Evolution, 49(6), 1995, pp. 1280–1283

AN EXACT TEST FOR POPULATION DIFFERENTIATION

MICHEL RAYMOND[1] AND FRANÇOIS ROUSSET[2]
Institut des Sciences de l'Evolution, URA CNRS 327, Laboratoire de Génétique et Environnement, CC 065, USTL, Place E. Bataillon, 34095 Montpellier cedex 05, France
[1]*E-Mail: Raymond@isem.univ-montp2.fr*
[2]*E-mail: Rousset@isem.univ-montp2.fr*

Key words.—Contingency table, F_{st}, Markov chain, Metropolis algorithm, population structure.

Received March 11, 1994. Accepted September 19, 1994.

Population differentiation is usually evaluated in computing various statistics from allelic frequencies in each sample, and applying a statistical test to tentatively reject the null hypothesis Ho (Ho: no differentiation among populations). The parameter F_{st}, defined by Wright (e.g. 1969), is most often used. It varies from 0 (absence of differentiation) to 1 (complete differentiation). Three statistics are commonly used to test whether populations are significantly differentiated. First, for one locus, a χ^2 statistics is computed on the allelic contingency table under the null hypothesis (Workman and Niswander 1970). Second, for each locus, an exact χ^2 or various other statistics are computed by permutation procedures (Hudson et al. 1992; Roff and Bentzen 1989). Third, for several loci, a 95% confidence interval of an F_{st} estimate is established with a bootstrap procedure over F_{st} estimates for each locus (Weir 1990a). The unbiased estimate of F_{st}, θ (Weir and Cockerham 1984) is required in this case. This last way of testing population differentiation will be referred hereafter as W90.

An alternative way of testing population differentiation is presented here, based on an exact nonparametric procedure.

THE PROBABILITY TEST FOR POPULATION DIFFERENTIATION

An exact probability test can be constructed by using the classical Fisher test for R × C contingency tables. Each row represents a population and each column an allele (table 1). N_{ij} is the number of copies of allele j in population i. The null hypothesis Ho, independent between row and column variables, corresponds to an absence of population differentiation. Given Ho, the probability Π of the observed table is:

$$\Pi = \frac{\prod_{i=1}^{p} (N_{i.}!) \cdot \prod_{j=1}^{k} (N_{.j}!)}{(N_{..}!) \prod_{i=1}^{p} \prod_{j=1}^{k} (N_{ij}!)},$$

where $N_{i.}$ refers to sample size of population i, N_{ij} refers to the total number of allele j, and $N_{..}$ to the sum of all cells.

The exact value (or P) of type-one error probability for rejecting Ho is computed by summing the probabilities of all tables that have the same or smaller probabilities and with the same row and column sums (e.g., Fisher 1935; Yates 1984). This exact test is impossible to perform for most data sets, because the total number of cases to consider increases rapidly (Gail and Mantel 1977). Even the most elaborate available software offering an exact Fisher test on R × C contingency table, based on the sophisticated network al-

TABLE 1. Contingency table for test of population differentiation. N_{ij} represents the number of allele j in population i.

	Alleles					
Populations	1	···	j	···	k	Total
1	N_{11}	···	N_{1j}	···	N_{1k}	$N_{1.}$
i	N_{i1}	···	N_{ij}	···	N_{ik}	$N_{i.}$
p	N_{p1}	···	N_{pj}	···	N_{pk}	$N_{p.}$
Total	$N_{.1}$		$N_{.j}$		$N_{.k}$	$N_{..}$

TABLE 2. Illustration of the Markov chain algorithm for a small imaginary data set. Only the three first elements of the chain are represented, generated when the random number generator of Marsaglia et al. (1990) was initiated with the seeds 4, 3, 2 and 1. The two selected cells of step 2 of the algorithm (see text) are in bold characters, and the two putative recipient cells are underlined. The value of ρ, *R*, and INF are indicated below each table. A bold arrow between two consecutive elements of the chain indicates that a new state has been reached. With Y = 100,000, the *P*-value of the observed table is 0.155 (SE = 0.005).

