Bayesian Hierarchical Models

Joseph Miller
Rutgers University

February 24, 2016

Background

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

Background

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

- Define:
  - $H =$ a set of hypotheses
  - $D =$ the dataset

Background

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

- Define:
  - $H =$ a set of hypotheses
  - $D =$ the dataset
- Conditional probability (Bayes Rule) provides the "correct" way to update beliefs, given the data collected (and a probability model).

ALGEBRA

$$\frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)} = \frac{P(D|H)P(H)}{\sum_H P(D \cap H)} = \frac{P(D|H)P(H)}{P(D)}$$

Algebra

$$\frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)} = \frac{P(D|H)P(H)}{\sum_H P(D \cap H)} = \frac{P(D|H)P(H)}{P(D)}$$

- $P(D)$ is the *marginal* distribution of the data, also sometimes called the *evidence*.

Algebra

$$\frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)} = \frac{P(D|H)P(H)}{\sum_H P(D \cap H)} = \frac{P(D|H)P(H)}{P(D)}$$

- $P(D)$ is the *marginal* distribution of the data, also sometimes called the *evidence*.
- But think of $\sum_H P(D|H)P(H)$ as a normalization factor so $P(D|H)P(H)$, (a function of H) is a valid PMF.

Algebra

$$\frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)} = \frac{P(D|H)P(H)}{\sum_H P(D \cap H)} = \frac{P(D|H)P(H)}{P(D)}$$

- $P(D)$ is the *marginal* distribution of the data, also sometimes called the *evidence*.
- But think of $\sum_H P(D|H)P(H)$ as a normalization factor so $P(D|H)P(H)$, (a function of H) is a valid PMF.
- If H is a vector, denominator can be difficult to compute.

Example: estimating the bias of a coin

Define:

- $\theta = P(X = \text{heads})$

Example: estimating the bias of a coin

Define:

- $\theta = P(X = \text{heads})$
- $\theta \sim \text{Unif}(0,1) \rightarrow f_\theta(\theta) = 1$
- $X =$ heads with probability θ and
  $X =$ tails with probability $1 - \theta$, so $P(D|\theta) = \theta^x(1-\theta)^{n-x}$
  where x is the number of heads and n is the size of the dataset.

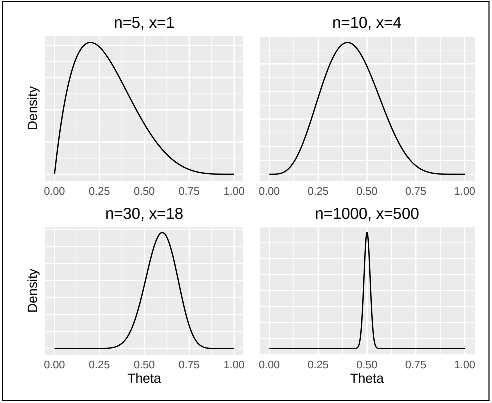
Example: estimating the bias of a coin

Define:

- $\theta = P(X = \text{heads})$
- $\theta \sim$ Unif(0,1) $\rightarrow f_\theta(\theta) = 1$
- $X =$ heads with probability θ and
  $X =$ tails with probability $1 - \theta$, so $P(D|\theta) = \theta^x(1-\theta)^{n-x}$
  where x is the number of heads and n is the size of the dataset.

$$\frac{P(D|\theta)P(\theta)}{\sum_\theta P(D|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times 1}{\sum_\theta \theta^x(1-\theta)^{n-x} \times 1} \propto \theta^x(1-\theta)^{n-x}$$

Example: posterior probability distributions for θ



Example: how does this relate to the conventional approach?

- $P(\theta|D) = \frac{\theta^x(1-\theta)^{n-x}\times 1}{\sum_\theta \theta^x(1-\theta)^{n-x}\times 1}$ has mode at the MLE and for a dataset of 18 Heads out of 30 flips, a 95% HDI of (0.427, 0.760).

