A Worked Example using the ‘HWEBayes’ Package

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington, Box 357232 Seattle, WA 98195–7232, USA

Email: jonno@u.washington.edu

1 Introduction

The methods described here are based on Wakefield (2009). We first give notation for the k allele case. Let p_{ij} be the frequency of genotype A_iA_j, and n_{ij} be the observed count, $i,j=1,...,k, j \geq i$. Under independence of sampling the likelihood is multinomial:

$$\Pr(\boldsymbol{n}|\boldsymbol{p}) = \frac{n!}{\prod_{i,j=1,j\geq i}^{k} n_{ij}!} \prod_{i,j=1,j\geq i}^{k} p_{ij}^{n_{ij}} \tag{1}$$

where $\boldsymbol{n} = (n_{11}, n_{12}, ..., n_{kk})$ and $\boldsymbol{p} = (p_{11}, p_{12}, ..., p_{kk})$ are $k(k+1)/2$-dimensional vectors and $n = \sum_{i,j=1,j\geq i}^{k} n_{ij}$. Under Hardy-Weinberg Equilibrium (HWE) $p_{ii} = p_i^2$, $i = 1,...,k$ and $p_{ij} = 2p_ip_j$, $i,j=1,...,k, i \neq j$.

We can parameterize the saturated model as $p_{ii} = p_i^2 + p_i \sum_{j \neq i} p_j f_{ij}$, $p_{ij} = 2p_ip_j(1-f_{ij})$ so that we have introduced a set of fixation indices f_{ij} (Weir 1996); $f_{ij} = 0$ for all $i \neq j$ recovers the HWE model. Under the HWE model the genotype frequencies arise as the product of the constituent allele frequencies, i.e. as $p_{ii} = p_i^2$, $p_{ij} = 2p_ip_j$. Hence with HWE we have just k parameters, the allele frequencies, $p_1, ..., p_k$.

We may examine posterior distributions of f_{ij} to discover the reasons for departure from HWE; a positive/negative f_{ij} indicates a deficiency/excess of heterozygotes of type A_iA_j. The fixation indices are on awkward ranges: $1 - \frac{1}{2p_ip_j} \leq f_{ij} \leq 1$ (so that the lower bound can extend below -1 which is not true for the model with a single f, see below), which can produce difficulties for frequentist inference.

An interesting sub-model corresponds to $f_{ij} = f$, and is known as the inbreeding model since all pairs of alleles frequencies are assumed to be equally perturbed. Under this model: $p_{ii} = p_i^2 + p_i(1-p_i)f$, $p_{ij} = 2p_ip_j(1-f)$, and $f_{\min} = \frac{-p_{\min}}{1-p_{\min}} \leq f \leq 1$ where $p_{\min}$ is the minimum of the allele frequencies. Under HWE the multinomial likelihood (1) takes the form

$$\Pr(\boldsymbol{n}|\boldsymbol{p}) = \frac{2^{\sum_{i=1,j>i}^{k} n_{ij}} n!}{\prod_{i,j=1,j\geq i}^{k} n_{ij}!} \prod_{i=1}^{k} p_i^{2n_{ii} + \sum_{j>i} n_{ij}}. \tag{2}$$

2 Methods

For Bayesian estimation we can specify conjugate Dirichlet priors under the null and under the saturated alternative that is parameterized in terms of the genotype frequencies. For $k > 2$ the single f model cannot be examined under a conjugate analysis, and even in the $k = 2$ case we cannot carry out a conjugate analysis if we wish to specify a prior for f directly. The prior we use for the single f model is of the form

$$\pi(\boldsymbol{p}, f) = \pi(\boldsymbol{p}) \times \pi(f|\boldsymbol{p})$$

where $\pi(f|\boldsymbol{p})$ allow us to specify a prior that gives $f_{\text{min}} < f < 1$. We choose to reparameterize as $\lambda = \log[(f - f_{\text{min}})/(1 - f)]$ and assume $\lambda \sim N(\mu_\lambda, \sigma_\lambda)$. We specify two quantiles of f, with associated probabilities, and then numerically solve for the prior parameters $\mu_\lambda, \sigma_\lambda$. For small k we choose a rejection algorithm for obtaining samples from the posterior, using samples from the prior. For larger values of k this algorithm becomes inefficient and we use MCMC and `WinBUGS`. The simplest way to see if the rejection algorithm is feasible is to run the algorithm and see how long it takes!

