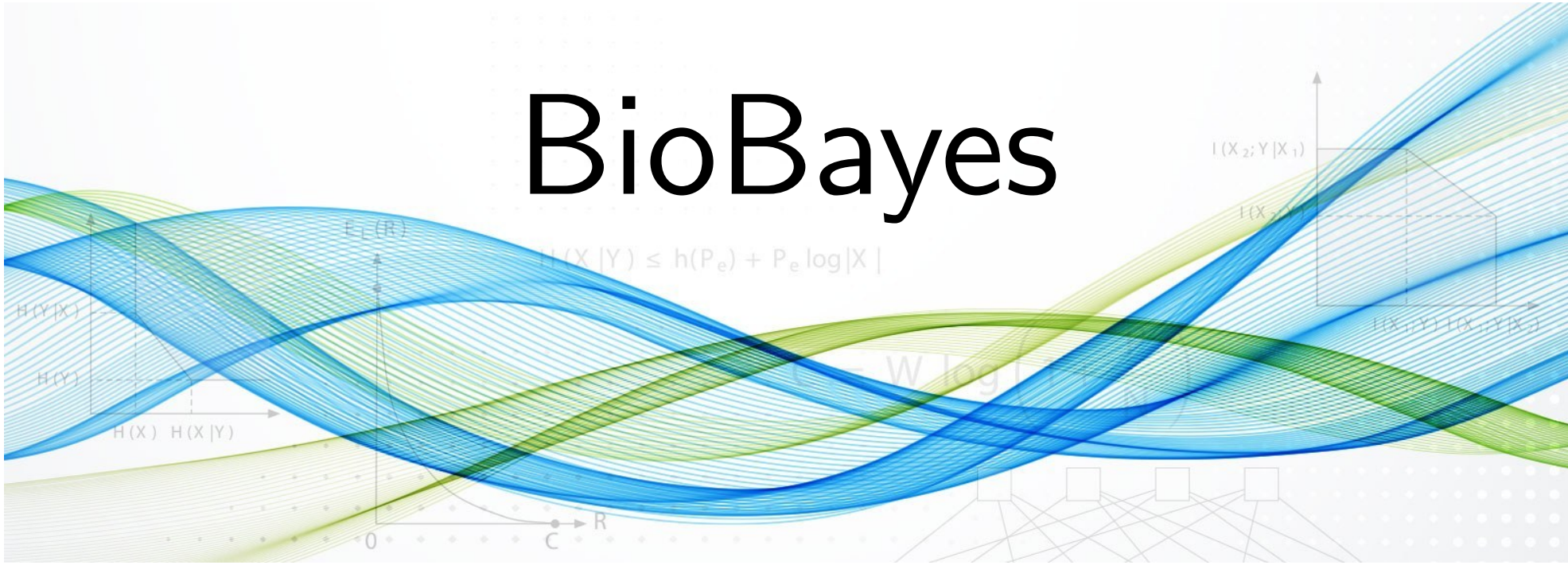


BioBayes

Workshop on Bayesian statistics with R

Centre for Developmental Neurobiology,
3-11 May 2018

Plan

Today	Introduction to Bayesian statistics
Tuesday 8 May	*RJags*: from statistical model design...
Friday 11 May	... to Bayesian inference

Introduction

- Understanding data variability
- Statistical confidence and probability
- Building statistical models
- Inference of “unknown” features from “known” data

Statistics uses the concept of **probability** to quantify **uncertainty**

Probability

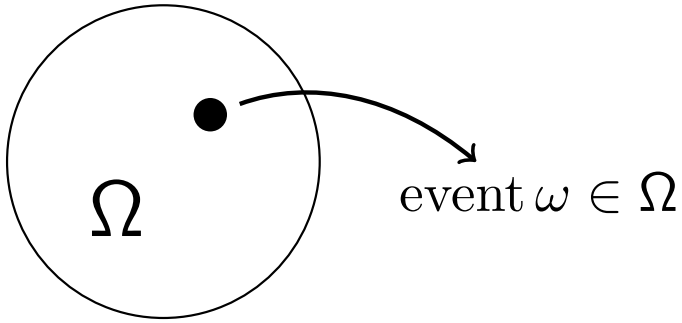
Probability is a well defined mathematical concept

Uncertainty

Uncertainty is a context-dependent concept which is characterized and interpreted differently according to different statistical approaches

Probability

- measure on a set of possible events



- normalization

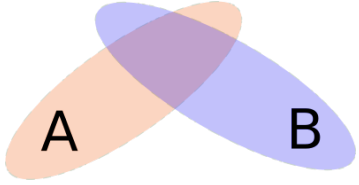
$$\rightarrow \sum_{\text{event}\in\Omega} P(\text{event}) = 1$$

Sum and product rules

For any event subset A and B of Ω

- **Sum rule:** (logic OR)

$$\text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \cap B)$$



- **Product rule:** (logic AND)

$$\text{Prob}(A \cap B) = \text{Prob}(A) \cdot \text{Prob}(B|A)$$

where $P(B|A)$ is the conditional probability of B "given" A

Exercise

What happens for mutually exclusive events?

Joint events

Given two events A and B the probability of observing A and B simultaneously is the "joint" probability

$$P(A, B) = \text{Prob}(A \text{ and } B)$$

A and B are statistically "independent" if

$$\text{Prob}(A \text{ and } B) = \text{Prob}(A) \cdot \text{Prob}(B)$$

Conditional events

If A and B are not independent, having already observed a given value of A changes ("conditions") the odds of observing B to

$$P(B|A) \equiv \frac{P(\mathrm{A\ and\ B})}{P(A)}$$

- $P(B|A)$ is normalized in the sense $\sum_B P(B|A) = 1$ for all A
- $P(A) = \sum_B P(\mathrm{A\ and\ B})$ is the "marginal" with respect to the variable B. In the probability jargon, "marginalizing" is the process of summing/integrating over one or more variables.

Statistical models

Statistical models provide a probabilistic description of how the data have been generated. We will consider three types of models

- **simple models:** data are modeled by a single known probability distribution (normal, exponential etc..)
- **mixed models:** data are distributed according to a linear combination of known distributions
- **hierarchical models:** the parameters are hierarchically organized

Statistical models can involve *latent*, unobserved, features which participate to the generative procedure.

