Imperial College London

# Bayesian statistics

*Introduction to Bayesian methods in (ecology) and evolution*

Matteo Fumagalli

13-17 February 2017

# Plan of the week (1)

## Monday

Bayesian thinking

- Bayesian ideas
- Bayes' Theorem
- Prior, likelihood and posterior distributions
- Practical: genome reconstruction from sequencing data

## Tuesday

Bayesian concepts

- More about prior distributions
- Inferences: point estimate, credible intervals, hypothesis testing
- Practical: estimation of population variation

# Plan of the week (2)

## Wednesday

Bayesian computation

- Asymptotic methods
- Noniterative sampling methods
- Markov chain Monte Carlo
- (Darwin's birthday @NHM)

## Thursday

Approximate Bayesian Computation

- Basic algorithms
- Model design
- Model assessment
- Practical: inference of speciation times

# Plan of the week (3)

## Friday

Experimental design

- Research lecture by Dr Pascale Gerbault
- Journal club

Figure: Dr Gerbault (UCL, UoWestminster)

# Useful information

- Scripts for examples, exercises, case studies, solutions: https://github.com/mfumagalli/BayesianMethods
- Completed exercises will guarantee a +3 bonus at the exam (for my question only!)
- Our case study will be the polar bear!

Figure: Polar bears: a Bayesian approach to understand their evolution.

# Bayesian thinking

Figure: Nessie, the Loch Ness Monster. True or fake?

# Bayesian thinking

- $T = \{0, 1\}$, whether I tell you I saw Nessie or not.
- $N = \{0, 1\}$, whether Nessie exists or not.

## Questions

- What are $p(T = 1|N = 1)$ and $p(T = 1|N = 0)$?
- What is a Maximum Likelihood Estimate of $N$?

# Bayesian thinking

Our inference on $N$ is driven solely by our observations, given by our likelihood function.

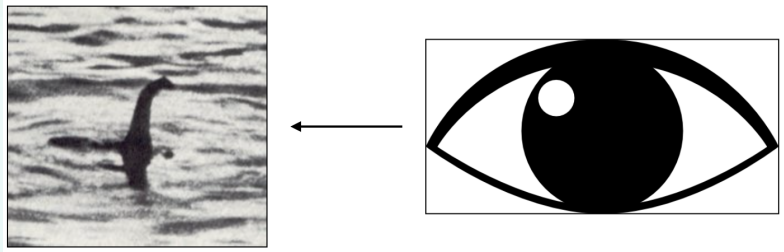


Figure: The eye: a "likelihood" organ.

# Bayesian thinking

In real life we take many decisions based not only on what we observe but also on some believes of ours.

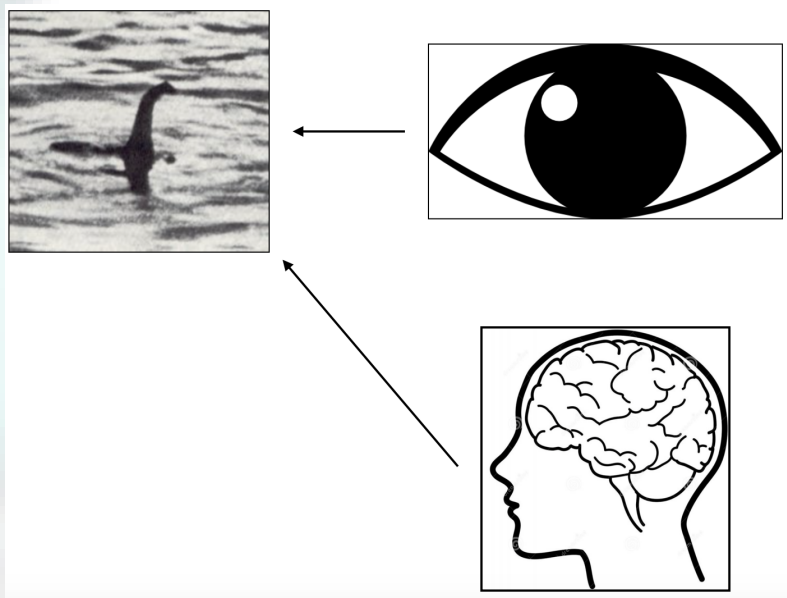


Figure: The brain: a "non-likelihood" organ.

# Bayesian thinking

- with "eyes only" our intuition is that $p(N|T) \approx p(T|N)$
- with "the brain" our intuition is that $p(N|T) \approx p(T|N)p(N)$

Our "belief" expresses the probability $p(N)$ **unconditional** of the data.

**Question**

How can we define $p(N)$?

# Bayesian thinking

The "belief" function $p(N)$ is called **prior probability** and the joint product of the likelihood $p(T|N)$ and the prior is proportional to the **posterior probability** $p(N|T)$.

The use of posterior probabilities for inferences is called Bayesian statistics.

# Bayesian thinking

Bayesian statistics is an alternative to frequentist approaches but without a definite division as in many cases the approach taken is **eclectic**.

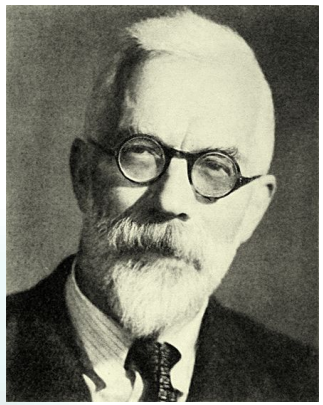


Figure: Ronald Fisher.

# Bayesian thinking

Example:
You submitted a manuscript for publication to a peer-reviewed journal and you want to assess its probability of being accepted and published.

**Question**

Which information would you use to make such inference?

# Bayesian thinking

Example:

You are measuring the biodiversity of some species on some rock shores in four different locations over three years.

Table: Biodiversity levels.

| Year | Loc. A | Loc. B | Loc. C | Loc. D |
|---|---|---|---|---|
| 2014 | 45 | 54 | 47 | 52 |
| 2015 | 41 | ? | 43 | 45 |
| 2016 | 32 | 38 | 37 | 35 |

**Question**

What is a reasonable value for the missing entry?

# Statistical inference

- Frequentist (repeated sampling from a model)
- Likelihoodist (as above using only observations)
- Bayesian (using prior information)
- Empirical Bayesian (observed data contribute to the prior)

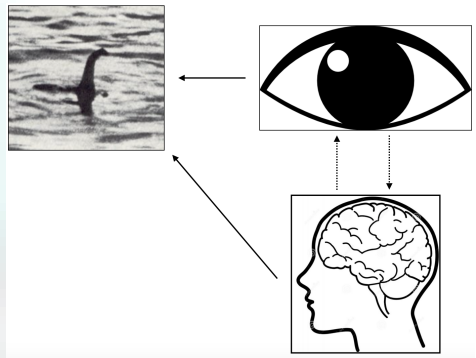


Figure: The brain **and** the eye: an Empirical Bayesian organ.

# Statistical inference

If $D$ is the data and $\theta$ is your unknown parameter, then

- the frequentist conditions on parameters and integrates over the data, $p(D|\theta)$,
- the Bayesian conditions on the data and integrates over the parameters, $p(\theta|D)$.

# Statistical inference

## Bayesian *vs.* Likelihoodist

- we derive "proper" probability distributions of our parameters rather than deriving a point estimate;
- a probability is assigned to a hypothesis rather than a hypothesis is tested;
- we can "accept" the null hypothesis rather than "fail to reject" it;
- parsimony imposed in model choice rather than correcting for multiple tests.

# History of Bayesian statistics

Figure: Thomas Bayes and Pierre-Simon, marquis de Laplace.

Figure: Pierre-Simon, marquis de Laplace

# Bayesian statistics

## John K. Kruschke (Revised 14 November 2010)

Statistical methods have been evolving rapidly, and many people think it's time to adopt modern Bayesian data analysis as standard procedure in our scientific practice and in our educational curriculum. Three reasons:
1. Scientific disciplines from astronomy to zoology are moving to Bayesian data analysis. We should be leaders of the move, not followers.
2. Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive. Bayesian analyses are readily computed with modern software and hardware.
3. Null-hypothesis significance testing (NHST), with its reliance on p values, has many problems. There is little reason to persist with NHST now that Bayesian methods are accessible to everyone.

