ST 740: Model Selection

Alyson Wilson

Department of Statistics
North Carolina State University

November 25, 2013

Formal Bayesian Model Selection

Bayesians are, well, sometimes Bayesian about model choice. A prior probability distribution on the possible models is chosen (usually uniform), and then Bayes Theorem is used to derive the posterior probability distribution.

Bayesian model choice is aimed at computing a posterior probability distribution over a set of hypotheses or models, in terms of their relative support from the data.

Note: Because this distribution is marginalized over the parameters, improper priors on the parameters cannot, in general, be used.

Bayes Factor

Suppose that we wish to compare models M_1 and M_0 for the same data. The two models may have different likelihoods, different numbers of unknown parameters, etc. (For example, we might want to compare two non-nested regression models.)

The Bayes factor in the general case is the ratio of the marginal likelihoods under the two candidate models. If θ_1 and θ_0 denote parameters under models M_1 and M_0 respectively, then

$$\begin{aligned} p(\mathbf{y} \mid M_1) &= \int \pi(\theta_1) f(\mathbf{y} \mid \theta_1) d\theta_1 \\ p(\mathbf{y} \mid M_0) &= \int \pi(\theta_0) f(\mathbf{y} \mid \theta_0) d\theta_0 \\ BF_{10} &= \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_0)} \end{aligned}$$

Bayes Factor

$$\frac{\pi(m=i\,|\,\mathbf{y})}{\pi(m=j\,|\,\mathbf{y})} = \frac{\pi(m=i)}{\pi(m=j)} \times \text{Bayes Factor}(m=i, m=j)$$

$$\begin{aligned}
\frac{\pi(m=i\,|\,\mathbf{y})}{\pi(m=j\,|\,\mathbf{y})} &= \frac{\frac{f(\mathbf{y}\,|\,m=i)\pi(m=i)}{f(\mathbf{y}\,|\,m=i)\pi(m=i)+f(\mathbf{y}\,|\,m=j)\pi(m=j)}}{\frac{f(\mathbf{y}\,|\,m=j)\pi(m=j)}{f(\mathbf{y}\,|\,m=i)\pi(m=i)+f(\mathbf{y}\,|\,m=j)\pi(m=j)}} \\
&= \frac{\pi(m=i)}{\pi(m=j)}\frac{f(\mathbf{y}\,|\,m=i)}{f(\mathbf{y}\,|\,m=j)} \\
&= \frac{\pi(m=i)}{\pi(m=j)}\frac{\int f(\mathbf{y}\,|\,\theta_i, m=i)\pi(\theta_i\,|\,m=i)d\theta_i}{\int f(\mathbf{y}\,|\,\theta_j, m=j)\pi(\theta_j\,|\,m=j)d\theta_j}
\end{aligned}$$

The *Bayes Factor* is the ratio of the *posterior odds* to the *prior odds*.

Guidelines for Bayes Factors

Bayes Factor($m = i, m = j$)	Interpretation
Under 1	Supports model j
1-3	Weak support for model i
3-20	Support for model i
20-150	Strong support for model i
Over 150	Very strong support for model i

Calculating Bayes Factors

It may be possible to do the integration and get an analytic expression for the Bayes Factor.

Assuming that we can't get an analytic expression, there are a variety of ways to calculate

$$f(\mathbf{y} \mid m = i) = \int f(\mathbf{y} \mid \theta_i, m = i)\pi(\theta_i \mid m = i)d\theta_i$$

General references

- Han and Carlin (2001), *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, JASA
- Kass and Raftery (1995), *Bayes Factors*, JASA
- Clyde and George (2004), *Model Uncertainty*, Statistical Science

Hypothesis Testing

Classical Approach

Most simple problems have the following general form.

There is an unknown parameter $\theta \in \Theta$, and you want to know whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

You make a set of observations $y_1, y_2, \ldots, y_n$ whose sampling density $f(\mathbf{y} \mid \theta)$ depends on θ.

