Genet. Res., Camb (1994), **64**, *pp.* 71–74

SHORT PAPER

An exact test for neutrality based on the Ewens sampling distribution

MONTGOMERY SLATKIN

Department of Integrative Biology, University of California, Berkeley, California 94720

(*Received 3 May 1994 and in revised form 1 June 1994*)

Summary

Using the Ewens sampling distribution of selectively neutral alleles in a finite population, it is possible to develop an exact test of neutrality by finding the probability of each configuration with the same sample size and observed number of allelic classes. The exact test provides the probability of obtaining a configuration with the same or smaller probability as the observed configuration under the null hypothesis. The results from the exact test may be quite different from those from the Ewens–Watterson test based on the homozygosity in the sample. The advantages and disadvantages of using an exact test in this and other population genetic contexts are discussed.

1. Introduction

Kimura (1968) proposed the neutral theory of molecular evolution in part to account for the extensive variability found in natural populations with the then recent application of electrophoretic methods. Soon after, there was a rich development of the population genetics theory designed to test the neutral theory using electrophoretic data. One class of tests was based on the work of Ewens (1972), who derived the sampling distribution of neutral alleles in a finite population. Watterson (1977) used Ewens' theory to propose a test of neutrality based on the observed homozygosity in the sample, and that test is now called homozygosity test or the 'Ewens–Watterson test' (Slatkin, 1982; Hartl & Clark, 1989). In this note I will show that Ewens' (1972) theory can also be the basis for an exact test of neutrality and that the exact test has somewhat different properties than the homozygosity test. I will also discuss the use of other exact tests in population genetics. The increased availability of fast computers and the development of new algorithms make exact tests much easier to use, but their use may not always be as appropriate as the name 'exact test' suggests.

2. Fisher's exact test

The statistical test described below was developed by analogy with Fisher's exact test for associations in an $r \times c$ contingency table (Weir, 1990). I will briefly review Fisher's exact test for a 2×2 table in order to make clear the relationship between it and the exact test proposed below. For a given set of entries in a 2×2 table, the set of 4 numbers $\{n_{11}, n_{12}, n_{21}, n_{22}\}$ is the observed 'configuration' of the table. The marginal sums are the numbers in each row ($n_{1\cdot} = n_{11}+n_{12}, n_{2\cdot} = n_{21}+n_{22}$) and the numbers in each column ($n_{\cdot 1} = n_{11}+n_{21}, n_{\cdot 2} = n_{12}+n_{22}$) and the total number of observations is $n = n_{1\cdot}+n_{2\cdot} = n_{\cdot 1}+n_{\cdot 2}$. Under the assumption that the entries of the table are randomly chosen but subject to the constraint that the marginal sums are fixed, the probability of obtaining any configuration, say (m_{ij}) is the hypergeometric distribution

$$P(m_{11}, m_{12}, m_{21}, m_{22}) = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!m_{11}!m_{12}!m_{21}!m_{22}!}.$$

Fisher's exact test gives the probability that a randomly generated configuration has a probability under the hypergeometric distribution equal to or less than the probability of the observed configuration.

In practice, there are two ways to carry out Fisher's exact test. The first is to generate all possible configurations, and then test each one to determine whether the hypergeometric probability is less than or equal to that for the observed configuration. The task can be time-consuming, even for fast computers, but new algorithms have been developed to make the process more efficient (e.g. Mehta & Patel, 1983). The alternative is to generate randomly a large number of configurations and then count the proportion with probabilities equal to or less then the probability for the observed configuration. That is the basis for both resampling and Markov chain algorithms that can in principle analyse any table (Guo & Thompson, 1992).

