STAT 535: Chapter 8: Checking Model Quality and Bayesian Hypothesis Testing

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

Checking Model Adequacy

- Checking the adequacy of a Bayesian model involves:
  1. determining how sensitive the posterior is to the specification of the prior and the likelihood
  2. checking that the values we obtain in our sample fit those we would expect to see, given our posterior knowledge
  3. checking robustness to individual data values

Sensitivity Analysis

- Checking the sensitivity to the specification of the data model/likelihood should be done regularly, but rarely is.
- We might examine the effect on the posterior of choosing related data models (e.g., Poisson vs. negative binomial for count data).
- Far more often, we check the sensitivity of the posterior to the **prior** specification.
- We might ask: What happens to the posterior when we:
  1. change the functional form of the prior?
  2. keep the same form, but change the parameter(s) of the prior?
- If the posterior is **robust** to such changes in the prior, we may be more comfortable with the posterior inferences we make.

Sensitivity Analysis

Example 1(a): Consider $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known.

- The conjugate prior for μ is $\mu \sim N(\delta, \tau^2)$.
- A noninformative prior for μ is $p(\mu) = 1$.
- Another choice of prior for μ might be a t-distribution centered at δ.
- How would the posterior change for these 3 prior choices?
- We could examine (1) plots of the posterior in each case, or (2) several posterior quantiles in each case.

Local Sensitivity Analysis

- Unfortunately, it may be too difficult to examine a large class of prior specifications, especially when the target parameter $\boldsymbol{\theta}$ is multidimensional.
- **Local** sensitivity analysis simply focuses on how changes in the hyperparameter value(s) affect the posterior.
- **Example 1(a)**: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, σ^2 known.
- Conjugate prior for μ: $\mu \sim N(\delta, \tau^2)$
- Compare resulting posterior (the plot and/or quantiles) to the posterior from these priors:

$$\mu \sim N(\delta - \tau, \tau^2)$$
$$\mu \sim N(\delta + \tau, \tau^2)$$
$$\mu \sim N(\delta, 0.5\tau^2)$$
$$\mu \sim N(\delta, 2\tau^2)$$

See R example.

Local Sensitivity Analysis

- **Example 1(b)**: $Y_1, \ldots, Y_{200}$ are annual deaths from horse kicks for 10 Prussian cavalry corps for each of 20 years.
- Let $Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, and let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior.
- Compare posteriors from these priors for λ:

$$\lambda \sim \text{Gamma}(2, 4)$$
$$\lambda \sim \text{Gamma}(4, 8)$$
$$\lambda \sim \text{Gamma}(1, 2)$$
$$\lambda \sim \text{Gamma}(0.1 \times 2, \sqrt{0.1} \times 4)$$
$$\lambda \sim \text{Gamma}(3 \times 2, \sqrt{3} \times 4)$$

See R example with Prussian horse kick data.

General recommendation when the posterior is highly sensitive to changes in prior specification:Choose a more "objective" prior (or be prepared to defend your prior knowledge!).

Posterior Predictive Distribution

- Recall that for a fixed value of θ, our data $\boldsymbol{Y}$ follow the distribution $p(\boldsymbol{Y}|\theta)$.
- However, the true value of θ is uncertain, so we should average over the possible values of θ to get a better idea of the distribution of $\boldsymbol{Y}$.
- **Before** taking the sample, the uncertainty in θ is represented by the prior distribution $p(\theta)$. So for some new data value y_{new}, averaging over $p(\theta)$ gives the **prior predictive distribution**:

$$p(y_{new}) = \int_{\Theta} p(y_{new}, \theta)\, \mathrm{d}\theta = \int_{\Theta} p(y_{new}|\theta)p(\theta)\, \mathrm{d}\theta$$

Posterior Predictive Distribution

- **After** taking the sample, we have a **better representation** of the uncertainty in θ via our posterior $p(\theta|\boldsymbol{y})$. So the **posterior predictive distribution** for a new data point y_{new} is:

$$p(y_{new}|\boldsymbol{y}) = \int_{\Theta} p(y_{new}|\theta, \boldsymbol{y}) p(\theta|\boldsymbol{y}) \, \mathrm{d}\theta$$

$$= \int_{\Theta} p(y_{new}|\theta) p(\theta|\boldsymbol{y}) \, \mathrm{d}\theta$$

(since y_{new} is independent of the sample data $\boldsymbol{y}$)

- This reflects how we would predict new data to behave / vary.
- If the data we **did observe** follow this pattern closely, it indicates we have chosen our model and prior well.

