OPERATING CHARACTERISTICS OF THE TEST FOR HARDY-WEINBERG EQUILIBRIUM: An M.S. Thesis Problem

BU-545-M

by

December, 1974

D. S. Robson

Abstract

With the development of serum protein analysis and allozyme analysis the statistical test for Hardy-Weinberg equilibrium in natural populations is seeing increased usage, frequently employing sample sizes which are only marginally adequate for the intended purpose. A need exists for user-oriented power curves, preferably transformed into sample size requirements, to summarize the operating characteristics of this statistical test procedure.

Paper No. BU-545-M in the Biometrics Unit Mimeo Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

OPERATING CHARACTERISTICS OF THE TEST FOR HARDY-WEINBERG EQUILIBRIUM: An M.S. Thesis Problem

BU-545-M

by

December, 1974

D. S. Robson

The recent development of allozyme (isozyme) analysis as a laboratory technique for identifying distinguishable phenotypes at the enzymatic level has led to increased application of statistical tests for Hardy-Weinberg equilibrium in natural populations. When genotypic frequencies in a field sample conform to the Hardy-Weinberg law, this is taken as evidence in support of the hypothesis that the organisms in the sample did in fact come from a single, interbreeding population. Such evidence is used, for example, as an aid in demarking the geographic boundaries of distinct stocks of an exploited animal population. In circumstances where the population structure is known a priori, departures from fit to the Hardy-Weinberg law serve as the starting point for investigating the causes of disequilibrium, such as differential fitness of the several genotypes.

This newly extended use of the Hardy-Weinberg test statistic frequently places excessive demands on the test, particularly in those cases where acceptance of the null hypothesis of equilibrium is used, in effect, as positive evidence of stock purity. Small sample sizes are not uncommon in these applications, due both to the difficulties in collecting field samples and to the expense of the laboratory analyses. A need therefore exists for a comprehensive study of the small sample operating characteristics of this statistical test to provide guidance to users.

Paper No. BU-545-M in the Biometrics Unit Mimeo Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

The simplest application of the test arises in the case of two identifiable alleles, when a sample of size n is observed to consist of phenotypes AA, AB, and BB with frequencies N_{AA}, N_{AB}, and N_{BB}, respectively, $N_{AA} + N_{AB} + N_{BB} = n$. At equilibrium the relative frequencies in the population are p_A^2, $2p_Ap_B$, p_B^2, respectively, where p_A and p_B are the relative frequencies of the two alleles in the population, $p_A + p_B = 1$. The likelihood of the sample is then

$$P\left(N_{AA} = n_{AA}, N_{AB} = n_{AB}, N_{BB} = n_{BB} \middle| N_{AA} + N_{AB} + N_{BB} = n\right)$$

$$= \frac{n!}{n_{AA}!n_{AB}!n_{BB}!} (p_A^2)^{n_{AA}} (2p_Ap_B)^{n_{AB}} (p_B^2)^{n_{BB}}$$

where $p_A = 1 - p_B$ is an unknown parameter. The statistic

$$T = 2N_{AA} + N_{AB}$$

is (minimal) sufficient with respect to this model, and the test of fit to the model is then constructed to have size α with respect to the conditional distribution of the sample configuration (N_{AA}, N_{AB}, N_{BB}) given the value of the sufficient statistic T. Thus, the critical region consists of the two tails of the conditional probability distribution

$$P\left(N_{AB} = n_{AB} \middle| N_{AA} + N_{AB} + N_{BB} = n,\ 2N_{AA} + N_{AB} = t\right).$$

This distribution is easily derived by noting that

$$P(T = t|n) = \binom{2n}{t} p_A^t p_B^{2n-t}$$

giving

$$P(N_{AB} = n_{AB}|n,t) = \frac{n!}{n_{AA}!n_{AB}!n_{BB}!} \frac{2^{n_{AB}}}{\binom{2n}{t}}, \quad \begin{cases} 2n_{AA} + n_{AB} = t \\ 2n_{BB} + n_{AB} = 2n-t \end{cases}$$

with mean

$$\mu_{AB \cdot t} = \frac{t(2n-t)}{2n-1}$$

and variance

$$\sigma^2_{AB \cdot t} = \frac{2}{2n-3} \mu_{AB \cdot t}(\mu_{AB \cdot t} - 1).$$

In practice this distribution is usually approximated by a normal density function, producing the one degree of freedom chi-square approximation for

$$\chi^2 = \frac{(n_{AB} - \mu_{AB \cdot t})^2}{\sigma^2_{AB \cdot t}} .$$

Alternatives to this equilibrium model may be conveniently and meaningfully parameterized in terms of "coefficients of selection"; e.g.,

$$P_n(n_{AA}, n_{AB}, n_{BB}) = \frac{n!}{n_{AA}!n_{AB}!n_{BB}!} \left[p_A^2(1 - s_A)\right]^{n_{AA}} \left[2p_A p_B\right]^{n_{AB}} \left[p_B^2(1 - s_B)\right]^{n_{BB}} ,$$

where the coefficients of selection s_A and s_B are familiar parameters to the user. Graphs and tables of both conditional (on T) and unconditional power of this test are needed to provide a clear picture of its operating characteristics.

Refinements on this test as well as extensions to multiple alleles and multiple loci also need to be studied for their power characteristics. Refinements are required to enhance power against particular classes of alternatives; for example, when k different local populations have been independently sampled the investigator may wish to test equilibrium against the alternative that selection favors heterozygotes in this collection of stocks. A method is therefore required for efficiently combining the k one-tailed tests into a single test, and the power of such a test should likewise be investigated to provide guidance in sample size determination.