11
Bayesian Inference

11.1 The Bayesian Philosophy

The statistical methods that we have discussed so far are known as **frequentist (or classical)** methods. The frequentist point of view is based on the following postulates:

F1 Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.

F2 Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.

F3 Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

There is another approach to inference called **Bayesian inference.** The Bayesian approach is based on the following postulates:

L. Wasserman, *All of Statistics*

B1 Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is .35. This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.

B2 We can make probability statements about parameters, even though they are fixed constants.

B3 We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian inference is a controversial approach because it inherently embraces a subjective notion of probability. In general, Bayesian methods provide no guarantees on long run performance. The field of statistics puts more emphasis on frequentist methods although Bayesian methods certainly have a presence. Certain data mining and machine learning communities seem to embrace Bayesian methods very strongly. Let's put aside philosophical arguments for now and see how Bayesian inference is done. We'll conclude this chapter with some discussion on the strengths and weaknesses of the Bayesian approach.

11.2 The Bayesian Method

Bayesian inference is usually carried out in the following way.

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

2. We choose a statistical model $f(x|\theta)$ that reflects our beliefs about x given θ. Notice that we now write this as $f(x|\theta)$ instead of $f(x;\theta)$.

3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the **posterior** distribution $f(\theta|X_1, \ldots, X_n)$.

To see how the third step is carried out, first suppose that θ is discrete and that there is a single, discrete observation X. We should use a capital letter

now to denote the parameter since we are treating it like a random variable, so let Θ denote the parameter. Now, in this discrete setting,

$$\begin{aligned}\mathbb{P}(\Theta = \theta | X = x) &= \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}\end{aligned}$$

which you may recognize from Chapter 1 as **Bayes' theorem.** The version for continuous variables is obtained by using density functions:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \tag{11.1}$$

If we have n IID observations $X_1, \ldots, X_n$, we replace $f(x|\theta)$ with

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) = \mathcal{L}_n(\theta).$$

NOTATION. We will write X^n to mean $(X_1, \ldots, X_n)$ and x^n to mean $(x_1, \ldots, x_n)$. Now,

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta) \tag{11.2}$$

where

$$c_n = \int \mathcal{L}_n(\theta)f(\theta)d\theta \tag{11.3}$$

is called the **normalizing constant**. Note that c_n does not depend on θ. We can summarize by writing:

> Posterior is proportional to Likelihood times Prior
>
> or, in symbols,
>
> $$f(\theta|x^n) \propto \mathcal{L}(\theta)f(\theta).$$

You might wonder, doesn't it cause a problem to throw away the constant c_n? The answer is that we can always recover the constant later if we need to.

What do we do with the posterior distribution? First, we can get a point estimate by summarizing the center of the posterior. Typically, we use the mean or mode of the posterior. The posterior mean is

$$\overline{\theta}_n = \int \theta f(\theta|x^n)d\theta = \frac{\int \theta\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta}. \tag{11.4}$$

We can also obtain a Bayesian interval estimate. We find a and b such that $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$. Let $C = (a, b)$. Then

$$\mathbb{P}(\theta \in C|x^n) = \int_a^b f(\theta|x^n)\, d\theta = 1 - \alpha$$

so C **is a** $1 - \alpha$ **posterior interval.**

11.1 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Suppose we take the uniform distribution $f(p) = 1$ as a prior. By Bayes' theorem, the posterior has the form

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{s+1-1}(1-p)^{n-s+1-1}$$

where $s = \sum_{i=1}^n x_i$ is the number of successes. Recall that a random variable has a Beta distribution with parameters α and β if its density is

$$f(p;\, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

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

$$f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^{(s+1)-1}(1-p)^{(n-s+1)-1}.$$

We write this as

$$p|x^n \sim \text{Beta}(s+1, n-s+1).$$

Notice that we have figured out the normalizing constant without actually doing the integral $\int \mathcal{L}_n(p)f(p)dp$. The mean of a $\text{Beta}(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$ so the Bayes estimator is

$$\overline{p} = \frac{s+1}{n+2}. \tag{11.5}$$

It is instructive to rewrite the estimator as

$$\overline{p} = \lambda_n \widehat{p} + (1 - \lambda_n)\widetilde{p} \tag{11.6}$$

where $\widehat{p} = s/n$ is the MLE, $\widetilde{p} = 1/2$ 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 $\int_a^b f(p|x^n)\, dp = .95$.

