Introduction: A Practical Course in Bayesian Modelling

Alexander Ly, Helen Steingroever, Dora Matzke and Eric-Jan Wagenmakers

Psychological Methods
University of Amsterdam

Amsterdam, 30 May 2016

Course logistics

- Goal: Introducing the Bayesian view on statistical modelling using JAGS/WinBUGS and R
- Prerequisite: R
- Literature: Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course
- Examination: Five assignments. Send them to Dora "D.Matzke [Ed] uva.nl" and hand in a printed version before the next lecture
- Website: `http://www.ejwagenmakers.com/BayesCourse/BayesCourse.html`

Overview

Definition of model

- A small object, usually built to scale, that represents in detail another, often larger object.
- As researchers we simplify reality (say, an experiment) and focus only on "the" details that we believe are vital in describing reality (the experiment)
- A preliminary construction that is used in testing or perfecting a final product
- Performing an experiment takes effort and time, while running a (computer) model is cheap
- A schematic description or representation of something, especially a system or phenomenon, that accounts for its properties and is used to study its characteristics.
- It is unethical to someone’s head crack open to study cognition, but it is okay to thinker with a model for cognition to gain knowledge of human cognition

Statistical models

Statistical model

A model embodies a set of statistical assumptions concerning the generation of data, either uncertain future data or already observed data.

These assumptions are the details that we believe to be vital in describing reality (the experiment)

Modelling strategy

1. List all possible outcomes y
2. Identify the parameters θ that generate these possible outcomes
3. Structurally link the parameter to the data $f(y \mid \theta)$

Example: "1-trial" binomial model

1. Outcomes and its Interpretation

Simplest model: Two possible outcomes "0" and "1".

- Coin: "0" means tails, "1" means heads
- Bag of candy: "1" means yellow candy, "0" not yellow,
- Item response theory: "0" the student answered the item incorrectly, "1" answered the item correctly

2. Parameter and its interpretation

We assume that the outcomes of Y are governed by a parameter θ

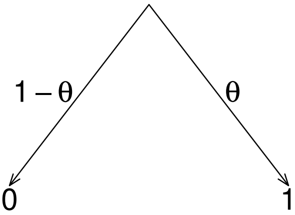
- Coin: θ represents the coin's propensity to fall heads
- Bag: θ represents the true proportion of yellow candies
- IRT: θ represents the student's ability

3. Schematic representation of the binomial distribution

Mathematically, there's a structural relationship $f(y \mid \theta)$

$$f(y \mid \theta) = \theta^y (1 - \theta)^{1-y} \tag{1}$$

that states how θ generates an outcome y



The binomial model consisting of *n*-trials

One observation/trial is not representative or informative.

- Coin: One observation is only representative for $\theta = 0$ or $\theta = 1$
- Bag of candy: One observation is not informative for the proportions of yellow candies
- IRT: It would be crude to decide on a student's ability based on only one item response

As experimenter we gain information by measuring repeatedly, say, *n* times.

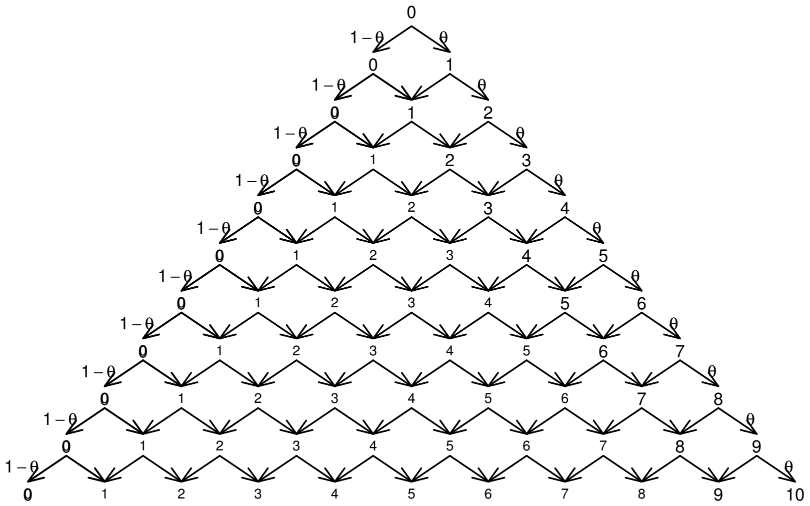
Same modelling strategy

Modelling strategy

1. List all possible outcomes y
2. Identify the parameters θ that generate these possible outcomes
3. Structurally link the parameter to the data $f(y \mid \theta)$

- Q: If we present a student with n items, what are the possible outcomes for the total number of correct responses?
- Answer: The student can respond $\{0, 1, \ldots, n\}$ items correctly

1. Outcomes of a binomial model



2. Interpretation of the parameter θ in *n*-trials

A number at the bottom $\{0, 1, \ldots, 10\}$ represent a possible future outcome y of the experiment Y. The assumption is that one and the same θ underlies the data generating process at each trial.

