


# *Introduction to Bayesian Data Analysis*

Mark Andrews

Psychology, Nottingham Trent University

`mark.andrews@ntu.ac.uk`

 `@xmjandrews`

# Introduction

- ▶ Bayesian data analysis is a general approach to statistical data analysis that is based on the extensive application of probability theory to all problems of inference and prediction.
- ▶ It is often contrasted with the more familiar *classical* or *frequentist* approach to data analysis.
- ▶ While these approaches have co-existed throughout 20th century statistics, until recently, the Bayesian approach was relatively marginalized.
- ▶ Since the early 1990s, however, there has been a remarkable rise in the prevalence of Bayesian methods.
- ▶ For example, we estimate that approximately 20% of recently published articles in high profile statistics journals are now on the topic of Bayesian methods. Likewise, there has been a concomitant rise in the prevalence of Bayesian methods in psychology.

# The Origin of Bayesian Data Analysis

- ▶ The origin of Bayesian inference can be traced to a posthumously published 1763 essay *Towards Solving a Problem in the Doctrine of Chances*, written some years earlier by Reverend Thomas Bayes (1701-1761).
- ▶ In this essay, Bayes focused on a problem that is mathematically identical to the problem of inferring the bias of a coin after observing the outcome of a sequence of flips of that coin.
- ▶ Bayes showed that the probability that the coin's bias is exactly  $\theta$  is proportional to the prior probability of  $\theta$  multiplied by the probability of the observed outcomes given  $\theta$ .

# Laplace and Inverse Inference

- ▶ Shortly after the publication of Bayes's essay, the French polymath Pierre-Simon Laplace (1749-1827) independently rediscovered Bayes' theorem and began to apply it to practical problems of data analysis.
- ▶ Laplace was the first to present Bayesian inference in what is now its modern form:

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{\int P(\mathcal{D}|\theta)P(\theta)d\theta}_{\text{marginal likelihood}}}$$

# Frequentism

- ▶ By the early 20th century, a frequentist definition of probability became increasingly widely adopted.
- ▶ Crucially, frequentism entails that the concept of probability only applies to the outcomes of a random *experiment*.
- ▶ As such, a parameter such as the bias of a coin, being a fixed but unknown quantity, can not be given a probabilistic interpretation.
- ▶ From this perspective, in the words of R. A. Fisher, Bayesian methods were "...founded upon an error, and must be wholly rejected." because "(i)nferences respecting populations, from which known samples have been drawn, cannot be expressed in terms of probability". Fisher (1925).

# *The Return of Bayesian Methods*

- ▶ Bayesian methods remained marginalized from the early to the late 20th century
- ▶ However, their potential practical advantage over classical methods was that they could be applied in principle to any problem where a probabilistic model of data could be specified.
- ▶ When computer power was minimal, any “in principle” advantages of Bayesian methods were not important.
- ▶ However, the roughly exponential growth of computational power from the 1970s onwards was the eventual catalyst for the adoption of Bayesian methods.
- ▶ By the late 1980's, Bayesian methods began returning to widespread use in science.

## Inference using Bayes's rule

- ▶ In any statistical model, we assume a probabilistic *generative model* of the data being analyzed.
- ▶ For example, in an independent samples t-test, we assume we have two data set drawn independently from two Normal distributions with fixed but unknown means.
- ▶ In general, Bayes's rule allows us to infer the probable values of the unknown variables in the probabilistic generative model on the basis of observed data by calculating

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{\int P(\mathcal{D}|\theta)P(\theta)d\theta}_{\text{marginal likelihood}}}$$

where  $\theta$  signifies the unknown variables, and  $\mathcal{D}$  is the observed data.