Observed data				Chain element no.												
	Allele															
Pop.	1	2	3	1				2				3				
1	10	6	3	10	6	3		10	6	3		<u>10</u>	6	**3**		
2	<u>44</u>	7	**4**	➡ **45**	**7**	3	➡	45	**7**	<u>3</u>	➡	45	6	4	➡	etc.
3	**5**	1	<u>0</u>	**4**	<u>1</u>	1		4	<u>1</u>	**1**		**4**	2	<u>0</u>		
Variables																
ρ		0			−0.81				−0.81				−0.95			
R		0.44			0.30				0.88				1.1			
INF		—			1				2				3			

gorithm of Mehta and Patel (1983), cannot handle data sets typical of most population studies.

Instead of computing this exact probability, an unbiased estimate can be obtained using a Markov chain method. The principle is to explore the space of all possible contingency tables with the same marginals and to determine for all such tables if its probability of occurrence is higher or not than Π. The proportion of tables with a lower or equal probability than Π is an unbiased estimate of *P*. More precisely the algorithm is the following (see table 2 for an illustration):

(1) Set variables ρ = 0 and INF = 0.

(2) Select at random two cells of the table on different rows and columns (cells $i1,j1$ and $i2,j2$).

(3) If at least one of these cells is empty ($N_{i1,j1} = 0$ or $N_{i2,j2} = 0$), go to step 5.

(4) The chain can potentially move to the new state (from n to $n + 1$) represented by the table where:

$$N_{i1,j1} = N_{i1,j1} - 1$$
$$N_{i2,j2} = N_{i2,j2} - 1$$
$$N_{i1,j2} = N_{i1,j2} + 1$$
$$N_{i2,j1} = N_{i2,j1} + 1$$

The ratio (or $R = N_{i1,j1} \cdot N_{i2,j2} / (N_{i2,j1} + 1) \cdot (N_{i1,j2} + 1)$, which corresponds to Π_{n+1}/Π_n) of the probability of the two tables is computed. If this ratio is equal to 1 or larger, the chain moves to the new state. If it is below one, the chain moves to the new state with probability *R*.

If a new state is reached, compute: ρ = ρ + Ln(*R*).

(5) if ρ (which corresponds to $\text{Ln}(\Pi_{n+1}/\Pi)$) is equal to or less than 0 then INF = INF + 1. INF memorizes the number of times the chain has encountered tables with a lower or equal probability than the observed one.

(6) Repeat Y times from step 2. The value of Y is determined by the user. A typical value could be 50,000.

(7) The unbiased approximation of *P* is INF/Y. The standard error of the estimate is calculated as described in Hastings (1970).

A dememorization process should be first performed to take into account that the chain is always starting from the observed sample (Guo and Thompson 1992). This is done by running first the algorithm without step 5 with, for example, Y = 1000.

The general principle of Markov chain methods to estimate a parameter is based on the algorithm of Metropolis et al. (1953). Its application to estimate the exact probability for rejecting the independence of a contingency table has been established by Guo and Thompson (1989, Technical report no. 187, Department of Statistics, University of Washington, Seattle), with a slightly different (and slower) algorithm.

We have built a computer program in Quick Basic to perform the unbiased estimation of the Fisher exact test on contingency tables, based on the above algorithm, and named STRUC. Pseudorandom numbers were generated according to Marsaglia et al. (1990). STRUC has been thoroughly tested for very simple cases by hand calculations and by comparison with exact probabilities computed for several contingency tables (e.g., Mehta and Patel 1983). This program has been used to compare some available methods to detect population differentiation.

COMPARISON WITH PERMUTATION METHODS

Like permutation methods (Hudson et al. 1992; Roff and Benzen 1989), the Markov chain is a way to generate an exact probability distribution under the null hypothesis, which is not biased by rare alleles of low sample size. Both can be used theoretically to compute a probability test, an exact χ^2 or any other relevant distribution statistics, and have the same advantage compared with asymptotic statistics such as the χ^2 of Workman and Niswander (1970).

The relative advantage of the various possible exact statistics seems to depend on which alternative hypothesis is used (e.g., Hudson et al. 1992), such that no general recommendation could be made, although Hudson et al. (1992) and Roff and Bentzen (1992) recommended the χ^2. However, they did not include the probability test in their power comparisons for various alternative hypotheses. Some comparisons with the exact χ^2 have been performed (table 3), by analyzing data already used by Hudson et al. (1992) and other accessible data sets (Dias et al. 1995). As suggested by Roff and Bentzen (1992), no lumping of cells of the contingency tables was done. The probability test provided equal or lower values in all examples, but this does not indicate necessarily its superiority. This aspect deserves further work, but is out of the scope of this paper.

