Xerxes Theoretical Background

Derivations and Statistical Details of the fstats and ras commands

Stephan Schiffels

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Started Writing: August 2023
Last Updated: September 2023

The FStats command

The program **`xerxes fstats`** within the poseidon-hs package serves as a one-stop shop for various SNP-based statistics that can be computed from genotype data, including F-Statistics, pairwise mismatch rate, FSt and likely more statistics in the future. F-Statistics were defined in a series of papers by Nick Patterson, David Reich and others, and comprehensively formally described in Patterson et al. 2012.

F-Statistics

The most basic F-Statistics available in **`fstats`** are:

- F2vanilla(A, B) = $\langle (a-b)^2 \rangle$
- F3vanilla(A, B; C)=$\langle (c-a)\,(c-b) \rangle$
- F4(A, B; C, D)=$\langle (a-b)\,(c-d) \rangle$

Here, capital letters A, B, C, D stand for individuals or groups of individuals, defined via a Mini-Language used in trident forge, or as group definitions in FStats-Configuration files described here: http://www.poseidon-adna.org/#/xerxes?id=input-via-a-configuraton-file.

Small letters a, b, c, d here stand for the observed allele frequencies (i.e. the fractions of observed non-missing alternative alleles over observed non-missing alleles within groups A, B, C, D). The average $\langle \cdot \rangle$ is an average over all available SNPs. For example, we have

$$\left\langle (a-b)^2 \right\rangle = \frac{\sum_{i=1}^{L} (a_i - b_i)^2 \, I[a_i, b_i]}{\sum_{i=1}^{L} I[a_i, b_i]} \tag{1}$$

where

$$I[a_i, b_i] = \begin{cases} 1 & \text{if } a_i \text{ and } b_i \text{ are both non-missing} \\ 0 & \text{else} \end{cases} \tag{2}$$

Here, a_i and b_i denote allele frequencies of entity A and B at SNP i. Non-missing here means that at least one allele in each groups is non-missing, such that a sample allele frequency is defined. Parts of group A and B could be missing, though.

Bias-correction

Many of the population genetic use-cases and properties of F-Statistics (described in Patterson et al. 2012) are derived from *population* allele frequencies, that is, precise frequencies of an entire population. In practice, however, we only have a finite number of samples (in the most extreme case, one individual) to estimate population frequences.

More precisely, given *population* allele frequencies a' and b', we consider *sample* allele frequencies a and b as imperfect *estimators* of the population allele frequencies.

To reason about biases of these estimators, we first denote a noise-model. Under the assumption that the n samples of population A are independent samples, the sampling probability for observing k non-reference alleles within a sample of n samples, given a population frequency a' is a binomial distribution:

$$p_n[k \mid a'] = \text{Binom}[k, n \mid a'] = \binom{n}{k} (a')^k (1-a')^{n-k} \tag{3}$$

We can then define the estimator

$$a = k / n \tag{4}$$

and can show that this estimator is indeed unbiased:

```
In[•]:= Expectation[k / n, k ≈ BinomialDistribution[n, aPrime]]
Out[•]=
aPrime
```

i.e. that its expectation is just the population allele frequency a' itself.

This fact can be used to show that F4-statistics using sample allele frequencies are unbiased estimators of the same statistics using population frequencies:

$$\begin{aligned} F4[a, b, c, d] &= \langle (a-b)(c-d) \rangle = \langle a\,c \rangle - \langle a\,d \rangle - \langle b\,c \rangle + \langle b\,d \rangle \\ &= \langle a \rangle \langle c \rangle - \langle a \rangle \langle d \rangle - \langle b \rangle \langle c \rangle + \langle b \rangle \langle d \rangle \\ &= a'c' - a'd' - b'c' + b'd' \\ &= F4[a', b', c'd'] \end{aligned} \tag{5}$$

where in the second step we have made used of the fact that because the terms in the average-brackets are based on separate samples of individuals, the average over their product factorises. So for example we have

$$\langle a\,c \rangle = \langle a \rangle \langle c \rangle = a'c'. \tag{6}$$

Now, the same is *not* true for quadratic terms. Concretely, consider the vanilla – F_2 statistic:

$$F_{2,\text{vanilla}}(a,b) = \left\langle (a-b)^2 \right\rangle = \left\langle a^2 - 2\,a\,b + b^2 \right\rangle = \left\langle a^2 \right\rangle - 2\langle a\,b \rangle + \left\langle b^2 \right\rangle \tag{7}$$

Here, the middle term is again unproblematic, see equation 6. But the quadratic terms are not unbiased. For example, we have

```
In[•]:= Expectation[(k / n)^2, k ≈ BinomialDistribution[n, aPrime]]
Out[•]=
aPrime - aPrime^2 + aPrime^2 n
----------------------------------
n
```

or

$$= \left(a'\right)^2 + \frac{a'\,(1-a')}{n} \tag{8}$$

In other words $\langle a^2 \rangle$ *overestimates* $(a')^2$ by a term inversely proportional to $1/n$, so the smaller the sample size the larger the bias.

