \text{ if } A = O \text{ is empty.}$$

We will use for the linkage distribution and their marginal distributions the customary notations or these new notations, whichever is more convenient.[6]

As an immediate consequence of our definitions we get the following properties of the l.d.

(i) *If* (8) *holds for any* A *than*

$$l_A(A_1) = l_A(A - A_1). \tag{11}$$

(ii) *As a consequence of* (8) *it follows* (*with the notation* (9)) *that*

$$l_i(1) = l_i(0) = \tfrac{1}{2}. \tag{9'}$$

(iii) If c_{ij} denotes the c.p. between i and j, then

$$c_{ij} = c_{ji} = l_{ij}(1,0) + l_{ij}(0,1) = 2l_{ij}(1,0) = 2l_{ij}(0,1). \tag{12}$$

(iv) *For any three subscripts* i, j, k *the "triangular" relation holds*

$$c_{ij} + c_{jk} \geqq c_{ik} \tag{13}$$

and

$$c_{ij} + c_{jk} + c_{ik} \leqq 2. \tag{14}$$

To prove this consider the marginal distribution $l_{ijk}(x_ix_jx_k)$. From (11) and (12) we conclude

$$\begin{aligned} c_{ij} &= 2[l_{ijk}(100) + l_{ijk}(010)] \\ c_{ik} &= 2[l_{ijk}(100) + l_{ijk}(001)] \\ c_{jk} &= 2[l_{ijk}(010) + l_{ijk}(001)] \\ 1 &= 2[l_{ijk}(000) + l_{ijk}(100) + l_{ijk}(010) + l_{ijk}(001)]. \end{aligned}$$

[6] It is easy to indicate experiments which should furnish the relative frequencies corresponding to the l.d.: If a homozygous female $(x_1, \cdots, x_m; x_1, \cdots, x_m)$ is mated to a homozygous male $(y_1, \cdots, y_m; y_1, \cdots, y_m)$ where each $x_i \neq y_i$, the resulting offsprings will all be of type $(x_1, \cdots, x_m; y_1, \cdots, y_m)$. If such an offspring is back crossed to $(y_1, \cdots, y_m; y_1, \cdots, y_m)$ there will be 2^m different genotypes of offsprings, viz. $(x_1, x_2, \cdots, x_m; y_1, y_2, \cdots, y_m)$, $(y_1, x_2, \cdots, x_m; y_1, y_2, \cdots, y_m)$, etc. whose frequencies are proportional to the 2^m values of the l.d., viz. to $l(0, 0, \cdots, 0)$, $l(1, 0, 0, \cdots, 0)$ etc. Such an experiment should give the same results for any two sets of x's and y's. (There is, of course, the statistical problem how to determine the "best" values of the l.p. from these observations.) In an analogous way a marginal distribution can be observed: Suppose we wish for $m = 5$, the $l_{123}(\epsilon_1, \epsilon_2, \epsilon_3)$. The offspring of a cross between females $(x_1, x_2, x_3, x_4, x_5; x_1, x_2, x_3, x_4, x_5)$ and males $(y_1, y_2, y_3, x_4, x_5; y_1, y_2, y_3, x_4, x_5)$ are of type $(x_1, x_2, x_3, x_4, x_5; y_1, y_2, y_3, x_4, x_5)$. If they are crossed to $(x_1, \cdots, x_5; x_1, \cdots, x_5)$ there will be eight different types of offsprings proportional to the eight values of $l_{123}(\epsilon_1, \epsilon_2, \epsilon_3)$. In this last setup the y_i should be dominant and in the experiment, described above, the y_i should be recessive in order to be able to distinguish between the phenotypes of the individuals.

Solving these equations with respect to the l-values we get

$$l_{ijk}(100) = \tfrac{1}{4}(c_{ij} + c_{ik} - c_{jk})$$

(15) $$l_{ijk}(010) = \tfrac{1}{4}(c_{ij} + c_{jk} - c_{ik})$$

$$l_{ijk}(001) = \tfrac{1}{4}(c_{ik} + c_{jk} - c_{ij})$$

(16) $$l_{ijk}(000) = \tfrac{1}{4}(2 - c_{ij} - c_{ik} - c_{jk}).$$

Thence (13) and (14) follow. The condition (14) is of course always fulfilled if $c_{ij} \leqq \frac{1}{2}$, but this restriction does not seem to be necessary. From (15) and (16) we deduce:

(v) *If* $m = 3$, *the set of three c.p.* c_{12}, c_{13}, c_{23} *for which the inequalities* (13), (14) *hold is equivalent to the l.d.* $l(x_1, x_2, x_3)$ *for which* (8) *holds.* For $m \geqq 4$ the c.p. are no longer equivalent to the l.d. Another necessary condition for the c.p. will be derived in section 8.

Now let us consider and characterize *some important particular cases of the l.d.*

(i) *Free assortment* (Mendel). In this case all 2^m values of the l.d are equal and therefore equal to $(\frac{1}{2})^m$.

(ii) *Complete linkage* (reported by Morgan and other authors). In terms of the l.d. this means

(17) $$l(1,1,\cdots,1,1) = l(0,0,\cdots,0,0) = \tfrac{1}{2} \text{ or } l_S(S) = \tfrac{1}{2}.$$

Consequently, all other values of the l.d. are zero. It follows that *all c.p. are zero* because all l_{ij} (1,0) are zero. (See also Theorem I, section 7.)

(iii) *Linkage groups* (Morgan). In terms of the l.d. this means that *the l.d. resolves into a product of several distributions*, e.g.

(18) $$l(x_1, x_2, \cdots, x_g) = f(x_1, x_2)g(x_3, x_4, x_5)h(x_6, x_7, x_8, x_9).$$

(There is no loss of generality in assuming that numerically consecutive characters form a linkage group.) As f, g, and h are distributions it follows with notation (9) that "within" the groups:

$$c_{12} = 2f(10), \quad c_{34} = 2g_{34}(10), \cdots, \quad c_{45} = 2g_{45}(10),$$

$$c_{67} = 2h_{67}(10), \cdots, \quad c_{89} = 2h_{89}(10)$$

these crossover values are quite arbitrary. On the other hand we have because of (9′)

$$f_i(1) = f_i(0) = g_j(1) = g_j(0) = h_k(1) = h_k(0) = \tfrac{1}{2},$$

$$(i = 1, 2;\ j = 1, 2, 3;\ k = 1, \cdots 4)$$

Hence for the c.p. "among" the groups

$$c_{13} = 2 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{2}, \text{ etc. Hence } c_{13} = c_{14} = c_{23} = \cdots = c_{59} = \tfrac{1}{2}$$

in exact accordance with Morgan's idea of linkage groups. If each group consists of only one element: $l(x_1, x_2, \cdots, x_m) = f(x_1)g(x_2) \cdots k(x_m)$ it follows

that $f(x_1) = g(x_2) = \cdots = k(x_m) = \frac{1}{2}$ for $x_i = 0, 1$, hence $l(x_1, \cdots, x_m) = (\frac{1}{2})^m$ for all combinations of the arguments and we have again free assortment.

(*iv*) *Groups of completely linked characters.* Combining and generalizing the ideas of (ii) and (iii) we may speak of i groups of completely linked characters if *within such a group no crossover takes place.* Then the m_i characters in each group act as one character. An example will suffice. Suppose $m = 9$ and three such groups, consisting of the characters 1,2, and 3,4,5, and 6,7,8,9 respectively. Assume that

$$l(11, 111, 1111) = l(00, 000, 0000) = a, \quad l(00, 111, 1111) = l(11, 000, 0000) = C_1$$

$$l(11, 000, 1111) = l(00, 111, 0000) = C_2, \quad l(11, 111, 0000) = l(00, 00, 1111) = C_3$$

where these four numbers are $\neq 0$ and with sum $\frac{1}{2}$; hence all other probabilities are zero. It follows that the c.p. "within" the groups are all zero: $c_{12} = c_{34} = \cdots = c_{45} = c_{67} = \cdots = c_{89} = 0$, but the "among" c.p. are different from zero, e.g. $c_{13} = c_{14} = c_{15} = c_{23} = c_{24} = c_{25} = 2C_1 + 2C_2$ and, with an obvious notation: $c_{I,II} = 2(C_1 + C_2)$, $c_{I,III} = 2(C_1 + C_3)$, $c_{II,III} = 2(C_2 + C_3)$.

A particular case (also a particular case of (iii)) arises if the l.d. resolves into a product of some distributions such that there is complete linkage in each of these. The "within" crossovers are then again zero but all the c.p. "among" the groups equal $\frac{1}{2}$.

5. The case $m = 2$. It will be easier for the reader if this case, though it has been investigated before by several authors [16], [7], [14], will be presented by means of explicit computations before attempting the general one where m and r are arbitrary.

If $m = r = 2$, the number of types $(x_1, x_2; y_1, y_2)$ equals ten. The l.d. is completely determined by the c.p. $c_{12} = c$ and v.v., because $l(10) = l(01) = c/2$, $l(00) = l(11) = (1 - c)/2$. Now let $p^{(n)}(x_1, x_2)$ be the probability that in the nth generation a male (or female) individual *transmits the genes* x_1, x_2; and denote by $p_1^{(n)}(x_1)$ and $p_2^{(n)}(x_2)$ the respective marginal distributions. The formula corresponding to (4) then becomes

$$\begin{aligned} p^{(n)}(1,1) = {} & v^{(n)}(1,1;1,1) + \tfrac{1}{2}v^{(n)}(1,1;1,2) + \tfrac{1}{2}v^{(n)}(1,1;2,1) \\ & + \frac{1-c}{2}v^{(n)}(1,1;2,2) + \frac{c}{2}v^{(n)}(1,2;2,1), \end{aligned} \tag{19}$$

and three analogous formulae. To understand this, consider e.g. the last term of (19); it is the probability that an individual be of type (1,2; 2,1) or (2,1; 1,2) and transmits the set (1,1). By (19) $p^{(n)}(x_1, x_2)$ is deduced from the given distribution $v^{(n)}$ of genotypes.