EXAMPLE: HOW DOES THIS RELATE TO THE CONVENTIONAL APPROACH?

- $P(\theta|D) = \frac{\theta^x(1-\theta)^{n-x}\times 1}{\sum_\theta \theta^x(1-\theta)^{n-x}\times 1}$ has mode at the MLE and for a dataset of 18 Heads out of 30 flips, a 95% HDI of (0.427, 0.760).
- Corresponding frequentist analysis yields a confidence interval of approximately (0.425, 0.775).

Example: how does this relate to the conventional approach?

- $P(\theta|D) = \frac{\theta^x(1-\theta)^{n-x}\times 1}{\sum_\theta \theta^x(1-\theta)^{n-x}\times 1}$ has mode at the MLE and for a dataset of 18 Heads out of 30 flips, a 95% HDI of (0.427, 0.760).
- Corresponding frequentist analysis yields a confidence interval of approximately (0.425, 0.775).
- Or, $P((X \leq 12) \cup (X \geq 18)|\theta = 0.5) = 0.362$ so do not reject $H_0 : \theta = 0.5$.

Example: how does this relate to the conventional approach?

- $P(\theta|D) = \frac{\theta^x(1-\theta)^{n-x}\times 1}{\sum_\theta \theta^x(1-\theta)^{n-x}\times 1}$ has mode at the MLE and for a dataset of 18 Heads out of 30 flips, a 95% HDI of (0.427, 0.760).
- Corresponding frequentist analysis yields a confidence interval of approximately (0.425, 0.775).
- Or, $P((X \leq 12) \cup (X \geq 18)|\theta = 0.5) = 0.362$ so do not reject $H_0 : \theta = 0.5$.
  - Q. Is this satisfactory?

Example: how does this relate to the conventional approach?

- $P(\theta|D) = \frac{\theta^x(1-\theta)^{n-x}\times 1}{\sum_\theta \theta^x(1-\theta)^{n-x}\times 1}$ has mode at the MLE and for a dataset of 18 Heads out of 30 flips, a 95% HDI of (0.427, 0.760).
- Corresponding frequentist analysis yields a confidence interval of approximately (0.425, 0.775).
- Or, $P((X \leq 12) \cup (X \geq 18)|\theta = 0.5) = 0.362$ so do not reject $H_0 : \theta = 0.5$.
  - Q. Is this satisfactory?
  - Q. What sort of prior does the decision *Do Not Reject* imply?

Conjugate prior

- Back to the likelihood: $\theta^x(1-\theta)^{n-x}$ is the kernel of a Beta$(1+x, n-1+x)$ density. Dividing it by the normalization factor, $\sum_\theta \theta^x(1-\theta)^{n-x}$, yields a Beta distribution, suggesting a new concept:

Conjugate prior

- Back to the likelihood: $\theta^x(1-\theta)^{n-x}$ is the kernel of a Beta$(1+x, n-1+x)$ density. Dividing it by the normalization factor, $\sum_\theta \theta^x(1-\theta)^{n-x}$, yields a Beta distribution, suggesting a new concept:
  - If we want to generalize our prior distribution (before, $P(\theta)=1$), may want to use a prior that results in a posterior, $P(\theta|x)$, that has the same form as the prior.

CONJUGATE PRIOR

- Back to the likelihood: $\theta^x(1-\theta)^{n-x}$ is the kernel of a Beta$(1+x, n-1+x)$ density. Dividing it by the normalization factor, $\sum_\theta \theta^x(1-\theta)^{n-x}$, yields a Beta distribution, suggesting a new concept:
    - If we want to generalize our prior distribution (before, $P(\theta)=1$), may want to use a prior that results in a posterior, $P(\theta|x)$, that has the same form as the prior.
- A *Conjugate Prior* with respect to a likelihood function is a distribution that yields a posterior with the same functional form as the prior.