- Coin: The same propensity θ in all trials. "Coin does not wear".
- Bag of candy: The proportion of θ of yellow stays the same. "Sampling with replacement".
- IRT: Every question is of equal difficulty, student's ability θ stays the same in an exam of n questions. "No learning effects".

3. Structural relationship between θ and the data

The Galton board:

`https://www.youtube.com/watch?v=6YDHBFVIvIs`

Important remarks

- The parameter θ is known: $\theta = 0.5$
- Possible outcomes: $y = 0, 1, \ldots, n$, say, $n = 10$
- Randomness: CANNOT predict where any SINGLE ball will go to
- Frequentist: Can predict the overall pattern for LOTS of balls. This overall behaviour is given by the probability density function (pdf).
- Data generation: The pdf is the structural relationship that links the parameter θ and n to a potential outcome y. Once θ and n are known, we can generate an outcome y

Galton board as a model for an experiment

Galton board is a mechanistic data generating device

- $\theta = 0.5$ is known, thus, "we know where to put the nails"
- n is known, thus, the number of layers and the collection of possible outcomes $\{0, 1, \ldots, n\}$ are known
- Note: One ball is one outcome y of the experiment Y

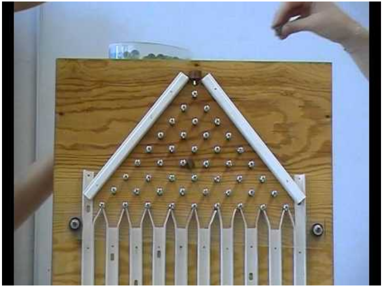
An experiment as a collaborative way of data generation

- A student has her own (fixed) ability θ
- We as experimenter choose the number of trials n
- Note: One student sitting through the exam yields one exam score y of the experiment Y
- Even if we know the student's ability θ exactly, we cannot predict the student's exam score exactly

The Super Academic: Francis Galton (1822 – 1911)

- Field expert: One of the inventors of genetics, psychology, psychometrics, statistics and more
- Mathematician: Strong theoretical background
- Carpenter: Capable of building a board to generate data

Data generation with wood, nails and balls



Modern student in this class

- Field expert: Psychology
- ~~Mathematician: Strong theoretical background~~ Not necessary, (use R) (Though, mathematics is preferred)
- ~~Carpenter: Capable of building a board to generate data~~ Just use R

Modern student in this class

- Field expert: Psychology
- ~~Mathematician: Strong theoretical background~~ Not necessary, (use R) (Though, mathematics is preferred)
- ~~Carpenter: Capable of building a board to generate data~~ Just use R

R equivalent of the Galton board

- Generate one ball

```
> rbinom(1, 10, prob=0.5)
[1] 4
```

- Generate twenty balls

```
> rbinom(20, 10, prob=0.5)
 [1] 4 6 3 4 5 2 5 7 4 5 5 5 4 6 6 6 3 6 5 6
```

Probability density functions (pdf)

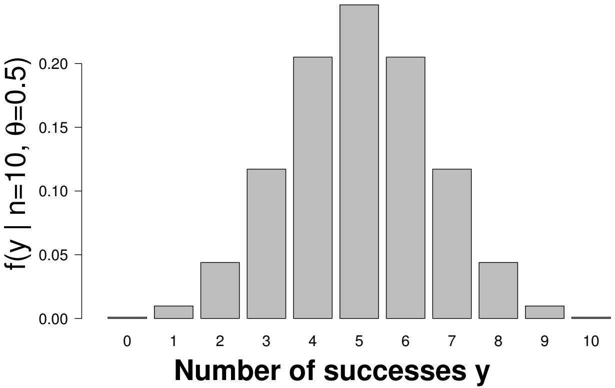
- To generate these data points R uses the probability density function (pdf) of the binomial model
- For n and θ are known, the pdf is

$$f(y \mid \theta, n) = \binom{n}{y} \theta^{y} (1 - \theta)^{n-y} \tag{2}$$

a function of the possible outcomes $y = 0, 1, \ldots, n$

- The pdf is actually the profile at bottom of the Galton's board. Thus, the pdf describes the overall ("lots of balls") behaviour of the outcomes of Y and we write $Y \sim \text{Bin}(\theta, n)$

