Notes on F_{st}

Nick Patterson

February 26, 2020

These are some old notes, explaining my version of F_{st}. There is an overlap with [1], but also some material never published.

1 What is F_{st}?

Suppose we have a biallelic marker in two populations in Hardy-Weinberg equilibrium. Choose the variant allele, and suppose that the allele has population frequency p_1, p_2 in populations 1 and 2 respectively Set $q_i = 1 - p_i$. Then we can define Wright's F_{st} as

$$F_{st} = N/D \tag{1}$$

where

$$\begin{aligned} N &= p_1(q_2 - q_1) + p_2(q_1 - q_2) && (2)\\ D &= p_1q_2 + q_1p_2 = N + p_1q_1 + p_2q_2 && (3) \end{aligned}$$

This is a definition of F_{st}, a parameter measuring divergence at a given locus, *not* a sample statistic. In this paper we are only interested in divergence measures of biallelic markers in two populations and will always assume the populations are themselves homogeneous. We are primarily interested in estimating divergence time since a split between two populations that diverged at some point in the past. Changing notation so that

$$\begin{aligned} x &= p_1\\ y &= p_2 \end{aligned}$$

we find:

$$\begin{aligned} N &= (x-y)^2\\ D &= x + y - 2xy \end{aligned}$$

Note that we can write D as

$$D = (x-y)^2 + x(1-x) + y(1-y)$$

which makes N/D look like a ratio of variances (or an ANOVA style statistic) and motivated Wright's original idea. This also makes it clear that $0 \leq N/D \leq 1$.

Suppose for now that x is the allele frequency at some past time, and that y is the frequency now at time τ later, where unit time is $2M$ generations and M is the effective population size. Fix x and consider the expectation of N and D conditional on x. D is easy. Population frequency under the Wright-Fisher diffusion is a martingale and so

$$E(D|x) = 2x(1-x) \tag{4}$$

There are numerous ways to compute $E(N|x)$. Here is a simple argument due to Simon Myers. We claim that

$$E(y(1-y)|x) = x(1-x)e^{-\tau}$$

Consider two chromosomes from the population today. The probability they are heterozygous, conditional on y is of course $2y(1-y)$. However the two samples are heterozygous if and only if they have not coalesced by time $-\tau$ and the ancestors at that time are heterozygous. The probability of no coalescence is just $e^{-\tau}$ and this proves our claim.

Now it follows that

$$\begin{aligned} E(N|x) &= E((x-y)^2|x) && (5)\\ &= -x^2 + E(y - y(1-y)|x) && (6)\\ &= x - x^2 - x(1-x)e^{-\tau} && (7)\\ = x(1-x)(1-e^{-\tau}) && (8) \end{aligned}$$

Another proof is to be found in Nei's book [6, Chapter 13], but the proof above is more conceptual. As a check, we see that $E(N|x) \approx x(1-x)\tau$ for τ small. On the other hand, as $\tau \to \infty$ then

$$E(N|x) \to x(1-x)$$

But the probability that y fixes at 1 is x (martingale again) and so more directly

$$E(N|x) \to x^2(1-x) + (1-x)^2 x = x(1-x)$$

The distribution of y, conditional on x is the Kimura function $K(x, y; t)$ given explicitly by Kimura ([5]). (See for example [7] for a recent treatment and references, or [4] for a textbook treatment.)

Instead of considering x as an allele frequency at some past time, suppose instead we have two populations with allele frequencies x in population 1 and

y in population 2. The populations diverged t generations ago, and effective population sizes are M_1 and M_2 respectively.

Set $\tau_1 = t/2M_1$ and $\tau_2 = t/2M_2$. The Wright-Fisher diffusion is reversible [8] [8] and it follows that the distribution of y conditional on x is $K(x, y; \tau)$ where $\tau = \tau_1 + \tau_2$. Hence

$$E(N|x) = x(1-x)(1-e^{-\tau}) \quad (9)$$

$$E(D|x) = 2x(1-x) \quad (10)$$

Thus defining the overall population F_{st} as the quotient of N and D each summed over many alleles, (an issue we discuss in much more detail in section 3) we get that

$$F_{st} = \frac{1 - e^{-(\tau_1+\tau_2)}}{2} \quad (11)$$

It is important to note that $E(N|x)/E(D|x)$ is independent of the allele frequency x. Thus it will not be critical exactly how the marker was ascertained (however see below). Nevertheless this is only true under simple demographies, and will *not* be the case in other circumstances, such as migration or merging of populations after original divergence. Thus independence of F_{st} on allele frequency is a test for a simple demographic history.

