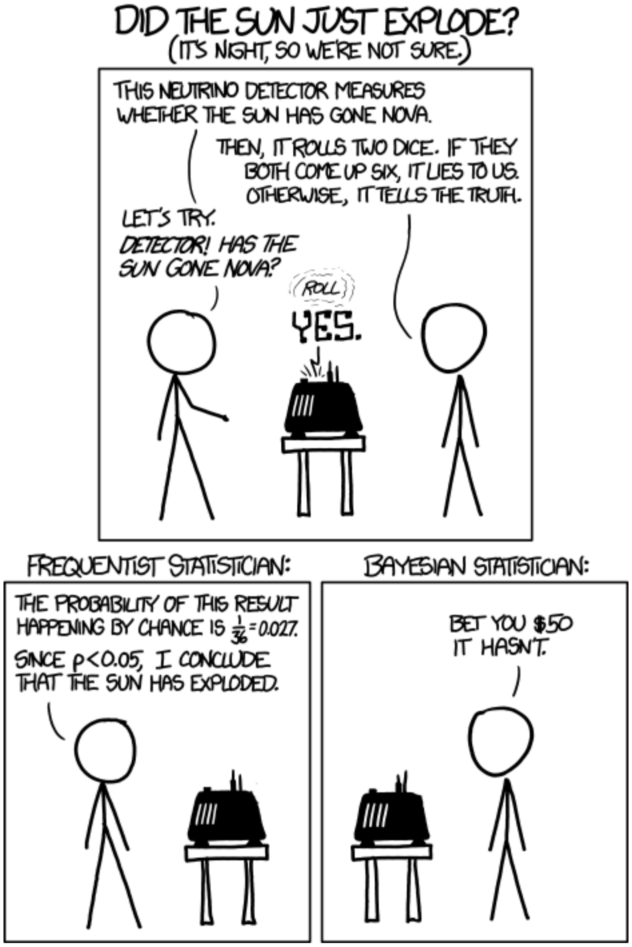
STATISTICS AND DATA SCIENCE 365 / 565
DATA MINING AND MACHINE LEARNING

# Notes on Bayesian Inference


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

xkcd.com/1132/

## 1. What's the Difference?

There are two main approaches to statistical inference, *frequentist* (or classical) methods and *Bayesian* methods. The key differences between the frequentist and Bayesian approaches are as follows:

| | frequentist | Bayesian |
|---|---|---|
| probability is: | limiting relative frequency | degree of subjective belief |
| parameter $\theta$ is a: | fixed constant | random variable |
| probability statements are about: | procedures | parameters |
| frequency guarantees? | yes | no |

To illustrate these differences, we consider the *interval estimation problem*. A typical Bayesian statement is: the subjective probability that a parameter $\theta$ is in a *credible interval* $C$ given the data $\mathcal{D}$ is 95 percent. A typical frequentist statement is: the frequency with which a *confidence interval* $C$ traps $\theta$ is at least 95 percent, no matter what the value of $\theta$ is. In symbols:

Idk what Θ is but given the data, it might be in C

Θ is something, and it might be in interval C, but that interval can change

$$\text{Bayesian}: \ \mathbb{P}(\theta \in C \,|\, \mathcal{D}) = 0.95, \qquad \text{frequentist}: \ \min_{\theta} \mathbb{P}_{\theta}(\theta \in C) = 0.95.$$

In the first statement, $\theta$ is random and $C$ is fixed. In the second statement, $\theta$ is fixed and $C$ is random. To elucidate the notion of frequentist coverage further, consider a sequence of statistical problems involving parameters $\theta_1, \theta_2, \ldots$. The problems can be completely unrelated and the parameters could refer to completely different quantities. If $C_1, C_2, \ldots$ are the corresponding confidence intervals with 95 percent coverage, then

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\theta_i \in C_i) \geq 0.95$$

The indicator function is going to give 1 or 0 / yes or no if Θ is in the interval
With a sequence of n parameters, surely some parameter will be part of a confidence interval
This probability will be greater than 0.95
Is it hinting at the danger of arriving at spurious conclusions like in the XKCD comic?

almost surely, for any sequence $\theta_1, \theta_2, \ldots$.

