Introduction to Statistics for Astronomers and Physicists

Section 3b: Priors and Introduction to Posterior Analysis

Dr Angus H Wright

2022-02-09

Section 3b: Introduction

Section 3: Bayesian Statistics (Weeks 7-9)

Bayes theorem led to a revolution in statistics, via the concepts of prior and posterior evidence. In modern astronomy and physics, applications of Bayesian statistics are widespread. We begin with a study of Bayes theorem, and the fundamental differences between frequentest and Bayesian analysis. We explore applications of Bayesian statistics, through well studied statistical problems (both within and outside of physics).

Specifying Priors

The computational difficulties of practical use of Bayes' Theorem generally arise when it is necessary to evaluate the normalisation constant in the denominator:

$$p(x) = \int p(\theta)p(x|\theta)\mathrm{d}\theta$$

For example, consider a dataset $\tilde{x}$ that consists of i.i.d. observations from a Poisson distribution with unknown rate parameter: $X \sim \mathrm{Pois}(\theta)$. Now suppose that our prior belief about θ is that θ is **definitely** within the range $0 \leq \theta \leq 1$, but all values in that range are equally likely. Our prior on θ is therefore:

$$p(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

In this case, the normalising constant in Bayes Theorem is:

$$p(x) = \int p(\theta)p(x|\theta)\mathrm{d}\theta \tag{1}$$

$$= \int_0^1 e^{-n\theta}\theta^{-\sum x_i}\mathrm{d}\theta \tag{2}$$

which can only be evaluated numerically!

Even simple priors can lead to awkward numerical problems.

Conjugate Priors

Fortunately, there are combinations of prior-likelihood pairs that are easy to analyse. These are the so-called **conjugate** distributions.

If F us a class of sampling distribution $p(x|\theta)$, and P is a class of prior distributions to θ, then the class P is **conjugate** to F if $P(\theta|x) \in P$ for all $P(.|\theta) \in F$ and $P(.) \in P$.

We saw such an example of such a pair of distributions last week with the binomial data and the beta distribution prior. It turns out that the *only* classes of data which have conjugate priors are those from the exponential familiy. That is:

$$p(x|\theta) = h(x)g(\theta)e^{t(x)c(\theta)}$$

for functions h, g, t, and c, such that:

$$\int p(x|\theta)\mathrm{d}x = g(\theta)\int h(x)e^{t(x)c(\theta)}\mathrm{d}x = 1$$

Distributions that satisfy these requirements are:

- The Exponential distribution
- The Poisson distribution
- The single-parameter Gamma distribution
- The Binomial distribution
- The Gaussian distribution with known variance

Provided that there is no direct conflict with our prior beliefs, and provided that our data follows one of the above distributions, conjugate priors provide us with mathematical simplicity.

Some prior-likelihood conjugate pairs are:

Likelihood	Prior	Posterior
$x \sim \mathrm{Bin}(n,\theta)$	$\mathrm{Be}(p,q)$	$\mathrm{Be}(p+x, q+n-x)$
$x_1,\ldots,x_n \sim \mathrm{Po}(\theta)$	$\mathrm{Ga}(p,q)$	$\mathrm{Ga}(p+\sum_{i=1}^n x_i, q+n)$
$x_1,\ldots,x_n \sim \mathrm{N}(\theta,\tau^{-1})$	$\mathrm{N}(b,c^{-1})$	$\mathrm{N}(\frac{cb+n\tau\bar{x}}{c+n\tau}, \frac{1}{c+n\tau})$
$x_1,\ldots,x_n \sim \mathrm{Ga}(k,\theta)$	$\mathrm{Ga}(p,q)$	$\mathrm{Ga}(p+nk, q+\sum_{i=1}^n x_i)$

(3)

where k and τ are known.

