ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
IX, 4 (1968)

L. BONEVA (Sofia)

A MATHEMATICAL MODEL OF AN ISOLATED POPULATION AT EQUILIBRIUM

1. Introduction. The present work aims at transfering some concepts of thermodynamics to population genetics. The concept of entropy plays an important role in thermodynamics. It is on this concept that modern information theory is based. The latter has already been applied to a description of the mechanism of transmitting parental characteristics to posterity. Since a population is a set of specimens, the concept of entropy should be used freely enough in population genetics to investigate the development of a population. Statical models, e. g. the Hardy-Weinberg model [13], are used in most treatises on genetics. In the present paper an isolated population is treated as an analogue to a closed thermodynamical system. There are given some definitions concerning isolated populations as well as the conditions for their development.

Under "isolated population" we mean a Mendel-type population isolated in the geographical sense ([3], [5], [6]), i. e. such a population in which the heredity factors are invariable with respect to time and which develops itself under constant environmental influence. Moreover, it is assumed that the isolated population consists of a sufficient number of specimens of a given species, their crossing being random. We shall consider such a population as a system developing in time.

We assume that the development of an isolated population is an irreversible process. Taking an arbitrary generation of the successive generations of a given population as the initial one, it will characterize the initial state of the population. Each successive generation will then represent the state of the population at the moment under consideration.

The state of a specimen is characterized by a number of features. Here we assume that each specimen is characterized by a finite number of constant quantitative characteristics. To put it another way, we limit our consideration to grown-up specimens which have accomplished their physical development but have not yet entered into the process of ageing.

Such characteristics could be the height, the length of different bones, the dimensions of the skull, etc., while in lower organisms the main characteristics are the dimensions of the cell. All these characteristics are constant for a given period of time and no internal change of specimens is observed. These characteristics are conditioned by the joint genetic effect of many genes situated at different loci. The separate effect of each of these genes is very insignificant which means that the features under investigation are continuously distributed within the population.

Thus the state of a separate specimen is determined by the values of its characteristics, while the state of the whole population is determined by the joint probability distribution of these characteristics. The change in these characteristics from generation to generation is the cause for each successive generation to enter into a new stage of development. However, we may assume that at some moment the population reaches such a state where the effect of all factors is equal to zero. This would be determined by isolation, the irreversibility of the development, and the tendency for stabilization. We shall call such a state of a given isolated population an *equilibrium state.* There are two reasons for the stability of a once reached equilibrium state. The first of them is the equilibrium between mutations and eliminations of the genes, while the second one is the equilibrium between favourable heterozygotes and unfavourable homozygotes ([3], [11]). Consequently, the equilibrium is a state of the population that cannot be changed except by external factors (such as migrations, changes of environment, etc.).

We should note, however, that in the development of an isolated population the results of the random crossing become more and more miscellaneous, due to mutations and recombinations. It follows that the state of the population becomes more and more indefinite. The entropy is a measure for the indefiniteness of a system. Thus, it could be assumed that the entropy of an isolated population grows from generation to generation. Irrespectively of the increase of entropy, the lack of external influences is noticeable in the stabilization of other parameters of the probability distribution of characteristics in an isolated population. It could be assumed that under conditions excluding external influences the mean values, the variances, and the covariances of the characteristics do not change from generation to generation. This would yield the statement that the entropy of the distribution of characteristics is bounded. We may expect that an isolated population reaches its equilibrium state if and only if the entropy of the joint distribution of characteristics reaches its maximum value. These assumptions will be considered in detail in the following section. Some rather interesting inferences of our assumptions will also be indicated.

2. The mathematical model. Let us consider a dynamic population determined by the sequence of probability distributions of n-dimensional random vectors $(X_1^{(k)}, X_2^{(k)}, \ldots, X_n^{(k)})$, $k = 0, 1, 2, \ldots$, having probability densities $f^{(k)}(x_1, x_2, \ldots, x_n)$. The density $f^{(k)}(x_1, x_2, \ldots, x_n)$ characterizes the k-th generation of the dynamic population, while

$$H^{(k)} = H(f^{(k)}) = E\{-\log f^{(k)}(x_1, x_2, \ldots, x_n)\},$$

provided that it exists, is the entropy the k-th generation.