The field of statistics puts more emphasis on frequentist methods although Bayesian methods certainly have a presence. The machine learning community embraces Bayesian methods more strongly. There are, in fact, many flavors of Bayesian inference. *Subjective Bayesians* interpret probability strictly as personal degrees of belief. *Objective Bayesians* try to find prior distributions that formally express ignorance with the hope that the resulting posterior is, in some sense, objective. *Empirical Bayesians* estimate the prior distribution from the data. *Frequentist Bayesians* are those who use Bayesian methods only when the resulting posterior has good frequency behavior.

## 2. The Bayesian Method

Also called generative model
Also called likelihood

Let $x_1, \ldots, x_n$ be $n$ observations sampled from a probability density $p(x \,|\, \theta)$. In this chapter, we write $p(x \,|\, \theta)$ if we view $\theta$ as a random variable and $p(x \,|\, \theta)$ represents the conditional probability density conditioned on $\theta$. In contrast, we write $p_\theta(x)$ if we view $\theta$ as a deterministic value. Bayesian inference is usually carried out in the following way.

1. We choose a probability density $\pi(\theta)$ — called the *prior distribution* — that expresses our beliefs about a parameter $\theta$ before we see any data.

Also called generative model / likelihood

2. We choose a statistical model $p(x \mid \theta)$ that reflects our beliefs about $x$ given $\theta$.
3. After observing data $\mathcal{D}_n = \{x_1, \dots, x_n\}$, we update our beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

By Bayes' theorem, the posterior distribution can be written as

Posterior Likelihood/ generative model Prior

$$p(\theta \mid x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n \mid \theta)\pi(\theta)}{p(x_1, \dots, x_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta) \qquad (1)$$

Evidence

where $\mathcal{L}_n(\theta) = \prod_{i=1}^n p(x_i \mid \theta)$ is the *likelihood function* and

Likelihood: for all trials, multiply them together
Each trial is the probability of seeing that X value given Θ

$$c_n = p(x_1, \dots, x_n) = \int p(x_1, \dots, x_n \mid \theta)\pi(\theta) d\theta = \int \mathcal{L}_n(\theta)\pi(\theta) d\theta$$

is the normalizing constant, which is also called *evidence*.

We can get a *Bayesian point estimate* by summarizing the center of the posterior. Typically, we use the mean or mode of the posterior distribution. The posterior mean is

You can get a prediction for Θ
By finding the two or three dimensional
Mean/center of the distribution

$$\overline{\theta}_n = \int \theta p(\theta \mid \mathcal{D}_n) d\theta = \frac{\int \theta \mathcal{L}_n(\theta)\pi(\theta) d\theta}{\int \mathcal{L}_n(\theta)\pi(\theta) d\theta}. \qquad (2)$$

We can also obtain a *Bayesian interval estimate*. For example, for $\alpha \in (0, 1)$, we could find $a$ and $b$ such that

$$\int_{-\infty}^{a} p(\theta \mid \mathcal{D}_n)\, d\theta = \int_{b}^{\infty} p(\theta \mid \mathcal{D}_n)\, d\theta = \alpha/2.$$

We can find a credible interval (in two dimensions)
To capture the prediction for Θ within some bound

Let $C = (a, b)$. Then

$$\mathbb{P}(\theta \in C \mid \mathcal{D}_n) = \int_{a}^{b} p(\theta \mid \mathcal{D}_n)\, d\theta = 1 - \alpha$$

so $C$ is a $1 - \alpha$ *Bayesian posterior interval* or *credible interval*. If $\theta$ has more than one dimension, the extension is straightforward and we obtain a *credible region*.

In three dimensions, it will be a region

Simple example where Θ can either be 1 or 0 because any data that it generates (i.e. from the generative model/likelihood) will be 0 or 1

*Example 2.1.* Let $X \sim \text{Bernoulli}(\theta)$ and we have observed data $\mathcal{D}_n = \{x_1, \dots, x_n\}$. Suppose we take the uniform distribution $\pi(\theta) = 1$ as a prior. By Bayes' theorem, the posterior is