# Bayesian statistics

Trouble with the prior?

## John K. Kruschke (Revised 14 November 2010)

Some people may have the mistaken impression that the advantages of Bayesian methods are negated by the need to specify a prior distribution. [...]
* It is inappropriate not to use a prior. [...]
* Priors are explicitly specified and must be agreeable to a skeptical scientific audience. [...] If skeptics disagree with the specification of the prior, then the robustness of the conclusion can be explicitly examined by considering other reasonable priors. In most applications, with moderately large data sets and reasonably informed priors, the conclusions are quite robust.
* [...]
* When different groups of scientists have differing priors, stemming from differing theories and empirical emphases, then Bayesian methods provide rational means for comparing the conclusions from the different priors.

# Bayesian statistics

Trouble with the p-value?

## John K. Kruschke (Revised 14 November 2010)

Although there are many difficulties in using p values, the fundamental fatal flaw of p values is that they are ill defined, because any set of data has many different p values. [...]
The literature is full of articles pointing out the many conceptual misunderstandings held by practitioners of NHST. For example, many people mistake the p value for the probability that the null hypothesis is true. Even if those misunderstandings could be eradicated, such that everyone clearly understood what p values really are, the p values would still be ill defined. Every fixed set of data would still have many different p values.

# Bayesian statistics

## Why is it becoming so commonly used?

- recent increased computing power
- good frequentist properties
- answers are easily interpretable by non-specialists
- already implemented in packages

# Bayesian statistics

## Why are WE using it?

Bayesian statistics is very used in many topics in life sciences:

- genetics (e.g. fine mapping of disease-susceptibility genes)
- ecology (e.g. agent-based models)
- evolution (e.g. inference of phylogenetic trees)
- bioinformatics (e.g. genome assembly)
- systems biology (e.g. gene networks)
- ...

# Case study

Figure: Polar bears and brown bears. What's their evolutionary history?

# Case study


Figure: Location of samples of polar and brown bears collected in the high Arctic.

# Case study

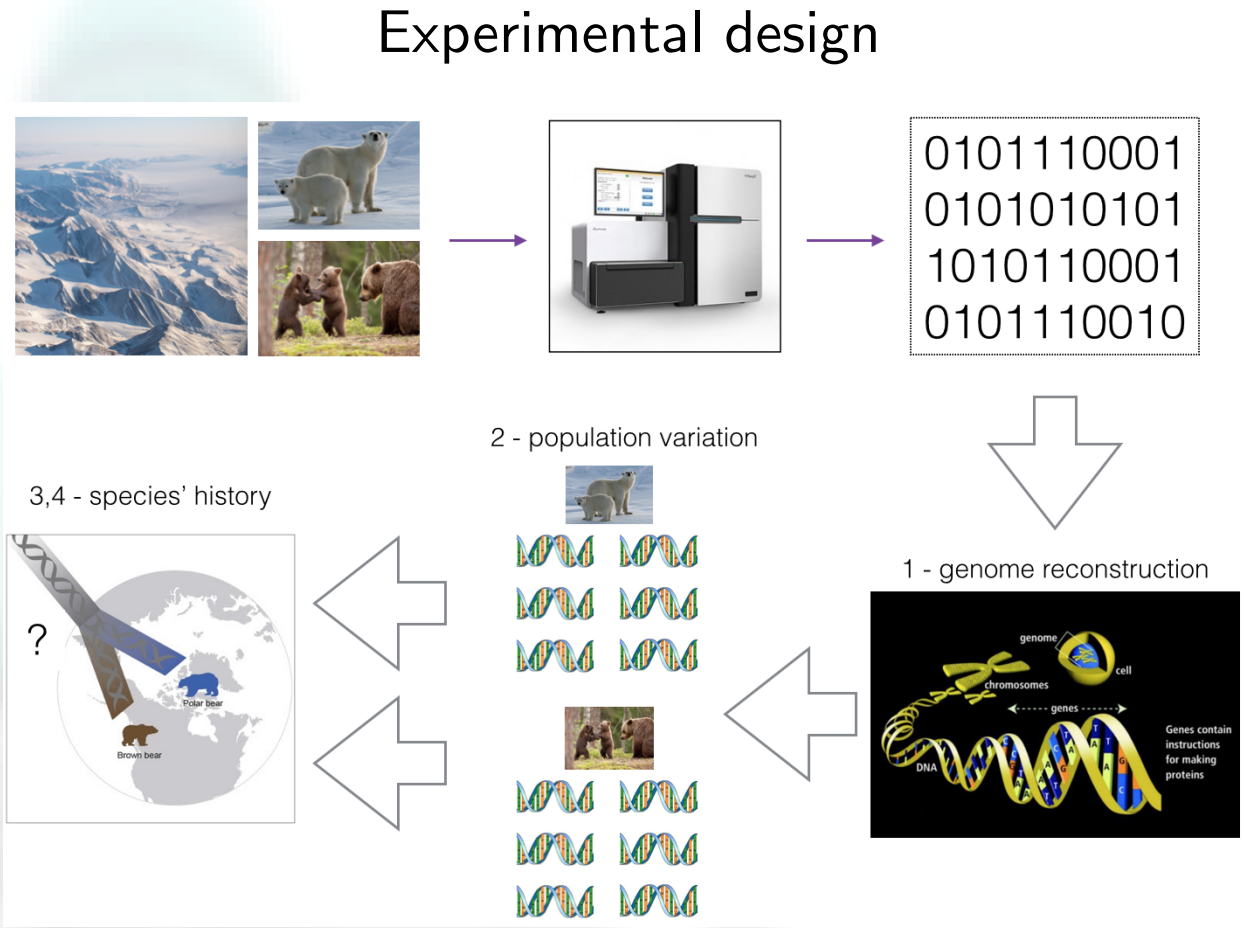



Figure: What is the evolutionary relationship between polar bears and brown bears? Insights from genomic data.

# Bayesian concepts

## The likelihood approach

- $Y$ is a random variable
- $f(\vec{y}|\vec{\theta})$ is a probability distribution (called the *likelihood*) representing the sampling model for the observed data $\vec{y} = (y_1, y_2, ..., y_n)$ given a vector of unknown parameters $\vec{\theta}$
- $\int f(\vec{y}|\vec{\theta})d\vec{\theta}$ is not necessarily $= 1$ or even finite
- It is possible to find the value of $\vec{\theta}$ that maximises the likelihood function: we can calculate a *maximum likelihood estimate* (MLE) for $\vec{\theta}$, as: $\hat{\vec{\theta}} = argmax_{\vec{\theta}} f(\vec{y}|\vec{\theta})$

# Bayesian concepts

## Bayes' Theorem

$$p(\vec{\theta}|\vec{y}) = \frac{f(\vec{y}|\vec{\theta})\pi(\vec{\theta})}{m(\vec{y})} = \frac{f(\vec{y}|\vec{\theta})\pi(\vec{\theta})}{\int f(\vec{y}|\vec{\theta})d\vec{\theta}} \quad (1)$$

- $\vec{\theta}$ is not a fixed parameter but a random quantity with prior distribution $\pi(\vec{\theta})$
- $p(\vec{\theta}|\vec{y})$ is the posterior probability distribution of $\vec{\theta}$
- $\int p(\vec{\theta}|\vec{y})d\vec{\theta} = 1$

# Bayesian concepts

Bayes' Theorem in action!

Figure: The chytrid fungus *Batrachochytrium dendrobatidis* is the most significant threat to amphibian populations.