We refer to

$$H_0 : \quad \theta \in \Theta_0 \text{ as the null hypothesis, and}$$
$$H_A : \quad \theta \in \Theta_1 \text{ as the alternative hypothesis.}$$

Hypothesis Testing

Classical Approach

A test is decided by a rejection region R where

$$R = \{\mathbf{y} : \text{observing } \mathbf{y} \text{ would lead to the rejection of } H_0.\}$$

Decisions between tests are based on the probability of Type I errors, $P(R \mid \theta)$ for $\theta \in \Theta_0$, and on the probability of Type II errors, $1 - P(R \mid \theta)$ for $\theta \in \Theta_1$.

Often R is chosen so that the probability of a Type II error is as small as possible subject to the requirement that the probability of a Type I error is always less than or equal to some fixed value α.

Example

The Likelihood Principle

The Likelihood Principle. In making inferences or decisions about θ after **y** is observed, all relevant experimental information is contained in the likelihood function for the observed **y**. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).

The Likelihood Principle

From Berger (1985). Suppose that we make 12 independent tosses of a fair coin. The results of the experiment are 9 heads and 3 tails.
If number of tosses was fixed at 12, then y (the number of heads) is Binomial(12,p), and

$$f(y \mid p, n) = \binom{n}{y} p^y (1-p)^{n-y} \propto p^9 (1-p)^3$$

If we tossed the coin until 3 tails appeared, then y is NegativeBinomial(3, p) and

$$f(y \mid p, r) = \binom{r+y-1}{r-1} p^y (1-p)^r \propto p^9 (1-p)^3$$

A classical test of the hypothesis $H_0 : p = \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$ has p-value 0.075 for binomial sampling and 0.0325 for negative binomial sampling. But data are the same: 9 heads and 3 tails!

Hypothesis Testing

Bayesian Approach

The Bayesian approach calculates the posterior probabilities

$$p_0 = P(\theta \in \Theta_0 \,|\, \mathbf{y})$$
$$p_1 = P(\theta \in \Theta_1 \,|\, \mathbf{y})$$

and decides between H_0 and H_A accordingly.

$p_0 + p_1 = 1$ since $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

Of course the posterior probability of our hypothesis depends on the prior probability of our hypothesis.

Example

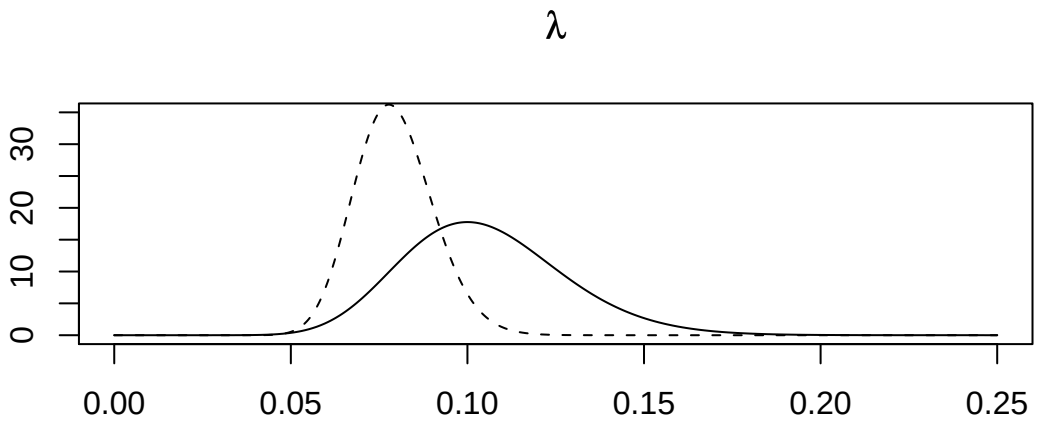
Suppose that we are modeling the expected lifetime of a set of light bulbs. We model the lifetimes using an Exponential(λ) sampling distribution so that