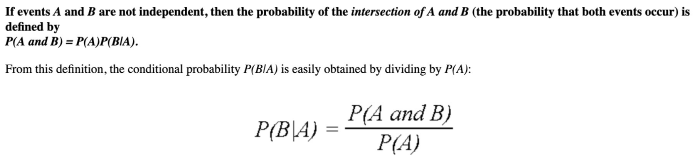
$$p(\theta \mid \mathcal{D}_n) \propto \pi(\theta)\mathcal{L}_n(\theta) = \theta^s(1 - \theta)^{n-s} = \theta^{s+1-1}(1 - \theta)^{n-s+1-1}$$

p(x == 1 | Θ) === p(x and Θ) / p(Θ)

where $s = \sum_{i=1}^n x_i$ is the number of successes. Recall that a random variable $\theta$ on the interval $(0, 1)$ has a Beta distribution with parameters $\alpha$ and $\beta$ if its density is

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

Conditional probability



**If events *A* and *B* are not independent, then the probability of the *intersection of A and B* (the probability that both events occur) is defined by**
***P(A and B) = P(A)P(B|A).***

From this definition, the conditional probability *P(B|A)* is easily obtained by dividing by *P(A)*:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

So the posterior is a beta distribution in this cas

We see that the posterior distribution for $\theta$ is a Beta distribution with parameters $s+1$ and $n-s+1$. That is,

This weird symbol is called the Gamma function

$$p(\theta \mid \mathcal{D}_n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}\theta^{(s+1)-1}(1-\theta)^{(n-s+1)-1}.$$

We write this as

$$\theta \mid \mathcal{D}_n \sim \mathrm{Beta}(s+1, n-s+1).$$

Notice that we have figured out the normalizing constant without actually doing the integral $\int \mathcal{L}_n(\theta)\pi(\theta)\, d\theta$. Since a density function integrates to one, we see that

$$\int_0^1 \theta^s(1-\theta)^{n-s} = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)}. \tag{3}$$

The mean of a $\mathrm{Beta}(\alpha,\beta)$ distribution is $\alpha/(\alpha+\beta)$ so the Bayes posterior estimator is

$$\overline{\theta} = \frac{s+1}{n+2}. \tag{4}$$

It is instructive to rewrite $\overline{\theta}$ as

$$\overline{\theta} = \lambda_n\widehat{\theta} + (1-\lambda_n)\tilde{\theta} \tag{5}$$

where $\widehat{\theta} = s/n$ is the maximum likelihood estimate, $\tilde{\theta} = 1/2$ is the prior mean and $\lambda_n = n/(n+2) \approx 1$. A 95 percent posterior interval can be obtained by numerically finding $a$ and $b$ such that $\int_a^b p(\theta \mid \mathcal{D}_n)\, d\theta = .95$.

But you can also have a beta prior instead of a uniform prior distribution

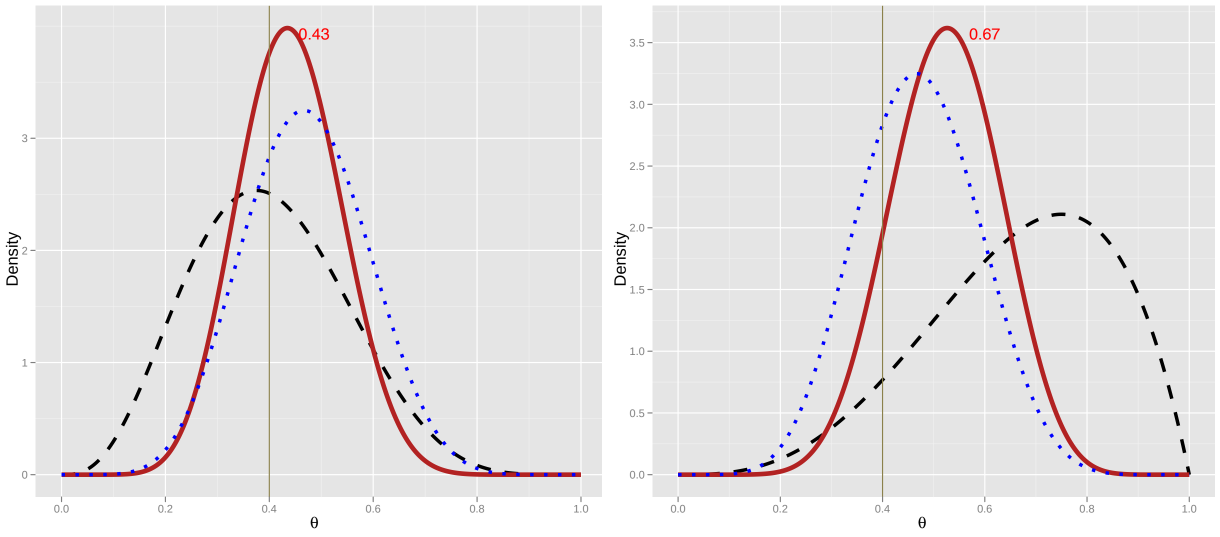
Suppose that instead of a uniform prior, we use the prior $\theta \sim \mathrm{Beta}(\alpha,\beta)$. If you repeat the calculations above, you will see that $\theta \mid \mathcal{D}_n \sim \mathrm{Beta}(\alpha+s, \beta+n-s)$. The flat prior is just the special case with $\alpha = \beta = 1$. The posterior mean in this more general case is