# Bayes' Theorem

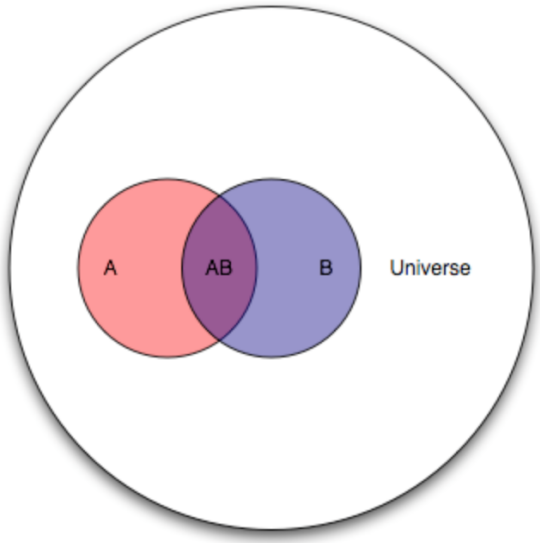
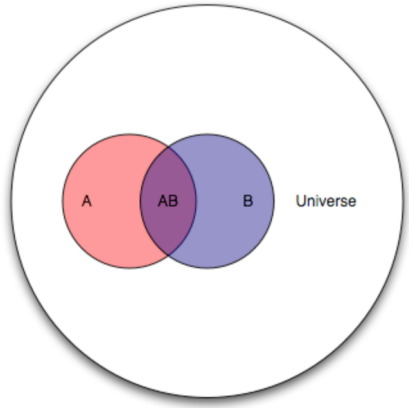



Figure: Sets $U$, $A$ (samples with infection), $B$ (samples with positive test result) and $A \cap B$ (or $AB$).

Given that the test is positive for a randomly selected sample, what is the probability that said sample is infected?

# Bayes' Theorem




If:

- 1% of samples collected are infected,
- 80% of samples with infection will get positive test,
- 9.6% of samples without infection will also get positive tests,

**Question**

If a sample had a positive test, what is the probability that the sample is actually infected?

# Bayes' Theorem

## A normal-normal model

If

$$f(y|\theta) = N(y|\theta, \sigma^2)$$
$$\pi(\theta) = N(\theta|\mu, \tau^2)$$

then

$$p(\theta|y) = N(\theta|\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}) \quad (2)$$

# Bayes' Theorem

## A normal-normal model

"Shrinking" factor $B$:

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2} \tag{3}$$

with $0 \leq B \geq 1$.
Then

$$E(\theta|y) = B\mu + (1 - B)y$$
$$Var(\theta|y) = (1 - B)\sigma^2 \equiv B\tau^2$$

Question:

- What if $\sigma^2 >> \tau^2$?
- What if $\sigma^2 << \tau^2$?

# Bayes' Theorem

## A normal-normal model

Assume that $y = 6$, $\sigma = 1$, $\mu = 2$ and $\tau = 1$.

Exercise:

- Open `R` and calculate and plot the prior, likelihood, and posterior distributions.
- Calculate the *maximum a posteriori probability* (MAP).
- What happens if we use a skewer (sharper) or wider prior?
- What happens if we have more observations?

# Monte Carlo sampling

Drawing random samples from it instead of directly calculating the parameters.

Figure: Monte Carlo and its famous casino.

Exercise:

- Open `R` and plot the posterior distribution of the previous example using Monte Carlo sampling.
- What happens if we use more or less draws?

# CASE STUDY (1)

Reconstructing genomes from sequencing data.

All scripts used in these examples and for the exercise are available at `https://github.com/mfumagalli/BayesianMethods/blob/master/Examples_1.R` and `https://github.com/mfumagalli/BayesianMethods/blob/master/Exercise_1.R`

# Prior distributions

How can we decide which prior distribution is more appropriate in our study?

- They are derived from past information or personal opinions from experts.
- They are typically distributed as commonly used distribution families.
- They can be limited to to bear little information.
- ...

# Prior distributions

## Elicited priors

- define the collection of $\theta$ which are "possible",
- assign some probability to each one of these cases,
- make sure that they sum up to 1.

# Prior distributions

Elicited priors, a **discrete** case.

Figure: How many baby rabbits per litter?

# Prior distributions

Elicited priors, a **continuous** case.

Figure: Bumpass Hell, hot springs and fumaroles at Lassen Volcanic National Park, California.

# Prior distributions

Elicited priors, a **parametric** solution: $\theta$ belongs to a parametric distributional family $\pi(\theta|\nu)$.

Advantages:

- reduces the effort to the elicitee;
- overcomes the finite support problem;
- may lead to simplifications in the computation of the posterior.

Disadvantage:

- impossible to find a distribution that perfectly matches the elicitee's beliefs.

# Prior distributions

Elicited priors, a **parametric** solution

$$\pi(\theta) = \begin{cases} 0 & \text{for } \theta < 80 \text{ or } \theta > 110 \\ N(\mu, \sigma^2) & \text{for } 80 \leq \theta \leq 110 \end{cases}$$

with $\mu = 88$ and $\sigma^2 = 10$

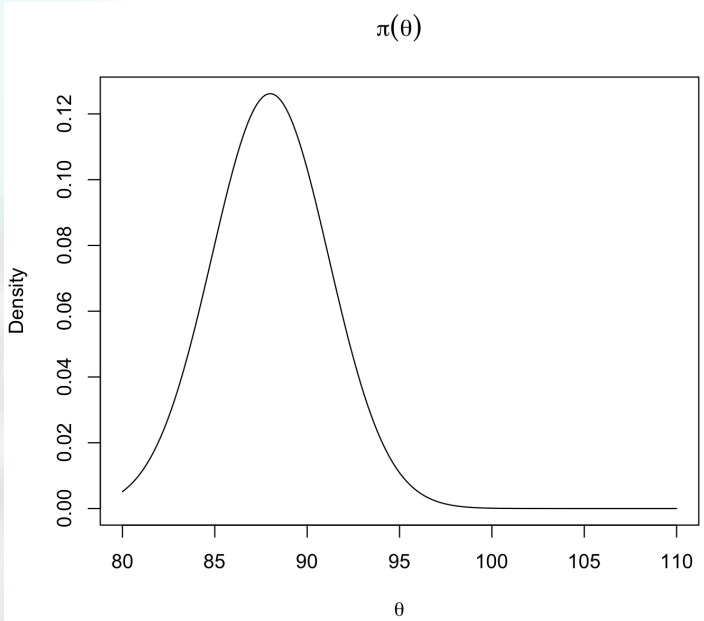



Figure: Elicited prior distribution of water temperature.

# Prior distributions

Elicited priors:

- focus on quantiles close to the middle of the distribution (e.g. the $50^{th}$, $25^{th}$ and $75^{th}$) rather than extreme quantiles (e.g. the $95^{th}$ and $5^{th}$),
- assess the symmetry,
- priors can be updated and reassessed as new information is available,
- useful for experimental design.

# Prior distributions

## Conjugate priors

$\pi(\theta)$ is member of a family which is *conjugate* with the likelihood $f(\vec{y}|\theta)$ so that the posterior distribution $p(\theta|\vec{y})$ belongs to the same distributional family as the prior.

# Prior distributions

## Conjugate priors

Example: $Y$ is the count of distinct elephant herds arriving at the pool in a day during the migration season.

Figure: Elephants drinking at the pool. What's the arrival rate for distinct herds?

# Prior distributions

## Conjugate priors

Poisson distribution is an appropriate model for $Y$ if:

1. $Y$ is the number of times an event occurs in an interval and it can take values any positive integer values including 0;
2. the occurrence of one event does not affect the probability that a second event will occur (i.e. events occur independently);
3. the rate at which events occur is constant (it cannot be higher in some intervals and lower in other intervals);
4. two events cannot occur at exactly the same instant;
5. the probability of an event in an interval is proportional to the length of the interval.

# Prior distributions

Conjugate priors

If $\theta$ is the event rate (rate parameter), then the probability of observing $y$ events in an interval is:

$$f(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!},\ y \in \{0, 1, 2, ...\},\ \theta > 0 \tag{4}$$

which is the probability mass function (pmf) for a Poisson distribution.