Table 1. *Illustration of all configurations for n = 16 and k = 7 and the calculations for the exact and homozygosity tests*

j	c_j	$\Pr(c_j \mid k)$	*F*
1	(3, 3, 2, 2, 2, 2, 2)	0·00111*	0·1484†
2	(3, 3, 3, 2, 2, 2, 1)	0·00986*	0·1562†
3	(3, 3, 3, 3, 2, 1, 1)	0·00986*	0·1641†
4	(4, 2, 2, 2, 2, 2, 2)	0·00042*	0·1562†
5	(4, 3, 2, 2, 2, 2, 1)	0·01664*	0·1641†
6	(4, 3, 3, 2, 2, 1, 1)	0·06658	0·1719†
7	(4, 3, 3, 3, 1, 1, 1)	0·01973*	0·1797†
8	(4, 4, 2, 2, 2, 1, 1)	0·02497*	0·1797†
9	(4, 4, 3, 2, 1, 1, 1)	0·06658	0·1875†
10	(4, 4, 4, 1, 1, 1, 1)	0·00832*	0·2031†
11	(5, 2, 2, 2, 2, 2, 1)	0·00399*	0·1797†
12	(5, 3, 2, 2, 2, 1, 1)	0·05326	0·1875†
13	(5, 3, 3, 2, 1, 1, 1)	0·07101	0·1953†
14	(5, 4, 2, 2, 1, 1, 1)	0·07989	0·2031†
15	(5, 4, 3, 1, 1, 1, 1)	0·05326	0·2109†
16	(5, 5, 2, 1, 1, 1, 1)	0·03196*	0·2266†
17	(6, 2, 2, 2, 2, 1, 1)	0·01664*	0·2109†
18	(6, 3, 2, 2, 1, 1, 1)	0·08877	0·2188†
19	(6, 3, 3, 1, 1, 1, 1)	0·02959*	0·2266†
20	(6, 4, 2, 1, 1, 1, 1)	0·06658	0·2344†
21	(6, 5, 1, 1, 1, 1, 1)	0·02130*	0·2578†
22	(7, 2, 2, 2, 1, 1, 1)	0·03804	0·2500†
23	(7, 3, 2, 1, 1, 1, 1)	0·07609	0·2578†
24	(7, 4, 1, 1, 1, 1, 1)	0·02283*	0·2734†
25	(8, 2, 2, 1, 1, 1, 1)	0·04993	0·2969†
26	(8, 3, 1, 1, 1, 1, 1)	0·02663*	0·3047†
27	(9, 2, 1, 1, 1, 1, 1)	0·03551*	0·3516†
28	(10, 1, 1, 1, 1, 1, 1)	0·01065*	0·4141

In this table, *j* is the number of the configuration in the order generated by the recursive algorithm described in the text. $\Pr(c_j \mid k)$ is the probability of obtaining c_j according to Equation (1) in the text and *F* is the computed heterozygosity from Equation (3) in the text. The (hypothetical) observed configuration, c_o, (9, 2, 1, 1, 1, 1, 1), is c_{27}. The asterisk (*) indicates that the $\Pr(c_j \mid k) \leqslant \Pr(c_o \mid k) = 0{\cdot}03551$ and the dagger (†) indicates that $F(c_j) \leqslant F(c_o) = 0{\cdot}3516$. For this c_o, the tail probability from the exact test is $P_E = 0{\cdot}29001$ (from Equation (2) in the text) and the tail probability from the homozygosity test is $P_H = 0{\cdot}98935$ (from Equation (4) in the text).

3. Ewens sampling distribution

Ewens (1972) showed that in a sample of *n* copies of a locus, the observed number of allelic classes, *k*, is a sufficient statistic to estimate the parameter $\theta = 4N\mu$, where *N* is the number of diploid individuals in a randomly mating population and μ is the rate of mutation. He showed further that for a given *k*, the probability of obtaining a particular configuration of alleles in the sample depends only on *k* and *n*. In this case, a configuration is the set of numbers of alleles in each class arranged in non-increasing order. For a configuration $c = \{r_i\} (i = 1, \ldots, k; r_1 \geqslant r_2 \geqslant \ldots \geqslant r_k \geqslant 1)$ the probability of obtaining that configuration is

$$\Pr\{r_i \mid k\} = \frac{n!}{|S_n^k| \, 1^{\alpha_1} 2^{\alpha_2} \mathrm{K} n^{\alpha_n} \alpha_1! \, \alpha_2 \, \mathrm{K} \alpha_n !}, \qquad (1)$$

where α_i is the number of values in the set $\{r_i\}$ equal to *i* and S_n^k is the Stirling number of the first kind (Ewens, 1979, Eq. 3.78). To simplify the notation, I will denote $\Pr\{r_i \mid k\}$ by $\Pr(c \mid k)$.