$$\overline{\theta} = \frac{\alpha+s}{\alpha+\beta+n} = \left(\frac{n}{\alpha+\beta+n}\right)\widehat{\theta} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\theta_0$$

where $\theta_0 = \alpha/(\alpha+\beta)$ is the prior mean.

An illustration of this example is shown in Figure 1. We use the Bernoulli model to generate $n = 15$ data with parameter $\theta = 0.4$. We observe $s = 7$. Therefore, the maximum likelihood estimate is $\widehat{\theta} = 7/15 = 0.47$, which is larger than the true parameter value $0.4$. The left plot of Figure 1 adopts a prior $\mathrm{Beta}(4,6)$ which gives a posterior mode 0.43, while the right plot of Figure 1 adopts a prior $\mathrm{Beta}(4,2)$ which gives a posterior mode 0.67.

*Example 2.2.* Let $\theta = (\theta_1, \ldots, \theta_K)$ be a $K$-dimensional parameter ($K > 1$). The multinomial model with a Dirichlet prior is a generalization of the Bernoulli model and Beta prior of the previous example. The Dirichlet distribution for $K$ outcomes is the exponential family distribution on



Your initial prior was good, so when you did the experiment and updated your posterior, you got close to the true value of 0.4

Since your initial prior (black-dashed) was bad, after conducting the experiment, your posterior / prediction got closer to the real answer of 0.4 but not by that much (still around 0.63)

Figure 1: Illustration of Bayesian inference on Bernoulli data with two priors. The three curves are prior distribution (black-dashed), likelihood function (blue-dotted), and the posterior distribution (red-solid). The true parameter value $\theta = 0.4$ is indicated by the vertical line.

the $K-1$ dimensional probability simplex[1] $\Delta_K$ given by

$$\pi_\alpha(\theta) \quad = \quad \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K \theta_j^{\alpha_j - 1} \tag{6}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ is a non-negative vector of scaling coefficients, which are the parameters of the model. We can think of the sample space of the multinomial with $K$ outcomes as the set of vertices of the $K$-dimensional hypercube $\mathbb{H}_K$, made up of vectors with exactly one $1$ and the remaining elements $0$:

$$x = \underbrace{(0, 0, \ldots, 0, 1, 0, \ldots, 0)}_{K \text{ places}}. \tag{7}$$

Let $x_i = (x_{i1}, \ldots, x_{iK}) \in \mathbb{H}_K$. If

$$\theta \sim \text{Dirichlet}(\alpha) \quad \text{and} \quad x_i \,|\, \theta \sim \text{Multinomial}(\theta) \;\; \text{for } i = 1, 2, \ldots, n \tag{8}$$

then the posterior satisfies:

$$p(\theta \,|\, x_1, \ldots, x_n) \propto \mathcal{L}_n(\theta)\, \pi(\theta) \propto \prod_{i=1}^n \prod_{j=1}^K \theta_j^{x_{ij}} \prod_{j=1}^K \theta_j^{\alpha_j - 1} = \prod_{j=1}^K \theta_j^{\sum_{i=1}^n x_{ij} + \alpha_j - 1}. \tag{9}$$

[1] The probability simplex $\Delta_K$ is defined as

$$\Delta_K = \left\{ \theta = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K \,|\, \theta_i \geq 0 \text{ for all } \; i \text{ and } \sum_{i=1}^K \theta_i = 1 \right\}.$$

We see that the posterior is also a Dirichlet distribution:

$$\theta \,|\, x_1, x_2, \ldots, x_n \quad \sim \quad \text{Dirichlet}(\alpha + n\overline{x}) \tag{10}$$

where $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \in \Delta_K$.