# Conjugate priors

## Likelihood: Poisson distribution

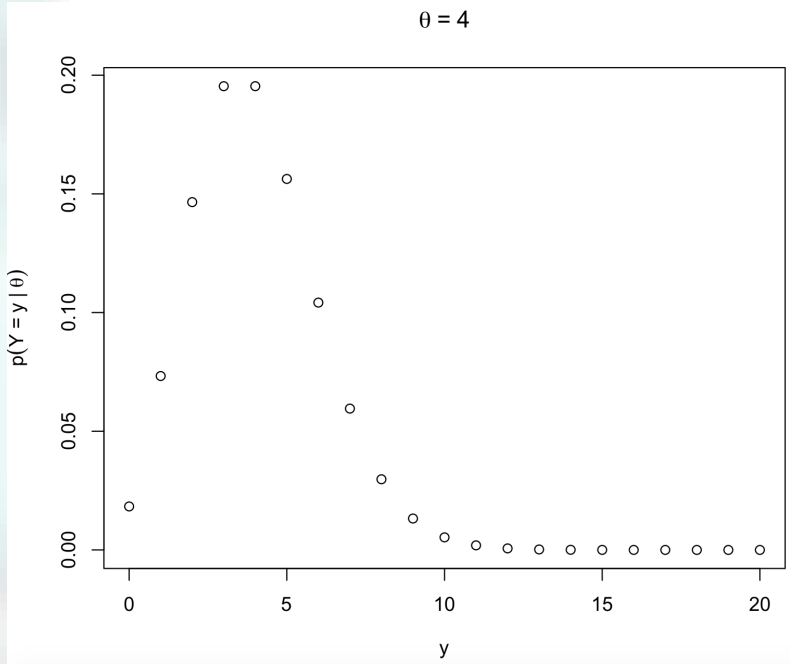



Figure: Poisson distribution for $\theta = 4$. This is the likelihood distribution for the number of herds per day with a rate of 4.

# Conjugate priors

Prior: Gamma distribution

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\ \theta > 0,\ \alpha > 0,\ \beta > 0 \qquad (5)$$

$$E(G(\alpha,\beta)) = \alpha\beta$$
$$Var(G(\alpha,\beta)) = \alpha\beta^2$$

# Conjugate priors

Prior: Gamma distribution

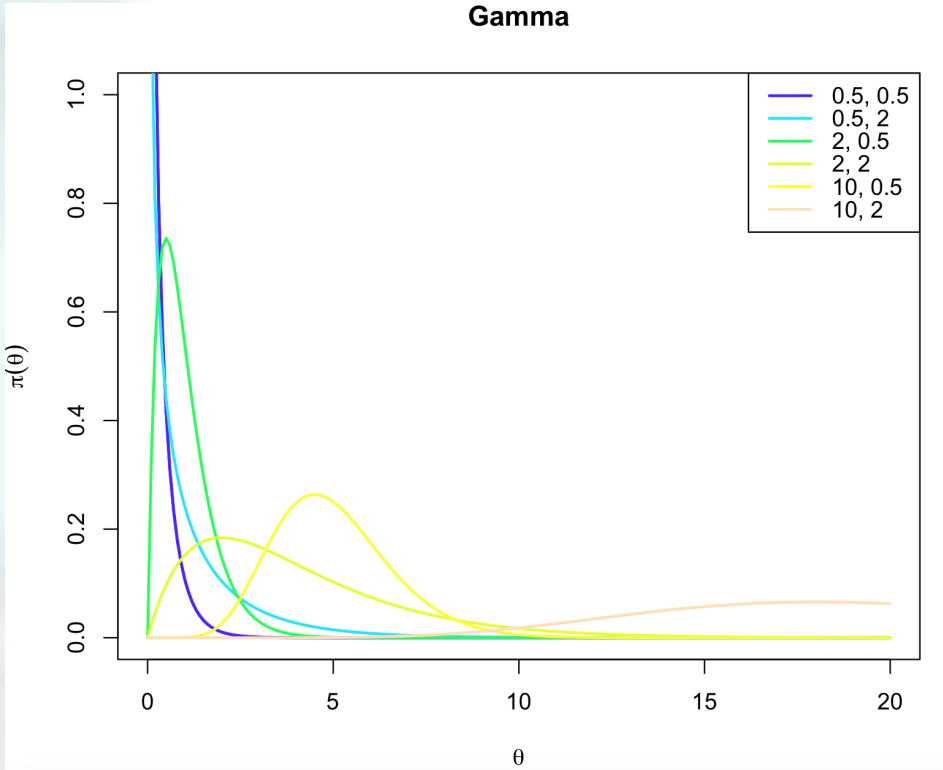



Figure: Gamma distribution for different values of shape and rate parameters.

# Conjugate priors

Posterior: (another) Gamma distribution

$$\begin{aligned} p(\theta|y) &\approx f(y|\theta)\pi(\theta) \\ &\approx (e^{-\theta}\theta^{y})(\theta^{\alpha-1}e^{-\theta/\beta}) \\ &= \theta^{y+\alpha-1}e^{-\theta(1+1/\beta)} \end{aligned}$$

$p(\theta|y) \sim G(\alpha', \beta')$ with $\alpha' = y + \alpha$ and $\beta' = (1 + 1/\beta)^{-1}$

Conjugate priors allow for posterior distributions to emerge without numerical integration.

# Conjugate priors

Example:

- we have some intuition that we expect to see 3 herds per day (prior)
- we observed 4 herds (data)

What is the posterior distribution of $\theta$, the average number of herds (Gamma-Poisson model)?

Exercise:
Calculate and plot the distribution in `R`.

# Hierarchical modelling

*Hyperpriors* define the density distribution of hyperparameters.

$$p(\vec{\theta}|\vec{y}) = \frac{\int f(\vec{y}|\vec{\theta})\pi(\vec{\theta}|\vec{\nu})h(\vec{\nu})d\vec{\nu}}{\int\int f(\vec{y}|\vec{\theta})\pi(\vec{\theta}|\vec{\nu})h(\vec{\nu})d\vec{\nu}d\vec{\theta}} \tag{6}$$

## Empirical Bayesian

*Estimated posterior* $p(\vec{\theta}|\vec{y},\hat{\vec{\nu}})$ by replacing $\vec{\nu}$ with an estimate $\hat{\vec{\nu}}$ obtained by maximising the marginal distribution $m(\vec{y}|\vec{\nu})$.

If $\vec{\nu} \sim h(\vec{\nu}|\vec{\lambda})$ with unknown parameters $\vec{\lambda}$, then we have a third-stage prior $g(\vec{\lambda})$.

# Non-informative priors

Can we use a Bayesian approach when no reliable prior information on $\vec{\theta}$ is available?

Yes, a *non-informative prior* distribution for $\vec{\theta}$ contains "no information" about $\vec{\theta}$ and all the information in the posterior will (mostly) arise from the data.

# Non-informative priors

Discrete case:

If $\vec{\Theta} = \{\theta_1, \theta_2, ..., \theta_n\}$, then

$$p(\theta_i) = \frac{1}{n},\ i = 1, 2, ..., n$$

with

$$\sum_{1}^{n} \frac{1}{n} = 1$$

# Non-informative priors

Continuous and bounded case:

If $\vec{\Theta} = [a, b]$ with $-\infty < a < b < +\infty$, then

$$p(\theta) = \frac{1}{b-a},\ a < \theta < b$$

# Non-informative priors

Continuous and unbounded case:

If $\vec{\Theta} = (-\infty, +\infty)$ then

$$p(\theta) = c, \text{ any } c > 0$$

is an *improper* distribution as

$$\int_{-\infty}^{+\infty} p(\theta) d\theta = +\infty$$

Bayesian inference is still possible under some circumstances.

# Non-informative priors

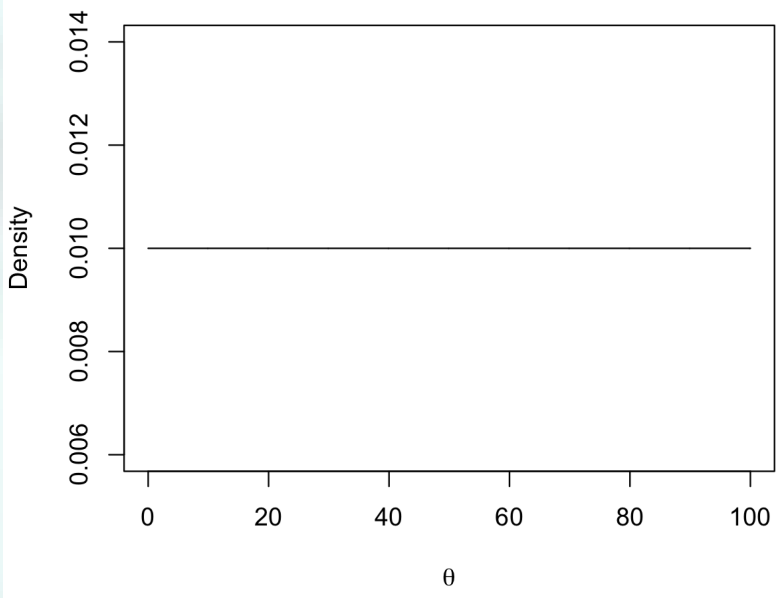



Figure: A uniform prior distribution for the arrival rate of elephant herds.

