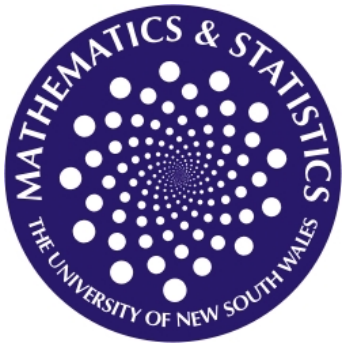


Lecture 1: Introduction to Bayesian inference

sem 2, 2018

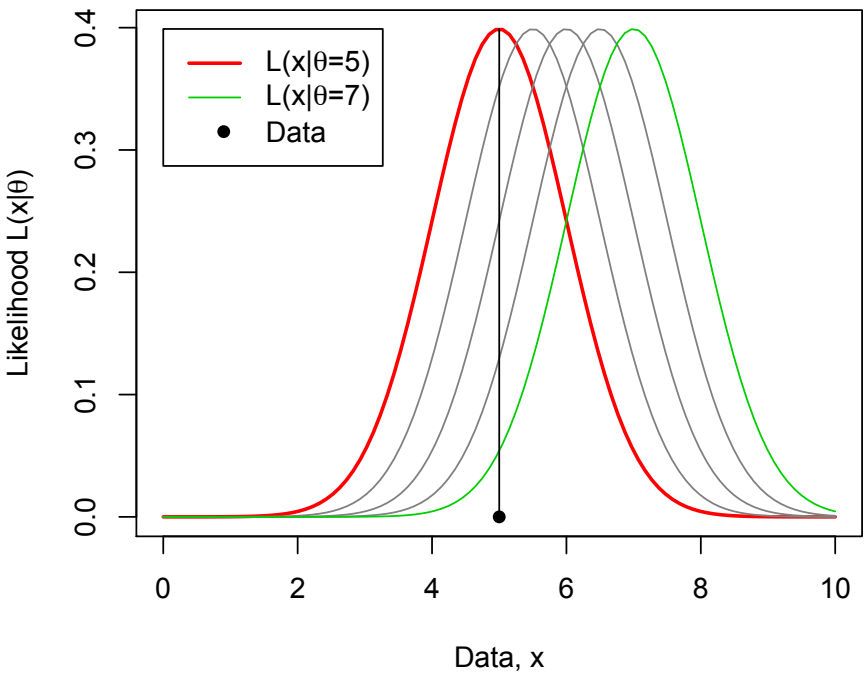
Outline

1. What is Bayesian statistics?
    - Bayes formula, prior and posterior probabilities
2. Monte Carlo integration basics
    - What is this, and why do we need it?

Outline

1. What is Bayesian statistics?
    - Bayes formula, prior and posterior probabilities
2. Monte Carlo integration basics
    - What is this, and why do we need it?

Classical statistics recap: Maximum Likelihood



Likelihood function: $L(x \mid \theta)$

- x is the random variable (not θ!)
- θ is fixed
- ** θ is considered as a constant **

MLE: $\hat{\theta} = \arg_\theta \max L(x \mid \theta)$

- Change θ so that x is given largest likelihood
- Treat as a function of θ (but it isn't)

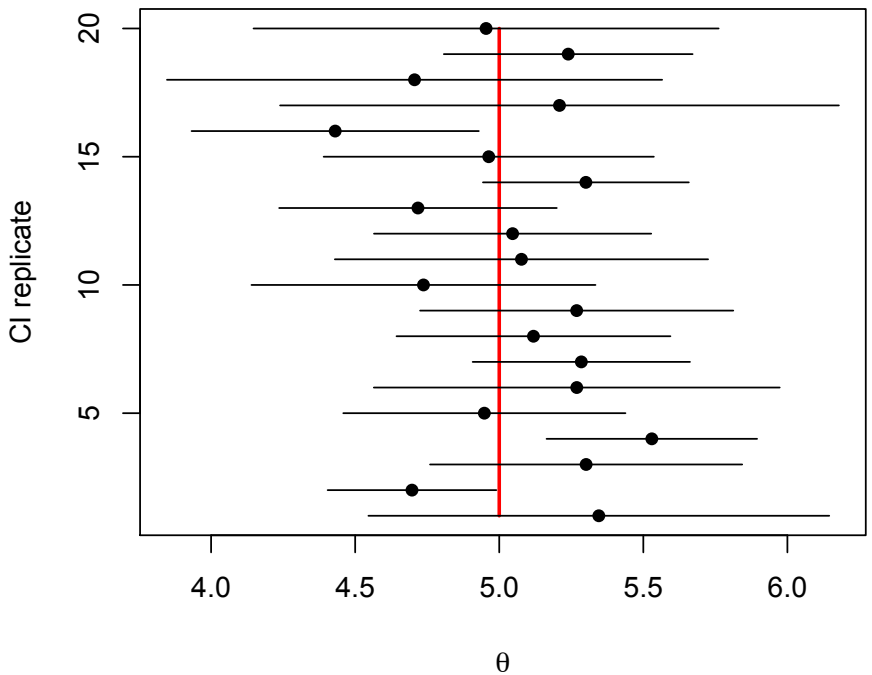
Model Prediction

- Distribution of future data y obtained via $Y \sim \pi(y \mid \hat{\theta})$

MLE properties

- In many circumstances $\hat{\theta} \sim \mathrm{N}(\theta, \sigma_0^2)$
- Allows construction of confidence intervals for $\hat{\theta}$ and hypothesis tests

Classical Statistics: Confidence Intervals



100(1 − α)% CI for pop. mean (θ):

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

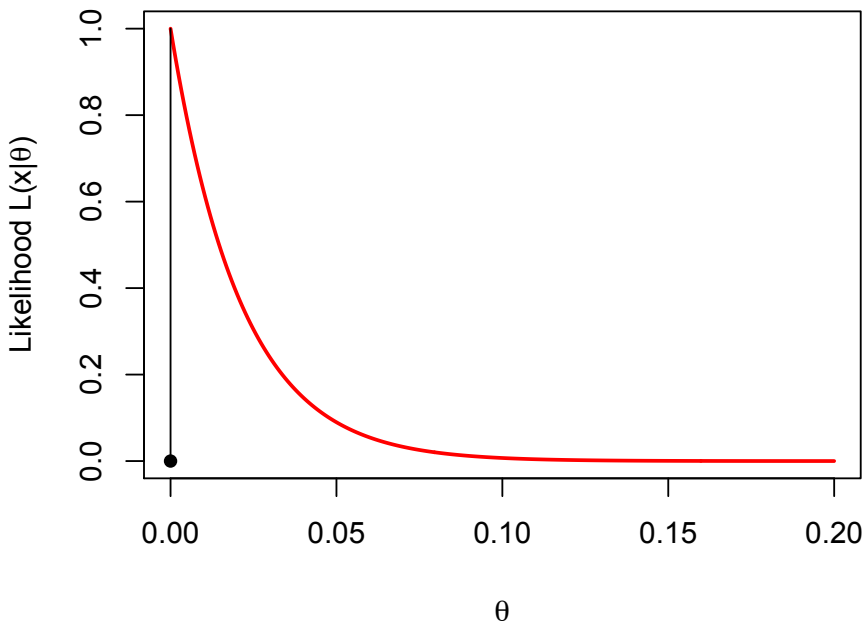
Interpretation:

- Very awkward!
- Would like to say e.g.:
  "there is a 95% probability that θ is in your (single) confidence interval"
- But we can't

Our (single) CI either contains θ (Pr = 1) or it doesn't (Pr = 0).

