Several methods or tests and various software are currently being developed for analyzing data in population genetics and ecology, which often rely on computer-intensive algorithms. The potential user is thus confronted with the painful experience of freedom and, in particular, has to make a priori choices between different methods. Using examples drawn from population genetics, we explain some of these recently developed tools.

François Rousset and Michel Raymond are at the Laboratoire Génétique et Environnement, Institut des Sciences de l'Évolution, Université de Montpellier II, 34095 Montpellier, France.

A test based on such a distribution is 'exact' in the sense that exact probability values can be computed, although p is unknown. Exactness of tests does not imply that they 'tell the truth' about some hypothesis, since the outcome of the analysis can only be a statement about probabilities of events. Examples of exact tests in population genetics are listed in Table 1.

Persistent doubts are expressed about the use of the probability distribution defined by eqn 2, and methods based on the estimation of unknown parameter values are still proposed particularly in the ecological literature. The logic underlying the definition of exact tests shows that (1) the choice of a distribution where marginal counts are fixed

\[
\frac{n_{1a} n_{2a} \cdot n_{1A} n_{2A} \cdot}{n_{1a} n_{2a} \cdot n_{1A} n_{2A} \cdot} (1 - p)^{n_{1a}} p^{n_{2a}}
\]

(1)

(where \(n = n_{1a} + n_{2a}\) and depends on the unknown frequency \(p\) of allele \(A\) in the population.

It is well known that whatever the frequency \(p\), the distribution of the chi-square statistic can be approximated by the chi-square distribution and this approximation improves as sample size increases. This illustrates the first traditional answer to this problem, where the probability computed is an asymptotic approximation to the true value that depends on \(p\). The other solution, more rarely applicable, is to find a probability distribution that is independent of the unknown parameter \(p\). In the present example, although the probability distribution of samples is dependent upon the allelic frequency, the conditional probability among all possible samples with the same observed total (marginal) allelic counts \(n_{1a}\) and \(n_{2a}\) (as in Box I) is independent of this frequency. This conditional probability is

\[
\frac{n_{1a} n_{2a} \cdot n_{1A} n_{2A} \cdot}{n_{1a} n_{2a} \cdot n_{1A} n_{2A} \cdot}
\]

(2)

Table 1. Some exact tests available in population genetics

<table>
<thead>
<tr>
<th>Type of test*</th>
<th>Probability distribution</th>
<th>Statistics</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardy-Weinberg</td>
<td>genotypic counts</td>
<td>allelic counts</td>
<td>various</td>
</tr>
<tr>
<td>Population structure</td>
<td>allelic counts</td>
<td>allelic counts</td>
<td>various</td>
</tr>
<tr>
<td>Linkage disequilibrium</td>
<td>genotypic counts</td>
<td>genotypic counts</td>
<td>various</td>
</tr>
<tr>
<td>Ewens–Watterson</td>
<td>allelic counts</td>
<td>number of alleles</td>
<td>Pr(S)</td>
</tr>
</tbody>
</table>

*Each test is characterized by a probability distribution generated under the null hypothesis, and a test statistic used to define the rejection zone and to compute the P-value. For each probability distribution, there are as many exact tests as possible statistics, and only some of them are indicated. Computer programs for most of these tests are available as described in the references given.

\(\Phi(S)\) is the probability of the sample under the null hypothesis.

\(\chi^2\) is the sample homogeneity.
must be based on explicit probability arguments, (2) it is not necessary for the marginal counts to be fixed in advance by the experimental set-up\(^3\), and (3) it is not necessary to estimate the unknown frequencies. These points are important to keep in mind because they are also the basic justification of permutation procedures: permutation techniques are relevant when the distribution generated by permutation is independent of the unknown parameters under the ‘null’ hypothesis.