The Role of Data

Consider a set of i.i.d. random observations $X_1,\ldots,X_n$ drawn from $\mathrm{N}(\theta,\sigma^2)$, where σ^2 is known.

$$P(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(x_i-\theta)^2\right]$$

The likelihood for these data is:

$$\therefore l(\theta,x) \propto \exp\left[-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right]$$

From the conjugate prior: $\theta \sim \mathrm{N}(b,c^2)$. So:

$$\therefore p(\theta|x) \propto \exp\left\{-\frac{1}{2c^2}\sum(b-\theta)^2\right\} \times \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right\}$$

$$= \exp\left\{-\frac{b^2-2b\theta+\theta^2}{2c^2} - \frac{\sum x_i^2 - 2\theta n\bar{x} + n\theta^2}{2\sigma^2}\right\} \tag{4}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{c^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{b}{c^2}-\frac{n\bar{x}}{\sigma^2}\right)\right]\right\} \tag{5}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{c^2}+\frac{n}{\sigma^2}\right]\left[\theta - \frac{\frac{b}{c^2}-\frac{n\bar{x}}{\sigma^2}}{\frac{1}{c^2}+\frac{n}{\sigma^2}}\right]^2\right\} \tag{6}$$

$$\tag{7}$$

$$\therefore \theta|x \sim \mathrm{N}\left(\frac{\frac{b}{c^2} - \frac{n\bar{x}}{\sigma^2}}{\frac{1}{c^2} + \frac{n}{\sigma^2}}, \frac{1}{\left[\frac{1}{c^2} + \frac{n}{\sigma^2}\right]}\right)$$

We can write this neater by defining $\tau = 1/\sigma^2$ and $d = 1/c^2$:

$$\theta|x \sim \mathrm{N}\left(\frac{bd - n\bar{x}\tau}{d + n\tau}, \frac{1}{d + n\tau}\right)$$

Given this, we can observe some important details:

1) Observe that the posterior expectation is $\mathbb{E}[\theta|x] = \gamma_n b + (1 - \gamma_n)\bar{x}$, where $\gamma_n = \frac{d}{d+n\tau}$. That is, the **posterior mean** is a weighted average of the **prior mean** and the **sample mean**. The weight parameter γ_n is determined by the relative strength of information contained within the data and the prior. If $n\tau$ is large relative to c, then $\gamma_n \to 0$, and the posterior expectation converges onto the sample mean.
2) The posterior precision (reciprocal of the variance) is equal to the prior precision plus $n\times$ the data precision.
3) as $n \to \infty$, then $\theta|x \sim \mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$; so **the prior has no effect in the limit of large** n.
4) as $c \to \infty$, or equivalently as $d \to 0$, we again obtain $\theta|x \sim \mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$. **If the data is much more precise than the prior, then the prior has no effect**.
5) the posterior distribution depends on the data only through $\bar{x}$, and not through the individual values of x_i. In mathematical terms: $\bar{x}$ **is *sufficient* for** θ**.**

We can use asymptotic to demonstrate that, if the true value of a parameter θ is θ_0, and the prior probability of θ_0 is not 0, then with increasing amounts of data the posterior probability density at $\theta = \theta_0$ tends to unity.

Improper Priors

The strength of the prior belief about θ in this Gaussian mean example is determined by the variance, or precision d, of the Gaussian prior. Large values of c correspond to strong prior beliefs, and small values correspond to weak prior beliefs (and/or information). If we let $d \to 0$ (i.e. very weak prior belief), then $\theta|x \sim \mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$.

We have produced a perfectly valid posterior distribution in the limit of $c \to 0$, however this limit does not produce a valid *probability* distribution. This is because

$$p(\theta) = \mathrm{N}(0, \infty) \propto 1$$

which cannot be a valid probability distribution because

$$\int_{\mathcal{R}} p(\theta)\mathrm{d}\theta = \infty.$$

So the posterior $\mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$ cannot be obtained by using a "proper" prior distribution. Instead it arises through the use of a prior specification $p(\theta) \propto 1$, which is an example of an **improper prior distribution**.

Jeffery's Prior

We saw from the normal mean example that attempting to represent ignorance within the framework of a standard conjugate analysis led to the concept of improper priors. But there is a more fundamental set of problems as well.

If we specify a prior for θ of the form $p(\theta) \propto 1$, and then consider the parameter $\phi = \theta^2$:

$$p(\phi) = p(\theta^2)\frac{\mathrm{d}\theta}{\mathrm{d}\phi} \tag{8}$$

$$\propto \frac{1}{\sqrt{\phi}} \tag{9}$$

However, if we are ignorant about the possible values of θ, then surely we ought to be equally ignorant about the possible values of ϕ. So we could equally justify a prior $p(\phi) \propto 1$.