- Rule out scenarios that are impossible in real life.
- Lack a conjugate model (sampling methods are required).
- Non-informative priors are related to *reference* priors.

# Bayesian inference

The posterior distribution of model parameters can be difficult to interpret.

We want to **summarise** the information enclosed in these distribution.

# Point estimation

Example: Arrival rate with $y = 1$ and prior $\sim G(0.5, 1)$.

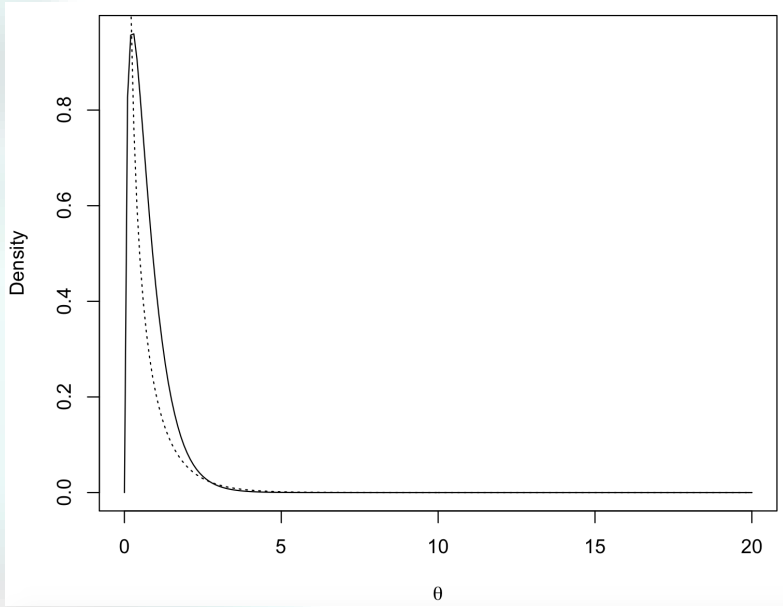



Figure: Posterior using a one-tailed gamma prior.

Question:
What is the (i) mean, (ii) mode and (iii) median of this resulting posterior distribution?

# Point estimation

- The mode is the easiest to calculate as we can work directly with the numerator.
- If the prior distribution is flat then the *posterior mode* will be equal to the maximum likelihood estimate.
- If the posterior distribution is symmetric, then the mean and the median are equivalent.
- For symmetric unimodal distributions, all these three features are equivalent.
- For asymmetric distributions, the median is often the best choice as it is less affected by outliers and it is an intermediate to the mode and the mean.

# Point estimation

If we want to obtain a measure of accuracy of a point estimate $\hat{\theta}(\vec{y})$, we can calculate the *posterior variance*:

$$E_{\theta|\vec{y}}(\theta - \hat{\theta})^2 \tag{7}$$

In the multivariate case the posterior mode is $\hat{\vec{\theta}}(\vec{y}) = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$.

# Credible intervals

A $100 \times (1 - \alpha)$ credible set for $\vec{\theta}$ is a subset $C$ of $\vec{\Theta}$ such that:

$$1 - \alpha \leq P(C|\vec{y}) = \int_C p(\vec{\theta}|\vec{y}) d\vec{\theta} \tag{8}$$

*"The probability that $\theta$ lies in $C$ given the observed data $y$ is at least $(1 - \alpha)$"*
e.g. $\alpha = 0.05$

# Credible intervals

In continuous settings we can calculate the *highest posterior density*, or **HPD**, credible set, defined as:

$$C = \{\theta \in \Theta : p(\theta|y) \geq k(\alpha)\} \tag{9}$$

where $k(\alpha)$ is the largest constant satisfying $P(C|y) \geq (1 - \alpha)$.

Example: $p(\theta|\vec{y}) \sim G(2, 1)$ and $k(\alpha) = 0.1$.

# Credible intervals

How can we summarise our results?

- the posterior mean
- several posterior percentiles (e.g. 0.025, 0.25, 0.50, 0.75, 0.975)
- a credible interval
- posterior probabilities $p(\theta > c|y)$ where $c$ is a notable point (e.g. 0, 1, depending on the problem)
- a plot of the distribution to check whether it is unimodal, multimodal, skewed, ...

# Hypothesis testing

In the **frequentist** approach,

1. one formulates a null hypothesis $H_0$ and an alternative hypothesis $H_a$,
2. an appropriate test statistic is chosen $T(\vec{Y})$,
3. one computes the *observed significance*, or *p-value*, of the test as the chance that $T(\vec{Y})$ is "more extreme" that $T(\vec{y}_{obs})$, where the "extremeness" is towards the alternate hypothesis,
4. if the p-value is less than some threshold, typically in the form of a pre-specified Type I error rate, $H_0$ is rejected, otherwise it is not.

# Hypothesis testing

Limits of frequentist approach:

1. only when two hypotheses are nested (e.g. $H_0$ is a simplification of $H_a$ and involves setting one parameter of $H_a$ to some known constant value)
2. evidence *against* the null hypothesis (e.g. a large p-value does not mean that the two models are equivalent, but only that we lack evidence of the contrary; we don't "accept the null hypothesis" but "fail to reject it")
3. no direct interpretation as weight of evidence (but only as a long-term probability; p-values are not the probability that $H_0$ is true!)

# Hypothesis testing

In the **Bayesian** approach,

1. one can test as many models as desired, $M_i, i = 1, ..., m$,
2. one calculates the posterior probability that each model is correct
3. one compares each pair of posterior probabilities.

# Hypothesis testing

Suppose we have two models $M_1$
and $M_2$ for data $Y$ and the two models have parameters $\vec{\theta}_1$ and $\vec{\theta}_2$.

With prior densities $\pi_i(\theta_i)$ and $i = 1, 2$, the marginal distributions of $Y$ are:

$$p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i \tag{10}$$

We can calculate the posterior probabilities $P(M_1|y)$ and $P(M_2|y) = 1 - P(M_1|y)$ for the two models.

# Bayes factors

A Bayes factor (BF) is used to summarise these results, and it is equal to the ration of posterior odds of $M_1$ to the prior odds of $M_1$:

$$BF = \frac{P(M_1|y)/P(M_2|y)}{P(M_1)/P(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} \tag{11}$$

If the two models are *a priori* equally probable then:

$$BF = p(M_1|y)/p(M_2|y) \tag{12}$$

which is the posterior odds of $M_1$.

# Bayes factors

## Interpretation

BF captures the change in the odds in favour of model 1 (vs. 2) as we move from the prior to the posterior.

| **BF** | **Strength of evidence** |
|---|---|
| 1 to 3 | not worth more than a bare mention |
| 3 to 20 | positive |
| 20 to 150 | strong |
| $> 150$ | very strong |

# Bayesian inference

We now have sequenced our bears' genomes and, using the method in Exercise 1, assigned each individual genotype.

What is the frequency of a certain allele at the **population** level?

# CASE STUDY (2)

Estimating population variation from genomic data.

All scripts used in these examples and for the exercise are available at `https://github.com/mfumagalli/BayesianMethods/blob/master/Examples_2.R` and `https://github.com/mfumagalli/BayesianMethods/blob/master/Exercise_2.pdf`

# Bayesian computation

The calculation of posterior distributions often involves the evaluation of complex high-dimensional integrals.

When a conjugate prior is not available or appropriate we can evaluate the posterior distribution with:

1. asymptotic methods for approximating the posterior density;
2. numerical integration.

# Asymptotic methods

## Bayesian Central Limit Theorem

When there are many data points $p(\theta|x)$ will be approximately normally distributed.

For large data points, the posterior can be approximated by a normal distribution with mean equal to the posterior mode and (co)variance (matrix) equal to minus the inverse of the second derivative matrix of the log posterior evaluated at the mode.

