Estimating *F*-Statistics: A Historical View

Bruce S. Weir*†

Characterizing the genetic structure of populations is of importance to evolutionary biology, to human disease gene mapping, and to forensic science. Sewall Wright introduced a set of *F*-statistics to describe population structure in 1951, and he emphasized that these quantities were ratios of variances. Responding to uncertainty over the best way to estimate *F*-statistics, Weir and Cockerham published a method-of-moments set of estimators in 1984. This paper continues to be widely cited, with over 7,000 citations to date. Some background to the publishing history of the Weir and Cockerham paper is given here, along with subsequent developments and a discussion of current uses of Wright's *F*-statistics.

1. Introduction. This article discusses the context and history of "Estimating *F*-Statistics for the Analysis of Population Structure" by B. S. Weir and C. C. Cockerham (1984). Apart from the material presented in this paper, there is interest because of its very high citation record: the Web of Science lists over 7,000 citations as of July 2012 and Google Scholar lists over 8,000. Although it addresses a topic in population genetics, it was included in a list of the 25 most-cited papers in statistics (Ryan and Woodall 2005).

Weir and Cockerham's paper was prepared in response to a question in 1983 from population geneticist John Doebley: "How should I estimate *F*-statistics?" At that time he, Weir, and Cockerham were faculty members at North Carolina State University. Doebley had noticed many alternative equations in the literature for estimating F_{ST}, one of the suite of quantities for describing population structure, and he noted competing claims for the properties of these estimates. Weir and Cockerham's paper starts as follows:

*To contact the author, please write to: Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232; e-mail: bsweir@uw.edu.

†This work was supported in part by National Institutes of Health grant GM 075091. John Doebley suggested the 1984 paper of Weir and Cockerham: that paper rested on previous work by the late C. Clark Cockerham. Joe Felsentein was the corresponding editor for the 1984 paper. Bill Hill and the late Alan Robertson provided helpful discussions for the 1984 paper.

Philosophy of Science, 79 (December 2012) pp. 637–643. 0031-8248/2012/7905-0024$10.00
Copyright 2012 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/667904 Published online by Cambridge University Press

> This journal [*Evolution*] frequently contains papers that report values of *F*-statistics estimated from genetic data collected from several populations. These parameters, F_{ST}, F_{IT}, and F_{IS}, were introduced by Wright (1951), and offer a convenient means of summarizing population structure. While there is some disagreement about the interpretation of the quantities, there is considerably more disagreement on the method of evaluating them. Different authors make different assumptions about sample sizes or numbers of populations and handle the difficulties of multiple alleles and unequal sample sizes in different ways. Wright himself, for example, did not consider the effects of finite sample size. (1984, 1358)

Sewall Wright introduced *F*-statistics in 1951 to augment the inbreeding coefficient, and his set of *F*-coefficients were devised for describing population structure in breeds of livestock and in natural populations (Wright 1951). Wright's coefficients were actually parameters rather than data functions, and his use of "statistics" has been reflected by subsequent authors defining the coefficients in terms of sample values. A central theme to Weir and Cockerham's work is that parameters can be defined as components of evolutionary theories and then appropriate estimating statistics found.

Wright defined the basic inbreeding coefficient, F_{IT} or F, as the correlation between genes on uniting gametes relative to the total array of those in random derivatives of the foundation stock. Similarly, F_{ST} or θ was introduced as the correlation between uniting gametes relative to those across all subdivisions. "Most importantly, F_{ST} is the ratio of the actual variance of gene frequencies of subdivisions to its limiting value, irrespective of their own structure" (Wright 1965, 402). The concept of "relative to" is not an easy one, and it has made the study of population structure difficult. As an illustration, consider this fanciful example: a large group of people, all of whom are related to each other as first cousins, establish a community and choose mates randomly within that community. The resulting set of children are tested for consistency with the Hardy-Weinberg law (i.e., for independence of their maternal and paternal alleles). Because their parents mated at random, the within-population inbreeding coefficient F_{IS} is zero. From the perspective of an outsider, however, the children are inbred, $F_{IT} = 1/16$, because their parents were cousins. The group of cousins are related one to another, relative to the general population, to an extent $F_{ST} = 1/16$.

For a formal statistical treatment, the meaning of correlation is made clearer by introducing variables x for each allele. For allele j in population subgroup i, x_{ij} is defined to be 1 if the allele has a particular type A and is 0 otherwise. Then

- the mean of x_{ij} over all alleles in subpopulation i is the subpopulation allele frequency p_i;

https://doi.org/10.1086/667904 Published online by Cambridge University Press

- the mean of x_{ij} over all alleles in all subpopulations is the population allele frequency p;
- the mean of p_i over all subpopulations is p;
- the correlation of x_{ij} and $x_{ij'}$ for different alleles j, j' is θ or F_{ST}; and
- the variance of p_i over all subpopulations i is $\theta p(1-p)$.

