Heredity (1981), **47** (2), 259-262 0018-067X/81/03450259$02.00

© 1981. The Genetical Society of Great Britain

MEAN AND VARIANCE OF THE NUMBER OF SAMPLES SHOWING HETEROZYGOTE EXCESS OR DEFICIENCY

PARTHA P. MAJUMDER and RANAJIT CHAKRABORTY

Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, Texas 77025, U.S.A.

Received 11.ii.1981

IN the study of population structure from data on gene frequencies, it is often necessary to examine deviations of observed proportions of heterozygotes from the corresponding Hardy–Weinberg proportions. Specifically, if we have gene frequency data at a single codominant di-allelic locus from a large number of random-mating populations, we ask the question: In the absence of disturbing evolutionary forces, how often should one expect the observed proportions of heterozygotes to deviate above (or below) the expectations under Hardy–Weinberg equilibrium? Usually, it is stated, without any formal justification, that this expectation is 50 per cent (Avise and Felley, 1979). We have examined the validity of this statement from a theoretical standpoint. It is worth mentioning also that the related problem of testing for significance of deviations from Hardy–Weinberg expected frequencies has very little statistical power (Lewontin and Cockerham, 1959; Ward and Sing, 1970; Chakraborty and Rao, 1972; Haber, 1980).

Suppose n diploid individuals are sampled from a population with gene frequencies p and q of the two alleles A and a at the di-allelic locus under consideration. Then, under the assumption of random-mating and in the absence of any "disturbing evolutionary forces", the expected proportions of AA, Aa and aa individuals are given by: p^2, $2pq$ and q^2, respectively, or by $p(2np-1)/(2n-1)$, $4npq/(2n-1)$ and $q(2nq-1)/(2n-1)$, respectively, if Levene's (1949) corrections are made. [Note that, following the usual convention, in practice the expected frequencies are calculated using the observed sample gene frequencies.] Suppose m independent samples of n individuals are drawn from this population. Then, we wish to know in how many of these m samples will there be an excess of heterozygote individuals over that expected under Hardy–Weinberg equilibrium. It is important to note here that the m independent samples are not drawn from the same population in practice, but are generally drawn from m different subpopulations (which may have different underlying gene frequencies). We shall deal with this more realistic situation later.

In a single random sample of n individuals, let n_1, n_2 and n_3 denote, respectively, the observed number of AA, Aa and aa individuals. We have to compute:

$$P(n_2 > E[Aa] \| p, q), \qquad (1)$$

where $E[Aa]$ is the expected number of Aa individuals under Hardy–Weinberg equilibrium, which, given n_1, n_2 and n_3, is computed as: $2(n_1+\frac{1}{2}n_2)(n_3+\frac{1}{2}n_2)/n$ [or, as: $(2n_1+n_2)(2n_3+n_2)/(2n-1)$, if Levene's corrections are made]. Now, given the gene frequencies, one can compute the probability of a particular outcome as:

$$\frac{n!}{n_1!n_2!n_3!}(p^2)^{n_1}(2pq)^{n_2}(q^2)^{n_3}. \qquad (2)$$

In order to obtain the required probability, one can compute all possible 3-partitions of n (for a fixed n) and add the probabilities (computed by (2)) for all those partitions satisfying (1). In table 1 are given the probabilities of excess heterozygotes for various values of p and n. From this table it is seen that as n increases the probability of excess heterozygotes in the sample converges to 0·50, but the rate of convergence is dependent on the underlying gene frequency in the population. Generally speaking, if the underlying population gene frequency is not extreme, then the rate of convergence is faster, and even in samples of moderate size one can expect the probability of observing excess heterozygotes to be about 0·50. If, however, the true gene frequency in the population has an extreme value this probability could be very different from 0·50.

The probability $P(n_2>E[Aa]|p, q)$ is also a symmetric function of the underlying population gene frequency (p) around $p=0{\cdot}5$. Furthermore, since $E(n_2-E[Aa])$ over all partitions of n equals zero, the average positive deviation, *i.e.*, $E(n_2-E[Aa]|n_2>E[Aa])$, is smaller than the average absolute negative deviation, $E(|n_2-E[Aa]|$ given $n_2<E[Aa])$, wherever $P(n_2>E[Aa]|p, q)>0{\cdot}50$. In other words, when gene frequencies are extreme, positive deviations, while very common, are on the average smaller in magnitude compared with their negative counterparts.

