

Bayesian Inference using WinBUGS

Instructor: Sujit K. Ghosh

COURSE OUTLINE

1. Fundamentals of Bayesian Inference.
2. Multinomial Genotypic Counts.
3. Introduction to WinBUGS.
4. Sample codes.

Slide 29

1. Bayesian Inference

Subjective opinion:

“We are all inherently Bayesian, but usually it is difficult to behave as one”

- You are about to make your first submission to journal “J”
- You assess the chances of publication in that journal.
- You have an opinion, but the true probability is unknown.
- You submit your article and it is accepted. *Congrats!*
- What is your revised opinion on the acceptance probability for papers like yours?
- Would you use “1”?
- YES or NO? (circle your answer)

Slide 30

Two contentious notions of the Bayesian school of statistical inference:

1. Probability is a measure of degree of belief (rather than, e.g., a measure of long-run frequency)
2. Statistical inference is the orderly ranking of beliefs along a correspondingly subjective probability scale (rather than, say, the measurement of evidence based on data).

Reference:

Vieland, V. J. (1998) Bayesian Linkage Analysis, or: How I learned to stop worrying and love the posterior probability of linkage, *Am. J. Hum. Genet.*, **63**, 947-954.

Slide 31

Can you identify this person? (-:



Qn.: What is your prior guess?

Ans.:

(write your answer above)

Slide 32

Bayes Theorem

If B_1, \dots, B_k is a sequence of mutually exclusive and exhaustive events from a sample space, $\Omega = B_1 \cup \dots \cup B_k$ and $B_j \cap B_i = \phi$, then for any event $A \subseteq \Omega$,

$$\Pr(A) = \sum_{j=1}^k \Pr(A|B_j) \Pr(B_j) \quad (1)$$

and hence

$$\begin{aligned} \Pr(B_j|A) &= \frac{\Pr(A|B_j) \Pr(B_j)}{\Pr(A)} \\ &= \frac{\Pr(A|B_j) \Pr(B_j)}{\sum_{j=1}^k \Pr(A|B_j) \Pr(B_j)} \end{aligned} \quad (2)$$

For genetic applications we need to extend the notion of probability to the general case of random variables.

Throughout the lectures we will follow some general notations.

Slide 33

General notation for statistical inference

Statistical inference is the science of making conclusions about a 'population' from 'sample' drawn from that population.

θ : unobservable vector of quantities or population *parameters* of interest.

y : the observed sample data drawn from the population.

\hat{y} : unknown, but potentially observable quantities. Future *predicted* values.

x : explanatory variables or *covariates*.

$p(\cdot|\cdot)$: conditional probability distribution with arguments determined by the context.

$p(\cdot)$: marginal distribution.

$\Pr(\cdot)$: Probability of an event.

The goal is to make probabilistic statements about unobserved quantities (e.g. θ, \hat{y} , etc.) given the observed quantities (e.g. y, x , etc.).

Slide 34

Bayes Theorem for random variables

Recall the definitions of the random variables (see Dr. Browning's lectures).

Within a Bayesian framework *both* data and parameters are considered random. We now extend the Bayes Theorem to more general cases.

Using an extension of (1), we get

$$p(y) = \int p(y|\theta)p(\theta)d\theta, \quad (3)$$

which provides the marginal density of the data.

And using an extension of (2), we get

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \end{aligned} \quad (4)$$

$$\propto p(y|\theta)p(\theta), \quad (5)$$

which provides the posterior density of θ in the light of the data y .

Slide 35

Interpretation of Bayes Theorem:

A prior is a probability statement of some *prior knowledge* about parameter (before observing any *data*).

The user need to elicit the prior distribution $p(\theta)$ either from past experiments or from the expert knowledge. In the worst case, one may use the so-called 'non-informative' priors.

Posterior is the probability statement of the *updated knowledge* about the parameter in the light of *current data*.

