Analyzing the genetic structure of populations: a Bayesian approach

Introduction

Our review of Nei's G_{st} and Weir and Cockerham's θ illustrated two important principles:

1. It's essential to distinguish *parameters* from *estimates*. *Parameters* are the things we're really interested in, but since we always have to make inferences about the things we're really interested in from limited data, we have to rely on *estimates* of those parameters.

2. This means that we have to identify the possible sources of sampling error in our estimates and to find ways of accounting for them. In the particular case of Wright's F-statistics we saw that, there are two sources of sampling error: the error associated with sampling only some individuals from a larger universe of individuals within populations (*statistical sampling*) and the error associated with sampling only some populations from a larger universe of populations (*genetic sampling*).[1]

It shouldn't come as any surprise that there is a Bayesian way to do what I've just described. As I hope to convince you, there are some real advantages associated with doing so.

The Bayesian model

I'm not going to provide all of the gory details on the Bayesian model. In fact, I'm only going to describe two pieces of the model.[2] First, a little notation:

$$\begin{aligned} n_{11,i} &= \#\text{ of } A_1A_1 \text{ genotypes} \\ n_{12,i} &= \#\text{ of } A_1A_2 \text{ genotypes} \\ n_{22,i} &= \#\text{ of } A_2A_2 \text{ genotypes} \end{aligned}$$

[1] The terms "statistical sampling" and "genetic sampling" are due to Weir [3].

[2] The good news is that to do the Bayesian analyses in this case, you don't have to write any `JAGS` code. I'll provide the code.

$$
\begin{aligned}
i &= \text{population index} \\
I &= \text{number of populations}
\end{aligned}
$$

These are the data we have to work with. The corresponding genotype frequencies are

$$
\begin{aligned}
x_{11,i} &= p_i^2 + f p_i (1 - p_i) \\
x_{12,i} &= 2 p_i (1 - p_i)(1 - f) \\
x_{22,i} &= (1 - p_i)^2 + f p_i (1 - p_i)
\end{aligned}
$$

So we can express the likelihood of our sample as a product of multinomial probabilities

$$
P(\mathbf{n}|\mathbf{p}, f) \propto \prod_{i=1}^{I} x_{11,i}^{n_{11,i}} x_{12,i}^{n_{12,i}} x_{22,i}^{n_{22,i}} \quad .
$$

Notice that I am assuming here that we have the same f in every population. It's easy enough to relax that assumption, but we won't worry about it here.

To complete the Bayesian model, all we need are some appropriate priors. Specifically, we so far haven't done anything to describe the variation in allele frequency among populations. Suppose that the distribution of allele frequencies among populations is well-approximated by a Beta distribution. A Beta distribution is convenient for many reasons, and it is quite flexible. Don't worry about what the formula for a Beta distribution looks like. All you need to know is that it has two parameters and that if these parameters are π and θ, we can set things up so that

$$
\begin{aligned}
\mathrm{E}(p_{ik}) &= \pi \\
\mathrm{Var}(p_{ik}) &= \pi(1 - \pi)\theta
\end{aligned}
$$

Thus π corresponds to $\bar{p}$ and θ corresponds to F_{st}.[3] Figure 1 illustrates the shape of the Beta distribution for different choices of π and θ. To complete the Bayesian model we need only to specify priors on π, f, and θ. In the absence of any prior knowledge about the parameters, a uniform prior on [0,1][4] is a natural choice.

The *Isotoma petraea* example

Here's the `JAGS` code to estimate f and θ:

[3] For any of you who happen to be familiar with the usual parameterization of a Beta distribution, this parameterization corresponds to setting $\nu = ((1 - \theta)/\theta)\pi$ and $\omega = ((1 - \theta)/\theta)(1 - \pi)$.