Actual interpretation: in the long run over all experiments/datasets 95% of them will contain the true parameter

Classical Statistics: Non-regular likelihoods



Hospital Example:

θ = $\mathbb{P}$(death of baby from operation)

r = Number of previous deaths

n = Number of previous operations

Observed data: $r = 0$, $n = 47$

$$r \sim \text{Bin}(n, \theta)$$

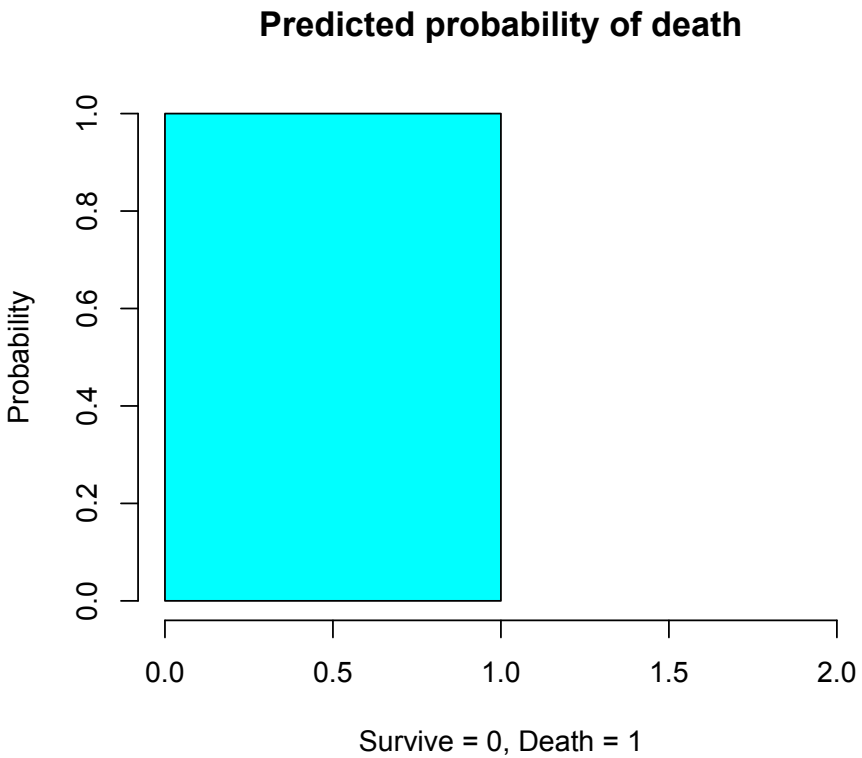
The likelihood:

- $\hat{\theta} = 0$ is still well defined
- But $\hat{\theta} \sim \text{N}(\theta, \sigma_0^2)$ fails
- How to get a CI for θ?
- How to do hypothesis test?

Standard likelihood asymptotics can fail even in simple situations

Need non-standard likelihood asymptotics – can be tricky.

Classical Statistics: Predictions



Hospital Example:

Predictive distribution is

$$\begin{aligned} Y &\sim \pi(y \mid \hat{\theta}) \\ &= \text{Ber}(\hat{\theta}) \end{aligned}$$

with $\hat{\theta} = 0$.

Predictive distribution will produce:

- 100% prediction for survival
- 0% prediction for death

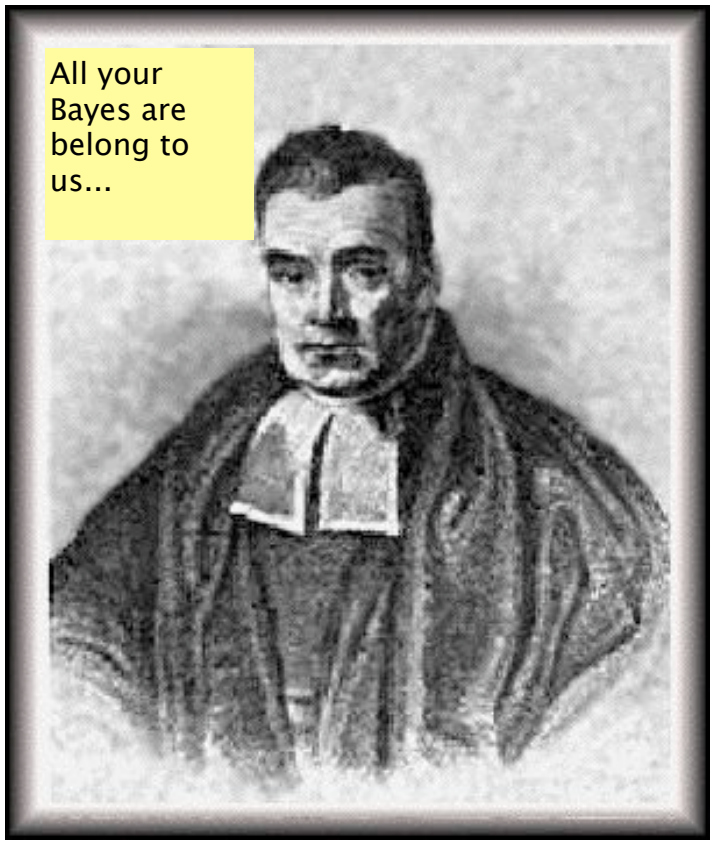
This is completely unrealistic

Performing model predictions using a fixed point estimate, $\hat{\theta}$, can cause problems

Does not take uncertainty about θ into account

Can use profile likelihoods – but is there an alternative approach?

What is Bayesian Statistics?



T. Bayes.

Wikipedia page:

`http://en.wikipedia.org/wiki/Thomas_Bayes`

Thomas Bayes (1701—1761)

- English mathematician and Presbyterian minister
- 1763 publication:
  An Essay towards solving a Problem in the Doctrine of Chances

  Contained a special case of what is now know as Bayes' Theorem:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

- Essentially a conditional probability statement

What is Bayesian Statistics?

Where does Bayes' formula come from?

$$\left.\begin{aligned}\mathbb{P}(A \text{ and } B) &= \mathbb{P}(A \mid B)\mathbb{P}(B)\\ \mathbb{P}(A \text{ and } B) &= \mathbb{P}(B \mid A)\mathbb{P}(A)\end{aligned}\right\} \Rightarrow \text{ Equate}$$

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

There are other derivations as well.

Classical:

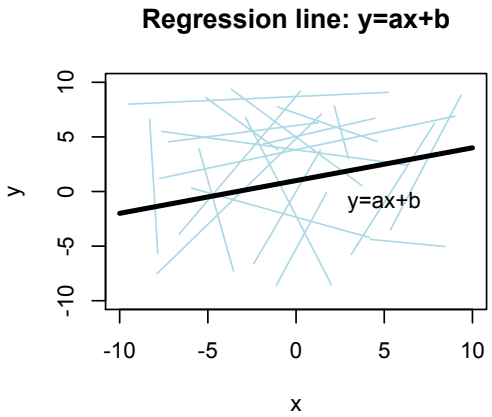
Parameter θ (though unknown) is a constant.

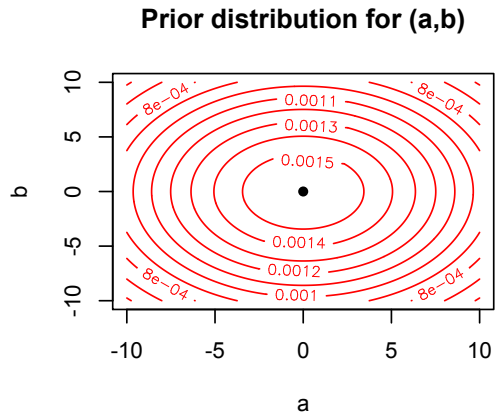
Bayesian:

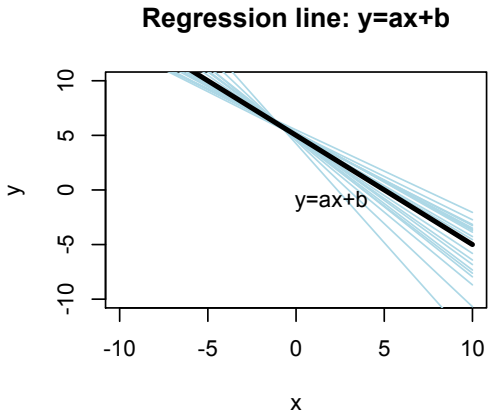
Parameter θ (though unknown) is random (i.e. a random variable).