Suppose that instead of a uniform prior, we use the prior $p \sim \text{Beta}(\alpha, \beta)$. If you repeat the calculations above, you will see that $p|x^n \sim \text{Beta}(\alpha + s, \beta +$

$n - s$). The flat prior is just the special case with $\alpha = \beta = 1$. The posterior mean is

$$\overline{p} = \frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{n}{\alpha + \beta + n}\right)\widehat{p} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)p_0$$

where $p_0 = \alpha/(\alpha + \beta)$ is the prior mean. ■

In the previous example, the prior was a Beta distribution and the posterior was a Beta distribution. When the prior and the posterior are in the same family, we say that the prior is **conjugate** with respect to the model.

11.2 Example. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. For simplicity, let us assume that σ is known. Suppose we take as a prior $\theta \sim N(a, b^2)$. In problem 1 in the exercises it is shown that the posterior for θ is

$$\theta | X^n \sim N(\overline{\theta}, \tau^2) \tag{11.7}$$

where

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

$$w = \frac{\frac{1}{\mathsf{se}^2}}{\frac{1}{\mathsf{se}^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\mathsf{se}^2} + \frac{1}{b^2},$$

and $\mathsf{se} = \sigma/\sqrt{n}$ is the standard error of the MLE $\overline{X}$. This is another example of a conjugate prior. Note that $w \to 1$ and $\tau/\mathsf{se} \to 1$ as $n \to \infty$. So, for large n, the posterior is approximately $N(\widehat{\theta}, \mathsf{se}^2)$. 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 $\mathbb{P}(\theta \in C | X^n) = .95$. We can do this by choosing c and d such that $\mathbb{P}(\theta < c | X^n) = .025$ and $\mathbb{P}(\theta > d | X^n) = .025$. So, we want to find c such that

$$\begin{aligned} \mathbb{P}(\theta < c | X^n) &= \mathbb{P}\left(\frac{\theta - \overline{\theta}}{\tau} < \frac{c - \overline{\theta}}{\tau} \,\middle|\, X^n\right) \\ &= \mathbb{P}\left(Z < \frac{c - \overline{\theta}}{\tau}\right) = .025. \end{aligned}$$

We know that $\mathbb{P}(Z < -1.96) = .025$. So,

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

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

11.3 Functions of Parameters

How do we make inferences about a function $\tau = g(\theta)$? Remember in Chapter 3 we solved the following problem: given the density f_X for X, find the density for $Y = g(X)$. We now simply apply the same reasoning. The posterior CDF for τ is

$$H(\tau|x^n) = \mathbb{P}(g(\theta) \leq \tau|x^n) = \int_A f(\theta|x^n)d\theta$$

where $A = \{\theta : \ g(\theta) \leq \tau\}$. The posterior density is $h(\tau|x^n) = H'(\tau|x^n)$.

11.3 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $f(p) = 1$ so that $p|X^n \sim \text{Beta}(s+1, n-s+1)$ with $s = \sum_{i=1}^n x_i$. Let $\psi = \log(p/(1-p))$. Then

$$\begin{aligned} H(\psi|x^n) &= \mathbb{P}(\Psi \leq \psi|x^n) = \mathbb{P}\left(\log\left(\frac{P}{1-P}\right) \leq \psi \ \middle| \ x^n\right) \\ &= \mathbb{P}\left(P \leq \frac{e^\psi}{1+e^\psi} \ \middle| \ x^n\right) \\ &= \int_0^{e^\psi/(1+e^\psi)} f(p|x^n)\, dp \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^{e^\psi/(1+e^\psi)} p^s(1-p)^{n-s}\, dp \end{aligned}$$

