84. *Probability-theoretic Investigations on Inheritance. I. Distribution of Genes.*

By Yûsaku KOMATU.

Department of Mathematics, Tokyo Institute of Technology and Department of Legal Medicine, Tokyo University.

(Comm. by T. FURUHATA, M.J.A., July 12, 1951.)

0. Introduction.

The inheritance phenomenon has found increasing application in various branches of biology and medicine. Its theoretical foundation, especially from probability-theoretic or statistical point of view, has together been attempted. Most of such studies has treated, however, the problems only separately corresponding to individual concrete cases. It is the purpose of this paper and its subsequent papers to investigate the phenomenon based upon a general mode of inheritance from probability-theoretic standpoint in detail and to develop a unified theory in a systematic manner.

1. Mode of inheritance to be discussed.

Let us consider a single inherited character. The character of every individual is composed of a pair of genes each of which has originated from each of its parents through their sexual cells. We suppose now, in general, that the character in question consists of m distinct genes denoted by A_i $(i = 1, \ldots, m)$, the inheritance of which is subject to Medelian law. The possible genotypes may then be denoted by A_iA_j $(i, j = 1, \ldots, m)$.

But, since the order of genes in a genotype is immaterial, both genotypes A_iA_j and A_jA_i for different suffices i, j must also be regarded as identical each other. Hence, introducing the abbreviated notation

$$A_{ij} = A_iA_j ,$$

the symmetry relations $A_{ij} = A_{ji}$ for any i and j follow immediately. In view of these relations, the number of different genotypes amounts to

$$\frac{1}{2}m(m+1) ; \tag{1.1}$$

there exist m homozygotes A_{ii} $(i = 1, \ldots, m)$ and $\frac{1}{2}m(m-1)$ heterozygotes A_{ij} $(i, j = 1, \ldots, m ; i < j)$.

If a gene A_j is recessive against A_i, then the genotypes A_{ii} and A_{ij} present both, as a phenotype, the character A_i, while the genotype A_{jj} alone presents the character A_j. This dominance

relation itself having no effect upon other genotypes, we conclude that the existence of a dominance relation diminishes the number of phenotypes by one compared with that of genotypes. But if genes A_i and A_j are equally intensive each other, three genotypes A_{ii}, A_{ij} and A_{jj} present, as phenotypes, the characters A_i, A_iA_j and A_j respectively, and hence the passage from genotypes to phenotypes does not cause the diminution of the number. Therefore, if there exist l dominance relations on the whole between m genes, the number of different phenotypes becomes

$$m^* = \tfrac{1}{2}m(m+1) - l \,. \tag{1.2}$$

2. Frequencies of genes.

Now, we denote by p_i the (relative) frequency, i.e. the probability of distribution, of the gene A_i in a population, while by $\bar{A}_{ii}$ and $\bar{A}_{ij}$ the frequencies of homozygote A_{ii} and heterozygote A_{ij} respectively. These quantities must evidently satisfy the *fundamental relations*

$$\sum_{i=1}^{m} p_i = 1 \qquad \text{and} \qquad \sum_{i=1}^{m} \bar{A}_{ii} + \sum_{i<j} \bar{A}_{ij} = 1 \,. \tag{2.1}$$

If the population is in an equilibrium state with respect to distribution of genes, the frequencies of genotypes are given by

$$\bar{A}_{ii} = p_i^2 \,, \qquad \bar{A}_{ij} = 2p_ip_j \quad (i, j = 1, \ldots, m \,;\, i<j) \,. \tag{2.2}$$

When dominance relations are existent, the frequency of phenotype A_μ^* possessing dominant gene A_i is then given by

$$\bar{A}_\mu^* = p_i^2 + 2p_ip_j + 2p_ip_k + \cdots \quad (\mu = 1, \ldots, m^*) \,, \tag{2.3}$$