As an illustration, we can consider Mantel tests, which are used increasingly in population studies\(^13-15\). Suppose that the ‘null’ hypothesis is that genetic differentiation follows an island model (Box 2), with each subsample assumed to be sampled from a different subpopulation. For any given parameter values (i.e., values of $4N_m N_p$, $4N_m q$ in Box 2), the probability distribution of each subsample\(^16\) is independent of the observed subsamples in other subpopulations so that the probability of observing subsamples $s_1$ and $s_2$ in geographic positions $a$ and $b$, respectively, is the same as that of the reverse observation under the ‘null’ hypothesis. It is then possible to test for a relationship between genetic data and geographic distance by permutation of the subsamples between the different geographic locations. Such a test will be ‘exact’. By contrast, no conclusion can be drawn from random permutations of pairwise statistics (such as pairwise $F_{ST}$ estimates) between the different pairs of locations (cf. Box 3).

Some extensions of Mantel tests have been proposed such as partial Mantel tests, where the ‘null’ hypothesis corresponds to a correlation of some pairwise parameter with one parameter and aims to test for a partial correlation with another one\(^14,15,17\), but the validity of the permutation procedures proposed in such cases remains to be formally investigated.

### The choice of the statistic

A statistic is computed from the data, and is used to rank all possible samples of the distribution. The $P$-value of the test is then computed by summing up the probabilities of all samples of one side of the ranking, starting from (and including) the value of the observed sample (Table 2). Thus, different test statistics define different rankings of possible samples, but the $P$-value is in all cases defined as a sum of exact probabilities of samples with more extreme ranks, so all tests are exact tests. The probability of the observed sample is used as a test statistic in many situations (thus defining an exact probability test), but this choice, which traces back to Fisher’s exact test for contingency tables\(^18\), has no general justification.

How do we choose the test statistics? A test statistic is generally chosen to maximize the power of the test when some specified alternative hypothesis is true. It is therefore pivotal to have an idea of what could be a (biologically) plausible alternative hypothesis in order to design a powerful test. When specific alternative hypotheses have been formulated and probability distributions under the alternative hypothesis are known, it is possible to define more efficient test statistics and more efficient estimators.

For example, phenomena such as selling and other types of inbreeding lead to a specific form of heterozygote deficiency which is often the appropriate alternative hypothesis to design tests of random mating\(^19\). Sometimes it is less easy. For example, a sampling distribution is known for the island model, and has been used to define an estimator of $N_m$ that is efficient for this model\(^20\), but there is no equivalent result for other possible models of population structure.

Likewise, departure from neutrality could be the result of selection for particular genotypes, which leads to various predictions concerning the allele frequency distribution or the genotypic constitution. In most cases, the kind of selection acting on the locus studied is not known a priori; thus the formulation of a biologically sound alternative hypothesis of a test of neutrality is difficult. This is illustrated by the low power of the Ewens–Watterson test (which uses the sample homozygosity as a test statistic) with respect to specific types of selection\(^21\).

### Computing or estimating $P$-values

To compute the $P$-value of an exact test, all possible samples of the probability distribution must be considered and ranked according to a particular test statistic (see details in Table 2). This complete enumeration is not always possible, either because algorithms generating this exhaustive
sampling are not presently available or because computation time is limiting due to the large number of possible configurations to be considered. As an illustration, the number of possible genotypic tables corresponding to a sample of 40 individuals with 4 equifrequent alleles is 6671, and is higher than $10^7$ if the sample size is increased fourfold. One solution is to estimate this $P$-value by considering only a ‘random’ subset of possible samples. This subset is generated using two possible sampling techniques: permutation or Markov Chain Monte Carlo (MCMC) algorithms. For both methods, the probability of sampling a particular element is equal to its probability of occurrence under the ‘null’ hypothesis. The principle of permutation is to generate independent random samples under this hypothesis. The principle of MCMC is explained in Box 1. The main weakness of MCMC is that the speed of convergence to the correct value, though usually much faster than that of permutation algorithms for many applications, cannot be predicted in advance. This issue is a current area of investigation.