All information about the parameter is contained in the posterior distribution given the current data.

Slide 36

Posterior distribution can be seen as a compromise between data and prior distribution.

$$\frac{\partial^2 \log(p(\theta|\mathbf{y}))}{\partial \theta^2} = \frac{\partial^2 \log(p(\mathbf{y}|\theta))}{\partial \theta^2} + \frac{\partial^2 \log(p(\theta))}{\partial \theta^2}$$

Posterior Info. = Data info. + Prior info.

$$E[\text{Var}(\theta|\mathbf{y})] = \text{Var}[\theta] - \text{Var}[E(\theta|\mathbf{y})]$$

The last equation implies that the posterior variance is on average smaller than the prior variance. This is quite intuitive, as the inclusion of data should increase the precision (by reducing the variance).

These features of the posterior distributions become extremely helpful when eliciting a prior distribution.

Slide 37

Sequential version of Bayes Theorem

Suppose we observe the data y_1, y_2, \dots, y_n sequentially one at a time.

Want to estimate θ . What happens if the experiments stops at m th step and we want an estimate of θ given the current data?

An alternative way to write the Bayes Theorem is

$$p(\theta|y_1, \dots, y_m) \propto p(y_m|\theta)p(\theta|y_1, \dots, y_{m-1}),$$

$m = 2, \dots, n$

where $p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$.

This is sometimes referred to as *Bayesian learning*.

“*Past posterior*” behaves like a “*present prior*” and is updated in the light of “*current data*.”

This feature can also be used to combine different experiments (meta-analysis).

Slide 38

Prediction:

Given data y , what is most likely future value of the data, \tilde{y} , if the sampling experiment is repeated?

This is answered using the *Posterior Predictive Distribution* of \tilde{y} given y ,

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y) d\theta \end{aligned} \quad (6)$$

The last equality is based on conditional independence of y and \tilde{y} given θ . Again, the above result is a consequence of axioms of probability.

Slide 39

Posterior summarization:

If possible a Bayesian will present the entire posterior distribution, $p(\theta|y)$. (graphically and/or analytically)

However, in most applications (especially when θ is of high dimension), various numerical summaries of the posterior distribution are desirable.

Posterior point estimates:

- Posterior mean: $E[\theta|y] = \int \theta p(\theta|y) d\theta$.
This is also called the Bayes estimate under the squared error loss.
- Posterior sd: $sd(\theta|y) = \sqrt{\text{Var}(\theta|y)}$,
where $\text{Var}(\theta|y) = E[\theta^2|y] - (E[\theta|y])^2$.
- Posterior percentiles: Fix $\alpha \in (0, 1)$.
Find θ_α , such that $\Pr(\theta \leq \theta_\alpha|y) = \alpha$. Then θ_α is called the 100α posterior percentile of θ , e.g., $\alpha = 0.5$ provides the posterior median of θ .

Slide 40

Posterior interval estimates:

- $100(1 - \alpha)\%$ equal-tail posterior interval for θ is obtained by finding the $100\frac{\alpha}{2}$ and $100(1 - \frac{\alpha}{2})$ percentiles of the posterior distribution.
- $100(1 - \alpha)\%$ Highest density region (HDR) for θ is obtained by finding a subset C of the form,

$$C = \{\theta : p(\theta|y) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant such that $Pr(C|y) = \int_C p(\theta|y)d\theta \geq 1 - \alpha$.

In most applications, it is easier to obtain $100(1 - \alpha)\%$ equal-tail posterior intervals. (e.g., using MCMC methods)

The software **WinBUGS** by default produces five-number summary of the posterior distribution, which consists of the *mean*, *median*, *sd*, *2.5%*, and *97.5% percentiles*.

Slide 41

Bayesian Hypothesis Testing:

Suppose that we want to test, $H_0 : \theta \in \Theta_0$ vs. $H_A : \theta \in \Theta_A$.