# Asymptotic methods

Example:
Recalling the beta-binomial model with flat prior,
$p(\theta|x) \approx \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: ...
2. take the derivative of $l(\theta)$ and set it to zero, obtaining $\hat{\theta}^\pi = ...$
3. take the second derivative evaluated at $\hat{\theta}$, ...
4. take the minus inverse, ...
5. $p(\theta|x) \sim ...$

# Asymptotic methods

Example:
Recalling the beta-binomial model with flat prior,
$p(\theta|x) \approx \theta^x(1-\theta)^{n-x}$.

The approximation is given by:

1. take the log: $l(\theta) = x \log \theta + (n-x) \log(1-\theta)$
2. take the derivative of $l(\theta)$ and set it to zero, obtaining $\hat{\theta}^\pi = \frac{x}{n}$
3. take the second derivative evaluated at $\hat{\theta}$, $-\frac{n}{\hat{\theta}} - \frac{n}{1-\hat{\theta}}$
4. take the minus inverse, $\frac{\hat{\theta}(1-\hat{\theta})}{n}$
5. $p(\theta|x) \sim N(\hat{\theta}^\pi, \frac{\hat{\theta}(1-\hat{\theta})}{n})$

# Asymptotic methods

*Model approximations* or *first order approximations*: the estimate $\theta$ by the mode and the error goes to 0 at a rate proportional to $1/n$.

The estimates of moments and quantiles may be poor if the posterior differ from normality.

The *Laplace's Method* provides a second order approximation to the posterior mean, with an error that decreases at a rate $1/n^2$.

# Asymptotic methods

Advantages:

- they replace numerical integration with numerical differentiation,
- they are deterministic (without elements of stochasticity).
- they reduce the computational complexity if any study of robustness (how sensitive are our conclusions to changes in the prior/likelihood?).

# Asymptotic methods

Disadvantages:

- they require that the posterior is unimodal,
- the require that the size of the data is large (how large is "large enough"?),
- for high high-dimensional parameters the calculation of Hessian matrices (second derivatives) are hard.

# Non-iterative Monte Carlo methods

If $\vec{\theta} \sim h(\vec{\theta})$ with $h(\vec{\theta})$ being a posterior distribution, we want to estimate $\gamma$ the posterior mean of $c(\vec{\theta})$, where
$\gamma \equiv E[c(\vec{\theta})] = \int c(\vec{\theta}) h(\vec{\theta}) d\vec{\theta}$.
If $\vec{\theta}_1, \vec{\theta}_2, ..., \vec{\theta}_N$ are independent and identically distributed (iid) as $h(\vec{\theta})$, then:

$$\hat{\gamma} = \frac{1}{N} \sum_{i=1}^{N} c(\vec{\theta}_i) \tag{13}$$

which converges to $E[c(\vec{\theta})]$ with probability 1 as $N \to \infty$.

The computation of **posterior expectations** requires only a sample of size $N$ from the posterior distribution.

# Non-iterative Monte Carlo methods

In the univariate case, a histogram of the sampled $\theta_i$s estimates the posterior itself.

An estimate of $p \equiv Pa < c(\theta) < b$ is given by

$$\hat{p} = \frac{\text{number of } c(\theta_i) \in (a, b)}{N} \tag{14}$$

In contrast to asymptotic methods, accuracy improves with $N$, the Monte Carlo sample size (which we can choose and have control upon) rather than $n$ the size of the data set (which can may not be able to control).

# Non-iterative Monte Carlo methods

What happens if we can't directly sample from this distribution?

There are methods for **indirect** sampling of the posterior distribution: (i) importance sampling, (ii) rejection sampling, (iii) weighted bootstrap.

# Non-iterative Monte Carlo methods

Rejection sampling:

If we identify an *envelope function* $g(\vec{\theta})$ and a constant $M > 0$ such that $L(\vec{\theta})\pi(\vec{\theta}) < Mg(\vec{\theta})$ for all $\vec{\theta}$, then:

1. Generate $\theta_i \sim g(\theta)$,
2. Generate $U \sim Uniform(0, 1)$,
3. If $MUg(\theta_i) < L(\theta_i)\pi(\theta_i)$ accept $\theta_i$ otherwise reject $\theta_i$.

If we repeat this procedure until $N$ samples are obtained, the members of this sample will be random variables from $h(\theta)$.

It is hard to sample from the true posterior but it is easier to sample from the envelope function.

# Markov chain Monte Carlo methods

- All previous methods are non-iterative as they draw a sample of fixed size $N$.
- There is no notion of "convergence" but rather we require $N$ to be sufficiently large.
- For many problems with high dimensionality it may be difficult to find an importance sampling density or an envelope function.

In these cases it is now standard practice to use *Markov chain Monte Carlo* (MCMC) methods.

# Markov chain Monte Carlo methods

MCMC can sequentially sample parameter values from a Markov chain.

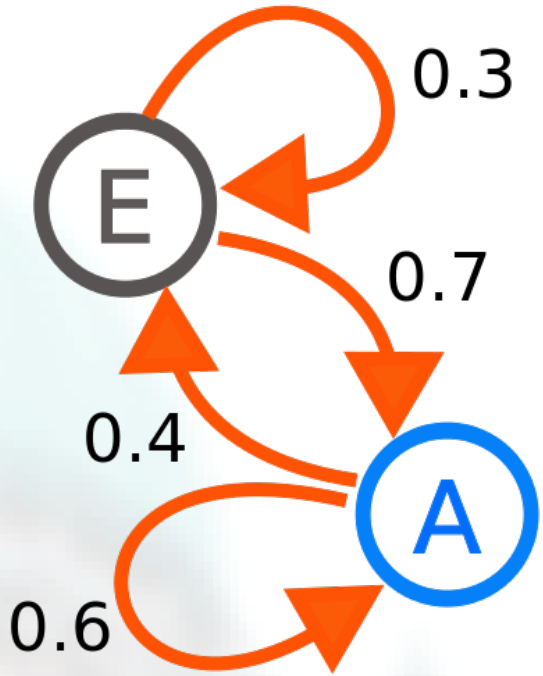



Figure: A diagram of a two-state Markov process, with the states labelled E and A. Each number represents the probability of the Markov process changing from one state to another state, with the direction indicated by the arrow. Image: Wikipedia.

# Markov process and chain

1. A mathematical object following a stochastic (or random) process, typically defined as a collection of random variables.
2. The next value of the process depends only on the current value, but it is independent of the previous values.
3. A Markov chain is a Markov process that has a particular type of state space, which dictates the possible values that a stochastic process can take.

# Markov chain Monte Carlo

## Stationary distribution

The probability distribution to which the process converges for large values of steps, or iterations.

The stationary distribution of an MCMC is the desired posterior distribution.
An assessment of *convergence* of the Markov chain to its stationary distribution is required.

The majority of Bayesian MCMC computation is based on two algorithms: the *Gibbs sampler* and the *Metropolis-Hastings (M-H)* algorithm.

# Gibbs sampler

Suppose our model has $k$ parameters $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_k)$.

We assume that we can sample from the full conditional distributions.

The collection of full conditional distributions uniquely determines the joint posterior distribution $p(\vec{\theta}, \vec{y})$ and therefore all marginal posterior distributions $p(\theta_i, \vec{y})$, for $i = 1, ..., k$.
...

# Gibbs sampler

...

Given an arbitrary set of starting $\{\theta_2^{(0)}, ..., \theta_2^{(k)}\}$, the algorithm, for $(t = 1, ..., T)$, is:

- Draw $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, \vec{y})$
- Draw $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_k^{(t-1)}, \vec{y})$
- ...
- Draw $\theta_k^{(t)}$ from $p(\theta_k|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{k-1}^{(t)}, \vec{y})$

$(\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_k^{(t)})$ converges to a draw from the true joint posterior distribution $p(\theta_1, \theta_2, ..., \theta_k|\vec{y})$.

For $t > t_0$ then $\{\vec{\theta^{(t)}}, t = t_0 + 1, ..., T\}$ is a correlated sample from the true posterior.

# Gibbs sampler

1. The parameter space must be fully *connected*, without "holes".
2. When $\theta$ and $\nu$ are highly correlated the chain will have a "slow mixing".
3. To ensure that all the full conditional distributions are available, the prior distribution of each parameter can be chosen to be conjugate with the corresponding likelihood.

# Gibbs sampler

A histogram of $\{\theta_i^{(t)}, t = t_0 + 1, ..., T\}$ provides an estimator of the marginal posterior distribution for $\theta_i$.