A_j, A_k, $\cdots$ being supposed to be the corresponding recessive genes. In view of (2.1) we obtain the corresponding identity

$$\sum_{\mu=1}^{m^*} \bar{A}_\mu^* = 1 \,. \tag{2.4}$$

In quantitative treatment of an inherited character, the directly observed quantities are merely the frequencies of phenotypes, and hence the relative frequencies $\bar{A}_\mu^*$ ($\mu = 1, \ldots, m^*$) are regarded as to be basic. In probability-theoretic study of inheritance phenomenon, however, the frequencies p_i ($i = 1, \ldots, m$) play strictly important role. These fundamental quantities must therefore be calculated suitably in such a manner that the above mentioned relations are satisfied. Now, the frequencies of genes and of phenotypes must, in fact, satisfy the relations (2.1) and (2.4). It is therefore necessary to determine the unknown quantities p_i from the simultaneous equations (2.3) under the condition (2.1), $\bar{A}_\mu^*$ being supposed to be given. Hence, the question is essentially to determine $m-1$ quantities from m^*-1 independent equations. As the number of equations exceeds that of the unknown quantities by

$$(2.5) \qquad (m^*-1)-(m-1)=\tfrac{1}{2}m(m-1)-l\,,$$

the observed frequencies of phenotypes must satisfy certain relations which are equal to the difference (2.5) in number and are independent of each other and also of the fundamental relation (2.4); i.e. m^*-m. In other words, in any population with equilibrium distribution, the frequencies of phenotypes may not be quite arbitrary, but certain relations of the above cited sort must be satisfied, which may and will in the sequel be called *equilibrium conditions*.

Although the existence of recessive genes makes generally impossible to know directly the explicit frequencies of genotypes, the similar circumstances as noticed above for phenotypes are also valid for genotypes. Here the necessary number of equilibrium conditions is $(\frac{1}{2}m(m+1)-1)-(m-1)=\frac{1}{2}m(m-1)$, which is equal to the number of heterozygotes. These conditions can explicitly be obtained in following manner. In fact, if the distribution is in equilibrium state, it is necessary that the relations $\bar{A}_{ij}^2=(2p_ip_j)^2=4p_i^2p_j^2=4\bar{A}_{ii}\bar{A}_{jj}$ hold for every pair of different i and j. Conversely, it is easily seen that these $\frac{1}{2}m(m-1)$ relations

$$(2.6) \qquad \bar{A}_{ij}^2=4\bar{A}_{ii}\bar{A}_{jj} \qquad (i,\ j=1,\ \ldots,\ m;\ i<j)$$

also suffice for the distribution to be in equilibrium.

In theoretical treatment of problems, there needs frequently to know frequencies of genes from those of genotypes. If the considered population is in an equilibrium state, various methods for this purpose will be possible according to the above mentioned circumstances. For instance, we may simply put

$$(2.7) \qquad p_i=\sqrt{\bar{A}_{ii}}\,,$$

or also

$$(2.8) \qquad p_i=1-\sum_{j\neq i}\sqrt{\bar{A}_{jj}}\,.$$

In the quantitative theory of inheritance the frequencies p_i $(i=1,\ \ldots,\ m)$ of genes are the quantities playing the definitively important role. Moreover, on the way of many calculations, the fundamental relation $\sum p_i=1$ will frequently be used very effectively. It is therefore not only disagreeable but also very inconvenient that this relation is satisfied by calculated values only within a certain extent of errors. It is in any case to be desired that we obtain formulae for calculating frequencies of genes in which this relation is always satisfied identically. From this point of view, the formulae (2.7) and (2.8) are both to be regarded incomplete. But, by combining these formulae, we can establish new formulae for which our requirement is fulfilled. Namely, we

can take the mean of both formulae with weight $(m-1):1$.

Such formulae are of course not necessarily unique. In fact, we can obtain another such formula also fulfilling our requirement. We will take here remarkable formulae which do not contain even a square-root process, as follows. In view of $\sum p_i = 1$ we obtain

$$\bar{A}_{ii} + \tfrac{1}{2}\sum_{j\neq i}\bar{A}_{ij} = p_i^2 + \tfrac{1}{2}\sum_{j\neq i} 2p_ip_j = p_i\left(p_i + \sum_{j\neq i} p_j\right) = p_i\,,$$

from which *we get the formulae for calculating frequencies of genes:*

$$(2.9)\quad p_i = \bar{A}_{ii} + \tfrac{1}{2}\sum_{j\neq i}\bar{A}_{ij} = \tfrac{1}{2}\left(\bar{A}_{ii} + \sum_{j=1}^{m}\bar{A}_{ij}\right) \quad (i = 1, \ldots, m)\,.$$

It is easy to verify that they satisfy identically the fundamental relation required.