Let the prior probabilities of the hypotheses be $\pi_0 = Pr(H_0) = Pr(\theta \in \Theta_0)$ and $\pi_A = Pr(H_A) = Pr(\theta \in \Theta_A)$

π_0/π_A is called the *prior odds* on H_0 against H_A .

In the light of data, a Bayesian would naturally update the prior odds to *posterior odds*,

$$p_0/p_A = Pr(H_0|y)/Pr(H_A|y)$$

and compare posterior odds to prior odds using the *Bayes factor (BF)*,

$$BF = \frac{p_0/p_A}{\pi_0/\pi_A} = \frac{\pi_A \int_{\Theta_0} p(y|\theta)p(\theta)d\theta}{\pi_0 \int_{\Theta_A} p(y|\theta)p(\theta)d\theta}$$

Note that $BF > 1$ supports H_0 as compared to H_A and vice versa. In practice, it is convenient to work with $\log_{10} BF$ or $2 \log_e BF$.

Slide 42

Kass & Raftery (1995) JASA:

Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another.

Jeffreys (1961) suggested the following scale:

| $\log_{10} BF$ | BF | Evidence against H_0 |
|----------------|-----------|------------------------------------|
| 0 to 1/2 | 1 to 3.2 | Not worth more than a bare mention |
| 1/2 to 1 | 3.2 to 10 | Substantial |
| 1 to 2 | 10 to 100 | Strong |
| >2 | >100 | Decisive. |

Slide 43

The main conceptual difference:

As a *frequentist*, the random variables are supposed to capture relevant features of the process of *sampling from a population*, whereas...

in the *Bayesian* approach the machinery of random variables is used to express your *uncertainty about unknown quantities* conditional on the observations.

Statistical analysis is *fundamentally* an *inversion* process, since it aims at retrieving the “causes” (i.e. unknown parameters) from the “effects” (i.e. observations).

The Bayesian approach provides a formal way to quantify the uncertainty about the *causes given the effects*.

Slide 44

In fact, $L(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}|\mathbf{y})$ if we use a vague (possibly) improper prior $p(\boldsymbol{\theta}) \equiv 1$. This is equivalent to the *fiducial approach* of Fisher (1956).

Indeed, it often turns out that using the Bayesian formalism with relatively vague priors, produces procedures which perform well using traditional frequentist criteria (e.g. low MSE over repeated sampling)!

Hence, for making statistical inference it is better to take the unified perspective of *Bayes+frequentist* and not Bayes vs. frequentist.

In most applied work, for instance, it is useful to reason in a Bayesian way when formulating inferences and predictions, and to reason in a frequentist way when evaluating their performances.

Slide 45

Some highlights of Bayesian inference

1. Ability to formally incorporate prior information.
2. The reason for stopping experimentation does not affect the inference.
3. Answers are more easily interpretable by non-specialists.
(e.g. credible intervals as opposed to confidence intervals)
4. All analyses follow *directly* from the posterior; no separate theories of estimation, testing, multiple comparisons, etc. are needed.
5. Inferences are conditional on the data that is actually observed.
6. Bayes procedures automatically possess numerous optimality properties.
(e.g. consistency, admissibility, minimaxity, parsimony in model choice, etc.)

Slide 46

An Example: Estimating the probability of a female birth

Let y be the number of girls in n recorded births. Let θ be the proportion of female births.

How would you estimate the ratio of male to female birth rates, $\phi = \frac{1-\theta}{\theta}$?

Here, $Y|\theta \sim \text{Binomial}(\theta, n)$, i.e.,

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

where $y = 0, 1, \dots, n$ and $0 \leq \theta \leq 1$.

To complete the probability modeling we need a prior distribution for θ . We discuss several alternatives.

