Ann. Hum. Genet., Lond. (1969), **33**, 23
Printed in Great Britain

Population subdivision with respect to multiple alleles

BY C. C. LI
Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Penna, 15213

In view of the current interest in studying human isolated populations and the fact that many human gene markers have multiple alleles, the writer thought that it would be helpful to have the problem of population subdivision discussed in more detail and hence stimulate further investigation. Although the problem under consideration seems at first sight purely genetical, an appropriate change of a few technical terms will render the problem identical with those of epidemiologists and sociologists studying the association of certain traits or diseases in specific (homogeneous) groups and in the combined (heterogeneous) group. It may be said that these problems are 'isomorphic'. Thus, an investigation in one area has similar implications in the others. We shall illustrate these ideas very briefly for the case of two alleles before introducing multiple alleles.

TWO ALLELES

Suppose there are K mendelian isolates, in each of which random mating is practiced. Let p_i and q_i be frequencies of the alleles A and a in the ith isolate; $p_i + q_i = 1$. These isolates in general are not of the same size; let w_i be the relative size of the ith isolate so that $\Sigma w_i = 1, i = 1, 2, \ldots, K$. If we view the K isolates as a whole (i.e. the total population), the frequency of allele A will be

$$\bar{p} = p = \Sigma p_i w_i, \tag{1}$$

the summation being taken over the isolates $i = 1, 2, \ldots, K$ throughout the paper. Note that we shall write simply p (without a bar or subscript) for the average frequency $\bar{p}$. The genotype frequencies of AA, Aa, aa in each isolate are $p_i^2, 2p_iq_i, q_i^2$, respectively. In the total population the genotype frequencies, being the sum of the separate ones, may be written as follows:

$$\begin{array}{ccccc} & A & a & \\ A & \Sigma p_i^2 w_i = p^2 + \sigma^2 & \Sigma p_i q_i w_i = pq - \sigma^2 & p \\ a & \Sigma p_i q_i w_i = pq - \sigma^2 & \Sigma q_i^2 w_i = q^2 + \sigma^2 & q \\ & p & q & 1 \end{array} \tag{2}$$

where σ^2 is the variance of the gene frequency among the isolates (subgroups of the total population). Formula (2) is known as Wahlund's principle (1928; see Wright, 1931, p. 128; Li, 1955, p. 298). Since there are only two alleles, the variance (σ_1^2) of the frequency of allele A is the same as that (σ_2^2) of the allele a, and further, the covariance (σ_{12}) is also of the same magnitude but negative in sign. Hence, in formula (2), there is only one parameter $\sigma^2 = \sigma_1^2 = \sigma_2^2 = -\sigma_{12}$.

The important feature of Wahlund's formula is that when the total of the random mating subgroups is viewed in terms of the average gene frequencies, the homozygotes have been increased in frequency at the expense of the heterozygotes over those expected on random mating. The effect of pooling these random mating subgroups with different p_i's is identical with that of

inbreeding in the total population. Using the inbreeding coefficient F, defined as the correlation coefficient between the uniting gametes (Wright, 1922, 1943) we obtain the equivalence

$$\text{due to subdivision,} \quad \sigma^2 = Fp(1-p), \quad \text{due to inbreeding.} \tag{3}$$

If we have no knowledge of the individual isolates and merely observe the total pooled population, there is no way to tell whether there is heterogeneity or inbreeding.

The analogy between the genetic situation (2) and that of epidemiological studies of the association of diseases is at once apparent by regarding A as a symbol for the presence of arthritis (say) and a for the absence of that ailment. Let the two rows represent the conditions of the wives and the two columns represent the conditions of their husbands. The average gene frequency p then becomes the general prevalence of arthritis in the total population. The various 'isolates' may be identified with the different age groups of the couples. Suppose that there is no association between husband and wife in acquiring arthritis in any age group ($\equiv$ random mating in each isolate). The pooling of the various age groups with different p's will create association between husband and wife with respect to arthritis; which is equivalent to inducing correlation between uniting gametes in the total population. The following numerical illustration should help to unify the problems in genetics and epidemiology.