Conjugate prior

- Back to the likelihood: $\theta^x(1-\theta)^{n-x}$ is the kernel of a Beta$(1+x, n-1+x)$ density. Dividing it by the normalization factor, $\sum_\theta \theta^x(1-\theta)^{n-x}$, yields a Beta distribution, suggesting a new concept:
  - If we want to generalize our prior distribution (before, $P(\theta) = 1$), may want to use a prior that results in a posterior, $P(\theta|x)$, that has the same form as the prior.
- A *Conjugate Prior* with respect to a likelihood function is a distribution that yields a posterior with the same functional form as the prior.
- The conjugate prior for a Bernoulli R.V. is Beta(α, β).

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$

  $\frac{\theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}} = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\Sigma_\theta \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}} =$

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$
$\frac{\theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}} = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\Sigma_\theta \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}} =$
$\frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{B(x+\alpha-1, n-x+\beta-1)} \rightarrow \theta|x \sim \text{Beta}(x+\alpha, n-x+\beta)$

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \to P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$
  $\frac{\theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}} = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\Sigma_\theta \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}} =$
  $\frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{B(x+\alpha-1, n-x+\beta-1)} \to \theta|x \sim \text{Beta}(x+\alpha, n-x+\beta)$
- Beyond the computational advantages, *conjugate priors* allow your prior to be interpreted as *past data*.

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$ $\frac{\theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}} = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\sum_\theta \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}} =$ $\frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{B(x+\alpha-1,n-x+\beta-1)} \rightarrow \theta|x \sim \text{Beta}(x+\alpha, n-x+\beta)$
- Beyond the computational advantages, *conjugate priors* allow your prior to be interpreted as *past data*.
  - If $P(\theta) = 1 \rightarrow \theta \sim \text{Unif(0,1)} \leftrightarrow \text{Beta}(1, 1)$, implying that you have seen a single Heads and a single Tails, but nothing more.

Conjugate Prior continued

- $\theta \sim \text{Beta}(\alpha, \beta) \rightarrow P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$
- $P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\sum_\theta P(x|\theta)P(\theta)} = \frac{\theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}} =$
  $\frac{\theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\sum_\theta \theta^x(1-\theta)^{n-x} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}} = \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\Sigma_\theta \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}} =$
  $\frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{B(x+\alpha-1,n-x+\beta-1)} \rightarrow \theta|x \sim \text{Beta}(x+\alpha, n-x+\beta)$
- Beyond the computational advantages, *conjugate priors* allow your prior to be interpreted as *past data*.
  - If $P(\theta) = 1 \rightarrow \theta \sim \text{Unif(0,1)} \leftrightarrow \text{Beta}(1, 1)$, implying that you have seen a single Heads and a single Tails, but nothing more.
- $\text{E}[\theta|x] = \frac{x+\alpha}{n+\alpha+\beta}$, $\text{var}[\theta|x] = \frac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta)^2(n+1+\alpha+\beta)}$

Interlude

Given $P(\theta|x)$, you can compute $P(x_{\text{new}}|x)$, the posterior predictive distribution:

Interlude

Given $P(\theta|x)$, you can compute $P(x_{\text{new}}|x)$, the posterior predictive distribution:

$$\begin{aligned}
P(x_{\text{new}}|x) &= \Sigma_\theta P(x_{\text{new}}, \theta|x) \\
&= \Sigma_\theta P(x_{\text{new}}|\theta, x) P(\theta|x) \\
&= \Sigma_\theta P(x_{\text{new}}|\theta) P(\theta|x)
\end{aligned}$$

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{7}{11}$	$\frac{?}{10}$

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{7}{11}$	$\frac{?}{10}$

- A few options:

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{7}{11}$	$\frac{?}{10}$

- A few options:
  - Use $x = \Sigma_j x_j$, $n = \Sigma_j n_j$ to get posterior distribution on θ and derive predictive distribution. Assumptions?

