STAT 237
Model Comparison

April 29-?, 2022

Colin Reimer Dawson

Outline

Model Selection and Bayesian Occam's Razor

Outline

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration
- With our batting average example, we may wonder:

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration
- With our batting average example, we may wonder:
  - Should κ_θ be allowed to take different values for each position?

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration
- With our batting average example, we may wonder:
  - Should κ_θ be allowed to take different values for each position?
  - Does it make sense to model each position separately, or would we obtain more robust predictions if we combined non-pitchers?

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration
- With our batting average example, we may wonder:
  - Should κ_θ be allowed to take different values for each position?
  - Does it make sense to model each position separately, or would we obtain more robust predictions if we combined non-pitchers?
- Each different way we answer these questions corresponds to a different **model structure**.

Model Selection

- In many cases we have not just more than one potential set of values of our **parameters**, but we may have more than one **model structure** under consideration
- With our batting average example, we may wonder:
  - Should κ_θ be allowed to take different values for each position?
  - Does it make sense to model each position separately, or would we obtain more robust predictions if we combined non-pitchers?
- Each different way we answer these questions corresponds to a different **model structure**.
- We may want to indicate which model structure we're using with a variable m, which takes values between 1 and M

A different type of hierarchical model

- Each model m will usually depend on some parameter vector, θ, but this may not be the same size across model structures

A different type of hierarchical model

- Each model m will usually depend on some parameter vector, θ, but this may not be the same size across model structures
- Conceptually we can imagine a **model of models**:

$$p(m, \theta, \mathbf{y}) = p(m)p(\theta \mid m)p(\mathbf{y} \mid \theta, m)$$

A different type of hierarchical model

- Each model m will usually depend on some parameter vector, θ, but this may not be the same size across model structures
- Conceptually we can imagine a **model of models**:

$$p(m, \theta, \mathbf{y}) = p(m)p(\theta \mid m)p(\mathbf{y} \mid \theta, m)$$

- To examine the **posterior plausibility of each model structure** (**averaging** over possible θs), we would be interested in

$$p(m \mid \mathbf{y}) = C_{\mathbf{y}} p(\mathbf{y} \mid m)p(m)$$

Marginal Likelihood

To find $p(m \mid \mathbf{y})$, we need $p(m)$ (which we specify as part of the prior), and $p(\mathbf{y} \mid m)$.

Marginal Likelihood

To find $p(m \mid \mathbf{y})$, we need $p(m)$ (which we specify as part of the prior), and $p(\mathbf{y} \mid m)$.

The latter is the **marginal likelihood** for model m:

Marginal Likelihood

The **marginal likelihood** for a dataset $\mathbf{y}$ given a model class, m, is

$$p(\mathbf{y} \mid m) = \int p(\mathbf{y} \mid \theta, m) p(\theta \mid m)\ d\theta$$

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $m = 1$, we set

$$p(\theta \mid m = 1) = I(\theta = 0.5)$$

(a "degenerate" PMF on θ)

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $m = 1$, we set

  $$p(\theta \mid m = 1) = I(\theta = 0.5)$$

  (a "degenerate" PMF on θ)
- For a biased coin ($m = 2$), we might put a Beta prior on θ, such as a Uniform

  $$p(\theta \mid m = 2) \cdot 1I(0 < \theta < 1)$$

  (a Uniform PDF on $[0, 1]$)

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $m = 1$, we set

$$p(\theta \mid m = 1) = I(\theta = 0.5)$$

(a "degenerate" PMF on θ)

- For a biased coin ($m = 2$), we might put a Beta prior on θ, such as a Uniform

$$p(\theta \mid m = 2) \cdot 1I(0 < \theta < 1)$$

(a Uniform PDF on $[0, 1]$)

- After 40 flips, we see 25 heads.

Example: Fair or Biased Coin?

- Suppose we don't know whether a coin is fair or not.
- For $m = 1$, we set

$$p(\theta \mid m = 1) = I(\theta = 0.5)$$

  (a "degenerate" PMF on θ)
- For a biased coin ($m = 2$), we might put a Beta prior on θ, such as a Uniform

$$p(\theta \mid m = 2) \cdot 1I(0 < \theta < 1)$$

  (a Uniform PDF on $[0, 1]$)
- After 40 flips, we see 25 heads.
- This gives conditional posteriors:

$$\mu \mid \mathbf{y}, m = 1 \sim I(\mu = 0.5)$$
$$\mu \mid \mathbf{y}, m = 2 \sim \text{Beta}(25 + 1, 15 + 1)$$