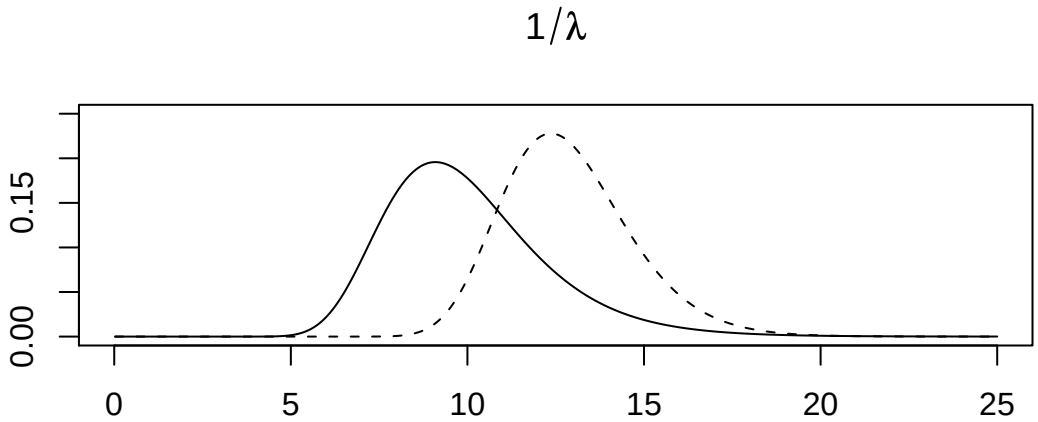
$$f(\mathbf{t} \mid \lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} t_i) \ .$$

If we use Gamma prior for λ, our posterior distribution is Gamma($\alpha + n, \beta + \sum_{i=1}^{n} t_i$), and our posterior distribution for $\frac{1}{\lambda}$, the expected lifetime, is InverseGamma($\alpha + n, \beta + \sum_{i=1}^{n} t_i$).

Example

Suppose that we choose a Gamma(21,200) prior distribution for λ and $H_0 : \frac{1}{\lambda} \leq 10$ and $H_A : \frac{1}{\lambda} > 10$. We observe $\sum_{i=1}^{30} t_i = 442.95$.





Example

Our prior probability that $\frac{1}{\lambda} \leq 10$ is 0.559. If we consider this in terms of odds, our prior odds on H_0 are $\frac{0.559}{1-0.559} = 1.27$.

Our posterior probability that $\frac{1}{\lambda} \leq 10$ is 0.038. If we consider this in terms of odds, our posterior odds on H_0 are $\frac{0.038}{1-0.038} = 0.040$.

We define the Bayes factor as the ratio of the posterior odds to the prior odds. Here

$$\text{BF} = \frac{0.040}{1.27} = 0.033$$

The Bayes factor is a measure of the evidence provided by the data in support of the hypothesis H_0. In this case, the data provides strong evidence against the null hypothesis.

Model Selection

Oxygen Uptake Example

Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program.

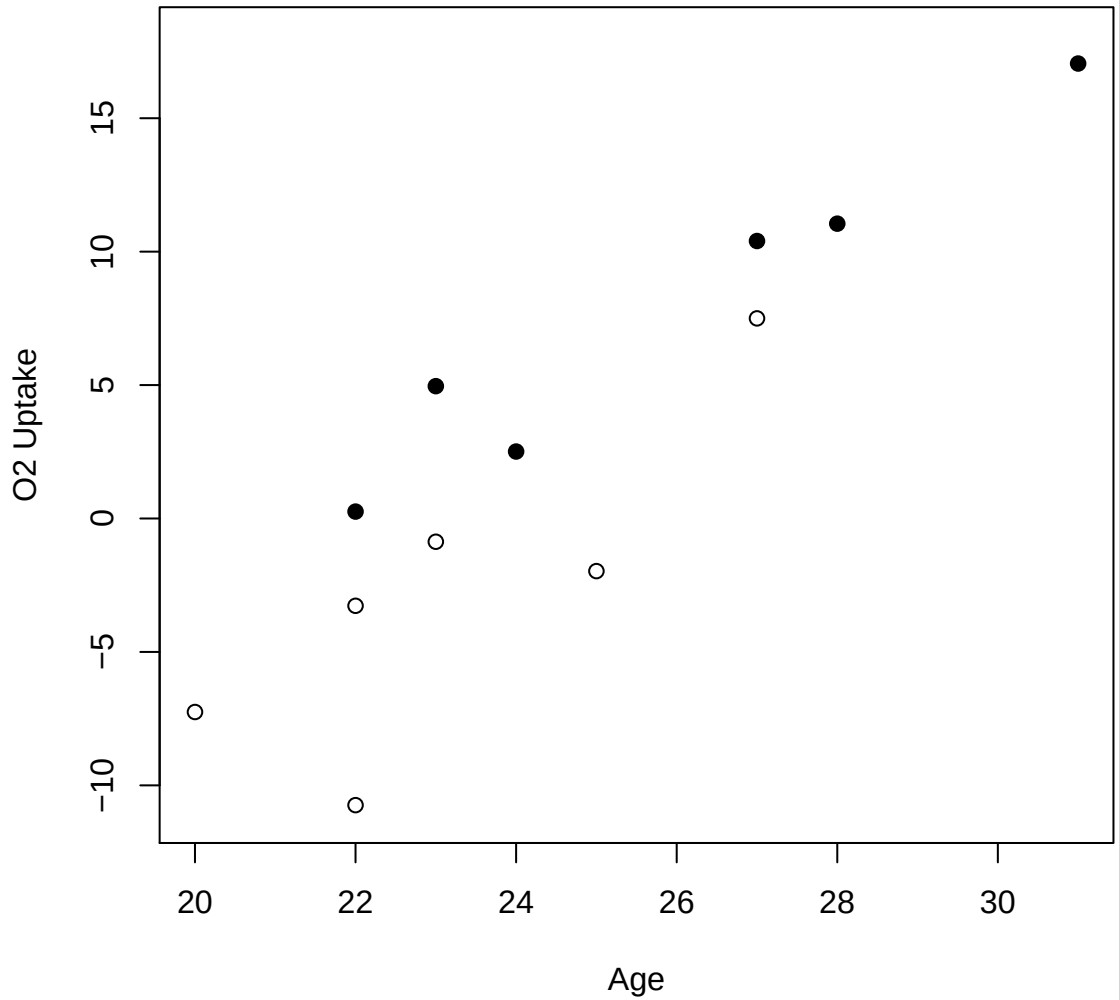
The maximum oxygen uptake of each subject was measured (in liters per minute) while running on an incline treadmill, both before and after the 12-week program. Of interest is how a subject's change in maximal oxygen upake depends on which program they were assigned to. However, we also want to account for the age of the subject, and we think there may be an interaction between age and the exercise program.