If, as before, x and y are written for x_1, x_2 and y_1, y_2 it is to be understood that $x = y$ means $x_1 = y_1$ and $x_2 = y_2$. The relation corresponding to (5′) takes then the form

$$\begin{aligned} v^{(n+1)}(x; y) &= p^{(n)}(x)p^{(n)}(x) \quad \text{if} \quad x = y \\ &= 2p^{(n)}(x)p^{(n)}(y) \quad \text{if} \quad x \neq y. \end{aligned} \tag{20}$$

Applying (19) to the $(n+1)$st generation and using (20) we get the recurrence formula

$$p^{(n+1)}(1,1) = [p^{(n)}(1,1)]^2 + p^{(n)}(1,1)p^{(n)}(1,2) + p^{(n)}(1,1)p^{(n)}(2,1) + (1-c)p^{(n)}(1,1)p^{(n)}(2,2) + cp^{(n)}(1,2)p^{(n)}(2,1). \tag{21}$$

Here the right side can be rewritten so as to give

$$p^{(n+1)}(1,1) = (1-c)p^{(n)}(1,1) + cp_1^{(n)}(1)p_2^{(n)}(1) \tag{22}$$

and three analogous formulae. Because of (7):

$$p^{(n+1)}(x_1, x_2) = (1-c)p^{(n)}(x_1, x_2) + cp_1^{(0)}(x_1)p_2^{(0)}(x_2). \tag{22'}$$

From this recurrence formula, which has the particularly simple property that the second term on the right side is independent of n, it is easy to derive step by step:

$$p^{(n)}(x_1, x_2) = (1-c)^n p^{(0)}(x_1, x_2) + [1-(1-c)^n]p_1^{(0)}(x_1)p_2^{(0)}(x_2). \tag{23}$$

Hence, if $c \neq 0$:

$$\lim_{n\to\infty} p^{(n)}(x_1, x_2) = p_1^{(0)}(x_1)p_2^{(0)}(x_2). \tag{24}$$

The preceding results were obtained by Robbins and Jennings. We will formulate a theorem after having studied the general case of arbitrary m and r.[7]

6. The general recurrence formula. Considering random mating and assuming general linkage, we now wish to find the relations which correspond to the formulae (19)–(22) in the case of m r-valued characters. It will turn out, that, by using the l.d., the following proof of the general case becomes surprisingly simple compared with older investigations of the particular case of free assortment, the values of the l.d. acting somehow as natural "separators" for certain groups of terms.

Denote by $w^{(n)}(x_1, \cdots, x_m; y_1, \cdots, y_m) \equiv w^{(n)}(x; y)$ the probability of a genotype whose maternal genes are the x and whose paternal genes the y. Then from (a''):

$$w^{(n)}(x; y) = w^{(n)}(y; x). \tag{25}$$

Writing $x = y$ if and only if $x_i = y_i$, $(i = 1, \cdots, m)$ we put just as in (2)

$$\begin{aligned} v^{(n)}(x; y) &= w^{(n)}(x; x), && \text{if } x = y \\ &= w^{(n)}(x; y) + w^{(n)}(y; x) = 2w^{(n)}(x; y), && \text{if } x \neq y. \end{aligned} \tag{25'}$$

[7] A suggestive remark, repeatedly made by Professor S. Wright states that (assuming random mating) there can be no equilibrium until all of the factors are combined at random. This is indeed a necessary condition for stability.

There are r^{2m} w-values and $\frac{1}{2}r^m(r^m + 1)$ v-values in each generation the respective sums being always equal to one. Denote by $p^{(n)}(x_1, \cdots, x_m)$ *the probability that a male (female) individual of the nth generation transmits the genes* x, and by

$$p_i^{(n)}(x_i),\ p_{ij}^{(n)}(x_i, x_j), \cdots, p_{12\cdots m}^{(n)}(x_1, \cdots, x_m) = p^{(n)}(x_1, \cdots, x_m)$$

the corresponding marginal distributions, defined as usual (see (9)). Sometimes it will be convenient to denote such a marginal distribution by $p_A(z_A) \equiv p_A(z)$ where $A \subset S$, and $p_A(z)$ is the sum of all $p(x)$ such that $x_i = z_i$ for all $i \in A$. Following convention the subscript will be omitted if $A = S$; hence $p_S(z) = p(z)$ and if A is empty, $A = 0$, the corresponding $p_0(z) = 1$.

To simplify the writing $p(x)$, $v(x; y)$, etc. will be written instead of $p^{(n)}(x)$, $v^{(n)}(x; y)$, etc. and $p'(x)$, $v'(x; y)$, etc. for $p^{(n+1)}(x)$, etc. Finally, remember that $l(A)$ is the probability that the paternal genes of A and the maternal genes of $A' = S - A$ will be transmitted and accordingly $l_A(A_1)$ is the (marginal) probability that the paternal genes of A_1 and the maternal genes of $A - A_1$ will be transmitted. $(S \supset A \supset A_1)$.

Let us derive $p'(z)$ from $p(z)$. From the meaning of the different distributions we gather that

$$p(z) = \Sigma l(A)w(x; y) \tag{26}$$

where A *is an arbitrary subset of* S *and* x *and* y *such that*

(a)
$$y_i = z_i \quad \text{for} \quad i \in A$$
$$x_i = z_i \quad \text{“} \quad i \in A'.$$

In fact, the set z will be transmitted if and only if an individual possesses these genes and also transmits them; now consider any $l(A)$ i.e. the probability to transmit the paternal genes of A; this probability is to be multiplied by all possible w-probabilities which contain as arguments the paternal genes of A and the maternal genes of A', as stated in (a). Now let us write (26) also for the $(n + 1)$st generation:

$$p'(z) = \Sigma l(A)w'(x; y). \tag{26'}$$

Next we have, just as always, [see (5), (20)]

$$w'(x; y) = w'(y; x) = p(x)p(y). \tag{27}$$

Hence from (26′) and (27) follows

$$p'(z) = \Sigma l(A)p(x)p(y) \tag{28}$$

with the condition of summation given by (a).

The right side of (28) contains $(2r)^m$ terms. Now we will write it in two different ways by collecting its terms under two different aspects: (i) arranged according to the marginal values of the l-distribution (ii) arranged according to the marginal values of the p-distribution. Let us begin with (i).

The genes $z_1, z_2, \cdots, z_m$ can be transmitted only by individuals which possess each z_i either before or after the semicolon or both; (either from the mother or from the father or from both parents). Hence, if A_1 and A_2 are two disjoint subsets of S, the type of such an individual is such that

$$\text{(b)} \qquad \begin{aligned} &x_i \neq z_i, \quad y_i = z_i \quad \text{for all} \quad i \in A_1 \\ &x_j = z_j, \quad y_j \neq z_j \quad \text{“} \quad \text{“} \quad j \in A_2 \\ &x_k = y_k = z_k \qquad \text{“} \quad \text{“} \quad k \in S - A_1 - A_2 . \end{aligned}$$

Hence the paternal genes of A_1 and the maternal genes of A_2 must be transmitted and for the remaining genes either choice is admissible. Consequently, each $w'(x; y)$ in (26′)—or, what is the same, each $p(x)p(y)$ in (28)—is multiplied by the probability that the paternal genes of A_1 and the maternal genes of A_2 are transmitted. Now writing $A_1 + A_2 = A$ this last probability is exactly the marginal probability $l_A(A_1) = l_A(A_2)$. Thence

$$p'(z) = \Sigma p(x)p(y)l_A(A_1) \tag{29}$$

where the sum is extended over all pairs x, y defined by (b). This is a first recurrence formula. If in (29) $w'(x; y)$ is written instead of $p(x)p(y)$ and then all accents are omitted we get

$$p(z) = \Sigma w(x; y)l_A(A_1) \tag{30}$$

with the summation according to (b). This formula is necessary in order to derive $p(z)$ from the given distribution $w(x; y)$ of genotypes. It corresponds to (19).

Now let us collect the terms of (28) in the second way. Let us determine the factor of any $l(A)$ in (28), e.g. of $l(1,1,0,0,0)$ (where $m = 5$ and A the subset 1,2). Any factor of $l(1,1,0,0,0)$ must be of the form $p(z_1, z_2, \cdot, \cdot, \cdot)$ $p(\cdot, \cdot, z_3, z_4, z_5)$ where all possible values of the variables must be written on the empty places marked by points, and the sum of all these products is to be taken. Now, as in each of the two p's on each of the free places all numbers between 1 and r have to be used, the sum of all these products resolves into the product of the respective sums of the p's. In such a sum each term, on the places belonging to A contains the same fixed values z_A and on the other places any possible value combination; hence such a sum is precisely the marginal probability $p_A(z_A) = p_A(z)$ and the same holds for the other sum of the p's and for $A' = S - A$. Thus we get the second, even more important recurrence formula

$$p'(z) = \sum_{(A)} l(A)p_A(z)p_{A'}(z) \tag{31}$$

where the sum is over all subsets A of S. This formula corresponds to (22) and the limit theorem which will be proved in the next section is an almost immediate consequence of (31). It is worth noticing that the derivations of

(31) and (29) from (28) are completely independent of each other and that only (31) is needed for the limit theorem

From (29) and (31) the interesting identity follows

$$\sum_{(A)} l(A)p_A(z)p_{A'}(z) = \Sigma p(x)p(y)l_A(A_1) \tag{32}$$

which somehow reminds us of a general Abel-transformation.

Let us summarize: (i) From a given distribution of genotypes $w^{(n)}$ (or $v^{(n)}$) the $p^{(n)}$ are derived by (30). (ii) From these $p^{(n)}$ the $w^{(n+1)}$ follow by (27) (or $v^{(n+1)}$ by (25′) and (27)). (iii) Instead of step (ii), from $p^{(n)}$ the consecutive $p^{(n+1)}$, $p^{(n+2)}$, $\cdots$, $p^{(n+\nu)}$ may be derived directly by means of (31). Finally, if desired, $w^{(n+\nu+1)}$ follows by (27).

