Bayesian Methods for Data Science (DATS 6450 - 11)

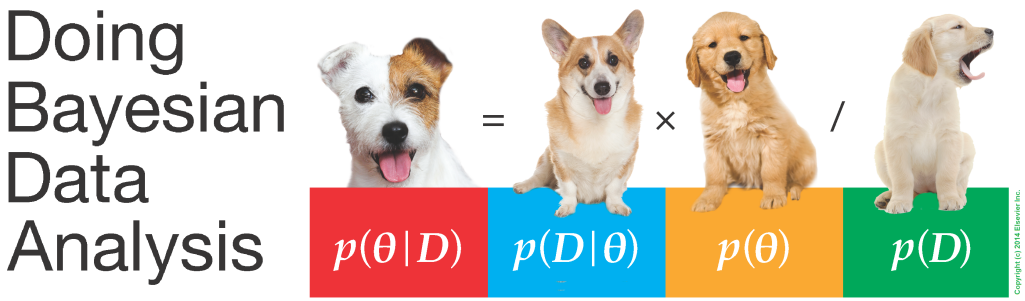
Model Comparison and Hierarchical Modeling

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University
yuxiaohuang@gwu.edu

November 5, 2019

Reference



Picture courtesy of the book website

- This set of slices is an excerpt of the book by Professor John K. Kruschke, with some trivial changes by the creator of the slides
- Please find the reference to and website of the book below:
  - Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier
  - `https://sites.google.com/site/doingbayesiandataanalysis/`

Overview

1. General formula and the bayes factor

2. Example: two factories of coins

Introduction

- Previously, bayesian inference reallocates credibility over parameters within a model
  - assuming that the bias of a coin follows a beta distribution
  - parameter estimation updates the posteriors of the bias given the data
- There are situations where different models compete to describe the same set of data
  - the data consist of the apparent positions of the planets and sun against the background stars
  - are these data best described by an earth-centric model or by a solar-centric model?
- Now, bayesian inference reallocates credibility over different models

Introduction

- Bayesian model comparison is really just a case of Bayesian parameter estimation applied to a hierarchical model in which the top-level parameter is an index for the models
- A central technical point is that a "model" consists of both its likelihood function and prior distribution, and model comparison can be extremely sensitive to the choice of prior, even if the prior is vague, unlike continuous parameter estimation within models

General formula and the bayes factor

- Let $\theta_1, \ldots, \theta_n$ be the set of parameters and m the model
- Then, Bayes' rule becomes

$$p(\theta_1, \ldots, \theta_n, m|D) = \frac{p(D|\theta_1, \ldots, \theta_n, m)p(\theta_1, \ldots, \theta_n, m)}{\sum_m \int \mathrm{d}_{\theta_m} p(D|\theta_1, \ldots, \theta_n, m)p(\theta_1, \ldots, \theta_n, m)}$$
$$= \frac{\prod_m p_m(D|\theta_m, m)p(\theta_m|m)p(m)}{\sum_m \int \mathrm{d}_{\theta_m} \prod_m p_m(D|\theta_m, m)p(\theta_m|m)p(m)}$$

- The factoring of the likelihood-times-prior into a chain of dependencies is the hallmark of a hierarchical model
- A hierarchical diagram illustrating these relations is shown in Figure 10.1 (see next page)