$$\mathbf{Y}_i = \beta_0 + \beta_1 x_1 + \beta_2 \text{age} + \beta_3(\text{age} * x_1) + \epsilon_i$$

where $x_1 = 0$ if subject i is on the running program and 1 if on the aerobic program.

Model Selection

Oxygen Uptake Example



Bayesian Model Selection

Using classical methods, we might use forward, backward, or stepwise selection to choose the variables to include in a regression model.

The Bayesian solution to model selection is conceptually straightforward: if we believe that some of the regression coefficients are potentially equal to zero, we specify a prior distribution that captures this. A convenient way to represent this is to write the regression coefficient for variable j as $\beta_j = z_j \times b_j$, where $z_j \in \{0, 1\}$ and b_j is a real number. We then write our regression model as

$$\mathbf{Y}_i = z_0 b_0 + z_1 b_1 x_1 + z_2 b_2 \text{age} + z_3 b_3 (\text{age} * x_1) + \epsilon_i$$

Each value of $\mathbf{z} = (z_0, \ldots, z_p)$ corresponds to a different model; in particular to a different collection of variables with non-zero regression coefficients.

Bayesian Model Selection

Since we haven't observed $\mathbf{z}$, we treat it just like the other parameters in the model. To do model selection, we want to look at the posterior probability for each possible value of $\mathbf{z}$.

$$\pi(\mathbf{z} \mid \mathbf{y}, \mathbf{X}) = \frac{\pi(\mathbf{z}) f(\mathbf{y} \mid \mathbf{X}, \mathbf{z})}{\sum_{\tilde{\mathbf{z}}} \pi(\tilde{\mathbf{z}}) f(\mathbf{y} \mid \mathbf{X}, \tilde{\mathbf{z}})}$$

So if we can calculate $f(\mathbf{y} \mid \mathbf{X}, \mathbf{z})$, then we can calculate the posterior distribution for $\mathbf{z}$. Remember from our discussion of Bayes factors,

$$\frac{\pi(\mathbf{z}_a \mid \mathbf{y}, \mathbf{X})}{\pi(\mathbf{z}_b \mid \mathbf{y}, \mathbf{X})} = \frac{\pi(\mathbf{z}_a)}{\pi(\mathbf{z}_b)} \times \frac{f(\mathbf{y} \mid \mathbf{z}_a, \mathbf{X})}{f(\mathbf{y} \mid \mathbf{z}_b, \mathbf{X})}$$