and

$$\begin{aligned} h(\psi|x^n) &= H'(\psi|x^n) \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{\partial\left(\frac{e^\psi}{1+e^\psi}\right)}{\partial\psi}\right) \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{1}{1+e^\psi}\right)^2 \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s+2} \end{aligned}$$

for $\psi \in \mathbb{R}$. ■

11.4 Simulation

The posterior can often be approximated by simulation. Suppose we draw $\theta_1, \ldots, \theta_B \sim p(\theta|x^n)$. Then a histogram of $\theta_1, \ldots, \theta_B$ approximates the posterior density $p(\theta|x^n)$. An approximation to the posterior mean $\overline{\theta}_n = \mathbb{E}(\theta|x^n)$ 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 $f(\theta|x^n)$, let $\tau_i = g(\theta_i)$. Then $\tau_1, \ldots, \tau_B$ is a sample from $f(\tau|x^n)$. This avoids the need to do any analytical calculations. Simulation is discussed in more detail in Chapter 24.

11.4 Example. Consider again Example 11.3. We can approximate the posterior for ψ without doing any calculus. Here are the steps:

1. Draw $P_1, \ldots, P_B \sim \text{Beta}(s+1, n-s+1)$.
2. Let $\psi_i = \log(P_i/(1-P_i))$ for $i = 1, \ldots, B$.

Now $\psi_1, \ldots, \psi_B$ are IID draws from $h(\psi|x^n)$. A histogram of these values provides an estimate of $h(\psi|x^n)$. ∎

11.5 Large Sample Properties of Bayes' Procedures

In the Bernoulli and Normal examples we saw that the posterior mean was close to the MLE. This is true in greater generality.

11.5 Theorem. *Let $\widehat{\theta}_n$ be the* MLE *and let $\widehat{\mathsf{se}} = 1/\sqrt{nI(\widehat{\theta}_n)}$. Under appropriate regularity conditions, the posterior is approximately Normal with mean $\widehat{\theta}_n$ and standard deviation $\widehat{\mathsf{se}}$. Hence, $\overline{\theta}_n \approx \widehat{\theta}_n$. Also, if $C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{\mathsf{se}}, \widehat{\theta}_n + z_{\alpha/2}\widehat{\mathsf{se}})$ is the asymptotic frequentist $1-\alpha$ confidence interval, then C_n is also an approximate $1-\alpha$ Bayesian posterior interval:*

$$\mathbb{P}(\theta \in C_n | X^n) \to 1-\alpha.$$

There is also a Bayesian delta method. Let $\tau = g(\theta)$. Then

$$\tau|X^n \approx N(\widehat{\tau}, \widetilde{\mathsf{se}}^2)$$

where $\widehat{\tau} = g(\widehat{\theta})$ and $\widetilde{\mathsf{se}} = \widehat{\mathsf{se}}\,|g'(\widehat{\theta})|$.

11.6 Flat Priors, Improper Priors, and "Noninformative" Priors

An important question in Bayesian inference is: where does one get the prior $f(\theta)$? One school of thought, called **subjectivism** says that the prior should

reflect our subjective opinion about θ (before the data are collected). This may be possible in some cases but is impractical in complicated problems especially if there are many parameters. Moreover, injecting subjective opinion into the analysis is contrary to the goal of making scientific inference as objective as possible. An alternative is to try to define some sort of "noninformative prior." An obvious candidate for a noninformative prior is to use a flat prior $f(\theta) \propto$ constant.

In the Bernoulli example, taking $f(p) = 1$ leads to $p|X^n \sim \text{Beta}(s+1, n-s+1)$ as we saw earlier, which seemed very reasonable. But unfettered use of flat priors raises some questions.