Slide 47

A class of “conjugate” prior distributions is given by, $\theta \sim \text{Beta}(b_1, b_2)$, i.e.

$$p(\theta) = \frac{\Gamma(b_1 + b_2)}{\Gamma(b_1)\Gamma(b_2)} \theta^{b_1-1} (1-\theta)^{b_2-1}$$

where $b_1 > 0, b_2 > 0$ are hyperparameters of the prior distribution and $\Gamma(b) = \int_0^\infty t^{b-1} e^{-t} dt$ denotes the *Gamma function*.

Using Bayes Theorem (5), the posterior distribution of θ is given by,

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^{y+b_1-1} (1-\theta)^{n-y+b_2-1} \end{aligned}$$

Hence, it follows that,

$$\theta|y \sim \text{Beta}(y + b_1, n - y + b_2).$$

Ex. Show that, if $b_1 = b_2 = b$, then $\phi|y \sim F_{2b, 2b}$, where $F_{2b, 2b}$ denotes the Fisher's F distribution.

Slide 48

How do we choose b_1, b_2 ?

- Uniform prior: $b_1 = b_2 = 1$.
This also known as reference prior.
- Jeffreys prior: $b_1 = b_2 = 0.5$.
- Vague proper prior: $b_1 = b_2 = \epsilon$, where $\epsilon > 0$ is small number (e.g., $\epsilon = 10^{-3}$, etc.)
- Haldane's prior: $b_1 = b_2 = 0$. (improper!)

Note that for all of these priors, $E[\theta] = 0.5$ and $Var[\theta] = 0.25/(2b + 1)$, where $b = 1, 0.5, \epsilon$ and 0 , respectively.

For all our illustrations, we will use $b_1 = b_2 = b$.

Another prior that is rarely used is Zellner's prior: $p(\theta) \propto \theta^\theta (1 - \theta)^{1-\theta}$. Closed form posterior distribution is not available for this prior.

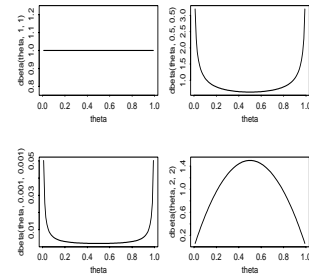


Figure 1: Beta priors with $b = 1, 0.5, 10^{-3}$ and 2

Summarizing posterior distribution

I. Point estimates:

The posterior mean is given by,

$$E[\theta|y] = \frac{y + b}{n + 2b} = b_n \frac{y}{n} + (1 - b_n) \cdot \frac{1}{2}$$

where $b_n = [1 + \frac{2b}{n}]^{-1}$.

Note that the Bayes estimate is a compromise between the MLE $\hat{\theta} = \frac{y}{n}$ and the prior mean $\frac{1}{2}$. What happens when $n \rightarrow \infty$ or $b \rightarrow 0$?

The posterior sd is given by,

$$sd(\theta|y) = \frac{1}{\sqrt{n + 2b + 1}} \sqrt{E(\theta|y)(1 - E(\theta|y))}$$

where $E[\theta|y]$ is given above.

What happens to $sd(\theta|y)$ as $n \rightarrow \infty$ or $b \rightarrow 0$? For what values of b , $sd(\theta|y) < sd(\theta)$?

II. Interval estimates:

Closed form expression are not available. Incomplete Beta functions are required to find interval estimates. A simple alternative is to use a Monte Carlo method.

A Monte Carlo method to obtain interval estimates:

```
for(i in 1:N){
  Generate theta[i] ~ Beta(y+b, n-y+b)
  Compute phi[i] = (1-theta[i])/theta[i] }
```

Arrange $\theta[i]$'s and $\phi[i]$'s in ascending order. Then $(\theta[0.025N], \theta[0.975N])$ provides a 95% equal-tail posterior interval for θ .

The Monte Carlo estimates can be made arbitrarily close to theoretical estimates, by choosing a large N (say $N = 10,000$).