# *Inference using Bayes's rule*

- ▶ We can understand Bayes's rule by way of simple probability calculations.
- ▶ Consider the “Boxes and Bulbs” problem<sup>1</sup>:

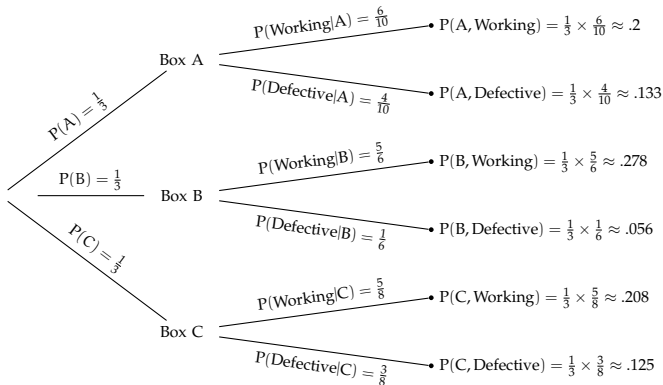
*Box A has 10 lightbulbs, of which 4 are defective. Box B has 6 lightbulbs, of which 1 is defective. Box C has 8 lightbulbs, of which 3 are defective. We randomly choose a box, and then randomly choose a lightbulb from that box. If we do choose a nondefective bulb, what is the probability it came from Box C?*

---

<sup>1</sup>This problem is taken from *Schaum's Outlines: Probability* (2nd ed., 2000), pages 87-88.



# Probabilistic tree diagram



- This tree structure gives you all the information you need to solve the problem. Note that on the right are given all the joint probabilities, which can be arranged to a joint probability table as follows:

	Working	Defective
Box A	.2	.133
Box B	.278	.056
Box C	.208	.125

- Now it is simple to see the overall probability of choosing a working bulb. It is probability of choosing Box A and choosing a working bulb *or* choosing Box B and choosing a working bulb *or* choosing Box C and choosing a working bulb. This is the sum of the first column, which is  $.2 + .278 + .208 \approx .686$ .
- Likewise, to get the probability of choosing Box C *given* that we have chosen a working bulb, we see what proportion of the total probability of choosing a working bulb, i.e. .686 is from when we choose Box C and a working bulb, i.e. .208. In other words, it is  $.208/.686 \approx .304$ .

## Using probability rules

- $P(\text{Bulb} = \text{Working})$  is equal to

$$\begin{aligned} P(\text{Working}) &= P(\text{Working}, A) + P(\text{Working}, B) + P(\text{Working}, C), \\ &= P(\text{Working}|A)P(A) + P(\text{Working}|C)P(B) + P(\text{Working}|C)P(C), \\ &= .2 + .278 + .208. \\ &\approx .686. \end{aligned}$$

- When asked for  $P(C|\text{Working})$  we use the rule of conditional probability being the joint probability divided by the marginal probability, i.e.

$$\begin{aligned} P(C|\text{Working}) &= \frac{P(\text{Working}, C)}{P(\text{Working})}, \\ &= \frac{P(\text{Working}|C)P(C)}{P(\text{Working}|A)P(A) + P(\text{Working}|C)P(B) + P(\text{Working}|C)P(C)}, \\ &= \frac{.208}{.686}, \\ &\approx .304. \end{aligned}$$

# Inference in a Bernoulli Distribution

*... Polish mathematicians Tomasz Gliszczynski and Wacław Zawadowski... spun a Belgian one euro coin 250 times, and found it landed heads up 140 times ... When tossed 250 times, the one euro coin came up heads 139 times and tails 111. ...*

*The Guardian, January 4, 2002<sup>2</sup>*

- ▶ A sample of  $n = 250$  coin tosses can be modelled as  $n$  independent and identically distributed Bernoulli random variables with parameter  $\theta$ .
- ▶ In other words, our probabilistic model is

$$x_i \sim \text{Bernoulli}(\theta), \quad \text{for } i \in \{1, 2, \dots, n\}.$$

and we would like to inference the probable values of  $\theta$  given an observation of  $m = 139$  (or  $m = 140$ , etc.).

---

<sup>2</sup>See <http://bit.ly/1B0Ku9b> for original story and <http://bit.ly/1B0Kx4Q> for discussion.

## *Inference in a Bernoulli Distribution*

- ▶ The probabilistic generative model of the Euro coin toss data is as follows:
  - ▶ The coin's bias corresponds to the fixed but unknown value of the parameter  $\theta$  of a Bernoulli random variable.
  - ▶ The observed outcomes  $x_1, x_2 \dots x_n$  are  $n$  iid samples from  $\text{Bernoulli}(\theta)$ .
- ▶ This generative model can be extended by assuming that  $\theta$  is itself drawn from a prior distribution  $P(\theta)$ :

$$\begin{aligned}\theta &\sim P(\theta), \\ x_i &\sim \text{Bernoulli}(\theta), \quad \text{for } i \in \{1, 2 \dots n\}.\end{aligned}$$

- ▶ In other words, we assume that a value for  $\theta$  was randomly drawn from  $P(\theta)$  and then  $n$  binary variables were sampled from  $\text{Bernoulli}(\theta)$ .