Posterior Predictive Distribution

Example 2 again: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$,

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$
$$\lambda|\boldsymbol{y} = \text{Gamma}(\sum y_i + \alpha, n + \beta)$$

Posterior predictive distribution is:

$$p(y_{new}|\boldsymbol{y}) = \int_0^\infty p(y_{new}|\lambda) p(\lambda|\boldsymbol{y})\, \mathrm{d}\lambda$$
$$= \int_0^\infty \left[\frac{\lambda^{y_{new}} e^{-\lambda}}{(y_{new})!}\right] \left[\frac{(n+\beta)^{\sum y_i + \alpha}}{\Gamma(\sum y_i + \alpha)} \lambda^{\sum y_i + \alpha - 1} e^{-(n+\beta)\lambda}\right] \mathrm{d}\lambda$$

Posterior Predictive Distribution

So

$$p(y_{new}|\boldsymbol{y}) = \frac{(n+\beta)^{\sum y_i+\alpha}}{\Gamma(\sum y_i+\alpha)\Gamma(y_{new}+1)} \int_0^\infty \lambda^{y_{new}+\sum y_i+\alpha-1} e^{-(n+\beta+1)\lambda}\, d\lambda$$

$$= \frac{(n+\beta)^{\sum y_i+\alpha}}{\Gamma(\sum y_i+\alpha)\Gamma(y_{new}+1)} \frac{\Gamma(y_{new}+\sum y_i+\alpha)}{(n+\beta+1)^{y_{new}+\sum y_i+\alpha}}$$

$$= \frac{\Gamma(y_{new}+\sum y_i+\alpha)}{\Gamma(\sum y_i+\alpha)\Gamma(y_{new}+1)} \left(\frac{n+\beta}{n+\beta+1}\right)^{\sum y_i+\alpha} \left(\frac{1}{n+\beta+1}\right)^{y_{new}}$$

which is a negative binomial with mean $\frac{\sum y_i+\alpha}{n+\beta}$ and variance $\frac{\sum y_i+\alpha}{(n+\beta)^2}(n+\beta+1)$.

Posterior Predictive Distribution

- $\Rightarrow$ The posterior predictive distribution has the same mean as the posterior distribution, but a **greater** variance (additional "sampling uncertainty" since we are drawing a **new** data value).
- See `R` example (Prussian army data).

More about Posterior Predictive Distribution

- **Example 1(a) again**: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, σ^2 known.
- Posterior for $\mu|\boldsymbol{y}$ is normal with mean

$$\mu_{\text{post}} = \frac{\delta/\tau^2 + n\bar{y}/\sigma^2}{1/\tau^2 + n/\sigma^2}$$

and variance

$$\sigma^2_{\text{post}} = \frac{\tau^2\sigma^2}{\sigma^2 + n\tau^2}.$$

- Note $y_{new}|\mu \sim N(\mu, \sigma^2)$, so the posterior predictive distribution is:

$$p(y_{new}|\boldsymbol{y}) = \int_{-\infty}^{\infty} p(y_{new}|\mu)p(\mu|\boldsymbol{y})\,\mathrm{d}\mu.$$

More about Posterior Predictive Distribution

- Sometimes the form of $p(y_{new}|\boldsymbol{y})$ can be derived directly, but it is often easier to sample from $p(y_{new}|\boldsymbol{y})$ using Monte Carlo methods:
- For $j = 1, \ldots, J$, sample
  1. $\mu^{[j]}$ from $p(\mu|\boldsymbol{y})$ and
  2. $y^{*[j]}$ from $p(y_{new}|\mu^{[j]})$
- Then $y^{*[1]}, \ldots, y^{*[J]}$ are an iid sample from $p(y_{new}|\boldsymbol{y})$.
- See R example with lead data.