The posterior mean can be estimated as the posterior mean:

$$\hat{E}(\theta_i|\vec{y}) = \frac{1}{T - t_0} \sum_{t=t_0+1}^{T} \theta_i^{(t)} \tag{15}$$

The time $0 <= t <= t_0$ is called the $burn - in$ period.

# Metropolis algorithm

Given:

- $p(\vec{\theta}|\vec{y}) \propto h(\vec{\theta}) \equiv f(\vec{y}|\vec{\theta})\pi(\vec{\theta})$
- a *candidate*, or *proposal*, symmetric density $q(\vec{\theta}^*|\vec{\theta}^{(t-1)})$ which satisfies $q(\vec{\theta}^*|\vec{\theta}^{(t-1)}) = q(\vec{\theta}^{(t-1)}|\vec{\theta}^*)$,
- a starting value $\vec{\theta^{(0)}}$ at iteration $t = 0$,

for $(t = 1, ..., T)$ the algorithm repeats:

1. Draw $\vec{\theta}^* = q(\cdot|\vec{\theta}^{(t-1)})$
2. Calculate $r = h(\vec{\theta}^*)/h(\vec{\theta}^{(t-1)}) =$
3. If $r \geq 1$, set $\vec{\theta}^{(t)} = \vec{\theta}^*$, otherwise set $\vec{\theta}^{(t)} = \vec{\theta}^*$ with probability $r$ or set $\vec{\theta}^{(t)} = \vec{\theta}^{(t-1)}$ with probability $1 - r$.

$\vec{\theta}^{(t)}$ converges in distribution to a draw from the true posterior density $p(\vec{\theta}|\vec{y})$.

# Metropolis algorithm

A usual candidate density is:

$$q(\vec{\theta}^*|\vec{\theta}^{(t-1)}) = N(\vec{\theta}^*|\vec{\theta}^{(t-1)}, \tilde{\Sigma}) \tag{16}$$

*random walk Metropolis*: symmetric and "self-correcting" distribution.

$\tilde{\Sigma}$, the posterior variance, can be empirically estimated from a preliminary run.

# Metropolis-Hastings algorithm

When $q(\vec{\theta} * | \vec{\theta}^{(t-1)}) \neq q(\vec{\theta}^{(t-1)} | \vec{\theta} *)$ the acceptance rate $r$ is:

$$r = \frac{h(\vec{\theta}^*) q(\vec{\theta}^{(t-1)} | \vec{\theta}^*)}{h(\vec{\theta}^{(t-1)}) q(\vec{\theta}^* | \vec{\theta}^{(t-1)})} \tag{17}$$

A draw $\vec{\theta}(t)$ converges in distribution to a draw from the true posterior density as $t \to \infty$.

# Hastings independence chain

If we set $q(\vec{\theta}^*|\vec{\theta}^{(t-1)}) = q(\vec{\theta}^*)$ then the proposal ignores the current value of the variable.

The acceptance rate is:

$$r = \frac{h(\vec{\theta}^*)/q(\vec{\theta}^*)}{h(\vec{\theta}^{(t-1)})/q(\vec{\theta}^{(t-1)})} \tag{18}$$

# MCMC algorithms

- *Langevin-Hastings* algorithm introduces a systematic drift in the candidate density.
- *Slice sampler* algorithm uses auxiliary variables to expand the parameter space.
- *Hybrid* forms combined multiple algorithm in a single problem.
- *Adaptive* algorithms use the early output from a chain to refine the sampling as it progresses.

# Convergence

Diagnostic strategy:

- run parallel chains with starting points from a wide distribution;
- visually inspect these chains;
- for each graph calculate the scale reduction factor;
- investigate crosscorrelations among parameters.

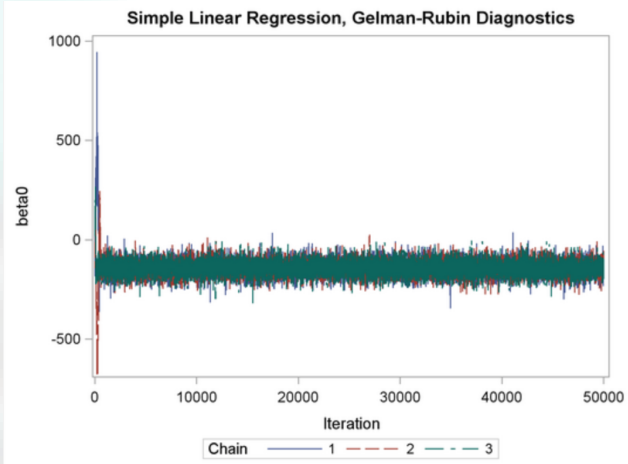



Figure: Three chains mixing for increasing $t$.

# Software

- `WinBUGS`
- `OpenBUGS`
- `BRugs`
- `JAGS`
- `rjags`
- ...

# Approximate Bayesian Computation

The posterior probability

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \tag{19}$$

which can be difficult as the marginal likelihood

$$p(x) = \int p(x|\theta)\pi(\theta)d\theta \tag{20}$$

can involve a high dimensional integral.

# Approximate Bayesian Computation

- If the likelihood can be evaluated up to a normalising constant, Monte Carlo methods can be used to sample from the posterior.

- If the likelihood function becomes difficult to define and compute, it is easier to *simulate* data samples from the model given the value of a parameter.

# Rejection algorithm

If data are **discrete** and of low dimensionality, given observation $y$, repeat the following until $N$ points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject $\theta_i$ if $x_i \neq y$

These are sampled from $p(\theta|x)$.

# Rejection algorithm

Example (elephants):

- We observe 4 herds arriving.
- The likelihood is Poisson-distributed and the prior is Gamma-shaped $G(3, 1)$.
- The posterior distribution is Gamma distributed with shape parameter $3 + 4 = 7$ and scale 0.5.

Let's assume that we can't evaluate the likelihood but we know how to *simulate* $y$ given a certain value of our parameter $\theta$.

Exercise:
Calculate the posterior distribution of $\theta$.

# Rejection algorithm

If data are **discrete** and of low dimensionality, given observation $y$, repeat the following until $N$ points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject $\theta_i$ if $\rho(x_i, y) > \epsilon$

where $\rho(\cdot)$ is a function measuring the distance between simulated and observed points.

# Rejection algorithm

Example (water temperature):

- $\theta$ is continuous with prior distribution $U(80, 110)$.
- we have a single observation $y = 91.3514$.
- as $\rho(\cdot)$ we use the Euclidean distance:

$$\rho(x_i, y) = \sqrt{(x_i - y)^2} \tag{21}$$

We can't evaluate the likelihood function but we can simulate observations that are distributed according to it.

Exercise:
Calculate the posterior distribution of $\theta$.

# Rejection algorithm

Alternatively, $\epsilon$ is the proportion of accepted simulations (ranked by distance with observations).

In this case one sets the number of simulations to be performed (not the number of accepted simulations).

# Rejection algorithm

Exercise (water temperature):

1. $Y = \{91.34, 89.21, 88.98\}$
2. $\theta$ has prior $N(\mu = 90, \sigma^2 = 20)$ for $80 \geq \theta \leq 110$
3. the simulating function is `simulate <- function(param) rnorm(n=1, mean=param, sd=sqrt(10))`
4. the distance function is $\rho(x_i, Y) = \frac{\sum_{j \in Y} \sqrt{(x_i - j)^2}}{|Y|}$
5. $N = 10,000$ and $\epsilon = 0.05$

Tasks:

1. plot the sampled prior distribution
2. plot the distribution of ranked distances with indication of 5% threshold
3. plot the posterior distribution
4. calculate notable quantiles and HPD 95% using `library(coda); HPDinterval(as.mcmc(x), prob=0.95))`

# Rejection algorithm

If data are of **high dimensionality**, given observation $y$, repeat the following until $N$ points have been accepted:

1. Draw $\theta_i \sim \pi(\theta)$
2. Simulate $x_i \sim p(x|\theta_i)$
3. Reject $\theta_i$ if $\rho(S(x_i), S(y)) > \epsilon$

with $S(y)$ being summary statistics.

