

Lecture Note 11: Bayesian Point Estimation (CB 7.2.3)

We have already discussed two general approaches to constructing point estimators of parameters, the method of moments and the maximum likelihood method. A third important class of estimators are the Bayesian estimators, so called because they make use of Bayes' Theorem.

Suppose we are interested in the probability that a coin comes up heads. Let P be the probability of heads, and suppose that P is chosen by "Nature," according to a uniform distribution on $(0, 1)$. We do not observe P , but we get to toss the coin once and see whether it comes up heads. Before seeing the outcome of the coin flip, we know that the *marginal* distribution of P is uniform on the unit interval. If we in fact observe heads, how should we "update" this distribution to reflect the new information?

The marginal density of P is:

$$f_P(p) = 1, \quad 0 \leq p \leq 1,$$

and the conditional density of X given $P = p$ is

$$f_{X|P}(x|p) = p^x \cdot (1 - p)^{1-x}.$$

Therefore we can calculate the joint density:

$$f_{X,P}(x,p) = f_{X|P}(x|p) \cdot f_P(p) = p^x \cdot (1 - p)^{1-x}.$$

Note that X can only take on values 0 or 1, so its marginal density (actually PMF) is

$$f_X(x) = \int_p f_{X,P}(x,p) dp = x \cdot \int_0^1 p dp + (1-x) \cdot \int_0^1 (1-p) dp = x \cdot \frac{1}{2} + (1-x) \cdot \frac{1}{2} = \frac{1}{2},$$

and the conditional distribution of P given X is:

$$f_{P|X}(p|x) = \frac{f_{X|P}(x|p) \cdot f_P(p)}{\int_0^1 f_{X|P}(x|p) \cdot f_P(p) dp} = 2p^x(1-p)^{1-x}.$$

This conditional distribution is what we are after: given the data (X), we want to know what the conditional distribution of the parameter (P) looks like. Let's calculate it for this example. Let $X = 1$ denote heads.

$$f_{P|X}(p|x=1) = \frac{f_{X|P}(x|p) \cdot f_P(p)}{\int_0^1 f_{X|P}(x|p) \cdot f_P(p) dp} = \frac{p \cdot 1}{\int_0^1 p \cdot dp} = 2p.$$

Thus the conditional density of p has a triangular shape, with more mass close to 1 than close to 0. We call the marginal distribution f_P the "prior" distribution to reflect the interpretation of P being chosen before X , and we call $f_{P|X}$ the "posterior" distribution. We see that the prior is modified based on the likelihood function $f_{X|P}$ to obtain the posterior.

Now let us look at a more general class of prior distributions. Suppose that the prior for P follows a Beta distribution:

$$f_P(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1},$$

with α and β known numbers. Notice that the case $\alpha = \beta = 1$ gives the uniform distribution, so that the previous analysis should be a special case. Recall that the mean and variance of the Beta distribution are

$$E[P] = \frac{\alpha}{\alpha + \beta},$$

and

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

respectively. Suppose we want the prior distribution to have mean $1/4$ and variance $1/100$. Then there is a Beta distribution corresponding to that, namely the Beta distribution with

$$\frac{\alpha}{\alpha + \beta} = \frac{1}{4},$$

and

$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{1}{100},$$

which corresponds to $\alpha = 71/16 \approx 4$ and $\beta = 213/16 \approx 13$. (More realistic values might be a mean of $1/2$ and a variance of $1/100$, but we will work with these numbers in this example.)

Again the data consist of just a single observation with $X = 1$. The joint distribution of P and X , at $X = 1$, is

$$f_P(p) \cdot f_{X|P}(x|p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \cdot p.$$

How do we figure out the conditional distribution of P given $X = 1$? Well, we know that all we have to do is find a constant such that $f_P(p) \cdot f_{X|P}(x = 1|p)$ integrates out to one as a function of p . Strip away the part of the function that does not depend on p , and we are left with the kernel of the conditional density:

$$f_{P|X}(p|x) \propto p^\alpha \cdot (1-p)^{\beta-1}.$$

This implies that the conditional distribution of P given $X = 1$ is a Beta distribution with parameters $\alpha + 1$ and β . The mean and variance of this distribution are

$$E[P|X = 1] = \frac{\alpha + 1}{\alpha + \beta + 1},$$

and

$$V(P|X = 1) = \frac{(\alpha + 1) \cdot \beta}{(\alpha + \beta + 1)^2(\alpha + \beta + 2)},$$

respectively. After observing $X = 1$, we update the distribution of P : the mean moves upwards $(\alpha + 1)/(\alpha + \beta + 1) = (71/16 + 1)/(71/16 + 213/16 + 1) = 0.29$, is slightly higher than the unconditional mean, $\alpha/(\alpha + \beta) = 1/4$, and the variance $(\alpha\beta)/((\alpha + \beta)^2(\alpha + \beta + 1)) = 0.0104$ is slightly higher than the prior variance of 0.01 . (This is somewhat unusual. Typically the posterior variance is lower than the prior variance, due to the extra information. Here the fact that the extra information is so far from the prior mean implies that the uncertainty is actually increased by the extra information.)