2 A Comparison with Cavalli-Sforza

Cavalli in his book [3], discusses F_{st}. He defines it (see equation 1.11.1) as

$$d = \frac{N}{2P(1-P)}$$

where N is the same as for us and P is the allele frequency at the divergence time of the two populations. Take the case that the effective population size of each population is M, and that the divergence time is τ where unit time is $2M$ generations. He then gives:

$$d = 1 - e^{-\tau} \quad (12)$$

while equation (11) gives for this situation,

$$d = \frac{1 - e^{-2\tau}}{2}$$

Who is right? First of all, for our denominator D, it is trivial to check (see (3)) that

$$E(D|P) = 2P(1-P)$$

so Cavalli's d and our F_{st} are *roughly* the same quantity. Further writing $x - y = (x - P) + (P - y)$ and noting that conditional on P, $x - P$ and $P - y$ are independent and mean 0, we get from equation (8) that

$$E(N|P) = 2P(1-P)(1-e^{-\tau})$$

which is enough to prove Cavalli's formula (12). I believe the difference is in ascertainment. Cavalli implicitly assumes that P is not 0 or 1, or in other words that the marker is polymorphic at the time of population divergence. I assume on the other hand that the marker is polymorphic in population 1. Thus alleles that have fixed differently in the two populations are excluded in our calculation of F_{st}. This explains how for very large time, Cavalli's d will be close to 1, as most polymorphism between two very divergent populations will be with markers fixed in both populations. On the other hand, our population F_{st} will not be larger than 1/2. Fortunately, for very small times τ the two equations both yield

$$F_{st} \approx \tau$$

In many practical circumstances, our calculation seems more realistic, as the status of a marker at divergence time is essentially unknowable. As an example of an application of current interest, many chimp SNP's have been ascertained as heterozygotes in Clint, a chimpanzee that is primarily from the western subpopulation. This is the kind of ascertainment considered here.

3 Estimation

Returning to equations (2, 3), Suppose we have a set S of markers $A_k (k = 1, \ldots M)$. For marker k we define now $N^{[k]}$ and $D^{[k]}$ in the obvious way. We now *define* $F(S) = F_{st}$ for the marker set S by

$$F(S) = \frac{N(S)}{D(S)} \tag{13}$$

where

$$N(S) = \frac{\sum_{k=1}^{M} N^{[k]}}{M} \tag{14}$$

$$D(S) = \frac{\sum_{k=1}^{M} D^{[k]}}{M} \tag{15}$$

Note that $F(S)$ is a function of the population allele frequencies and is *not* a statistic (function of the observations). We wish to compute an estimator of F. Given the form of equation (13) it is highly desirable to find unbiased estimators of $N^{[k]}, D^{[k]}$ else the bias will eventually dominate the estimate. Fix for now, marker k, and suppose the population frequencies are p_1, p_2 for the variant allele, and we observe allele counts a_1, a_2 for the variant allele, b_1, b_2 for the reference allele. Take $n_i = a_i + b_i,\ i = 1, 2$. $N = N^{[k]}$ is defined as $(p_1 - p_2)^2$. A naive estimator for N is

$$X = (a_1/n_1 - a_2/n_2)^2$$

We calculate the bias of X. Writing

$$X = ((a_1/n_1 - p_1) - (a_2/n_2 - p_2) + (p_1 - p_2))^2$$

Then

$$\begin{aligned} E(X) &= (p_1 - p_2)^2 + Var(a_1/n_1|p_1) + Var(a_2/n_2|p_2) && (16) \\ &= (p_1 - p_2)^2 + p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2 && (17) \end{aligned}$$

Define $h = p_1(1-p_1)$ ($2h$ is the heterozygosity at the marker for population 1). Then a natural estimator for h is

$$\hat{h} = \frac{a_1(n_1 - a_1)}{n_1(n_1 - 1)} \tag{18}$$

It is easy to check that $\hat{h}$ is unbiased. Similarly define h' for population 2, with a corresponding estimator $\hat{h'}$. This is enough to show that:

$$\hat{N} = (a_1/n_1 - a_2/n_2)^2 - \hat{h}/n_1 - \hat{h'}/n_2 \tag{19}$$

is an unbiased estimator for N. Now

$$D = N + h + h'$$

which shows

$$\hat{D} = \hat{N} + \hat{h} + \hat{h'} \tag{20}$$

is an unbiased estimator for D.
By the Lehmann-Scheffé theorem [2, Theorem 4.2.2] $\hat{N}$ and $\hat{D}$ are uniformly minimum variance unbiased estimators. No longer fixing a marker and writing $\hat{N}^{[k]}$ for our estimator of $N^{[k]}$, and so on, we see that a natural estimator for $F(S)$ is

$$\hat{F} = \frac{\sum_k \hat{N}^{[k]}}{\sum_k \hat{D}^{[k]}} \tag{21}$$

Another view of this is that given a pair of allele frequencies $F = (f_1, f_2)$ we define

$$\begin{aligned} N(F) &= (f_1 - f_2)^2 \\ D(F) &= N(F) + f_1(1-f_1) + f_2(1-f_2) \end{aligned}$$

