Population genetics snippets for genepop

Peter Beerli

August 30, 2015

Contents

0.1Basics **1**

0.2Exact test **2**

0.3Fixation indices **4**

0.4Isolation by Distance **5**

0.5Further Reading **8**

0.6References **8**

0.7Disclaimer **9**

1 Basics

Genepop delivers basic population genetic statistics. For example, test on the deviation from Hardy-Weinberg equilibrium or FST estimations. One of the key measurements are the number or ratio of heterozygotes in a population. We can test whether the observed numbers match the expectation of the Hardy-Weinberg proportions.

For example, if we have a single locus with two alleles A and B, then we can recognize heterozygotes as having a genotype of AB. Homozygotes will have AA or BB.

Parents AA and BB will have offspring AB. Parents AA and AB will have AA, AB when we assume random pairing of gametes. Parents AB and AB will have AA, 2AB, BB, again, assuming random pairing.

We can construct a table with the alleles and calculate the expected frequencies from the marginals of the table, assuming that the frequency of A is p and the frequency of B is $q = 1 - p$.

	A	B
A	p^2	pq
B	pq	q^2

The expected homozygosity is $p^2 + q^2$, and the expected heterozygosity is $2pq$. We could now wonder whether the observed genotypes match the expected and do a HWE test. This could be done with an exact test or a χ^2 approximation.

2 Exact test

Assume we have collected frogs in two ponds and caught these genotypes:

Pond	AA	AB	BB	AC	AD	marginals
1	1	1	1	1	0	4
2	0	0	0	2	1	3
marginal	1	1	1	3	1	

Of interest would be whether the two ponds belong to the same population or whether they two are differentiated so that we should recognize them as two populations. We are interested whether the pattern we see is likely by chance alone or whether there is something going on. We could now calculate the probability seeing this exact table, which is

$$\frac{n_1!n_2!n_{AA}!n_{AB}!n_{BB}!n_{AC}!n_{AD}!}{n!n_{1AA}!n_{1AB}!n_{1BB}!n_{1AC}!n_{1AD}!n_{2AA}!n_{2AB}!n_{2BB}!n_{2AC}!n_{2AD}!}.$$

This is the frequency of all the number of patterns being generated from the marginals given enumeration of the observation pattern, using the numbers in the table we get:

$$\frac{4!3!1!1!1!3!1!}{7!\quad 1!1!1!1!0!\quad 0!0!0!2!1!} = \frac{4!3!3!}{7!2!} = \frac{3}{35}$$

Enumerating all possible patterns with fixed marginals lead to this list with its frequencies:

Pond	AA	AB	BB	AC	AD	Frequency
1	1	1	1	1	0	$\frac{3}{35}$
2	0	0	0	2	1	
1	1	1	1	0	1	$\frac{1}{35}$
2	0	0	0	3	0	
1	1	1	0	1	1	$\frac{3}{35}$
2	0	0	1	2	0	
1	1	1	0	2	0	$\frac{3}{35}$
2	0	0	1	1	1	
1	1	0	0	2	1	$\frac{3}{35}$
2	0	1	1	1	0	
1	1	0	0	3	0	$\frac{1}{35}$
2	0	1	1	0	1	
1	1	0	1	1	1	$\frac{3}{35}$
2	0	1	0	2	0	
1	1	0	1	2	0	$\frac{3}{35}$
2	0	1	0	1	1	
1	0	1	1	1	1	$\frac{3}{35}$
2	1	0	0	2	0	
1	0	1	1	2	0	$\frac{3}{35}$
2	1	0	0	1	1	
1	0	1	0	2	1	$\frac{3}{35}$
2	1	0	1	1	0	
1	0	1	0	3	0	$\frac{1}{35}$
2	1	0	1	0	1	
1	0	0	1	2	1	$\frac{3}{35}$
2	1	1	0	1	0	
1	0	0	1	3	0	$\frac{1}{35}$
2	1	1	0	0	1	
1	0	0	0	3	1	$\frac{1}{35}$
2	1	1	1	0	0	

The probabilities add to 1.0, we can calculate a p-value as the sum of p_i if p_i is smaller or equal to the observed p.

$$\mathrm{p-value} = \sum_{p_i \le p_{\mathrm{obs}}} p_i$$

For our problem here the p-value is 1.0.