Since the mean of a Dirichlet distribution $\pi_\alpha(\theta)$ is given by

$$\mathbb{E}(\theta) = \left( \frac{\alpha_1}{\sum_{i=1}^{K} \alpha_i}, \ldots, \frac{\alpha_K}{\sum_{i=1}^{K} \alpha_i} \right), \tag{11}$$

the posterior mean of a multinomial with Dirichlet prior is

$$\mathbb{E}(\theta \,|\, x_1, \ldots, x_n) = \left( \frac{\alpha_1 + \sum_{i=1}^{n} x_{i1}}{\sum_{i=1}^{K} \alpha_i + n}, \ldots, \frac{\alpha_K + \sum_{i=1}^{n} x_{iK}}{\sum_{i=1}^{K} \alpha_i + n} \right). \tag{12}$$

This again can be viewed as smoothing out the maximum likelihood estimate by allocating some additional probability mass to low frequency observations. The parameters $\alpha_1, \ldots, \alpha_K$ act as "virtual counts" that don't actually appear in the observed data.

An illustration of this example is shown in Figure 2. We use the multinomial model to generate $n = 20$ data points with parameter $\theta = (0.2, 0.3, 0.5)$. We adopt a prior $\text{Dirichlet}(6, 6, 6)$. The contours of the prior, likelihood, and posterior with $n = 20$ observed data are shown in the first three plots in Figure 2. As a comparison, we also provide the contour of the posterior with $n = 200$ observed data in the last plot. From this experiment, we see that when the number of observed data is small, the posterior is affected by both the prior and the likelihood; when the number of observed data is large, the posterior is mainly dominated by the likelihood.

In the previous two examples, the prior was a Dirichlet distribution and the posterior was also a Dirichlet. When the prior and the posterior are in the same family, we say that the prior is *conjugate* with respect to the model; this will be discussed further below.

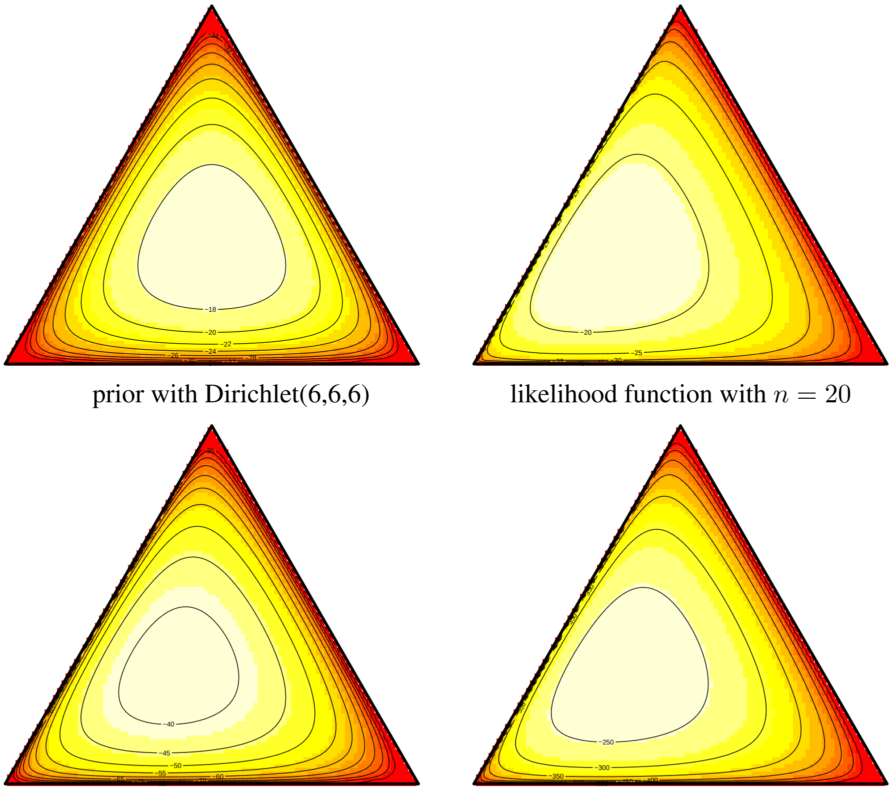
*Example 2.3.* Let $X \sim N(\theta, \sigma^2)$ and $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be the observed data. For simplicity, let us assume that $\sigma$ is known and we want to estimate $\theta \in \mathbb{R}$. Suppose we take as a prior $\theta \sim N(a, b^2)$. Let $\overline{x} = \sum_{i=1}^{n} x_i / n$ be the sample mean. It can be shown that the posterior for $\theta$ is