[4] `dunif(0,1)` in `JAGS` notation

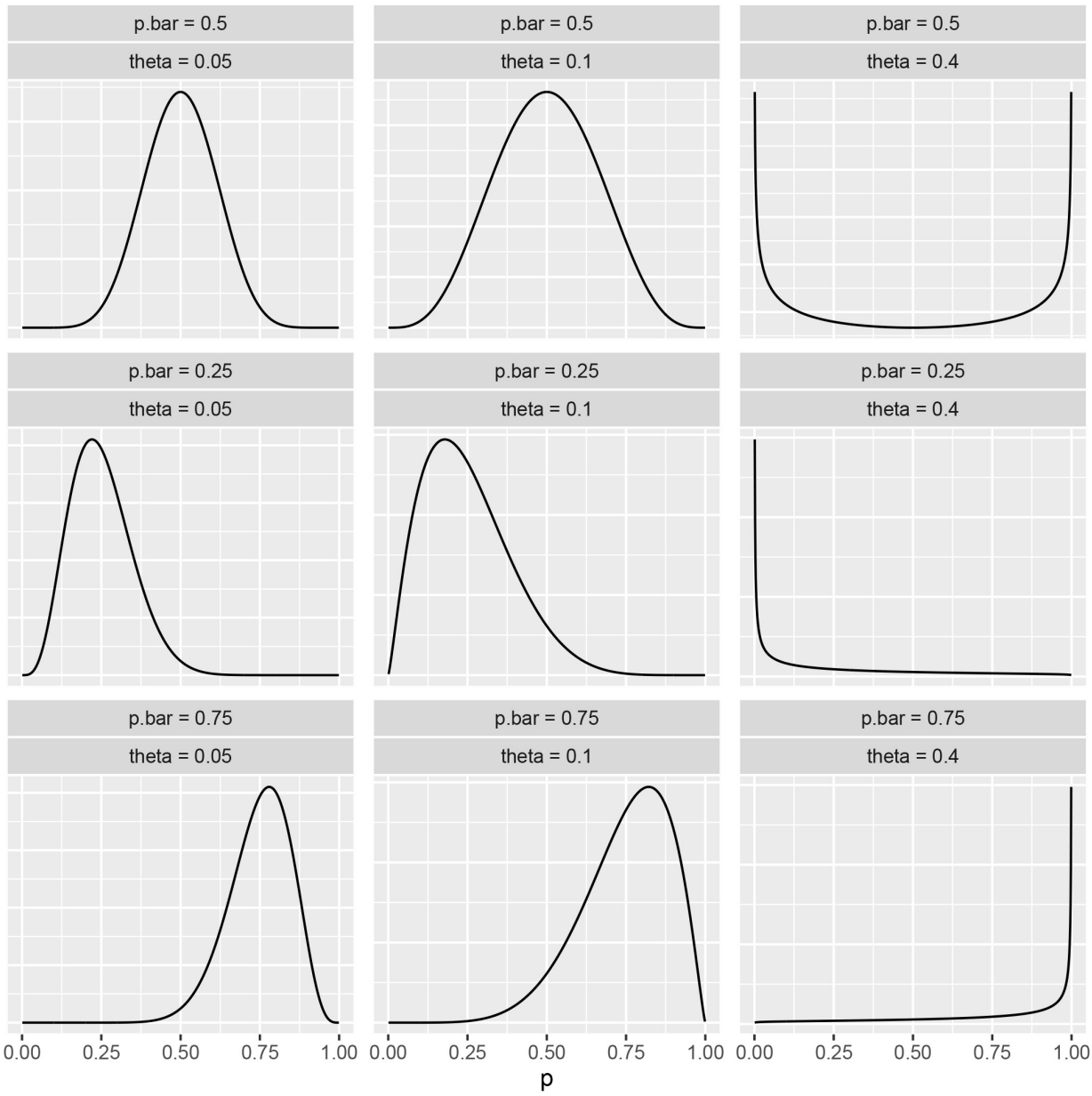

Figure 1: Shapes of the Beta distribution for different choices of π and θ. In the figure captions "p.bar" corresponds to π, and "theta" corresponds to θ.

```
model {
   ## genotype frequencies
   ##
   for (i in 1:n.pops) {
     for (j in 1:n.loci) {
       x[i,j,1] <- p[i,j]*p[i,j] + f*p[i,j]*(1-p[i,j])
       x[i,j,2] <- 2*(1-f)*p[i,j]*(1 - p[i,j])
       x[i,j,3] <- (1-p[i,j])*(1-p[i,j]) + f*p[i,j]*(1-p[i,j])
     }
   }

   ## likelihood
   ##
   for (i in 1:n.pops) {
     for (j in 1:n.loci) {
       n[i,j,1:3] ~ dmulti(x[i,j,], N[i,j])
     }
   }

   ## priors
   ##
   ## allele frequencies within populations
   ##
   for (i in 1:n.pops) {
     for (j in 1:n.loci) {
       p[i,j] ~ dbeta(alpha, beta)
     }
   }
   ## inbreeding coefficient within populations
   ##
   f ~ dunif(0, 1)

   ## theta (Fst)
   ##
   theta ~ dunif(0,1)

   ## pi
   ##
```

```
  for (i in 1:n.loci) {
    pi[j] ~ dunif(0,1)
  }

  ## parameters of the beta distribution
  ## the weird constraints are to ensure that both of them
  ## lie in [1, 1.0e4]
  ##
  for (i in 1:n.loci) {
    alpha[i] <- max(1, min(((1-theta)/theta)*pi[i], 1.0e4))
    beta[i]  <- max(1, min(((1-theta)/theta)*(1-pi[i]), 1.0e4))
  }
}
```