As an illustration of these formulae let us write (31) for $m = 3,4,5$:

$$\begin{aligned} p'(x_1, x_2, x_3) = 2[&l(000)p(x_1, x_2, x_3) + l(100)p_1^{(0)}(x_1)p_{23}(x_2, x_3) \\ &+ l(010)p_2^{(0)}(x_2)p_{13}(x_1, x_3) \\ &+ l(001)p_3^{(0)}(x_1)p_{12}(x_1, x_2)] \end{aligned} \tag{31′}$$

$$\begin{aligned} p'(x_1, x_2, x_3, x_4) = 2[&l(0000)p(x_1, x_2, x_3, x_4) \\ &+ l(1000)p_1^{(0)}(x_1)p_{234}(x_2, x_3, x_4) + \cdots \\ &+ l(1100)p_{12}(x_1, x_2)p_{34}(x_3, x_4) \\ &+ l(1010)p_{13}(x_1, x_3)p_{24}(x_2, x_4) \\ &+ l(1001)p_{14}(x_1, x_4)p_{23}(x_2, x_3)] \end{aligned} \tag{31″}$$

$$\begin{aligned} p'(x_1, x_2, x_3, x_4, x_5) = 2[&l(00000)p(x_1, \cdots, x_5) \\ &+ l(10000)p_1^{(0)}(x_1)p_{2345}(x_2, x_3, x_4, x_5) + \cdots \\ &+ l(11000)p_{12}(x_1, x_2)p_{345}(x_3, x_4, x_5) + \cdots]. \end{aligned} \tag{31‴}$$

In the last formula the last group contains ten terms. As an illustration of (30) we write e.g. for $m = 3$, $r = 2$, with $p^{(n)} = p$ and $v^{(n)} = v$:

$$\begin{aligned} p(x_1, x_2, x_3) \equiv v(x_1, x_2, x_3; x_1, x_2, x_3) &+ \tfrac{1}{2}[v(x_1, x_2, x_3; y_1, x_2, x_3) \\ &+ v(x_1, x_2, x_3; x_1, y_2, x_3) + \cdots] \\ &+ [l_{12}(00)v(x_1, x_2, x_3; y_1, y_2, x_3) \\ &+ l_{13}(00)v(x_1, x_2, x_3; y_1, x_2, y_3) + \cdots] \\ &+ l(000)v(x_1, x_2, x_3; y_1, y_2, y_3) \\ &+ [l(100)v(y_1, x_2, x_3; x_1, y_2, y_3) \\ &+ l(010)v(x_1, y_2, x_3; y_1, x_2, y_3) + \cdots]. \end{aligned} \tag{30′}$$

7. Limit theorems. In order to find $\lim_{n\to\infty} p^{(n)}(x_1, \cdots, x_m)$ we write the recurrence formula (31) in the form

$$p^{(n+1)}(x) - 2l(00\cdots0)p^{(n)}(x) = \sum_{(A)}' l(A)p_A^{(n)}(z)p_{A'}^{(n)}(z). \tag{33}$$

Here $\sum_{(A)}'$ means a sum over all subsets A of S which are neither void nor equal to S. If we write $q_m^{(n)}$ for the right side of (33) and $p^{(n)}(x) = p_m^{(n)}$, $2l(0, \cdots, 0) = \alpha_m$ the last equation takes the form

$$p_m^{(n+1)} - \alpha_m p_m^{(n)} = q_m^{(n)}. \tag{34}$$

Consider first the case $\alpha_m = 1$, or $l(0, \cdots, 0) = l(1, \cdots, 1) = \frac{1}{2}$, i.e. *complete linkage*, as defined in section 3. In this case all $l(A)$-values on the right side of (33) are zero, hence $q_m^{(n)} = 0$ and

$$p_m^{(n+1)} = p_m^{(n)} \qquad (n = 0, 1, 2, \cdots). \tag{35}$$

This is exactly the same result as (7): All $p_m^{(n)}$ are equal to $p_m^{(0)}$ and because of (27) also

$$w^{(n)}(x; y) = w^{(1)}(x; y) \quad \text{or} \quad v^{(n)}(x; y) = v^{(1)}(x; y) \quad (n = 1, 2, \cdots). \tag{36}$$

In fact, if the characters are completely linked, they act as one character. Hence we have

THEOREM I. *If the m Mendelian characters are completely linked, the distribution of genotypes reaches the stationary state in the first filial generation.*

Now consider (34) in the general case where $0 \leqq \alpha_m < 1$. Then the following *lemma* will be used: *If in a recurrence formula of the form* (34), $|\alpha_m| < 1$ *and* $\lim_{n\to\infty} q_m^{(n)} = q_m$ *exists, then* $\lim_{n\to\infty} p_m^{(n)} = p_m = q_m/(1 - \alpha_m)$. This can be proved directly in various simple ways. It may also be regarded as a consequence of well-known general convergence theorems. See also [15].

In order to apply the lemma let us first notice that q_2 exists. In fact, $p^{(n+1)}(x_1, x_2) - 2l(00)p^{(n)}(x_1, x_2) = 2l(01)p_1^{(0)}(x_1)p_2^{(0)}(x_2)$ and as the right side is independent of n, q_2 certainly exists. Hence, it follows from the lemma that p_2 exists. For $m = 3$ the recurrence formula (31′) shows that $q_3^{(n)}$ contains no marginal distribution of p of an order higher than two; therefore each of the terms of $q_3^{(n)}$ approaches a limit, hence $q_3 = \lim_{n\to\infty} q_3^{(n)}$ exists, and consequently, because of the lemma, p_3 exists. We may continue in this way because in (33) all marginal distributions of p on the right side are of an order $\leqq m - 1$. Hence for every m the $q_m^{(n)}$ approaches a limit and consequently the $\lim_{n\to\infty} p_m^{(n)}$ exists. Finally, in order to find p_m we notice that $q_2 = (1 - \alpha_2)p_1^{(0)}(x_1)p_2^{(0)}(x_2)$, hence $p_2 = p_1^{(0)}(x_1)p_2^{(0)}(x_2)$. Then, assuming that $p_{m-1} = p_1^{(0)}(x_1) \cdots p_{m-1}^{(0)}(x_{m-1})$ we see from (33), using (8), that $q_m = (1 - \alpha_m)p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)$. (See also (31′) (31″), (31‴).) Thence

$$\lim_{n\to\infty} p^{(n)}(x_1, x_2, \cdots, x_m) = p_1^{(0)}(x_1)p_2^{(0)}(x_2) \cdots p_m^{(0)}(x_m). \tag{37}$$

The last formula contains the limit theorem we wished to prove. It can be stated as follows:

THEOREM II. *If m characters are arbitrarily linked, with the one exception of "complete linkage", the distribution of transmitted genes* $p^{(n)}(x_1, \cdots, x_m)$ *"converges towards independence." The limit distribution is the product of the m marginal distributions of the first order* $p_i^{(0)}(x_i)$, *which are derived from* $p^{(0)}(x_1, \cdots, x_m)$, *the distribution of gametes in the initial generation.*

If, however, the initial distribution $p^{(0)}(x_1, \cdots, x_m)$ shows particular features, the stationary state may be reached already for a finite value of n. This happens with $n = 0$ and for every l.d. if $p^{(0)}(x_1, \cdots, x_m) = p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)$. In other particular cases it may happen under particular assumptions for the l.d.

Let us express the general result also in terms of the distribution of genotypes. It follows from (37) and (27) that

$$\lim_{n\to\infty} w^{(n+1)}(x; y) = \lim_{n\to\infty} p^{(n)}(x)p^{(n)}(y)$$

$$= p_1^{(0)}(x_1) \cdots p_m^{(0)}(x_m)p_1^{(0)}(y_1) \cdots p_m^{(0)}(y_m) = \prod_{i=1}^{m} [p_i^{(0)}(x_i)p_i^{(0)}(y_i)].$$

Now consider a product like $p_i^{(0)}(x_i)p_i^{(0)}(y_i)$. By definition of $p_1^{(0)}(x_1)$ and applying (27) we find

$$p_1^{(0)}(x_1)p_1^{(0)}(y_1) = \sum_{x_2} \cdots \sum_{x_m} p^{(0)}(x_1, \cdots, x_m) \sum_{y_2} \cdots \sum_{y_m} p^{(0)}(y_1, \cdots, y_m)$$

$$= \sum_{x_2,\cdots,x_m} \sum_{y_2,\cdots,y_m} p^{(0)}(x)p^{(0)}(y)$$

Introducing then in a natural way the marginal distribution:

$$w_i^{(n)}(x_i ; y_i)$$

$$(38) \qquad = \sum_{x_1,\cdots,x_{i-1}x_{i+1},\cdots,x_m} \sum_{y_1,\cdots,y_{i-1}y_{i+1},\cdots,y_m} w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m)$$

it is seen that

$$(39) \qquad p_i^{(0)}(x_i)p_i^{(0)}(y_i) = w_i^{(1)}(x_i ; y_i).$$

Thence the result

$$(40) \qquad \lim_{n\to\infty} w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m) = w_1^{(1)}(x_1 ; y_1) \cdots w_m^{(1)}(x_m ; y_m)$$

which may be stated as follows:

THEOREM III. *In case of m arbitrarily linked Mendelian characters the distribution of the genotypes in the nth generation,* $w^{(n)}(x_1, \cdots, x_m ; y_1, \cdots, y_m)$, *"approaches independence" as* $n \to \infty$. *The limit distribution is the product of the m marginal distributions* $w_i^{(1)}(x_i ; y_i)$ *of the ith character* $(i = 1, \cdots, m)$ *in the first filial generation.*

This theorem, which may be regarded as a corollary to THEOREM II, holds for any type of linkage, except "complete linkage" as defined in (17) where (36) is valid.

8. Solution of the recurrence equations (31). Formula (31) expresses $p^{(n)}(x_1, \cdots, x_m)$ in terms of $p^{(n-1)}(x_1, \cdots, x_m)$ (and all marginal distributions of $p^{(n-1)}$) and of the l.d. It seems desirable to try to express $p^{(n)}(x)$ in terms of $p^{(0)}(x)$. Now (31) is not a single equation but rather a complex system of difference equations with constant coefficients because for each marginal distribution of order $i < m$ the respective recurrence formula (31) of order i has to be used. (Or, if it is preferred to consider the marginal distributions as sums of p-values of order m, then all these p-values appear simultaneously and there is again a complicated system of difference equations.) In this situation it is not to be expected that the integration will yield simple explicit formulae, particularly as long as the l.d. is left arbitrary. However, the construction of the following formulae is clear. They reduce to simpler expressions in particular cases.