Example 1: $Y \sim \text{Bin}(\theta = 0.5, n = 10)$ (Plot)



Example 1: $Y \sim \text{Bin}(\theta = 0.5, n = 10)$ (R-Code)

The heights of the bars can be found by typing

```
> data.frame(row.names=0:10,
    chance=dbinom(0:10, 10, 0.5))
         chance
0  0.0009765625
1  0.0097656250
2  0.0439453125
3  0.1171875000
4  0.2050781250
5  0.2460937500
6  0.2050781250
7  0.1171875000
8  0.0439453125
9  0.0097656250
10 0.0009765625
```

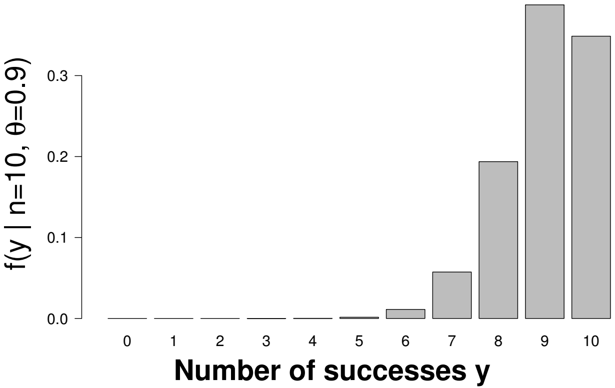
Example 1: $Y \sim \text{Bin}(\theta = 0.5, n = 10)$ (Maths)

The heights of the bars can also be calculated by hand using the pdf

$$f(y \mid \theta = 0.5, n = 10) = \binom{10}{y} 0.5^y (1 - 0.5)^{10-y} \tag{3}$$

where $y = 0, 1, \ldots, 10$.

Example 2: $Y \sim \text{Bin}(\theta = 0.9, n = 10)$ (Plot)



Example 2: $Y \sim \text{Bin}(\theta = 0.9, n = 10)$ (R-Code)

The heights of the bars can be found by typing

```
> data.frame(row.names=0:10,
    chance=dbinom(0:10, 10, 0.9))
         chance
0   0.0000000001
1   0.0000000090
2   0.0000003645
3   0.0000087480
4   0.0001377810
5   0.0014880348
6   0.0111602610
7   0.0573956280
8   0.1937102445
9   0.3874204890
10  0.3486784401
```

Example: Bin($\theta = 0.9, n = 10$) (Maths)

The heights of the bars can also be calculated by hand using the pdf

$$f(y \mid \theta = 0.9, n = 10) = \binom{10}{y} 0.9^y 0.1^{10-y} \tag{4}$$

where $y = 0, 1, \ldots, 10$.

Summary: Data generative view of a model

Statistical model

A model embodies a set of statistical assumptions concerning the generation of data, either uncertain future data or already observed data.

- Data generative view is useful when planning an experiment, that is, before data are observed
- To do so, the structural relationship $f(y \mid \theta, n)$ and the parameters θ and n are supposed to be known
- Thus, the data y are (still) unknown, therefore, random
- We can play around with different values of n and θ to see what we can expect about the overall ("lots of balls") behaviour of the data without actually making a Galton board or recruiting people for an experiment

Explanatory view of a model

Statistical model

A model embodies a set of statistical assumptions concerning the generation of data, either uncertain future data or already observed data.

Inference

- After the data collection, we have the observations y_{obs} which are *not* random
- Likelihood: The functional relationship $f(y_{\text{obs}} \mid \theta)$ is used as an explanatory model of how the data came about
- Goal of inference: Discover which θ is responsible for the observed data y_{obs}, thus, y_{obs} and n are known, while θ is unknown.

Two inference strategies: Estimation and testing

Goal of inference: Discover which θ is responsible for the observed data y_{obs}, thus, y_{obs} and n are known, while θ is unknown.

