11 Parametric Inference III: Bayesian Inference.

11.1 The Bayesian Philosophy

The statistical theory and methods that we have discussed so far are known as *frequentist (or classical)* inference. The frequentist point of view is based on the following postulates:
(i) Probability refers to limiting relative frequencies.
(ii) Probabilities are objective properties of the real world.
(iii) Parameters are fixed, (usually unknown) constants. Because they are not fluctuating, no probability statements can be made about parameters.
(iv) Statistical procedures should be designed to have well defined probabilistic properties, in the sense described in (i). For example, a 95 per cent confidence interval should trap the true value of the parameter with limiting frequency at least 95 per cent.

There is another approach to inference called *Bayesian inference.* The Bayesian approach is based on the following:
(i) 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.
(ii) Given (i), we can make probability statements about parameters, even though they are fixed constants.
(iii) The correct way to make inferences about a parameter θ, is to produce a probability distribution for θ. Inferences, such as point estimates and interval estimates may then be extracted from this distribution.

11.2 The Bayesian Method

Bayesian inference is usually carried out in the following way. We start by expressing degrees of beliefs about a parameter θ before we see any data. Let $f(\theta)$ denote this prior density function for θ. Now suppose we observe data $X_1, \ldots, X_n \sim f(x;\theta)$. In this context, $f(x;\theta)$ should be interpreted as our beliefs about the data given θ so we now write $f(x|\theta)$ instead of $f(x;\theta)$. The joint density of the data given θ is $f(x_1,\ldots,x_n|\theta) = \prod_i f(x_i|\theta)$. Once we see the data, we want to compute the distribution for θ give the observed data. We call this the *posterior distribution.* How do we compute this posterior distribution?

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,

$$P(\Theta=\theta|X=x) = \frac{P(X=x,\Theta=\theta)}{P(X=x)} = \frac{P(X=x|\Theta=\theta)P(\Theta=\theta)}{\sum_\theta P(X=x|\Theta=\theta)P(\Theta=\theta)}$$

which you may recognize from earlier in the course as *Bayes' theorem.* The version for continuous variables is obtained by using density functions:

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

Now if we have n iid observations $X_1,\ldots,X_n$, we should replace $f(x|\theta)$ with $f(x_1,\ldots,x_n|\theta) = \prod_i f(x_i|\theta)$ to get

$$f(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)f(\theta)}{\int f(x_1,\ldots,x_n|\theta)f(\theta)d\theta}.$$

Note that $f(x_1,\ldots,x_n|\theta) = \prod_i f(x_i|\theta) = \mathcal{L}_n(\theta)$, the likelihood function, so we can rewrite this as

$$f(\theta|x_1,\ldots,x_n) = \frac{\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta}.$$

Finally, note that $\int \mathcal{L}_n(\theta)f(\theta)d\theta$ is a constant that does not depend on θ; we call this quantity the *normalizing constant.* So we can write the last equation as

$$f(\theta|x_1,\ldots,x_n) \propto \mathcal{L}_n(\theta)f(\theta)$$

or

"posterior is proportional to likelihood times prior."

We will make one more notational simplification. We will write X^n to mean $(X_1,\ldots,X_n)$ and x^n to mean $(x_1,\ldots,x_n)$. With this notation, Bayes' theorem is $f(\theta|x^n) \propto \mathcal{L}_n(\theta)f(\theta)$. You might wonder, doesn't it cause a problem to throw away the constant $\int \mathcal{L}_n(\theta) f(\theta) d\theta$? The answer is that we can always figure out what the constant is since we know that $\int f(\theta|x^n) d\theta = 1$. Hence, we often omit the constant until we really need it.

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

$$\overline{\theta}_n = \int \theta f(\theta|x^n) d\theta.$$

We can also obtain a Bayesian interval estimate. 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

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

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

EXAMPLE 11.1 *Let $X_1, \ldots, X_n \sim Ber(p)$. 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 $s = \sum_i x_i$ is the number of heads. Recall that random variable has a Beta distribution with parameters α and β if its density is

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

So 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 Beta (α, β) is $\alpha/(\alpha+\beta)$ (see Chapter 3) so the Bayes estimator is

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

It is instructive to rewrite the estimator as

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

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 per cent posterior interval can be obtained by numerically finding a and b such that* $\int_a^b f(p|x^n)dp = .95$. *We can do this in R as follows:*

```
### Let's generate some Bernouliis from p=.3, say
n <- 20
x <- rbinom(n,1,.3)
print(x)
s <- sum(x)
grid <- seq(0,1,length=1000)
posterior <- dbeta(grid,s+1, n-s+1)     ### compute the posterior density
###plot it
plot(grid,posterior,type="l",xlab="p",ylab="posterior density")
###find interval
left <- qbeta(.025,s+1,n-s+1)
right <- qbeta(.975,s+1,n-s+1)
interval <- c(left,right)
print(interval)
```

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*.