III. Prediction:

The posterior predictive distribution (using (6)) is given by,

$$\tilde{y}|y \sim \text{Bernoulli}(E[\theta|y])$$

In other words, given the data y , the chance of a female birth is given by

$$\Pr(\tilde{y} = 1|y) = \frac{y + b}{n + 2b}$$

This distribution can be used to check the performance of a statistical model by comparing the future observation to the predicted observation \tilde{y} .

IV. Hypothesis Testing:

Test $H_0 : \theta < 0.5$ vs. $H_A : \theta \geq 0.5$.

Under the prior $\theta \sim \text{Beta}(b, b)$, the Bayes Factor is given by,

$$BF = \frac{\int_0^{0.5} \theta^{y+b-1} (1-\theta)^{n-y+b-1} d\theta}{\int_{0.5}^1 \theta^{y+b-1} (1-\theta)^{n-y+b-1} d\theta}$$

which can be computed easily using Incomplete Beta functions.

The natural logarithm of Bayes Factor is called the *weight of evidence* (See Good, 1950, 1983).

Application to a classic problem of Laplace:

Ex. Suppose that a total of 241,945 girls and 251,527 boys were born in Paris from 1745 to 1770.

- Compute the Bayes estimate and Bayes factors under different Beta priors, with $b = 1, 0.5, 10^{-3}$ and 0, respectively.
- What would you conclude about the proportion of female birth, θ ?
- What about the ratio ϕ ?

Although we don't need [winBUGS](#), see the code in Section 4.

(The above problem was originally solved by Thomas Bayes and Laplace 'independently' in the early 18th century)

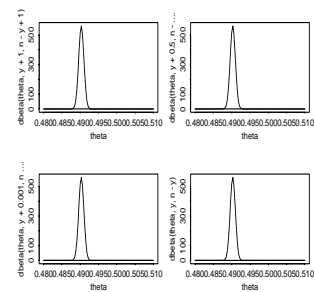


Figure 2: Posterior densities of θ under Beta priors with $b = 1, 0.5, 10^{-3}$ and 0

V. Extensions to a robust class of priors:

In the above example we have used a Beta prior with different hyperparameters. This may be questionable.

One way to robustify the class of prior distribution is to use a mixture of conjugate priors. E.g., we could use

$$\theta \sim p\text{Beta}(b_1, b_2) + (1 - p)\text{Beta}(c_1, c_2)$$

where p is the mixing parameter.

It easily follows that posterior distribution of θ is also a mixture of Beta distributions.

Another extension is to let the hyperparameter p be unknown and use a hyperprior $p \sim U(0, 1)$.

In most applications, with non-conjugate priors, we need *Markov Chain Monte Carlo (MCMC)* methods, implemented using [WinBUGS](#) to obtain posterior inference.

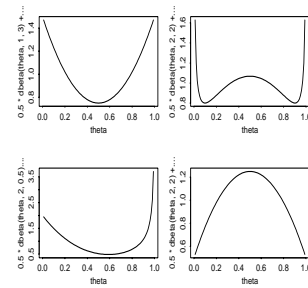


Figure 3: Mixture of Beta distributions as priors.

Choosing a “good” prior ($p(\theta)$):

There is no guaranteed way to do this, just as there is no guaranteed way to arrive at a “good” frequentist model.

(“where does the likelihood come from?”)

Some general strategies..

Informative priors

(a) *Histogram approach:* Assign probability masses to the “possible” values in such a way that their sum is unity, and their relative contributions reflect the experimenter’s prior beliefs as closely as possible.

e.g. for the previous example, you may elicit the prior by asking some expert, “What is the chance that proportion of female birth is less than 0.1, 0.2, ..., 0.9 respectively?” This would provide an empirical cdf for θ .

Can be awkward for continuous and/or unbounded θ .

(b) *Matching a functional form:* Assume that the prior belongs to a parametric family $p(\theta|\eta)$, choosing hyper-parameter η so that the result matches the elicitee’s true prior beliefs as nearly as possible.

e.g. for the female birth study we may consider a *Beta* distribution for θ , i.e. $\theta \sim \text{Beta}(b_1, b_2)$ and choose (b_1, b_2) that roughly matches our belief.