3. Constancy of equilibrium concerning generations.

Let us consider again a population which has an equilibrium distribution of the inherited character. Then, the frequencies of genotypes are given by the formulae (2.2), i.e.

$$(3.1)\qquad \bar{A}_{ii} = p_i^2\,, \qquad \bar{A}_{ij} = 2p_ip_j \qquad (i \neq j)\,.$$

Suppose now that matings in the population take place at random with respect to the character; in other words, that the rate of matings $A_{ij} \times A_{hk}$ is equal to the probability of selecting the corresponding pair of A_{ij} and A_{hk} in the population. The number of possible different matings regarding genotypes is then equal to $\frac{1}{8}m(m+1)(m^2+m+2)$.

Probability of each type of matings and probability of producing each genotype multiplied by mating-probability are listed in the additional table; here the suffices i, j, h, k being supposed to be different each other. Making use of this table we shall calculate the distribution of the character in the next generation. First, all the possible matings which can produce the homozygote A_{ii} are those in which the both contain the gene A_i. Such types of matings are

Mating	Probability of mating	Probability of producing each type						
		A_{ii}	A_{ij}	A_{ih}	A_{ik}	A_{jj}	A_{jh}	A_{jk}
$A_{ii}\times A_{ii}$	p_i^4	p_i^4	—	—	—	—	—	—
$A_{ii}\times A_{ik}$	$4p_i^3p_k$	$2p_i^3p_k$	—	—	$2p_i^3p_k$	—	—	—
$A_{ii}\times A_{hh}$	$2p_i^2p_h^2$	—	—	$2p_i^2p_h^2$	—	—	—	—
$A_{ii}\times A_{hk}$	$4p_i^2p_hp_k$	—	—	$2p_i^2p_hp_k$	$2p_i^2p_hp_k$	—	—	—
$A_{ij}\times A_{ij}$	$4p_i^2p_j^2$	$p_i^2p_j^2$	$2p_i^2p_j^2$	—	—	$p_i^2p_j^2$	—	—
$A_{ij}\times A_{ik}$	$8p_i^2p_jp_k$	$2p_i^2p_jp_k$	$2p_i^2p_jp_k$	—	$2p_i^2p_jp_k$	—	—	$2p_i^2p_jp_k$
$A_{ij}\times A_{hk}$	$8p_ip_jp_hp_k$	—	—	$2p_ip_jp_hp_k$	$2p_ip_jp_hp_k$	—	$2p_ip_jp_hp_k$	$2p_ip_jp_hp_k$

$$(3.2) \qquad A_{ii} \times A_{ii}\,, \quad A_{ii} \times A_{ik}\,, \quad A_{ij} \times A_{ij}\,, \quad A_{ij} \times A_{ik}\,;$$

the total number of these matings is equal to $\frac{1}{2}m(m+1)$.

The probability of producing the type A_{ii} is then given by

$$p_i^4 + 2p_i^3 \sum_{k \neq i} p_k + p_i^2 \sum_{j \neq i} p_j^2 + 2p_i^2 \sum_{j,\ k \neq i;\ j<k} p_j p_k = p_i^2\,.$$

Next, the possible matings which give rise to the heterozygote A_{ij} $(i \neq j)$ are those in which the one contains A_i and the other contains A_j. Such types of matings are

$$(3.3) \qquad \begin{array}{l} A_{ii} \times A_{ij}\,, \quad A_{jj} \times A_{ji}\,, \quad A_{ii} \times A_{jk}\,, \quad A_{jj} \times A_{ik}\,, \quad A_{ii} \times A_{jj}\,, \\ A_{ij} \times A_{ij}\,, \quad A_{ij} \times A_{ik}\,, \quad A_{ji} \times A_{jk}\,, \quad A_{ki} \times A_{kj}\,, \quad A_{ih} \times A_{jk}\,, \end{array}$$

the total number of which is $\frac{1}{2}(m-1)(m+6)$.