Let us use a method of indeterminate coefficients. To simplify the writing denote $p^{(0)}(x_1, \cdots, x_m)$ and its marginal distributions $p_i^{(0)}(x_i)$, $p_{ij}^{(0)}(x_i, x_j)$, etc. by $p_{12,\ldots,m}, p_i, p_{ij}$, etc. From genetical as well as mathematical considerations we gather the general form of $p_{12}^{(n)\ldots}{}_m$ in terms of $p_{12\ldots m}$ and its marginal distributions; that this is indeed the general form will be verified by our very computations. Consider the set S consisting of the m numbers $1, 2, \cdots m$ and divide S in every possible way in two disjoint parts A_1 and A_2, none of them being empty, so that $A_1 + A_2 = S$, then divide S in every possible way into three disjoint parts so that $A_1 + A_2 + A_3 = S$, and finally S is divided into m disjoint parts each consisting of one single element. Denoting the unknown coefficients in a corresponding way by $\alpha_S^{(n)}$, $\alpha_{A_1,A_2}^{(n)}$, $\alpha_{A_1,A_2,A_3}^{(n)}$, etc. and writing $p_S^{(n)}$ and p_S for $p_{12}^{(n)\ldots}{}_m$ and $p_{12}^{(0)\ldots}{}_m$ the general form of p_S will be

$$p_S^{(n)} = \alpha_S^{(n)} p_S + \sum_{(A_1)} \alpha_{A_1,A_2}^{(n)} p_{A_1} p_{A_2} + \sum_{(A_1,A_2)} \alpha_{A_1,A_2,A_3}^{(n)} p_{A_1} p_{A_2} p_{A_3} + \cdots + \alpha_{1,2,3,\ldots,m}^{(n)} p_1 p_2 p_3 \cdots p_m . \tag{41}$$

This holds for every m. We get e.g. for $m = 4$

$$p_{1234}^{(n)} = \alpha_{1234}^{(n)} p_{1234} + (\alpha_{1,234}^{(n)} p_1 p_{234} + \alpha_{2,134}^{(n)} p_2 p_{134} + \cdots) + (\alpha_{12,34}^{(n)} p_{12} p_{34} + \cdots) + (\alpha_{12,3,4}^{(n)} p_{12} p_3 p_4 + \cdots) + \alpha_{1,2,3,4}^{(n)} p_1 p_2 p_3 p_4 . \tag{41'}$$

For $m = 6$, e.g., there are eleven different types: One term $\alpha_{1\ldots6}^{(n)} p_{1\ldots6}$; then 6 terms of the form $\alpha_{1,2\ldots6}^{(n)} p_1 p_{23\ldots6}$; 15 terms like $\alpha_{12,3456}^{(n)} p_{12} p_{3456}$; 10 terms like $\alpha_{123,456}^{(n)} p_{123} p_{456}$; 15 terms like $\alpha_{1,2,3456}^{(n)} p_1 p_2 p_{3456}$; 60 terms like $\alpha_{1,23,456}^{(n)} p_1 p_{23} p_{456}$; 15 terms like $\alpha_{12,34,56}^{(n)} p_{12} p_{34} p_{56}$; 20 terms as $\alpha_{1,2,3,456}^{(n)} p_1 p_2 p_3 p_{456}$; 15 terms as $\alpha_{12,34,5,6}^{(n)} p_{12} p_{34} p_5 p_6$; 15 terms as $\alpha_{1,2,3,4,56}^{(n)} p_1 p_2 p_3 p_4 p_{56}$; and one final term $\alpha_{1,2,3,4,5,6}^{(n)} p_1 p_2 p_3 p_4 p_5 p_6$.

In (41) the $\alpha_{\ldots}^{(n)}$ are unknown constants depending on n and on the l.d. In order to find them consider (31) and write for the values of the l.d. v_A^m instead of $2l(A)$ (no confusion is possible because no marginal distribution of the l.d. occurs in (31)). With this notation (31″) e.g. reads:

$$p_{1234}^{(n+1)} = v_0^4 p_{1234}^{(n)} + (v_1^4 p_1 p_{234}^{(n)} + \cdots) + (v_{12}^4 p_{12}^{(n)} p_{34}^{(n)} + \cdots). \tag{31″}$$

If there is no ambiguity the upper m in v_A^m may even be omitted. Now assume the equations (41) to be written for $\mu = 2$, $\mu = 3$, $\cdots$, $\mu = m$. Introduce into the left side of (31) the expression (41) for $p_S^{(n+1)}$ and in the same way replace on the right side of (31) all $p_i^{(n)}$, $p_{ij}^{(n)}$, $\cdots$, $p_{12\cdots m}^{(n)}$ by their respective expressions (41). In this way an equality is obtained from which recurrence formulae for the unknown coefficients may be deduced by collecting all groups of terms which contain the same products of p's.

If this is carried out, e.g. for $m = 4$, the recurrence formulae are

$$
\begin{aligned}
\alpha_{1234}^{(n+1)} &= v_0\alpha_{1234}^{(n)} \\
\alpha_{123,4}^{(n+1)} &= v_0\alpha_{123,4}^{(n)} + v_4\alpha_{123}^{(n)} \\
\alpha_{12,34}^{(n+1)} &= v_0\alpha_{12,34}^{(n)} + v_{12}\alpha_{12}^{(n)}\,\alpha_{34}^{(n)} \qquad \text{etc.} \\
\alpha_{12,3,4}^{(n+1)} &= v_0\alpha_{12,3,4}^{(n)} + v_{12}\alpha_{12}^{(n)}\alpha_{3,4}^{(n)} + v_3\alpha_{12,4}^{(n)} + v_4\alpha_{12,3}^{(n)} \\
\alpha_{1,2,3,4}^{(n+1)} &= v_0\alpha_{1,2,3,4}^{(n)} + v_1\alpha_{2,3,4}^{(n)} + \cdots + v_{12}\alpha_{1,2}^{(n)}\alpha_{3,4}^{(n)} + \cdots
\end{aligned} \tag{42}
$$

In general, i.e. for any m, these recurrence formulae are of a clear structure the first one being particularly simple, namely

$$\alpha_S^{(n+1)} = v_0\alpha_S^{(n)}. \tag{43}$$

It can be solved immediately and gives

$$\alpha_S^{(n)} = v_0^n\,. \tag{43'}$$

The other recurrence formulae are all of the form

$$x_{n+1} = v_0x_n + f(n) \text{ with } x_0 = 0, \tag{44}$$

where $f(n)$ is a given function of n whose general form is still to be investigated. The solution of (44) is

$$x_n \equiv \sum_{\nu=0}^{n-1} f(\nu)v_0^{n-1-\nu}. \tag{44'}$$

With the notations used in (41) the equation (44) may be written:

$$\alpha_{A_1,A_2,\cdots,A_\mu}^{(n+1)} = v_0\alpha_{A_1,A_2,\cdots,A_\mu}^{(n)} + A_{A_1,A_2,\cdots,A_\mu}^{(n)}\,. \tag{44''}$$

We have to determine $A_{A_1,A_2,\cdots,A_\mu}$. For reasons of symmetry and homogeneity let us introduce constants $\alpha_1^{(n)} = \alpha_2^{(n)} = \cdots = \alpha_m^{(n)} = 1$. With that notation e.g. the last term in the second line in (42) reads $v_4\alpha_{123}^{(n)}\alpha_4^{(n)}$ or the third term to the right in the fourth line of (42): $v_3\alpha_{12,4}^{(n)}\alpha_3^{(n)}$ etc.

The construction of $A_{A_1,A_2,\cdots,A_\mu}^n$ may then be described as follows: Each $A_{A_1,A_2,\cdots,A_\mu}^n$ is a sum of $2^{\mu-1} - 1$ terms, each term being a product of one v-value and two α's. The set consisting of the μ elements A_1, A_2, $\cdots$, A_μ is to be divided in all possible ways into two non-empty, disjoint, complementary parts which form the subscripts of the two α's in question; the subscript of v is equal to the subscript of either of these two α-values; it makes no difference which,

because of the specific symmetry (8′) of the l.d.; it should be noted that in the subscripts of v no comma occurs. As an example let us write $A^{(n)}_{1234,567,8}$ for $m = 8$. We get: $A^{(n)}_{1234,567,8} = v_8\alpha^{(n)}_8\alpha^{(n)}_{1234,567} + v_{567}\alpha^{(n)}_{567}\alpha^{(n)}_{1234,8} + v_{1234}\alpha^{(n)}_{1234}\alpha^{(n)}_{567,8}$. Or if we wish $A^{(n)}_{12,34,5,6}$ for $m = 6$: $A^{(n)}_{12,34,5,6} = v_6\alpha^{(n)}_6\alpha^{(n)}_{12,34,5} + v_5\alpha^{(n)}_5\alpha^{(n)}_{12,34,6} + v_{12}\alpha^{(n)}_{12}\alpha^{(n)}_{34,5,6} + v_{34}\alpha^{(n)}_{34}\alpha^{(n)}_{12,5,6} + v_{56}\alpha^{(n)}_{5,6}\alpha^{(n)}_{12,34} + v_{125}\alpha^{(n)}_{12,5}\alpha^{(n)}_{34,6} + v_{126}\alpha^{(n)}_{12,6}\alpha^{(n)}_{34,5}$.

Hence, in principle our "integration" problem, where n is the variable, is completely solved: First $p^{(n)}_s$ is given by (41). Then, in order to find any $\alpha^{(n)}_{A_1,A_2,\ldots A_\mu}$, we first determine the corresponding $A_{A_1,A_2,\ldots A_\mu}$ by the rule just explained and illustrated, and then it follows from (44′) that

$$\alpha^{(n)}_{A_1,A_2,\cdots,A_\mu} = \sum_{\nu=0}^{n-1} v_0^{n-1-\nu} A^{(\nu)}_{A_1,A_2,\cdots,A_\mu}. \tag{44‴}$$

This whole procedure, although in principle very simple, may of course be lengthy if m is not small and if no specific assumption for the l.d. is considered; for in the expression of $A_{A_1,A_2,\ldots,A_\mu}$ many different α-values appear,—each however with less than m subscripts—which play the role of abbreviations for complicated expressions; in other words the explicit solution for $m = 6$, for instance requires the solutions for $m < 6$, all these solutions being however completely given by our formulae, down to $m = 2$, where $\alpha^{(n)}_{12}$ and $\alpha^{(n)}_{1,2}$ are given by (23).