Improper Priors. Let $X \sim N(\theta, \sigma^2)$ with σ known. Suppose we adopt a flat prior $f(\theta) \propto c$ where $c > 0$ is a constant. Note that $\int f(\theta)d\theta = \infty$ so this is not a probability density in the usual sense. We call such a prior an **improper prior.** Nonetheless, we can still formally carry out Bayes' theorem and compute the posterior density by multiplying the prior and the likelihood: $f(\theta) \propto \mathcal{L}_n(\theta)f(\theta) \propto \mathcal{L}_n(\theta)$. This gives $\theta|X^n \sim N(\overline{X}, \sigma^2/n)$ and the resulting point and interval estimators agree exactly with their frequentist counterparts. In general, improper priors are not a problem as long as the resulting posterior is a well-defined probability distribution.

Flat Priors are Not Invariant. Let $X \sim \text{Bernoulli}(p)$ and suppose we use the flat prior $f(p) = 1$. This flat prior presumably represents our lack of information about p before the experiment. Now let $\psi = \log(p/(1-p))$. This is a transformation of p and we can compute the resulting distribution for ψ, namely,

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}$$

which is not flat. But if we are ignorant about p then we are also ignorant about ψ so we should use a flat prior for ψ. This is a contradiction. In short, the notion of a flat prior is not well defined because a flat prior on a parameter does not imply a flat prior on a transformed version of the parameter. Flat priors are not **transformation invariant.**

Jeffreys' Prior. Jeffreys came up with a rule for creating priors. The rule is: take

$$f(\theta) \propto I(\theta)^{1/2}$$

where $I(\theta)$ is the Fisher information function. This rule turns out to be transformation invariant. There are various reasons for thinking that this prior might be a useful prior but we will not go into details here.

11.6 Example. Consider the Bernoulli (p) model. Recall that

$$I(p) = \frac{1}{p(1-p)}.$$

Jeffreys' rule says to use the prior

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}.$$

This is a Beta (1/2,1/2) density. This is very close to a uniform density. ∎

In a multiparameter problem, the Jeffreys' prior is defined to be $f(\theta) \propto \sqrt{|I(\theta)|}$ where $|A|$ denotes the determinant of a matrix A and $I(\theta)$ is the Fisher information matrix.

11.7 Multiparameter Problems

Suppose that $\theta = (\theta_1, \ldots, \theta_p)$. The posterior density is still given by

$$f(\theta|x^n) \propto \mathcal{L}_n(\theta) f(\theta). \tag{11.8}$$

The question now arises of how to extract inferences about one parameter. The key is to find the marginal posterior density for the parameter of interest. Suppose we want to make inferences about θ_1. The marginal posterior for θ_1 is

$$f(\theta_1|x^n) = \int \cdots \int f(\theta_1, \cdots, \theta_p|x^n) d\theta_2 \ldots d\theta_p. \tag{11.9}$$

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

$$\theta^1, \ldots, \theta^B \sim f(\theta|x^n)$$

where the superscripts index the different draws. Each θ^j is a vector $\theta^j = (\theta_1^j, \ldots, \theta_p^j)$. Now collect together the first component of each draw:

$$\theta_1^1, \ldots, \theta_1^B.$$

These are a sample from $f(\theta_1|x^n)$ and we have avoided doing any integrals.

11.7 Example (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 want to estimate $\tau = g(p_1, p_2) = p_2 - p_1$. Then,

$$X_1 \sim \text{Binomial}(n_1, p_1) \quad \text{and} \quad X_2 \sim \text{Binomial}(n_2, p_2).$$

If $f(p_1, p_2) = 1$, the posterior is

$$f(p_1, p_2|x_1, x_2) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2}.$$

Notice that (p_1, p_2) live on a rectangle (a square, actually) and that

$$f(p_1, p_2|x_1, x_2) = f(p_1|x_1) f(p_2|x_2)$$

where

$$f(p_1|x_1) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} \quad \text{and} \quad f(p_2|x_2) \propto p_2^{x_2}(1-p_2)^{n_2-x_2}$$

which implies that p_1 and p_2 are independent under the posterior. Also, $p_1|x_1 \sim \text{Beta}(x_1+1, n_1-x_1+1)$ and $p_2|x_2 \sim \text{Beta}(x_2+1, n_2-x_2+1)$. If we simulate $P_{1,1}, \ldots, P_{1,B} \sim \text{Beta}(x_1+1, n_1-x_1+1)$ and $P_{2,1}, \ldots, P_{2,B} \sim \text{Beta}(x_2+1, n_2-x_2+1)$, then $\tau_b = P_{2,b} - P_{1,b}$, $b = 1, \ldots, B$, is a sample from $f(\tau|x_1, x_2)$. ■