For testing, the calculation of Bayes factors under conjugate priors is straightforward. For the non-conjugate models we use importance sampling, with the proposal taken as either the prior (for small k), or a normal distribution based on moments from an MCMC run. Wakefield (2009) contains details of all of the above.

3 Illustration: Estimation

We illustrate using the four allele data previously analyzed by a number of authors (Guo and Thompson 1992; Wakefield 2009). The data are given in Table 1.

A_1	0	3	5	3
A_2		1	18	7
A_3			1	5
A_4				2
	A_1	A_2	A_3	A_4

Table 1: Data on four alleles.

We first illustrate the use of the function `DirichSampHWE` which can be used to simulate samples from the prior or from the posterior when the prior is Dirichlet (so that we have a conjugate analysis) under the HWE model.

We first simulate under the from the Dir(1,1,1,1) prior under the HWE model. Figure 1 gives histogram representations of the (marginal) posteriors of the four allele frequencies — these are theoretically identical, but differ due to sampling variability.

```
library(HWEBayes)
# Four allele example
bvec0 <- c(1,1,1,1)
nvec0 <- rep(0,10)
# First sample from the prior under the null
priorsampH0 <- DirichSampHWE(nvec0,bvec0,nsim=1000)
par(mfrow=c(2,2))
hist(priorsampH0$pvec[,1],xlab=expression(p[1]),main="")
hist(priorsampH0$pvec[,2],xlab=expression(p[2]),main="")
hist(priorsampH0$pvec[,3],xlab=expression(p[3]),main="")
hist(priorsampH0$pvec[,4],xlab=expression(p[4]),main="")
```

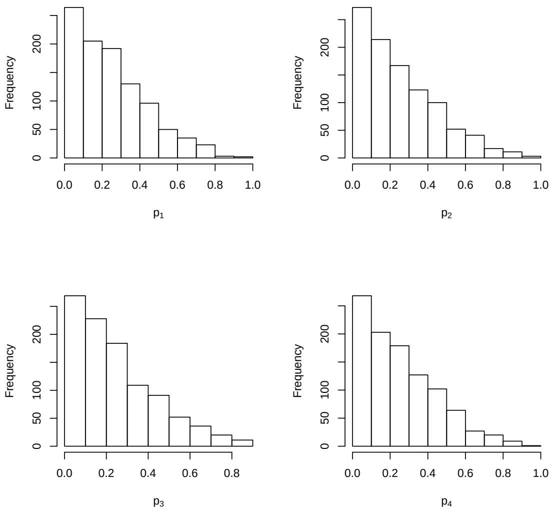

Figure 1: Samples from the prior Dir(1,1,1,1) under HWE.

We next obtain samples, again using `DirichSampHWE`, from the posterior. The function `HWEmodelsMLE` obtains the MLEs under the HWE, single f and saturated models. In Figure 2 we give histograms of the posteriors of the allele frequencies, along with the MLEs ($\widehat{p}_1 = 0.12, \widehat{p}_1 = 0.33, \widehat{p}_1 = 0.33, \widehat{p}_1 = 0.21$). As expected for this prior (which contains little information for estimation, relative to the data), the MLEs are close to the center of the posteriors. Notice the way the data are input: $n_{11}, n_{12}, n_{13}, n_{14}, n_{22}, n_{23}, n_{24}, n_{33}, n_{34}, n_{44}$.

```
data(DiabRecess)
nvec <- DiabRecess
postsampH0 <- DirichSampHWE(nvec,bvec0,nsim=1000)
MLE4 <- HWEmodelsMLE(nvec)
par(mfrow=c(2,2))
hist(postsampH0$pvec[,1],xlab=expression(p[1]),main="")
abline(v=MLE4$qhat[1],col="red")
hist(postsampH0$pvec[,2],xlab=expression(p[2]),main="")
abline(v=MLE4$qhat[2],col="red")
hist(postsampH0$pvec[,3],xlab=expression(p[3]),main="")
abline(v=MLE4$qhat[3],col="red")
hist(postsampH0$pvec[,4],xlab=expression(p[4]),main="")
abline(v=MLE4$qhat[4],col="red")
```