Now let us do this more systematically. There are two ingredients to a Bayesian analysis. First a model for the data given some unknown parameters. In our example that model was $f_{X|P}(x|p) = p^x \cdot (1-p)^{1-x}$. Second, a prior distribution for the parameters. In our case that is the Beta distribution with parameters α and β . This prior distribution is known to the researcher. Then, using Bayes' theorem we calculate the conditional distribution of the parameters given the data, also known as the posterior distribution,

$$f_{P|X}(p|x) = \frac{f_{X,P}(x,p)}{f_X(x)} = \frac{f_{X|P}(x|p) \cdot f_P(p)}{\int f_{X|P}(x|p) \cdot f_P(p) dp}.$$

In this step we often use a shortcut. First note that, as a function of p , the conditional density of P given X is proportional to

$$f_{P|X}(p|x) \propto f_{X|P}(x|p) \cdot f_P(p).$$

Once we calculate this product, all we have to do is find the constant that makes this expression integrate out to one as a function of the parameter. At that stage it is sometimes easy to recognize the distribution and figure out through that route what the constant is.

Example: Let us look at a second example. Suppose the conditional distribution of X given the parameter μ is normal with mean μ and variance 1. The prior distribution for μ is normal with mean zero and variance 100. What is the posterior distribution of μ given $X = x$? The posterior distribution is proportional to

$$\begin{aligned} f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot 100}\mu^2\right) \\ &= \exp\left(-\frac{1}{2}\left(x^2 - 2x\mu + \mu^2 + \mu^2/100\right)\right) \\ &\propto \exp\left(-\frac{1}{2(100/101)}\left(\mu - (100/101)x\right)^2\right). \end{aligned}$$

This implies that the conditional distribution of μ given $X = x$ is normal with mean $(100/101)x$ and variance $100/101$. \square .

Point Estimates from Posterior Distributions

The method of moments and maximum likelihood estimators return a single point estimate for a given data set. In contrast, the Bayesian posterior is an entire *distribution* over the parameter space. We can turn this in to a point estimate by taking some measure of central tendency, such as the conditional mean of the parameter given the data.¹ For example, the conditional mean in the previous normal example is $(100/101)x$. We see that the Bayes estimator is not necessarily unbiased: since $X \sim N(\mu, 1)$, the mean of $\hat{\mu} \equiv (100/101)X$ for a fixed μ is $E[\hat{\mu}|\mu] = (100/101)\mu$. On the other hand, the bias is quite small, and if we had a larger prior variance, the bias would be even smaller.

Interpreting the Prior and Posterior Distributions

¹Another reasonable choice would be the conditional median.

So far, we have imagined that some hypothetical other player called Nature is selecting the parameter according to some fixed, known prior distribution. This obviously greatly limits the application of the Bayesian approach. However, we can apply it more broadly if we imagine the prior as a *subjective* assessment, before seeing the data, of how likely different possible parameter values are. If we regard the prior distribution as reflecting our beliefs about the unknown parameter, then the posterior distribution is just an updated version of our beliefs, using the laws of conditional probability to reflect the additional information in the observed data X .

This interpretation has been suggested by Ramsey (1931), and developed axiomatically by Savage (1954). (A nice discussion of subjective expected utility theory is given in Kreps (1988).) Savage develops an extension of the von Neumann-Morgenstern expected utility theory, which says that under certain axioms of coherency, a decision-maker faced with an uncertain decision problem, must ask as if she assigns subjective probabilities to the unknown states of nature, updates them according to Bayes' theorem, and maximizes expected utilities where the expectations are with respect to the posterior distributions.

Another interpretation is to just imagine the prior distribution as a device to generate procedures. Hopefully the procedures can then be evaluated and we can see if they produce sensible results. We might want to pick a particularly easy or relatively "uninformative" form for the prior, calculate the corresponding posterior, and study its properties. For example, it often turns out that the mean of the posterior distribution is similar to the maximum likelihood estimator, provided the prior distribution is relatively smooth. This approximation is particularly good when the sample size is large.

References

- Kreps, D., (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.
- Ramsey, F. P. (1931), "Truth and Probability," in *The Foundations of Mathematics and Other Logical Essays*, ed. by R.B. Braithwaite, pp. 156-98. Routledge and Kegan Paul, London.
- Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, New York. (Reissued in 1972 by Dover.)