You'll also find an `R` script that calls this code. The relevant function has the very creative name of `analyze.data()`. It requires data in a very particular format, namely a list that consists of four named elements:

1. `n.pops`: The number of populations in the sample.
2. `n.loci`: The number of loci scored in the sample.
3. `n`: A `n.pops`×`n.loci`×3 matrix of genotype counts where in the final dimension the first entry corresponds to the number of $A1A1$ homozygotes, the second entry corresponds to the number of $A1A2$ heterozygotes, and the third entry corresponds to the number of $A2A2$ homozygotes.
4. `N`: An `n.pops`×`n.loci` matrix of sample sizes at each locus. This could be calculated automatically by `analyze.data()`, but I haven't written that code yet.

It's not too hard to get data into that format. `f-statistics.R` also provides a set of functions to construct that list from a CSV file in which each line corresponds with an individual, the first column (`pop`) is the population from which that individual was collected, and the remaining columns are the genotype (scored as 0, 1, 2) of the individual at a particular locus.

The *Isotoma petraea* data come to us in a somewhat different format, so there's also a script (`isotoma.R`) that constructs the necessary input list and calls `analyze.data()`. If you look at the code, you'll see that I've specified `n.chains=5`. That allows me to check convergence by looking at `Rhat`. If you run the code, here's what you'll get (except for MCMC error):

```
> print(fit)
Inference for Bugs model at "f-statistics.txt", fit using jags,
 5 chains, each with 30000 iterations (first 25000 discarded), n.thin = 5
 n.sims = 5000 iterations saved
          mu.vect sd.vect   2.5%    25%    50%    75%  97.5%  Rhat n.eff
f           0.527   0.097  0.327  0.464  0.533  0.595  0.698 1.001  5000
theta       0.112   0.051  0.024  0.076  0.108  0.143  0.223 1.003  3000
deviance   46.679   4.848 38.924 43.096 46.095 49.594 57.610 1.001  3900

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 11.8 and DIC = 58.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

It's easy to modify the code to consider two special cases:

- $f = 0$: This corresponds to the case when genotypes within populations are in Hardy-Weinberg proportions. Implemented in `f-statistics-f0.jags`.
- $\theta = 0$: This corresponds to the case when population allele frequencies are identical across populations, i.e., there is no genetic differentiation. Implemented in `f-statistics-t0.jags`

Before we go any further though and start comparing these models, I need to explain what that DIC number is.

The Deviance Information Criterion

A widely used statistic for comparing models in a Bayesian framework is the Deviance Information Criterion. `R2jags` calculates an estimate of it for us automatically, but you need to know that if you're serious about model comparison, yous shouldn't rely on the DIC calculation from `R2jags` unless you've verified it.[5] Let's pretend for a moment that I never wrote that last sentence and talk about why the DIC statistic may be useful

The `deviance` is a measure of how well the model fits the data, specifically -2 times the average of the log likelihood values calculated from the parameters in each sample from

[5] If you're interested in learning more, feel free to ask, but I'm afraid both the explanation and the solution are a little complicated.

Model	deviance	pD	DIC
"Full" model	46.8	11.7	58.5
$f = 0$ model	72.3	11.2	83.5
$\theta = 0$ model	61.7	1.9	63.6

Table 1: DIC calculations for the *Isotoma petrae* example.

the posterior. `pD` is a measure of model complexity. Roughly speaking it is the number of parameters in the model.[6] DIC is a composite measure of how well the model does. It's a compromise between fit and complexity, and smaller DICs are preferred. A difference of more than 7-10 units is regarded as strong evidence in favor of the model with the smaller DIC.

In this case the minimum difference in DIC values is nearly 15 units, meaning that the full model is clearly preferred. We'll get to what that means biologically in just a moment. Why? I don't trust the DIC statistics I just reported here.

Back to F-statistics

This is one of those cases where I don't trust the calculation of DIC in `JAGS`. Why? Because I calculated it from scratch and got a very different result:[7]