Overcomes finite support problem inherent in the histogram approach. However, it may not be possible to “shoehorn” his or her prior beliefs into any standard parametric forms.

If possible, obtain expert opinions (or records from previous study) to elicit your prior belief.

(c) *Conjugate family*: Defined as one that leads to a posterior distribution $p(\theta|\mathbf{y})$ belonging to the same distributional family as the prior.

Conjugate priors for exponential family is readily obtained.

Suppose $p(y_i|\theta) = f(y_i)g(\theta) \exp\{t(y_i)^T c(\theta)\}$.

It follows that,

$$p(\mathbf{y}|\theta) \propto g(\theta)^n \exp\{s(\mathbf{y})^T c(\theta)\},$$

where $s(\mathbf{y}) = \sum_{i=1}^n t(y_i)$.

Hence the conjugate prior density is given by

$$p(\theta|\eta) \propto g(\theta)^{\eta_1} \exp\{\eta_2^T c(\theta)\}.$$

Notice that,

-in higher dimensions, priors that are *conditionally* conjugate are often available (and helpful)

-a finite mixture of conjugate priors may be sufficiently flexible (allowing multimodality, heavier tails, etc.) while still enabling simplified posterior calculations.

Slide 61

Non-informative priors

(a) Uniform prior: Suppose the parameter space,

$$\Theta = (\theta_1, \dots, \theta_m) \Rightarrow p(\theta_i) = \frac{1}{n}, i = 1, \dots, n$$

Suppose $\Theta = [a, b] \Rightarrow p(\theta) = \frac{1}{b-a}$ (uniform)

(b) *Improper* priors:

$\Theta = (-\infty, \infty) \Rightarrow p(\theta) = c$, any $c > 0$. Still legitimate if $\int p(\mathbf{y}|\theta)d\theta < \infty$.

(c) Jeffreys prior: Invariant to monotone transformations. Suppose $\phi = h(\theta)$ where $h(\cdot)$ is a monotone function. Can we find $p(\theta)$ such the $p(\phi)$ is equivalent to $p(\theta)$?

This is achieved if we choose, $p(\theta) \propto [I(\theta)]^{1/2}$, where $I(\theta)$ is the expected Fisher Information in the model, i.e.

$$I(\theta) = E_{\mathbf{y}|\theta} \left[\left(\frac{\partial}{\partial \theta} \log p(\mathbf{y}|\theta) \right)^2 \right].$$

Notice that this prior can often be improper.

Slide 62

(d) Vague priors: Use a proper density with very large variance but centered at a reasonable value of θ . e.g. $\theta \sim N(0, 10^6)$ or $\theta \sim \text{Gamma}(10^{-3}, 10^{-3})$ etc.

However there are some difficulties associated with noninformative priors. When using improper priors one MUST check that the posterior is proper.

In almost every real problem, the data analyst will have more information than can be conveniently included in the model.

Slide 63

Exercise: Suppose $y_i|\theta \sim N(\theta, 1)$, for $i = 1, \dots, n$.

1. Compute $p(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$.
2. What is the MLE of θ ?
3. What is the distribution of the MLE?
4. Suppose now we assign a prior, $\theta \sim N(0, \tau^2)$. Compute the posterior density $p(\theta|\mathbf{y})$. using Bayes Theorem.
 - (a) What happens to the posterior distribution when $\tau^2 \rightarrow 0$?
 - (b) What happens to the posterior distribution when $\tau^2 \rightarrow \infty$?
5. Convince yourself that the posterior distribution is a compromise between the prior and the likelihood.