11.8 Bayesian Testing

Hypothesis testing from a Bayesian point of view is a complex topic. We will only give a brief sketch of the main idea here. The Bayesian approach to testing involves putting a prior on H_0 and on the parameter θ and then computing $\mathbb{P}(H_0|X^n)$. Consider the case where θ is scalar and we are testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

It is usually reasonable to use the prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ (although this is not essential in what follows). Under H_1 we need a prior for θ. Denote this prior density by $f(\theta)$. From Bayes' theorem

$$\begin{aligned}
\mathbb{P}(H_0|X^n = x^n) &= \frac{f(x^n|H_0)\mathbb{P}(H_0)}{f(x^n|H_0)\mathbb{P}(H_0) + f(x^n|H_1)\mathbb{P}(H_1)} \\
&= \frac{\frac{1}{2}f(x^n \mid \theta_0)}{\frac{1}{2}f(x^n \mid \theta_0) + \frac{1}{2}f(x^n \mid H_1)} \\
&= \frac{f(x^n \mid \theta_0)}{f(x^n \mid \theta_0) + \int f(x^n \mid \theta) f(\theta) d\theta} \\
&= \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta) f(\theta) d\theta}.
\end{aligned}$$

We saw that, in estimation problems, the prior was not very influential and that the frequentist and Bayesian methods gave similar answers. This is not

the case in hypothesis testing. Also, one can't use improper priors in testing because this leads to an undefined constant in the denominator of the expression above. Thus, if you use Bayesian testing you must choose the prior $f(\theta)$ very carefully. It is possible to get a prior-free bound on $\mathbb{P}(H_0|X^n = x^n)$. Notice that $0 \leq \int \mathcal{L}(\theta)f(\theta)d\theta \leq \mathcal{L}(\widehat{\theta})$. Hence,

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \mathcal{L}(\widehat{\theta})} \ \leq \ \mathbb{P}(H_0|X^n = x^n) \ \leq \ 1.$$

The upper bound is not very interesting, but the lower bound is non-trivial.

11.9 Strengths and Weaknesses of Bayesian Inference

Bayesian inference is appealing when prior information is available since Bayes' theorem is a natural way to combine prior information with data. Some people find Bayesian inference psychologically appealing because it allows us to make probability statements about parameters. In contrast, frequentist inference provides confidence sets C_n which trap the parameter 95 percent of the time, but we cannot say that $\mathbb{P}(\theta \in C_n|X^n)$ is .95. In the frequentist approach we can make probability statements about C_n, not θ. However, psychological appeal is not a compelling scientific argument for using one type of inference over another.

In parametric models, with large samples, Bayesian and frequentist methods give approximately the same inferences. In general, they need not agree.

Here are three examples that illustrate the strengths and weakness of Bayesian inference. The first example is Example 6.14 revisited. This example shows the psychological appeal of Bayesian inference. The second and third show that Bayesian methods can fail.

11.8 Example (Example 6.14 revisited)**.** We begin by reviewing the example. Let θ be a fixed, known real number and let X_1, X_2 be independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Now define $Y_i = \theta + X_i$ and suppose that you only observe Y_1 and Y_2. Let

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2 \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

This is a 75 percent confidence set since, no matter what θ is, $\mathbb{P}_\theta(\theta \in C) = 3/4$.

Suppose we observe $Y_1 = 15$ and $Y_2 = 17$. Then our 75 percent confidence interval is $\{16\}$. However, we are certain, in this case, that $\theta = 16$. So calling

this a 75 percent confidence set, bothers many people. Nonetheless, C is a valid 75 percent confidence set. It will trap the true value 75 percent of the time.