Under simple assumptions for the l.d. the explicit expressions for the α become simple. Two extreme cases are complete linkage and free assortment. In the first case $p^{(n)}_{12\ldots m} = p^{(0)}_{12\ldots m}$ and nothing remains to be done. The case of free assortment where all $v = (\frac{1}{2})^{m-1}$ can be dealt with directly by induction, or we may evaluate the general formulae given above which in this case become quite simple. We have[8]

$$2^{mn}\alpha^{(n)}_{A_1,A_2,\ldots,A_\mu} = 2^n(2^n - 1)\cdots(2^n - \mu + 1). \tag{45}$$

That shows that the values of the coefficients $\alpha^{(n)}_{\ldots}$ depend only on the number of elements A_i which appear as subscripts. Thus we find e.g. for $m = 6$, if we write in each line of (45′) one typical value:

$$\begin{aligned}
&\alpha^{(n)}_{123456} = 2^n/2^{6n} = 1/2^{5n}\\
&\alpha^{(n)}_{1,23456} = \alpha^{(n)}_{12,3456} = \alpha^{(n)}_{123,456} = (2^n - 1)/2^{5n}\\
&\alpha^{(n)}_{1,2,3456} = \alpha^{(n)}_{1,23,456} = \alpha^{(n)}_{12,34,56} = (2^n - 1)(2^n - 2)/2^{5n}\\
&\alpha^{(n)}_{1,2,3,456} = \alpha^{(n)}_{12,34,5,6} = (2^n - 1)(2^n - 2)(2^n - 3)/2^{5n}\\
&\alpha^{(n)}_{1,2,3,4,56} = (2^n - 1)(2^n - 2)\cdots(2^n - 4)/2^{5n}\\
&\alpha^{(n)}_{1,2,3,4,5,6} = (2^n - 1)(2^n - 2)\cdots(2^n - 5)/2^{5n}.
\end{aligned} \tag{45′}$$

Thus in the simple case of independent assortment the explicit solution is very simple too. It confirms the fact that $\lim_{n\to\infty} \alpha^{(n)}_{1,2,3,4,5,6} = 1$ while all other α's approach

[8] The values on the right side of (45) are indicated in [1]; but the solution for free assortment reported in this article does not seem to coincide with ours.

zero. To prove this, however, without recurring to computations, was the purpose of the preceding section.

9. Crossover distribution and crossover probabilities. The limit theorem of §7 as well as the computations of the preceding section, in short, all investigations and concepts considered so far, are valid for any l.d. We shall now define and use a *crossover distribution*, (c.d.), which is completely equivalent to the l.d. but preferable for the study of certain particular cases. Apparently biologists have not considered the general concept of the c.d. but only the c.p. c_{ij}. This concept is basic and tangible but not sufficient for a complete description of the linkage mechanism when $m \geqq 4$, as was seen in the preceding sections.

It is obvious that, from our point of view, a mathematical theory of linkage must be based on the properties of and a set of assumptions on the l.d., or the c.d. The *linear theory* will be considered from this standpoint. This theory is, of course, still compatible with a variety of particular assumptions. In the last section some simple particular cases will be presented and studied with a special view to *interference*.

The probability that an individual transmits the set of "paternal genes" belonging to A and the set of "maternal genes" belonging to $A' = S - A$ is denoted by $l(A)$, where $l(A) = l(A')$; e.g. with $m = 8$: $l(1,0,1,1,1,0,0,1) = l(0,1,0,0,0,1,1,0)$. Considering here the succession of arguments we see that in either set of eight arguments: The first and the second are from different sets, the second and the third are again from different sets, the third and the fourth are from the same set, $\cdots$ the seventh and eighth are from different sets. Writing 0 for "same" and 1 for "different" and using these numbers to correspond to the $(m - 1)$ consecutive intervals between the m genes, we introduce:

$$l(10111001) + l(01000110) = \pi(1100101).$$

Here $\pi(\eta_1, \eta_2, \cdots, \eta_{m-1})$ where $\eta_i = 0$ or 1, is an $(m - 1)$-variate alternative. The relation between the l.d. and this new distribution may be written in the form

$$(46) \quad 2l(\epsilon_1, \epsilon_2, \cdots, \epsilon_m) = \pi(|\epsilon_1 - \epsilon_2|, |\epsilon_2 - \epsilon_3|, \cdots |\epsilon_{m-1} - \epsilon_m|), \quad \epsilon_i = 0 \text{ or } 1.$$

In this definition no fixed "order" of the genes is implied so far. The numbers $1,2 \cdots m$ are used like names.

But it seems to be admitted today by leading biologists that a certain natural order of the genes exists. If this is so the numbers $1,2, \cdots m$ should be used in agreement with this order. Let us note, however, that the situation is in reality slightly different: Only the genes *within each linkage group* (§4) are assumed to be ordered, whereas no order exists among the groups. Let us for the moment disregard this circumstance and assume that all genes under consideration belong to the same linkage group.

Within such a linkage group a one-dimensional or linear order prevails, to be understood in the geometric sense of "location". Some more precise definitions

concerning this linear order will be considered later. For the moment we simply imagine that each of the two sets of genes belonging to an individual is arranged like m consecutive discrete points on a line segment.[9] The crossover distribution $\pi(\eta_1, \eta_2, \cdots \eta_{m-1})$, introduced in (46) becomes more meaningful under this assumption where, now, the numbering corresponds to this linear order. Then the argument 0 in this distribution can be interpreted as "coherence" and the argument 1 as "interchange" or "crossing over" and the "intervals" as intervals in the geometric sense. Whether this "crossing over", which means transition from the maternal to the paternal set or vice versa, is to be conceived as a "break" (Janssen's chiasmatypie) does not matter for the above definitions. If however, the idea is that between two neighboring genes not more than one break is possible then the "event," which we call crossover, would be at the same time a break; if, biologically, more than one break between i and $(i + 1)$ is not excluded, then the event "crossover within $(i, i + 1)$" means "odd number of breaks within this interval."

Now, let us consider the relation between the c.d. and the c.p. It has been repeatedly remarked that the c.p. are not equivalent to the l.d., hence they are not equivalent to the c.d. either. There are $\frac{1}{2} \cdot m(m - 1)$ c.p. but $2^{m-1} - 1$ l-values, or π-values. If $m \geqq 4$ the second number is greater than the first. Besides, the l-values are absolutely arbitrary probabilities. For the c.p. in section 4 some restrictions were derived. Let us derive another *set of restrictions* by considering four numbers i, j, k, l which we may denote by 1, 2, 3, 4. (The following computation has nothing to do with linear order. It applies *if* $m = 4$ to the l.d. $l(\epsilon_1\epsilon_2\epsilon_3\epsilon_4)$ and if $m > 4$ to the respective four-dimensional marginal distributions of the l.d.) Write $v(\epsilon_1\epsilon_2\epsilon_3\epsilon_4) = 2l(\epsilon_1\epsilon_2\epsilon_3\epsilon_4)$ and let us add up the six c.p. corresponding to these four numbers. From $c_{ij} = 2l_{ij}(1,0) = v_{ij}(1,0)$ we get

$$\begin{aligned} c_{12} + c_{13} + \cdots + c_{34} &= 3v(1000) + 3v(0100) + 3v(0010) + 3v(0001) \\ &\quad + 4v(1001) + 4v(1010) + 4v(1100) \\ &= 4 - 4v(0000) - v(1000) - v(0100) - v(0010) - v(0001) \leqq 4. \end{aligned} \tag{47}$$

Hence as by (14) $c_{12} + c_{23} + c_{13} \leqq 2$, it follows that

$$c_{il} + c_{jl} + c_{kl} \leqq 2 \tag{14'}$$

is another necessary condition for the c.p. The limit "2" can be reached, as we see for $v(0000) = v(1000) = v(0100) = v(0010) = v(0001) = 0$; then

$$\begin{aligned} c_{12} &= c_{34} = v(1001) + v(1010) \\ c_{23} &= c_{14} = v(1100) + v(1010) \\ c_{13} &= c_{24} = v(1001) + v(1100) \end{aligned}$$

[9] "The genes are represented as lying in a line like beads on a string. The numerical data from crossing over show in fact that this arrangement is the only one that is consistent with the results obtained" [11]. This is but one of many statements in favor of the linear theory.

and

$$c_{12} + c_{23} + c_{13} = c_{14} + c_{24} + c_{34} = 2.$$

To summarize the facts about the c.p.: *In case of m characters there are* $\frac{1}{2}m(m-1)$ c.p. $c_{ij} = 2l_{ij}(10) = 2l_{ij}(0,1)$. *These values must satisfy the following necessary conditions* (besides $0 \leqq c_{ij} \leqq 1$):

(13) $c_{ij} + c_{jk} \geqq c_{ik}$ for any three subscripts

(14) $c_{ij} + c_{jk} + c_{ik} \leqq 2$ " " " "

(14′) $c_{il} + c_{jl} + c_{kl} \leqq 2$ " " four "

If in an analogous way five or more subscripts are considered no new condition turns up. It has, however, not been proved that the above given necessary conditions are sufficient for a consistent system of c.p. If we wish to be sure of consistency the starting point must be a l.d. or a c.d. from which the c_{ij} are deduced.

[This question of consistency belongs in the same class as the following problem: "Under what conditions does a set of $\binom{m}{2}$ distributions $V_{ij}(x_i, x_j)$ form the marginal distributions of second order of an m-dimensional distribution $V(x_1, \cdots, x_m)$?" Here $V(x_1, \cdots, x_m)$ is the probability that the first result is $\leqq x_1$, the second $\leqq x_2$, the last $\leqq x_m$. An analogous question arises for the set of $\binom{m}{3}$ distributions $V_{ijk}(x_i, x_j, x_k)$, etc.[10]]

In the following it will be necessary to know *the expressions of the c.p. in terms of the c.d.* Put $m - 1 = n$ and denote by p_i, p_{ij}, etc. in the usual way the following probabilities derived from the c.d.: p_i is the probability of "success" in the i-th trial, p_{ij} the probability of success in both the i-th and j-th trial, etc. It has to be kept in mind that for the c.d. and all magnitudes derived from it the "i-th trial" is associated with the i-th interval, i.e. with the interval $(i, i+1)$ "and success in the i-th trial" means cross over in this interval. [Whereas in the l.d. and in magnitudes derived from it, like $c_{ij} \equiv l_{ij}(1, 0)$ the subscript i denotes the i-th gene. (See (46)).] Now denote by S_1 the sum of all probabilities p_i, by S_2 the sum of all p_{ij}, $\cdots$. Besides, let $P_{1,\ldots,i}(x)$ be the probability of exactly x successes in the first i trials ($i = 1, 2, \cdots, n$), and analogously, $P_{2\ldots j}(x)$ the probability of x successes in the $j-1$ trials 2, 3, $\cdots$, j, etc. Then the desired formulae follow easily: First we have obviously

$$c_{i,i+1} = p_i \qquad (i = 1, 2, \cdots, n).$$

Because $c_{i,i+1}$ is the probability of one interchange between the genes i and $i+1$ i.e. of an interchange in the i-th interval, of "success in the i-th trial".