Parameters and likelihood

Given a statistical model M characterized by a set of parameters $\boldsymbol{\theta}$ the data **likelihood**

$$L_M(\boldsymbol{\theta}) = \mathrm{Prob}(\mathit{data}|\boldsymbol{\theta}, M)$$

measures the probability to generate the data from model M using the given parameter set $\boldsymbol{\theta}$.
In the presence of latent variables **z** alongside the observed variables **x**, the likelihood is expresse as the sum over the unobserved variables

$$L_M(\boldsymbol{\theta}) = \sum_{\mathbf{z}} \mathrm{Prob}(\mathbf{x} = \mathit{data}, \mathbf{z}|\boldsymbol{\theta}, M)$$

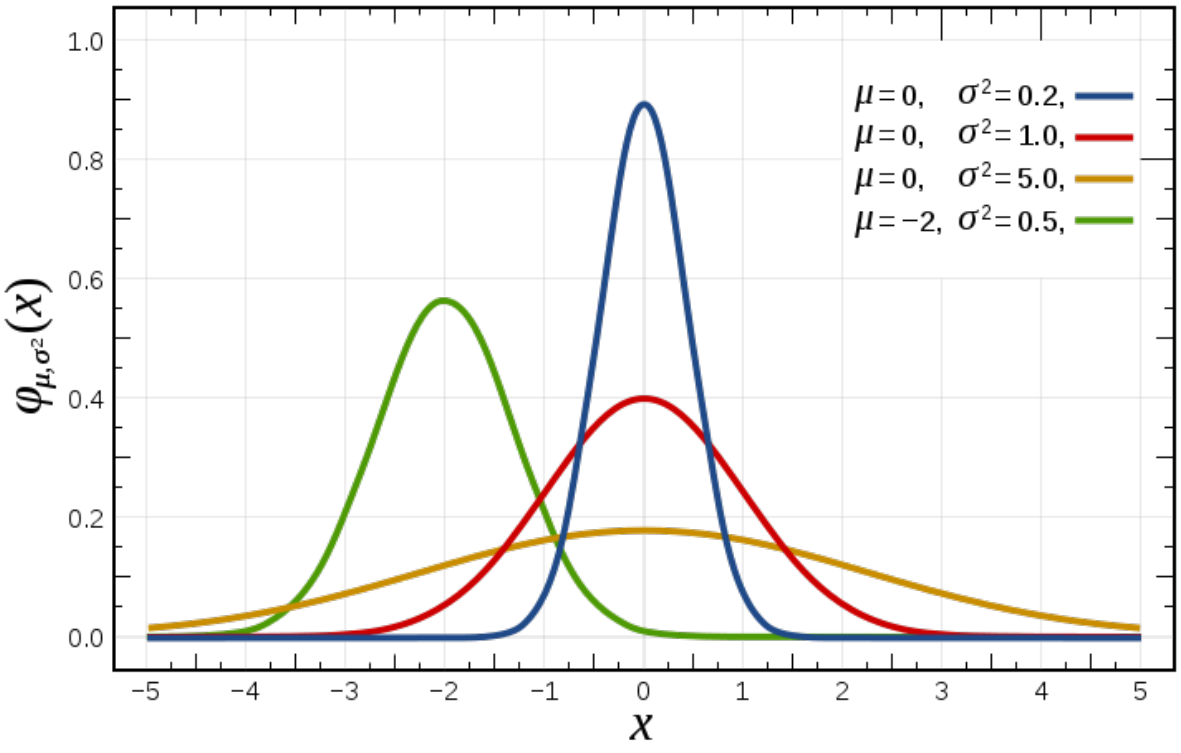
Support of a distribution

Statistical models are built by considering **stochastic processes** that might have generated the data and the numerical range (**support**) of the data.

1. **Discrete support**: positive integers, signed integers, integers within a range.
2. **Continuous support**: real numbers, positive reals, intervals.

Based on the numerical range of the data we can choose suitable probability distributions to describe the data.

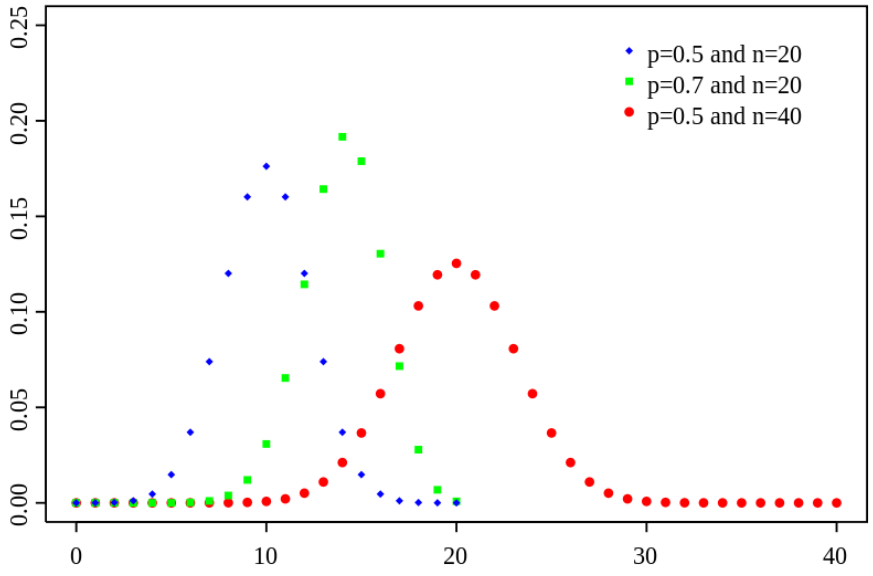
Common Distributions



Normal distribution

- range: real numbers
- parameters: mean (μ) and variance (σ)