	Isolate I (young)			Isolate II (old)			Pooled total population		
	A	a	Total	A	a	Total	A	a	Total
A	0·01	0·09	0·10	0·49	0·21	0·70	0·25	0·15	0·40
a	0·09	0·81	0·90	0·21	0·09	0·30	0·15	0·45	0·60
Total	0·10	0·90	1·00	0·70	0·30	1·00	0·40	0·60	1·00
	Correlation = 0			Correlation = 0			Correlation = 0·375		

(4)

Here we have $\sigma^2 = Fp(1-p) = 0{\cdot}25 - 0{\cdot}16 = 0{\cdot}09$. The tables above may be read either as epidemiological data or as genetical data, depending upon the meanings attached to the symbols. Further discussions of the common problem in genetics and epidemiology may be found in Li (1961, Chapters 6 and 10).

One important simplifying feature for two alleles (or 2×2 contingency tables) is that there is no problem in assigning a scale to the alleles (or conditions) A and a. We always obtain the same correlation coefficient no matter what numerical values are given to A and a. In general, however, this is not true for tables larger than 2×2. And this is the difficulty with multiple alleles, as we shall see.

INBREEDING COEFFICIENT, INDEPENDENT OF SCALE

In this section we introduce multiple alleles with inbreeding but not subdivision. Although the methodology and procedures adopted in the following will remain the same for any number of alleles, we shall use only three alleles for brevity and concreteness. Also, in order to avoid double subscripts, we let p, q, r be the frequencies of the alleles A_1, A_2, A_3, respectively; $p+q+r = 1$. If the inbreeding system in the population is such that the correlation between the uniting gametes is F, then the genotype frequencies will be as follows:

	A_1	A_2	A_3	
A_1	$p^2+Fp(1-p)$	$pq-Fpq$	$pr-Fpr$	p
A_2	$pq-Fpq$	$q^2+Fq(1-q)$	$qr-Fqr$	q
A_3	$pr-Fpr$	$qr-Fqr$	$r^2+Fr(1-r)$	r
	p	q	r	1

(5)

In order to be able to calculate a correlation coefficient from any two-way table, we must have numerical values assigned to the marginal variables. In the situation shown in (5), it is legitimate to speak of correlation between uniting gametes, because no matter what numerical values be arbitrarily assigned to the alleles A_1, A_2, A_3, the correlation coefficient for the distribution (5) will turn out to be F. In other words, the value of the correlation coefficient of distribution (5) is invariant of scale. The reader may satisfy himself that this is so by giving A_1, A_2, A_3 any arbitrary numbers, say, Y_1, Y_2, Y_3, or simply, 0, 1, Y, as correlation is independent of origin and unit. A simpler way of seeing the invariant property of the correlation coefficient is to split the frequencies (5) into two components: one of complete random mating and one of complete homozygosis.

	Component size $(1-F)$				Component size F				
	A_1	A_2	A_3		A_1	A_2	A_3		
A_1	p^2	pq	pr	p	p	0	0	p	(6)
A_2	pq	q^2	qr	q	0	q	0	q	
A_3	pr	qr	r^2	r	0	0	r	r	
	p	q	r	1	p	q	r	1	

These component tables are independent of scale. With any set of arbitrary values assigned to A_1, A_2, A_3, the correlation coefficient of the first component is always zero and that of the second component is always unity. Since these two component tables have the same marginal distributions, they may be combined to yield an average correlation (Li and Sacks, 1954, p. 355); which is

$$(1-F)(0)+F(1)=F.$$

In conclusion, we may legitimately speak of correlation between uniting gametes to describe the effect of inbreeding regardless of multiple alleles.