[10] For one-dimensional distributions $V_i(x)$ the question is trivial because any set of m distributions $V_i(x)$ can be considered as the marginal distributions of first order of $V(x_1, \cdots, x_m) = V_1(x_1) \cdots V_m(x_m)$.

Then $c_{i,i+2}$ is the probability of one interchange between i and $i+2$, i.e. of either an interchange in the first of the two intervals numbered i and $i+1$, and no interchange in the second; or of an interchange in the second but none in the first. Hence $c_{i,i+2} = P_{i,i+1}(1)$, $(i = 1, \cdots, n-1)$, because $P_{i,i+1}(1)$ is just the probability of exactly one "success" in the two trials numbered i and $i+1$. In the same way we get $c_{i,i+3} = P_{i,\ldots i+2}(1) + P_{i,\ldots i+2}(3)$, $(i = 1, \cdots, n-2)$, because an interchange between i and $i+3$ means either exactly one or exactly three interchanges in the three intermediate intervals. Hence we get altogether, with $n = m - 1$:

$$
\begin{aligned}
c_{i,i+1} &= p_i \\
c_{i,i+2} &= P_{i,i+2}(1) \\
c_{1m} &= P_{12\ldots n}(1) + P_{12\ldots n}(3) + \cdots P_{12\ldots n}(\bar{n}), \text{ where } \bar{n} = n \text{ if } n \text{ odd}, \\
&\qquad = n - 1 \text{ if } n \text{ even}.
\end{aligned} \tag{48}
$$

Let us also express the c_{ij} in terms of the S_i. It is well known (see e.g. [3]) that

$$
P_{1,\ldots,n}(x) = \sum_{\nu=x}^{n} (-1)^{\nu+x} S_\nu \qquad (x = 0, 1, \cdots n). \tag{49}
$$

Applying these to (48) we easily find the convenient expressions:

$$
\begin{aligned}
c_{12} &= p_1 \text{, etc.} \\
c_{13} &= (p_1 + p_2) - 2p_{12} \equiv (S_1 - 2S_2)_{12} \text{, etc.} \\
c_{14} &= (p_1 + p_2 + p_3) - 2(p_{12} + p_{13} + p_{23}) + 4p_{123} \\
&\qquad \equiv (S_1 - 2S_2 + 4S_3)_{123} \text{, etc.} \\
c_{15} &= (S_1 - 2S_2 + 4S_3 - 8S_4)_{1\ldots4} \text{, etc.} \\
&\cdots\cdot \\
c_{1,m} &= S_1 - 2S_2 + 4S_3 + \cdots + (-2)^m S_{m-1} .
\end{aligned} \tag{50}
$$

10. The linear theory. Consider a linkage group of size m and assume for the moment that $c_{ij} \neq c_{ik}$ for all i, j, and k. It seems that the main mathematical content of the linear theory can be summarized as follows: *It is possible to establish in a unique way an order or a succession of the genes, such that for the* $\binom{m}{2} = \frac{m(m-1)}{2}$ *c.p. the* $(m-1)(m-2)$ *inequalities*

$$
\begin{aligned}
c_{ij} &< c_{i,j+1} \qquad (i = 1, 2, \cdots, m-2) \\
c_{ij} &< c_{i-1,j} \qquad (i = 2, 3, \cdots, m-1)
\end{aligned} \qquad (i < j) \tag{51}
$$

hold. In this succession j will be between i and k if c_{ik} is greater than the two other c.p. c_{ij} and c_{jk}. The two arrangements $1, 2, \cdots m$ and $m, m-1, \cdots,$

$\cdots$ 1 are considered as corresponding to the same order. *Furthermore, this order is a straight-line-succession for which an additive distance relation holds* (cf. also [4a] and [13]). Instead of the restriction $c_{ij} \neq c_{ik}$ it is sufficient to assume *the weaker restriction, that in any triple* c_{ij}, c_{ik}, c_{kj} *one is greater than the two others.* Without such a restriction uniqueness of the order no longer holds. E.g. in case of independent assortment where all c_{ij} equal $\frac{1}{2}$ any of the m: possible numberings of the genes is equally admissible from the point of view of the linear theory. In the case of complete linkage where all c.p. are zero it will be logical to consider all m genes as located in the same point. Obviously there are all kinds of intermediate cases. We shall come back to this point at the end of this section.

Now consider again the case of "*different*" c.p. (in the above defined sense). Let us prove that there can be *not more than one succession* for which (51) holds. In fact it follows from (51) that also:

$$(51') \qquad \begin{matrix} c_{ij} < c_{ik} & (i = 1, 2, \cdots, m-2) & \text{for all} & k > j \\ \text{and} \quad c_{ij} < c_{kj} & (i = 2, 3, \cdots, m-1) & \text{for all} & k < i. \end{matrix} \qquad i < j$$

These are all together $M = 2\cdot 1 + 3\cdot 2 + \cdots + (m-1)(m-2) = 2\cdot\binom{m}{3}$ inequalities. On the other hand there are all together $\binom{m}{3} = M/2$ "between"-relations for m numbers, each of them being defined by two inequalities as $c_{ij} < c_{ik}$ and $c_{jk} < c_{ik}$ (if j is between i and k); hence on the whole M such inequalities. But these are the same as (51'), as we see by changing i, j, k into j, k, i in the second equation (51'). Thus it is not possible to find two different successions which both satisfy (51).

As to the *metric* of the problem, Morgan proposed originally that the value of the c.p. c_{ij} should be used as the distance between i and j. It has, however been objected repeatedly that this distance would not be additive; this is obvious since the triangular relation (13) holds for three subscripts (see also (50).[11] The equality $c_{ij} + c_{jk} = c_{ik}$ holds only in the exceptional cases where multiple crossingover is excluded. It seems, however that an adequate definition of distance is available if we try to formulate in terms of probability theory what the biologist had in mind. Let $j \geqq i$. *The distance* $d_{i,j+1}$ *between* i *and* $j+1$ *may be defined as the mathematical expectation of the number of crossingovers in* $(i, j+1)$, i.e. in the $j + 1 - i$ intervals between i and $j + 1$. Hence if

[11] For a geometric equivalent of m points with $m(m-1)/2$ *arbitrary* distances we would have to turn to an $(m-1)$-dimensional space. In fact it is well known that there are between k points in the plane only $S_2 = 2k - 3$ arbitrary distances, in space only $S_3 = 3k - 6$, in r-space $S_r = rk - r(r+1)/2$. Hence for $r = m - 1$ and $k = m$: $S_{m-1} = m(m-1)/2$.

$P_{i,\cdots,j}(x)$ denotes, as before, the probability of exactly x crossovers in these $(j + 1 - i)$ intermediate intervals the formula holds

$$d_{i,j+1} = \sum_{x=0}^{j+1-i} xP_{i\cdots j}(x). \tag{52}$$

Of course, an appropriate unit may be used such that in practical use the distance becomes *proportional* to the d_{ij} introduced above.

The mean value to the right in (52) is well known for any distribution $\pi(x_1, \cdots, x_n)$ whether an "independent" or a general distribution; (i.e. in our case: with or without "interference"). Denoting in the usual way by $\pi_i(x_i)$ the marginal distributions of first order of $\pi(x_1, \cdots, x_n)$ and putting $\pi_i(1) = p_i$ = the probability of success in the i-th trial, we get:

$$d_{i,j+1} = p_i + p_{i+1} + \cdots + p_j \tag{53}$$

and in the same way with $k > j$

$$d_{j+1,k+1} = p_{j+1} + p_{j+2} + \cdots + p_k$$

$$d_{i,k+1} = p_i + p_{i+1} + \cdots + p_k$$

hence $d_{i,j+1} + d_{j+1,k+1} = d_{i,k+1}$, or in general:

$$d_{ij} + d_{jk} = d_{ik} \qquad (i < j < k). \tag{54}$$

It may be mentioned that the additive property of the mathematical expectation which was used here is very well known (particularly for independent events) but not always correctly proved. The proof is contained in the transformation expressed in the following equalities:

$$\begin{aligned} d_{i,j+1} &\equiv \sum_{x=0}^{j+1-i} xP_{i\cdots j}(x) \\ &= \sum_{x_i} \sum_{x_{i+1}} \cdots \sum_{x_j} (x_i + x_{i+1} + \cdots + x_j)\pi_{i,i+1,\cdots,j}(x_i, \cdots, x_j) \\ &= \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} (x_i + x_{i+1} + \cdots + x_j)\pi(x_1, x_2, \cdots, x_n). \end{aligned} \tag{55}$$

(For general distributions Stieltjes integrals replace the sums.) In (55) $\pi(x_1, \cdots, x_n)$ is the given n-variate distribution, $P_{i,\cdots,j}$ the probability of exactly x successes in the successive trials numbered $i, \cdots, j$ and $\pi_{i,i+1,\cdots,j}(x_i, \cdots x_j)$ is the respective marginal distribution of $\pi(x_1, \cdots, x_n)$. The first equality in (55) is not obvious, while the second is rather trivial. From the second or third form of $d_{i,j+1}$ in (55), follows (53). The last expression in (55) shows that the expectation of any such sum as $(x_i + x_{i+1} + \cdots x_j)$ can be computed with respect to one and the same distribution $\pi(x_1, \cdots, x_n)$. Therefore *the distance $d_{i,j+1}$ may also be defined as the expectation of $(x_i + x_{i+1} + \cdots + x_j)$ with respect to the c.d.*

Because of the first equation (48) we get from (53)

$$d_{ij} = c_{i,i+1} + c_{i+1,i+2} + \cdots + c_{j-1,j}\,. \tag{53'}$$

Hence the distance d_{ij} is equal to the sum of the $j - i$ intermediate c.p. No difficulty arises for us from the obvious fact that always

$$c_{ij} \leqq d_{ij}\,;\ \text{and in general}\ c_{ij} < d_{ij}\,, \tag{53''}$$

because the distance d_{ij} is defined by (52), or (55) and not as c_{ij}.