TABLE 3. Comparison of the probability test, the exact χ^2, and the asymptotic χ^2. Two data sets from the literature were used. N is the total size of the contingency table, k is the number of alleles. P values ± SE are given.

Data set	N	k	P-values		
			Probability test	Exact χ^2	Asymptotic χ^2
Dias et al. 1995					
locus *MS1*	186	21	$<5\ 10^{-6}$	0.0069 ± 0.0020	0.0018 (df = 140)
locus *MS3*	180	22	0.035 ± 0.0057	0.13 ± 0.012	0.11 (df = 147)
Krietman and Aguadé 1986					
Adh^S	53	15	0.45 ± 0.011	0.47* ± 0.01	0.44 (df = 14)
Adh^F	34	24	0.55 ± 0.006	0.55* ± 0.006	0.46 (df = 23)
All	87	39	0.27 ± 0.011	0.33* ± 0.01	0.35 (df = 38)

* These values are different from those of table 5 of Hudson et al. (1992) because they pooled some cells of the contingency table.

For very small data sets, permutation procedure are faster than the Markov chain method for a same precision of the estimate. However, as the data set increases to a size commonly encountered in population studies, the Markov chain method outcompetes permutation procedures (see Guo and Thompson 1992). This is because the time required to permute each element is proportional to the total size of the contingency table, whereas for the Markov chain method, computation time does not increase much with the number of subpopulations, the number of alleles, or the total size of all samples (results not shown).

TESTING THE W90 METHOD

The W90 method was tested by analyzing the complete data file published by Weir (1990a) to illustrate the bootstrap method for the diploid case. To compare the multilocus bootstrap method with the exact test on each locus, a global test for all loci was obtained by Fisher's combined probability test (Fisher 1970 in Manly 1985). Opposite conclusions were reached. The bootstrap result, which indicates a significant structure is supported neither by exact results for each locus nor by their combination (table 4).

It is easy to demonstrate that the use of bootstrap to build a confidence interval is incorrect in this case, because only a few loci are available. Let us consider two loci (for simplicity) in a nonstructured population. The true value of F_{st} is 0, and each of the unbiased θ_1 and θ_2 estimates will be positive or negative, in both cases with probability 1/2 (Weir 1990a). There are two situations:

1) θ_1 and θ_2 are both positive or both negative (with probability 1/2). Resampling of loci with replacement (bootstrap) will produce a θ distribution, which never includes the true value, and no confidence interval will include it.

2) θ_1 and θ_2 have opposite signs (with probability 1/2). Resampling of loci with replacement will produce a q distribution from which all confidence intervals (constructed as in Weir 1990a) representing more than 50% of the values will include the true value (zero).

The overall expected percentage of intervals generated at any significance level from 50% to 100% is therefore 50% under the null hypothesis. The bootstrap method produces an inaccurate nonparametric confidence interval in this case, as it does more generally (Schenker 1985; Efron 1987) because it converges to the correct confidence interval only when the number of independent estimates of F_{st} (i.e. loci) increases. Thus, as stated by Weir (1990a, p. 390), it is necessary to sample several loci. However, his own example shows that a dubious conclusion can be reached with five loci. In this example, two loci were uninformative, and the three others yielded positive estimates (see table 3). In such a case, it is not possible to find a confidence interval including the value 0 by bootstrapping over loci. However, under the null hypothesis, this will occur with probability $(1/2)^n$ with n informative locus.

TABLE 4. Comparison of W90 and the probability test. Data on θ are those published by Weir (1990) to illustrate his method on the diploid case. Multiloci test refers to the bootstrap of W90 or the combination of independent tests (Fisher's method). All exact tests were done with a Markow chain of 50,000 steps, and 1000 steps of dememorization (see text for explanations).