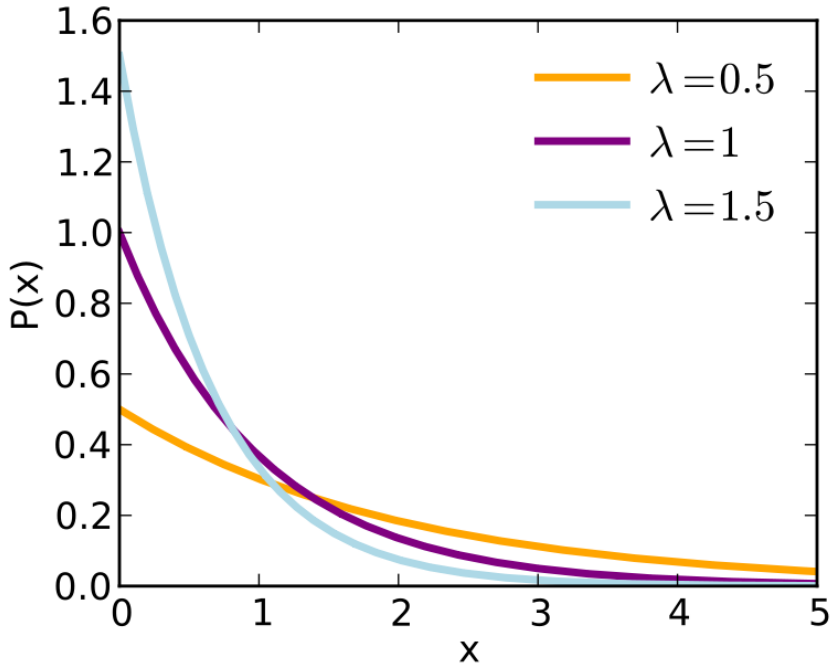
Common Distributions



Binomial distribution

- Given an event which occurs with probability p on a single trial, it models the probability of observing k times the same event over N trials.
- range: positive integers
- parameters: probability (p), number of trials (N)

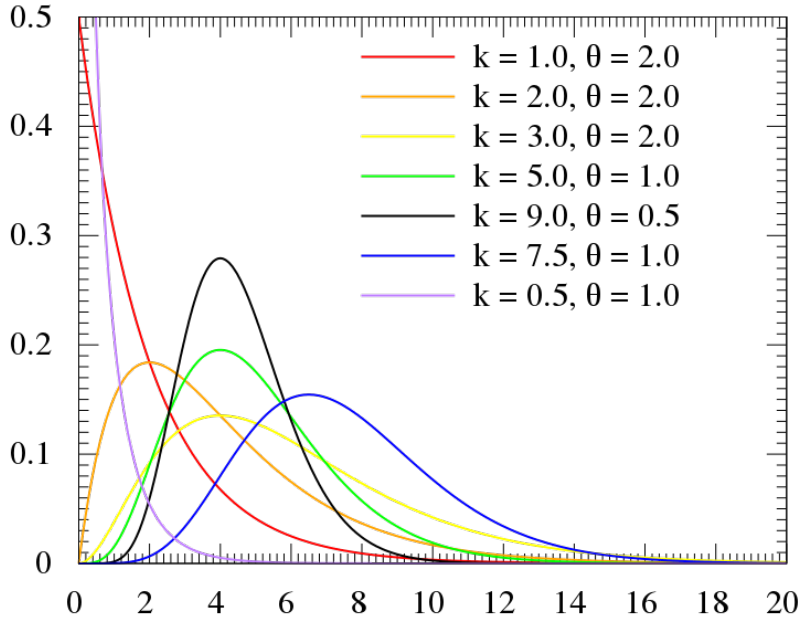
Common Distributions



Exponential distribution

- It models the waiting time between two events of a Poisson process occurring with rate λ
- range: positive reals
- parameters: rate λ

Common Distributions



Gamma distribution

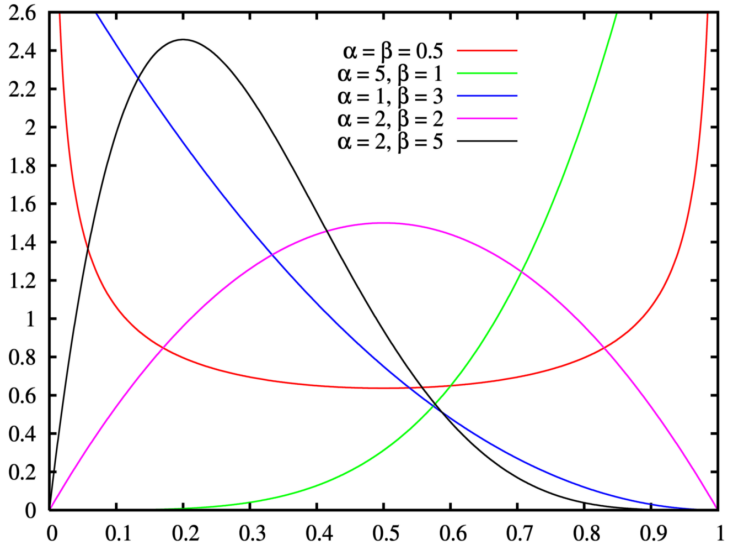
- It provides a flexible model of positive real numbers (gene expression levels, waiting times, effect sizes etc.).
- range: positive reals
- parameters: shape (k) and scale (θ)

Common Distributions

Wishart distribution

- Models multidimensional positive real numbers (combinatorial gene expression levels, covariance matrices, ect.).
- range: positive reals
- parameters: scale (S), degrees of freedom (ν)

Common Distributions



Beta distribution

- It models the probability of a single event.
- range: $[0, 1]$
- parameters: pseudo-observations α, β

Common Distributions

Dirichlet distribution

- Generalization of the beta distribution: it models the probabilities of K mutually exclusive events.
- range: $[0, 1]^K$
- parameters: pseudo-observations $\alpha_1, \cdots, \alpha_K$, $\beta_1, \cdots, \beta_K$

Common Distributions

Multinomial distribution

- Generalization of the binomial distribution: it models the number of observations of K mutually exclusive events occurring with probabilities $p_1, \cdots, p_K$ from N trials.
- range: positive intergers
- parameters: probabilities $p_1, \cdots, p_K$

Exercises

- **1** In an experiment we can observe two types of cells in a tissue, which distribution can we use to model the proportions of the two cell types?
- **2** How can we model the length of the axon initial segment of a single cell type?
- **3** How can we model the joint expression of 2 genes in a tissue?
- **4** Imagine we want to characterize the frequency of some repetitive animal behaviour. Which probability distribution can we use?
- **5** In a simplified behavioural model we consider four states: hunt, rest, explore or escape. Which distribution can we use to characterize the behavioral frequencies?