Figure 10.1

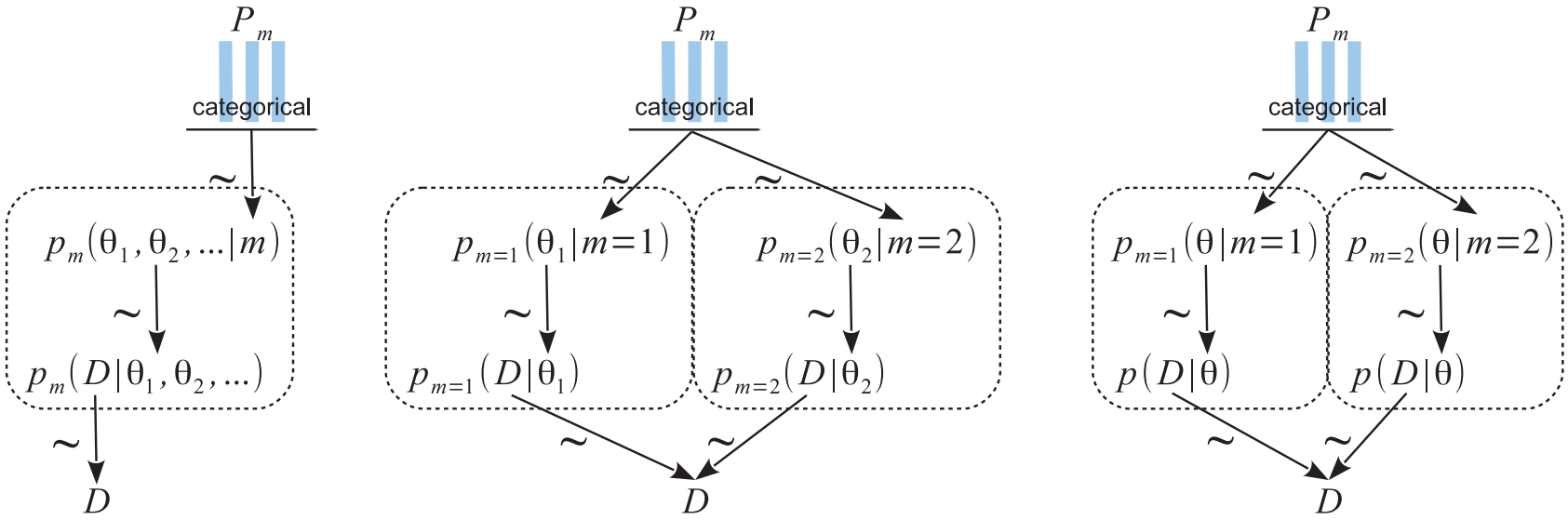


Figure 10.1: Model comparison as a single hierarchical model. Dashed boxes enclose the sub-models being compared. Left panel shows general conception, with parameters θ_m for all sub-models in a joint space. Middle panel shows the usual case in which the likelihood and prior reduce to functions of only θ_m for each m. Right panel shows the special case in which the likelihood function is the same for all m, and only the form of the prior is different for different m. (Middle and right panels depict only two sub-models, but there can be many.) Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

General formula and the bayes factor

- When we want to know the relative credibilities of models overall, we can also apply Bayes' rule to the model index alone:

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_m p(D|m)p(m)},$$

where the probability of the data D, given the model m, $p(D|m)$, is

$$p(D|m) = \int \mathrm{d}\theta_m p_m(D|\theta_m, m) p_m(\theta_m|m)$$

General formula and the bayes factor

- The relative posterior probability of two models, m_1 and m_2, is

$$\frac{p(m=1|D)}{p(m=2|D)} = \underbrace{\frac{p(D|m=1)}{p(D|m=2)}}_{\text{BF}} \frac{p(m=1)}{p(m=2)} \underbrace{\frac{/\sum_m p(D|m)p(m)}{/\sum_m p(D|m)p(m)}}_{=1}$$

- Here "BF" is the bayes factor for models 1 and 2, which is the ratio of the probabilities of the data in the two models
- One convention for converting the magnitude of the BF to a discrete decision about the models is that:
  - there is "substantial" evidence for model m $= 1$ when BF > 3.0
  - there is "substantial" evidence for model m $= 2$ when BF $< 1/3$

Prior

- Suppose a coin was built by one of the two factories, m_1 and m_2
- Factory m_1 generates coins with biases distributed as a beta with mode $\omega_1 = 0.25$ and concentration $\kappa = 12$

$$\theta \sim \text{beta}(\theta|\omega_1(\kappa-2)+1,(1-\omega_1)(\kappa-2)+1) = \text{beta}(3.5, 8.5)$$

- Factory m_2 generates coins with biases distributed as a beta with mode $\omega_2 = 0.75$ and concentration $\kappa = 12$

$$\theta \sim \text{beta}(\theta|\omega_2(\kappa-2)+1,(1-\omega_2)(\kappa-2)+1) = \text{beta}(8.5, 3.5)$$

- The hierarchical models are shown in Figure 10.2 (see next page)