Wright (1951, 325) said, "More generally, F_{ST} in the broad sense can always be obtained, *at least empirically* for the variance of distribution of gene frequencies even in cases involving selection" (emphasis added). This has led to a common definition of F_{ST} in terms of sample mean $\bar{p}$ and sample variance s^2 of allele frequencies over a set of samples from different populations:

$$F_{ST} = \frac{s^2}{\bar{p}\,(1-\bar{p})}.$$

This definition is unfortunate as it depends on a particular sample, and the parameter being estimated by this quantity will be different even for different samples from the same population.

2. The Work of Cockerham and Nei. Cockerham (1969, 1973) gave a series of analyses of the indicator variables x in an analysis of variance framework and showed how to estimate F and θ. These papers were written for a single allele: "I thought I had set a proper stage for analysis of gene frequencies, leaving the pooling over alleles to other people. I much prefer to set the groundwork and let others hammer out the details. This does not seem to be what happens, however" (C. Clark Cockerham, personal communication, November 28, 1983). Part of the subsequent success of Weir and Cockerham (1984) is that the details were included. In reviewing the first submission of the paper, Peter Smouse said of Cockerham's work that "the 1969 paper is virtually unreadable. . . . The 1973 paper is considerably more readable, but is still heavy going for most of us. [I had trouble with it, and I'm supposed to know what I'm doing.]" In one of the papers motivating Weir and Cockerham, Nei and Chesser (1983, 253) said that "Cockerham (1973) developed statistical methods . . . but his methods are quite complicated. . . . Furthermore he defines allelic correlations in reference to an imaginary population."

In series of papers from 1972 to 1983, Masatoshi Nei gave a framework different from that of Wright and Cockerham. Rather than working with correlations and a reference population, he worked with "gene diversity" or the sums of squares of allele frequencies in populations. There is no longer an evolutionary framework: "gene diversity is defined by using the gene frequencies at the present generation, so that no assumption is required about the pedigrees of individuals, selection, and migration in the past" (Nei 1977, 225). As a description of current frequencies, Nei's approach is appropriate, but there is then no basis for an evolutionary interpretation of esti-

https://doi.org/10.1086/667904 Published online by Cambridge University Press

mates and no justification for making statements about divergence from ancestral populations, the effects of natural selection, or the extent of migration, for example. The work in these papers does not need to distinguish allele frequencies in a sampled (sub)population and the average over all populations. The work of Nei and his colleagues loses the "relative to" perspective of Wright and Cockerham.

3. The Work of Reynolds et al., 1983. In his PhD thesis directed by Cockerham and Weir, John Reynolds gave a multiple-allele extension of Cockerham's 1969 and 1973 papers to provide an estimate of θ. The thrust of the resulting paper (Reynolds, Weir, and Cockerham 1983) was to show that the evolutionary framework allowed θ to be regarded as a measure of the time since the sampled populations diverged from an ancestral population: this allowed the reconstruction of trees linking populations. For a set of populations, all of size N, and mating at random, θ has the predicted value of $1 - (1 - 1/2N)^t \approx t/2N$ if t generations have elapsed since the populations diverged from a common ancestral population.

The interpretation of θ as a measure of evolutionary distance required the implicit assumption of no (or low) mutation, as is appropriate for short-term evolution within a species. There is an explicit assumption that θ is changing over time. This is in contrast to the situation with Nei's distance, which is appropriate for between-species studies, does require mutation, and assumes that θ is constant within populations.

Reynolds et al. (1983) also surveyed existing methods of estimating F_{ST} and of estimating genetic distance. Their simulations confirmed the advantages of the Cockerham estimation approach for the case of population divergence by drift. The stage was set for a comprehensive examination of F-statistics by Weir and Cockerham.

4. Weir and Cockerham (1984). In responding to the initial (June 1983) submission to *Evolution*, the editor, Douglas Futuyma, said in September 1983, "Reviewer 1 felt it 'is very caustic in places. I'm afraid that it, in its present form, will invite a lot of correspondence that will cloud the picture instead of unifying the methodologies which, I believe, is the real purpose of the paper.' A similar point is made by Rev. 2 [Peter Smouse] and by Felsenstein, who is concerned that *Evolution* not serve as a forum for an endless and confusing set of replies from some predictable quarters. I agree." There is little justification in general for authors to be "caustic," and from a distance of almost 30 years and the success of the 1984 article, it seems even less justifiable.

The reviewers for the original submission had some technical comments that were addressed by additional theoretical work, included in the appendix to the article, and by simulations. The revision process was aided by discus-

https://doi.org/10.1086/667904 Published online by Cambridge University Press

sions Weir had with W. G. Hill and A. Robertson while he was on sabbatical leave at the University of Edinburgh.

The revised paper was submitted in March 1984, accepted in May, and published in December. Each year since then it has received an increasing number of citations as different areas of application are explored. The estimator has proven to be very robust and amenable to incorporation into computer packages. As mentioned earlier, a key point in the paper is the necessity to define a parameter of interest and then seek ways to estimate this quantity.

Extensions to Weir and Cockerham (1984). The concepts of "relative to" are not trivial, and the algebra leading to the *F*-statistic estimators in the study is somewhat dense. Even further complexity in the paper was avoided, but at the expense of biological and statistical limitations.