# Inference in a Bernoulli Distribution

- We aim to calculate

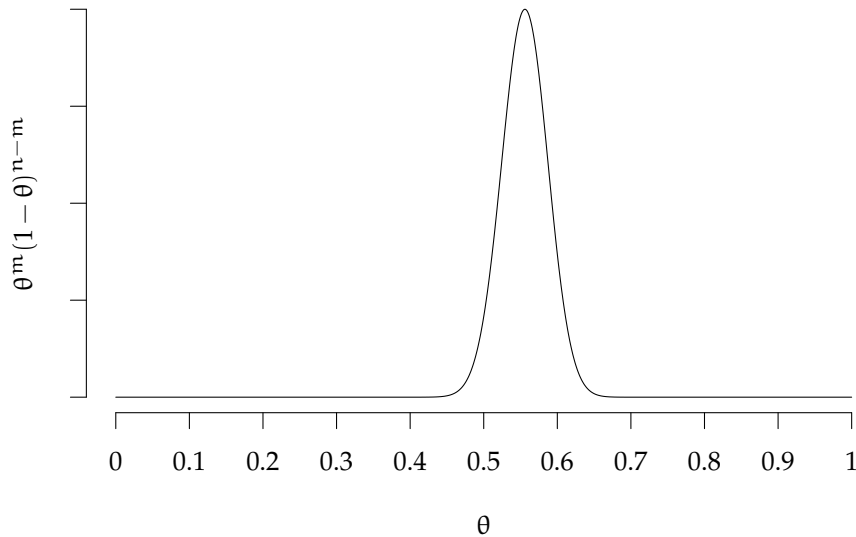
$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{\int P(\mathcal{D}|\theta)P(\theta)d\theta}_{\text{marginal likelihood}}}$$

where  $\theta$  signifies the coin's bias, and  $\mathcal{D}$  is the observed coin toss data.

- The *likelihood* of  $\theta$  gives the probability of the observed outcomes as a function of  $\theta$ :

$$\begin{aligned} P(x_1, x_2 \cdots x_n | \theta) &= \prod_{i=1}^n P(x_i | \theta), \\ &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}, \\ &= \theta^m (1 - \theta)^{n-m}. \end{aligned}$$

## *Inference in a Bernoulli Distribution*



The likelihood function for  $n = 250$  and  $m = 139$ .

# Conjugate Priors

- ▶ For a given likelihood function, a *conjugate prior* distribution is a prior probability distribution that leads to a posterior distribution of the same parameteric family.
- ▶ Using conjugate priors allows Bayesian inference and other probabilistic calculations to be performed analytically.
- ▶ Only a small subset of probabilistic models have conjugate priors.
- ▶ However, conjugate priors play a vital role in Monte Carlo methods like Gibbs sampling even in complex models.