3.1 Estimators in the presence of inbreeding

Now we extend the theory to the case where some of our samples have excess homozygosity due to inbreeding. We give estimators of N, D that are unbiased, without explicitly estimating the inbreeding coefficients. Let x_0, x_1, x_2 be the number of samples of population 1 with 0, 1, 2 copies of the variant allele. Let y_0, y_1, y_2 be the corresponding numbers for population 2. Let

$$\begin{aligned} s &= x_0 + x_1 + x_2 \\ t &= y_0 + y_1 + y_2 \end{aligned}$$

We will require that $s, t > 1$. In the notation of the previous section:

$$\begin{aligned} a_1 &= x_1 + 2x_2 \\ a_2 &= y_1 + 2y_2 \\ n_1 &= 2s \\ n_2 &= 2t \end{aligned}$$

which will lead to estimators for N, D. In the presence of inbreeding, these estimators are incorrect. Note however that if we pick alleles randomly from each diplotype, then we will obtain valid unbiased estimators. We can of course then obtain more efficient estimators by averaging over our choice of alleles. Define $n_1 = a, n_2 = b$. Select an allele at random from each diploid genotype. Let u be the allele count for population 1, and v be the count for population 2. From equation (19) we want to compute expected values of:

$$\begin{aligned} X &= (u/s - v/t)^2 \\ \hat{h} &= \frac{u(s-u)}{s(s-1)} \\ \hat{h'} &= \frac{v(t-v)}{t(t-1)} \end{aligned}$$

when our estimator for N is

$$\hat{N} = E(X) - E(\hat{h})/s - E(\hat{h'})/t \tag{22}$$

For X, we see that u has mean $x_1/2 + x_2$ and variance $x_1/4$. Similarly v has mean $y_1/2 + y_2$ and variance $y_1/4$. It follows that

$$E(X) = \left(\frac{x_1 + 2x_2}{2s} - \frac{y_1 + 2y_2}{2t}\right)^2 + \frac{x_1}{4s^2} + \frac{x_2}{4t^2}$$

For $E(\hat{h})$ we need the expected value of $u(s-u)$. Standard binomial coefficent identities show that

$$E(u(s-u)) = x_0 x_2 + (x_0 + x_2)x_1/2 + x_1(x_1 - 1)/4$$

Now it follows that:

$$E(\hat{h}) = \frac{x_0 x_2 + (x_0 + x_2)x_1/2 + x_1(x_1 - 1)/4}{s(s-1)} \tag{23}$$

$$E(\hat{h'}) = \frac{y_0 y_2 + (y_0 + y_2)y_1/2 + y_1(y_1 - 1)/4}{t(t-1)} \tag{24}$$

We now can apply equation (22) to obtain $\hat{N}$. For $\hat{D}$ we have, using $D = N + h + h'$, the equation

$$\hat{D} = \hat{N} + E(\hat{h}) + E(\hat{h'}) \tag{25}$$

The same ideas lead to a simple estimator of the inbreeding coefficient, p_I. the probability, in a sample from a population, that the two alleles at a locus are identical by descent (IBD). For our case, with a assumed homogeneous population, this is the same as Wright's fixation index F. (See [6, page 154]). Consider population 1, with the same notation as above. Let H be the probability that two alleles from an individual are heterozygous. Then

$$H = (1 - p_I)h$$

so that $p_I = (h - H)/h$. An unbiased estimator of H is

$$\hat{H} = \frac{x_1}{s}$$

Thus we obtain a natural estimate of p_I:

$$\hat{p}_I = \frac{\sum(\hat{h} - \hat{H})}{\sum \hat{h}} \tag{26}$$

where we sum over all SNPs in our data.

4 Ascertainment

Now given an *ascertainment scheme* $\mathcal{A}$ which we think of as a probability distribution $\mathcal{A}(F)$ on a pair $F = (f_1, f_2)$ of allele frequencies then we define $N(\mathcal{A}) = E_{\mathcal{A}}(N(F))$ and $D(\mathcal{A}) = E_{\mathcal{A}}(D(F))$. Then $F_{st}(\mathcal{A})$ is naturally defined as

$$F_{st}(\mathcal{A}) = N(\mathcal{A})/D(\mathcal{A}) \tag{27}$$

Note that this is a definition, not a means of computing F_{st} unless the distribution $\mathcal{A}(F)$ is known explicitly. However if we observe a sample S of size M sampled with probability distribution $\mathcal{A}$ then it follows from the law of large numbers that

$$\begin{aligned} N(S)/M &\rightarrow N(\mathcal{A}) \\ D(S)/M &\rightarrow D(\mathcal{A}) \end{aligned}$$