If in a given dynamic population the densities $f^{(k)}(x_1, x_2, \ldots, x_n)$ are equal for all generations, thus if

$$f^{(k)}(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n) \tag{1}$$

holds, this population will be called a *population at equilibrium.*

If (1) is valid from some $k > 0$ onwards only, we say that the population reaches its *equilibrium at the k-th generation.*

A dynamic population is called *isolated* if for $k = 0, 1, 2, \ldots$ the following conditions are satisfied:

(a) $$EX_i^{(k)} = m_i, \quad i = 1, 2, \ldots, n;$$

(b) $$E[(X_i^{(k)} - m_i)(X_j^{(k)} - m_j)] = \lambda_{ij}, \quad i, j = 1, 2, \ldots, n;$$

(c) $$H^{(k)} \leqslant H^{(k+1)};$$

(d) equality in (c) holds if and only if $H^{(k)}$ is the maximum entropy value $H_{\max}$ in the class of probability distributions with first and second moments given by (a) and (b).

The following theorem may be proved for the maximum entropy:

THEOREM. *If $f = f(x_1, x_2, \ldots, x_n)$ satisfies conditions* (a) *and* (b) *in the definition of an isolated population, the equality $H(f) = H_{\max}$ is satisfied if and only if f is the n-dimensional normal density.*

To prove the theorem we will use the following

LEMMA. *If the functions $f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n)$ and $f_0(\boldsymbol{x}) = f_0(x_1, x_2, \ldots, x_n)$ defined in R_n satisfy the conditions*

(i) $$f(\boldsymbol{x}) \geqslant 0 \quad \text{and} \quad f_0(\boldsymbol{x}) > 0 \quad \text{for all} \quad \boldsymbol{x} \in R_n,$$

(ii) $$\int_{R_n} f(\boldsymbol{x})\,d\boldsymbol{x} = \int_{R_n} f_0(\boldsymbol{x})\,d\boldsymbol{x} = 1,$$

(iii) $$\int_{R_n} f(\boldsymbol{x}) \log f_0(\boldsymbol{x})\,d\boldsymbol{x} = \int_{R_n} f_0(\boldsymbol{x}) \log f_0(\boldsymbol{x})\,d\boldsymbol{x},$$

then

$$\int_{R_n} f(\boldsymbol{x}) \log f(\boldsymbol{x})\,d\boldsymbol{x} \geqslant \int_{R_n} f_0(\boldsymbol{x}) \log f_0(\boldsymbol{x})\,d\boldsymbol{x}, \tag{2}$$

where the equality holds if and only if $f(\boldsymbol{x}) \equiv f_0(\boldsymbol{x})$ *almost everywhere.*

Proof of the lemma. Let us note at the beginning that the first two derivatives of the function $\varphi(x) = x\log x$ are:

$$\varphi'(x) = \log x+1, \qquad \varphi''(x) = 1/x.$$

As for $x > 0$ the second derivative is positive, $\varphi(x)$ is a convex function and for arbitrary $u \geqslant 0$ and $v > 0$ we have the inequality (here we assume that $\varphi(0) = 0$):

$$\varphi(u)-\varphi(v) = u\log u - v\log v \geqslant (u-v)(\log v+1). \tag{3}$$

In (3) equality holds if and only if $u = v$.