Expectations

Given a probability distribution $P(x)$ we can compute the average of x with respect to $P(x)$ as

$$\mathbb{E}[x] = \langle x \rangle_P = \sum_{x \in \Omega} P(x) \cdot x$$

More generally,

$$\mathbb{E}[g(x)] = \langle g(x) \rangle_P = \sum_{x \in \Omega} P(x) \cdot g(x)$$

Examples

- mean: $g(x) = x$
- variance: $g(x) = x^2$

Sampling and Montecarlo approximation

To calculate expectations from the formulas above we need to perform sums or integrals (often in many dimensions) $\rightarrow$ HARD!
A much easier way to compute expectations is to generate first a large "sample" of values of x drawn from its probability distribution $\{x_1, \cdots, x_N\}$ and use the **Montecarlo approximation**

Montecarlo approximation

$$x_i \sim P(x), \quad i = 1, \cdots, N$$

$$\mathbb{E}[g(x)] \approx \frac{1}{N} \sum_{i=1}^{N} g(x_i)$$

where the symbol $\sim$ stands for "drawn from"

Exercise 6

We collected data from 100 patients to study the effect of a gene G on a certain desease. G can be found in the two variants A and B. Given the data

	A	B
sick	40	5
healty	20	35

- Calculate $P(\mathrm{sick}|A)$ and $P(\mathrm{sick}|B)$.
- Calculate the probability of being sick or healty.

Statistical inference

Back to our previous exercise on conditional probabilities. Although we can get the empirical estimate of the probabilities of each of the four combinations (gene variant A/B, sick/healthy) we can consider the four probabilities as parameters of a multinomial distribution and try to compute the probability

$$\mathrm{Prob}(p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B} | \textit{data table})$$

where s and h refer to sick and healty state of the patient.
NOTE: The multinomial model tells us how to generate data given parameters $p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}$.

Statistical inference corresponds to solve the inverse problem of obtaining the distribution of the parameters of the model given the data.

Bayes theorem

Assume we have a set of data and a statistical model with parameters θ. We are interested in the "posterior" probability

$$\mathrm{Prob}(\theta|\mathrm{data}, \mathrm{Model})$$

Meaning: how does my uncertainty about the model parameters change after observing the data?

$$\begin{aligned} P(\theta, \mathrm{data}|\mathrm{Model}) &= P(\mathrm{data}|\mathrm{Model}) \cdot P(\theta|\mathrm{data}, \mathrm{Model}) \\ &= P(\theta|\mathrm{Model}) \cdot P(\mathrm{data}|\theta, \mathrm{Model}) \end{aligned}$$

Posterior distribution

$$P(\theta|\mathrm{data}, \mathrm{Model}) = \frac{\overbrace{P(\mathrm{data}|\theta, \mathrm{Model})}^{\text{likelihood}} \cdot \overbrace{P(\theta|\mathrm{Model})}^{\text{prior}}}{\underbrace{P(\mathrm{data}|\mathrm{Model})}_{\text{marginal likelihood}}}$$

Bayesian computation

- Statistical inference corresponds to the computation of the posterior probabilities of the parameters we are interested in. Although in most of the cases we cannot obtain closed expressions (formulas) for the posterior distributions, we can still generate samples from these distributions using Montecarlo algorithms.
- Samples can be then used to obtain the parameter distributions or to calculate "posterior expectations" of quantities of interest (means, variances, correlations...).
- **Good news:** We still need to focus on (1) how to generate data from the model and (2) choose prior distribution of the parameters that reflect our knowledge of the parameters prior acquiring the data.

More good news

Montecarlo sampler are already available. We will focus on RJAGS which allows us to obtain posterior samples after setting up generative model and prior distributions.

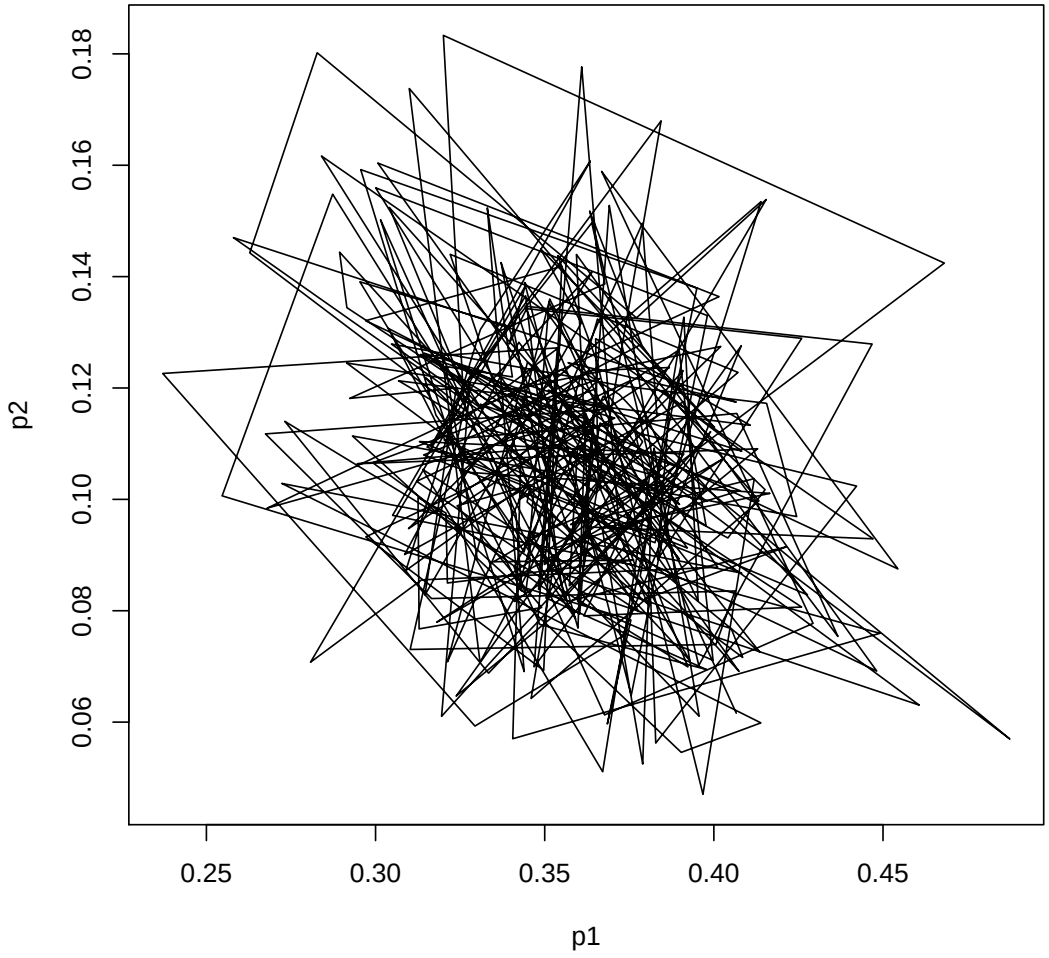
Example

In the case of exercise 6 we simply have

$$\text{Model} : \{N_{s,A}, N_{h,A}, N_{s,B}, N_{h,B}\} \sim \mathcal{M}(\{p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}\}; N = 100)$$
$$\text{Prior} : \{p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}\} \sim \mathit{Dirichlet}(\vec{\alpha})$$