Inference

- Estimate θ: Guess θ based on the observations y_{obs}
- Hypothesis test: Postulate that θ is known, say, $\theta = \theta_0$ do a prediction about the overall behaviour of the data y and compare this to the observations y_{obs}.
- Note: A prediction should be done before one observe the outcome. Thus, pre-register your hypotheses.

Estimation

A (point) estimate is a best guess for θ based on the data.

Examples

- Estimate θ based on $n = 10$ and $y_{\text{obs}} = 7$
- Uncertainty quantification: How certain are we about this best guess?
- Estimate θ based on $n = 100$ and $y_{\text{obs}} = 70$
- Uncertainty quantification: How certain are we about this best guess?

A Bayesian posterior can give you both a point estimate and a method to quantify the uncertainty about this estimate simultaneously.

Bayesian estimation

- Because θ is unknown, a Bayesian says that her knowledge about θ is random. Hence, θ has a "prior" distribution $\pi(\theta)$
- The prior $\pi(\theta)$ allows us to backtrack θ conditioned on the observations y_{obs} using Bayes' rule.

Bayes' rule

$$\pi(\theta \mid y_{\text{obs}}) = \frac{f(y_{\text{obs}} \mid \theta)\pi(\theta)}{\int f(y_{\text{obs}} \mid \theta)\pi(\theta)\mathrm{d}\theta} \tag{5}$$

Bayes' rule reads as

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalisation constant}} \tag{6}$$

Bayes' rule

$$\pi(\theta \mid y_{\text{obs}}) = \frac{f(y_{\text{obs}} \mid \theta)\pi(\theta)}{\int f(y_{\text{obs}} \mid \theta)\pi(\theta)\mathrm{d}\theta} \tag{7}$$

- Likelihood $f(y_{\text{obs}} \mid \theta)$: The model relates the observations y_{obs} back to the parameter θ
- Prior $\pi(\theta)$: Our knowledge about θ before any datum is observed. $\theta \sim \pi(\theta)$
- Posterior $\pi(\theta \mid y_{\text{obs}})$: Updated knowledge about θ conditioned on the observations y_{obs}
- Normalisation constant $\int f(y_{\text{obs}} \mid \theta)\pi(\theta)\mathrm{d}\theta$ secures that the posterior is a distributions that sums to one
- Note: The observations y_{obs} is known, the likelihood and prior are chosen by the experimenter. To calculate the normalisation constant use WinBUGS and JAGS within R

Prior selection strategy

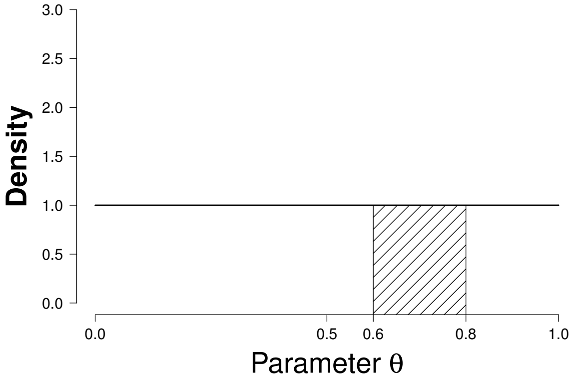
Bayesian statistics requires a prior, which also boils down to choosing a distribution

Prior selection strategy

1. List all possible outcomes for the parameter θ
2. Choose a density $\pi(\theta)$ for θ
3. Robustness check: See how the conclusions change when the prior is changed

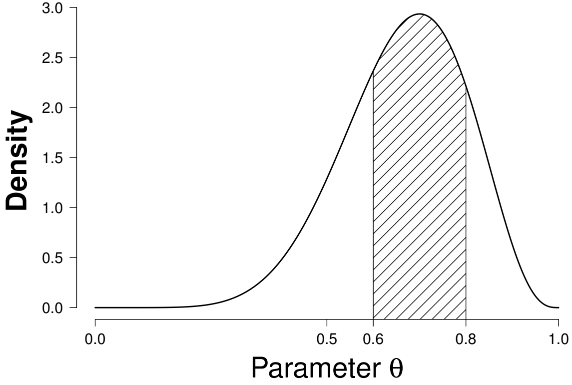
For continuous variables that take values in the bounded interval $(0, 1)$ we typically use a so-called beta distribution, see the shinyApp. Hence, $\theta \sim \text{Beta}(\alpha, \beta)$. When $\alpha = \beta = 1$ this is the uniform distribution on $(0, 1)$.