## The beta distribution

- For the binomial likelihood function

$$\theta^m(1-\theta)^{n-m}$$

a conjugate prior is the beta distribution

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

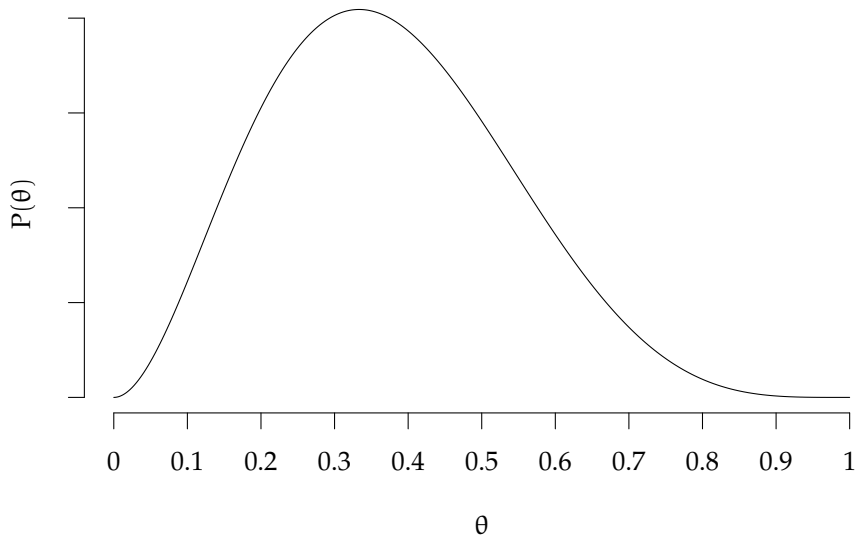
- The *normalizing constant* term is

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{1}{B(\alpha, \beta)},$$

where  $B(\alpha, \beta)$  is the beta function:

$$B(\alpha, \beta) = \int \theta^{\alpha}(1-\theta)^{\beta-1} d\theta.$$

## *The beta distribution*



The beta distribution with  $\alpha = 3$  and  $\beta = 5$ .

## Posterior distribution

- Denoting the observed data by  $D = (n, m)$ , with the beta prior, the posterior distribution is

$$\begin{aligned} P(\theta|D, \alpha, \beta) &= \frac{P(D|\theta)P(\theta|\alpha, \beta)}{\int P(D|\theta)P(\theta|\alpha, \beta) d\theta'} \\ &\propto \overbrace{\theta^m(1-\theta)^{n-m}}^{\text{likelihood}} \times \overbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{prior}}, \\ &\propto \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}, \\ &= \text{Beta}(m + \alpha, n - m + \beta). \end{aligned}$$

where the normalizing constant is the reciprocal of the beta function

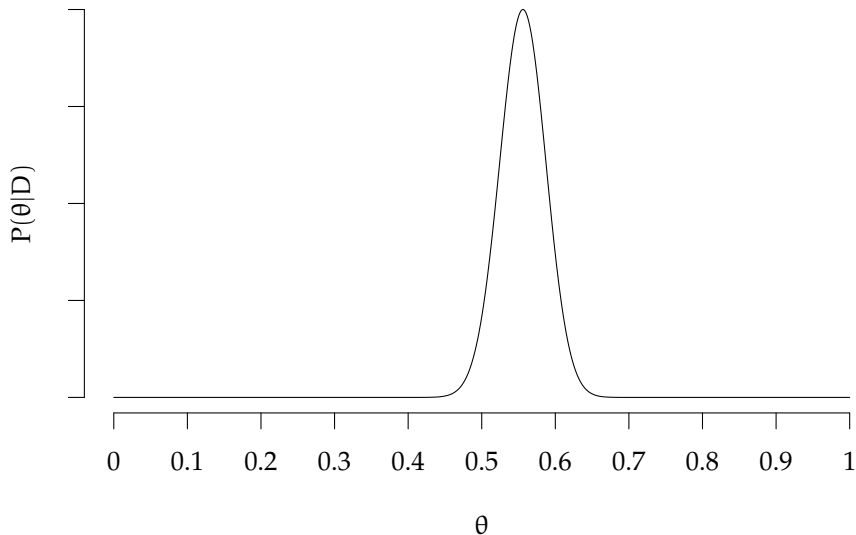
$$\begin{aligned} \frac{\Gamma(m + \alpha)\Gamma(n - m + \beta)}{\Gamma(n + \alpha + \beta)} &= \int \theta^{\alpha+m-1}(1-\theta)^{\beta+n-m-1} d\theta. \\ &= B(\alpha + m, \beta + n - m). \end{aligned}$$

## *Posterior distribution*

- ▶ For our Euro coin example, our observed data are  $n = 250$  and  $m = 139$ .
- ▶ A noninformative uniform prior on  $\theta$  is  $\text{Beta}(\alpha = 1, \beta = 1)$ .
- ▶ With this prior, the posterior distribution is

$$\begin{aligned}\text{Beta}(m + \alpha, n - m + \beta) &= \text{Beta}(139 + 1, 250 - 139 + 1), \\ &= \text{Beta}(140, 112)\end{aligned}$$

## *Posterior distribution*



The posterior distribution when  $n = 250$ ,  $m = 139$ ,  $\alpha = 1$  and  $\beta = 1$ .