The probability of producing the type A_{ij} is then given by

$$2p_i^3p_j + 2p_j^3p_i + 2p_i^2p_j \sum_{k \neq i,\ j} p_k + 2p_j^2p_i \sum_{k \neq i,\ j} p_k + 2p_i^2p_j^2 + 2p_i^2p_j^2$$
$$+ 2p_i^2p_j \sum_{k \neq i,\ j} p_k + 2p_j^2p_i \sum_{k \neq i,\ j} p_k + 2p_ip_j \sum_{k \neq i,\ j} p_k^2 + 2p_ip_j \sum_{h,\ k \neq i,\ j;\ h \neq k} p_hp_k = 2p_ip_j\,.$$

Thus, frequencies of genotypes remaining constant concerning generations, those of phenotypes remain also constant. Hence:

Supposing that matings take place at random, the equilibrium distribution of the inherited character remains constant concerning generations.

4. Buffer effect against disturbances and stability of equilibrium.

The constancy of equilibrium distribution concerning generations, stated in the preceding section, may be discussed from a more general stand-point. Consider now a population whose distribution of the inheritance character deviates from equilibrium state. It is a matter of indifference what is a cause of such deviation; for instance, it may be caused by migration of another races possessing different distributions. A general problem of such a cross-breeding will also be discussed in a subsequent chapter.

We first notice here that, *when two or more races with different distributions are unified, the resulting distribution in new race with mosaic composition can generally never present an equilibrium state, even if each component race has been in equilibrium state.* In fact, considering g races $X^{(\nu)}$ $(\nu = 1, \ldots, g)$ with respective frequencies $p_i^{(\nu)}$ of genes A_i, the frequencies of genotypes of such races are given by

$$(4.1) \qquad \bar{A}_{ii}^{(\nu)} = p_i^{(\nu)2}\,, \qquad \bar{A}_{ij}^{(\nu)} = 2p_i^{(\nu)}p_j^{(\nu)} \quad (i \neq j)\,.$$

Suppose now that the races $X^{(\nu)}$ are mixed at the rate $\lambda^{(\nu)}$ (> 0) with $\sum_{\nu=1}^{g} \lambda^{(\nu)} = 1$, the distribution of the new-composed race then becomes

$$\bar{A}_{ii} = \sum_{\nu=1}^{g} \lambda^{(\nu)} \bar{A}_{ii}^{(\nu)}, \qquad \bar{A}_{ij} = \sum_{\nu=1}^{g} \lambda^{(\nu)} \bar{A}_{ij}^{(\nu)} \tag{4.2}$$

for which we have

$$\begin{aligned} 4\bar{A}_{ii}\bar{A}_{jj} - \bar{A}_{ij}^2 &= 4\sum_{\nu=1}^{g} \lambda^{(\nu)} \bar{A}_{ii}^{(\nu)} \sum_{\nu=1}^{g} \lambda^{(\nu)} \bar{A}_{jj}^{(\nu)} - \left(\sum_{\nu=1} \lambda^{(\nu)} \bar{A}_{ij}^{(\nu)}\right)^2 \\ &= 4\sum_{\nu=1}^{g} \lambda^{(\nu)} p_i^{(\nu)2} \sum_{\nu=1}^{g} \lambda^{(\nu)} p_j^{(\nu)2} - \left(2\sum_{\nu=1}^{g} \lambda^{(\nu)} p_i^{(\nu)} p_j^{(\nu)}\right)^2 \\ &= 4\sum_{\mu<\nu} \lambda^{(\mu)} \lambda^{(\nu)} (p_i^{(\mu)} p_j^{(\nu)} - p_j^{(\mu)} p_i^{(\nu)})^2. \end{aligned} \tag{4.3}$$

From this we conclude immediately that the equilibrium condition (2.6) can be satisfied by the composed population if and only if all the component races have an identical distribution. It will be noted by the way that the fact $4\bar{A}_{ii}\bar{A}_{jj} - \bar{A}_{ij}^2 > 0$ provided $p_i^{(\mu)} \neq p_i^{(\nu)}$ $(i = 1, \ldots, m)$ at least for a pair of μ, ν, follows simply from a well-known inequality due to Cauchy.