Ewens (1972) derived this distribution under the assumption that the population followed the Wright–Fisher model of reproduction, but later theory, summarized by Ewens (1979), has shown that (1) holds under more general assumptions about reproduction as well, as long as the allelic classes are selectively equivalent. Besides the assumption of neutrality, (1) depends on the assumption of the infinite alleles model of mutation and the assumption that the population is of constant size.

Given (1) it is easy to describe the exact test. Let *C* be the total number of configurations for a given *n* and *k*, and let $c_j, j = 1, \ldots, C$, be the *j*th configuration. I will show later how to compute *C* and assign numbers to the configurations. For each c_j, (1) gives us the probability of obtaining that configuration $\Pr(c_j \mid k)$. Let c_o be the observed configuration. By analogy with Fisher's exact test for an $r \times c$ table, we can find the probability that each configuration, *c*, has a probability of occurrence equal to or less than $\Pr(c_o \mid k)$. That is, we can define

$$P_E = \sum_{c_j \ni \Pr(c_j \mid k) \leqslant \Pr(c_o)} \Pr(c_j \mid k), \qquad (2)$$

where the subscript *E* denotes the exact test, which will later be contrasted with the homozygosity test.

The cumulative probability, P_E, provides the basis for the exact test of the null hypothesis. We could, for example, choose a significance level γ and say that we will reject the null hypothesis (that the sample was drawn from a neutral locus following the infinite alleles mutation model in a population of constant size) if $\gamma/2 < P_E < 1 - \gamma/2$ (a two-tailed test).

We can compare the performance of the exact test with the homozygosity test. For each configuration, we can define the computed homozygosity,

$$F(c) = \sum_{i=1}^{k} r_i^2 / n^2. \qquad (3)$$

In this notation, the homozygosity test proposed by Watterson (1977) is based on computing the cumulative probability P_H, defined to be

$$P_H = \sum_{c_j \ni F(c_j) \leqslant F(c_o)} \Pr(c_j \mid k). \qquad (4)$$

4. Generating configurations

All that is needed to apply either of these tests to data is a method for generating configurations. As in the case of Fisher's exact test, there are two possibilities. One could use a Monte Carlo method to generate random configurations that follow (1). Each randomly generated configuration would be tested to determine whether it satisfied the necessary criterion (either

$\Pr(c \mid k) \leqslant \mathrm{P}(c_o \mid k)$ or $F(c) \leqslant F(c_o)$) and then the results would be accumulated to find the proportion of all randomly generated configurations that pass the test. That is how the homozygosity test has been applied (Watterson, 1978; Fuerst *et al.* 1977; Hartl & Clark, 1989, p. 141). The second possibility is to generate all configurations and compute the sums in (2) and (4).

I have developed a recursive algorithm to count and generate all configurations for a given n and k. I found that there were not as many configurations as my intuition suggested, so it is possible to examine all configurations even for realistic sample sizes. For example, with $n = 89$ and $k = 15$, as in the data set of Keith *et al.* (1985), there are 3014304 different configurations. Although that is too many to process by hand, they can be examined by even a small computer in a few minutes.