**The bootstrap: an alternative?**

The principle underlying the bootstrap is that if we do not know the true distribution of the difference between an unknown parameter $\theta$ and some estimator of this parameter, we can approximate it by the distribution of the difference between an estimate $\hat{\theta}$ obtained from a particular data set and estimates computed from different imaginary data sets generated using information provided by the sample. Instructive examples are given in the context of mark-recapture experiments in Ref. 23 and in the context of phylogenetic analysis in Ref. 24.

There are two ways to generate such imaginary data sets. In the nonparametric bootstrap, they are obtained by sampling with replacement within the real data set. For example, the distribution of the difference between an estimator of $F_{ST}$ and the true value of $F_{ST}$ could be approximated by the distribution of the difference between an estimate $\hat{F}_{ST}$ obtained from several loci and further estimates computed from different resamples of loci from this data set. The parametric bootstrap is similar except that one generates imaginary data sets according to a probability distribution whose parameters are estimated from the real data set. For example, the data may be used to estimate the parameters $p$ and $Nm$ of the distribution of subpopulation gene frequency $x$ in an island model (Box 2). Then imaginary subpopulations can be generated by sampling from the distribution with estimated parameter values $\hat{p}$ and $\hat{Nm}$, and imaginary data sets can be generated by sampling from such imaginary populations. In both cases it is possible to define confidence intervals for the parameter values (see Ref. 25 for a recent review and discussion of the different methods).

Though superficially similar to permutation methods, bootstrap methods are based on large sample approximations and are not an alternative answer to problems that can be solved by exact methods. Rather, they can address other problems. Like other asymptotic techniques, they may be the only ones available for most questions of interest. However, they can be misleading and may need laborious studies to be validated. There are as yet few investigations of possible uses of the bootstrap for population genetic analyses and it has not always been found accurate. However, these studies did not consider improved uses of the bootstrap such as those discussed in Ref. 25.

**Applying the bootstrap to population structure**

How can we define bootstrap confidence intervals for the value of $F_{ST}$? Consider the example of an island model with migration rate $m$ between subpopulations, and two possible alleles $A$ and $a$. The probability distribution of the allele frequency would follow Wright’s distribution (see Box 2). In this model, $F_{ST} = 1 / (1 + 4Nm)$. Now suppose that $n$ individuals are sampled in each of two subpopulations, and we wish to estimate $Nm$. What will happen if we resample individuals? $F_{ST}$ is estimated from sample allele frequencies, and the larger the sample size $n$, the closer we will be to some value, which will
Box 2. The island model and the definition of $F_{ST}$

In the island model, migrants into a subpopulation are supposed to come with identical frequencies $x$/n from any other of a number of $n$ subpopulations. If $n$ is large and the fraction of migrants is $m < 1$, and considering only two alleles for simplicity, the probability distribution of the gene frequency $x$ in a particular subpopulation is approximately $f(x) = Cx^4(1-x)^3$, where $C$ is a constant, $p = 1 - q$ are the allele frequencies in the total collection of subpopulations, and $N$ is the subpopulation size. Thus the probability of identity $Q_{scr}$ of a pair of genes sampled within a subpopulation is the expectation of $x^2 = 1 - q^2$:

$$Q_{scr} = \int_0^1 \left( x^2 + (1-x)^2 \right) f(x) dx = \frac{1 + 4Nm^2 - 2q^2}{1 + 4Nm}$$

and the probability of identity of a pair of genes sampled between subpopulations is the sum of squares of expected allele frequencies, $Q_{mag} = p^2 + q^2$. Hence $F_{ST}$ defined as $Q_{mag} - Q_{scr}/(1 - Q_{scr})$ has value $1/(1 + 4Nm)$. This is no more than Wright’s $F_{IS}$, formulated in terms of probabilities of identity in state for pairs of genes, rather than identity by descent or ‘standardized variance of gene frequencies’. With this definition, it is easier to grasp basic properties of both the parameter $F_{ST}$ and its estimators in more complex models.

