Testing Hardy-Weinberg

Introduction

Because the Hardy-Weinberg principle tells us what to expect concerning the genetic composition of a sample when no evolutionary forces are operating, one of the first questions population geneticists often ask is "Are the genotypes in this sample present in the expected, i.e., Hardy-Weinberg, proportions?" We ask that question because we know that if the genotypes are *not* in Hardy-Weinberg proportions, at least one of the assumptions underlying derivation of the principle has been violated, i.e., that there is some evolutionary force operating on the population, and we know that we can use the magnitude and direction of the departure to say something about what those forces might be. In particular, we now know that inbreeding leads to a deficiency of heterozygotes, and we know that the extent of that deficiency can be measured by f.[1]

What we haven't talked about is (a) how to estimate f from data and (b) how to tell whether we have good evidence that the estimate is positive (meaning that there's a deficiency of heterozygotes in the population) or negative. Both (a) and (b) pose more of a challenge than you might initially think. After all we also know that the numbers in our sample may differ from expectation just because of random sampling error. For example, Table 1 presents data from a sample of 1000 English blood donors scored for MN phenotype. M and N are co-dominant, so that heterozygotes can be distinguished from the two homozygotes. Clearly the observed and expected numbers don't look very different.[2] The differences semm likely to be attributable purely to chance, but we need some way of assessing that "likeliness."

---

[1] Quiz question: Which definition of f is relevant for determining whether there is a deficiency of heterozygotes?

[2] For the time being, I simply calculated the expected numbers in the way you'd tell your students in introductory biology to do it: (1) Use the sample frequency of M to estimate its population frequency. (This is a maximum-likelihood estimate, by the way. (2) Calculate the expected frequency of each genotype from the Hardy-Weinberg proportions. (3) Calculated the expected numbers of each genotype by multiplying the expected frequency of each by the total sample size.

		Observed	Expected
Phenotype	Genotype	Number	Number
M	mm	298	294.3
MN	mn	489	496.3
N	nn	213	209.3

Table 1: Adapted from Table 2.4 in [2] (from [1])

Testing Hardy-Weinberg

One approach to testing the hypothesis that genotypes are in Hardy-Weinberg proportions is quite simple. We can simply do a χ^2 or G-test for goodness of fit between observed and predicted genotype (or phenotype) frequencies, where the predicted genotype frequencies are derived from our estimates of the allele frequencies in the population.[3] There's only one problem. To do either of these tests we have to know how many degrees of freedom are associated with the test. How do we figure that out? In general, the formula is

$$
\begin{aligned}
\text{d.f.} \quad = \quad & (\text{\# of categories in the data -1 }) \\
& -(\text{\# number of parameters estimated from the data}) \quad .
\end{aligned}
$$

For this problem we have

$$
\begin{aligned}
\text{d.f.} \quad = \quad & (\text{\# of phenotype categories in the data - 1}) \\
& -(\text{\# of allele frequencies estimated from the data}) \\
= \quad & (3 - 1) - 1 \\
= \quad & 1 \quad .
\end{aligned}
$$

In the ABO blood group we have 4 phenotype categories, and 3 allele frequencies. That means that a test of whether a particular data set has genotypes in Hardy-Weinberg proportions will have $(4 - 1) - (3 - 1) = 1$ degrees of freedom for the test. Notice that this also means that if you have completely dominant markers, like RAPDs or AFLPs, you can't determine whether genotypes are in Hardy-Weinberg proportions because you have 0 degrees of freedom available for the test.

---

[3]If you're not familiar with the χ^2 or G-test for goodness of fit, consult any introductory statistics or biostatistics book, and you'll find a description. In fact, you probably don't have to go that far. You can probably find one in your old genetics textbook. Or you can just boot up your browser and head to Wikipedia: http://en.wikipedia.org/wiki/Goodness_of_fit.

Phenotype	A	AB	B	O	Total
Observed	862	131	365	702	2060

Table 2: Data on variation in ABO blood type.

An example

Table 2 exhibits data drawn from a study of phenotypic variation among individuals at the ABO blood locus:

The maximum-likelihood estimate of allele frequencies, assuming Hardy-Weinberg, is:[4]

$$
\begin{aligned}
p_a &= 0.281 \\
p_b &= 0.129 \\
p_o &= 0.590 \quad,
\end{aligned}
$$

giving expected numbers of 846, 150, 348, and 716 for the four phenotypes. $\chi^2_1 = 3.8$, $0.05 < p < 0.1$.