I have written a *C* program that implements this algorithm and found that it provides answers in a reasonable time even for moderately large sample sizes ($n \leqslant 100$). The algorithm is based on being able to count and generate all configurations for $k = 2$. For $k = 2$, there are $[n/2]$ distinct configurations, where [.] indicates the largest integer equal to or less than the quantity enclosed (i.e. $[13/2] = [12/2] = 6$). Those configurations can be easily generated: $\{[(n+1)/2], [n/2]\} \ldots \{n-1, 1\}$. For a given n and $k > 2$, the program then generates all configurations permissible for each possible value of r_1 (i.e. all configurations with $r_2 \leqslant r_1$), with the sample size for those configurations equal to $n - r_1$ and $k = k - 1$, and the process continues until the $k = 2$ case is reached. The program starts with the most even configuration and works upwards to larger values of r_1. An example is given in Table 1. I will distribute copies of this program, which runs under UNIX and on a PC, without cost. Please send inquiries about the program by electronic mail to me at monty@kaline.berkeley.edu.

There is no reason to think that this algorithm is optimal, and more efficient ones can probably be found. This algorithm does work and the program that implements it provides a strong verification that it works correctly. In Equation (1), the values of $n!$ and S_n^k are the same for every configuration with a given n and k, so it is not necessary to calculate them. Instead, the program computes the rest of (1) for each configuration and sums those values. The results are then normalized by dividing by this sum, which avoids the ugly problem of computing the Stirling number. In numerous tests, I found that the resulting sum was equal to $S_n^k/n!$, which ensures that all configurations were counted and that the probabilities were correctly computed.

5. The exact vs. the homozygosity test

The exact test may give quite different results than the homozygosity test and it is important to understand why. Consider, for example, the data of Keith *et al.* (1985) for the Xdh locus in a California population of *Drosophila pseudoobscura*. Hartl & Clark (1989, p. 141) use this data set as the example of the application of the homozygosity test. In this data set, $n = 89, k = 15$, and the observed configuration is (52, 9, 8, 4, 4, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1). My program gave values of $P_E = 0{\cdot}7198$ and $P_H = 0{\cdot}9910$. The value of P_H is consistent with the results from the Monte Carlo analysis described by Hartl & Clark (1989) and indicates that the homozygosity in this sample is much larger than would be expected in a random draw from the Ewens sampling distribution. Hence, the homozygosity test allows us to reject the null hypothesis. The observed configuration is too 'uneven' or there are 'too many rare alleles'. The value of P_E indicates, however, that the observed configuration is not especially unlikely, so the exact test does not allow us to reject the null hypothesis. This situation is similar to the one illustrated in Table 1, for a much smaller (hypothetical) data set, and we can see why the tests may give different results. If $k > 2$, there is not necessarily a close relationship between the evenness of the distribution as measured by F or other criteria and the probability of its occurrence, so the two tests are focusing attention on different regions of the sample space. Alternatively, we can view $\Pr(c \mid k)$ as another test statistic, and one that is not necessarily highly correlated with F when $k > 2$.

6. Which test should we believe?

Watterson (1977) suggested the homozygosity test after he found that the quantity F arose naturally in the analysis of a model of overdominant selection, in which all heterozygotes have a fitness of $1 + s$ relative to all homozygotes. Overdominance would tend to create more even configurations (small F) and underdominance would tend to create more uneven configurations (large F). In a subsequent paper, Watterson (1978) showed that the homozygosity test had somewhat less power when testing for deleterious mutations but that F was still an appropriate test statistic. Thus the homozygosity test was designed to test against specific alternative hypotheses about the maintenance of genetic diversity. In contrast, the exact test does not test against a particular alternative hypothesis and instead uses the probability of each configuration to indicate which regions of the sample space allow the rejection of the null hypothesis.

It seems that the choice of which test to believe cannot be made on purely statistical grounds. Instead, the inherent biological plausibility of the alternative models must be taken into account. The highly symmetric models of selection examined by Watterson (1977, 1978) are plausible and simple representations of the biological intuition that, on average, heterozygotes are superior or that, on average, new mutants are deleterious. What is less clear is whether F would remain the appropriate test statistic for less

https://doi.org/10.1017/S0016672300032560 Published online by Cambridge University Press

symmetric models describing the same selection processes.