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{7}{11}$	$\frac{?}{10}$

- A few options:
  - Use $x = \Sigma_j x_j$, $n = \Sigma_j n_j$ to get posterior distribution on θ and derive predictive distribution. Assumptions?
  - Weight *experiments* equally, e.g. $x \approx n \times \frac{1}{5}\Sigma(\frac{x_j}{n_j})$ and use x and n (same as before) to derive predictive distribution. What problem does this solve?

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{7}{11}$	$\frac{?}{10}$

- A few options:
  - Use $x = \Sigma_j x_j$, $n = \Sigma_j n_j$ to get posterior distribution on θ and derive predictive distribution. Assumptions?
  - Weight *experiments* equally, e.g. $x \approx n \times \frac{1}{5}\Sigma(\frac{x_j}{n_j})$ and use x and n (same as before) to derive predictive distribution. What problem does this solve?
- Q. Define $\theta_j = \mathrm{E}_x[\frac{x_j}{n_j}]$. Suppose $\theta_4 = 0.3$, how does this influence your belief about θ_5?

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{3}{3}$	$\frac{5}{7}$	$\frac{7}{11}$	$\frac{?}{10}$
Parameters	θ_1	θ_2	θ_3	θ_4	θ_5

Hyperparameters α, β

Solution:

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{3}{3}$	$\frac{5}{7}$	$\frac{7}{11}$	$\frac{?}{10}$
Parameters	θ_1	θ_2	θ_3	θ_4	θ_5
Hyperparameters α, β					

Solution:

- Allow α and β to be random draws from a (noninformative) *hyperprior* distribution.

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{3}{3}$	$\frac{5}{7}$	$\frac{7}{11}$	$\frac{?}{10}$
Parameters	θ_1	θ_2	θ_3	θ_4	θ_5

Hyperparameters α, β

Solution:

- Allow α and β to be random draws from a (noninformative) *hyperprior* distribution.
- Imagine that θ_j are random draws from a prior distribution with parameters (α, β). Random draws from a $\text{Bin}(n_j, \theta_j)$ distribution then generate x_j.

Hierarchical Models

Suppose we have the following (incomplete) binomial data:

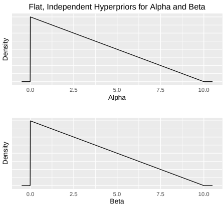
Experiment (j)	1	2	3	4	5
Data ($\frac{x_j}{n_j}$)	$\frac{5}{10}$	$\frac{3}{3}$	$\frac{5}{7}$	$\frac{7}{11}$	$\frac{?}{10}$
Parameters	θ_1	θ_2	θ_3	θ_4	θ_5

Hyperparameters α, β

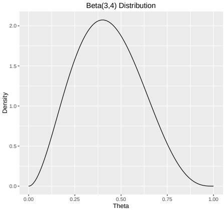
Solution:

- Allow α and β to be random draws from a (noninformative) *hyperprior* distribution.
- Imagine that θ_j are random draws from a prior distribution with parameters (α, β). Random draws from a $\text{Bin}(n_j, \theta_j)$ distribution then generate x_j.
- Q. Can you give an example where it would be useful to insist that whatever the true value of α, $\alpha = \beta$?

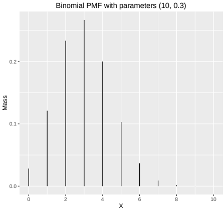
STRUCTURE OF THE MODEL



(a) $P(\phi)$



(b) $P(\theta|\phi = (3,4))$



(c) $P(x|\theta = 0.3)$

$$P(\phi, \theta|x) \propto P(\phi)P(\theta|\phi)P(x|\theta)$$

Hierarchical Models

- First, $P(\phi, \theta) = P(\phi)P(\theta|\phi)$

Hierarchical Models

- First, $P(\phi, \theta) = P(\phi)P(\theta|\phi)$
- $P(\phi, \theta|x) \propto P(\phi, \theta)P(x|\phi, \theta) = P(\phi, \theta)P(x|\theta) = P(\phi)P(\theta|\phi)P(x|\theta)$

Hierarchical Models

- First, $P(\phi, \theta) = P(\phi)P(\theta|\phi)$
- $P(\phi, \theta|x) \propto P(\phi, \theta)P(x|\phi, \theta) = P(\phi, \theta)P(x|\theta) = P(\phi)P(\theta|\phi)P(x|\theta)$
- So, $P(\phi, \theta|x) \propto P(\phi)P(\theta|\phi)P(x|\theta)$ is the joint posterior.