Hypothesis Testing

- Recall that classical hypothesis testing emphasizes the **p-value**: The probability (under H_0) that a test statistic would take a value as (or more) favorable to H_a as the observed value of this test statistic.
- For example, given iid data $\boldsymbol{y} = y_1, \ldots, y_n$ from $f(y|\theta)$, where $-\infty < \theta < \infty$, we might test $H_0 : \theta \leq 0$ vs. $H_a : \theta > 0$ using some test statistic $T(\boldsymbol{Y})$ (a function of the data).
- Then if we calculated $T(\boldsymbol{y}) = T^*$ for our observed data $\boldsymbol{y}$, the p-value would be:

$$\begin{aligned} \text{p-value} &= P[T(\boldsymbol{Y}) \geq T^* | \theta = 0] \\ &= \int_{T^*}^{\infty} f_T(t|\theta = 0)\, \mathrm{d}t \end{aligned}$$

where $f_T(t|\theta)$ is the distribution (density) of $T(\boldsymbol{Y})$.

Issues with Classical Hypothesis Testing

- This p-value is an average over T values (and thus sample values) that **have not occurred** and are **unlikely to occur**.
- Since the inference is based on "hypothetical" data rather than **only** the **observed** data, it violates the Likelihood Principle.
- Also, the idea of conducting many repeated tests that motivate "Type I error" and "Type II error" probabilities is not sensible in situations where our study is not repeatable.

The Bayesian Approach

- A simple approach to testing finds the posterior probabilities that θ falls in the null and alternative regions.
- We first consider one-sided tests about θ of the form:

$$H_0 : \theta \leq c \quad \text{vs.} \quad H_a : \theta > c$$

for some constant c, where $-\infty < \theta < \infty$.

- We may specify prior probabilities for θ such that

$$p_0 = P[-\infty < \theta \leq c] = P[\theta \in \Theta_0]$$

and

$$p_1 = 1 - p_0 = P[c < \theta < \infty] = P[\theta \notin \Theta_0]$$

where Θ_0 is the set of θ-values such that H_0 is true.

The Bayesian Approach

- Then the **posterior probability** that H_0 is true is:

$$P[\theta \in \Theta_0 | \boldsymbol{y}] = \int_{-\infty}^{c} p(\theta|\boldsymbol{y})\,\mathrm{d}\theta$$

$$= \frac{\int_{-\infty}^{c} p(\boldsymbol{y}|\theta)p_0\,\mathrm{d}\theta}{\int_{-\infty}^{c} p(\boldsymbol{y}|\theta)p_0\,\mathrm{d}\theta + \int_{c}^{\infty} p(\boldsymbol{y}|\theta)p_1\,\mathrm{d}\theta}$$

by Bayes' Law (note the denominator is the marginal distribution of $\boldsymbol{Y}$).

The Bayesian Approach

- Commonly, we might choose an uninformative prior specification in which $p_0 = p_1 = 1/2$, in which case $P[\theta \in \Theta_0 | \boldsymbol{y}]$ simplifies to

$$\frac{\int_{-\infty}^{c} p(\boldsymbol{y}|\theta) p_0 \, \mathrm{d}\theta}{\int_{-\infty}^{\infty} p(\boldsymbol{y}|\theta) p_0 \, \mathrm{d}\theta} = \frac{\int_{-\infty}^{c} p(\boldsymbol{y}|\theta) \, \mathrm{d}\theta}{\int_{-\infty}^{\infty} p(\boldsymbol{y}|\theta) \, \mathrm{d}\theta}$$

Hypothesis Testing Example

- **Example 1** (Coal mining strike data): Let Y = number of strikes in a sequence of strikes before the cessation of the series.
- Suppose we have data $Y_1, \ldots, Y_{11}$ for 11 such sequences in France.
- The Poisson model would be natural, but for these data, the variance greatly exceeds the mean.
- We choose a geometric(θ) model

$$f(y|\theta) = \theta(1-\theta)^y$$

where θ is the probability of cessation of the strike sequence, and y_i= number of strikes before cessation.

- We will use a prior for θ of $p(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$.

Hypothesis Testing Example

- So the posterior is:

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto L(\theta|\boldsymbol{y})p(\theta) \\ &= \theta^n(1-\theta)^{\sum y_i}\theta^{-1}(1-\theta)^{-1/2} \\ &= \theta^{n-1}(1-\theta)^{\sum y_i - 1/2} \end{aligned}$$

  which is a beta$(n, \sum y_i + 1/2)$ distribution.
- We will test $H_0 : \theta \leq 0.05$ vs. $H_a : \theta > 0.05$.
- Then $P[\theta \leq 0.05|\boldsymbol{y}] = \int_0^{0.05} p(\theta|\boldsymbol{y})\, d\theta$, which is the area to the left of 0.05 in the beta$(n, \sum y_i + 1/2)$ density.
- This can be found directly (or via Monte Carlo methods).
- See R example with coal mining strike data.