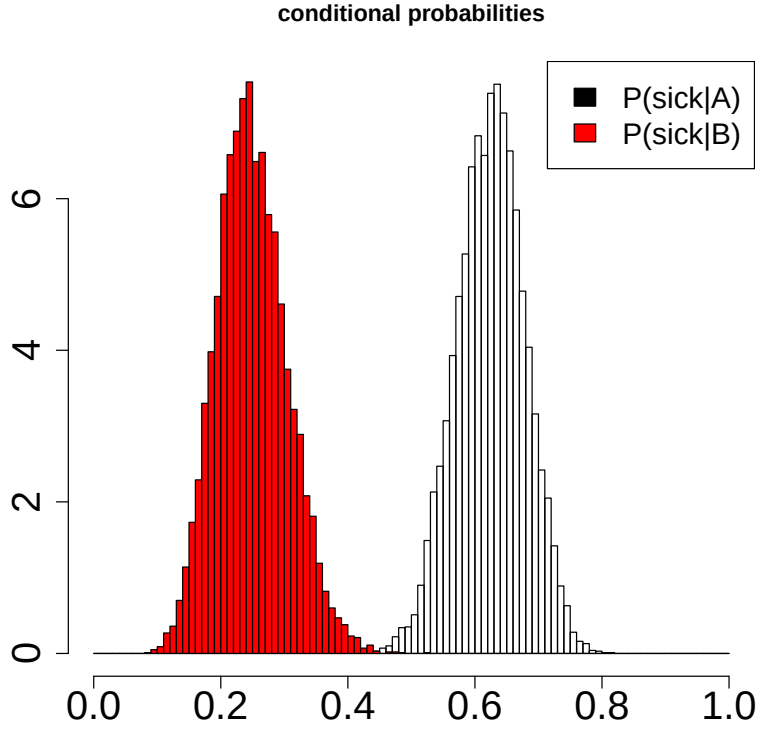
Montecarlo samplers

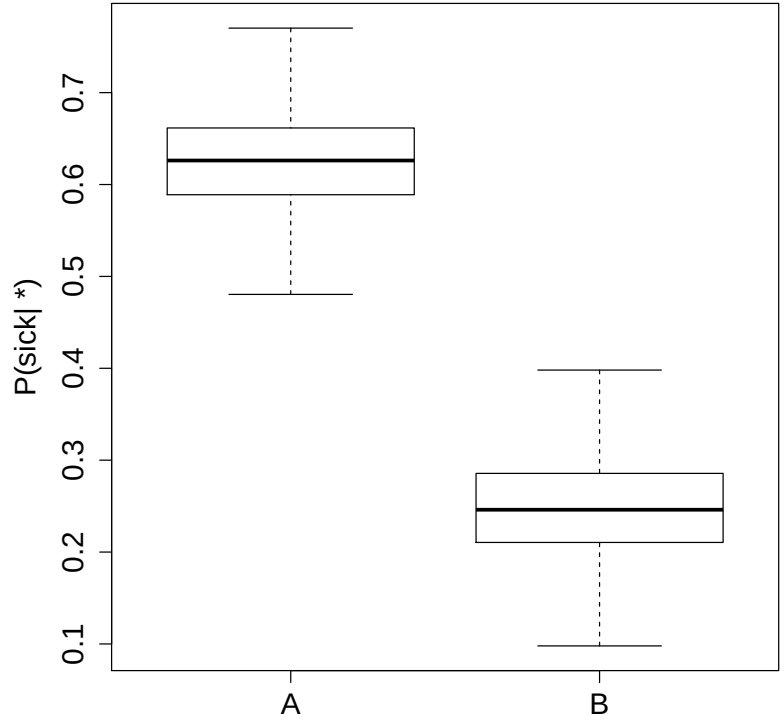
Montecarlo samplers generate Markov chain sequences which explore the parameter space by visiting more often regions of higher posterior probability.



After sampling $p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}$ from their posterior distribution we can compute the conditional probabilities as