7. Other exact tests in population genetics

Fisher's exact test and its descendants carry an unfortunate name. It is difficult not to think that because a test is exact it is also better. Yet all properly defined tests are just as exact: the value of P_H is exactly the probability that the observed value of F is less than or equal to the value for a randomly generated configuration. The question is whether there are plausible and interesting alternative models that suggest useful test statistics. Several exact tests have been proposed for population genetics. Fisher's exact test for an $r \times c$ table has long been used to test for significant nonrandom association (i.e. linkage disequilibrium) between alleles at different loci (Weir, 1990), and I have argued that the tail probability from Fisher's exact test provides a useful way to characterize the extent of nonrandom association when there are more than two alleles per locus (Slatkin, 1994). A slight modification of Fisher's exact test can be used to test for significant deviations from Hardy–Weinberg genotypic proportions (Weir, 1990; Guo & Thompson, 1992). More recently, Raymond & Rousset (in prep.) have discussed the use of Fisher's exact test for testing for significant differentiation of local populations. Guo & Thompson (1992) provide an elegant algorithm based on a Markov chain that makes it possible to carry out Fisher's exact test for arbitrarily large tables, and Raymond & Rousset (1994) have implemented a variant of that algorithm in a program package that will perform the exact test for a variety of problems.

These theoretical developments are important and their implementation will be useful for population geneticists. But as the results presented here show, an exact test may not help identify deviations from the null hypothesis that are expected under biologically plausible conditions. For example, the value of Wright's inbreeding coefficient, F_{IS}, arises naturally in the analysis of neutral alleles in a self-fertilizing species. It may well be more appropriate to use F_{IS} as a test statistic when the interest is determining whether or not self-fertilization is occurring. A similar remark applies to F_{ST} as a measure of population differentiation. It arises naturally in models of differentiation caused by drift and gene flow and hence may be an appropriate test statistic when that is the alternative hypothesis of interest.

The situation is somewhat different for the study of linkage disequilibrium when there are more than two alleles per locus. At present there is no theory in which a particular test statistic arises naturally, and the most commonly used test statistics are proportional to the χ^2 statistic for the $r \times c$ table (Weir, 1990). But χ^2 just provides information that is equivalent to that in the Fisher's exact test for the same table, so at this point there seems no way to improve on the exact test for linkage disequilibrium.

8. Conclusions

My goal in writing this note was not to say that the exact test for neutrality is necessarily better than the homozygosity test or other tests of neutrality. Instead it was to point out that an exact test is computationally possible and in fact no harder to implement than the homozygosity test, and that its application raises important issues about how statistical tests are used in population genetics.

I thank F. Bonhomme, W. J. Ewens, L. Excoffier, and M. Raymond for helpful discussions of this topic. This research was supported in part by NIH grant GM 40282, and in part by CNRS while a guest in the laboratory of F. Bonhomme in Montpellier, France.

References

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Fuerst, P. A., Chakraborty, R. & Nei, M. (1977). Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* **86**, 455–483.

Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

Hartl, D. & Clark, A. G. (1989). *Principles of Population Genetics, Second Edition*. Sunderland, Mass.: Sinauer Associates.

Keith, T. P., Brooks, L. D., Lewontin, R. C., Martinez-Cruzado, J. C. & Rigby, D. L. (1985). Nearly identical distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. *Molecular Biology and Evolution* **2**, 206–216.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Mehta, C. R. & Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association, Series B* **78**, 427–434.

Raymond, M. & Rousset, F. (1994). GENEPOP (ver. 1.1), a population genetics software package for exact tests and oecumenism. *Journal of Heredity*, in press.

Slatkin, M. (1982). Testing neutrality in a subdivided population. *Genetics* **100**, 533–545.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.

Watterson, G. A. (1977). Heterosis or neutrality. *Genetics* **85**, 789–814.

Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland, Mass.: Sinauer Assoc.