Putting $u = f(\boldsymbol{x})$ and $v = f_0(\boldsymbol{x})$, and integrating both sides of (3) we may use the assumptions of the lemma and obtain

$$\int_{R_n} [f(\boldsymbol{x})\log f(\boldsymbol{x}) - f_0(\boldsymbol{x})\log f_0(\boldsymbol{x})]\,d\boldsymbol{x} \geqslant$$

$$\geqslant \int_{R_n} [f(\boldsymbol{x})-f_0(\boldsymbol{x})][\log f_0(\boldsymbol{x})+1]\,d\boldsymbol{x} =$$

$$= \int_{R_n} f(\boldsymbol{x})\log f_0(\boldsymbol{x})\,d\boldsymbol{x} - \int_{R_n} f_0(\boldsymbol{x})\log f_0(\boldsymbol{x})\,d\boldsymbol{x} +$$

$$+ \int_{R_n} f(\boldsymbol{x})\,d\boldsymbol{x} - \int_{R_n} f_0(\boldsymbol{x})\,d\boldsymbol{x} = 0.$$

Therefrom immediately follows inequality (2). Finally, as in (3) the equality sign is valid if and only if $u = v$, in (2) we have an equality only when $f(\boldsymbol{x}) \equiv f_0(\boldsymbol{x})$ almost everywhere. Thus the lemma is proved.

Proof of the theorem. Let Λ be the determinant of the symmetric matrix $\{\lambda_{ij}\}$ and Λ_{ij} the cofactor of the element λ_{ij}. The function

$$f_0(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Lambda|}} \exp\left[-\frac{1}{2\Lambda}\sum \Lambda_{ij}(x_i-m_i)(x_j-m_j)\right]$$

is the probability density of a n-dimensional normal distribution and it satisfies conditions (a) and (b) of the definition of an isolated population.

Let $f(\boldsymbol{x})$ be any arbitrary n-dimensional density which also satisfies conditions (a) and (b). In order to prove the theorem it is sufficient to show that the functions $f_0(\boldsymbol{x})$ and $f(\boldsymbol{x})$ satisfy the conditions of the lemma.

Conditions (i) and (ii) are implied by the type of the function $f_0(\boldsymbol{x})$ and by the fact that both functions are probability density functions. In order to check condition (iii) we compute

$$\log f_0(\boldsymbol{x}) = -\frac{1}{2\Lambda}\sum \Lambda_{ij}(x_i-m_i)(x_j-m_j)+C$$

where $C = -\log[(2\pi)^{n/2}\sqrt{|A|}]$.

Hence

$$\int_{R_n} f(x)\log f_0(x)\,dx =$$

$$= -\frac{1}{2A}\sum A_{ij}\int_{R_n}(x_i-m_i)(x_j-m_j)f(x)\,dx + C\int_{R_n} f(x)\,dx =$$

$$= -\frac{1}{2A}\sum A_{ij}\lambda_{ij} + C = -\frac{n}{2} + C,$$

and, analogically,

$$\int_{R_n} f_0(x)\log f_0(x)\,dx = -\frac{n}{2} + C.$$

Consequently, condition (iii) is also satisfied and the lemma may be applied. This ends the proof of the theorem.

It follows from the theorem just proved, that an isolated population is at equilibrium if and only if $f^{(0)}(x_1, x_2, \ldots, x_n)$ is the n-dimensional normal density. On the other hand, two requirements:

1° $f^{(k-1)}(x_1, x_2, \ldots, x_n) \neq f^{(k)}(x_1, x_2, \ldots, x_n)$;

2° $f^{(k)}(x_1, x_2, \ldots, x_n)$ to be the density of a n-dimensional normal distribution,

form a necessary and sufficient condition for an isolated population to be at equilibrium, starting from the k-th generation ($k > 0$).

3. Discussion and biological inferences. Under the assumption that entropy increases with the development of the isolated population, while the mean values and the second moments remain constant, we have proved that the distribution of the characteristics at a population equilibrium is a normal distribution. This may be one of the possible explanations of the fact that most characteristics really have a normal distribution. This explanation is different from the usually given explanations based on the Moivre-Laplace theorem or on the central limit theorem of probability theory.