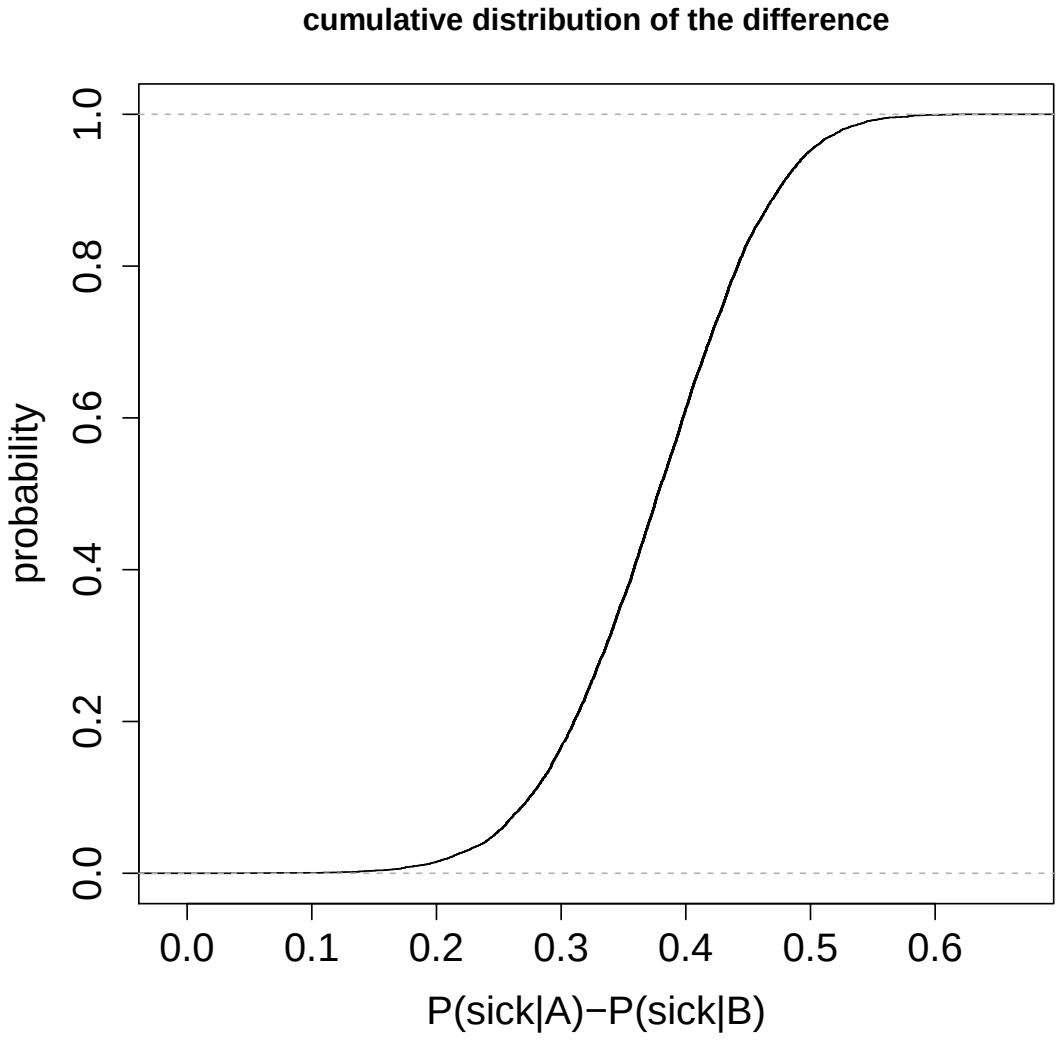
$$P(sick|A) = \frac{p_{s,A}}{p_{s,A} + p_{h,A}}; \quad P(sick|A) = \frac{p_{s,B}}{p_{s,B} + p_{h,B}}$$





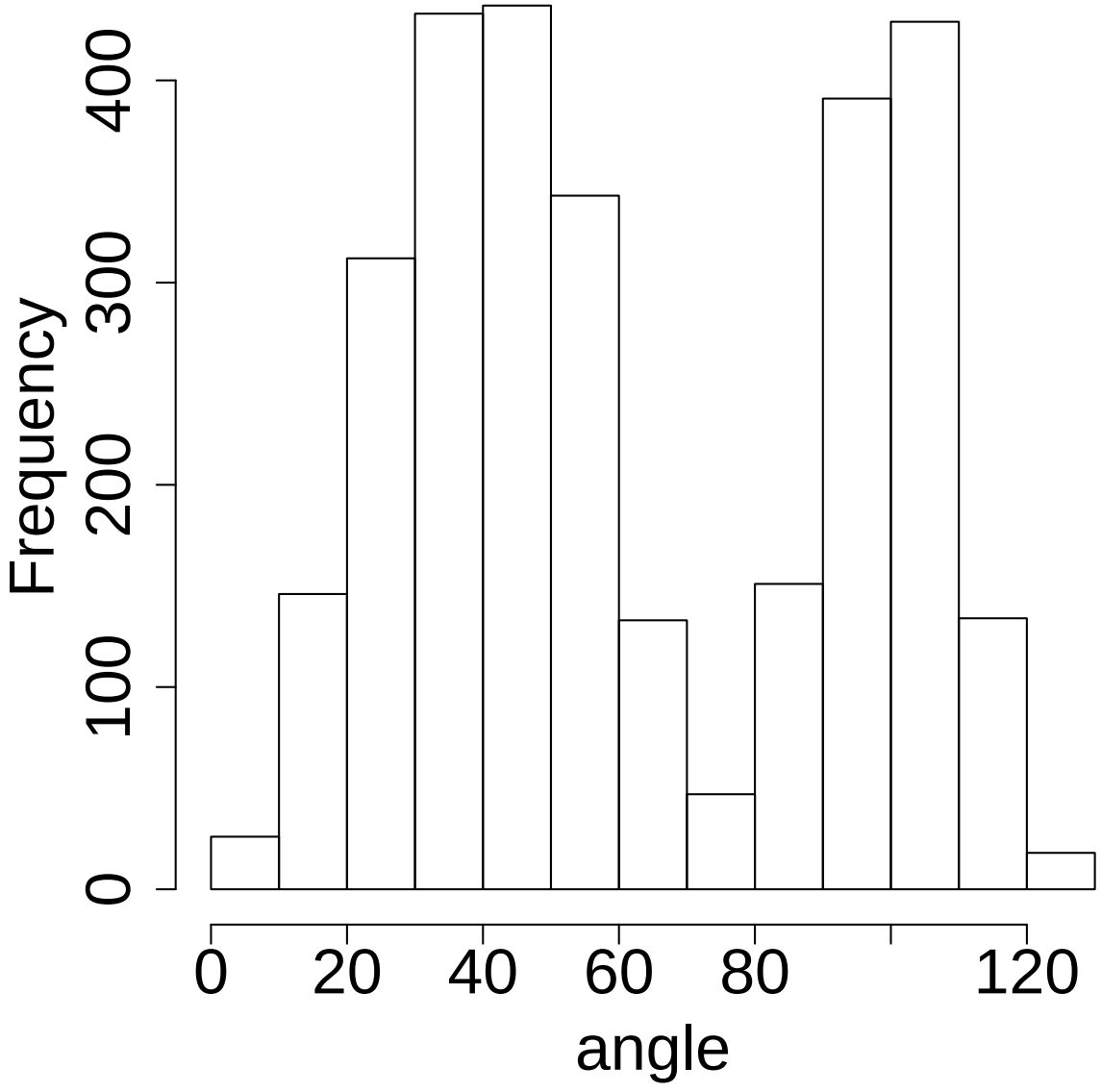
Exercise

Is the difference between $P(sick|A)$ and $P(sick|B)$ is significant?



Mixed models and latent variables

Consider the following distribution of cell orientation in a tissue



We can model the angle distribution with a mixture of two normal distributions.

Model

The angle distribution is characterized by two cell types. The orientation distribution of each cell type is normally distributed.