$$\theta \,|\, \mathcal{D}_n \sim N(\overline{\theta}, \tau^2) \tag{13}$$

where

$$\overline{\theta} = w\widehat{\theta} + (1 - w)a,$$

$$\widehat{\theta} = \overline{x}, \quad w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2},$$



prior with Dirichlet(6,6,6)

likelihood function with $n = 20$

posterior distribution with $n = 20$

posterior distribution with $n = 200$

Figure 2: Illustration of Bayesian inference on multinomial data with the prior $\mathrm{Dirichlet}(6, 6, 6)$. The contours of the prior, likelihood, and posteriors are plotted on a two-dimensional probability simplex (Starting from the bottom left vertex of each triangle, clock-wisely the three vertices correspond to $\theta_1, \theta_2, \theta_3$). We see that when the number of observed data is small, the posterior is affected by both the prior and the likelihood; when the number of observed data is large, the posterior is mainly dominated by the likelihood.

and $se = \sigma/\sqrt{n}$ is the standard error of the maximum likelihood estimate $\widehat{\theta}$. This is another example of a conjugate prior. Note that $w \to 1$ and $\tau/se \to 1$ as $n \to \infty$. So, for large $n$, the posterior is approximately $N(\widehat{\theta}, se^2)$. The same is true if $n$ is fixed but $b \to \infty$, which corresponds to letting the prior become very flat.

Continuing with this example, let us find $C = (c, d)$ such that $\mathbb{P}(\theta \in C \,|\, \mathcal{D}_n) = 0.95$. We can do this by choosing $c$ and $d$ such that $\mathbb{P}(\theta < c \,|\, \mathcal{D}_n) = 0.025$ and $\mathbb{P}(\theta > d \,|\, \mathcal{D}_n) = 0.025$. More specifically, we want to find $c$ such that

$$\mathbb{P}(\theta < c \,|\, \mathcal{D}_n) = \mathbb{P}\left(\frac{\theta - \overline{\theta}}{\tau} < \frac{c - \overline{\theta}}{\tau} \,\middle|\, \mathcal{D}_n\right) = \mathbb{P}\left(Z < \frac{c - \overline{\theta}}{\tau}\right) = 0.025$$

where $Z \sim N(0, 1)$ is a standard Gaussian random variable. We know that $\mathbb{P}(Z < -1.96) =$

0.025. So,

$$\frac{c - \overline{\theta}}{\tau} = -1.96$$

implying that $c = \overline{\theta} - 1.96\tau$. By similar arguments, $d = \overline{\theta} + 1.96\tau$. So a 95 percent Bayesian credible interval is $\overline{\theta} \pm 1.96\,\tau$. Since $\overline{\theta} \approx \widehat{\theta}$ and $\tau \approx se$ when $n$ is large, the 95 percent Bayesian credible interval is approximated by $\widehat{\theta} \pm 1.96\ se$ which is the frequentist confidence interval.

# 3. Bayesian Prediction

After the data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ have been observed, the Bayesian framework allows us to predict the distribution of a future data point $x$ conditioned on $\mathcal{D}_n$. To do this, we first obtain the posterior $p(\theta \,|\, \mathcal{D}_n)$. Then

$$\begin{aligned} p(x \,|\, \mathcal{D}_n) &= \int p(x, \theta \,|\, \mathcal{D}_n) d\theta & (14) \\ &= \int p(x \,|\, \theta, \mathcal{D}_n) p(\theta \,|\, \mathcal{D}_n) d\theta & (15) \\ &= \int p(x \,|\, \theta) p(\theta \,|\, \mathcal{D}_n) d\theta. & (16) \end{aligned}$$

Where we use the fact that $p(x \,|\, \theta, \mathcal{D}_n) = p(x \,|\, \theta)$ since all the data are conditionally independent given $\theta$. From the last line, the predictive distribution $p(x \,|\, \mathcal{D}_n)$ can be viewed as a weighted average of likelihood $p(x \,|\, \theta)$. The weights are determined by the posterior distribution of $\theta$.