Figure 10.2

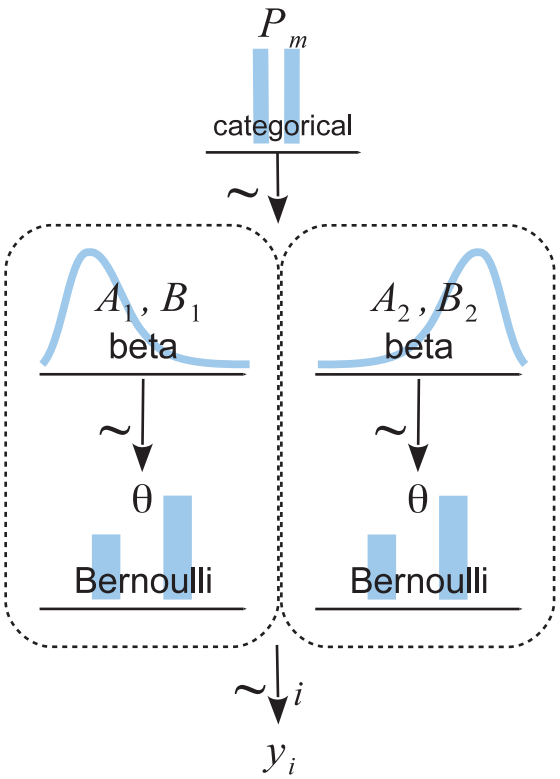


Figure 10.2: Hierarchical diagram for two models of a coin. Model 1 is a tail-biased mint; model 2 is a head-biased mint. This diagram is a specific case of the right panel of Figure 10.1, because the likelihood function is the same for both models and only the priors are different. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Data and potential solutions

- Suppose we flip the coin 9 times and get 6 heads:

$$z = 6 \quad \text{and} \quad N = 9$$

- **Q:** Given those data, what are the posterior probabilities of the coin coming from the head-biased or tail-biased factories?
- **Q:** What can we do to find this?

Data and potential solutions

- Suppose we flip the coin 9 times and get 6 heads:

$$z = 6 \quad \text{and} \quad N = 9$$

- **Q:** Given those data, what are the posterior probabilities of the coin coming from the head-biased or tail-biased factories?
- **Q:** What can we do to find this?
- **A:**
  - formal analysis
  - grid approximation
  - MCMC

Solution by formal analysis

- For beta$(\theta|a,b)$ prior and bern$(z,N|\theta)$ likelihood, what is the posterior?

Solution by formal analysis

- For beta$(\theta|a,b)$ prior and bern$(z,N|\theta)$ likelihood, what is the posterior?

$$\text{beta}(\theta|z+a, N-z+b)$$

- Then based on Bayes' rule:

$$\frac{\theta^{z+a-1}(1-\theta)^{N-z+b-1}}{B(z+a,N-z+b)} = \frac{\theta^{z}(1-\theta)^{N-z}}{p(z,N)}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

- The numerators on both sides can be canceled out:

$$\frac{\cancel{\theta^{z+a-1}(1-\theta)^{N-z+b-1}}}{B(z+a,N-z+b)} = \frac{\cancel{\theta^{z}(1-\theta)^{N-z}}}{p(z,N)}\frac{\cancel{\theta^{a-1}(1-\theta)^{b-1}}}{B(a,b)}$$

- Then

$$p(z,N) = \frac{B(z+a,N-z+b)}{B(a,b)}$$

Solution by formal analysis

- Suppose:

$$p(m=1) = p(m=2) = 0.5,$$

then we can rewrite the equation for model comparison as

$$\begin{aligned}
\frac{p(m=1|D)}{p(m=2|D)} &= \frac{p(D|m=1)}{p(D|m=2)} \frac{p(m=1)}{p(m=2)} \\
&= \frac{B(z+a_1, N-z+b_1)/B(a_1,b_1)}{B(z+a_2, N-z+b_2)/B(a_2,b_2)} \frac{0.5}{0.5} \\
&= \frac{B(9.5, 12.5)/B(3.5, 8.5)}{B(14.5, 7.5)/B(8.5, 3.5)} \\
&= 0.213
\end{aligned}$$

- As a result:

$$p(m=1|D) = 17.6\% \quad \text{and} \quad p(m=2|D) = 82.4\%$$

Solution by grid approximation

- The overall model has two parameters, the bias θ and central tendency ω
- We can use grid approximation to estimate the joint posterior, $p(\theta, \omega|D)$, and in turn the marginal posteriors, $p(\theta|D)$ and $p(\omega|D)$
- The results are shown in Figure 10.3 (see next page)