A Bayesian approach

We've already seen how to use JAGS to provide allele frequency estimates from phenotypic data at the ABO locus.

```
model {
   # likelihood
   pi[1] <- p.a*p.a + 2*p.a*p.o
   pi[2] <- 2*p.a*p.b
   pi[3] <- p.b*p.b + 2*p.b*p.o
   pi[4] <- p.o*p.o
   x[1:4] ~ dmulti(pi[],n)

   # priors
   a1 ~ dexp(1)
   b1 ~ dexp(1)
   o1 ~ dexp(1)
```

---

[4]Take my word for it, or run the EM algorithm on these data yourself.

```
    p.a <- a1/(a1 + b1 + o1)
    p.b <- b1/(a1 + b1 + o1)
    p.o <- o1/(a1 + b1 + o1)

    n <- sum(x[])
}
list(x=c(862, 131, 365, 702))
```

As you may recall, this produced the following results:

```
> source("multinomial.R")
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size: 20

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
  |**************************************************| 100%
Inference for Bugs model at "multinomial.txt", fit using jags,
 5 chains, each with 2000 iterations (first 1000 discarded)
 n.sims = 5000 iterations saved
         mu.vect sd.vect   2.5%    25%    50%    75%  97.5%  Rhat n.eff
p.a        0.282   0.008  0.266  0.276  0.282  0.287  0.297 1.001  5000
p.b        0.129   0.005  0.118  0.125  0.129  0.133  0.140 1.001  5000
p.o        0.589   0.008  0.573  0.584  0.589  0.595  0.606 1.001  5000
deviance  27.811   2.007 25.830 26.363 27.229 28.577 33.245 1.001  4400

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 = 2.0 and DIC = 29.8
DIC is an estimate of expected predictive error (lower deviance is better).
>
```

Now that we know about inbreeding coefficients and that they allow us to measure the departure of genotype frequencies from Hardy-Weinberg proportions, we can modify this a

bit and estimate allele frequencies without assuming that genotypes are in Hardy-Weinberg proportions.

```
model {
    # likelihood
    pi[1] <- p.a*p.a + f*p.a*(1-p.a) + 2*p.a*p.o*(1-f)
    pi[2] <- 2*p.a*p.b*(1-f)
    pi[3] <- p.b*p.b + f*p.b*(1-p.b) + 2*p.b*p.o*(1-f)
    pi[4] <- p.o*p.o + f*p.o*(1-p.o)
    x[1:4] ~ dmulti(pi[],n)

    # priors
    a1 ~ dexp(1)
    b1 ~ dexp(1)
    o1 ~ dexp(1)
    p.a <- a1/(a1 + b1 + o1)
    p.b <- b1/(a1 + b1 + o1)
    p.o <- o1/(a1 + b1 + o1)

    f ~ dunif(0,1)

    n <- sum(x[])
}
```

This simple change produces the following results:

```
> source("abo-inbreeding.R")
Compiling model graph
    Resolving undeclared variables
    Allocating nodes
    Graph Size: 30

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
  |**************************************************| 100%
Inference for Bugs model at "abo-inbreeding.txt", fit using jags,
 5 chains, each with 2000 iterations (first 1000 discarded)
 n.sims = 5000 iterations saved
```

```
          mu.vect sd.vect    2.5%     25%     50%     75%   97.5%   Rhat n.eff
f           0.403   0.139   0.059   0.326   0.429   0.505   0.599  1.013   550
p.a         0.349   0.027   0.290   0.332   0.352   0.368   0.392  1.006   960
p.b         0.161   0.014   0.132   0.152   0.162   0.171   0.186  1.006   840
p.o         0.490   0.039   0.429   0.461   0.485   0.514   0.577  1.006  1000
deviance   25.200   2.416  22.249  23.411  24.716  26.342  31.206  1.007   470

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 = 2.9 and DIC = 28.1
DIC is an estimate of expected predictive error (lower deviance is better).
>
```

Notice that the allele frequency estimates have changed quite a bit and that the posterior mean of f is about 0.40. On first appearance, that would seem to indicate that we have lots of inbreeding in this sample. **BUT** it's a human population. It doesn't seem very likely that a human population is really that highly inbred.

Indeed, take a closer look at *all* of the information we have about that estimate of f. The 95% credible interval for f is between 0.06 and 0.60. That suggests that we don't have much information at all about f from these data.[5] How can we tell if the model with inbreeding is better than the model that assumes genotypes are in Hardy-Weinberg proportions?

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.[6] Fortunately, in this case, the results are fairly reliable.[7] The results of the DIC calculations for our two models are summarized in Table 3.

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 the

---

[5]That shouldn't be too surprising, since any information we have about f comes indirectly through our allele frequency estimates.

[6]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.

[7]You'll just have to trust me on this unless you asked the last question.

Model	deviance	pD	DIC
$f > 0$	25.2	2.9	28.1
$f = 0$	27.8	2.0	29.9

Table 3: DIC calculations for the ABO example.

posterior. `pD` is a measure of model complexity, roughly speaking the number of parameters in the model.[8] 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 difference in DIC values is only about 0.8, so we have very little evidence for $f > 0$ model for these data. This is consistent with the weak evidence for a departure from Hardy-Weinberg that was revealed in the χ^2 analysis.

References

[1] T E Cleghorn. MNS gene frequencies in English blood donors. *Nature*, 187:701, 1960.

[2] P W Hedrick. *Genetics of Populations*. Jones and Bartlett Publishers, Sudbury, MA, 2nd ed. edition, 2000.

Creative Commons License

---

[8]Notice that we estimated 2 parameters in the $f = 0$ model (2 allele frequencies) and 3 parameters in the $f > 0$ model (2 allele frequencies plus the inbreeding coefficient).