This demonstrates that prior ignorance *as represented by uniformity* is not translated across scales/transformations.

One particular point of view is that ignorance of priors ought to be consistent across $1-1$ parameter transformations. This idea leads us to the concept of **Jeffery's Prior**, which is based on the concept of Fisher Information:

$$I(\theta) = -\mathbb{E}\left\{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log\left[p(x|\theta)\right]\right\} \tag{10}$$

$$= \mathbb{E}\left\{\frac{\mathrm{d}}{\mathrm{d}\theta}\log\left[p(x|\theta)\right]\right\}^2. \tag{11}$$

In the case of a vector parameter θ, $I(\theta)$ is the matrix that is formed as minus the expectation of the matrix of second-order partial derivatives of $\log\left[p(x|\theta)\right]$.

Jeffery's prior is then defined as:

$$P_0(\theta) \propto |I(\theta)|^{\frac{1}{2}}$$

The consistency of the prior under transformations can be varified simply. Suppose $\phi = g(\theta)$ is a $1-1$ parameter transformation of θ. Then, by change of variables:

$$p(\phi) \propto p_0(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right| \tag{12}$$

$$= I(\theta)^{\frac{1}{2}}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|. \tag{13}$$

By definition:

$$I(\phi) = I(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|$$

and so:

$$p(\phi) \propto I(\phi)^{\frac{1}{2}}$$

as required.

We can see the Jeffery's prior in action using a few simple examples.

A Binomial Sample

$$x|\theta \sim \mathrm{Bin}(mn, \theta) \tag{14}$$

$$\therefore p(x|\theta) \propto \theta^x(1-\theta)^{n-x} \tag{15}$$

$$\log\left[p(x|\theta)\right] = x\log\theta + (n-x)\log(1-\theta) \tag{16}$$

$$\frac{\partial^2}{\partial\theta^2}\left\{log\left[p(x|\theta)\right]\right\} = \frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \tag{17}$$

So:

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\left\{log\left[p(x|\theta)\right]\right\}\right) = \frac{-n\theta}{\theta^2} - \frac{n-n\theta}{(1-\theta)^2} \tag{18}$$

$$= \frac{-n}{\theta} - \frac{n}{1-\theta} \tag{19}$$

$$\tag{20}$$

$$I(\theta) = \frac{n(1-\theta)+n\theta}{\theta(1-\theta)} \tag{21}$$

$$= \frac{n}{\theta(1-\theta)} \tag{22}$$

$$\therefore p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

Gaussian Distribution with both μ and σ unknown

Looking at a vector case with more than one unknown:

$$p(x|\theta) \propto \theta^{-n} \exp\left\{-\frac{n(s+(\bar{x}-\mu))}{2\sigma^2}\right\}$$

where $s = \frac{1}{n}(x_i - \bar{x})^2$, we have the Fisher Matrix:

$$I(\theta) = \mathbb{E}\left\{\begin{pmatrix} \frac{\partial^2}{\partial\mu^2} & \frac{\partial}{\partial\mu}\frac{\partial}{\partial\sigma} \\ \frac{\partial}{\partial\sigma}\frac{\partial}{\partial\mu} & \frac{\partial^2}{\partial\sigma^2} \end{pmatrix} \log\left[p(x|\theta)\right]\right\}$$

Going through each term one at a time:

$$\frac{\partial^2}{\partial\mu^2}\log\left[p(x|\theta)\right] = n\sigma^2$$

$$\frac{\partial}{\partial\sigma}\frac{\partial}{\partial\mu}\log\left[p(x|\theta)\right] = \frac{\partial}{\partial\mu}\frac{\partial}{\partial\sigma}\log\left[p(x|\theta)\right] \tag{23}$$

$$= \frac{n(\bar{x}-\mu)}{\sigma^4} \tag{24}$$

$$\tag{25}$$

$$\frac{\partial^2}{\partial\sigma^2}\log\left[p(x|\theta)\right] = \frac{-n}{2\sigma^4} + \frac{n(s+(\bar{x}-\mu)^2)}{2\sigma^6}$$

So:

$$\therefore I(\theta) = \begin{pmatrix} n\sigma^2 & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \propto \frac{1}{\sigma^4}$$

So the Jeffery's Prior for the Gaussian likelihood with unknown mean and variance is:

$$P_0(\mu,\sigma^2) \propto \frac{1}{\sigma^2}$$

Objections to Jeffery's Prior

There are a number of issues that have been raised with Jeffery's Prior. The most relevant of these is that the prior relies on the likelihood, and so depends on the form of the data. The prior distribution should only represent prior information, and should not be influenced by what data are to be collected.