Bayesian estimation with a uniform prior $\theta \sim \text{Beta}(1, 1)$



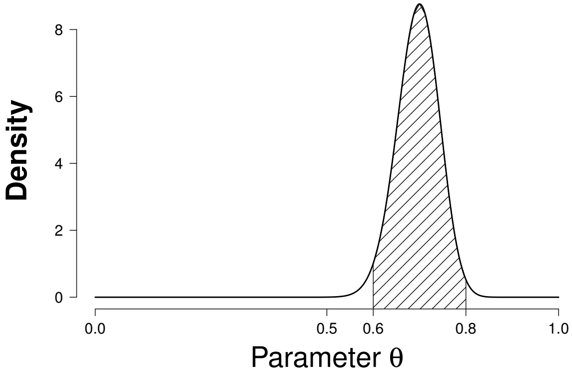
Prior probability of finding θ in $(0.6, 0.8)$ is 20 %.

Posterior given $y_{\text{obs}} = 7$ successes in $n = 10$ and the uniform prior



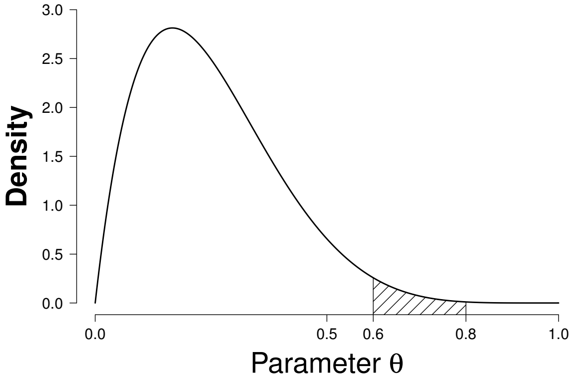
Posterior probability of finding θ in $(0.6, 0.8)$ is 54.2 %.

Posterior given $y_{\text{obs}} = 70$ successes in $n = 100$ and the uniform prior



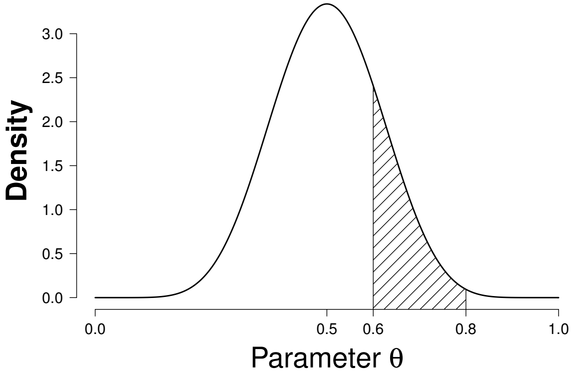
Posterior probability of finding θ in $(0.6, 0.8)$ is 97.2 %.

Bayesian estimation with a Beta(2, 6) prior



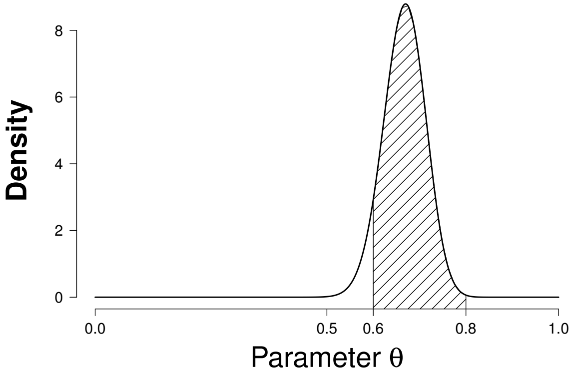
Prior probability of finding θ in $(0.6, 0.8)$ is 1.8 %.

Posterior given $y_{\text{obs}} = 7$ successes in $n = 10$ and a Beta$(2, 6)$ prior



Posterior probability of finding θ in $(0.6, 0.8)$ is 19.6 %.

Posterior given $y_{\text{obs}} = 70$ successes in $n = 100$ and a Beta($\alpha = 2, \beta = 6$) prior



Posterior probability of finding θ in $(0.6, 0.8)$ is 92.6 %.