## Summarizing the posterior distribution

- The mean, variance and modes of any beta distribution are as follows:

$$\langle \theta \rangle = \frac{\alpha}{\alpha + \beta},$$

$$V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

$$\text{mode}(\theta) = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

- Thus, in our case of  $\text{Beta}(140, 112)$ , we have

$$\langle \theta \rangle = 0.5556,$$

$$V(\theta) = 0.001, \quad \text{sd}(\theta) = 0.0312,$$

$$\text{mode}(\theta) = 0.556.$$

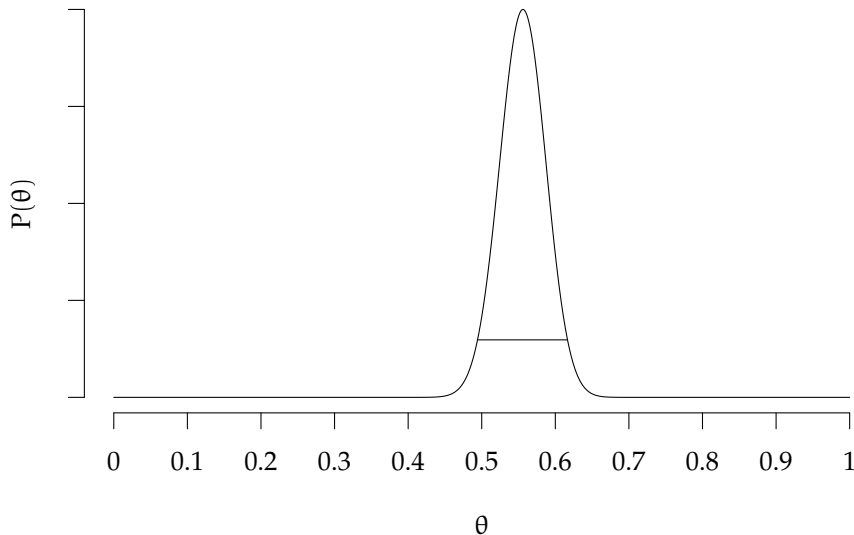
## *High posterior density (HPD) intervals*

- ▶ HPD intervals provide ranges that contain specified probability mass. For example, the 0.95 HPD interval is the range of values that contain 0.95 of the probability mass of the distribution.
- ▶ The  $\varphi$  HPD interval for the probability density function  $P(x)$  is computed by finding a probability density value  $p^*$  such that

$$P(\{x: P(x) \geq p^*\}) = \varphi.$$

- ▶ In other words, we find the value  $p^*$  such that the probability mass of the set of points whose density is greater than  $p^*$  is exactly  $\varphi$ .
- ▶ In general, the HPD is not trivial to compute but in the case of symmetric distributions, it can be easily computed from the cumulative density function.

## *The 0.95 HPD interval*



The posterior distribution, with its 0.95 HPD, when  $n = 250$ ,  $m = 139$ ,  $\alpha = 1$  and  $\beta = 1$ . In this case, the HPD interval is  $(0.494, 0.617)$ .



## Posterior predictive distribution

- ▶ Given that we have observed  $m$  heads in  $n$  coin tosses, what is the probability that the *next* coin toss is heads.
- ▶ This is given by the *posterior predictive* probability that  $x = 1$ :

$$\begin{aligned} P(x = 1|D, \alpha, \beta) &= \int P(x = 1|\theta) \overbrace{P(\theta|D, \alpha, \beta)}^{\text{Posterior}} d\theta, \\ &= \int \theta \times P(\theta|D, \alpha, \beta) d\theta, \\ &= \langle \theta \rangle, \\ &= \frac{\alpha + m}{\alpha + \beta + n}. \end{aligned}$$

- ▶ Thus, given 139 heads in 250 tosses, the predicted probability that the next coin will also be heads is  $\approx 0.5556$ .