The Bayesian solution is more satisfying to many. For simplicity, assume that θ is an integer. Let $f(\theta)$ be a prior mass function such that $f(\theta) > 0$ for every integer θ. When $Y = (Y_1, Y_2) = (15, 17)$, the likelihood function is

$$\mathcal{L}(\theta) = \begin{cases} 1/4 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Applying Bayes' theorem we see that

$$\mathbb{P}(\Theta = \theta | Y = (15, 17)) = \begin{cases} 1 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\mathbb{P}(\theta \in C | Y = (15, 17)) = 1$. There is nothing wrong with saying that $\{16\}$ is a 75 percent confidence interval. But is it not a probability statement about θ. ■

11.9 Example. This is a simplified version of the example in Robins and Ritov (1997). The data consist of n IID triples

$$(X_1, R_1, Y_1), \ldots, (X_n, Y_n, R_n).$$

Let B be a finite but very large number, like $B = 100^{100}$. Any realistic sample size n will be small compared to B. Let

$$\theta = (\theta_1, \ldots, \theta_B)$$

be a vector of unknown parameters such that $0 \leq \theta_j \leq 1$ for $1 \leq j \leq B$. Let

$$\xi = (\xi_1, \ldots, \xi_B)$$

be a vector of **known** numbers such that

$$0 < \delta \leq \xi_j \leq 1 - \delta < 1, \quad 1 \leq j \leq B,$$

where δ is some, small, positive number. Each data point (X_i, R_i, Y_i) is drawn in the following way:

1. Draw X_i uniformly from $\{1, \ldots, B\}$.
2. Draw $R_i \sim \text{Bernoulli}(\xi_{X_i})$.
3. If $R_i = 1$, then draw $Y_i \sim \text{Bernoulli}(\theta_{X_i})$. If $R_i = 0$, do not draw Y_i.

The model may seem a little artificial but, in fact, it is caricature of some real **missing data** problems in which some data points are not observed. In this example, $R_i = 0$ can be thought of as meaning "missing." Our goal is to estimate

$$\psi = \mathbb{P}(Y_i = 1).$$

Note that

$$\begin{aligned} \psi &= \mathbb{P}(Y_i = 1) = \sum_{j=1}^{B} \mathbb{P}(Y_i = 1 | X = j)\mathbb{P}(X = j) \\ &= \frac{1}{B}\sum_{j=1}^{B} \theta_j \equiv g(\theta) \end{aligned}$$

so $\psi = g(\theta)$ is a function of θ.

Let us consider a Bayesian analysis first. The likelihood of a single observation is

$$f(X_i, R_i, Y_i) = f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i}.$$

The last term is raised to the power R_i since, if $R_i = 0$, then Y_i is not observed and hence that term drops out of the likelihood. Since $f(X_i) = 1/B$ and that Y_i and R_i are Bernoulli,

$$f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} = \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i}.$$

Thus, the likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{n} f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} \\ &= \prod_{i=1}^{n} \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i} \\ &\propto \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i}. \end{aligned}$$

We have dropped all the terms involving B and the ξ_j's since these are known constants, not parameters. The log-likelihood is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} Y_i\, R_i \log \theta_{X_i} + (1 - Y_i)\, R_i\, \log(1 - \theta_{X_i}) \\ &= \sum_{j=1}^{B} n_j \log \theta_j + \sum_{j=1}^{B} m_j \log(1 - \theta_j) \end{aligned}$$

where

$$\begin{aligned} n_j &= \#\{i : \ Y_i = 1, R_i = 1, X_i = j\} \\ m_j &= \#\{i : \ Y_i = 0, R_i = 1, X_i = j\}. \end{aligned}$$

Now, $n_j = m_j = 0$ for most j since B is so much larger than n. This has several implications. First, the MLE for most θ_j is not defined. Second, for most θ_j, the posterior distribution is equal to the prior distribution, since those θ_j do not appear in the likelihood. Hence, $f(\theta|\text{Data}) \approx f(\theta)$. It follows that $f(\psi|\text{Data}) \approx f(\psi)$. In other words, the data provide little information about ψ in a Bayesian analysis.