Bayesian Model Selection

$$
\begin{aligned}
f(\mathbf{y} \mid \mathbf{X}, \mathbf{z}) &= \int \int f(\mathbf{y}, \beta, \sigma^2 \mid \mathbf{X}, \mathbf{z}) d\beta d\sigma^2 \\
&= \int \int f(\mathbf{y} \mid \beta, \sigma^2, \mathbf{X}, \mathbf{z}) \pi(\beta \mid \mathbf{X}, \mathbf{z}, \sigma^2) \pi(\sigma^2 \mid \mathbf{X}, \mathbf{z}) d\beta d\sigma^2
\end{aligned}
$$

The g-prior is convenient because you can do the integration in closed form.

$$
f(\mathbf{y} \mid \mathbf{z}, \mathbf{X}) = \pi^{-n/2} \frac{\Gamma((\nu_0 + n)/2)}{\Gamma(\nu_0/2)} (1+g)^{-p_z/2} \frac{(\nu_0 \sigma_0^2)^{\nu_0/2}}{(\nu_0 \sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}}
$$

Other Prior Distributions

g-prior

The *g-prior* starts from the idea that parameter estimation should be invariant to the scale of the regressors.

Suppose that $\mathbf{X}$ is a given set of regressors and $\tilde{\mathbf{X}} = \mathbf{XH}$ for some $p \times p$ matrix H. If we obtain the posterior distribution of β from $\mathbf{y}$ and $\mathbf{X}$ and $\tilde{\beta}$ from $\mathbf{y}$ and $\tilde{\mathbf{X}}$, then, according to this principle of invariance, the posterior distributions of β and $\mathbf{H}\tilde{\beta}$ should be the same.

$$\begin{aligned} \pi(\beta \,|\, \sigma^2) &\sim \text{MVN}(0, k(\mathbf{X}^T\mathbf{X})^{-1}) \\ \pi(\sigma^2) &\sim \text{InverseGamma}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

A popular choice for k is $g\sigma^2$. Choosing g large reflects less informative prior beliefs.

Other Prior Distributions

g-prior

In this case, we can show that

$$\pi(\beta \,|\, \sigma^2, \mathbf{y}, \mathbf{X}) \sim \text{MVN}\left(\frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\right)$$

$$\pi(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}) \sim \text{InverseGamma}\left(\frac{n+\nu_0}{2}, \frac{\nu_0\sigma_0^2 + SSR_g}{2}\right)$$

where $SSR_g = \mathbf{y}^T(I - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$.

Another popular "noninformative" choice is $\nu_0 = 0$.

Assumptions

- $\beta_j = 0$ for any j where $z_j = 0$
- p_z is the number of non-zero elements in $\mathbf{z}$
- $\mathbf{X}_z$ is the $n \times p_z$ matrix corresponding to the variables j for which $z_j = 1$

$$SSR_g^z = \mathbf{y}^T(I - \frac{g}{g+1}\mathbf{X}_z(\mathbf{X}_z^T\mathbf{X}_z)^{-1}\mathbf{X}_z)\mathbf{y}$$

Bayesian Model Selection

From Hoff Table 9.1, assuming $g = n$ and equal a priori probabilities for each value of $\mathbf{z}$:

$\mathbf{z}$	Model	$\pi(\mathbf{z} \mid \mathbf{y}, \mathbf{X})$
(1,0,0,0)	β_0	0.00
(1,1,0,0)	$\beta_0 + \beta_1 x_1$	0.00
(1,0,1,0)	$\beta_0 + \beta_2 \text{age}$	0.18
(1,1,1,0)	$\beta_0 + \beta_1 x_1 + \beta_2 \text{age}$	0.63
(1,1,1,1)	$\beta_0 + \beta_1 x_1 + \beta_2 \text{age} + \beta_3 \text{age} * x_1$	0.19

Which model has the highest posterior probability?
What evidence is there for an age effect? (What is the sum of the posterior probabilities of the models that include age?)

Bayesian Methods for Point Null Hypotheses

We can't use a continuous prior density to conduct a test of $H_0 : \theta = \theta_0$ because that gives a prior probability of 0 that $\theta = \theta_0$ and hence a posterior probability of 0 to the null hypothesis.