Statistically, the method of moments adopted in the paper has the advantage of giving estimates with low bias but little else. There is not a general expression for the sampling distribution of the estimates, and expressions for estimate variances are quite complex. The reason is that the sampling distributions of allele or genotype frequencies over subpopulations are not known in general, unlike the known multinomial distribution for genotype counts over repeated samples from the same population. If distributions can be assumed, however, there is scope for maximum likelihood and Bayesian methods (Holsinger and Weir 2009).

By assuming that allele frequencies are normally distributed over populations, Weir and Hill (2002) found maximum likelihood estimates for θ. These estimates have chi-square distributions that allow a discussion of sampling properties in a better way than the numerical resampling methods described in the original Weir and Cockerham paper.

Genetically, the paper suffers from the "star phylogeny" assumption of all sampled populations descending independently from a single ancestral population and having the same value of θ. These assumptions were removed by Weir and Hill (2002), who gave moment estimates of population-specific values of θ. This later work was extended by Browning and Weir (2010), who worked with haplotype clusters instead of individual genetic markers and gave improved estimates of population-specific θ values (still relative to the whole set of populations).

5. Applications of F_{ST}. The growing number of citations to Weir and Cockerham (1984) indicates the wide range of areas in which knowledge of θ is important. Among the 20 most recent citations, as of February 2012, eight were for studies on fish and shellfish, two on humans, two on grapevine pathogens, two on termites, and one each on chestnut trees, partridges, chickens, ants, and frogs. One application with societal benefits is in forensic science and the calculation of "match probabilities" for DNA profiles. If a suspect has a profile that matches that of a crime scene sample, the relevant

https://doi.org/10.1086/667904 Published online by Cambridge University Press

question is, Given that the suspect/defendant has the profile, what is the chance that someone else in the population also has the profile? The answer depends on θ (Weir 2007): increasing values of θ indicate increasing dependencies among alleles in the same subpopulation, greater match probabilities, and less probative value to a match. There is current debate on corresponding arguments for Y-chromosome profiles (Buckleton, Krawczak, and Weir 2011).

6. Conclusion. The Weir and Cockerham (1984) paper on estimating population structure parameters did not have an easy passage to publication, but it has emerged as one of the most-cited papers in population genetics. Its longevity reflects the importance of the quantities being estimated in many aspects of genetics. Two key features have contributed to its methodology being still in use: there are relatively few assumptions made about the underlying evolutionary process, and the focus on parameters rather than statistics means that different sampling designs can be accommodated.

It should not be thought, however, that the paper is held in universal high regard. In a personal communication to the author, Masatoshi Nei said,

> Actually I am asked to write a historical paper about F_{ST}. Of course, I have not written anything yet, because I am still working on my next book. I have already spent more than seven years, but it will take a few more months to finish it. However, I plan to write something for AHG, because I have promised for them. My view on F_{ST} is conceptually different from yours. So it would be interesting to see different views on the same issue. My recent view is written in Chapter 12 of my 2000 book "Molecular Evolution and Phylogenetics." (November 4, 2010)

REFERENCES

Browning, Sharon R., and Bruce S. Weir. 2010. "Population Structure with Localized Haplotype Clusters." *Genetics* 185:1337–44.

Buckleton, John S., Michael Krawczak, and Bruce S. Weir. 2011. "The Interpretation of Lineage Markers in Forensic DNA Testing." *Forensic Science International: Genetics* 5:78–83.

Cockerham, C. Clark. 1969. "Variance of Gene Frequencies." *Evolution* 23:72–84.

———. 1973. "Analyses of Gene Frequencies." *Genetics* 74:679–700.

Holsinger, Kent E., and Bruce S. Weir. 2009. "Genetics in Geographically Structured Populations: Defining, Estimating, and Interpreting F_{ST}." *Nature Reviews Genetics* 10:639–50.

Nei, Masatoshi. 1977. "F-Statistics and Analysis of Gene Diversity in Subdivided Populations." *Annals of Human Genetics* 41:225–33.

Nei, Masatoshi, and Ronald K. Chesser. 1983. "Estimation of Fixation Indices and Gene Diversities." *Annals of Human Genetics* 47:253–59.

Reynolds, John, Bruce S. Weir, and C. Clark Cockerham. 1983. "Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance." *Genetics* 105:767–79.

Ryan, Thomas P., and William H. Woodall. 2005. "The 25 Most-Cited Papers in Statistics." *Journal of Applied Statistics* 32:461–74.

Weir, Bruce S. 2007. "The Rarity of DNA Profiles." *Annals of Applied Statistics* 1:358–70.

https://doi.org/10.1086/667904 Published online by Cambridge University Press

Weir, Bruce S., and C. Clark Cockerham. 1984. "Estimating *F*-Statistics for the Analysis of Population Structure." *Evolution* 38:1358–70.
Weir, Bruce S., and William G. Hill. 2002. "Estimating F-Statistics." *Annual Review of Genetics* 36:721–50.
Wright, Sewall. 1951. "The Genetical Structure of Populations." *Annals of Eugenics* 15:323–54.
———. 1965. "The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating." *Evolution* 19:395–420.

https://doi.org/10.1086/667904 Published online by Cambridge University Press