Fair Coin: Prior and Posterior

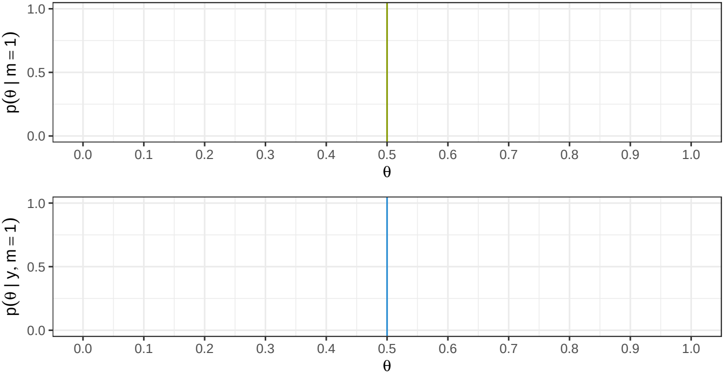


Figure: Top: Prior on θ, conditioned on the coin being fair. Bottom: Posterior on θ, conditioned on the coin being fair. Note that conditioning on the coin being fair makes the data irrelevant for inferring θ

Biased Coin: Prior and Posterior

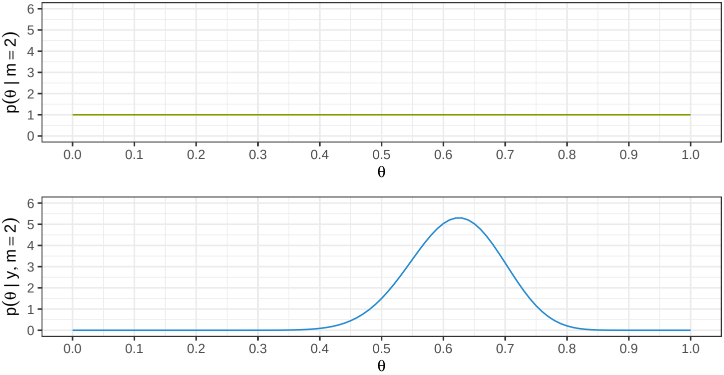


Figure: Top: Prior on θ, conditioned on the coin being biased. Bottom: Posterior on θ, conditioned on the coin being biased. When the coin can have any bias, the posterior concentrates mass near the observed proportion of heads

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$p(y \mid m = 2) = \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta \\
&= \int_0^1 \binom{40}{25} \theta^y (1 - \theta)^{40-y} \times 1 \; d\theta
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta \\
&= \int_0^1 \binom{40}{25} \theta^y (1-\theta)^{40-y} \times 1 \; d\theta \\
&= \binom{40}{25} \int_0^1 \theta^{26-1} (1-\theta)^{16-1} \; d\mu
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta \\
&= \binom{40}{25} \int_0^1 \theta^{26-1}(1-\theta)^{16-1} \; d\mu \\
&= \binom{40}{25} \frac{\Gamma(26)\Gamma(16)}{\Gamma(42)}
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$\begin{aligned} p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta \\ &= \binom{40}{25} \frac{\Gamma(26)\Gamma(16)}{\Gamma(42)} \\ &= \frac{40!}{25!15!} \frac{25!15!}{41!} \end{aligned}$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$\begin{aligned} p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \; d\theta \\ &= \frac{40!}{25!15!} \frac{25!15!}{41!} \\ &= 1/41 \end{aligned}$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$
\begin{aligned}
p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2)\ d\theta \\
&= 1/41 \\
&= \mathbf{0.0243}
\end{aligned}
$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$\begin{aligned} p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2) \, d\theta \\ &= \mathbf{0.0243} \end{aligned}$$

If the coin is fair (i.e., $\theta = 0.5$ with probability 1), then the marginal likelihood is just

$$p(y \mid m = 1) = \binom{40}{25} (1/2)^{25} (1/2)^{15} = \mathbf{0.0366}$$

Example: Fair or Biased Coin?

The marginal likelihood for the biased coin (average probability of 25 heads out of 40) is:

$$\begin{aligned} p(y \mid m = 2) &= \int_0^1 p(y \mid \theta, m = 2) p(\theta \mid m = 2)\, d\theta \\ &= \mathbf{0.0243} \end{aligned}$$

If the coin is fair (i.e., $\theta = 0.5$ with probability 1), then the marginal likelihood is just

$$p(y \mid m = 1) = \binom{40}{25} (1/2)^{25} (1/2)^{15} = \mathbf{0.0366}$$

and so the "fair coin hypothesis" yields a higher **marginal likelihood** than the "biased coin hypothesis" with a uniform prior.