# Marginal likelihood

- The posterior distribution is

$$P(\theta|D, \alpha, \beta) = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}} \overbrace{P(\theta|\alpha, \beta)}^{\text{Prior}}}{\underbrace{\int P(D|\theta)P(\theta|\alpha, \beta) d\theta}_{\text{Marginal likelihood}}}.$$

where the *marginal likelihood* gives the likelihood of the model given the observed data:

$$\int P(D|\theta)P(\theta|\alpha, \beta) d\theta \stackrel{\text{def}}{=} P(D|\alpha, \beta).$$

- In this example, it has a simple analytical form:

$$P(D|\alpha, \beta) = B(\alpha + m, \beta + n - m) = \frac{\Gamma(m + \alpha)\Gamma(n - m + \beta)}{\Gamma(n + \alpha + \beta)}.$$

## Model comparison

- ▶ Given  $D$ , we can compare the probability of model  $M_1$  relative to model  $M_0$  as follows:

$$\frac{P(M_1|D)}{P(M_0|D)} = \underbrace{\frac{P(D|M_1)}{P(D|M_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(M_1)}{P(M_0)}}_{\text{Priors odds}}.$$

- ▶ When both models are equally probable a priori, then the relative posterior probabilities is determined by the Bayes factor.
- ▶ We can compare our model  $M_1$ , i.e. with  $\alpha = \beta = 1$ , with the  $M_0$  model that  $\theta = \frac{1}{2}$ .

$$\frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(D|\theta)P(\theta|\alpha = 1, \beta = 1) d\theta}{\int P(D|\theta)\delta(\theta - \frac{1}{2}) d\theta}.$$

- ▶ This effectively compares a model that assumes a completely random coin machine, i.e., coin biases are generated according to  $\text{Beta}(\alpha = 1, \beta = 1)$ , to a completely perfect coin machine, i.e., coin biases are generated according to  $\delta(\theta - \frac{1}{2})$ .

## Model comparison

- We can compare our model  $M_1$ , i.e. with  $\alpha = \beta = 1$ , with the  $M_0$  model that  $\theta = \frac{1}{2}$ .

$$\begin{aligned}\frac{P(D|M_1)}{P(D|M_0)} &= \frac{\int P(D|\theta)P(\theta|\alpha = 1, \beta = 1) d\theta}{\int P(D|\theta)\delta(\theta - \frac{1}{2}) d\theta}, \\ &= \frac{\Gamma(\alpha + m)\Gamma(\beta + n - m)}{\Gamma(\alpha + \beta + n)} \Big/ \frac{1}{2}^m (1 - \frac{1}{2})^{n-m}, \\ &= \frac{m!(n - m)!}{(n + 1)!} \Big/ \frac{1}{2^n}.\end{aligned}$$

If  $n = 250$ ,  $m = 139$ , then

$$= \frac{139!111!}{251!} \Big/ \frac{1}{2^{250}} = 0.38.$$

- This is a factor of 2.65 in favour of the unbiased coin hypothesis.
- Note that the classical statistics null hypothesis test gives a p-value of  $p = 0.0875$ .

# The role of Markov Chain Monte Carlo

- In general, given observed data  $D$  and a model  $\Omega$ , the posterior distribution over the parameters  $\theta$  of the model is

$$P(\theta|D) = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}} \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{\int P(D|\theta)P(\theta) d\theta}_{\text{Marginal likelihood}}}.$$

where the *marginal likelihood* gives the likelihood of the model given the observed data.

- Given the posterior distribution  $P(\theta|D)$ , our aim is often to characterise this distribution in terms of e.g. its mean, variance, etc.
- Likewise, we may aim to calculate *posterior predictive* distributions such as

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D) d\theta.$$

## *Sampling from posterior distributions*

- ▶ In only rare situations can we determine the characteristics of the posterior distribution, or calculate posterior predictive distributions, in closed form.
- ▶ However, in general, if we can draw samples from  $P(\theta|D)$  then we can approximate, e.g., the mean of the distribution by

$$\langle \theta \rangle = \int \theta P(\theta|D) \approx \frac{1}{N} \sum_{i=1}^N \tilde{\theta}_i,$$

or the posterior predictive distribution by

$$P(x_{\text{new}}|D) = \int P(x_{\text{new}}|\theta)P(\theta|D) \, d\theta \approx \frac{1}{N} \sum_{i=1}^N P(x_{\text{new}}|\tilde{\theta}_i),$$

where

$$\tilde{\theta}_1, \tilde{\theta}_2 \dots \tilde{\theta}_N$$

are samples from  $P(\theta|D)$ .