		Probability test	
	θ	Probability	SE
Locus 2	0.0213	0.23	0.00662
Locus 3	0.0136	0.21	0.014
Locus 4	0.0160	0.12	0.012
Overall statistics	0.0158	Not applicable	
Multiloci test	$P < 0.05$	$P > 0.1$	

CONCLUSION

Exact tests (either exact χ^2 or probability test) for populations differentiation has several advantages over previous methods. First, they are accurate and unbiased, even for very small samples or low-frequency alleles. Second, they provide test results for each locus (in contrast to W90), which allows for the possibility of detecting aberrant loci (e.g., selected loci). Multi-loci statistics can be obtained by Fisher's combined probability test. Third, they are independent of the ploidy level, although random mating is required for diploids or higher ploidy levels.

The Markov chain method is faster than permutation procedures for realistic data sets, and computation time remains within reasonable limits even for very big data set (e.g., $N > 1000$). Most tests using a Markov chain will give acceptable results (SE ≤ 0.01) in less than 30 s when run on a PC with Intel 486 processor (25 MHz) computer. The STRUC program is available by anonymous ftp (at FTP.CNRS-MOP.FR in the MSDOS directory) or upon E-mail request,

and is also included in the GENEPOP population genetics software package (Raymond and Rousset 1995).

The bootstrap method of W90 is only asymptotically correct and its validity has apparently not been previously determined by comparison with an exact test or by simulation. Exact test methods do not replace the need to compute F_{st}, as this parameter is used for other purposes (e.g., estimation of the number of migrants). The probability test can be used for a wide variety of biological data, as it represents an unbiased estimation of the Fisher exact test on contingency table. For example, the gametic or genotypic linkage disequilibrium (Weir 1990b) can be analyzed with this method.

Acknowledgments

We thank D. Bourguet, C. Chevillon, P. David, P. Jarne, J. D. Lebreton, N. Pasteur, M. Slatkin, and B. Weir for helpful comments. This work was financed in part by the MESR (grant no. 91C0939), the INRA (AIP toxicology), and the Servive Commun de Biosystématique de Montpellier. The 121^e RT is acknowledged for the stimulating environment it provided to F.R. This is contribution 94-089 of the Institut des Sciences de l'Evolution (URA CNRS 327).

Literature Cited

Dias, P., G. R. Verheyen, and M. Raymond. 1995. Source-sink populations in Mediterranean blue tits: evidence using single-locus minisatellite probes. Journal of Evolutionary Biology, submitted.

Efron, B. 1987. Better bootstrap confidence intervals. Journal of the American Statistical Association 82:171–185.

Fisher, R. A. 1935. The logic of inductive inference. Journal of the Royal Statistical Society 98:39–54.

Gail, M., and N. Mantel. 1977. Counting the number of contingency tables with fixed margins. Journal of the American Statistical Association 72:859–862.

Guo, S. W., and E. A. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics 48:361–372.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Molecular Biology and Evolution 9:138–151.

Manly, B. F. J. 1985. The statistics of natural selection. Chapman and Hall, London.

Marsaglia, G., A. Zaman, and W. W. Tsang. 1990. Toward a universal random number generator. Statistics and Probability Letters 8:35–39.

Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in r × c contingency tables. Journal of the American Statistical Association 78:427–434.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Raymond, M., and F. Rousset. 1995. GENEPOP (ver. 1.2): A population genetics software for exact test and ecumenicism. Journal of Heredity 86:248–249.

Roff, D. A., and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Molecular Biology and Evolution 6:539–545.

———. 1992. Detecting geographical subdivision: A comment on a paper by Hudson et al. Molecular Biology and Evolution 9: 968.

Schenker, N. 1985. Qualms about bootstrap confidence intervals. Journal of the American Statistical Association 80:360–361.

Weir, B. S. 1990a. Intraspecific differentiation, Pp. 373–410 *in* D. M. Hillis and C. Moritz, eds. Molecular systematics. Sinauer, Sunderland, Mass.

———. 1990b. Frenetic data analysis. Sinauer, Sunderland, Mass.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.

Workman, P. L., and J. D. Niswander. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. American Journal of Human Genetics 22:24–49.

Wright, S. 1969. Evolution and the genetics of populations, Vol. 2. The theory of gene frequencies. University of Chicago Press, Chicago.

Yates, F. 1984. Tests of significance for 2 × 2 contingency tables. Journal of the Royal Statistical Society, series A 147:426–463.

Corresponding Editor: J. B. Walsh