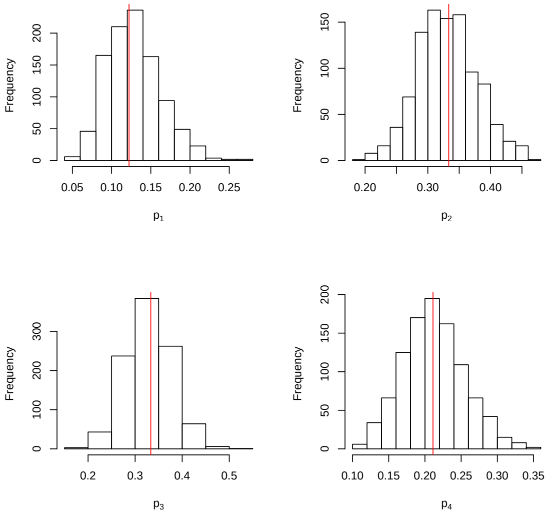

Figure 2: Posterior samples under HWE and a Dir(1,1,1,1) prior. MLEs are shown as vertical red lines.

We now turn to estimation under the saturated alternative, via the function `DirichSampSat`. We first simulate from the prior Dir(1,1,1,1,1,1,1,1,1,1) and display a number of summaries in Figure 3. Specifically, for illustration, we give p_{11}, p_{12}, p_{22}, p_1, p_2 and f_{12}.

```
bvec1 <- rep(1,10)
nvec1 <- rep(0,10)
```

```
priorsampH1sat <- DirichSampSat(nvec=nvec1,bvec1,nsim=1000)
par(mfrow=c(2,3))
hist(priorsampH1sat$pvec[,1],xlab=expression(p[11]),main="")
hist(priorsampH1sat$pvec[,2],xlab=expression(p[12]),main="")
hist(priorsampH1sat$pvec[,3],xlab=expression(p[22]),main="")
hist(priorsampH1sat$pmarg[,1],xlab=expression(p[1]),main="")
hist(priorsampH1sat$pmarg[,2],xlab=expression(p[2]),main="")
hist(priorsampH1sat$fixind[,2,1],xlab=expression(f[12]),main="")
```

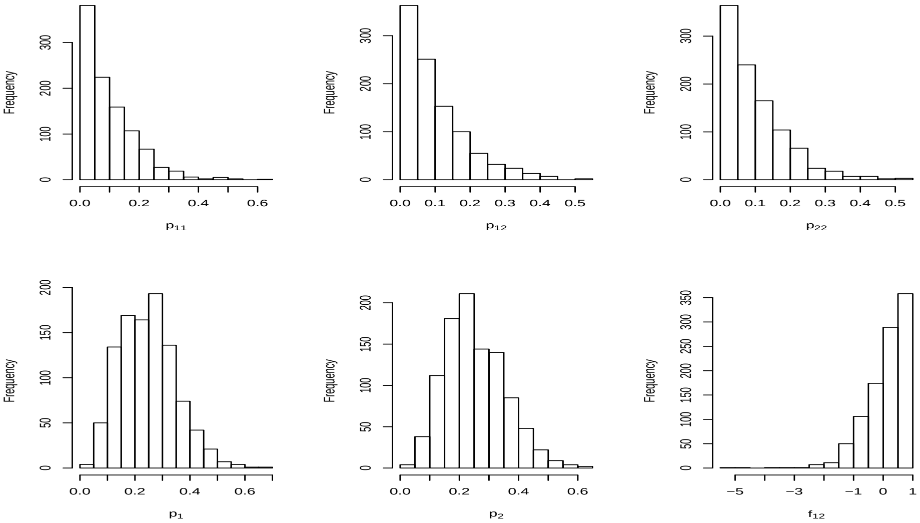

Figure 3: Prior samples under the saturated model and a Dir(1,1,1,1,1,1,1,1,1) prior.

The posterior samples are obtained in similar fashion with the `DirichSampSat` function. Figure 4 gives various summaries, again with the MLEs indicated.

```
# Sample from the saturated posterior for the 4 allele data
postsampH1sat <- DirichSampSat(nvec,bvec1,nsim=1000)
par(mfrow=c(2,3))
hist(postsampH1sat$pvec[,1],xlab=expression(p[11]),main="")
abline(v=MLE4$phat[1,1],col="red")
hist(postsampH1sat$pvec[,2],xlab=expression(p[12]),main="")
abline(v=MLE4$phat[1,2],col="red")
hist(postsampH1sat$pvec[,3],xlab=expression(p[22]),main="",xlim=c(0,.3))
abline(v=MLE4$phat[2,2],col="red")
```

```
hist(postsampH1sat$pmarg[,1],xlab=expression(p[1]),main="")
abline(v=MLE4$qhat[1],col="red")
hist(postsampH1sat$pmarg[,2],xlab=expression(p[2]),main="")
abline(v=MLE4$qhat[2],col="red")
hist(postsampH1sat$fixind[,2,1],xlab=expression(f[12]),main="")
abline(v=MLE4$fixind[1,2],col="red")
```