Bayes Factor

- How does this data affect the plausibility that the coin is biased?

Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$
\begin{aligned}
\frac{p(m=2 \mid y)}{p(m=1 \mid y)} &= \frac{p(m=2)p(y \mid m=2)}{p(m=1)p(y \mid m=1)} \\
&= \frac{p(m=2)}{p(m=1)} \times \frac{0.0243}{0.0366} \\
&= \frac{p(m=2)}{p(m=1)} \times 0.663
\end{aligned}
$$

Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$\frac{p(m=2 \mid y)}{p(m=1 \mid y)} = \frac{p(m=2)p(y \mid m=2)}{p(m=1)p(y \mid m=1)}$$
$$= \frac{p(m=2)}{p(m=1)} \times \frac{0.0243}{0.0366}$$
$$= \frac{p(m=2)}{p(m=1)} \times 0.663$$

- Thus, relative to what we believed before seeing the data, our **subjective odds** that the coin is biased **should go down** after seeing 25 heads out of 40! (with the "uniform" notion of what "bias" looks like)

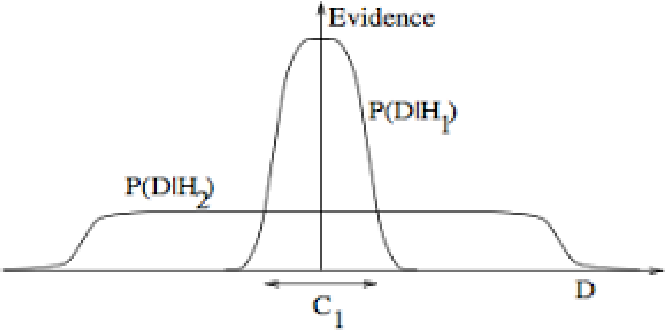
Bayes Factor

- How does this data affect the plausibility that the coin is biased?
- Consider the ratio of the posterior plausibilities of the two model classes:

$$\frac{p(m=2 \mid y)}{p(m=1 \mid y)} = \frac{p(m=2)p(y \mid m=2)}{p(m=1)p(y \mid m=1)}$$
$$= \frac{p(m=2)}{p(m=1)} \times \frac{0.0243}{0.0366}$$
$$= \frac{p(m=2)}{p(m=1)} \times 0.663$$

- Thus, relative to what we believed before seeing the data, our **subjective odds** that the coin is biased **should go down** after seeing 25 heads out of 40! (with the "uniform" notion of what "bias" looks like)
- The ratio of marginal likelihoods, by which our "belief ratio" is scaled, is called the **Bayes Factor**

Conservation of Explanatory Power



Marginal likelihood "rewards" specific predictions

Conservation of Explanatory Power

Probabilistic Occam's Razor

Bayesian Occam's Razor

A "possible world" consists of a model m, along with a (possibly trivial) parameter-setting, θ

$$p(m|\mathbf{y}) = \int \frac{p(m,\theta)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})} d\theta \; d\theta$$

$p(\mathbf{y}|m,\theta)$ Rewards specific predictions by (m,θ)

Bayesian Occam's Razor

A "possible world" consists of a model m, along with a (possibly trivial) parameter-setting, θ

$$p(m|\mathbf{y}) = \int \frac{p(m,\theta)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})} d\theta$$
$$= \int \frac{p(m)p(\theta|m)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})}\ d\theta$$

$p(\mathbf{y}\|m,\theta)$	Rewards specific predictions by (m,θ)
$p(\theta\|m)$	Penalizes flexibility of the model class

Bayesian Occam's Razor

$$p(m|\mathbf{y}) = \int \frac{p(m,\theta)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})} d\theta$$

$p(\mathbf{y}\|m,\theta)$	Rewards specific predictions by (m,θ)
$p(\theta\|m)$	Penalizes flexibility of the model class

Bayesian Occam's Razor

$$
\begin{aligned}
p(m|\mathbf{y}) &= \int \frac{p(m,\theta)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})} d\theta \\
&= \int \frac{p(m)p(\theta|m)p(\mathbf{y}|m,\theta)}{p(\mathbf{y})}\ d\theta
\end{aligned}
$$

$p(\mathbf{y}\|m,\theta)$	Rewards specific predictions by (m,θ)
$p(\theta\|m)$	Penalizes flexibility of the model class