# 4 Multiparameter Problems

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be the observed data. Suppose that $\theta = (\theta_1, \ldots, \theta_d)$ with some prior distribution $\pi(\theta)$. The posterior density is still given by

$$p(\theta \,|\, \mathcal{D}_n) \propto \mathcal{L}_n(\theta)\pi(\theta). \tag{17}$$

The question now arises of how to extract inferences about one single parameter. The key is to find the marginal posterior density for the parameter of interest. Suppose we want to make inferences about $\theta_1$. The marginal posterior for $\theta_1$ is

$$p(\theta_1 \,|\, \mathcal{D}_n) = \int \cdots \int p(\theta_1, \cdots, \theta_d \,|\, \mathcal{D}_n) d\theta_2 \ldots d\theta_d. \tag{18}$$

In practice, it might not be feasible to do this integral. Simulation can help: we draw randomly from the posterior:

$$\theta^1, \ldots, \theta^B \sim p(\theta \,|\, \mathcal{D}_n)$$

where the superscripts index different draws. Each $\theta^j$ is a vector $\theta^j = (\theta_1^j, \ldots, \theta_d^j)$. Now collect together the first component of each draw: $\theta_1^1, \ldots, \theta_1^B$. These are a sample from $p(\theta_1 \,|\, \mathcal{D}_n)$ and we have avoided doing any integrals. One thing to note is, sampling $B$ data from a multivariate distribution $p(\theta \,|\, \mathcal{D}_n)$ is challenging especially when the dimensionality $d$ is large. We will discuss this topic further in the chapter on computing.

*Example 4.1.* (Comparing Two Binomials) Suppose we have $n_1$ control patients and $n_2$ treatment patients and that $x_1$ control patients survive while $x_2$ treatment patients survive. We assume the Binomial model:

$$X_1 \sim \text{Binomial}(n_1, \theta_1) \;\text{ and }\; X_2 \sim \text{Binomial}(n_2, \theta_2).$$

We want to estimate $\tau = g(\theta_1, \theta_2) = \theta_2 - \theta_1$.

If $\pi(\theta_1, \theta_2) = 1$, the posterior is

$$p(\theta_1, \theta_2 \,|\, x_1, x_2) \propto \theta_1^{x_1}(1-\theta_1)^{n_1 - x_1}\theta_2^{x_2}(1-\theta_2)^{n_2-x_2}.$$

Notice that $(\theta_1, \theta_2)$ live on a rectangle (a square, actually) and that

$$p(\theta_1, \theta_2 \,|\, x_1, x_2) = p(\theta_1 \,|\, x_1)p(\theta_2 \,|\, x_2)$$

where

$$p(\theta_1 \,|\, x_1) \propto \theta_1^{x_1}(1-\theta_1)^{n_1-x_1} \;\text{ and }\; p(\theta_2 \,|\, x_2) \propto \theta_2^{x_2}(1-\theta_2)^{n_2-x_2}$$

which implies that $\theta_1$ and $\theta_2$ are independent under the posterior. Also, $\theta_1 \,|\, x_1 \sim \text{Beta}(x_1+1, n_1 - x_1 + 1)$ and $\theta_2 \,|\, x_2 \sim \text{Beta}(x_2+1, n_2 - x_2 + 1)$. If we simulate $\Theta_1^1, \ldots, \Theta_1^B \sim \text{Beta}(x_1+1, n_1 - x_1 + 1)$ and $\Theta_2^1, \ldots, \Theta_2^B \sim \text{Beta}(x_2+1, n_2-x_2+1)$, then $\tau_b = \Theta_2^b - \Theta_1^b$, $b = 1, \ldots, B$, is a sample from $p(\tau \,|\, x_1, x_2)$.