Known genetic models are built predominantly on the basis of Laplace's theory and the Mendel laws. The first adequate genetic model was given by Hardy and Weinberg in 1908. Later on there appeared many other models based on the same probability schemes but much more developed and elaborate. Such are the models of Bernstein [1], Fisher [5], [6], Feller [4], Kempthorne [8], and other ones. These models refer to isolated characteristics, i. e. a simple genetic characteristic is assumed

to be any characteristic which may be treated as a separate unit (cf. [7]). The laws of Mendel refer also to simple characteristics.

According to the model of Hardy and Weinberg the population reaches its equilibrium already at its second generation, i. e. the frequences of the genotypes remain constant from the second generation onwards. This model could be generalized to the case with more genes under investigation but this would require the additional assumption of the independence of genes. That means that the loci at which these genes are situated should not be linked. Otherwise, the more linked the revelant loci are, the more distant in time the population equilibrium is (cf. [13]). Hence it follows that on one hand the equilibrium of the population would depend on the number and the situation of the observed genes, while on the other hand these models do not reflect the changes in time and the development of the characteristics in question.

In the model presented in this paper the normal distribution of characteristics results from the equilibrium state of a given isolated population. This seems to fall in line fairly well with the observed reality. When investigating a given population, observations are to be carried out on separate specimens. On the ground of obtained measurements the parameters of the population and the distribution of the characteristics could be found. This allows for an immediate checking of the accepted assumptions. Strictly speaking, this enables us to check the adequateness of the model describing the development and the equilibrium state of a population.

The following biological inferences could be made:

1° An isolated population has not reached its equilibrium if we can find among the observed characteristics a characteristic with entirely non-normal distribution.

2° The maximum entropy of a population is

$$H_{\max} = \log[(2\pi e.)^{n/2}\sqrt{|\Lambda|}],$$

where $|\Lambda|$ is the generalized variance or the determinant of the matrix of second moments. It follows that the maximum entropy is determined by the generalized variance of the observed characteristics only. Therefore the conditions in the definition of an isolated population might be weakened accordingly; one could assume that only the generalized variance of the distribution of characteristics does not change from generation to generation. It could also be noted that between the determinant Λ and the determinant P of the correlation matrix the following relation holds $\Lambda = \sigma_1^2\sigma_2^2 \ldots \sigma_n^2 P$. For $n = 1$ the determinant Λ is the ordinary variance σ^2, while for $n = 2$ it takes the form $\Lambda = \sigma_2^2\sigma_2^2(1-\varrho^2)$, where ϱ is the correlation coefficient.

3° The greater the maximum entropy of the population, the greater the mean amount of information and, for the case of fixed variances, the more uncorrelated the characteristics.

4° Different populations could be compared on the basis of the preceding remark. That means that a population with greater entropy is more advanced towards stabilization. It is evident that in the compared populations the same characteristics are to be considered.

5° On the basis of 3° it is possible to compare two groups of characteristics of one and the same population. If it is proved that one of the groups has a smaller entropy, the characteristics would be more more correlated and so some of them could be eliminated.

References

[1] С. Н. Бернштейн, *Собрание сочинений*, том 4, Москва 1964.

[2] H. Cramér, *Mathematical methods of statistics*, Princeton 1949, (Polish ed., Warszawa 1958).

[3] T. Dobzhanski, *Dynamik der menschlichen Evolution*, Stuttgart 1965.

[4] W. Feller, *An introduction to probability theory and its applications*, New York 1950, (Polish ed., Warszawa 1960).

[5] R. A. Fisher, *The correlation between relatives on the supposition of Mendelian inheritance*, Trans. Roy. Soc. Edinburgh (1918), pp. 399-433.

[6] —, *The genetical theory of natural selection*, Oxford 1930.

[7] K. Gumiński, *Termodynamika procesów nieodwracalnych*, Warszawa 1962.

[8] O. Kempthorne, *An introduction to genetic statistics*, New York 1957.