What we do instead is assign a probability p_0 to $\theta = \theta_0$ and a probability density $p_1\pi^*(\theta)$ to values $\theta \neq \theta_0$, where $p_1 = 1 - p_0$.

In mathematics, we assign a prior distribution

$$\pi(\theta) = p_0\delta(\theta - \theta_0) + p_1 I(\theta \neq \theta_0)\pi^*(\theta)$$

Dirac Delta

- $\delta(x) = 0$ for $x \neq 0$.
- $\delta(x)$ is a singular measure giving point mass to 0.
- Cdf is a step function that changes from 0 to 1 at $x = 0$.
- $$\int_{-\infty}^{\infty} f(x)\,\delta(x)dx = f(0)$$

Bayesian Methods for Point Null Hypotheses

We know that

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta}$$

Let

$$f^*(y) = \int f(\mathbf{y} \mid \theta)\pi^*(\theta)d\theta$$

be the predictive distribution of $\mathbf{y}$ under the alternative hypothesis. Then

$$\begin{aligned}
&\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta \\
&\quad = \int p_0\delta(\theta - \theta_0)f(\mathbf{y} \mid \theta)d\theta + \int p_1 I(\theta \neq \theta_0)\pi^*(\theta)f(\mathbf{y} \mid \theta)d\theta \\
&\quad = p_0 f(\mathbf{y} \mid \theta_0) + p_1 f^*(\mathbf{y})
\end{aligned}$$

Bayesian Methods for Point Null Hypotheses

Then

$$\begin{aligned}
\pi(\theta \mid \mathbf{y}) &= \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta} \\
&= \frac{p_0\delta(\theta - \theta_0)f(\mathbf{y} \mid \theta) + p_1 I(\theta \neq \theta_0)\pi^*(\theta)f(\mathbf{y} \mid \theta)}{p_0 f(\mathbf{y} \mid \theta_0) + p_1 f^*(\mathbf{y})} \\
&= \frac{p_0 f(\mathbf{y} \mid \theta_0)}{p_0 f(\mathbf{y} \mid \theta_0) + p_1 f^*(\mathbf{y})}\delta(\theta - \theta_0) \\
&\quad + \frac{p_1 f^*(\mathbf{y})}{p_0 f(\mathbf{y} \mid \theta_0) + p_1 f^*(\mathbf{y})}\pi^*(\theta \mid \mathbf{y})I(\theta \neq \theta_0)
\end{aligned}$$

This means that the posterior probability that $\theta = \theta_0$ is

$$\frac{p_0 f(\mathbf{y} \mid \theta_0)}{p_0 f(\mathbf{y} \mid \theta_0) + p_1 f^*(\mathbf{y})}$$

Bayesian Methods for Point Null Hypotheses

Example

From Albert (2007). Suppose that we are interested in assessing the fairness of a coin, and that we're interested in testing $H_0 : p = 0.5$ against $H_A : p \neq 0.5$. We're indifferent between the two hypotheses (fair coin and not fair coin), so we assign $p_0 = 0.5$ and $p_1 = 0.5$.

If the coin is fair, our entire prior probability distribution is concentrated at 0.5. If the coin is unfair, we need to specify our beliefs about what the unfair coin probabilities will look like. Suppose that we choose a $\pi^*(\theta)$ to be a Beta(10,10). (This has mass from about 0.2 to 0.8.)

Bayesian Methods for Point Null Hypotheses

Example

Suppose that we observe $y = 5$ heads in $n = 20$ trials. To determine our posterior probability that the coin is fair, we need to calculate

$$\begin{aligned} f^*(\mathbf{y}) &= \int f(\mathbf{y} \mid \theta)\pi^*(\theta)d\theta \\ &= \int \frac{p^5(1-p)^{15}p^{10-1}(1-p)^{10-1}}{B(10,10)}dp \\ &= \frac{B(15,25)}{B(10,10)} \end{aligned}$$

This means that the posterior probability that $p = 0.5$ is

$$\frac{(0.5)(0.5)^5(1-0.5)^{15}}{(0.5)(0.5)^5(1-0.5)^{15} + 0.5\frac{B(15,25)}{B(10,10)}} = 0.2802215$$