Figure 10.3

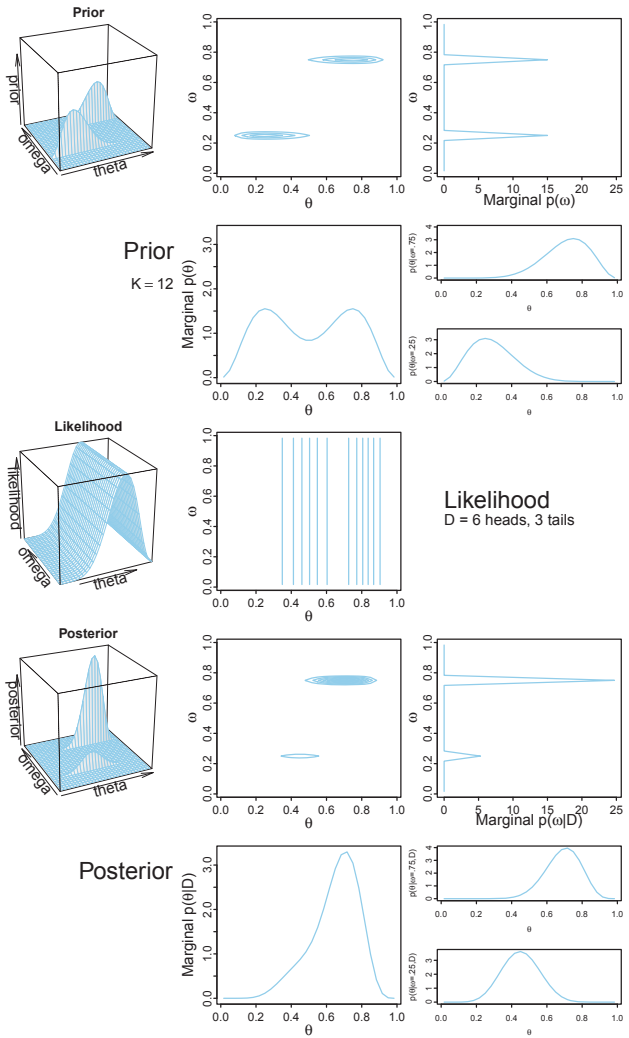


Figure 10.3: A representation of the joint ω, θ parameter space when the mode parameter, ω, is allowed only two discrete values. (For an example with a continuous distribution on ω, compare with Figure 9.2, p. 224.) Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Solution by MCMC

- For large and complex models (where formal analysis and grid approximation will not work), we approximate the posterior probabilities using MCMC
- The idea is that, we put the models together into a hierarchy, and the MCMC process will visit different values of the model index proportionally to their posterior probabilities

Hierarchical MCMC computation of relative model probability

- See `Jags-Ydich-Xnom1subj-MbernBetaModelComp.R` for details
- Figure 10.4 (see next page) shows the prior and posterior distributions of m, θ_1, and θ_2

Figure 10.4

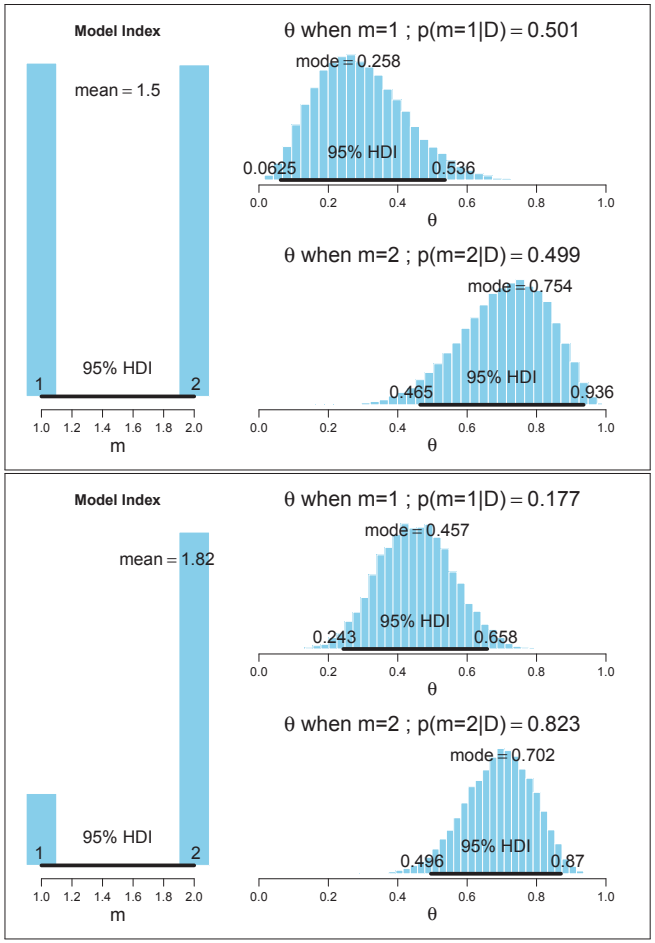


Figure 10.4: The prior and posterior distributions for script `Jags-Ydich-Xnom1subj-MbernBetaModelComp.R`. The upper frame, which shows the prior distribution, has labels that indicate $p(\theta|D)$ but the data set, D, is empty. The lower frame shows the posterior distribution. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Prediction: model averaging