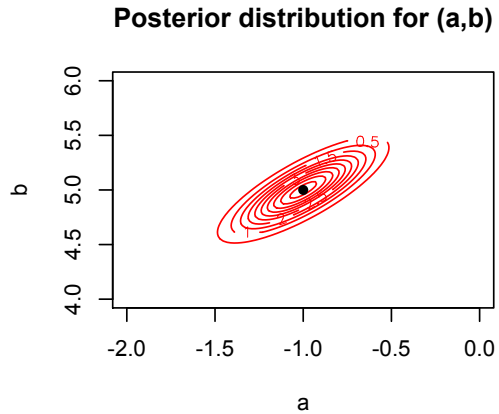
Allows easy descriptive inference in terms of direct probabilities.

What's the idea behind Bayesian inference?









Fitting a linear regression:
y=ax+b

Start with prior guess of (a,b) ** before seeing data ** in distribution form e.g.

$$\begin{pmatrix} a \\ b \end{pmatrix} = \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}\right)$$

After fitting the model to data prior guess is updated to posterior belief about (a,b)

Still in distributional form

$$\begin{pmatrix} a \\ b \end{pmatrix} = \mathrm{N}\left(\hat{\mu}, \hat{\Sigma}\right)$$

Key point:

Parameters are represented by distributions, not point estimates

How does Bayes Theorem help us do this?

If we set:

$A = x$ = observed data

$B = \theta$ = unknown model parameters

Then:

$$\begin{aligned} \mathbb{P}(\theta \,|\, x) &= \frac{\mathbb{P}(x \,|\, \theta)\mathbb{P}(\theta)}{\mathbb{P}(x)} \\ \pi(\theta \,|\, x) &= \frac{L(x \,|\, \theta)\pi(\theta)}{\pi(x)} \\ \pi(\theta \,|\, x) &\propto L(x \,|\, \theta)\pi(\theta) \end{aligned}$$

Here:

$$\begin{aligned} \pi(\theta) &= \text{Prior distribution} \\ L(x \,|\, \theta) &= \text{Likelihood function} \\ \pi(x) &= \text{Normalisation constant (usually ignorable)} \\ \pi(\theta \,|\, x) &= \text{Posterior distribution} \end{aligned}$$

Prior and Posterior

$$\pi(\theta \mid x) \propto L(x \mid \theta)\pi(\theta)$$

$\pi(\theta)$ = Prior distribution:
Describes our knowledge about the model parameters:

- in distributional form
- before we have seen the data.

Can be "uninformative" (e.g. "flat") or involve expert elicitation

$\pi(\theta \mid x)$ = Posterior distribution:
Describes our knowledge about the model parameters:

- in distributional form
- after we have seen the data.

Going from $\pi(\theta) \rightarrow \pi(\theta \mid x)$ reflects how our beliefs about the model parameters, θ, have changed by observing the data, x.

Do we really have prior knowledge?

Example:

We have a model $X \mid \theta \sim \text{Bin}(10, \theta)$ and observe $x = 10$.
Hypothesis $H_0 : \theta \leq 0.5$ is rejected in favour of $H_1 : \theta > 0.5$ each time.

3 different scenarios:

- A tea-drinker claims she can detect whether milk was added before or after the tea, just by taste. She does so correctly for 10 cups.
- A music expert claims he can distinguish between a page of Haydn's and Mozart's work. He correctly categorises 10 pieces.
- A drunk friend claims she can predict the outcome of tossing a fair coin, and does so correctly for 10 tosses.

In terms of data only, we must draw the same conclusion in each case.

But our *prior beliefs* suggest:

- being highly skeptical about the drunk friend;
- slightly impressed about the tea-drinker;
- not at all surprised about the music expert.

Bayesian validity

Argument for priors:

It is sensible to include prior information together with inference, as we invariably have some knowledge about the model parameters.

Argument against priors:

Different prior beliefs lead to different inferences.

Whether you see the argument against priors as a good or bad thing will determine your acceptability of the Bayesian approach.

Things to bear in mind:

- Prior knowledge is used to construct a likelihood model;
- Prior knowledge is used to determine test significance levels (can choose a high level to ensure rejection of "unlikely" hypothesis);
- Classical statistics maximises the likelihood surface. Bayesian statistics averages over it (see later). The prior just weights the averaging. Weighting is completely uncontroversial in classical statistics (think weighted regression).

Summary: Bayesian updating

There are 4 key steps in the Bayesian approach:

- Specification of a likelihood model $L(x \mid \theta)$;
- Determiniation/elicitation of a suitable prior distribution $\pi(\theta)$;
- Calculation of the posterior distribution via Bayes' Theorem

$$\pi(\theta \mid x) = \frac{L(x \mid \theta)\pi(\theta)}{\int_{\Theta} L(x \mid \theta)\pi(\theta)\mathrm{d}\theta}$$

- Draw inference from the posterior.

How to compute the posterior distribution?

There are two practical ways of deriving the posterior distribution

1) Algebraically:

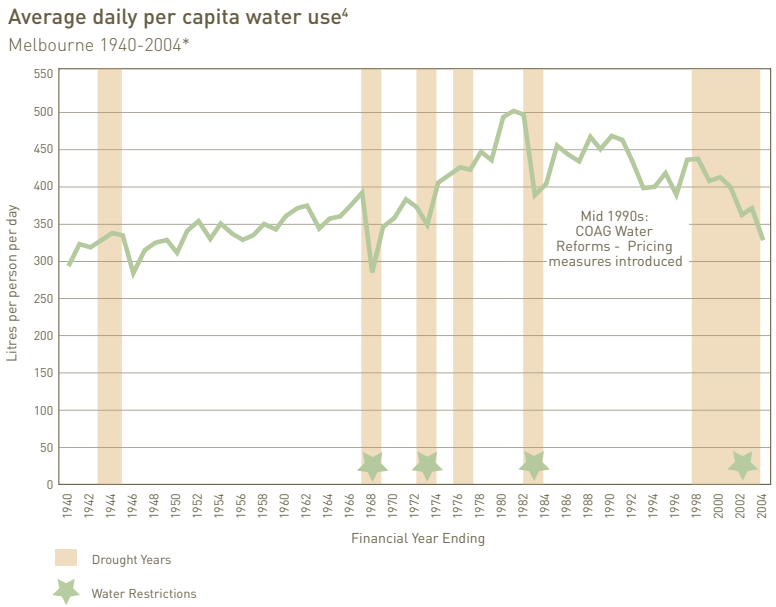
- All calculations are exact
- Calculations can be difficult
- Not always possible to obtain exact results (integrations not always possible)

2) Computationally:

- All calculations are are approximate
- Calculations are much simpler
- Can always obtain results

We cover both options in this course.

Example: Water Consumption



Melbourne average daily per capita water use (litres) from 1940-2004

Data is discrete, so one possible model is $x_i \sim \text{Poisson}(\theta)$, $i = 1, \ldots, n$.

Independence, and constant rate θ is clearly wrong, but it simplifies things for now (we will return to this example later).

$\sum_{i=1}^{n} x_i = 24890,\ n = 65$

Year	1940	1941	1942	1943	1944	1945	1946	...
Litres (x_i)	300	320	315	330	340	335	290	...