Hierarchical Models

- First, $P(\phi, \theta) = P(\phi)P(\theta|\phi)$
- $P(\phi, \theta|x) \propto P(\phi, \theta)P(x|\phi, \theta) = P(\phi, \theta)P(x|\theta) = P(\phi)P(\theta|\phi)P(x|\theta)$
- So, $P(\phi, \theta|x) \propto P(\phi)P(\theta|\phi)P(x|\theta)$ is the joint posterior. $P(\phi|x) = \Sigma_\theta P(\phi, \theta|x) = P(\phi, \theta|x)/P(\theta|x)$ and $P(\theta|x) = \Sigma_\phi P(\phi, \theta|x)$ are the marginal posteriors.

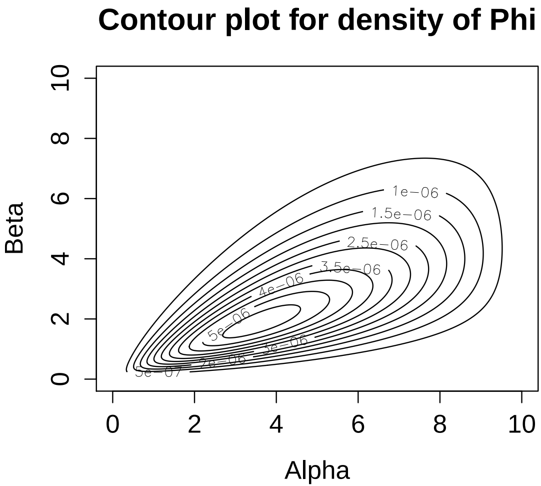
Hierarchical Models

- First, $P(\phi, \theta) = P(\phi)P(\theta|\phi)$
- $P(\phi, \theta|x) \propto P(\phi, \theta)P(x|\phi, \theta) = P(\phi, \theta)P(x|\theta) = P(\phi)P(\theta|\phi)P(x|\theta)$
- So, $P(\phi, \theta|x) \propto P(\phi)P(\theta|\phi)P(x|\theta)$ is the joint posterior.
  $P(\phi|x) = \Sigma_\theta P(\phi, \theta|x) = P(\phi, \theta|x)/P(\theta|x)$ and
  $P(\theta|x) = \Sigma_\phi P(\phi, \theta|x)$ are the marginal posteriors.
  $P(x_{new}|x) = \Sigma_\phi \Sigma_\theta P(\phi|x)P(\theta|\phi, x)P(x_{new}|\phi, \theta, x)$.