On the right side in (53') stands the sum of certain c.p. We have repeatedly remarked that there may be hitherto unknown restrictions for a consistent system of c.p. Hence it is important to notice that *there are no restrictions for the particular* $(m - 1)$ *c.p.* $c_{12}, c_{23}, \cdots, c_{m-1,m}$. *They can be quite arbitrarily chosen* because of $c_{i,i+1} = p_i$. Hence *any geometric representation of m genes arranged on a straight line in arbitrary distances* $d_{i,i+1}$ $(i = 1, 2, \cdots m - 1)$ *is surely consistent.* E.g. m consecutive genes may be arranged with equal distances $d_{12} = d_{23} = \cdots = d_{m-1,m}$. Or some distances may be zero; then the respective genes are localized in the same point, etc.

Finally, let us briefly consider the case of *several linkage groups.* According to §4 the l.d. then resolves into a product of several distributions; e.g. with $m = 12$:

$$\begin{aligned} l(\epsilon_1\epsilon_2, \cdots, \epsilon_{12}) &= f_1(\epsilon_1\epsilon_2\epsilon_3\epsilon_4) f_2(\epsilon_5\epsilon_6\epsilon_7) f_3(\epsilon_8\epsilon_9\epsilon_{10}) f_4(\epsilon_{11}\epsilon_{12}) \\ &= (\tfrac{1}{2})^4 \pi_1(|\,\epsilon_1 - \epsilon_2\,|\,, |\,\epsilon_2 - \epsilon_3\,|\,, |\,\epsilon_3 - \epsilon_4\,|) \\ &\qquad \pi_2(|\,\epsilon_5 - \epsilon_6\,|\,, |\,\epsilon_6 - \epsilon_7\,|) \cdots \pi_4(|\,\epsilon_{11} - \epsilon_{12}\,|). \end{aligned} \tag{56}$$

Then, as postulated by Morgan, *the linear order holds within each of the k groups, whereas all c.p. among the groups are equal to* $\frac{1}{2}$.

Let us conclude this section by transforming the basic conditions (51) of the linear theory by means of (48). This will be needed in the following section. Consider e.g. the condition $c_{13} < c_{14}$, i.e.

$$P_{12}(1) < P_{123}(1) + P_{123}(3) \tag{57'}$$

or $c_{24} < c_{14}$ yields $P_{23}(1) < P_{123}(1) + P_{123}(3)$. Or in the same way:

$$\begin{aligned} P_{123}(1) + P_{123}(3) &< P_{1234}(1) + P_{1234}(3) \\ &< P_{1,\ldots,5}(1) + P_{1,\ldots,5}(3) + P_{1,\ldots,5}(5), \text{ etc.} \end{aligned} \tag{57''}$$

Thus we may express the content of (51) as follows: *The probability that the "event" happens an odd number of times in a set T_i of i consecutive trials is less than the probability that the event happens an odd number of times in the set T_{i+1} or in the set T'_{i+1} each consisting of $i + 1$ consecutive trials where T_{i+1} and T'_{i+1} denotes respectively the sum of T_i and either the immediately following or the immediately preceding trial.* In this form we see again that the linear theory is an assumption, suggested by observations, and by no means logically necessary.

11. Some models of c.d.'s based on the linear theory. The simplest and very important example which has been suggested repeatedly is that of independent crossovers:

(*i*) *Independence.* The crossovers do not influence each other, i.e.

$$p_{ij} = p_i p_j, \qquad p_{ijk} = p_i p_j p_k, \cdots. \tag{58}$$

That this distribution is consistent is well known; hence only the specific inequalities (48) or (57) have to be considered. Here the expressions $P_{12\ldots i}(x)$, used in (57) become very simple, e.g. with $p_i + q_i = 1$:

$$P_{1\ldots4}(1) = p_1q_2q_3q_4 + q_1p_2q_3q_4 + q_1q_2p_3q_4 + q_1q_2q_3p_4.$$

Then a simple computation shows:

$$\begin{aligned} c_{i,j+1} - c_{ij} &= (q_i - p_i) \cdots (q_{j-1} - p_{j-1})p_j \\ c_{i-1,j} - c_{ij} &= p_{i-1}(q_i - p_i) \cdots (q_{j-1} - p_{j-1}). \end{aligned} \tag{59}$$

These differences will be positive if all $q_i - p_i > 0$ or all $p_i < \frac{1}{2}$. Hence: *A consistent c.d. which fulfils the conditions* (51) *of the linear theory is the distribution of "independent crossovers" with basic probabilities* $p_i = c_{i,i+1}$ $(i = 1, 2, \cdots m - 1)$, *with the one restriction*

$$c_{i,i+1} = p_i \leqq \tfrac{1}{2}. \tag{60}$$

The distribution is completely determined by (58). If all $p_i = p = \frac{1}{2}$, we have the particular case of free assortment.

Although this independence is more general than Mendel's original assumption, Morgan, Haldane and others reported observations, not in accordance with this hypothesis. One crossingover seems to prevent others in a certain "neighborhood". This phenomenon was named *interference.* It suggests that we have to consider the c.d. as a distribution of dependent rather than of independent events. This will be done in the following pages. First consider the limit-case of:

(*ii*) *Complete interference or disjoint events.* In this case we have

$$p_{ij} = p_{ijk} = \cdots = p_{12,\ldots,m-1} = 0. \tag{61}$$

Thence it follows that we have simply

$$\begin{aligned} c_{i,i+1} &= p_i \\ c_{i,i+2} &= p_i + p_{i+1} \\ c_{i,i+3} &= p_i + p_{i+1} + p_{i+2}, \text{ etc.} \end{aligned} \tag{62}$$

In this particular case the c.p. are additive $c_{ij} = d_{ij}$. It is obvious from (62) that *in this case the conditions* (51) *of the linear theory are fulfilled.* On the other hand it follows from (49), (for $x = 0$) and (61) that *the system is consistent if and only if*

$$S_1 \equiv p_1 + p_2 + \cdots + p_n \leqq 1 \qquad (n = m - 1). \tag{63}$$

This is in accordance with the fact that nearly or exactly additive c.p. have been observed always in connection with very small p_i-values.

The most striking observation leading to the concept of interference was that $p_{ij} \leqq p_i p_j$, i.e. that double crossovers appeared less frequently than one would have assumed for independent crossovers, but that nevertheless they did appear sometimes. A particularly simple model of dependence or interference which starts with this fact, preserving however, the main structure of independence, is the following:

(iii) One-parametric model of partial interference. Assume as before $(m - 1)$ basic probabilities p_i and put

$$p_{ij} = \epsilon p_i p_j, \qquad p_{ijk} = \epsilon p_i p_j p_k, \cdots, \text{etc.} \qquad (0 \leqq \epsilon \leqq 1). \tag{64}$$

There is independence if $\epsilon = 1$, complete interference for $\epsilon = 0$ and partial interference for intermediate values of ϵ. Let us first investigate conditions for the consistency of this distribution. Necessary and sufficient conditions for a consistent distribution of arbitrarily linked events are well known (see e.g. [3] (b) p. 239). Write $m - 1 = n$. A system of $p_i, p_{ij}, \cdots, p_{1\ldots n}$ is consistent if it is possible to compute from these $(2^n - 1)$ values, 2^n *non-negative values* $\pi(\eta_1, \eta_2, \cdots \eta_n)$ $(\eta_i = 0 \text{ or } 1)$ *which have the sum one* and are given by the formulae:

$$\begin{aligned}
&\pi(11\cdots1) = p_{12\cdots n}\\
&\pi(11\cdots10) = p_{12,\cdots,(n-1)} - p_{12,\cdots,n}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\pi(110\cdots0) = p_{12} - \sum_{(n_1)} p_{12n_1} + \sum_{n_1}\sum_{n_2} p_{12n_1n_2} - \cdots \pm p_{12\cdots n}\\
&\cdots\cdots\cdots\cdots\cdots\\
&\pi(00\cdots0) = 1 - \sum_{n_1} p_{n_1} + \sum_{n_1}\sum_{n_2} p_{n_1n_2} - \cdots \pm p_{12\cdots n}.
\end{aligned} \tag{65}$$

Because of the symmetry of (64) it will be sufficient to check (65) by means of the relations (49) which can be obtained from (65) by collecting groups of equations such that the corresponding $\pi(\eta_1, \cdots, \eta_n)$ show all the same number of 1's as arguments. Write $P_{1\ldots n}(x) = P_x$ for the probability of x successes in the n trials where independent events with basic probabilities p_i are considered, and P'_x for the analogous probability corresponding to the distribution (64) and introduce in the same way S_i and S'_i where as before, $S_2 = \Sigma p_{ij}$, $S_3 = \Sigma p_{ijk}$, etc. We then find:

$$\begin{aligned}
P'_0 &= 1 - S_1 + \epsilon S_2 - \epsilon S_3 + \cdots\\
&= P_0 + (1 - \epsilon)(- S_2 + S_3 - S_4 + \cdots) = P_0\epsilon + (1 - S_1)(1 - \epsilon).
\end{aligned} \tag{66}$$

It follows that the expression $P_0\epsilon + (1 - S_1)(1 - \epsilon)$ must be $\geqq 0$. For $\epsilon = 0$ this condition reduces to (63), whereas for $\epsilon = 1$, there is no restriction at all. On the other hand this is the only restriction of this kind, because we find

$$\begin{aligned}
P'_1 &= S'_1 - 2S'_2 + 3S'_3 - 4S'_4 + \cdots = S_1 - 2\epsilon S_2 + 3\epsilon S_3 - 4\epsilon S_4 + \cdots\\
&= P_1 + (1 - \epsilon)(2S_2 - 3S_3 + 4S_4 \cdots) = P_1\epsilon + S_1(1 - \epsilon).
\end{aligned}$$

This last expression is always $\geqq 0$ because of $P_1 \geqq 0$, $S_1 \geqq 0$, $\epsilon \leqq 1$. Furthermore we find for $i \geqq 2$ that $P_i' = P_i\epsilon$, hence always non negative. Therefore our system is consistent under the one condition (66).