Example: Water Consumption

$\pi(x \mid \theta)$ = Likelihood function:

$x_i \sim \text{Poisson}(\theta)$: $\quad \pi(x \mid \theta) = \frac{\theta^x}{x!}\exp(-\theta) \qquad i = 1, \ldots, n,\ \theta > 0.$

$$L(x \mid \theta) = \prod_{i=1}^{n}\left(\frac{\theta^{x_i}}{x_i!}\exp(-\theta)\right) = \frac{\theta^{\sum_i x_i}\exp(-n\theta)}{\prod_i x_i!}$$

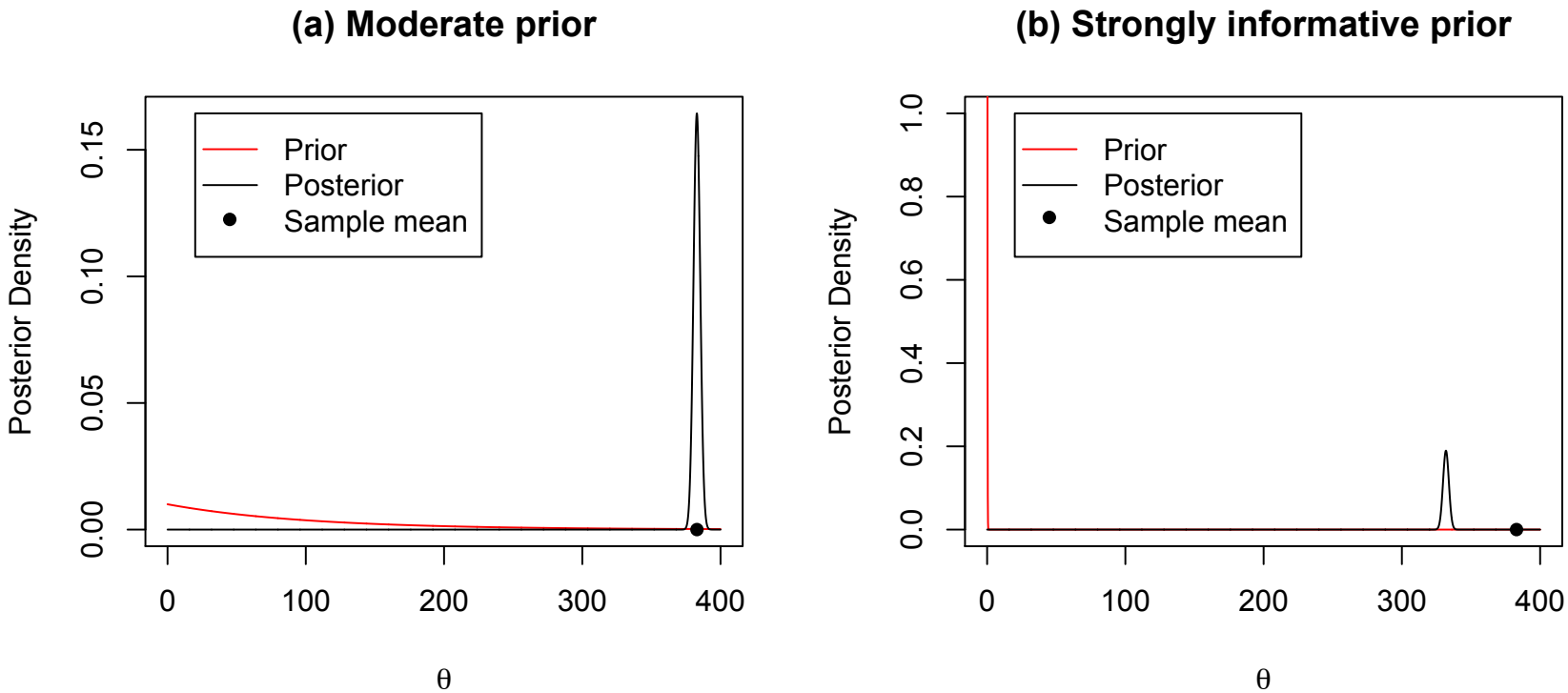
$\pi(\theta)$ = Prior distribution:

$\theta \sim \text{Gamma}(\alpha, \beta)$: $\quad \pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \qquad \alpha, \beta > 0$

$\pi(\theta \mid x)$ = Posterior distribution:

$$\begin{aligned}
\pi(\theta \mid x) &\propto \frac{\theta^{\sum_i x_i}\exp(-n\theta)}{\prod_i x_i!} \times \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \\
&\propto \theta^{\sum_i x_i}\exp(-n\theta) \times \theta^{\alpha-1}\exp(-\beta\theta) \\
&= \theta^{(\alpha+\sum_i x_i)-1}\exp(-\theta(\beta+n)) \\
&\propto \text{Gam}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)
\end{aligned}$$

Example: Water Consumption



Prior specification: $\alpha = 1, \beta = 0.01$ $\alpha = 1, \beta = 10$

Posterior is directly influenced by prior.

Prior choice:

- Can choose "uninformative" priors (see next week)
- Better to carefully elicit (justifiable) prior information
- Enough data can overwhelm prior information.

Posterior Inference

All information needed for further inference is contained in the posterior distribution

Examples:

- Point estimate (mean) of θ:

$$\bar{\theta} = \mathbb{E}_\pi[\theta] = \int_\Theta \theta \pi(\theta \,|\, x) \mathrm{d}\theta = \ldots = \frac{\alpha + \sum_i x_i}{\beta + n}$$

- Variance of θ:

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\theta) &= \int_\Theta \theta^2 \pi(\theta \,|\, x) \mathrm{d}\theta - \left(\int_\Theta \theta \pi(\theta \,|\, x) \mathrm{d}\theta \right)^2 \\ &= \ldots = \frac{\alpha + \sum_i x_i}{(\beta + n)^2} \end{aligned}$$

- 95% credible interval $[a, b]$ for θ: Solve (numerically)

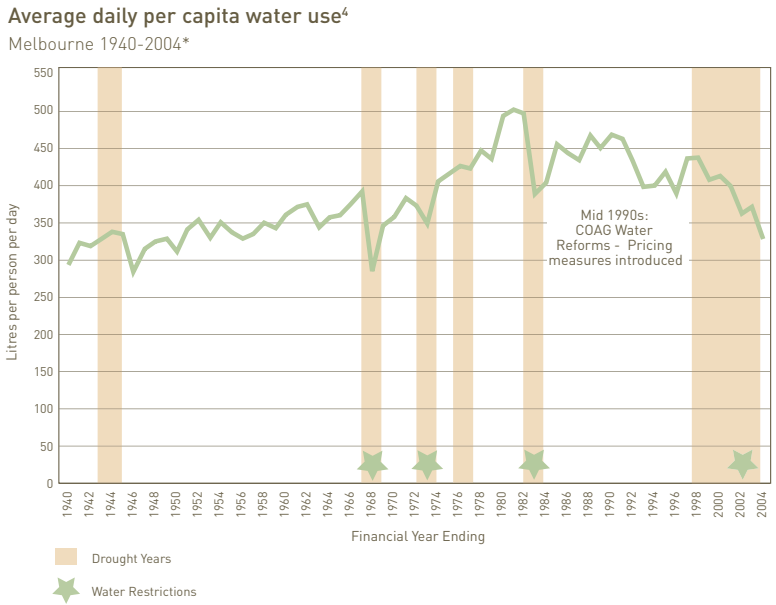
$$\int_0^a \pi(\theta \,|\, x) \mathrm{d}\theta = 0.025 \qquad \text{and} \qquad \int_0^b \pi(\theta \,|\, x) \mathrm{d}\theta = 0.975$$