Mixtures of Priors

Conjugate families of prior distributions are mathematically convenient. However they should only be used if a suitable member can be found which reflects the prior information at hand. In some situations, the natural conjugate family can be too restrictive.

Consider the case of a simple coin toss.

When a 'normal' coin is thrown it has almost invariably a probability of coming up heads 50% of the time. However, if the same coin is spun on a table, slight imperfections in the coin can cause it to prefer either heads or tails.

Taking this prior information into account, we may wish to specify a prior on θ that preferentially favours the unfair outcomes, *but agnostic to which one*. That is, we may wish to specify a bimodal distribution with peaks at $\theta = 0.3$ and $\theta = 0.7$.

Our likelihood for the number of heads in n spins of a coin will be Binomial: $X|\theta \sim \text{Bin}(n, \theta)$, and the conjugate distribution for the binomial distribution is the Beta distribution. However, no Beta distribution allows for bi- or multi-modality within the range $0 < \theta < 1$.

One solution is to use mixtures of conjugate distributions.

Suppose $p_1(\theta), \ldots, p_k(\theta)$ are all the conjugate distributions for θ, which lead to posterior distributions $p_1(\theta|x), \ldots, p_k(\theta|x)$. We can specify a family of multimodal prior distributions using weighted combinations of the different conjugate priors:

$$p(\theta) = \sum_{i=1}^{k} w_i p_i(\theta)$$

Such that:

$$\begin{aligned} p(\theta|x) &\propto p(\theta)p(x|\theta) && (26)\\ &= \sum_{i=1}^{k} w_i p_i(\theta) \times p(x|\theta) && (27)\\ &\propto \sum_{i=1}^{k} w_i^{\star} p_i(\theta|x) && (28) \end{aligned}$$

So the posterior is in the same family of conjugates, but with different relative weights.

Multi-parameter Models

For our purposes in the natural sciences, the examples that we've been using up until this point are not particularly useful. We have been looking largely at examples that analyse a single variable, such as the binomial coin-toss, a Gaussian distribution with known variance, etc. However in practice all problems that we will encounter will involve more than one variable.

This is where another aspect of Bayesian statistics is much more straight-forward than classical statistics. For highly complex multi-parameter models, no knew methods are required.

We now have a vector $\vec{\theta} = \{\theta_1, \ldots, \theta_k\}$ of unknown parameters which we wish to make inference about. We specify a multivariate model prior distribution $p(\vec{\theta})$ for $\vec{\theta}$, and combine the likelihood $p(x|\vec{\theta})$ via Bayes Theorem to obtain:

$$p(\vec{\theta}|x) = \frac{p(x|\vec{\theta})p(\vec{\theta})}{p(x)}$$

exactly as before.

We often want to draw conclusions about one or more parameters at the same time. These **marginal distributions** can be obtained in a straight-forward manner using probability calculations on the joint distributions.

For example, the marginal distribution of θ_1 is obtained by integrating out all of the other components of $\vec{\theta}$.

$$p(\theta_1|x) = \int_{\theta_2} \cdots \int_{\theta_k} p(\vec{\theta}|x) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_k$$

Equivalently, though, we can use **simulation** to draw samples from the joint distribution and then look at the parameters of interest (i.e. ignore the values of the other parameters).

Inference about multiparameter models creates the following complications:

1) **Prior Specification**: priors are now multivariate distributions. This means that we need to express dependencies between parameters as well, which is often complicated.
2) **Computation**: we now have even more complicated integrals to evaluate, which creates the necessity for complex numerical techniques.
3) **Interpretation**: the structure of the posterior distribution may be highly complex, which causes difficulties in interpretation.