Now in general, given populations A, B, F_{st} depends on the ascertainment scheme. Only in the case in which the ancestral population to A was panmictic and constant size (for all time, including time more remote than the split at the root), does F_{st} not depend on the ascertainment. However it is easy to see that F_{st} and indeed the joint frequency spectrum for A, B using alleles ascertained as polymorphic in A, will depend only on the drift time since the split of B at the root, and on the demography for A. As an example, in the simple case in which the allele frequency is in equilibrium at the root, which in

drift time is distant 0.3 from A and 0 from B we calculate, for a scaled mutation rate r which is 1 for time more remote than the split, and with an approximate numerical calculation:

r	F_{st}
0	.119
1	.128
10	.217

so that ascertainment can affect F_{st} by a factor of nearly 2.

So to summarize we define F_{st} to mean one of 4 things, with the context (hopefully) making it clear:

What F_{st} Can mean

Single SNP	N/D
Multiple SNPs	$\frac{\sum_i N_i}{\sum_i D_i}$
Ascertainment	$E(N)/E(D)$
Simple Demography.	$E(N)/E(D)$ (independent of ascertainment)

In no case is F_{st} a statistic, but a parameter that we can estimate.

5 Extreme F_{st}

It is worth mentioning that our estimates of F_{st} make sense, and are asymptotically consistent as the number of markers tends to infinity, even when the population sample sizes are very small. The minimal set seems to be ascertaining heterozygotes in population 1 by using 2 chromosomes from one individual, and then estimating F_{st} from 2 more chromosomes in population 1 and 2 in population 2. For a diploid species and autosomal data this requires sampling a total of 3 individuals. We then can estimate F_{st} from equations (19, 20, 21). This sample size is minimal, at least for our methods, as we need to be able to estimate the heterozygosity of each population, after ascertainment.

6 A related problem

We as before have population frequencies p_1, p_2, but now introduce an admixed population with frequency p_0. We have:

$$p_0 = \alpha p_1 + (1 - \alpha) p_2$$

for some $0 < \alpha < 1$. We wish to estimate α. Of course $p_0 - p_2 = \alpha(p_1 - p_2)$ and so

$$\alpha = \frac{E(p_0 - p_2)(p_1 - p_2)}{E(p_1 - p_2)^2} = T/B$$

B is the N we estimated earlier. So an unbiased estimator for B is

$$\hat{B} = (a_1/n_1 - a_2/n_2)^2 - \hat{h}/n_1 - \hat{h'}/n_2$$

where

$$\hat{h} = \frac{a_1(n_1 - a_1)}{n_1(n_1 - 1)}$$

with a similar expression for $\hat{h'}$. For T we write

$$T = ((p_0 - a_0/n_0) - (p_2 - a_2/n_2) + (a_0/n_0 - a_2/n_2))((p_1 - a_0/n_0) - (p_2 - a_2/n_2) + (a_1/n_1 - a_2/n_2))$$

Let

$$T' = (a_0/n_0 - a_2/n_2)(a_1/n_1 - a_2/n_2)$$

Then T' is an estimator of T and the bias of T' is

$$E(p_2 - a_2/n_2)^2 = p_2(1 - p_2)/n_2$$

We have already shown that an estimate of this is $\hat{h'}/n_2$. Summarizing:

$$\hat{T} = (a_0/n_0 - a_2/n_2)(a_1/n_1 - a_2/n_2) - \hat{h'}/n_2 \quad (28)$$

$$\hat{B} = (a_1/n_1 - a_2/n_2)^2 - \hat{h}/n_1 - \hat{h'}/n_2 \quad (29)$$

are unbiased estimates of T, B respectively. Summing $\hat{T}, \hat{B}$ over many loci will estimate α. This estimate is obviously asymptotically consistent.

Acknowledgment:
Conversations with Steven Schaffner and Alkes Price were very helpful in educating me here.

References

[1] Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L Price. Estimating and interpreting fst: the impact of rare variants. *Genome research*, 23(9):1514–1521, 2013.

[2] P. J. Bickel and K.A. Doksum. *Mathematical statistics: Basic Ideas and selected topics*. Holden-Day, 1977.

[3] L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes*. Princeton University Press, 1994.

[4] S. Karlin and H.M. Taylor. *A second course in stochastic processes*. Academic Press., 1981.

[5] M. Kimura. Solution of a process of random genetic drift with a continuous model. *PNAS*, 41:144–150, 1955.

[6] M. Nei. *Molecular evolutionary genetics*. Columbia University Press, 1987.

[7] N.J. Patterson. How old is the most recent ancestor of two copies of an allele? *Genetics*, 169:1093–1104, 2005.

[8] G.A. Watterson. Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.*, 10:239–253, 1976.