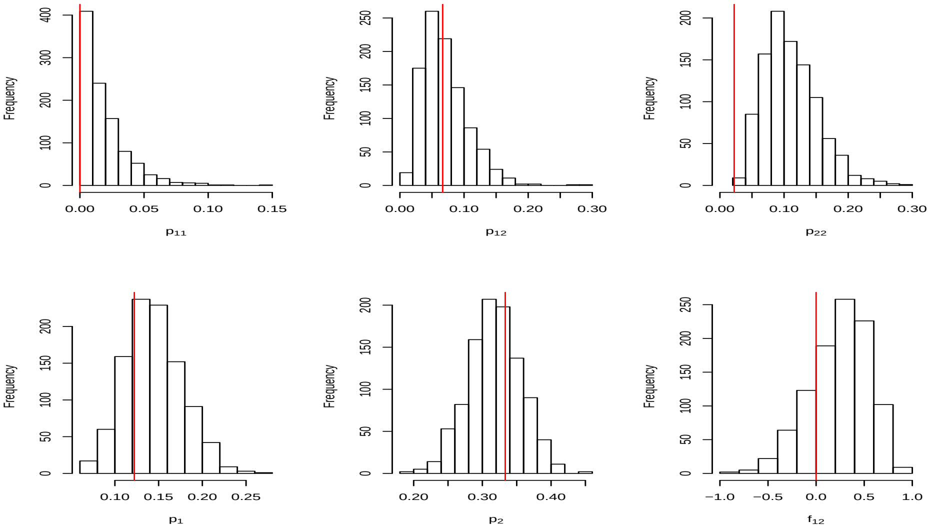

Figure 4: Posterior samples under the saturated model and a Dir(1,1,1,1,1,1,1,1,1) prior. MLEs are shown as vertical red lines.

We now carry out the single f example. We specify the 50% and 95% points of the prior for f as 0 and 0.26, and then numerically find μ_λ and σ_λ. Next we sample from the posterior using a rejection algorithm and in Figure 5 plot the resultant posteriors for p_1, p_2, p_3, p_4 and f, along with the MLEs.

```
# Single f example
bvec <- c(1,1,1,1)
# Find the parameters for the prior for f
init <- c(-3,log(1.1)) # Good starting values needed
lampr <- LambdaOptim(nsim=10000,bvec=bvec,f1=0,f2=0.26,p1=0.5,p2=0.95,init)
nsim <- 100
postsampf1 <- SinglefReject(nsim,bvec,lambdamu=lampr$lambdamu,
                             lambdasd=lampr$lambdasd,nvec)
```

```
par(mfrow=c(2,3))
hist(postsampf1$psamp[,1],xlab=expression(p[1]),main="")
abline(v=MLE4$fqhat[1],col="red")
hist(postsampf1$psamp[,2],xlab=expression(p[2]),main="")
abline(v=MLE4$fqhat[2],col="red")
hist(postsampf1$psamp[,3],xlab=expression(p[3]),main="")
abline(v=MLE4$fqhat[3],col="red")
hist(postsampf1$psamp[,4],xlab=expression(p[4]),main="")
abline(v=MLE4$fqhat[4],col="red")
hist(postsampf1$fsamp,xlab="f",main="")
abline(v=MLE4$fsingle,col="red")
```