Slide 64

2. Multinomial Genotypic Counts

Assume that every individual in a sample has the same probability of being a particular genotype. Then the probability in a sample of size n , there are y_i members in i th category is given by,

$$p(y_1, \dots, y_k | \theta_1, \dots, \theta_m) = n! \prod_{i=1}^m \frac{\theta_i^{y_i}}{y_i!} \quad (7)$$

where θ_i is the probability that an individual has i th genotype, such that $\sum_{i=1}^m \theta_i = 1$ and $\sum_{i=1}^m y_i = n$.

In most applications, the genotypic frequencies θ_i 's are functions of allele frequencies p_j 's, $j = 1, \dots, k$. It is of interest to estimate the allele frequencies p_j 's or some functions (e.g., disequilibrium coefficient, inbreeding coefficient etc.). So we need to specify prior distribution of the allele frequencies (p_1, \dots, p_k) .

Slide 65

An example: Allele frequency estimation

The MN blood groups are recognized by agglutination with anti-M or anti-N serum and can be regarded as being controlled by a single locus with two alleles, M and N . The three genotypes can all be distinguished.

| Genotype | Father | Mother | Total |
|----------|--------|--------|-------|
| MM | 26 | 27 | 53 |
| MN | 44 | 51 | 95 |
| NN | 23 | 15 | 38 |
| Total | 93 | 93 | 186 |

Here, we have $m = 3$ genotypes, with $\theta_1 = P_{MM}$, $\theta_2 = P_{MN}$ and $\theta_3 = P_{NN}$.

Assume that these 93 couples are a random sample of $n = 186$ individuals from a population.

We want to estimate the $k = 2$ allele frequencies, $p_1 = p_M$ and $p_2 = p_N$ ($p_N = 1 - p_M$).

Note that the observed genotypic frequencies are given by $y_1 = 53$, $y_2 = 95$ and $y_3 = 38$.

Slide 66

Case-I: Posterior estimation under Hardy-Weinberg equilibrium

$$\begin{aligned} P_{MM} &= p_M^2 \\ P_{MN} &= 2p_M p_N \\ P_{NN} &= p_N^2 \end{aligned}$$

where $p_M + p_N = 1$.

Substituting, the above quantities in eq. (7), we obtain the likelihood,

$$p(y_1, y_2, y_3 | p_M) \propto p_M^{2y_1 + y_2} (1 - p_M)^{y_2 + 2y_3}$$

Prior distribution for p_M :

For simplicity, we assume a conjugate prior,

$$p_M \sim \text{Beta}(b, b).$$

Posterior distribution for p_M :

Using Bayes Rule (5) we obtain,

$$p_M | y_1, y_2, y_3 \sim \text{Beta}(2y_1 + y_2 + b, 2y_3 + y_2 + b)$$

Slide 67

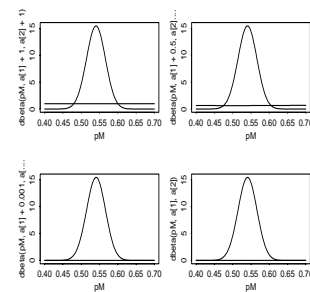


Figure 4: Posterior density of p_M under Beta priors with $b = 1, 0.5, 10^{-3}$ and 0

Slide 68

Case-II: Posterior estimation under Hardy-Weinberg disequilibrium

$$\begin{aligned} P_{MM} &= p_M^2 + p_M p_N f \\ P_{MN} &= 2p_M p_N (1 - f) \\ P_{NN} &= p_N^2 + p_M p_N f \end{aligned}$$

where $p_M + p_N = 1$ and f is called the *inbreeding coefficient*. Note that f must satisfy the constraint,

$$-\min\left\{\frac{p_M}{p_N}, \frac{p_N}{p_M}\right\} \leq f \leq 1. \quad (8)$$

Substituting, the above quantities in equation (7),

$$p(y_1, y_2, y_3 | p_M, f) \propto p_M^{y_1 + y_2} (1 - p_M)^{y_2 + y_3} (1 - f)^{y_2} (p_M + (1 - p_M)f)^{y_1} (1 - p_M + p_M f)^{y_3}$$

where the parameters p_M and f satisfy the constraint (8).