Manufacturing a problem

Suppose a machine is either satisfactorily made ($x = 1$) or not ($x = 2$). The probability of $x = 1$ depends on the room temperature during manufacturing, which is either cool ($\theta_1 = 0$) or hot ($\theta_1 = 1$), and the humidity, which is either dry ($\theta_2 = 0$) or humid ($\theta_2 = 1$). The probabilities of satisfactorily making a machine given the environmental conditions are governed by the likelihood:

$P(x = 1\|\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.6	0.8
$\theta_2 = 1$	0.7	0.6

(29)

The joint prior distribution of θ_1, θ_2 is given by:

$P(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.3	0.2
$\theta_2 = 1$	0.2	0.3

(30)

We can therefore calculate the numerator of Bayes theorem as:

$P(x = 1\|\theta_1, \theta_2) \times P(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.18	0.16
$\theta_2 = 1$	0.14	0.18

(31)

And so the joint posterior distribution is:

$P(\theta_1, \theta_2\|x)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	$\frac{18}{66}$	$\frac{16}{66}$
$\theta_2 = 1$	$\frac{14}{66}$	$\frac{18}{66}$

(32)

If we are only interested in the marginal distributions of the probability that it was hot/cool or dry/humid, we can simply sum along the axis of interest to get the **marginal distributions**:

$$P(\theta_1 = 0|x, \theta_2) = 32/66 \tag{33}$$
$$P(\theta_1 = 1|x, \theta_2) = 34/66 \tag{34}$$
$$P(\theta_2 = 0|x, \theta_1) = 34/66 \tag{35}$$
$$P(\theta_2 = 1|x, \theta_1) = 32/66 \tag{36}$$

We can show the same methods for more complex distributions, however the main question of interest is how do we summarise complex posteriors sensibly and accurately.

Summarising Posterior Information

The posterior distribution is a complete summary of the inference about θ. In essence, the posterior distribution **is** the inference about θ. However, for many applications, we wish to summarise the information contained within the posterior into some digestible quantity.

Often such a quantity is the "best" estimate of the value of θ; the **point estimate** of the unknown parameter. But this raises the question: how does one define what is "best"?

We can do this by specifying some form of **Loss Function** $L(\theta, a)$, which measures the penalty in estimating a value of θ given an individual datum at a.

There are a range of potential loss functions that are common to use, and the particular choice of the loss function will often depend on the problem being analysed. The most common loss functions (some of which we have already seen!) are:

1) Quadratic Loss: $L(\theta, a) = (\theta - a)^2$
2) Absolute Error Loss: $L(\theta, a) = |\theta - a|$
3) $0 - 1$ Loss:

$$L(\theta, a) = \begin{cases} 0 & \text{if } |a - \theta| \leq \epsilon \\ 1 & \text{if } |a - \theta| > \epsilon \end{cases}$$

4) Linear Loss: for specified $g, h > 0$

$$L(\theta, a) = \begin{cases} g(a - \theta) & \text{if } a > \theta \\ h(a - \theta) & \text{if } a \leq \theta \end{cases}$$

In each case, by minimising the posterior expected loss, we obtain a particular point estimate of θ for that particular choice of the loss function.

Calculating the expected loss $\rho(a, x)$ given a posterior distribution $p(\theta|x)$ requires calculating:

$$\rho(a, x) = \int L(\theta, a) p(\theta|x) \mathrm{d}\theta$$

Quadratic Loss

$$\begin{aligned}
\rho(a, x) &= \int L(\theta, a) p(\theta|x) \mathrm{d}\theta && (37) \\
&= \int (\theta - a)^2 p(\theta|x) \mathrm{d}\theta && (38) \\
&= \int (\theta - \mathbb{E}(\theta|x) + \mathbb{E}(\theta|x) - a)^2 p(\theta|x) \mathrm{d}\theta && (39) \\
&= \int (\theta - \mathbb{E}(\theta|x))^2 p(\theta|x) \mathrm{d}\theta - && (40) \\
&\quad 2 \int (\theta - \mathbb{E}(\theta|x))(\mathbb{E}(\theta|x) - a) p(\theta|x) \mathrm{d}\theta + && (41) \\
&\quad \int (\mathbb{E}(\theta|x) - a)^2 p(\theta|x) \mathrm{d}\theta && (42) \\
&= \mathrm{Var}(\theta|x) + \mathbb{E}(\theta|x) - a)^2 && (43)
\end{aligned}$$