SUBDIVISION WITH RESPECT TO MULTIPLE ALLELES

Let p_i, q_i, r_i be the frequencies of alleles A_1, A_2, A_3 in the ith isolate of relative size w_i. Under random mating each isolate has the genotype frequencies $p_i^2 A_1A_1$, $2p_iq_iA_1A_2$, etc. In the total population the genotype frequencies will be as follows:

	A_1	A_2	A_3	
A_1	$p^2+\sigma_1^2$	$pq+\sigma_{12}$	$pr+\sigma_{13}$	p
A_2	$pq+\sigma_{12}$	$q^2+\sigma_2^2$	$qr+\sigma_{23}$	q
A_3	$pr+\sigma_{13}$	$qr+\sigma_{23}$	$r^2+\sigma_3^2$	r
	p	q	r	1

(7)

where $\sigma_1^2=\Sigma p_i^2 w_i - p^2$ is the variance of A_1 frequency among the isolates, and $\sigma_{12}=\Sigma p_iq_iw_i-pq$ is the covariance of the frequencies of A_1 and A_2, etc. These variances and covariances are related on account of the restriction $p_i+q_i+r_i=1$. For instance, the variance of A_1 frequency is equal to the variance of the (A_2+A_3) combined frequency, as $p_i=1-(q_i+r_i)$;

$$\sigma_1^2=\sigma^2_{(2+3)}=\sigma_2^2+\sigma_3^2+2\sigma_{23}.$$

Thus, all covariances may be expressed in terms of the variances:

$$2\sigma_{12}=\sigma_3^2-\sigma_1^2-\sigma_2^2,\quad 2\sigma_{13}=\sigma_2^2-\sigma_1^2-\sigma_3^2,\quad 2\sigma_{23}=\sigma_1^2-\sigma_2^2-\sigma_3^2 \tag{8}$$

and for each row of (7), the variance and covariances add up to zero.

$$\sigma_1^2+\sigma_{12}+\sigma_{13}=0,\quad \sigma_{12}+\sigma_2^2+\sigma_{23}=0,\quad \sigma_{13}+\sigma_{23}+\sigma_3^2=0. \tag{9}$$

Given the joint distribution (7), our present problem is how to construct an index to measure the overall degree of association between the uniting gametes. Now, we can no longer assign arbitrarily any numerical values to the alleles, because the correlation coefficient varies with scale. If we regard A_1, A_2, A_3 as symbols for three qualitative classes of a trait, it becomes clear that we can only speak of association instead of coefficient of correlation.

Since F has been defined as the correlation between uniting gametes, we shall adopt a different symbol, to denote the degree of association for the two-way distribution (7). Analogous to F in (3), we may define quantities ϕ_1, ϕ_2, ϕ_3, such that

$$\sigma_1^2 = \phi_1 p(1-p), \quad \sigma_2^2 = \phi_2 q(1-q), \quad \sigma_3^2 = \phi_3 r(1-r). \tag{10}$$

As an overall measurement for the degree of association between uniting gametes as caused by subdivision of a population, Yasuda (1968) uses the following index

$$\phi(11) = \phi_1 p + \phi_2 q + \phi_3 r = \frac{\sigma_1^2}{1-p} + \frac{\sigma_2^2}{1-q} + \frac{\sigma_3^2}{1-r}. \tag{11}$$

Although Yasuda (1968) retains the symbol F, it is to be emphasized that the value of $\phi(11)$ does not have the meaning or property of a correlation coefficient.

Before we consider some other indices, it would be well at this stage to examine a numerical example (Table 1) to clarify the meaning of the various symbols as well as to facilitate comparison with other results later on. Most of the calculations in Table 1 are self-evident, and only a few explanatory remarks are necessary. In the last column of the upper portion of Table 1,

$$\phi_1 = \sigma_1^2/p(1-p) = 0{\cdot}02/0{\cdot}24 = 0{\cdot}0833, \text{ etc.},$$

according to (10). The lower portion of Table 1 shows the genotype frequencies of the five random isolates, the last one (lower right) being the average of the preceding five. Thus, the frequency of A_1A_1, in the total population is $(0{\cdot}04+0{\cdot}09+0{\cdot}16+0{\cdot}25+0{\cdot}36)/5 = 0{\cdot}18$. This 'summary' table corresponds to joint distribution (7). Again, the frequency of A_1A_1 in the total population is $p^2+\sigma_1^2 = 0{\cdot}16+0{\cdot}02 = 0{\cdot}18$. It is the association between the uniting gametes shown in this table that we wish to measure. Yasuda's proposal yields

$$\phi(11) = 0{\cdot}0833\,(0{\cdot}40) + 0{\cdot}25\,(0{\cdot}40) + 0{\cdot}25\,(0{\cdot}20) = 0{\cdot}1833.$$

This is an average of the individual ϕ's defined by (10), and now we shall compare it with some other measurements of association.