3 Fixation indices

We saw in Basics that the heterozygosities can be calculated from the allele frequencies and also look at observed heterozygosity. In the Basics we handled only a single population, but there may be many subpopulations that add complications, for example 'Is a system with subpopulations more or less diverse than one without? A thought experiment may help: think of a population of mice and some cats, the population of mice is structured in that there are mice on the left and mice on the right and the cats are in the middle:

mice (all AA) cat mice (all BB)

If we treat the mice as a single population we can calculate the **heterozygosity H** as 0.5 assuming there are the same number of mice left and right. The frequency of allele A and B is p=0.5, q=0.5, respectively. This leads to an expected genotype composition of AA 0.25 and AB 0.5 and BB 0.25. The observed heterozygosity H is 0.0. But if we look only within the subpopulations we will get correct answers about the allele frequencies and heterozygosities (p=1.0, q=0.0, H=2pq=0.0).

We can define:

H_I mean observed heterozygosity of an individual with subpopulation

H_S mean expected heterozygosity of an individual in a randomly mating subpopulation

H_T expected heterozygosity of randomly mating whole population

We can now define a fixation index F so that it gives us a scaled version of these heterozygosities so that we can compare them among different studies (Wright 1951).

- $F_{IS} = \frac{H_S - H_I}{H_S}$
- $F_{IT} = \frac{H_T - H_I}{H_T}$
- $F_{ST} = \frac{H_T - H_S}{H_T}$

F_{IS} measures how different the heterozygosity of an individual is compared to a subpopulation; the range of the index goes from -1 to 1, from all individuals are heterozygote to no heterozygotes detected in the sample. Example: if all individuals are heterzygote AB then the allele frequency is 0.5, thus the $H_S = 2pq = 2 \times 0.5 \times 0.5 = 0.5$, leading to 0.5-1.0/0.5 = -1., With no heterozygote detected we may have different allele frequencies p because we could see AA, BB at various combination in the sample, but filling in we still get: $2p(1-p) - 0.0/(2p(1-p) = 1.0$

F_{IT} measures the observed heterozygosity with respect to the total population.

F_{ST} measures the expected heterozygosity with respect to the total population. It has a range of 0 to 1, with 0 as there is no differentiation among the subpopulations (they are a panmictic unit) and with 1 where the subpopulations are completely separated from each other with independent histories. Example: We encounter the same number of AA in subpopulation 1 and BB in subpopulation 2: $H_T = 2pq = 0.5$, $H_{S1} = H_{S2} = 2p_iq_i = 0.0$; leading to $(0.5 - 0.0)/0.5 = 1.0$, the opposite with AB in population 1 and AB in population 2: $(0.5 - ((0.5 + 0.5)/2))/0.5 = 0.0$. But if we have all AA in population 1 and AB in population 2 we get: $(2 * 2/3 * 1/3 - ((2 * 1.0 * 0.0 + 2 * 0.5 * 0.5)/2))/(2 * 2/3 * 1/3) = 7/16$.

These F-indices are reated to each by this equality

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$$

Several other equalities have been established by several authors: for example:

$$F_{ST} = \frac{V_a}{pq}$$

relating FST to the variance in allele frequencies, assuming a two allele system. Or relating it to the loss of heterozygosity in a population (no mutation):

$$F_{ST} = 1 - (1 - \frac{1}{2N})^t$$

where t is time in generations, N is the population size. And, last example, the relation of FST to the coalescent times

$$F_{ST} = (t - t_0)/t$$

where t is the time of coalescence of all individuals (the most recent common ancestor to all individuals) and t_0 is the average time of the last coalescent in all subpopulations. This formula suggests that we can correlate heterozygosity, population size, and genealogy with the observed variance of alleles in the population.

4 Isolation by Distance

We can use pairwise F_{ST} to get some idea how subpopulations are connected to each other, for example we would assume the the number of migrants may depend on the distance between subpopulations, so we can correlate F_{ST} with geographic distance. It turns out that Rousset () advocates a variation of this by using

$$\frac{F_{ST}}{1 - F_{ST}}$$

and commonly use the log distance instead of linear distance if the dataset contains distances that vary strongly in magnitude, the goal is then to fit a linear model:

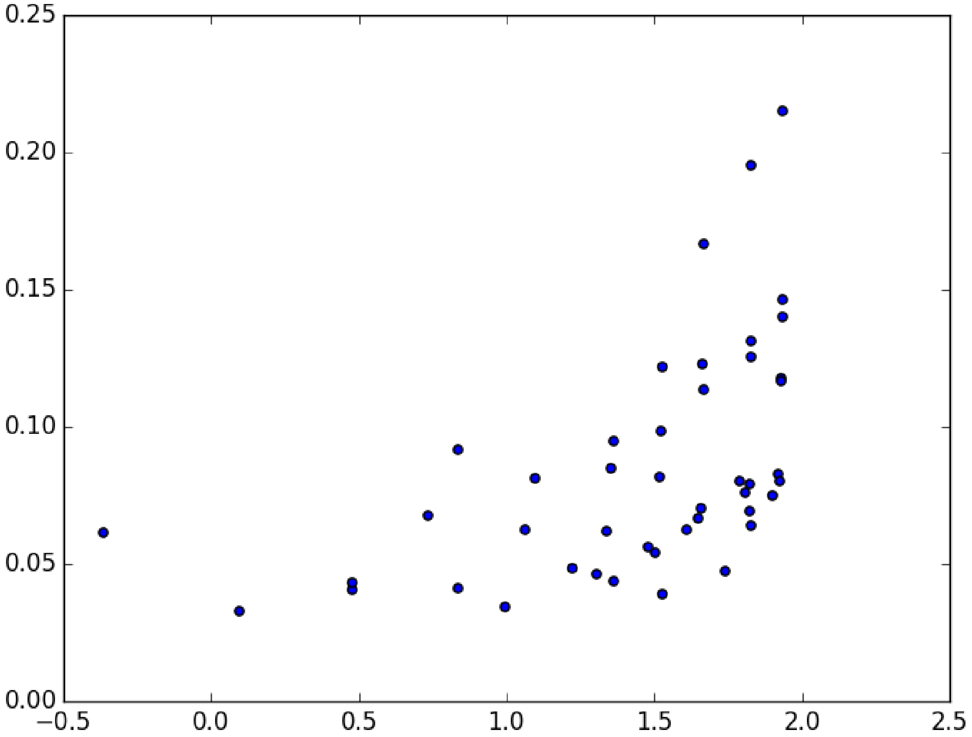
$$\frac{F_{ST}^{(i,j)}}{1 - F_{ST}^{(i,j)}} = b + aD^{(i,j)}$$

Running simulated data that was generated using a linear stepping stone model with 10 loci and 10 subpopulations we get the following tables and estimates of a and b.

```
Output file from the program Isolde
-----------------------------------

10 populations (/home/wbiomed/genepop/cgi-bin/tmp/215302/215302)

Fst/(1-Fst) estimates:

 0.0614
 0.0408  0.0329
 0.0918  0.0679  0.0430
 0.0812  0.0625  0.0345  0.0410
 0.0946  0.0847  0.0620  0.0465  0.0484
 0.1221  0.0982  0.0820  0.0540  0.0563  0.0437
 0.1665  0.1137  0.1228  0.0704  0.0668  0.0626  0.0389
 0.1952  0.1312  0.1254  0.0695  0.0790  0.0758  0.0800  0.0474
 0.2151  0.1461  0.1400  0.1166  0.1178  0.0802  0.0826  0.0748  0.0642

Ln(distance):

-0.3665
 0.4759  0.0940
 0.8340  0.7321  0.4759
 1.0972  1.0614  0.9962  0.8340
 1.3641  1.3536  1.3368  1.3053  1.2241
 1.5272  1.5228  1.5160  1.5040  1.4775  1.3641
 1.6674  1.6655  1.6626  1.6577  1.6473  1.6116  1.5272
 1.8269  1.8263  1.8253  1.8236  1.8203  1.8098  1.7903  1.7411
 1.9326  1.9324  1.9319  1.9312  1.9297  1.9252  1.9173  1.8998  1.8269

Fitting Fst/(1-Fst) to a + b ln(distance)
a =  0.0272397, b =  0.04018382
1000 permutations
(statistic: Spearman Rank correlation coefficient):

Test of isolation by distance (One tailed Pvalue):
 Pr(correlation > observed correlation) =0.00000 under null hypothesis

Other one tailed Pvalue:
```

```
Pr(correlation < observed correlation) =1.00000 under null hypothesis
```

The slope and test and lotting the points shows clearly that there is isolation by distance.



5 Further Reading

- Kent Holsinger's lecture on F-statistics:

6 References

To prepare this summary, I used the Shane's Simple Guide to F-statistics () , Raymond and Rousset (1995); and the Genepop manual.

7 Disclaimer

This text was written by Peter Beerli, Florida State University for a course on practical population genetics inference, Fall 2015. These notes are licensed under the Creative Commons Attribution-ShareAlike License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.