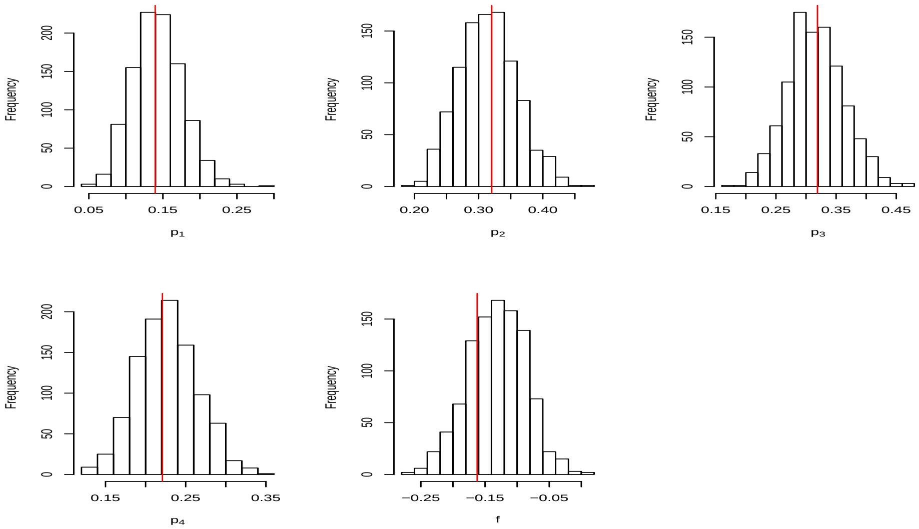

Figure 5: Posterior samples under the single f model. MLEs are shown as vertical red lines.

4 Illustration: Hypothesis Testing

We now consider the previous example but move from estimation to hypothesis testing, using Bayes factors:

$$\frac{\Pr(\boldsymbol{n}|H_0)}{\Pr(\boldsymbol{n}|H_1)}$$

A Bayes factor above (below) 1 indicates that the data are more (less) likely under the null than the alternative. Under conjugate Dirichlet priors the required normalizing constants are available in closed form. The following code evaluates the normalizing constant under the null (PrnH0) and under the saturated alternative (PrnH1sat), to give the Bayes factor (BFH0H1sat). Here $\Pr(\boldsymbol{n}|\text{ HWE }) = 1.39 \times 10^{-11}$ and $\Pr(\boldsymbol{n}|\text{ saturated}) = 1.88 \times 10^{-10}$ to give a Bayes factor of 0.074. Hence the data are $1/0.074 = 13.5$ times more likely under the saturated alternative than the null. For the single f model we obtain 1.4×10^{-10} (from the use of the singlefreject function above) so that the data are 10 times more likely than under the null, bit slightly less likely than under the saturated model.

```
PrnH0 <- DirichNormHWE(nvec,bvec0)
PrnH1sat <- DirichNormSat(nvec,bvec1)
BFH0H1sat <- PrnH0/PrnH1sat
```

We now evaluate the normalizing constant under the single f model using importance sampling. There are two possibilities for proposals, either using a normal distribution with user-specified moments, or from the prior. Note that the prior proposal estimate is far more variable, and so more samples are needed. When I ran the code I obtained an estimate of the normalizing constant of 1.31×10^{-10} $(1.29 \times 10^{-10}, 1.33 \times 10^{-10})$ using the normal proposal, and 1.31×10^{-10} $(9.79 \times 10^{-11}, 1.35 \times 10^{-10})$ using the prior proposal. Hence the data are slightly less likely to have come from the single f model than the saturated model, but there is little difference.

```
alpha <- rep(1,4)
# First simulate from a normal proposal using mean vector and covariance
# matrix from a WinBUGS run
gmu <- c(-0.4633092,0.3391625,0.3397936,-3.5438008)
gsigma <- matrix(c(
0.07937341,0.02819656,0.02766583,0.04607996,
0.02819656,0.07091320,0.04023827,0.01657028,
0.02766583,0.04023827,0.07042278,0.01752266,
0.04607996,0.01657028,0.01752266,0.57273683),nrow=4,ncol=4)
est1 <- HWEImportSamp(nsim=5000,nvec,ischoice=1,lambdamu=lampr$lambdamu,
             lambdasd=lampr$lambdasd,alpha=alpha,gmu,gsigma)
# Now let's evaluate using the prior
est2 <- HWEImportSamp(nsim=20000,nvec,ischoice=2,lambdamu=lampr$lambdamu,
             lambdasd=lampr$lambdasd,alpha=alpha,gmu,gsigma)
```

5 Discussion

Testing for HWE is routinely carried out in controls in genome-wide association studies, as a quality control method. In this context SNP data are the norm with 100s of thousands SNPs being examined. A Bayes factor may be calculated to examine the evidence for departures from HWE. In Wakefield (2009) the function `HWEDirichBF2` was used to calculate the Bayes factors, with conjugate Dirichlet priors under the null and alternative, with parameters (1,1) and (1,1,1), respectively.

`WinBUGS` code to carry out estimation for the single f model may be found at

`http://faculty.washington.edu/jonno/software.html`

References

Guo, S.W. and Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics 48*, 361–372.

Wakefield, J. (2009). Bayesian methods for examining Hardy-Weinberg equilibrium. *Biometrics*.

Weir, B.S. (1996). *Genetic Data Analysis II.* Sunderland MA: Sinauer.