We define a "cell type" variable (latent) $t = 1, 2$ so that we have

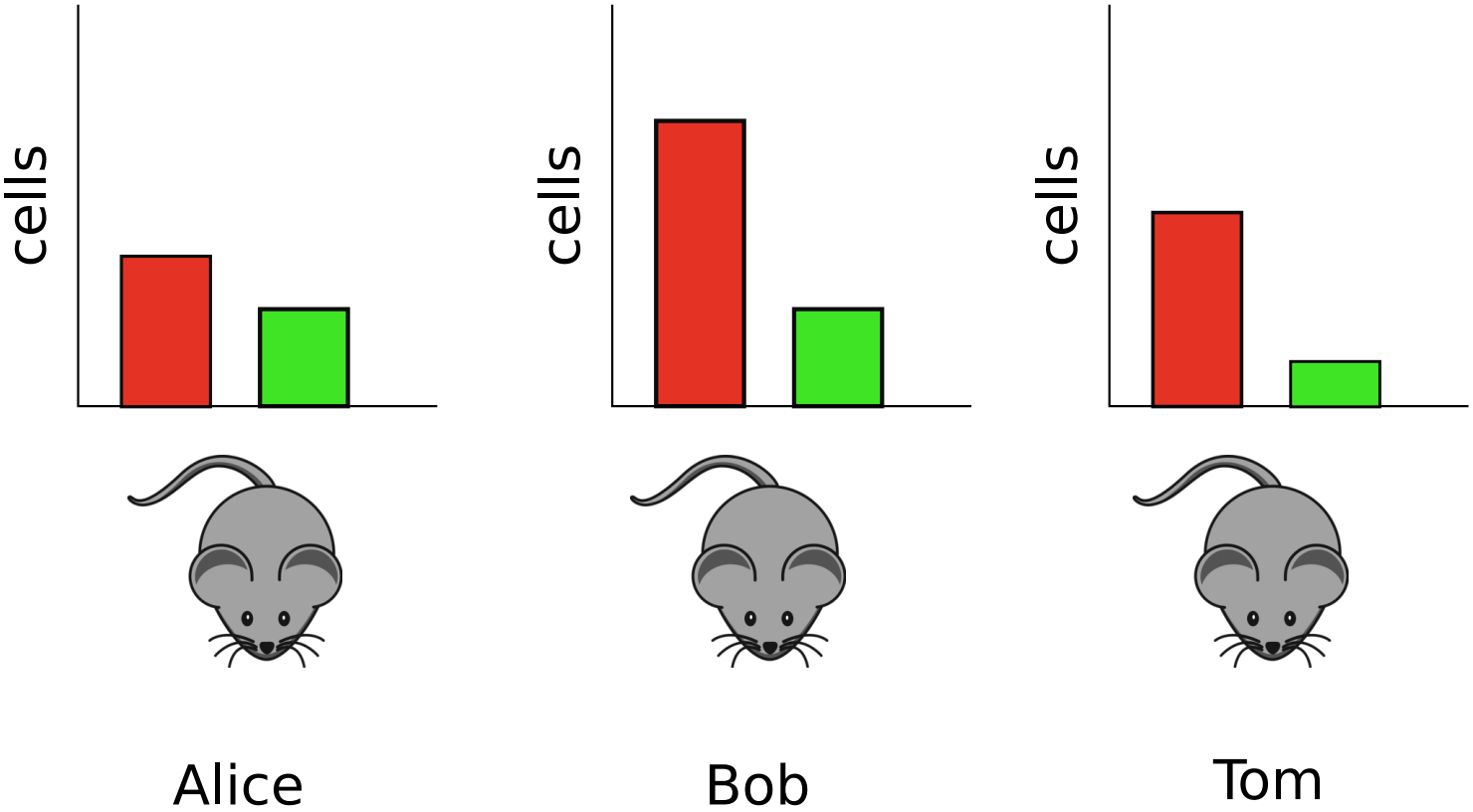
$$
\begin{aligned}
angle_i &\sim Normal(\mu_i, t_i) \\
t_i &\sim Binomial(1/2, N) \\
\mu_1, \mu_2 &\sim Normal(0, 0.001) \\
\sigma_1, \sigma_2 &\sim Gamma(2, 1)
\end{aligned}
$$

Exercise

Calculate the conditional distribution $P(t_i|\mu_1, \mu_2, \sigma_1, \sigma_2)$

Hierarchical models: motivation

- Consider an idea experiment where we want to study the cell size. Assume we can mark big (red) and small (green) cells. We count the cells in three trials.

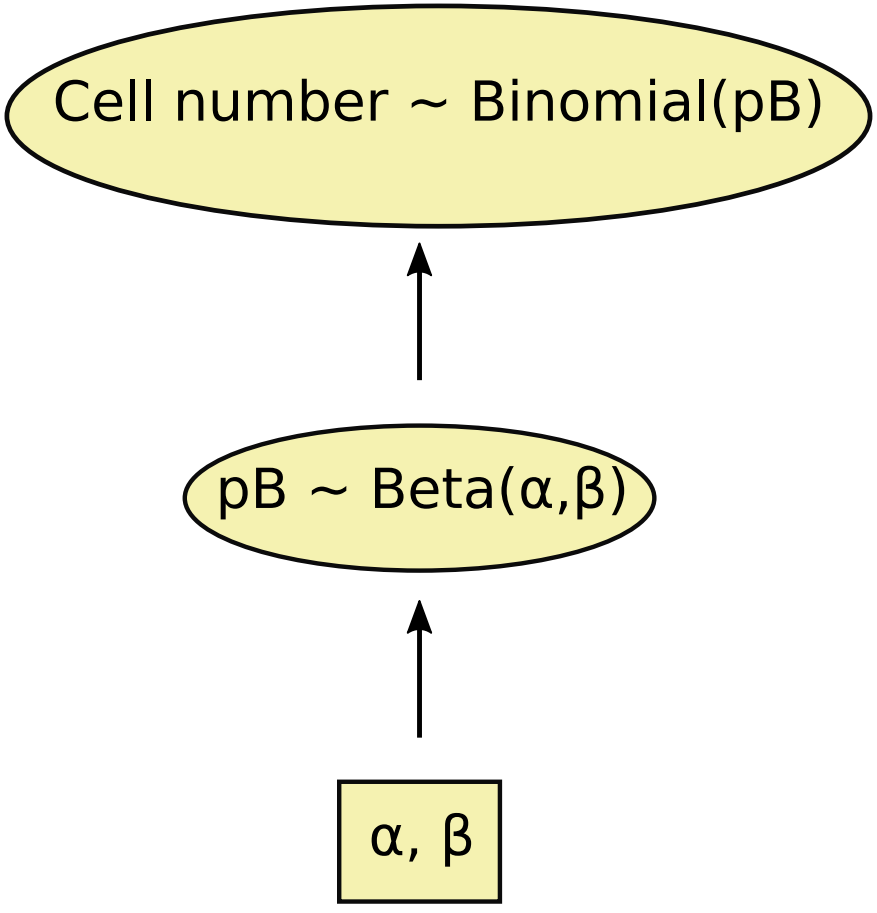


- How can we combine the data from multiple trials?
- How to compare multiple conditions taking into account trial variability?