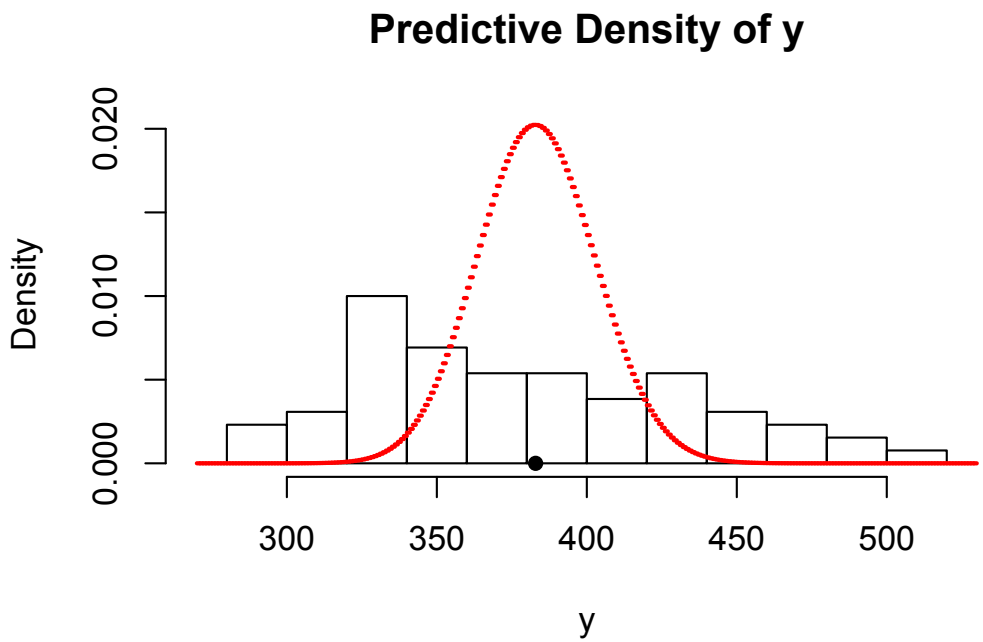
Posterior Inference

Examples: Predictive distribution for future data:

$$p(y \mid x) = \int_{\Theta} \pi(y \mid \theta)\pi(\theta \mid x)\mathrm{d}\theta$$

$$= \ldots = \text{NegBin}\left(y \,\middle|\, \alpha + n\bar{x}, \frac{1}{1+\beta+n}\right)$$





Note: Maths possible, though laborious. Q: is there an easier way?
(Answer: Yes – see later!)

Example: Water Consumption (cont.)

More realistic models

$x_i \sim \text{Poisson}(\theta)$ is clearly not realistic

Alternatives: $x_i \sim \text{Poisson}(\theta_i)$

- Linear trend:
  $\theta_i = \theta_0 + \theta_1 \text{year}_i$
  $(\theta_0, \theta_1) \sim \pi(\theta_0, \theta_1)$
- Piecewise linear trend:
  $\theta_i = \theta_0 + \theta_1 \text{year}_i + \theta_2 (\text{year}_i - \kappa)_+$ with e.g. $\kappa = 1982$.
  $(\theta_0, \theta_1, \theta_2) \sim \pi(\theta_0, \theta_1, \theta_2)$
- Piecewise linear trend (unknown κ):
  $\theta_i = \theta_0 + \theta_1 \text{year}_i + \theta_2 (\text{year}_i - \kappa)_+$
  $(\theta_0, \theta_1, \theta_2, \kappa) \sim \pi(\theta_0, \theta_1, \theta_2, \kappa)$

Clearly it is simple to construct more realistic models.

How to do: Bayesian inference for multi-parameter models?

Posterior Inference (multivariate)

All posterior inference proceeds via integration over the posterior.
Any parameters not of interest are integrated out.

Examples: Assume model parameters $(\theta_0, \theta_1, \theta_2, \kappa)$

- Marginal distribution of change-point κ:

$$\pi(\kappa \,|\, x) = \int_{\theta_0} \int_{\theta_1} \int_{\theta_2} \pi(\theta_0, \theta_1, \theta_2, \kappa \,|\, x, \text{year}) d\theta_0 d\theta_1 d\theta_2$$

- Point estimate (mean) of θ_2:

$$\begin{aligned} \bar{\theta}_2 &= \int_{\theta_2} \theta_2 \pi(\theta_2 \,|\, x, \text{year}) \mathrm{d}\theta_2 \\ &= \int_{\theta_2} \theta_2 \int_{\theta_0} \int_{\theta_1} \sum_{\kappa} \pi(\theta_0, \theta_1, \theta_2, \kappa \,|\, x, \text{year}) \mathrm{d}\theta_0 d\theta_1 \mathrm{d}\theta_2 \end{aligned}$$

- 95% credible interval $[a, b]$ for θ_2: Solve (numerically)

$$\int_0^a \pi(\theta_2 \,|\, x, \text{year}) d\theta = 0.025 \quad \text{and} \quad \int_0^b \pi(\theta_2 \,|\, x, \text{year}) d\theta = 0.975$$

Posterior Inference (multivariate)

Examples:

- Predictive distribution for future data:

$$
\begin{aligned}
p(y \mid x, \text{year}) &= \int_{\theta} \pi(y \mid \theta, \text{year}) \pi(\theta \mid x, \text{year}) d\theta \\
&= \int_{\theta_0} \int_{\theta_1} \int_{\theta_2} \sum_{\kappa} \pi(y \mid \theta_0, \theta_1, \theta_2, \kappa) \\
&\qquad \times \pi(\theta_0, \theta_1, \theta_2, \kappa \mid x, \text{year}) d\theta_0 d\theta_1 d\theta_2 \\
&= \ldots ?
\end{aligned}
$$

Summary:
Procedure is same as previously, and conceptually simple.

However, lots of difficult algebra even for simple models.
Closed solutions sometimes not possible.

This is the reason Bayesian statistics was not a viable procedure until the early 1990s.

Outline

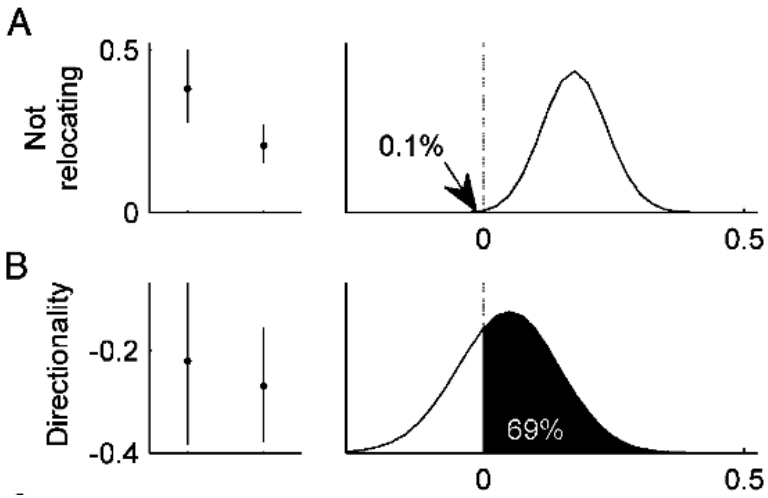
1. What is Bayesian statistics?
    - Bayes formula, prior and posterior probabilities
2. Monte Carlo integration basics
    - What is this, and why do we need it?

What is Monte Carlo integration?

We want to compute probabilities. E.g. Posterior mean or $\mathbb{P}(\mu_1 - \mu_2 > 0 \,|\, x)$

This requires integration of posterior

But it can be difficult(!?) to analytically integrate the posterior distribution to compute these.