Bayesian estimation example summary

Probability of finding θ in $(0.6, 0.8)$

y/n (success / trials)	$0/0$	$7/10$	$70/100$
Uniform prior $\theta \sim$ beta$(1, 1)$	20 %	54.2 %	97.2 %
Subjective prior $\theta \sim$ beta$(2, 6)$	1.8 %	19.6 %	92.6 %

Bayesian hypothesis testing: Model comparison

ShinyApp:
http://87.106.45.173:3838/felix/BayesLessons/BayesianLesson1.Rmd

- Think of a person write down the name
- Null hypothesis $\mathcal{H}_0$: The proportion of women is 50%. We, thus, presuppose that the proportion is known and to equal to $\theta = 0.5$
- Alternative hypothesis $\mathcal{H}_1$: The proportion of women can be anything within $(0, 1)$. Use a uniform prior of θ on $(0, 1)$
- The Bayes factor $\text{BF}_{10}(y_{\text{obs}})$ quantifies the evidence in the observations y_{obs} in favour of the alternative hypothesis against the null hypothesis.
- Classroom result: $\text{BF}_{10}(y_{\text{obs}}) = 1.61$, thus, $\text{BF}_{01}(y_{\text{obs}}) = 1/\text{BF}_{10}(y_{\text{obs}}) = 0.621$ with $y_{\text{obs}} = 5$ and $n = 18$

Bag of candy

- Each person takes out a candy and replace it with a new candy of the same type.
- Estimate the true proportion θ of yellow candy in the bag
- Posterior gives both point estimate and an uncertainty quantification
- Note: The true proportion θ is fixed throughout the process and is not random. Our knowledge about θ changes
- Classroom result: True proportion $\theta = 7/(7+11) = 0.39$, samples: 8 yellow and 10 black

Pros and cons of Bayesian statistics

Pros

- Allows for dynamic updating and a natural method to include prior knowledge
- Posterior gives both a point estimate and a credible interval to quantify our uncertainty about the estimate
- Only depends on the data that were actually observed y_{obs}

Cons

- Requires a prior. Choosing a prior can be hard.
- Sensitivity analysis: Check the results under different priors
- To calculate the posterior, need to be able to solve an integral
- No need for hard mathematics anymore, use WinBUGS or JAGS within R

WinBUGS, JAGS and MCMC

- In the olden days, Bayesian statistics was inaccessible due to the normalisation constant in Bayes' rule

$$\pi(\theta \mid y_{\text{obs}}) = \frac{f(y_{\text{obs}} \mid \theta)\pi(\theta)}{\int f(y_{\text{obs}} \mid \theta)\pi(\theta)\mathrm{d}\theta} \tag{8}$$

- Now WinBUGS or JAGS calculates the normalisation constant for you using MCMC sampling (next class)
- All we need to do is specify a likelihood $y \sim f(y \mid \theta)$ and a prior $\theta \sim \pi(\theta)$
- Recall that y_{obs} is not random, but that θ is random
- In fact, WinBUGS or JAGS exploit the fact that θ is random and actually generate samples of θ to calculate the posterior (next class)

Bayesian modelling strategy

Modelling strategy, but use the model as an explanatory device

1. List all possible outcomes for the data y
2. Identify the parameters θ that generate these possible outcomes
3. Structurally link the parameter to the data $f(y \mid \theta)$

Prior selection strategy

4. List all possible outcomes for the parameter θ
5. Choose a density $\pi(\theta)$ for θ
6. Robustness check: See how the conclusions change when the prior is changed

JAGS model file:

In both cases, we end up with a distribution

Data part: Likelihood

```
k ~ dbin(theta, n)
```

(In these slides, I used y instead k)

Prior part

```
theta ~ dbeta(1, 1)
```

Data part: Likelihood

```
data <- list("k", "n")
```

in "Rate_1.R" tells JAGS that y is not random

```
y ~ dbin(theta, n)
```

thus, the binomial distribution is used as a exploratory model

Prior part

```
myinits <- list(list(theta = 0.1)... )
```

in "Rate_1.R" tells JAGS that theta is random,

```
theta ~ dbeta(1, 1)
```

thus, JAGS knows that it needs to "sample" from θ to calculate the normalisation constant in Bayes' rule

Homework

1. Read and do the ShinyApp. (Getting started with Bayesian statistics using an app)
2. Read Chapters 1 and 2 of the book. (Learn to program the same app in R)

Write down any questions you have for EJ and bring them to the next class.

Slides can be found on

`http://www.alexander-ly.com/teaching/`

ShinyApp:

`http://87.106.45.173:3838/felix/BayesLessons/BayesianLesson1.Rmd`

Website:

`http://www.ejwagenmakers.com/BayesCourse/BayesCourse.html`