We now turn to a single population with distribution which is *not in equilibrium state*, and denote the frequencies of genotypes A_{ij} especially by

$$P_{ij} \ (i, j = 1, \ldots, m;\ i \leqq j) \qquad (P_{ij} = P_{ji});$$

here the relation $\sum_{i \leqq j} P_{ij} = 1$ which corresponds to the second relation (2.1), being of course assumed to be satisfied. Our main object is to determine the distribution in the next generation. It will be shown that *if the matings take place at random in the same sense as stated at the beginning part of the preceding section, the distribution will arrive at an equilibrium state soon in the next generation.*

For this purpose, we consider first all the possible matings in the initial generation. All the kinds of matings are just as in the table in the preceding section, and those which can produce the homozygote A_{ii} are also the same as stated there in (3.3). But in our present case, we must take, as the probabilities of matings, the new quantities P_{ii}^2, $2P_{ii}P_{ik}$, P_{ij}^2, $2P_{ij}P_{ik}$, instead of p_i^4, $4p_i^3p_k$, $4p_i^2p_j^2$, $8p_i^2p_jp_k$ in previous case, respectively. Making use of those probabilities we obtain the frequency of A_{ii} in the next generation as follows:

$$\begin{aligned} \bar{A}_{ii} &= P_{ii}^2 + P_{ii}\sum_{k \neq i} P_{ik} + \tfrac{1}{4}\sum_{j \neq i} P_{jj}^2 + \tfrac{1}{2}\sum_{j,\, k \neq i;\, i<k} P_{ij}P_{ik} \\ &= \left(P_{ii} + \tfrac{1}{2}\sum_{k \neq i} P_{ik}\right)^2. \end{aligned} \tag{4.4}$$

In quite a similar manner, we calculate the frequency of the heterozygote A_{ij} $(i \neq j)$ in the next generation, namely

$$\bar{A}_{ij} = P_{ii}P_{ij}+P_{jj}P_{ji}+P_{ii}\sum_{k\neq i,\,j}P_{jk}+P_{jj}\sum_{k\neq i,\,j}P_{ik}+2P_{ii}P_{jj}+\tfrac{1}{2}P_{ij}^2$$
$$+\tfrac{1}{2}P_{ij}\sum_{k\neq i,\,j}P_{ik}+\tfrac{1}{2}P_{ji}\sum_{k\neq i,\,j}P_{jk}$$
$$+\tfrac{1}{2}\sum_{k\neq i,\,j}P_{ki}P_{kj}+\tfrac{1}{2}\sum_{h,\,k\neq i,\,j\,;\,h\neq k}P_{ih}P_{jk}$$
$$=\Big(P_{ii}+\tfrac{1}{2}\sum_{k\neq i}P_{ik}\Big)\Big(P_{jj}+\tfrac{1}{2}\sum_{k\neq j}P_{jk}\Big).$$

Thus, if we now put

$$(4.6)\qquad p_i = P_{ii}+\tfrac{1}{2}\sum_{k\neq i}P_{ik}\qquad (i=1,\ \ldots,\ m)\,,$$

the above formulae (4.4) and (4.5) may also be expressed in the form

$$(4.7)\qquad \bar{A}_{ii}=p_i^2\,,\qquad \bar{A}_{ij}=2p_ip_j\qquad (i\neq j)\,.$$

This shows that the distribution in the next generation is just in an equilibrium state.

At this occasion a remarkable significance of the formulae (2.10) may be well recongnized. In fact, it is noticed that the definition (4.6) for p_i may be obtained from (2.10) merely by substituting P's instead of the corresponding $\bar{A}$'s.

Finally, one can really imagine that the genes will be redistributed uniformly according to the assumption of randomness of matings with respect to the inherited character. So the above obtained result will also be previously expected as a matter of course. But, by means of the foregoing analysis we have actually seen at any rate that *against any disturbances caused on distribution a buffer effect will act such that the equilibrium state of distribution remains constant.* In other words, *the equilibrium has a tendency to hold stability.*