Author date: 2021-12-07

A formal definition of F_{ST}

E. Castedo Ellerman (castedo@castedo.com)

NOTE: A newly published article on F_{ST} [1] looks like a promising source of a better formal definition.

The name F_{ST} was used as early as [2] for a certain measure between populations. In the many decades since, the name has been used to have slightly different meanings. The recent publication [3] covers a long list of different F_{ST} meanings in many papers over the decades. Most of the papers cover additional concepts which are not required to define and gain intuition on F_{ST}. And the mathematical details behind a formal definition are spread across many papers.

This documents gives a simple formal definition to F_{ST}, equivalent to [3] and [4]. This simple definition also makes clear how F_{ST} is precisely a ratio of variances.

The Definition

Random variables A_S, A_T and D model uncertainty for F_{ST}:

- A_S and A_T for the allele found in a random gamete from the "**S**ub-population" and "**T**op" population, respectively
- D for the random **d**ecent, **d**rift, or **d**ivergence of the "sub-population" from the "top" population

"Top" population can mean ancestral population (as in [3]), or it can mean "total" population (the original meaning in [2]).

Given assumptions

- A_S and A_T take values of 0 or 1
- A_T and D are independent
- $\mathrm{E}(A_S) = \mathrm{E}(A_T)$

the definition follows

$$F_{ST} := \frac{\mathrm{Var}(\mathrm{E}(A_S|D))}{\mathrm{Var}(A_T)}$$

Convenient Expectations

Conveniently, expectations of allele variables are allele frequencies. The following variable definition will be convenient

$$p := \mathrm{E}(A_T)$$

Due to the assumptions of F_{ST} the following are conveniently true

$$\begin{aligned} p = \mathrm{E}(A_S) \qquad & p = \mathrm{E}\left(A_T^2\right) \\ \mathrm{E}(A_S) = \mathrm{E}\left(A_S^2\right) \quad & \mathrm{E}(A_S|D) = \mathrm{E}\left(A_S^2|D\right) \end{aligned}$$

and

$$\mathrm{Var}(A_T) = \mathrm{E}\left(A_T^2\right) - \mathrm{E}(A_T)^2 = p - p^2 = p(1-p)$$

F_{ST} as variance explained or uncertainty reduced

In light of the following theorem, F_{ST} can be interpreted as allele variance explained by random descent/drift/divergence. Alternatively, an interpretation can also be allele uncertainty reduced by knowing descent/drift/divergence.

Theorem 1

$$\mathrm{Var}(A_T) = \mathrm{Var}(\mathrm{E}(A_S|D)) + \mathrm{E}(\mathrm{Var}(A_S|D))$$

Proof

$$\begin{aligned} \mathrm{Var}(\mathrm{E}(A_S|D)) &= \mathrm{E}\Big(\mathrm{E}(A_S|D)^2\Big) - \mathrm{E}(\mathrm{E}(A_S|D))^2 \\ &= \mathrm{E}\Big(\mathrm{E}(A_S|D)^2\Big) - p^2 \\ \mathrm{E}(\mathrm{Var}(A_S|D)) &= \mathrm{E}\Big(\mathrm{E}\left(A_S^2|D\right) - \mathrm{E}(A_S|D)^2\Big) \\ &= p - \mathrm{E}\Big(\mathrm{E}(A_S|D)^2\Big) \end{aligned}$$

Unbiased Estimators

Consider observing two independent random descents/drifts/divergences D_1 and D_2 under the assumptions for D of F_{ST}. Furthermore, for each $j \in \{1,2\}$, consider observing n_j independent random gametes within each resulting sub-population. Define $n_1 + n_2$ independent observed alleles $A_{S,j,i}$ with i indexing sampled gametes within each sampled sub-populations resulting from the independent descents/drifts/divergences. For convenience define the following:

$$\hat{p}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} A_{S,1,i} \quad \hat{p}_2 := \frac{1}{n_2}\sum_{i=1}^{n_2} A_{S,2,i}$$

The "Hudson" estimator of F_{ST} is defined in [3] as

$$\frac{(\hat{p}_1 - \hat{p}_2)^2 - \frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}{\hat{p}_1(1-\hat{p}_2) + \hat{p}_2(1-\hat{p}_1)}$$

We first show that the denominator is an unbiased estimator of $2\operatorname{Var}(A_T)$. With $\hat{p}_1$ and $\hat{p}_1$ independent it follows:

$$\begin{aligned}\operatorname{E}(\hat{p}_1(1-\hat{p}_2)+\hat{p}_2(1-\hat{p}_1)) &= \operatorname{E}(\hat{p}_1)\operatorname{E}(1-\hat{p}_2)+\operatorname{E}(\hat{p}_2)\operatorname{E}(1-\hat{p}_1)\\ &= 2p(1-p)\\ &= 2\operatorname{Var}(A_T)\end{aligned}$$

We now show the "Hudson" numerator is an unbiased estimator of $2\operatorname{Var}(\operatorname{E}(A_S|D))$. For each $j \in \{0,1\}$ define:

$$\hat{v}_j := \frac{1}{n_j-1}\sum_{i=1}^{n_j}(A_{S,j,i}-\hat{p}_i)^2$$

It follows as a classic unbiased estimator of variance [5] that:

$$\operatorname{E}(\hat{v}_j) = \operatorname{E}(\operatorname{Var}(A_S|D_1))$$

Since $\operatorname{E}(\operatorname{Var}(A_S|D)) = \operatorname{Var}(A_T) - \operatorname{E}(\operatorname{Var}(A_S|D))$, it follows that an unbiased estimator of $2\operatorname{Var}(\operatorname{E}(A_S|D))$ is

$$\hat{p}_1(1-\hat{p}_2)+\hat{p}_2(1-\hat{p}_1)-\hat{v}_1-\hat{v}_2$$

We now show this is equivalent to the "Hudson" numerator. Note that

$$\begin{aligned}\hat{p}_1(1-\hat{p}_2)+\hat{p}_2(1-\hat{p}_1) &= \hat{p}_1+\hat{p}_2-2\hat{p}_1\hat{p}_2\\ \hat{v}_i &= \hat{p}_i-\hat{p}_i^2+\frac{\hat{p}_i(1-\hat{p}_i)}{n_i-1}\\ (\hat{p}_1-\hat{p}_2)^2 &= \hat{p}_1^2+\hat{p}_2^2-2\hat{p}_1\hat{p}_2\end{aligned}$$

it follows that

$$\hat{p}_1(1-\hat{p}_2)+\hat{p}_2(1-\hat{p}_1)-\hat{v}_1-\hat{v}_2 = (\hat{p}_1-\hat{p}_2)^2-\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1}-\frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}$$

which is the numerator in the "Hudson" estimator in [3].

References

1. Ochoa A, Storey JD. Estimating FST and kinship for arbitrary population structures. Feldman MW, editor. PLOS Genetics. 2021;17: e1009241–. doi:10.1371/journal.pgen.1009241
2. Wright S. THE GENETICAL STRUCTURE OF POPULATIONS. Annals of Eugenics. 1949;15: 323–354. doi:10.1111/j.1469-1809.1949.tb02451.x
3. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. Genome Research. 2013;23: 1514–1521. doi:10.1101/gr.154831.113
4. Patterson NJ. Notes on Fst. Available: https://reich.hms.harvard.edu/sites/reich.hms.harvard.edu/files/inline-files/fstnote_1.pdf
5. DeGroot MH, Schervish MJ. Probability and statistics. 3rd ed. Boston: Addison-Wesley; 2002.