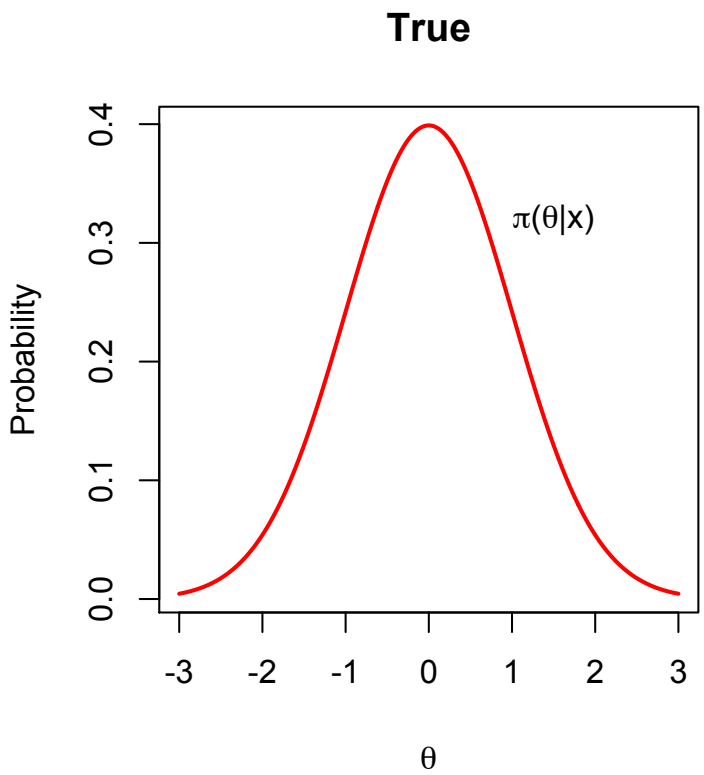


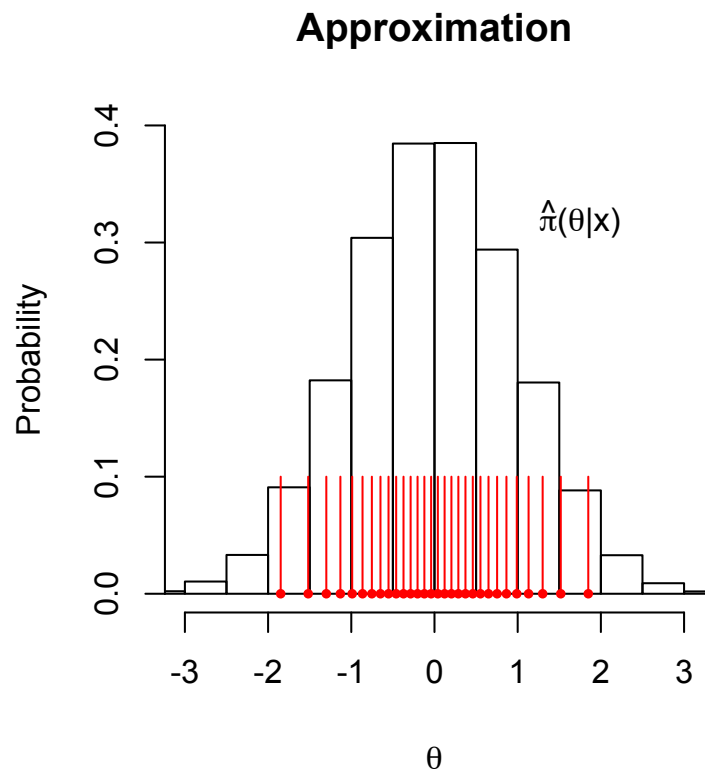
Monte Carlo refers to randomness.

Think of 'random' casino games of chance. So Monte Carlo integration refers to integration using random numbers.

How does this work?

A sample approximation to the posterior





Posterior mean

$$= \int_\theta \theta \pi(\theta \mid x) d\theta$$

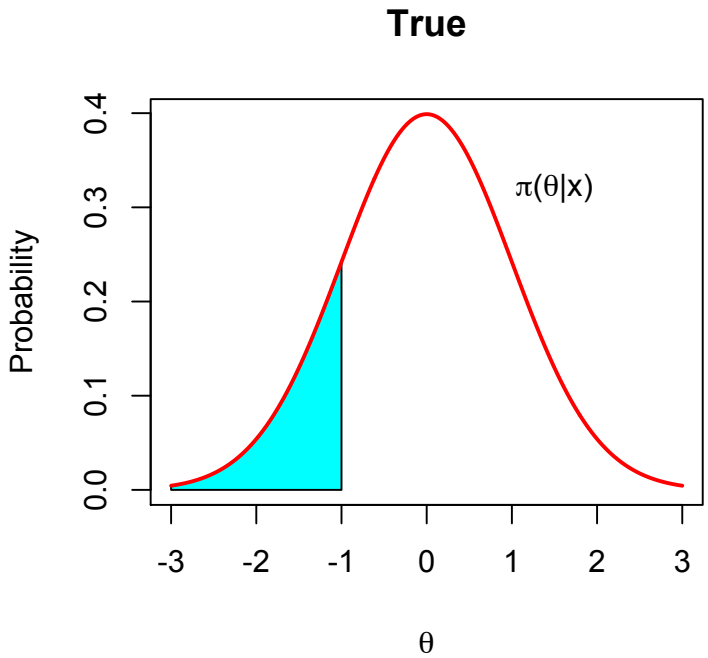
$$= \mathbb{E}_\pi[\theta]$$

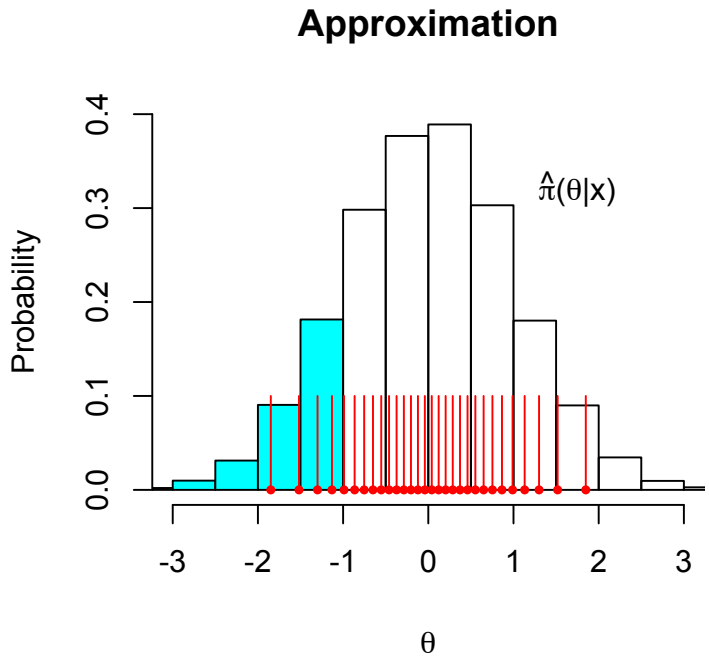
Posterior mean (*approximation*)

$$\approx \frac{1}{N} \sum_{i=1}^{N} \theta^{(i)}$$

where $\theta^{(1)}, \ldots, \theta^{(N)} \sim \pi(\theta \mid x)$.

A sample approximation to the posterior





$$\begin{aligned}\mathbb{P}(\theta < -1) &= \\ \int_\theta \mathbb{I}[\theta < -1]\pi(\theta \mid x)\mathrm{d}\theta &= \\ \mathbb{E}_\pi[\mathbb{I}(\theta < -1)]\end{aligned}$$

$$\mathbb{P}(\theta < -1) \approx \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\theta^{(i)} < -1]$$

where $\theta^{(1)}, \ldots, \theta^{(N)} \sim \pi(\theta \mid x)$.

Monte Carlo posterior predictive distributions

How to generate samples from $p(y \mid x)$?

$$p(y \mid x) = \int_{\Theta} \pi(y \mid \theta)\pi(\theta \mid x) d\theta$$

By inspection of formula:

- We already have samples $\theta^{(i)} \sim \pi(\theta \mid x)$
- For each $\theta^{(i)}$ we can generate $y^{(i)} \sim \pi(y \mid \theta^{(i)})$
- This gives us joint samples $(\theta^{(i)}, y^{(i)}) \sim \pi(y \mid \theta)\pi(\theta \mid x)$
- To obtain samples from $p(y)$, "integrate out" θ (i.e. discard the $\theta^{(i)}$ values) to leave $y^{(i)} \sim p(y)$ only

Hugely simpler than calculating exact algebraic expression!