# Rejection algorithm

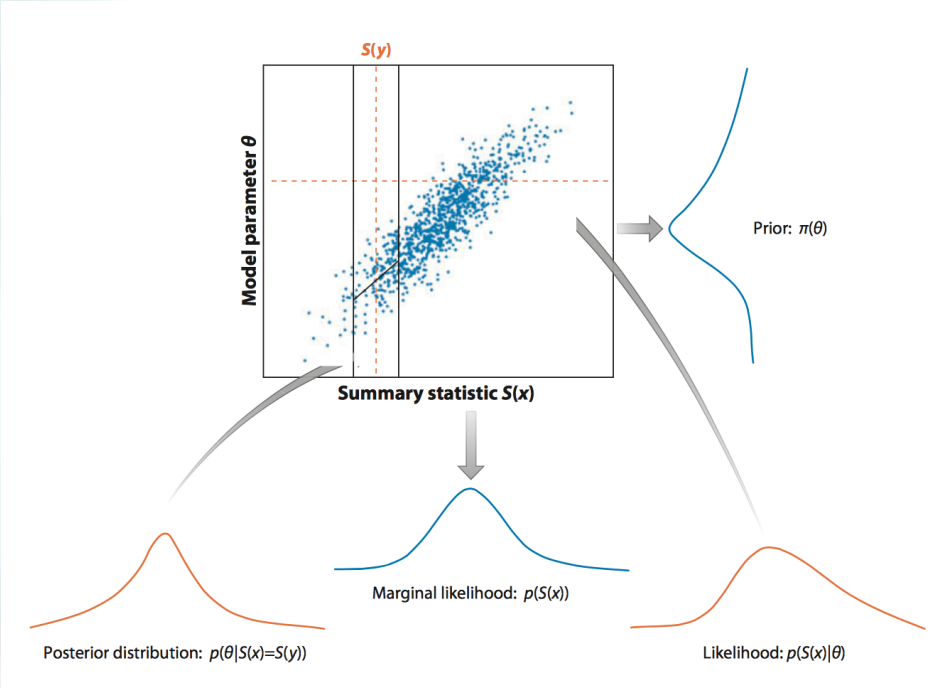



Figure: From Beaumont 2010 Annu Rev Ecol Evol Syst. Rejection- and regression-based approximate Bayesian computation (ABC).

# Summary statistics

- The choice of summary statistics is a mapping from a high dimension to a low dimension.
- Some information is lost, but with enough summary statistics much of the information is kept.
- The aim for the summary statistics is to satisfy the Bayes' sufficiency:

$$p(\theta|x) = p(\theta|S(x)) \tag{22}$$

Issues?
Solutions:

1. use a wider acceptance tolerance
2. perform a better sampling from the prior

# Regression-based ABC

1. Given observation $y$ repeat the following until $M$ points have been generated: A. Draw $\theta_i \sim \pi(\theta)$; B. Simulate $x_i \sim p(x|\theta_i)$
2. Calculate $S_j(x)$ for all $j$ and $k_j$
3. $\rho(S(x), S(y)) : \sqrt{\sum_{j=1}^{s} (\frac{S_j(x)}{k_j} - \frac{S_j(y)}{k_j})^2}$
4. Choose $\epsilon$ such that the proportion of accepted points $P_\epsilon = \frac{N}{M}$
5. Weight the simulated points $S(x_i)$ using $K_\epsilon(\rho(S(x_i), S(y)))$

$$K_\epsilon(t) = \begin{cases} \epsilon^{-1}(1 - (t/\epsilon)^2) & \text{for } t \leq \epsilon \\ 0 & \text{for } t > \epsilon \end{cases}$$

6. Apply weighted linear regression to the $N$ points that have nonzero weight to obtain an estimate of $\hat{E}(\theta|S(x))$
7. Adjust $\theta_i^* = \theta_i - \hat{E}(\theta|S(x)) + \hat{E}(\theta|S(y))$
8. The $\theta_i^*$ with weights $K_\epsilon(\rho(S(x_i), S(y)))$ are random draws from an approximation of $p(\theta|y)$.

# MCMC-ABC

Initialise by sampling $\theta^{(0)} \sim \pi(\theta)$.
At iteration $t \geq 1$,

1. Simulate $\theta' \sim K(\theta|\theta^{(t-1)})$ where $K(\cdot)$ is a proposal distribution that depends on the current value of $\theta$
2. Simulate $x \sim p(x|\theta')$.
3. If $\rho(S(x), S(y)) < \epsilon$ (rejection step),
   - $u \sim U(0,1)$,
   - if $u \leq \pi(\theta')/\pi(\theta^{(t-1)}) \times K(\theta^{(t-1)}|\theta')/K(\theta'|\theta^{(t-1)})$, update $\theta(t) = \theta'$;
   - otherwise $\theta(t) = \theta^{(t-1)}$;
4. otherwise $\theta(t) = \theta^{(t-1)}$.

# Model assessment

## Model choice

Given a series of model $\mu_1, \mu_2, ..., \mu_N$ with prior probabilities $\sum_i \pi(\mu_i) = 1$, it is of interest to calculate Bayes factors between two models $i$ and $j$:

$$\frac{p(\mu_i|x)}{p(\mu_j|x)} \div \frac{p(\mu_i)}{p(\mu_j)} \tag{23}$$

# Choice of summary statistics

The more the merrier?


Figure: Choosing summary statistics: the issue of pulling a short blanket.

# Choice of summary statistics

1. One could calculate the ratio of posterior density with or without a particular summary statistic. Departures greater than a threshold are suggestive that the excluded summary statistic is important.
2. Different summary statistics can be weighted differently according to their correlation with some model parameters.
3. The number of summary statistics can also be reduced via multivariate dimensional scaling summary statistics should be scaled in order to have equal mean and variance, if normally distributed.
4. Even if there is no need of a strong theory relating summary statistics to model parameters, it is suitable to have some expectations.

# Model validation

Validation is the assessment of goodness-of-fit of the model and comparing alternative models, to distinguish errors due to the approximation from errors caused by the choice of the model.

1. The distributions of simulated summary statistics are visualised and compared to the corresponding target statistic. If the target is outside, then this could be a problem in the model.
2. The observations are compared with the posterior predictive distribution. This can be done by simulating data with parameters drawn randomly from the current posterior distribution.

# Applications of ABC in ecology and evolution

Population genetics, agent-based models, protein interaction networks, speciation rates under a neutral ecological model, extinction rates from phylogenetic data, epidemiology, ...

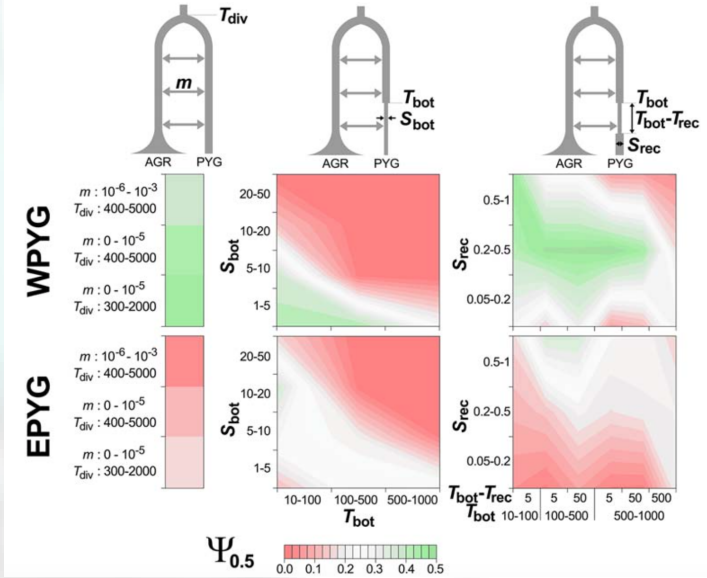



Figure: Patin et al. (2009). Different models simulating the demographic regime of the African groups and the mean proportion of small distances ($\Omega_{0.5}$) obtained in comparisons with simulated statistics.

# ABC or not ABC?

- When a likelihood function is known and can be efficiently evaluated, then there is not advantage to use ABC.
- When the likelihood function is known but difficult to evaluate in practise, the ABC is a valid option.
- Many scenarios in evolutionary biology or ecology can be generated by simulations.
- ABC can be useful for initial exploratory phase.

Please do not redistribute. These slides may contain material protected by copyright.