Now we consider a frequentist solution. Define

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\xi_{X_i}}. \tag{11.10}$$

We will now show that this estimator is unbiased and has small mean-squared error. It can be shown (see Exercise 7) that

$$\mathbb{E}(\widehat{\psi}) = \psi \quad \text{and} \quad \mathbb{V}(\widehat{\psi}) \leq \frac{1}{n\delta^2}. \tag{11.11}$$

Therefore, the MSE is of order $1/n$ which goes to 0 fairly quickly as we collect more data, no matter how large B is. The estimator defined in (11.10) is called the **Horwitz-Thompson** estimator. It cannot be derived from a Bayesian or likelihood point of view since it involves the terms ξ_{X_i}. These terms drop out of the log-likelihood and hence will not show up in any likelihood-based method including Bayesian estimators.

The moral of the story is this. Bayesian methods are tied to the likelihood function. But in high dimensional (and nonparametric) problems, the likelihood may not yield accurate inferences. ■

11.10 Example. Suppose that f is a probability density function and that

$$f(x) = cg(x)$$

where $g(x) > 0$ is a known function and c is unknown. In principle we can compute c since $\int f(x)\,dx = 1$ implies that $c = 1/\int g(x)\,dx$. But in many cases we can't do the integral $\int g(x)\,dx$ since g might be a complicated function and x could be high dimensional. Despite the fact that c is not known, it is often possible to draw a sample $X_1, \ldots, X_n$ from f; see Chapter 24. Can we use the sample to estimate the normalizing constant c? Here is a frequentist solution:

Let $\widehat{f}_n(x)$ be a consistent estimate of the density f. Chapter 20 explains how to construct such an estimate. Choose any point x and note that $c = f(x)/g(x)$. Hence, $\widehat{c} = \widehat{f}(x)/g(x)$ is a consistent estimate of c. Now let us try to solve this problem from a Bayesian approach. Let $\pi(c)$ be a prior such that $\pi(c) > 0$ for all $c > 0$. The likelihood function is

$$\mathcal{L}_n(c) = \prod_{i=1}^n f(X_i) = \prod_{i=1}^n cg(X_i) = c^n \prod_{i=1}^n g(X_i) \ \propto \ c^n.$$

Hence the posterior is proportional to $c^n\pi(c)$. The posterior does not depend on $X_1, \ldots, X_n$, so we come to the startling conclusion that, from the Bayesian point of view, there is no information in the data about c. Moreover, the posterior mean is

$$\frac{\int_0^\infty c^{n+1}\pi(c)\,dc}{\int_0^\infty c^n\pi(c)\,dc}$$

which tends to infinity as n increases. ■

These last two examples illustrate an important point. Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference.

What should we conclude from all this? The important thing is to understand that frequentist and Bayesian methods are answering different questions. To combine prior beliefs with data in a principled way, use Bayesian inference. To construct procedures with guaranteed long run performance, such as confidence intervals, use frequentist methods. Generally, Bayesian methods run into problems when the parameter space is high dimensional. In particular, 95 percent posterior intervals need not contain the true value 95 percent of the time (in the frequency sense).

11.10 Bibliographic Remarks

Some references on Bayesian inference include Carlin and Louis (1996), Gelman et al. (1995), Lee (1997), Robert (1994), and Schervish (1995). See Cox (1993), Diaconis and Freedman (1986), Freedman (1999), Barron et al. (1999), Ghosal et al. (2000), Shen and Wasserman (2001), and Zhao (2000) for discussions of some of the technicalities of nonparametric Bayesian inference. The Robins-Ritov example is discussed in detail in Robins and Ritov (1997) where it is cast more properly as a nonparametric problem. Example 11.10 is due to Edward George (personal communication). See Berger and Delampady (1987)

and Kass and Raftery (1995) for a discussion of Bayesian testing. See Kass and Wasserman (1996) for a discussion of noninformative priors.

11.11 Appendix

Proof of Theorem 11.5.