```
Dbar: 46.6
Dhat: 40.7
pD:   5.9
DIC:  52.4
```

If you compare these results to what's reported from `R2jags`, you'll see that `Dbar` in my calculation corresponds to the average deviance in the `R2jags` output.[8] That's because they're both calculated as -2.0 times the log likelihood of the data, averaged across all posterior samples. The difference is in the estimates for `pD`. My version calculates it according to the

[6]Notice that there we estimated one more parameter in the full model than in the $f = 0$ model, and that there are only two parameters (π and θ) in the $\theta = 0$ model. Notice also if we just count parameters in the model, there are more than 12 parameters in the full model: 12 allele frequencies (one for each population), π, and θ. `pD` is less than that for complicated reasons that I'd be happy to explain if you're really interested. Otherwise, just accept it as a well-known fact that counting the number of parameters in a hierarchical model isn't as straightforward as you might have guessed.

[7]Just include `DIC=TRUE` in the arguments to `analyze.data()` and you'll get a printout of these results.

[8]Except for rounding error.

Model	Dbar	Dhat	pD	DIC
Full	46.6	40.7	5.9	52.4
$f = 0$	72.3	67.1	5.2	77.6
$\theta = 0$	61.8	59.8	2.0	63.8

Table 2: DIC statistics for the *Isotoma petraea* data.

Method	F_{is}	F_{st}	F_{it}
Direct	0.14	0.21	0.32
Nei	0.31	0.24	0.47
Weir & Cockerham	0.54	0.04	0.56
Bayesian	0.52 (0.34, 0.69)	0.11 (0.02, 0.23)	

Table 3: Comparison of F_{is} and F_{st} estimates calculated in different ways.

original definition [2] as the difference between Dbar and Dhat, -2.0 times the log likelihood of the data at the posterior mean of the parameters. R2jags calculates it differently. In both cases, DIC is just Dbar + pD, but since the estimates of pD are different so are the estimates of DIC.

In any case, it's easy to compare DIC from the three models simply by adding model="f0" (for the $f = 0$ model) or model="t0" (for the $\theta = 0$ model) to the argument list of analyze.data(). Table 2 summarizes the results.

The $f = 0$ has a much larger DIC than the full model, a difference of more than 20 units. Thus, we have strong evidence for inbreeding in these populations of *Isotoma petraea*.[9] The $\theta = 0$ model also has a DIC substantially larger than the DIC for the full model, a difference of more than 10 units. Thus, we also have good evidence for genetic differentiation among these populations.[10]

It's useful to look back and think about the different ways we've used the data from *Isotoma petraea* (Table 3). Several things become apparent from looking at this table:

- The direct calculation is very misleading. A population that has only one individual sampled carries as much weight in determining F_{st} and F_{is} as populations with samples of 20-30 individuals.

[9]Much stronger than the evidence we had for inbreeding in the ABO blood group data, by the way.

[10]It's important to remember that this conclusion applies *only* to the locus that we analyzed. Strong differentiation at this locus need not imply that there is strong differentiation at other loci.

- By failing to account for genetic sampling, Nei's statistics significantly underestimate F_{is}, while Weir & Cockerham's estimate is indistinguishable from the Bayesian estimates.

- It's not illustrated here, but when a reasonable number of loci are sampled, say more than 8-10, the Weir & Cockerham estimates and the Bayesian estimates are even more similar. But the Bayesian estimates allow for more convenient comparisons of different estimates, and the credible intervals don't depend either on asymptotic approximations or on bootstrapping across a limited collection of loci. The Bayesian approach can also be extended more easily to complex situations. You'll see an example of this in Project #2 when I ask you to identify F_{ST} outliers.

References

[1] K E Holsinger and L E Wallace. Bayesian approaches for the analysis of population structure: an example from *Platanthera leucophaea* (Orchidaceae). *Molecular Ecology*, 13:887–894, 2004.

[2] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[3] B S Weir. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA, 1996.

Creative Commons License