Hierarchical models

- Collapsing our three trials into one single graph would correspond to assume that the probability of a cell to belong to each population does not depend on the experiment.
- Implying that cell sizes are independent identically distributed random variable.
- In the vast majority of cases biological variability breaks IID-ness.
- The Bayesian solution to this problem is to formulate s multi-level statistical model which takes into account the biological variability in the within-trial models (binomial)
- Hierarchical structure allows data from different trials to "borrow strength".

Hierarchical models



Number of large and small cells are sampled from pB's

Single trial proportions are sampled using α and β

superparameters characterizing biological variability on the proportion of large and small cells

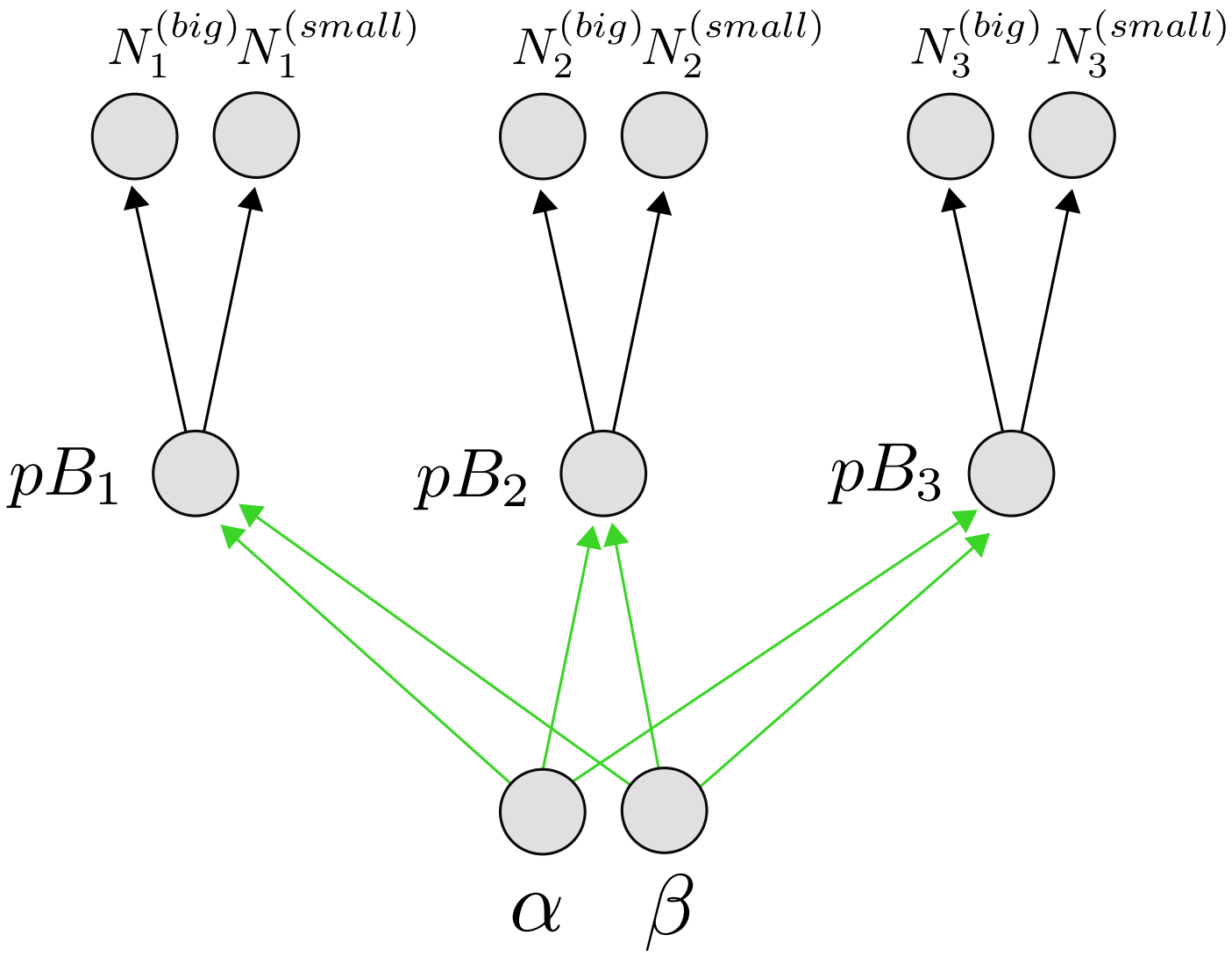
for each trial i

$$N_i^{(big)} \sim \text{Binomial}\left(p_i^{(big)}, N_i^{(tot)}\right) \tag{1}$$

$$p_i^{(big)} \sim \text{Beta}(\alpha, \beta) \tag{2}$$

$$\alpha, \beta \sim \text{Exponential}(\lambda) \tag{3}$$

Graphical representation



Bayesian model comparison

- Model selection and hypothesis testing in the classical (frequentist) framework is often an issue.
- Usually we have a null hypothesis $H0$ which we accept or reject based on the likelihood of observing the data given $H0$
- Comparing models with different number of parameters, avoiding over- and under-fitting issues implies the use of cross-validation techniques which are not guaranteed to work, especially for small datasets.
- Model selection in a Bayesian context is straightforward!

Given Model 1 and Model 2 to be compared we evaluate the Bayes factor

$$K = \frac{\mathrm{Prob}(\mathrm{data}|\mathrm{Model\ 1})}{\mathrm{Prob}(\mathrm{data}|\mathrm{Model\ 2})}$$

where $\mathrm{Prob}(\mathrm{data}|\mathit{Model})$ are the marginal likelihood, *i.e.* averaged over the model parameters.