It can be shown that the effect of the prior diminishes as n increases so that $f(\theta|X^n) \propto \mathcal{L}_n(\theta)f(\theta) \approx \mathcal{L}_n(\theta)$. Hence, $\log f(\theta|X^n) \approx \ell(\theta)$. Now, $\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta}) = \ell(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta})$ since $\ell'(\widehat{\theta}) = 0$. Exponentiating, we get approximately that

$$f(\theta|X^n) \propto \exp\left\{-\frac{1}{2}\frac{(\theta - \widehat{\theta})^2}{\sigma_n^2}\right\}$$

where $\sigma_n^2 = -1/\ell''(\widehat{\theta}_n)$. So the posterior of θ is approximately Normal with mean $\widehat{\theta}$ and variance σ_n^2. Let $\ell_i = \log f(X_i|\theta)$, then

$$\begin{aligned}
\frac{1}{\sigma_n^2} &= -\ell''(\widehat{\theta}_n) = \sum_i -\ell_i''(\widehat{\theta}_n) \\
&= n\left(\frac{1}{n}\right)\sum_i -\ell_i''(\widehat{\theta}_n) \approx n\mathbb{E}_\theta\left[-\ell_i''(\widehat{\theta}_n)\right] \\
&= nI(\widehat{\theta}_n)
\end{aligned}$$

and hence $\sigma_n \approx \mathsf{se}(\widehat{\theta})$. ■

11.12 Exercises

1. Verify (11.7).
2. Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$.
   (a) Simulate a data set (using $\mu = 5$) consisting of n=100 observations.
   (b) Take $f(\mu) = 1$ and find the posterior density. Plot the density.
   (c) Simulate 1,000 draws from the posterior. Plot a histogram of the simulated values and compare the histogram to the answer in (b).
   (d) Let $\theta = e^\mu$. Find the posterior density for θ analytically and by simulation.
   (e) Find a 95 percent posterior interval for μ.
   (f) Find a 95 percent confidence interval for θ.

3. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $f(\theta) \propto 1/\theta$. Find the posterior density.

4. Suppose that 50 people are given a placebo and 50 are given a new treatment. 30 placebo patients show improvement while 40 treated patients show improvement. Let $\tau = p_2 - p_1$ where p_2 is the probability of improving under treatment and p_1 is the probability of improving under placebo.

   (a) Find the MLE of τ. Find the standard error and 90 percent confidence interval using the delta method.

   (b) Find the standard error and 90 percent confidence interval using the parametric bootstrap.

   (c) Use the prior $f(p_1, p_2) = 1$. Use simulation to find the posterior mean and posterior 90 percent interval for τ.

   (d) Let
   $$\psi = \log\left(\left(\frac{p_1}{1-p_1}\right) \div \left(\frac{p_2}{1-p_2}\right)\right)$$
   be the log-odds ratio. Note that $\psi = 0$ if $p_1 = p_2$. Find the MLE of ψ. Use the delta method to find a 90 percent confidence interval for ψ.

   (e) Use simulation to find the posterior mean and posterior 90 percent interval for ψ.

5. Consider the Bernoulli(p) observations

   0 1 0 1 0 0 0 0 0 0

   Plot the posterior for p using these priors: Beta(1/2,1/2), Beta(1,1), Beta(10,10), Beta(100,100).

6. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

   (a) Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior. Show that the posterior is also a Gamma. Find the posterior mean.

   (b) Find the Jeffreys' prior. Find the posterior.

7. In Example 11.9, verify (11.11).

8. Let $X \sim N(\mu, 1)$. Consider testing
   $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

Take $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. Let the prior for μ under H_1 be $\mu \sim N(0, b^2)$. Find an expression for $\mathbb{P}(H_0|X = x)$. Compare $\mathbb{P}(H_0|X = x)$ to the p-value of the Wald test. Do the comparison numerically for a variety of values of x and b. Now repeat the problem using a sample of size n. You will see that the posterior probability of H_0 can be large even when the p-value is small, especially when n is large. This disagreement between Bayesian and frequentist testing is called the Jeffreys-Lindley paradox.