As:

$$\mathrm{Var}(x) = \mathbb{E}\left[(x - \mu)^2\right]$$

So the Quadratic Loss is minimized when $a = \mathbb{E}(\theta|x)$. Hence the posterior mean minimizes the quadartic loss, and the expected loss is the posterior variance $\mathrm{Var}(\theta|x)$.

Linear Loss

We can play the same game and demonstrate the expected loss using other loss functions. The case of absolute error loss is a special case of the linear loss (where $g = h = 1$). Looking at the linear loss, we can let q be the $h/(g+h)$ quartile of the posterior distribution:

$$\frac{g}{g+h} = \int_{-\infty}^{q} p(\theta|x)\mathrm{d}\theta$$

and suppose that $a > q$, then:

$$L(\theta, q) - L(\theta, a) = \begin{cases} g(q-a) & \text{if} \quad \theta \leq q \\ (g+h)\theta - hq - ga & \text{if} \quad q < \theta < a \\ h(a-q) & \text{if} \quad a \leq \theta \end{cases}$$

but for $q < \theta < a$:

$$(g+h)\theta - hq - ga < h(a_q)$$

so that:

$$L(\theta, q) - L(\theta, a) \leq \begin{cases} g(q-a) & \text{if} \quad \theta \leq q \\ h(a-q) & \text{if} \quad q \leq \theta \end{cases}$$

So

$$\mathbb{E}\left[L(\theta, q) - L(\theta, a)\right] \leq g(q-a)\frac{h}{g+h} + h(a-q)\left(1 - \frac{h}{g+h}\right) = 0$$

That is, $\rho(a, x) \geq \rho(q, a)$ for all a, so by setting $a = q$, the $h/(g+h)$ quartile of the posterior distribution minimises the linear loss. If $g = h = 1$, $h/(g+h) = 0.5$; that is, we are looking at the **median** of the distribution. Therefore, the absolute error loss is minimised by the posterior *median*.

0-1 Loss

In this case:

$$\rho(a, x) = P\left\{|\theta - a| > \epsilon|x\right\} = 1 - P\left\{|\theta - a| \leq \epsilon|x\right\}$$

We can define the **modal interval** of length 2ϵ as the interval $[x - \epsilon, x + \epsilon]$ which has the highest probability, then the midpoint of this interval is the distribution mode. If we chose ϵ to be arbitrarily small, the procedure will identify the posterior mode as the point estimate with this loss function.

Credability Intervals

The idea of **credability intervals** is analagous to confidence intervals in classical statistics. The reasoning is that point estimates give no measure of accuracy, so it is prefereable to specify the region in which a parameter is **likely** to reside.

This causes problems in classical statistics as parameters are not regarded as random, and so it is not possible to give an interval in which we interpret as having a certain *probability* of the parameter resideing there. There is no such difficulty in Bayesian Statistics, because the parameters are treated as random.

We can therefore define a region $C_\alpha(x)$ which is the $100(1-\alpha)\%$ credible interval for θ if:

$$\int_{C_\alpha(x)} P(\theta|x)\mathrm{d}\theta = 1 - \alpha$$

That is, there is a probability of $1 - \alpha$, based on the posterior distribution, that θ lies within $C_\alpha(x)$. However this leads to an obvious problem; the interval is not uniquely defined! Any region containing $1 - \alpha$ in probability will satisfy the equation.

However, as we generally want to find the **most probable** values of the parameter, we can naturally impose additional constraints on the credible interval. E.g. that the credible interval ought to be as small as possible.

This amounts to a region of the form:

$$C_\alpha(x) = \{\theta : P(\theta|x) \geq \gamma\}$$

where γ is chosen to ensure that:

$$\int_{C_\alpha(x)} P(\theta|x)\mathrm{d}\theta = 1 - \alpha$$

Such a region is called the **highest posterior density region**, and in general we find the region numerically.