OTHER MEASUREMENTS OF ASSOCIATION

When there is true inbreeding in a population as shown in (5), Li & Horvitz (1953) have proposed a number of methods for estimating the inbreeding coefficient F. Since we wish to have a single index to measure the association between uniting gametes in a subdivided population, the same measures may be adopted. The simplest one is based on the total proportion of heterozygotes in the entire population. Let H_0 be the total proportion of heterozygotes in a random mating population and H_d be that in the subdivided population (7). Then an overall measurement of association may be taken as

$$\begin{aligned}\phi(12) &= \frac{H_0 - H_d}{H_0} = \frac{\sigma_1^2+\sigma_2^2+\sigma_3^2}{2(pq+pr+qr)} \\ &= \frac{\phi_1 p(1-p)+\phi_2 q(1-q)+\phi_3 r(1-r)}{p(1-p)+q(1-q)+r(1-r)}.\end{aligned} \tag{12}$$

It is seen that this is also a weighted average of the individual ϕ's. From the total population in our numerical example (lower right of Table 1) this measurement yields

$$\phi(12) = \frac{0{\cdot}64 - 0{\cdot}52}{0{\cdot}64} = \frac{0{\cdot}12}{0{\cdot}64} = 0{\cdot}1875.$$

Another simple measure, opposite in a sense to the one indicated above, is based on the various homozygote frequencies. Let x_{11} be the frequency of homozygote A_1A_1, etc., in an inbreeding population. Li (1953) noted that

$$\frac{x_{11}}{p} + \frac{x_{22}}{q} + \frac{x_{33}}{r} = 1 + 2F.$$

In the case of a subdivided population, we have, by substituting (7) and (10) in the expression above and replacing F by ϕ,

$$\begin{aligned}\phi(13) &= \tfrac{1}{2}[(\phi_1 + \phi_2 + \phi_3) - (\phi_1 p + \phi_2 q + \phi_3 r)] \\ &= \tfrac{1}{2}[\phi_1(1-p) + \phi_2(1-q) + \phi_3(1-r)] \qquad (13)\end{aligned}$$

which is another weighted mean of the individual ϕ's. Applied to our numerical example it gives

$$\phi(13) = \tfrac{1}{2}[0{\cdot}5833 - 0{\cdot}1833] = 0{\cdot}2000.$$

Table 1. *Gene frequencies and their variances among five random mating isolates of equal size* ($w_i = \frac{1}{5}$).

Gene frequency		Isolates (1)	(2)	(3)	(4)	(5)	Total population	variance of gene freq.	ϕ (F analogue)
A_1	p_i	0·20	0·30	0·40	0·50	0·60	$p = 0{\cdot}40$	$\sigma_1^2 = 0{\cdot}02$	$\phi_1 = 0{\cdot}0833$
A_2	q_i	0·70	0·60	0	0·40	0·30	$q = 0{\cdot}40$	$\sigma_2^2 = 0{\cdot}06$	$\phi_2 = 0{\cdot}2500$
A_3	r_i	0·10	0·10	0·60	0·10	0·10	$r = 0{\cdot}20$	$\sigma_3^2 = 0{\cdot}04$	$\phi_3 = 0{\cdot}2500$
Total		1	1	1	1	1	1	0·12	0·5833