Two-Sided Tests

- Two-sided tests about θ have the form:

$$H_0 : \theta = c \text{ vs. } H_a : \theta \neq c$$

  for some constant c.
- We cannot test this using a continuous prior on θ, because that would result in a prior probability $P[\theta \in \Theta_0] = 0$ and thus a posterior probability $P[\theta \in \Theta_0 | \boldsymbol{y}] = 0$ for **any** data set $\boldsymbol{y}$.
- We could place a prior probability mass on the point $\theta = c$, but many Bayesians are uncomfortable with this since the value of this point mass is impossible to judge and is likely to greatly affect the posterior.

Two-Sided Tests

- **One solution**: Pick a small value $\epsilon > 0$ such that if θ is within ϵ of c, it is considered "practically indistinguishable" from c.
- Then let $\Theta_0 = [c - \epsilon, c + \epsilon]$ and find the posterior probability that $\theta \in \Theta_0$.
- **Example 1 again**: Testing $H_0 : \theta = 0.10$ vs. $H_a : \theta \neq 0.10$. Letting $\epsilon = 0.003$, then $\Theta_0 = [0.097, 0.103]$ and

$$P[\theta \in \Theta_0 | \boldsymbol{y}] = \int_{.097}^{.103} p(\theta|\boldsymbol{y})\, \mathrm{d}\theta = .033$$

  from R.
- **Another solution** (mimicking classical approach): Derive a $100(1-\alpha)\%$ (two-sided) HPD credible interval for θ. Reject $H_0 : \theta = c$ "at level α" if and only if c falls outside this credible interval.

Two-Sided Tests

- **Note**: Bayesian **decision theory** attempts to specify the **cost** of a wrong decision to conclude H_0 or H_a through a **loss function**.
- We might evaluate the *Bayes risk* of some decision rule, i.e., its **expected loss** with respect to the posterior distribution of θ.

The Bayes Factor

- The **Bayes Factor** provides a way to formally compare two **competing models**, say M_1 and M_2.
- It is similar to testing a "full model" vs. "reduced model" (with, e.g., a likelihood ratio test) in classical statistics.
- However, with the **Bayes Factor**, one model **does not have to be nested** within the other.
- Given a data set $\boldsymbol{y}$, we compare models

$$M_1 : f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) \text{ and } M_2 : f_2(\boldsymbol{y}|\boldsymbol{\theta}_2)$$

- We may specify prior distributions $p_1(\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{\theta}_2)$ that lead to prior probabilities for each model $p(M_1)$ and $p(M_2)$.

The Bayes Factor

By Bayes' Law, the **posterior odds** in favor of Model 1 versus Model 2 is:

$$\begin{aligned}
\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} &= \frac{\int_{\Theta_1} \frac{p(M_1) f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) p_1(\theta_1)\, \mathrm{d}\boldsymbol{\theta}_1}{p(\boldsymbol{y})}}{\int_{\Theta_2} \frac{p(M_2) f_2(\boldsymbol{y}|\boldsymbol{\theta}_2) p_2(\theta_2)\, \mathrm{d}\boldsymbol{\theta}_2}{p(\boldsymbol{y})}} \\
&= \frac{p(M_1)}{p(M_2)} \cdot \frac{\int_{\Theta_1} f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) p_1(\theta_1)\, \mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_2} f_2(\boldsymbol{y}|\boldsymbol{\theta}_2) p_2(\theta_2)\, \mathrm{d}\boldsymbol{\theta}_2} \\
&= [\text{prior odds}] \times [\text{Bayes Factor } B(\boldsymbol{y})]
\end{aligned}$$

The Bayes Factor

Rearranging, the Bayes Factor is:

$$B(\boldsymbol{y}) = \frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} \times \frac{p(M_2)}{p(M_1)}$$
$$= \frac{p(M_1|\boldsymbol{y})/p(M_2|\boldsymbol{y})}{p(M_1)/p(M_2)}$$

(the ratio of the posterior odds for M_1 to the prior odds for M_1).