In the present context $F_{ST}$ is not defined as a function of the allele frequencies in some subpopulations, say $x_1$ and $x_2$, in two subpopulations under study. These frequencies as well as probabilities of identities for these particular subpopulations – the so called ‘actual values’, $Q_{mag} = x_1^2 + (1-x_1)^2$ and $Q_{scr} = x_1x_2 + (1-x_1)(1-x_2)$ – are random variables that will differ for each pair of subpopulations. By contrast, $F_{ST}$ is a parameter, that is, it is a property of a probability distribution rather than of a particular realization of the process considered.

be determined by the frequencies of the $A$ allele in the two subpopulations. But this value will not be $1/(1 + 4Nm)$. This is readily seen by the fact that if we had taken another pair of subpopulations, the frequency of $A$ in each of them would be different from those in the previous pair, so when the sample size increases, the $F_{ST}$ estimates would become closer to another value; we do not get closer to the parameter value, but to a random value different for each pair of subpopulations. Consequently, a test based on a confidence interval constructed in this way should almost always reject the correct parameter value for large sample sizes.

The appropriate question is whether the value of the variable considered in some population (not its estimate on a sample) is a random variable the distribution of which is determined by the parameter value, or a fixed value identical for different populations exposed to the same process. This is the question to consider in order to choose between fixed and random models in statistics. In models where genetic drift plays a role, quantities such as the subpopulation frequency $x$ (Box 2) are random variables (sometimes called ‘actual values’). Since $F_{ST}$ is mainly used to measure the effects of genetic drift relative to other evolutionary forces, resampling individuals will generally not be appropriate for inferences about $F_{ST}$ values. Resampling loci may not be a perfect solution. Because typical population genetics studies rarely consider more than twenty loci, studies are needed to check that bootstrap confidence intervals generated by resampling loci are satisfying. But confidence intervals obtained in this way can be free of the problem of convergence to an incorrect parameter value.

**Persistent problems**

As new statistical techniques are proposed, the need to understand traditional concepts of probability and statistics is stronger. We have illustrated this need in some of the simplest problems encountered in population genetics. The recent development of efficient algorithms for exact testing and the large access to computer power provide a new input for the analysis of population data. More efficient tests and estimators can save much effort, but the more general problem of constructing confidence intervals for estimates of standard population genetic parameters has not yet any answer.

Traditional parameters such as $F_{ST}$ may be used to detect differences in levels of differentiation of populations exposed to different evolutionary processes, for example if a group of populations is expected to display a higher $F_{ST}$ than another group, or they may be used as measures that allow for another test.
the estimation of demographic parameters such as neighborhood sizes. In the first case, one would prefer a statistic whose distributions under the two situations we wish to distinguish do not overlap, such as an estimator with low variance. In the second case an estimator with low bias may be preferred.

However, in both cases we need realistic models to assess the power of different statistics or the efficiency of different estimators. Many studies have argued that there are discrepancies between direct and indirect estimates of gene flow parameters implying that the models generally considered are inappropriate, or inappropriately interpreted, or that present statistical analyses are inefficient. Whatever the reason(s), a better understanding of real populations will be necessary to show that many of the currently developed computer-intensive methods improve the efficiency of statistical analyses.

Acknowledgements

We thank F. Cézilly, C. Chevillon, M. Lascoux, N. Pasteur and J. Rouf fort for helpful advice and criticisms of the manuscript. This work was partly supported by the Centre de Systématique de Montpellier. This is contribution ISM 97-058 of the Institut des Sciences de l’Évolution (CNRS, UMR 5554).

References

12 Tocher, K.D. (1950) Extension of the Neyman–Pearson theory of tests to discontinuous variables, Biometrika 37, 130–144
45 Saltikov, M. (1994) Linkage disequilibrium in growing and stable populations, Genetics 137, 331–336