# 5. Simulation

Since the posterior distribution $p(\theta \,|\, \mathcal{D}_n)$ generally involves high dimensional integrals, it is generally approximated by simulation. Suppose we draw $\theta^1, \ldots, \theta^B \sim p(\theta \,|\, \mathcal{D}_n)$. Then a histogram of $\theta^1, \ldots, \theta^B$ approximates the posterior density $p(\theta \,|\, \mathcal{D}_n)$. An approximation to the posterior mean $\overline{\theta}_n = \mathbb{E}(\theta \,|\, \mathcal{D}_n)$ is $B^{-1}\sum_{j=1}^B \theta^j$. The posterior $1-\alpha$ interval can be approximated by $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$ where $\theta_{\alpha/2}$ is the $\alpha/2$ sample quantile of $\theta^1, \ldots, \theta^B$.

Once we have a sample $\theta^1, \ldots, \theta^B$ from $p(\theta \,|\, \mathcal{D}_n)$, let $\tau^i = g(\theta^i)$. Then $\tau^1, \ldots, \tau^B$ is a sample from $p(\tau \,|\, \mathcal{D}_n)$. This avoids the need to do any analytical calculations. Simulation techniques are discussed in more detail in a later chapter on statistical computing.

*Example 5.1.* Consider again Example **??**. We can approximate the posterior for $\psi$ without doing any calculus. Here are the steps:

1. Draw $\theta^1, \ldots, \theta^B \sim \mathrm{Beta}(s+1, n-s+1)$.
2. Let $\psi^i = \log(\theta^i/(1-\theta^i))$ for $i = 1, \ldots, B$.

Now $\psi^1, \ldots, \psi^B$ are i.i.d. draws from the posterior density $p(\psi \,|\, \mathcal{D}_n)$. A histogram of these values provides an estimate of $p(\psi \,|\, \mathcal{D}_n)$.

# 6. Bayesian Linear Models

Many frequentist methods can be viewed as the *maximum a posterior* (MAP) estimator under a Bayesian framework. As an example, we consider Gaussian linear regression:

$$Y = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon, \;\; \epsilon \sim N(0, \sigma^2). \tag{19}$$

Here we assume that $\sigma$ is known. Let $\mathcal{D}_n = \big\{(x_1, y_1), \ldots, (x_n, y_n)\big\}$ be the observed data points. The conditional likelihood of $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$ can be written as

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p(y_i \,|\, x_i, \beta) \propto \exp\bigg(-\frac{\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{d}\beta_j x_{ij}\Big)^2}{2\sigma^2}\bigg). \tag{20}$$

Using a Gaussian prior $\pi_\lambda(\beta) \propto \exp\left(-\lambda\|\beta\|_2^2/2\right)$, the posterior of $\beta$ can be written as

$$p(\beta \,|\, \mathcal{D}_n) \propto \mathcal{L}(\beta)\pi_\lambda(\beta). \tag{21}$$

The MAP estimator $\widehat{\beta}^{\mathrm{MAP}}$ takes the form

$$\widehat{\beta}^{\mathrm{MAP}} = \arg\max_{\beta} p(\beta \,|\, \mathcal{D}_n) = \arg\min_{\beta}\bigg\{\sum_{i=1}^{n}\big(y_i - \beta_0 - \sum_{j=1}^{d}\beta_j x_{ij}\big)^2 + \lambda\sigma^2\|\beta\|_2^2\bigg\}. \tag{22}$$

This is exactly the ridge regression with the regularization parameter $\lambda' = \lambda\sigma^2$. If we adopt the Laplacian prior $\pi_\lambda(\beta) \propto \exp\left(-\lambda\|\beta\|_1/2\right)$, we get the Lasso estimator

$$\widehat{\beta}^{\mathrm{MAP}} = \arg\min_{\beta}\bigg\{\sum_{i=1}^{n}\big(y_i - \beta_0 - \sum_{j=1}^{d}\beta_j x_{ij}\big)^2 + \lambda\sigma^2\|\beta\|_1\bigg\}. \tag{23}$$

Instead of using the MAP point estimate, a full Bayesian inference aims at obtaining the whole posterior distribution $p(\beta \,|\, \mathcal{D}_n)$. In general, $p(\beta \,|\, \mathcal{D}_n)$ does not have an analytic form and we need to resort to simulation to approximate the posterior.