The Bayes Factor

- **Note**: If the prior model probabilities are equal, i.e., $p(M_1) = p(M_2)$, then the Bayes Factor equals the posterior odds for M_1.
- **Note**: If the parameter spaces Θ_1 and Θ_2 are the same, then the Bayes Factor reduces to a **likelihood ratio**.
  Note that:

$$B(\boldsymbol{y}) = \frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} \times \frac{p(M_2)}{p(M_1)} = \frac{\frac{p(M_1,\boldsymbol{y})}{p(\boldsymbol{y})p(M_1)}}{\frac{p(M_2,\boldsymbol{y})}{p(\boldsymbol{y})p(M_2)}}$$

$$= \frac{\frac{p(M_1,\boldsymbol{y})}{p(M_1)}}{\frac{p(M_2,\boldsymbol{y})}{p(M_2)}} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}$$

The Bayes Factor

- Clearly a Bayes Factor much greater than 1 supports Model 1 over Model 2.
- Jeffreys proposed the following rules, if Model 1 represents a null model:

Result	**Conclusion**
$B(\boldsymbol{y}) \geq 1 \rightarrow$	Model 1 supported
$0.316 \leq B(\boldsymbol{y}) < 1 \rightarrow$	Minimal evidence against Model 1
(Note $0.316 = 10^{-1/2}$)	
$0.1 \leq B(\boldsymbol{y}) < 0.316 \rightarrow$	Substantial evidence against Model 1
$0.01 \leq B(\boldsymbol{y}) < 0.1 \rightarrow$	Strong evidence against Model 1
$B(\boldsymbol{y}) < 0.01 \rightarrow$	Decisive evidence against Model 1

- Clearly these labels are fairly arbitrary.

The Bayes Factor

- In the case when there are only **two possible models**, M_1 and M_2, then given the Bayes Factor $B(\boldsymbol{y})$, we can calculate the posterior probability of Model 1 as:

$$\begin{aligned} P(M_1|\boldsymbol{y}) &= 1 - P(M_2|\boldsymbol{y}) = 1 - \frac{P(\boldsymbol{y}|M_2)P(M_2)}{P(\boldsymbol{y})} \\ &= 1 - \frac{P(\boldsymbol{y}|M_1)}{B(\boldsymbol{y})}\frac{P(M_2)}{P(\boldsymbol{y})} \\ \Rightarrow P(M_1|\boldsymbol{y}) &= 1 - \left\{\frac{1}{B(\boldsymbol{y})}\frac{P(M_2)}{P(M_1)}\right\}P(M_1|\boldsymbol{y}) \\ \Rightarrow 1 &= \left[1 + \left\{\frac{1}{B(\boldsymbol{y})}\frac{P(M_2)}{P(M_1)}\right\}\right]P(M_1|\boldsymbol{y}) \\ \Rightarrow P(M_1|\boldsymbol{y}) &= \frac{1}{1 + \left\{\frac{1}{B(\boldsymbol{y})}\frac{P(M_2)}{P(M_1)}\right\}} \end{aligned}$$

Example: Comparing Two Means

Example 2(a): Comparing Two Means (Bayes Factor Approach)

- **Data**: Blood pressure reduction was measured for 11 patients who took calcium supplements and for 10 patients who took a placebo.
- We model the data with normal distributions having common variance:

$$\text{Calcium data} : Y_{1j} \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2), \quad j = 1, \ldots, 11$$

$$\text{Placebo data} : Y_{2j} \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2), \quad j = 1, \ldots, 10$$

Consider the two-sided test for whether the mean BP reduction differs for the two groups:

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_a : \mu_1 \neq \mu_2$$

Example: Comparing Two Means

- We will place a prior on the difference of standardized means

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

with specified prior mean μ_Δ and variance σ^2_Δ.

- Consider the classical two-sample t-statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}/\sqrt{n^*}},$$

where $n^* = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}$.

Example: Comparing Two Means

- H_0 and H_a define two specific models for the distribution of T.
- Under H_0, $T \sim$ (central) t with $(n_1 + n_2 - 2)$ degrees of freedom.
- Under H_a, $T \sim$ noncentral t.
- With this prior, the Bayes Factor for H_0 over H_a is:

$$B(\boldsymbol{y}) = \frac{t_{n_1+n_2-2}(t^*, 0, 1)}{t_{n_1+n_2-2}(t^*, \mu_\Delta \sqrt{n^*}, 1 + n^* \sigma_\Delta^2)}$$

  where t^* is the observed t-statistic.
- See `R` example to get $B(\boldsymbol{y})$ and $P[H_0|\boldsymbol{y}]$.