Multivariate Monte Carlo

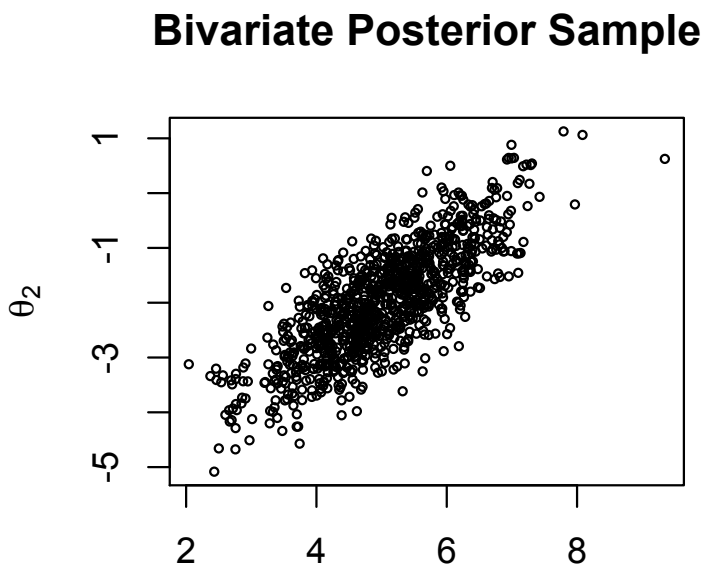
How to compute marginal distributions from joint distribution?

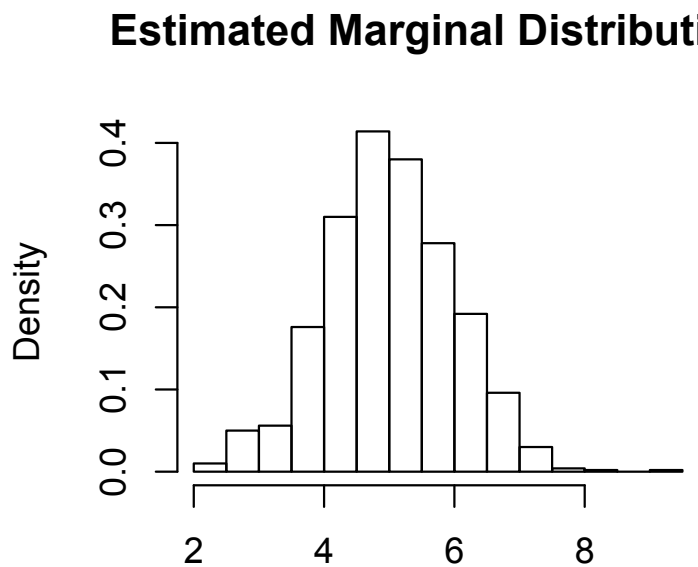
$$p(\theta_1 \mid x) = \int_{\theta_2} \pi(\theta_1, \theta_2 \mid x) d\theta_2$$

- Suppose we have samples $(\theta_1^{(i)}, \theta_2^{(i)}) \sim \pi(\theta_1, \theta_2 \mid x)$
- Integrating over θ_2: $\Rightarrow$ discard the θ_2 part of the sample

$$(\theta_1^{(i)}, \theta_2^{(i)}) \to \theta_1^{(i)}$$

- Construct histogram of $\theta_1^{(1)}, \ldots, \theta_1^{(n)}$





Multivariate Monte Carlo

How to compute distributions of functions of parameters?

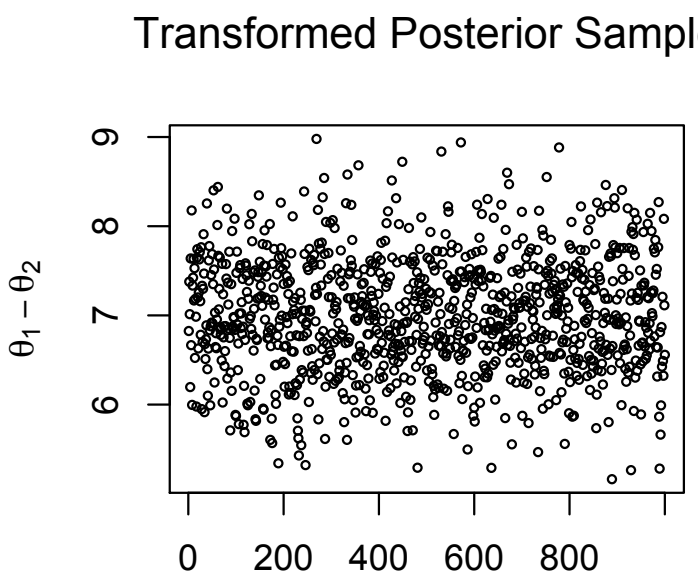
$$p(\theta_1 - \theta_2 \,|\, x) = \int_{\phi_2} \pi(\theta_1, \theta_2 \,|\, x) \left| \frac{d(\theta_1, \theta_2)}{d(\phi_1, \phi_2)} \right| d\phi_2$$

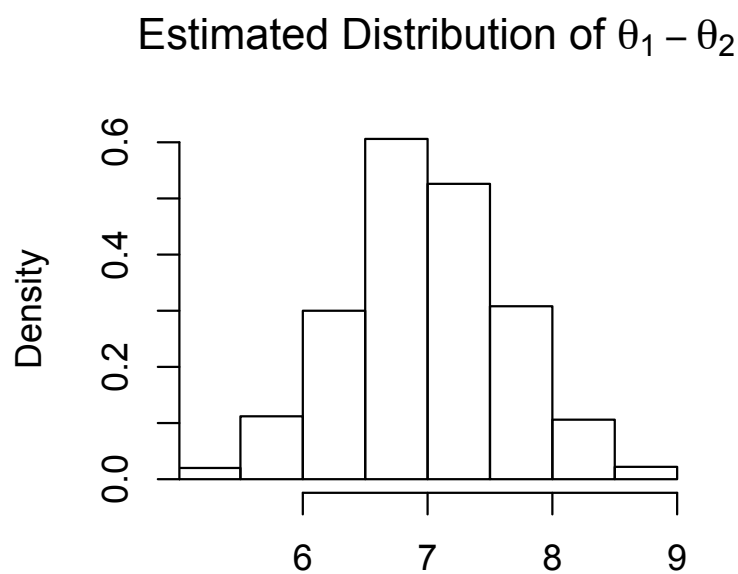
where $\phi = (\phi_1, \phi_2)$, with $\phi_1 = \theta_1 - \theta_2$ and e.g. $\phi_2 = \theta_1 + \theta_2$

- Given samples $(\theta_1^{(i)}, \theta_2^{(i)}) \sim \pi(\theta_1, \theta_2 \,|\, x)$, compute

$$(\theta_1^{(i)}, \theta_2^{(i)}) \to \theta_1^{(i)} - \theta_2^{(i)}$$

- Construct histogram of $\theta_1^{(1)} - \theta_2^{(1)}, \ldots, \theta_1^{(n)} - \theta_2^{(n)}$





Why does this work?

