SHORT PAPER

A correction to the exact test based on the Ewens sampling distribution

MONTGOMERY SLATKIN
Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

(*Received 4 July 1996 and in revised form 26 August 1996*)

Summary

The exact test for neutrality based on the Ewens sampling distribution described previously (Slatkin, 1994) is not correct. The problem is that the test as described is based on the probability of the ordered configuration of numbers of alleles, while it should be based on the probability of the unordered configuration. The correctly implemented exact test leads to results that are similar to those from the homozygosity test proposed by Watterson (1977) for relatively small sample sizes but can still differ substantially for larger sample sizes. Programs to perform the exact test are available from the author.

In an earlier paper, I described a test for selective neutrality based on the Ewens sampling distribution (Slatkin, 1994). I have realized that, although the test described in that paper is a legitimate statistical test, it does not have the properties of an exact test analogous to Fisher's exact test for an $r \times c$ contingency table. The test as described is based on a sample of n copies of a locus at which k different alleles are found with numbers $r_1, \ldots, r_k$. The probability of an ordered configuration, that is, one with $r_1 \geqslant r_2 \geqslant \ldots r_k \geqslant 1$, given the values of n and k, is

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

where α_i is the number of the r_j which take the value i and S_n^k is Stirling's number of the first kind (Ewens, 1972, 1979). Equation (1) is the correct version of Eq. (1) in Slatkin (1994), which contained typographical errors. The exact test as proposed summed the probabilities of all configurations for which $\Pr\{r_i \mid k\}$ is less than or equal to the probability of the observed configuration.

The problem with using (1) as the basis for the exact test is that there are different numbers of unordered configurations corresponding to each ordered configuration. The probability of each unordered configuration is obtained from (1) by removing the factorial terms in the numerator and denominator (Ewens, 1972; Stewart, 1977). The ratio of factorials in (1) is in fact the number of unordered configurations consistent with each ordered configuration. Because each unordered configuration represents a possible outcome of the random evolutionary process, the probability of an unordered configuration should be the basis for the exact test.

This change makes a substantial difference. To illustrate, consider the example in table 1 of Slatkin (1994) for which $k = 7$ and $n = 16$. Configuration 9 (c_9), (4, 4, 3, 2, 1, 1, 1), is consistent with $7!/(2!\,3!) = 420$ unordered configurations, while configuration 27 (c_{27}) (9, 2, 1, 1, 1, 1, 1) is consistent with $7!/5! = 42$ unordered configurations. Using Eq. (1), $\Pr(c_9) = 0{\cdot}06658 > \Pr(c_{27}) = 0{\cdot}03551$ and hence c_9 would seem to be more likely than c_{27}. For that reason c_9 was not included in the set of configurations less likely than c_{27}, the hypothetical observed configuration (see table 1, Slatkin, 1994). Yet each unordered configuration consistent with c_9 is less likely than each unordered configuration consistent with c_{27} because $0{\cdot}06658/420 < 0{\cdot}03551/42$. Therefore, in the correct exact test, c_9 would be included in the set of configurations which are less likely than the observed configuration. In fact, in the example in table 1 in Slatkin (1994), the correctly formulated exact test leads to the same result as the homozygosity test. Both tests give the same tail probability: $P_E = P_H = 0{\cdot}98935$.

For larger data sets, the correct exact test and the homozygosity test are not identical. In some cases I have examined, particularly those with relatively small values of n, they lead to nearly the same conclusions, but in others, with much larger values of n, the conclusions based on the two tests can be quite different. For example, with the data from Keith *et al.* (1985) used by Slatkin (1994), the observed configuration is (52, 9, 8, 4, 4, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1). The homozygosity test gives a tail probability of $P_H =$

https://doi.org/10.1017/S0016672300034236 Published online by Cambridge University Press

0·990998 and the exact test now gives a tail probability of $P_E = 0{\cdot}990330$. There is no consistent difference between the two tests. Of the 3014304 ordered configurations for the Keith *et al.* data set, 1928 of them had $\Pr(c) < \Pr(c_o)$ and $F(c) > F(c_o)$, where c_o is the observed configuration and $F(.)$ is the homozygosity, and 12180 had the inequalities reversed.

For n much larger than 100, the program to examine all configurations is too slow to be practical. I wrote a simulation program implementing Stewart's (1977) algorithm for generating random configurations that follow the Ewens sampling distribution. Both the exact and homozygosity tests can be carried out for each simulated configuration to provide an estimate of the tail probabilities, P_E and P_H. For the data set (30, 62, 97, 15, 53, 18, 55, 35, 57, 14866, 160, 439, 18, 356, 165, 40, 41, 14, 27, 36, 39, 23, 120, 209) ($n = 16975$ and $k = 24$), $P_H = 0{\cdot}99802$ and $P_E = 0{\cdot}28207$, and for the data set (7, 173, 3, 27, 16, 120, 29) ($n = 375$ and $k = 11$), $P_H = 0{\cdot}24552$ and $P_E = 0{\cdot}10999$. In these cases, the tail probabilities are based on 100000 replicates. Copies of computer programs (written in C) to carry out these analyses are available from the World Wide Web site, http://mw511.biol.berkeley.edu/homepage.html or on request from the author (slatkin@garnet.berkeley.edu).

I thank W. J. Ewens and S. Tavaré for helpful discussions of this topic. This research was supported in part by NIH grant GM 40282.

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.

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.

Slatkin, M. (1994). An exact test for neutrality based on the Ewens sampling distribution. *Genetical Research* **64**, 71–74.

Stewart, F. M. (1977). Computer algorithm for obtaining a random set of allele frequencies for a locus in an equilibrium population. *Genetics* **86**, 482–483.

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

https://doi.org/10.1017/S0016672300034236 Published online by Cambridge University Press