Example: Comparing Two Means

Example 2(a): Comparing Two Means (Gibbs Sampling Approach)

- Same data set, but suppose our interest is in testing whether the calcium yields a **better** BP reduction than the placebo:

$$H_0 : \mu_1 \leq \mu_2 \text{ vs. } H_a : \mu_1 > \mu_2$$

- We set up the sampling model:

$$Y_{1j} = \mu + \tau + \epsilon_{1j}, j = 1, \ldots, 11$$
$$Y_{2j} = \mu - \tau + \epsilon_{2j}, j = 1, \ldots, 10$$

  where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

- Thus $\mu_1 = \mu + \tau$ and $\mu_2 = \mu - \tau$.

Example: Comparing Two Means

We can assume independent priors for μ, τ, and σ^2:

$$\mu \sim N(\mu_\mu, \sigma^2_\mu)$$
$$\tau \sim N(\mu_\tau, \sigma^2_\tau)$$
$$\sigma^2 \sim IG(\nu_1/2, \nu_1\nu_2/2)$$

Then it can be shown that the full conditional distributions are:

$$\mu|\boldsymbol{y}_1, \boldsymbol{y}_2, \tau, \sigma^2 \sim \text{Normal}$$
$$\tau|\boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \sigma^2 \sim \text{Normal}$$
$$\sigma^2|\boldsymbol{y}_1, \boldsymbol{y}_2, \mu, \tau \sim IG$$

where the appropriate parameters are given in the R code.

Example: Comparing Two Means

- R **example**: Gibbs Sampler can obtain approximate posterior distributions for μ and (especially of interest) for τ.
- Note $P[\mu_1 > \mu_2 | \boldsymbol{y}] = P[\tau > 0 | \boldsymbol{y}]$.
- We can also find the **posterior predictive** probability $P[Y_1 > Y_2]$.

Issues with Bayes Factors

- **Note**: When an **improper prior** (one that does not integrate to a finite number over its support) is used for θ, the Bayes Factor is not well-defined.
- Note $B(\boldsymbol{y}) = \dfrac{\text{Posterior odds for } M_1}{\text{Prior odds for } M_1}$, and the "prior odds" is meaningless for an improper prior.
- There are several methods (Local Bayes factors, Intrinsic Bayes Factors, Partial Bayes Factors, Fractional Bayes Factors), none of them ideal, to define types of Bayes Factors with improper priors.
- One criticism of Bayes Factors is the (implicit) assumption that one of the competing models (M_1 or M_2) is correct.
- Another criticism is that the Bayes Factor depends heavily on the choice of prior.

The Bayesian Information Criterion

- The Bayesian Information Criterion (BIC) can be used (as a substitute for the Bayes factor) to compare two (or more) models.
- Conveniently, the BIC does **not** require specifying priors.
- For parameters $\boldsymbol{\theta}$ and data $\boldsymbol{y}$:

$$BIC = -2 \ln L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + p \ln(n)$$

  where p is the number of free parameters in the model, and $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ is the **maximized likelihood**, given observed data $\boldsymbol{y}$.
- Good models have relatively small BIC values:
  - A small value of $-2 \ln L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ indicates good fit to the data;
  - a small value of the "overfitting penalty" term $p \ln(n)$ indicates a simple, parsimonious model.

The Bayesian Information Criterion

- To compare two models M_1 and M_2, we could calculate

$$\begin{aligned} S &= -\frac{1}{2}[BIC_{M_1} - BIC_{M_2}] \\ &= \ln L(\hat{\boldsymbol{\theta}}_1|\boldsymbol{y}) - \ln L(\hat{\boldsymbol{\theta}}_2|\boldsymbol{y}) - \frac{1}{2}(p_1 - p_2)\ln(n) \end{aligned}$$

- A small value of S would favor M_2 here and a large S would favor M_1.
- As $n \to \infty$,

$$\frac{S - \ln(B(\boldsymbol{y}))}{\ln(B(\boldsymbol{y}))} \to 0$$

and for large n,

$$BIC_{M_1} - BIC_{M_2} = -2S \approx -2\ln(B(\boldsymbol{y})).$$

The Bayesian Information Criterion

- Note that differences in *BIC*'s can be used to compare several nonnested models.
- They should be trusted as a substitute for Bayes Factors only when (1) no reliable prior information is available and (2) when the sample size is **quite large**.
- See `R` examples: (1) Calcium data example and (2) Regression example on Oxygen Uptake data set.