Monte Carlo integration:

Estimate posterior quantity of interest using sample from posterior

Justification for expectations:

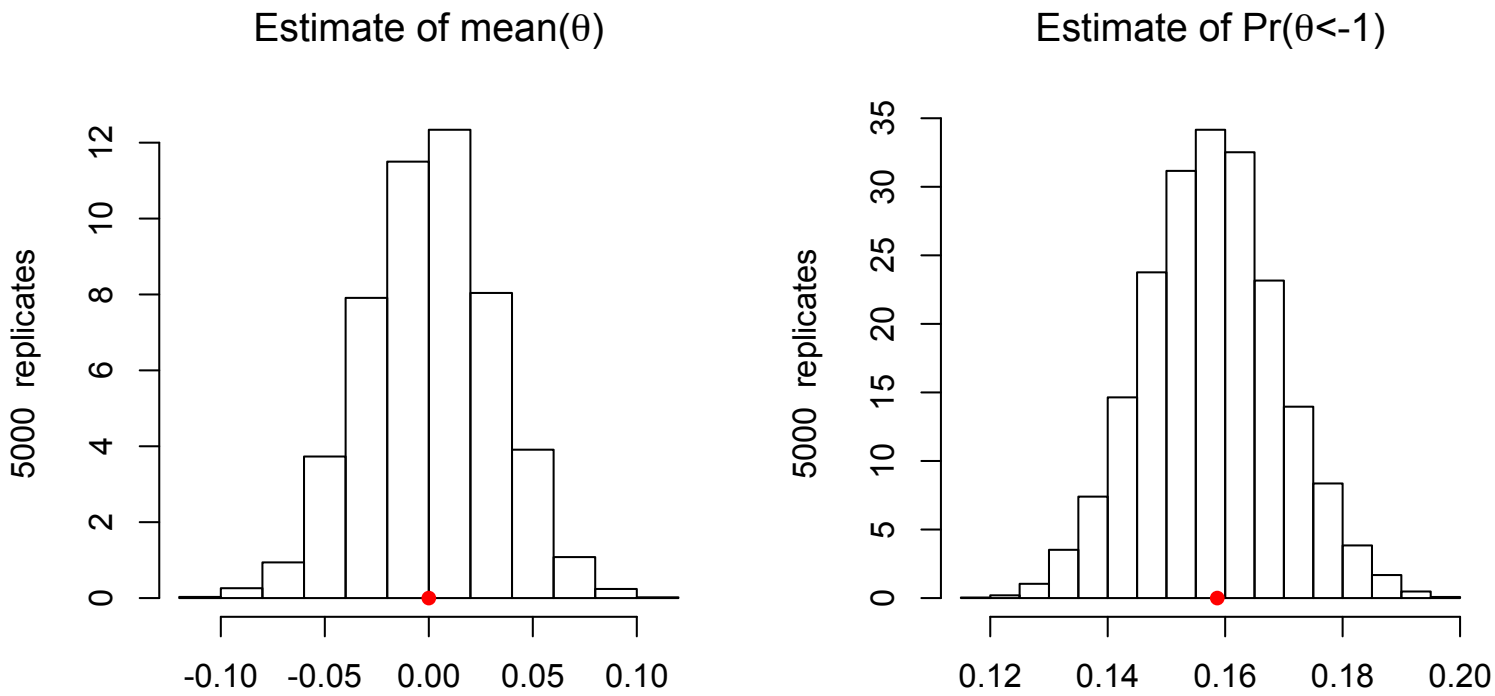
$$\pi(\theta \mid x) \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^{(i)}}(\theta)$$ indicator function

$\delta_A(b)$ is Dirac mass so that $\delta_A(b) = 1$ if $b \in A$, 0 otherwise.

Computing expectations:

$$\begin{aligned}
\mathbb{E}_\pi[g(\theta)] &= \int_\theta g(\theta)\pi(\theta \mid x)d\theta \\
&\approx \int_\theta g(\theta)\frac{1}{N}\sum_{i=1}^{N}\delta_{\theta^{(i)}}(\theta)d\theta \\
&= \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)})
\end{aligned}$$

Monte Carlo error



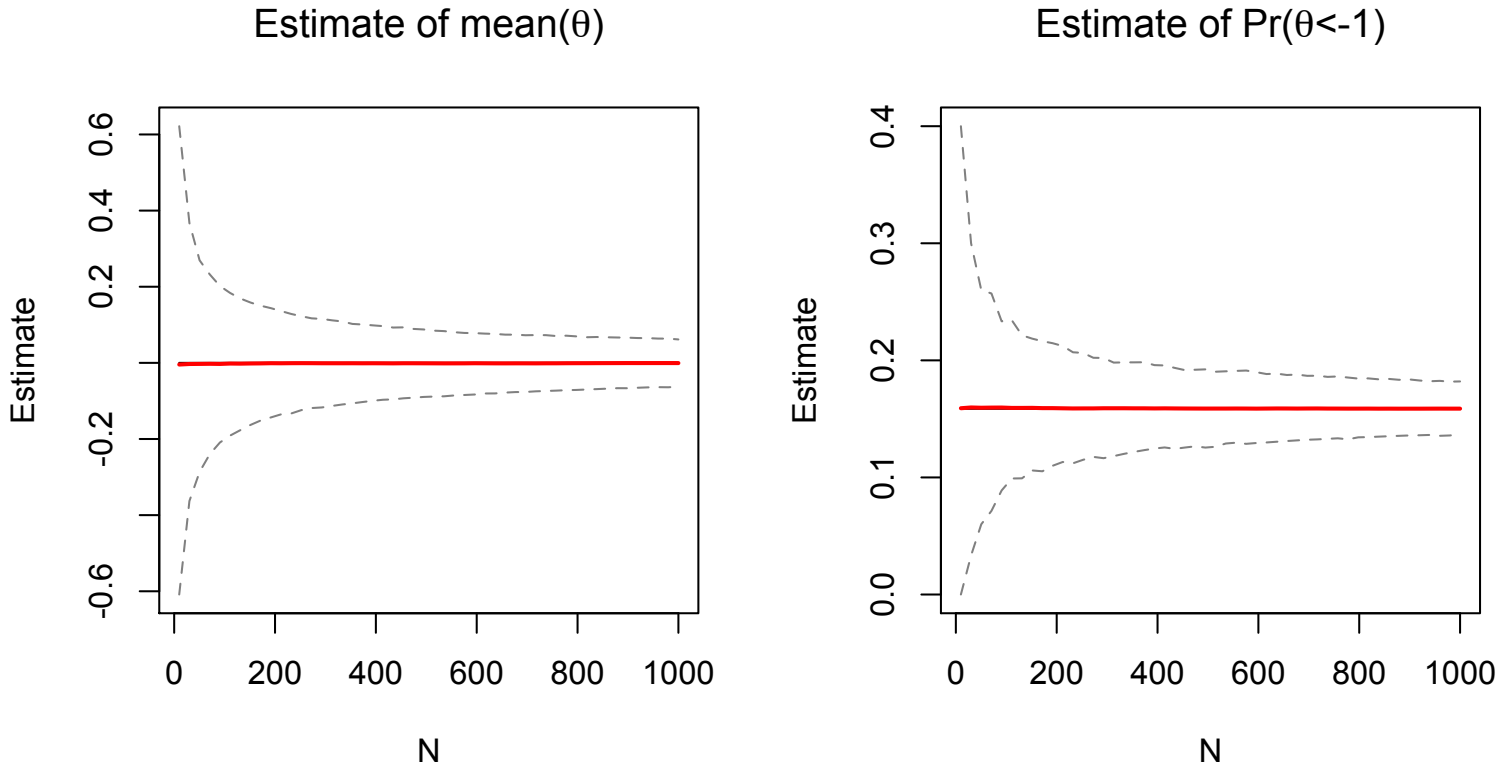
Estimating quantities with random samples gives random estimates
Estimate variability is known as Monte Carlo error.

$n = 1000, \theta^{(i)} \sim \mathrm{N}(0, 1)$

Reporting estimates of posterior quantities is commonly accompanied by (estimates of!) Monte Carlo error (recursion!)

[Not to be confused with posterior credible intervals . . . see later]

Monte Carlo error changes with sample size, N



Increasing N reduces the Monte Carlo error

For greater precision of estimates, increase number of samples

Given a target precision, continue to sample until achieved

Equivalent concept to classical "sample size computations" to find significant experimental results

Monte Carlo integration: Summary

1763 – 1984	Bayesian statistics is restricted to very simple analyses where exact integrations are possible
	Almost useless as a general statistical method
1984 – ∞	Complex integrations trivialised due to advent of cheap computing power and powerful Monte Carlo methods
	Bayesian statistics becomes mainstream statistical technique

Thanks to Monte Carlo methods, Bayesian inference is TRIVIAL to implement*, even in very complex models.

* Once you have the posterior samples!