EXAMPLE 11.2 *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 the homework, you will be asked to show that the posterior for* θ *is* $N(\overline{\theta}, \tau^2)$ *where*

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

where

$$w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}} \quad \text{and} \quad \frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2}$$

and $se = \sigma/\sqrt{n}$ *is the standard error of the mle* $\overline{X}$*. This is another example of a conjugate prior. Note that* $w \to 1$ *and* $\tau/se \to 1$ *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 is find $C = [c, d]$ *such that* $Pr(\theta \in C|X^n) = .95$*. We can do this by choosing* c *such that* $Pr(\theta < c|X^n) = .025$ *and* $Pr(\theta > d|X^n) = .025$*. So, we want to find* c *such that*

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

Now, we know that $P(Z < -1.96) = .025$*. 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 per cent Bayesian interval is* $\overline{\theta} \pm 1.96\,\tau$*. Since* $\overline{\theta} \approx \widehat{\theta}$ *and* $\tau \approx se$*, the 95 per cent Bayesian interval is approximated by* $\widehat{\theta} \pm 1.96\,se$ *which is the frequentist confidence interval.*

11.3 Functions of Parameters

How do we make inferences about $\tau = g(\theta)$? Remember earlier we solved the following "change of variables" problem: given the density f_X for X, find the density for $Y = g(X)$. We simply apply the same reasoning. The posterior cdf for τ is

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

where $A = \{\theta : \; g(\theta) \leq \tau\}$. The posterior density is $h(\tau|x^n) = H'(\tau|x^n)$. There is nothing new except Greek letters instead of Latin letters.

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 $\overline{\theta}_n = E(\theta|x^n)$ is

$$\frac{1}{B}\sum_{j=1}^{B}\theta_j.$$

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 $\theta_1, \ldots, \theta_B$.

Simulation makes the change-of-variables problem easier. Once you 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.

EXAMPLE 11.3 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$ *and* $f(\theta) = 1$. *The posterior is* $\text{Beta}(s+1, n-s+1)$ *where* $s = \sum_i x_i$. *We can draw samples from the posterior as follows:*

```
B <- 10000
theta <- rbeta(B, s+1, n-s+1)
hist(theta)
theta.bar <- mean(theta)
interval <- quantile(theta, c(.025,.975))
```

If we want to make inferences about $\tau = \log(\theta/(1-\theta))$ *we just do this:*

```
tau <- log(theta)
hist(tau)
tau.bar <- mean(tau)
interval <- quantile(tau, c(.025,.975))
```

11.5 Large Sample Properties of Bayes' Procedures.

In the last example, we saw that the posterior mean was close to the mle and the posterior interval was similar to the confidence interval. This is true in greater generality.

THEOREM 11.1 *Under weak conditions, we have that the posterior is approximately* $N(\widehat{\theta}, se^2)$. *Hence,* $\overline{\theta}_n \approx \widehat{\theta}_n$. *Also, if* $C = [\widehat{\theta}_n - z_{\alpha/2} se, \widehat{\theta}_n + z_{\alpha/2} se]$ *is the usual* $1-\alpha$ *mle-based confidence interval, then*

$$Pr(\theta \in C | X^n) \to 1 - \alpha.$$

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

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

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

SUMMARY. The frequentist large sample result is:

$$\widehat{\theta}_n \approx N(\theta, se^2).$$

The Bayesian result is

$$\theta | X^n \approx N(\widehat{\theta}, se^2) \approx N(\widehat{\theta}, se^2).$$

The interpretations are different but the estimates and intervals are approximately the same.

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

A big question in Bayesian inference is: where do you 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 seems less plausible in complicated problems especially if there are many parameters. 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 $p | X^n \sim Beta(s+1, n-s+1)$ as we saw earlier which seemed very reasonable. But unfettered use of flat priors raises some questions.

IMPROPER PRIORS. Consider the $N(\theta, 1)$ example. 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 real probability density in the usual sense. We call such a prior an *improper prior.* Nonetheless, we can still carry out Bayes' theorem and compute the posterior density $f(\theta) \propto \mathcal{L}_n(\theta) f(\theta) \propto \mathcal{L}_n(\theta)$. In the normal example, 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. Go back to the Bernoulli example and consider using the flat prior $f(p) = 1$. Recall that a flat prior presumably represents our lack of information about p before the experiment. Now let $\psi = \log(p/(1-p))$. This is a transformation and we can compute the resulting distribution for ψ. It turns out that

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

But one could argue that if we are ignorant about p then we are also ignorant about ψ so shouldn't we use a flat prior for ψ? This contradicts the prior $f_\Psi(\psi)$ for ψ that is implied by using a flat prior for p. 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.

EXAMPLE 11.4 *Consider the Bernoulli (p). Recall that*

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

Jeffrey's 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 $f(\theta) \propto \sqrt{\det I(\theta)}$ where $\det(A)$ denotes the determinant of a matrix A.

11.7 Multiparameter Problems

In principle, multiparameter problems are handled the same way. Suppose that $\theta = (\theta_1, \ldots, \theta_p)$. The posterior density is still given by

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

The question now arises of how to extract inferences about one parameter. The key is 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

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

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.

EXAMPLE 11.5 (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,*

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

Let us adopt the prior $f(p_1, p_2) = 1$. 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

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

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 $p_2|x_2 \sim \text{Beta}(x_2 + 1, n_2 - x_2 + 1)$. We can simulate from the posterior for τ as follows:

```
B <- 10000
p1 <- rbeta(B,x1+1, n1-x1+1)
p2 <- rbeta(B,x2+1, n2-x2+1)
tau <- p2 - p1
```

11.8 Appendix

Proof of Theorem 11.1.

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, $\log f(\theta|X^n) \approx \ell(\theta)$ where $\ell(\theta) = \log \mathcal{L}_n(\theta)$ is the log-likelihood function. 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

$$\sigma_n^2 = -\frac{1}{\ell''(\widehat{\theta}_n)}.$$

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

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

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