In practice, in m samples from m different populations with possibly very different true gene frequencies, finding the expected number of samples showing excess number of heterozygotes is difficult. If, however, the estimated frequencies of the gene in the m populations are "moderate", and if the sample sizes on the basis of which these gene frequencies have been estimated are rather "large", then one might expect $m/2$ of these populations to show an excess number of heterozygotes even if the populations from which these samples were drawn are panmictic. On the other hand, if one or more of the estimated gene frequencies in the populations turn out to be extreme, then one has to compute the required probability of heterozygote excess by complete enumeration (as we have done for constructing table 1), assuming that the estimated gene frequencies are the true population values. But, in such a case the role of individual sample sizes drawn from the populations becomes extremely important.

In the ith population, let H_i and $\hat{H}_i$ denote the expected and observed proportion of heterozygotes. Also, let $P_i=\text{Prob}(\hat{H}_i>H_i)$. Then the number of populations (N) in which the observed heterozygosity exceeds the expected is:

$$E(N)=\sum_{i=1}^{m} P_i.$$

Since all the m samples are independent,

$$V(N)=\sum_{i=1}^{m} P_i(1-P_i).$$

TABLE 1

Probabilities of excess heterozygotes for various values of gene frequency* (p) *and sample size* (n)

p	n = 50	100	200	250	500
0·05	0·84	0·78	0·62	0·56	0·55
0·10	0·63	0·54	0·53	0·52	0·52
0·20	0·53	0·50	0·51	0·51	0·50
0·30	0·53	0·51	0·51	0·50	0·50
0·40	0·50	0·50	0·50	0·50	0·50
0·50	0·47	0·48	0·49	0·49	0·50
0·60	0·50	0·50	0·50	0·50	0·50
0·70	0·53	0·51	0·51	0·50	0·50
0·80	0·53	0·50	0·51	0·51	0·50
0·90	0·63	0·54	0·53	0·52	0·53
0·9999	0·00005	0·0002	0·0008	0·001	0·005

* Calculated using Levene's corrections.

The detection of significance of deviation of the observed excess over the expected can easily be done by a standard goodness-of-fit chi-square test, which in the present case has 1 degree of freedom.

An example:

Gershowitz *et al.* (1972) have presented data on *MN* blood group frequencies of Yanomama Indians living in 37 separate villages located in Brazil and Venezuela (table 1 of their paper). For this body of data, we calculated the P_i values ($i=1,2,\dots,37$) by the method of complete enumeration as described above. In this case, $E(N)=\sum_{i=1}^{37} P_i=18{\cdot}505$. The observed value of N is 12. The χ^2 value for testing the significance of deviation of the observed from the expected turned out to be 2·287, which is insignificant at the 5 per cent level with 1 df. Thus, no significant excess of heterozygotes could be detected at the *MN* locus for the 37 Yanomama villages.

Acknowledgements.—We are grateful to Dr N. Ryman for bringing this problem to our notice and to Dr J. Felsenstein for comments on an earlier draft of the paper. The work is supported by U.S. Public Health Service Grants GM20293 and CA19311 from U.S. National Institutes of Health.

REFERENCES

AVISE, J. C., AND FELLEY, J. 1979. Population structure of freshwater fishes. I. Genetic variation of bluegill (*Lepomis macrochirus*) populations in man-made reservoirs. *Evolution, 33*, 15-26.

CHAKRABORTY, R., AND RAO, D. C. 1972. Detection of inbreeding coefficient from ABO blood-group data. *American Journal of Human Genetics, 24*, 352-354.

GERSHOWITZ, H., LAYRISSE, M., LAYRISSE, Z., NEEL, J. V., CHAGNON, N., AND AYRES, M. 1972. The genetic structure of a tribal population, the Yanomama Indians. II. Eleven blood-group systems and the ABH-Le secretor traits. *Annals of Human Genetics, 35*, 261-269.

HABER, M. 1980. Detection of inbreeding effects by the χ^2 test on genotypic and phenotypic frequencies. *American Journal of Human Genetics, 32*, 754-760.

LEVENE, H. 1949. On a matching problem arising in genetics. *Annals of Mathematical Statistics*, *20*, 91-94.

LEWONTIN, R. C., AND COCKERHAM, C. C. 1959. The goodness-of-fit test for detecting natural selection in random mating populations. *Evolution*, *13*, 561-564.

WARD, R. H., AND SING, C. F. 1970. A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. *American Naturalist*, *104*, 355-366.