The additional restrictions corresponding to the linear theory have still to be considered. A simple computation yields the result

$$(67) \qquad \begin{aligned} c_{i,j+1} - c_{ij} &= (1 - 2\epsilon p_i)(1 - 2\epsilon p_{i+1}) \cdots (1 - 2\epsilon p_{j-1})p_j \\ c_{i,j+1} - c_{i+1,j+1} &= p_i(1 - 2\epsilon p_{i+1}) \cdots (1 - 2\epsilon p_j). \end{aligned}$$

These differences are $\geqq 0$ if $p_i \leqq \frac{1}{2\epsilon}$ which is, for $\epsilon < 1$, less strong than (60). Hence we sum up: *A consistent model of partial interference with one parameter* ϵ *to fit the observations can be obtained on the basis of* $n = m - 1$ *probabilities* $p_1, p_2, \cdots p_n$ *by means of* (64), *if the condition*

$$(68) \qquad P_0\epsilon + (1 - S_1)(1 - \epsilon) \geqq 0 \quad \text{or} \quad S_1 \leqq 1 + \frac{\epsilon}{1 - \epsilon} P_0$$

holds and the additional restriction required by the "linear theory"

$$(69) \qquad p_i \leqq \frac{1}{2\epsilon}$$

is satisfied. For $\epsilon = 1$ *this reduces to "independent events" or "no interference" with no restriction* (68), *and* (69) *reducing to* (60). *For* $\epsilon = 0$ *our model yields "complete interference" or "disjoint events" with restriction* (68) *becoming* (63) *and no restriction* (69). If we say that this model contains one parameter only, the idea is that the p_i are to be identified with the basic c.p. $c_{i,i+1}$. It might, however, seem adequate to consider ϵ and $p_1, \cdots, p_{m-1}$ as m available parameters which may be determined from the observations by some appropriate method.

(*iv*) *An* $(m - 1)$*-parametric model of partial interference.* Numerical data show (see particularly [4]) that interference is particularly marked i.e. $p_{ij} < p_ip_j$, if the corresponding p_i, p_j are very small, whereas for greater values of the p_i we have more nearly the pattern of independence. This is rather a striking fact, and seems to be well confirmed by observation. In these final pages a model will be studied which takes into account the circumstance that the amount of interference seems to depend on the magnitudes of the p_i. It contains $(m - 1)$ parameters, is therefore rather flexible, but nevertheless very simple.

Assume $m - 1 = n$ numbers ϵ_i where $0 \leqq \epsilon_i \leqq 1$ and form by means of n probabilities p_i :

$$(70) \qquad \epsilon_i p_i = \bar{p}_i \qquad (0 \leqq \epsilon_i \leqq 1) \qquad (i = 1, 2, \cdots, n).$$

We may choose ϵ_i small if the corresponding p_i is small and larger if it is large; if the p's are all of the same order of magnitude the ϵ's need not differ much either. Then we simply define:

$$(71) \qquad p_{ij} = \bar{p}_i\bar{p}_j, \qquad p_{ijk} = \bar{p}_i\bar{p}_j\bar{p}_k, \cdots, \qquad p_{12\ldots n} = \bar{p}_1\bar{p}_2 \cdots \bar{p}_n.$$

Let us investigate the consistency of this model. In analogy to (66) we form with $S_1 = \Sigma p_i$, $\bar{S}_1 = \Sigma \bar{p}_i$, $\bar{S}_2 = \Sigma \bar{p}_i \bar{p}_j$, etc.:

$$(72) \qquad \begin{aligned} P_0' &= 1 - S_1 + \bar{S}_2 - \bar{S}_3 + \cdots \\ &= (1 - \bar{S}_1 + \bar{S}_2 - \bar{S}_3 + \cdots) - \sum_{i=1}^{n} (1 - \epsilon_i) p_i = \bar{P}_0 - \sum_{i=1}^{n} (1 - \epsilon_i) p_i \end{aligned}$$

where P_0' and $\bar{P}_0$ are the probabilities for zero successes for the model under consideration and for independent events with basic probabilities $\bar{p}_i$ respectively; hence $\bar{P}_0 = \prod_{i=1}^{n} (1 - \epsilon_i p_i)$ and we get the condition:

$$(73) \qquad \prod_{i=1}^{n} (1 - \epsilon_i p_i) \geqq \sum_{i=1}^{n} (1 - \epsilon_i) p_i \quad \text{or:} \quad \sum_{i=1}^{n} p_i \leqq \sum_{i=1}^{n} \bar{p}_i + \prod_{i=1}^{n} (1 - \bar{p}_i).$$

If all $\epsilon_i = 1$ there is no restriction (73), while for $\epsilon_i = 0$ we find again (63). The consideration of P_1', P_2', $\cdots$ yields no new condition, because we get, denoting by $\bar{P}_i$ the probability of i successes for the independent events with basic probabilities $\bar{p}_i$:

$$P_1' = S_1 - 2\bar{S}_2 + 3\bar{S}_3 - \cdots \pm n\bar{S}_n = \bar{S}_1 - 2\bar{S}_2 + \cdots \pm n\bar{S}_n + \sum_{i=1}^{n} p_i(1 - \epsilon_i)$$

$$= \bar{P}_1 + \sum_{i=1}^{n} p_i(1 - \epsilon_i) \geqq 0 \quad \text{and:}$$

$$P_i' = \bar{P}_i \geqq 0 \qquad (i \geqq 2).$$

As for the restrictions imposed by the linear theory we find:

$$(74) \qquad \begin{aligned} c_{i,j+1} - c_{ij} &= (1 - 2\bar{p}_i)(1 - 2\bar{p}_{i+1}) \cdots (1 - 2\bar{p}_{j-1})\bar{p}_j + p_j(1 - \epsilon_j) \\ c_{i,j+1} - c_{i+1,j+1} &= \bar{p}_i(1 - 2\bar{p}_{i+1}) \cdots (1 - 2\bar{p}_j) + p_i(1 - \epsilon_i). \end{aligned}$$

Thus the conditions of the linear theory are satisfied if

$$(75) \qquad \bar{p}_i \leqq \tfrac{1}{2} \quad \text{or} \quad p_i \leqq \frac{1}{2\epsilon_i}.$$

Hence summarizing: *On the basis of $m - 1$ probabilities p_i a consistent model of partial interference is obtained by means of* (70) *and* (71) *if the condition of consistency* (73) *and the conditions* (75) *are satisfied.*

It may be that the four simple models described in this section will seem too crude for the description of the complex mechanism of linkage. They could, of course, be combined and modified in various ways in order to serve at least as an approximation to the theoretical picture of reality we wish to construct. But, while these particular attempts may be inadequate, it seems to the author that the underlying principle is not wrong: that a mathematical theory of linkage must finally consist in statements on the l.d. (or the equivalent c.d.). The consideration of the c.p. is not sufficient for this purpose. The mathematical instrument for a theory of linkage seems to be the probability theory of the linkage distribution.

References

[1] F. Bernstein, *Variations- und Erblichkeitsstatistik.* Handbuch der Vererbungswissenschaft, Bd. 1, pp. 1–96.

[2] Kai Lai Chung, "On fundamental systems of probabilities of a finite number of events," *Ann. of Math. Stat.*, Vol. 14 (1943), pp. 234–37.

[3] H. Geiringer, (a) On the probability theory of arbitrarily linked events. *Ann. of Math. Stat.*, Vol. 9 (1938), pp. 260–271.

(b) "A note on the probability of arbitrary events," *Ann. of Math. Stat.*, Vol. 13 (1942) pp. 238–245.

[4] J. B. S. Haldane, (a) "The combination of linkage values and the calculation of distances between the loci of linked factors," *Jour. of Genetics*, Vol. 8 (1919), pp. 299–308.

(b) "Theoretical genetics of autopolyploids," *Jour. of Genetics*, Vol. 22 (1930), pp. 359–372.

[5] G. H. Hardy, "Mendelian proportions in a mixed population," *Science*, Vol. 28, (1908), p. 49–50.

[6] J. S. Huxley (editor), *The New Systematics*, Oxford 1940. (See articles by *S. Wright*, p. 161–183 and *H. J. Muller*, p. 158–268).

[7] H. S. Jennings, (a) "The numerical results of diverse systems of breeding with respect to two pairs of characters, etc.," *Genetics*, Vol. 12 (1917), pp. 97–154.

(b) "The numerical relations in the crossing over of the genes with a critical examination of the theory that the genes are arranged in a linear series," *Genetics*, Vol. 8 (1923), p. 393.

[8] K. v. Körösy, *Versuch einer Theorie der Genkoppelung.* (Bibliotheca Genetica), Leipzig 1929.

[9] K. Mather, *The Measurement of Linkage in Heredity*, London 1938.

[10] G. Mendel, "Versuche über Pflanzenhybriden." *Verh. des Naturforschd. Vereines in Brünn*, IV. Bd., *Abhandlungen* Brünn, 1866 pp. 3–47.

[11] T. H. Morgan, *The Theory of the Gene.* New Haven, 1928.

[12] K. Pearson, "On a generalized theory of alternative inheritance with special reference to Mendel's laws," *Phil. Trans. Royal Soc.* (A), Vol. 203 (1904), pp. 53–86.

[13] H. Rademacher, "Mathematische Theorie der Genkoppelung unter Berücksichtigung der Interferenz", 105. *Jahresber.* (1932), Schles. Ges. f. vaterländische Kultur. pp. 1–8.

[14] R. B. Robbins, (a) "Applications of mathematics to breeding problems, II." *Genetics*, Vol. 3 (1918), pp. 73–92.

(b) "Some applications of mathematics to breeding problems III." *Genetics*, Vol. 3 (1918), pp. 375–389.

[15] H. Tietze, "Über das Schicksal gemischter Populationen nach den Mendelschen Vererbungsgesetzen," *Zs. Angew. Math. u. Mech.*, Bd. 3, (1923), pp. 362–393.

[16] W. Weinberg, "Über Vererbungsgesetze beim Menschen." Zs. f. induktive Abstammungs- und Vererbungslehre, Vol. 1 (1909) p. 277–330.

[17] S. Wright, "Statistical genetics and evolution," *Bull. Amer. Math. Soc.*, Vol. 48 (1942), pp. 223–246.