JOINT AND MARGINAL POSTERIORS

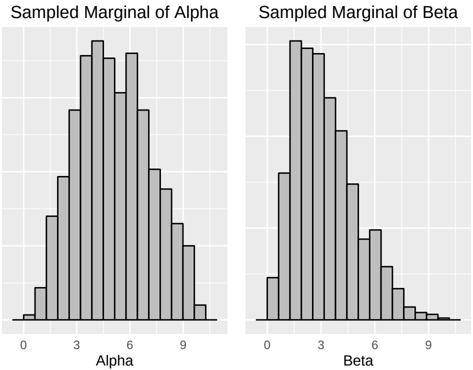
$$
\begin{aligned}
P(\phi, \theta|x) &\propto P(\phi)P(\theta|\phi)P(x|\theta) \\
&\propto P(\phi)\Pi_{j=1}^{4}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\frac{1}{B(\alpha,\beta)}\Pi_{j=1}^{4}\theta_j^{x_j}(1-\theta_j)^{n_j-x_j} \\
&= e^{-\alpha-\beta}\Pi_{j=1}^{4}\theta_j^{\alpha-1+x_j}(1-\theta_j)^{\beta-1+n_j-x_j}\frac{1}{B(\alpha,\beta)} \\
P(\theta|\phi, x) &= \Pi_{j=1}^{4}\theta_j^{\alpha-1+x_j}(1-\theta_j)^{\beta-1+n_j-x_j}\frac{1}{B(\alpha+x_j,\beta+n_j-x_j)} \\
P(\phi|x) &= \frac{P(\phi,\theta|x)}{P(\theta|\phi,x)} \propto e^{-\alpha-\beta}\Pi_{j}^{4}\frac{B(\alpha+x_j,\beta+n_j-x_j)}{B(\alpha,\beta)}
\end{aligned}
$$

Contour plot of $P(\phi|x)$

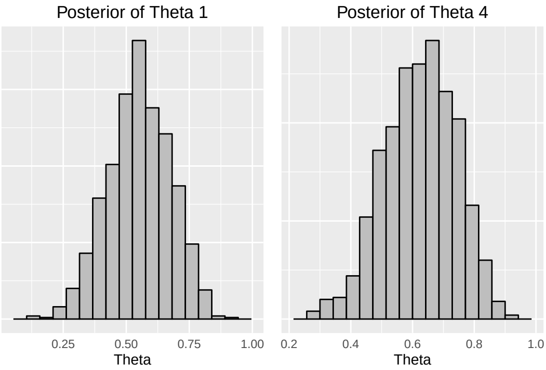


SAMPLED POINTS FROM JOINT PMF OF $\phi|x$

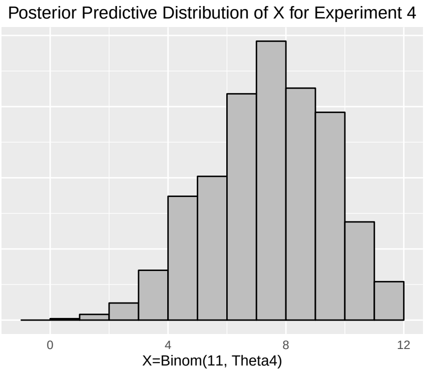
By rejection sampling from the numerically computed joint posterior of ϕ, I find parameter values that my posteriors of interest depend on:



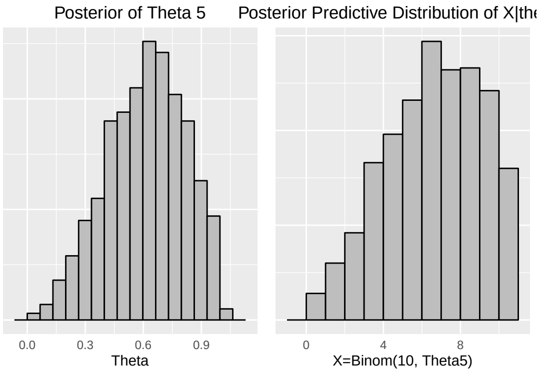
Empirical $P(\theta|x)$ for θ_1 and θ_4



Empirical $P(x_{new}|x)$ for experiment 4 (θ_4)



EMPIRICAL $P(x_{new}|x)$ FOR EXPERIMENT 5 (θ_5)



Given my (too cautious) hyperpriors, not much can be said about experiment five.

MORE CONCEPTS/QUESTIONS

- Posterior predictive checks

More concepts/Questions

- Posterior predictive checks
- Sensitivity analysis

More concepts/Questions

- Posterior predictive checks
- Sensitivity analysis
- Proper/improper priors $\rightarrow$ proper/improper posteriors

More concepts/Questions

- Posterior predictive checks
- Sensitivity analysis
- Proper/improper priors $\rightarrow$ proper/improper posteriors

Questions?

Fin.