- In many applications of model comparison, the analyst wants to identify the best model and then base predictions of future data on that single best model, denoted with index b
- In this case, predictions of future $\hat{y}$ are based exclusively on the likelihood function $p_b(\hat{y}|\theta_b, m = b)$ and the posterior distribution $p_b(\theta_b|D, m = b)$ of the winning model:

$$p(\hat{y}|D, m = b) = \int \mathrm{d}\theta_b p_b(\hat{y}|\theta_b, m = b) p_b(\theta_b|D, m = b)$$

- But the full model of the data is actually the complete hierarchical structure that spans all the models being compared

Prediction: model averaging

- Therefore, if the hierarchical structure really expresses our prior beliefs, then the most complete prediction of future data takes into account all the models, weighted by their posterior credibilities
- In other words, we take a weighted average across the models, with the weights being the posterior probabilities of the models
- Instead of conditionalizing on the winning model, we have

$$p(\hat{y}|D) = \sum_m p(\hat{y}|D,m)p(m|D)$$
$$= \sum_m \int \mathrm{d}\theta_m p_m(\hat{y}|\theta_m,m)p_m(\theta_m|D,m)p(m|D),$$

  which is called model averaging
- Figure 10.3 (see next page) illustrates the difference between

$$p_b(\theta_b|D,m=b) \quad \text{and} \quad \sum_m p_m(\theta_m|D,m)p(m|D)$$

Figure 10.3

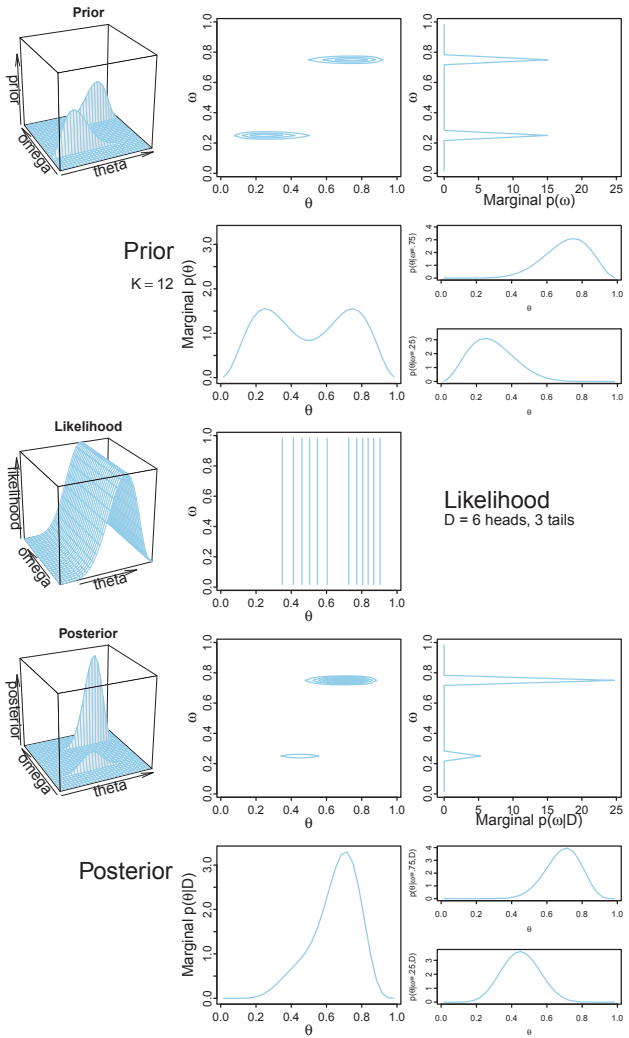


Figure 10.3: A representation of the joint ω, θ parameter space when the mode parameter, ω, is allowed only two discrete values. (For an example with a continuous distribution on ω, compare with Figure 9.2, p. 224.) Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Model complexity naturally accounted for