[9] M. Lerner, *The genetic basis of selection*, Edinburgh 1963.

[10] I. S. Penrose, *Outline of human genetics*, London 1962.

[11] M. Planck, *Einführung in die theoretische Physik*, Leipzig 1930.

[12] Э. Шенон, *Работы по теории информации и кибернетике*, Москва 1963.

[13] F. Vogel, *Lehrbuch der allgemeinen Humangenetik*, Berlin 1961.

[14] S. Wright, *Evolution in Mendelian populations*, Genetics 16 (1931), pp. 97-159.

INSTITUTE OF MATHEMATICS, BULGARIAN ACADEMY OF SCIENCES
SOFIA, BULGARY

Received on 5. 5. 1966

L. BONEVA (Sofia)

MATEMATYCZNY MODEL POPULACJI IZOLOWANEJ W STANIE RÓWNOWAGI

STRESZCZENIE

Praca niniejsza stanowi próbę zbudowania matematycznego modelu rozkładu cech w danej populacji izolowanej, znajdującej się w stanie równowagi.

W tym celu wykorzystuje się pewne pojęcia termodynamiki. Zakłada się, że izolowana populacja jest analogiem zamkniętego systemu termodynamicznego, a roz-

wój populacji jest procesem nieodwracalnym. Ponieważ populacja stanowi zbiór indywiduów, wygodnie jest posłużyć się pojęciem entropii przy badaniu rozwoju danej populacji.

Przy budowie samego modelu zakłada się, że dynamiczna populacja jest określona przez ciąg rozkładów n-wymiarowych ciągłych zmiennych losowych $(X_1^{(k)}, X_2^{(k)}, \ldots, X_n^{(k)})$ o gęstościach $f^{(k)}(x_1, x_2, \ldots, x_n)$ $k = 0, 1, 2, \ldots$, przy czym gęstość $f^{(k)}(x_1, x_2, \ldots, x_n)$ charakteryzuje k-te pokolenie, a $H^{(k)}$ jest entropią k-tego pokolenia.

Dowodzi się, że przy pewnych założeniach entropia populacji jest maksymalna wtedy i tylko wtedy, gdy $f(x_1, x_2, \ldots, x_n)$ jest gęstością rozkładu n-wymiarowego rozkładu normalnego.

W pracy wysnuwa się ponadto pewne wnioski o zachowaniu się cech w populacji, będącej w stanie równowagi, oraz podaje się pewne wnioski o charakterze biologicznym.

Л. БОНЕВА (София)

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ИЗОЛИРОВАННОЙ ПОПУЛЯЦИИ В СОСТОЯНИИ РАВНОВЕСИЯ

РЕЗЮМЕ

Настоящая работа является попыткой построения математической модели распределении признаков данной изолированной популяции, находящейся в состоянии равновесия.

С этой целью используются некоторые понятия термодинамики. Принимается, что изолированная популяция является аналогом замкнутой термодинамической системы, а развитие популяции — необратимым процессом. Так как популяция является совокупностью индивидуумов, то при исследовании развития данной популяции выгодно использовать понятие энтропии.

При построении самой модели принимается, что динамическая популяция определена последовательностью распределений n-мерных непрерывных случайных величин $(X_1^{(k)}, X_2^{(k)}, \ldots, X_n^{(k)})$, с плотностями $f^{(k)}(x_1, x_2, \ldots, x_n)$, $k = 0, 1, 2, \ldots$, при чем k-тое поколение характеризуется плотностью $f^{(k)}(x_1, \ldots, x_n)$, а $H^{(k)}$ — энтропия k-того поколения.

При некоторых ограничениях доказывается, что энтропия популяции является максимальной тогда и только тогда, когда $f(x_1, \ldots, x_n)$ — плотность n-мерного нормального распределения.

Далее высказываются некоторые предположения о поведении признаков популяции в равновесии и делаются некоторые биологические выводы.