Slide 69

Prior distribution for (p_M, f) :

The nonlinear constraint (8) makes it harder to specify the prior. For simplicity we use a reference prior

$$(p_M, f) \sim \text{Uniform}(A)$$

where $A = \{(p_M, f) : -\min\{\frac{p_M}{1-p_M}, \frac{1-p_M}{p_M}\} \leq f \leq 1, 0 \leq p_M \leq 1.\}$

How do we obtain the joint posterior distribution?

How do we obtain the posterior distribution of $D_M = P_{MM} - p_M^2$?

The joint posterior distribution of (p_m, f) is non-trivial and requires numerical techniques such as MCMC.

We will use **WinBUGS** to obtain samples from this posterior distribution. See the **WinBUGS** in Section 4.

(stay tuned for an intro to WinBUGS)

Slide 70

3. Introduction to WinBUGS

The software **WinBUGS** provides a method to implement the MCMC methods.

The software can be downloaded from

<http://www.mrc-bssu.cam.ac.uk/bugs/winbugs/contents.shtml> Once you downloaded the files you need to email the BUGS project for a key that will let you use the full version.

All the documentation specific to **WinBUGS** is available on-line and is packaged with the program.

The BUGS website also contains some slides of a tutorial introduction by Spiegelhalter and many other useful links.

Slide 71

Starting WinBUGS

1. Click on **WinBUGS** icon or program file. Read the license statement and explore the menus.
2. **HELP** - you will find manual and examples. Try some of the examples.
3. **FILE** - allows you to open existing files, or to start a new file to program your own example in BUGS language. This is what we will do in lab.
4. **Specification Tool** - allows you to 'check model' is the code is syntactically correct, 'load data' for your analysis, specify 'num of chains' for number of parallel chains, 'compile' to check the consistency of code and data and finally 'load inits' to provide initial values of the parameters.

Slide 72

5. Update Tool - allows you to 'update' required number of iterations.
6. Sample Monitor Tool - allows you to monitor selected (or all) parameters of the chain. After the updating is done, you may check out the 'trace', 'density', five-number summary 'stats' and selected 'quantiles.' First you should check for convergence using 'GR diag', 'autoC' and 'history'.

Slide 73

4. Sample codes

I. Female birth estimation

```

model{
y ~ dbin(theta, n)
theta ~ dbeta(b[1], b[2])
phi <- (1-theta)/theta
}

Data:
list(y=241945,n=493472,b=c(0.5,0.5))

Inits:
list(theta=0.5)
list(theta=0.1)
list(theta=0.9)

```

Slide 74

II. Allele frequency estimation

```

model{
y[1:3] ~ dmulti(theta [],n)

#Under HWE:
theta[1] <- p*p
theta[2] <- 2p*(1-p)
theta[3] <- (1-p)*(1-p)
p ~ dbeta(b[1], b[2])

#Under HWD:
#theta[1] <- p*p + p*(1-p)*f
#theta[2] <- 2p*(1-p)*(1-f)
#theta[3] <- (1-p)*(1-p) + p*(1-p)*f
#p ~ dunif(lower.p, 0.5)
#f ~ dunif(lower.f, 1)
#lower.p <- max(0,-f/(1-f))
#lower.f <- -p/(1-p)
#D <- theta[1] - p*p
}

```

Slide 75

```

Data:
Under HWE
list(y=c(53,95,38),n=186,b=c(0.5,0.5))

Under HWD
list(y=c(53,95,38),n=186)

Inits:
Under HWE
list(p=0.5)
list(p=0.1)
list(p=0.9)

Under HWD
list(p=0.25, f=0)
list(p=0.1, f=-0.1)
list(p=0.5, f=0.1)

```

Slide 76