- One of the nice qualities of Bayesian model comparison is that it naturally compensates for model complexity
- A complex model (usually) has an inherent advantage over a simpler model because the complex model can find some combination of its parameter values that match the data better than the simpler model
- Without some way of accounting for model complexity, the presence of noise in data will tend to favor the complex model

Model complexity naturally accounted for

- Bayesian model comparison compensates for model complexity by the fact that each model must have a prior distribution over its parameters, and more complex models must dilute their prior distributions over larger parameter spaces than simpler models
- Thus, even if a complex model has some particular combination of parameter values that fit the data well, the prior probability of that particular combination must be small because the prior is spread thinly over the broad parameter space

Example

- Consider again the case of two factories that mint coins
- Simple model:

$$\omega_s = 0.5 \quad \text{and} \quad \kappa_s = 1000$$

- Complex model:

$$\omega_c = 0.5 \quad \text{and} \quad \kappa_c = 2$$

Example

- Suppose that we flip a coin $N = 20$ times and observe $z = 15$ heads
- In this case, complex model wins. **Q:** Why?

Example

- Suppose that we flip a coin $N = 20$ times and observe $z = 15$ heads
- In this case, complex model wins. **Q:** Why?
- **A:**
  - the simple model loses because it has no θ value sufficiently close to the data proportion
  - the complex model has available θ values that exactly match the data proportion
- Suppose that we flip a coin $N = 20$ times and observe $z = 11$ heads
- In this case, simple model wins. **Q:** Why?

Example

- Suppose that we flip a coin $N = 20$ times and observe $z = 15$ heads
- In this case, complex model wins. **Q:** Why?
- **A:**
  - the simple model loses because it has no θ value sufficiently close to the data proportion
  - the complex model has available θ values that exactly match the data proportion
- Suppose that we flip a coin $N = 20$ times and observe $z = 11$ heads
- In this case, simple model wins. **Q:** Why?
- **A:**
  - the simple model has large prior probability on θ values sufficiently near the data proportion to be credible
  - the complex model loses because it pays the price of having a small prior probability on the values of θ near the data proportion

Caveats regarding nested model comparison

- A frequently encountered special case of comparing models of different complexity occurs when a *restricted model* (with various restrictions of those parameters) "nested" within the *full model* (that implements all the meaningful parameters)
- Many restricted models have essentially zero prior probability, in turn have zero posterior probability, even if the Bayes factor favors the model
- Such restricted models should not be accepted even if they "won" a model comparison

Extreme sensitivity to prior distribution

- When doing Bayesian estimation of continuous parameters within models and using realistically large amounts of data, the posterior distribution on the continuous parameters is typically robust against changes in vague priors
- However, when doing Bayesian model comparison, the form of the prior is crucial because the Bayes factor integrates the likelihood function weighted by the prior distribution
- Therefore, the posterior probabilities of the models, and the Bayes factors, can be extremely sensitive to the choice of prior distribution
  - if the prior distribution happens to place a lot of probability mass where the likelihood distribution peaks, then the marginal likelihood (i.e., $p(D|m)$) will be large
  - but if the prior distribution happens to place little probability mass where the likelihood distribution is, then the marginal likelihood will be small

Example

- For data $z = 65$ and $N = 100$ and model 1 beta(500, 500)
  - for model 2 beta(1, 1), i.e, uniform prior:

    $$pD(z, N, a = 500, b = 500)/pD(z, N, a = 1, b = 1) = 0.1,$$

    that is, model 2 is favored
  - for model 2 beta(0.01, 0.01), i.e., Haldane prior:

    $$pD(z, N, a = 500, b = 500)/pD(z, N, a = 0.01, b = 0.01) = 5.7,$$

    that is, model 1 is favored
- What can be done to ameliorate the problem?

Priors of different models should be equally informed

- One useful approach is to inform the priors of all models with a small set of representative data (the same for all models)
- The data could come from previous research, or the data could be a small percentage of the data (e.g., 10%) from the research at hand
- Suppose that the 10% subset has 6 heads in 10 flips
  - for model 2 beta(1, 1), i.e, uniform prior:

$$\frac{pD(z, N, a = 500 + 6, b = 500 + 4)}{pD(z, N, a = 1 + 6, b = 1 + 4)} = 0.05$$

  - for model 2 beta(0.01, 0.01), i.e., Haldane prior:

$$\frac{pD(z, N, a = 500 + 6, b = 500 + 4)}{pD(z, N, a = 0.01 + 6, b = 0.01 + 4)} = 0.05$$

- Thus, with the two models equally informed by a small amount of representative data, the Bayes factor is stable