Fortunately, we can *correct* this bias, by using a different estimator for $(a')^2$. Specifically, we can try to simply subtract the term itself from the estimator, using a instead of a':

```
In[•]:= Expectation[k^2 / n^2 - k / n (1 - k / n) / n, k ≈ BinomialDistribution[n, a]]
Out[•]=
(a - a^2 + a^2 n^2) / n^2
```

which turns out not to be quite right yet. It turns out, replacing n by $n-1$ in the denominator does the job:

```
In[•]:= Expectation[k^2 / n^2 - k / n (1 - k / n) / (n - 1), k ≈ BinomialDistribution[n, a]]
Out[•]=
a^2
```

So we now have an unbiased estimator for the square of the allele frequency

$$\left\langle a^2 - \frac{a\,(1-a)}{n} \right\rangle = \left(a'\right)^2 \tag{9}$$

which is what we need to describe unbiased estimators for F2 and F3 statistics. The following are the estimators proposed in Patterson et al. 2012 (in Appendix A, page 1089):

$$\begin{aligned} F_2(a,b) &= \left\langle (a-b)^2 - \frac{a(1-a)}{n_a - 1} - \frac{b(1-b)}{n_a - 1} \right\rangle \\ F_3(a,b,c) &= \left\langle (c-a)(c-b) - \frac{c(1-c)}{n_c - 1} \right\rangle \end{aligned} \tag{10}$$

We can easily show that they are indeed unbiased:

$$\begin{aligned} F_2(a,b) &= \left\langle (a-b)^2 - \frac{a(1-a)}{n_a - 1} - \frac{b(1-b)}{n_a - 1} \right\rangle \\ &= \left\langle a^2 - 2ab + b^2 - \frac{a(1-a)}{n_a - 1} - \frac{b(1-b)}{n_a - 1} \right\rangle \\ &= \left\langle a^2 - \frac{a(1-a)}{n_a - 1} \right\rangle + \left\langle b^2 - \frac{b(1-b)}{n_a - 1} \right\rangle - 2\langle ab \rangle \\ &= (a')^2 - 2a'b' + (b')^2 \\ &= (a' - b')^2 \\ &= F_2(a', b') \end{aligned} \tag{11}$$

where we have used the two identities defined above in equations 6 and 9.

Similarly, we have

$$F_3(a,b,c) = \left\langle (c-a)(c-b) - \frac{c(1-c)}{n_c - 1} \right\rangle$$

$$= \left\langle c^2 - a\,c - b\,c - a\,b - \frac{c(1-c)}{n_c - 1} \right\rangle$$

$$= \left\langle c^2 - \frac{c(1-c)}{n_c - 1} \right\rangle - \langle a\,c \rangle - \langle b\,c \rangle - \langle a\,b \rangle$$

$$= (c')^2 - a'\,c' - b'\,c' - a'\,b'$$

$$= (c' - a')\,(c' - b').$$

Pseudo-haploid sample sizes and bias-correction

So, as we see, for estimators to be unbiased, we need to subtract terms from F2 and F3 statistics, which depend on the sample size. More specifically, on the number of chromosomes sampled to obtain allele frequency estimates. The usual case with high-quality diploid genetic data, is that given a sample from n_d diploid individuals, there are $2\,n_d$ chromosomes. However, with low-quality ancient DNA, a common scheme to obtain genotype estimates is a simple random-sampling scheme of sequencing reads aligning to a given SNP position.

In such a haploid sampling scheme, the resulting number of “chromosomes” that one uses for esitmating allele frequencies is n_d instead of $2\,n_d$.

In xerxes, we make use of the “Genotype_Ploidy” column in the Janno File, to compute the correct number of chromosomes, since the genotype file itself does not provide this information (it is always coded as pseudo-diploid, even if the calling is haploid). If this column is missing, or contains missing data, xerxes will output a warning and assume that the data is fully diploid!

Other Statistics

Jackknife-Estimation of the standard error

Frequency-Ascertainment (experimental)

The RAS command

References

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. “Ancient Admixture in Human History.” Genetics 192 (3): 1065–93.