## Rejection sampling

- ▶ Rejection sampling is one of the simplest methods for sampling from posterior distributions.
- ▶ Let us denote  $P(\theta|D)$  by  $f(\theta)$ .
- ▶ First we sample from  $\tilde{\theta}$  from another (simpler) distribution  $g(\theta)$ .
- ▶ The distribution  $g(\theta)$  can be any function so long as there exists a constant  $M$  such that

$$M \cdot g(\theta) \geq f(\theta),$$

for all possible values of  $\theta$ .

- ▶ Then draw  $u \sim U(0, 1)$ , a random sample from a uniform distribution between 0 and 1.
- ▶ If

$$u \leq \frac{f(\tilde{\theta})}{M \cdot g(\tilde{\theta})}$$

then keep  $\tilde{\theta}$ .

- ▶ Continue until  $N$  samples are collected.

# Gibbs sampling

- In a multivariate probability distribution, e.g.

$$P(x, y, z),$$

the univariate conditional distributions, e.g.  $P(x|y = \tilde{y}, z = \tilde{z})$ , may be straightforward to sample from.

- In Gibbs sampling, we set e.g.  $y$  and  $z$  to initial values  $\tilde{y}_0$  and  $\tilde{z}_0$  and then sample

$$\tilde{x}_0 \sim P(x|y = \tilde{y}_0, z = \tilde{z}_0),$$

$$\tilde{y}_1 \sim P(y|x = \tilde{x}_0, z = \tilde{z}_0),$$

$$\tilde{z}_1 \sim P(z|x = \tilde{x}_0, y = \tilde{y}_1),$$

and so on.

- After convergence, the samples e.g.  $\{\tilde{x}_N, \tilde{y}_N, \tilde{z}_N\}$  are draws from  $P(x, y, z)$ .



# Metropolis Hastings

- ▶ Let us denote  $P(\theta|D)$  by  $f(\theta)$ .
- ▶ We sample from a symmetric *proposal* distribution  $Q(\cdot|\cdot)$ .
- ▶ We start with an initial  $\tilde{\theta}_0$ , and sample

$$\tilde{\theta} \sim Q(\theta|\tilde{\theta}_0).$$

- ▶ We then accept  $\tilde{\theta}$  with probability

$$\alpha = \min \left( 1.0, \frac{f(\tilde{\theta})}{f(\tilde{\theta}_0)} \right).$$

- ▶ After convergence, the accepted samples are draws from the distribution  $f(\theta)$ .
- ▶ For Metropolis Hastings, the distribution  $f(\theta)$  need be only known up to a proportional constant.

# *The Advantages of Bayesian Methods*

- ▶ Bayesian methods provide full probabilistic interpretations of all inferences and predictions that form the statistical analysis.
- ▶ This allows for a more meaningful and concise interpretation of testing or estimation procedures.
- ▶ For example, Bayesian hypothesis tests allow us explicitly state the probability, contingent on the modelling assumptions, of one hypothesis relative to another in light of observed data. Likewise, Bayesian HPD's allow us to explicitly state the probability that an unknown quantity's value lies within a certain range of values.
- ▶ These contrast with classical p-values and confidence intervals.
- ▶ For example, a p-value for a hypothesis test gives the probability a statistic as or more extreme than that obtained could have happened were the hypothesis true. The meaningfulness of such a statement is less clear and p-values are routinely misinterpreted.

# *The Advantages of Bayesian Methods*

- ▶ In any probabilistic model that can be formally specified, inference concerning any variable is always possible in principle.
- ▶ In other words, the Bayesian approach provides a universal method of inference. The probable values of any variables in the model — partially observed variables, latent variables, parameters, hyper-parameters, etc. — can always be inferred in principle, once the model has been defined.
- ▶ This contrasts starkly with classical methods where, beyond a small set of special cases, inference proceeds in a case by case manner, often with each problem requiring its own solution and even simple problems being surprisingly intractable.
- ▶ The practical consequence of this is that Bayesian methods will be applicable to arbitrary probabilistic models, not simply those with convenient properties or those that have been studied intensely and solved.
- ▶ This allows the scientist or researcher to choose or develop a model based on its scientific meaning rather than on mathematical convenience.