

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.

Of course 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. The differences seem likely to be attributable purely to chance, but we need some way of assessing that “likeliness.”

Phenotype	Genotype	Observed Number	Expected Number
M	MM	298	294.3
MN	MN	489	496.3
N	NN	213	209.3

Table 1: From Table 2-3 in Hedrick, *Genetics of Populations*, 2nd ed., Jones & Bartlett Publishers, New York, 2000

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

Table 2: Data on variation in ABO blood type.

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.¹ 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

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

For this problem we have

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

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.

An example

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

¹If 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.

The maximum-likelihood estimate of allele frequencies, assuming Hardy-Weinberg, is:²

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

giving expected numbers of 846, 150, 348, and 716 for the four phenotypes. $\chi_1^2 = 3.8$, $0.05 < p < 0.1$.

A Bayesian approach

We saw last time how to use WinBUGS 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)
  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 results in Figure 1.

²Take my word for it, or run the EM algorithm on these data yourself.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
p.a	0.2813	0.007668	1.237E-4	0.2664	0.2812	0.2962	1001	5000
p.b	0.1293	0.005405	5.948E-5	0.119	0.1292	0.1404	1001	5000
p.o	0.5894	0.008327	1.242E-4	0.5734	0.5894	0.6059	1001	5000

Figure 1: Results from WinBUGS analysis of the ABO data assuming genotypes are in Hardy-Weinberg proportions.

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[])
}

```

```
list(x=c(862, 131, 365, 702))
```

This produces the results in Figure 2

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
f	0.3991	0.1362	0.009539	0.07008	0.429	0.5907	1001	5000
p.a	0.3474	0.0258	0.001707	0.2905	0.3514	0.3885	1001	5000
p.b	0.1612	0.0139	8.712E-4	0.1321	0.1624	0.1861	1001	5000
p.o	0.4914	0.03765	0.002573	0.4327	0.4854	0.5758	1001	5000

Figure 2: Results from WinBUGS analysis of the ABO data relaxing the assumption that genotypes are in Hardy-Weinberg proportions.

Model	Dbar	Dhat	pD	DIC
$f > 0$	24.900	22.319	2.581	24.480
$f = 0$	27.827	25.786	2.041	29.869

Table 3: DIC calculations for the ABO example.

Notice that the allele frequency estimates have changed quite a bit and that the posterior mean of f is about 0.41. On first appearance, that would seem to indicate that we have lots of inbreeding in this sample. **BUT** it's a human population. That doesn't seem very likely. Take a closer look. The 95% credible interval for f is between 0.06 and 0.55. That suggests that we don't have much information at all about f from these data.³ 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. It can be calculated automatically in WinBUGS, just by clicking the right button. The results of the DIC calculations for our two models are summarized in Table 3.

Dbar and Dhat are measures of how well the model fits the data. Dbar is the posterior mean log likelihood, i.e., the average of the log likelihood values calculated from the parameters in each sample from the posterior. Dhat is the log likelihood at the posterior mean, i.e., the log likelihood calculated when all of the parameters are set to their posterior mean. pD

³That shouldn't be too surprising, since any information we have about f comes indirectly through our allele frequency estimates.

is a measure of model complexity, roughly speaking the number of parameters in the model. 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 about 5.5, so we have some evidence for $f > 0$ model for these data, even though they are from a human population. But the evidence is not very strong. This is consistent with the weak evidence for a departure from Hardy-Weinberg that was revealed in the χ^2 analysis.

Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.