Genotype frequencies of random mating isolates and the total population

(1)	A_1	A_2	A_3		(2)	A_1	A_2	A_3		(3)	A_1	A_2	A_3	
A_1	0·04	0·14	0·02	0·20		0·09	0·18	0·03	0·30		0·16	0	0·24	0·40
A_2	0·14	0·49	0·07	0·70		0·18	0·36	0·06	0·60		0	0	0	0
A_3	0·02	0·07	0·01	0·10		0·03	0·06	0·01	0·10		0·24	0	0·36	0·60
(4)					(5)					Total				
A_1	0·25	0·20	0·05	0·50		0·36	0·18	0·06	0·60		0·18	0·14	0·08	0·40
A_2	0·20	0·16	0·04	0·40		0·18	0·09	0·03	0·30		0·14	0·22	0·04	0·40
A_3	0·05	0·04	0·01	0·10		0·06	0·03	0·01	0·10		0·08	0·04	0·08	0·20

We may also adopt a 'least square' procedure by regarding the frequencies in the subdivided population (7) as the 'observed' values and the frequencies in (5) as the 'expected' values (F replaced by ϕ). Then the sum of squares of 'deviations' is

$$Q = [p(1-p)(\phi_1 - \phi)]^2 + [2\sigma_{12} + 2\phi pq]^2 + \ldots,$$

where $2\sigma_{12} = \phi_3 r(1-r) - \phi_1 p(1-p) - \phi_2 q(1-q)$ by (8) and (10). Setting $dQ/d\phi = 0$ and solving, we obtain

$$\phi(14) = \frac{C_1\phi_1 + C_2\phi_2 + C_2\phi_3}{C_1 + C_2 + C_3}, \qquad (14)$$

where

$$C_1 = p(1-p)[3p(1-p)-2qr],$$

$$C_2 = q(1-q)[3q(1-q)-2pr],$$

$$C_3 = r(1-r)[3r(1-r)-2pq],$$

$$C_1+C_2+C_3 = 6(p^2q^2+p^2r^2+q^2r^2)+2pqr.$$

In our numerical example, $C_1 = C_2 = 0{\cdot}1344$, $C_3 = 0{\cdot}0256$, and

$$\phi(14) = \frac{0{\cdot}0512}{0{\cdot}2944} = 0{\cdot}1739.$$

Finally, the unweighted mean of the ϕ's may also serve as an overall index for the degree of association.

$$\phi(15) = \tfrac{1}{3}(\phi_1+\phi_2+\phi_3). \tag{15}$$

It equals $0{\cdot}5833/3 = 0{\cdot}1944$ in our example. Note that index $\phi(13)$ is a type of combination of indices $\phi(11)$ and $\phi(15)$. It is not entirely without justification to use the unweighted mean. When the three alleles are regarded as two alleles, A_1 and (A_2+A_3), the 'inbreeding coefficient' in the total population would be ϕ_1. Similarly, for the cases A_2 vs. (A_1+A_3) and A_3 vs. (A_1+A_2), the inbreeding coefficients would be ϕ_2 and ϕ_3, respectively. If these three pooling systems are equally important with respect to the total population, then the simple average (15) would be a good measure of the over-all association.

DISCUSSION

The main point of this communication is that when a population is subdivided into a number of random mating isolates with respect to multiple alleles, each allele frequency has its own variance among the isolates and there is no unique index or coefficient to measure the degree of association between the uniting gametes as caused by the subdivision of the population. The basic difficulty is the lack of a numerical scale that is applicable to the alleles. Without a quantitative scale, no correlation coefficient between the uniting gametes can be calculated. Consequently, a number of simple indices or coefficients have been mentioned. This situation is analogous to the ordinary two-way table with qualitative classifications, for which there is also no unique and universally agreed upon measure of association.

In the example shown in Table 1, the first three indices of association are

$$\phi(11) = 0{\cdot}1833, \quad \phi(12) = 0{\cdot}1875, \quad \phi(13) = 0{\cdot}2000.$$

This is not to be interpreted as meaning $\phi(11) < \phi(12) < \phi(13)$ in general. Their magnitudes depend on how the alleles are distributed among the isolates. Nor is the covariance of the frequencies of two alleles necessarily negative. In our numerical example (Table 1),

$$\sigma_{12} = -0{\cdot}02, \quad \sigma_{13} = 0, \quad \sigma_{23} = -0{\cdot}04,$$

a covariance could well be positive, as shown in the following isolates.

Gene frequency		Isolates (1)	(2)	(3)	(4)	(5)	Total population	Variance	Covariance
A_1	p_i	0·10	0·20	0·30	0·40	0·50	$p = 0{\cdot}30$	$\sigma_1^2 = 0{\cdot}020$	$\sigma_{12} = +0{\cdot}010$
A_2	q_i	0·10	0·15	0·20	0·25	0·30	$q = 0{\cdot}20$	$\sigma_2^2 = 0{\cdot}005$	$\sigma_{13} = -0{\cdot}030$
A_3	r_i	0·80	0·65	0·50	0·35	0·20	$r = 0{\cdot}50$	$\sigma_3^2 = 0{\cdot}045$	$\sigma_{23} = -0{\cdot}015$

The corresponding indices of association in the total population are:

$$\phi(11) = 0{\cdot}1248, \quad \phi(12) = 0{\cdot}1129, \quad \phi(13) = 0{\cdot}0908.$$

It is seen that $\phi(11) > \phi(12) > \phi(13)$ in this example. At the present stage, the writer has no particular preference for any one of the measurements proposed in this communication. It is hoped that further investigation will enable us to choose from the existing ones or to construct new indices that would reflect the nature and degree of the association accurately.

Nei (1965) investigated a slightly different problem. He assumed inbreeding for multiple alleles in each subpopulation with a single inbreeding coefficient $f^{(i)}$ for all alleles in the ith isolate. This is correct for true inbreeding. Under 'Discussion', however, he noted: 'Another factor which complicates the situation is the fact that $f^{(i)}$ is not necessarily the same for all genotypes under non-random differentiation. In other words, the value of $f^{(i)}$ for A_jA_j may not be the same as that for A_kA_k or A_jA_k.' This is the problem discussed in the present communication.

Likewise, Jain & Workman (1967) have also considered multiple alleles with inbreeding and selection and use F exclusively as the inbreeding coefficient or the fixation index and do not discuss the difference between inbreeding and subdivision for multiple alleles. Hence it is thought better to introduce a different symbol ϕ for subdivision to be distinguished from the inbreeding coefficient F.

SUMMARY

With multiple alleles, the situation for population subdivision is no longer identical with that of inbreeding. In the latter case, all heterozygote frequencies are decreased to the same extent. In the former, a heterozygote frequency may be decreased or increased, or remains the same as that of a random mating population without subdivision, as the covariance of the frequencies of the Alleles A_i and A_j may be negative, positive, or zero. Also, no correlation coefficient can be calculated for the case of population subdivision, as no natural numerical values can be assigned to the alleles A_i. Several simple indices have been proposed to serve as an overall measurement of the degree of association between the uniting gametes and illustrated by numerical examples. The isomorphism of this problem with that of association for qualitative traits has been pointed out.

References

Jain, S. K. & Workman, P. L. (1967). Generalized F-statistics and the theory of inbreeding and selection. *Nature, Lond.* **214**, 674–678.

Li, C. C. (1953). On an equation specifying equilibrium populations. *Science N.Y.* **117**, 378–379.

Li, C. C. (1955). *Population Genetics*; Chapter 21, Subdivision and migration. University of Chicago Press, Chicago, Ill.

Li, C. C. (1961). *Human Genetics, Principles and Methods*. New York: McGraw-Hill.

Li, C. C. & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* **5**; 107–117.

Li, C. C. and Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**, 347–360.

Nei, M. (1965). Variation and covariation of gene frequencies in subdivided populations. *Evolution* **19**, 256–258.

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelations-erscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* **61**, 330–338.

Wright, S. (1931). Evolution in mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

Yasuda, N. (1968). An extension of Wahlund's principle to evaluate mating type frequency. *Am. J. Hum. Genet.* **20**, 1–23.