Probability and Statistics

Fall 2000

Lect. 9: Bayesian Inference I, CB 7.2.3

In the remainder of the course we look at statistical inference. The basic setting is the following: we have a random variable X, which we know has a probability density or probability mass function equal to

$$f_X(x;\theta).$$

What we are fundamentally interested in is the same type of things as before: what is the probability of some event happening, i.e., the probability that $X \in A$ for some set A, or the expected value of some function of X. However, the problem is complicated by the fact that we do not know the actual value of the parameter θ. Determining plausible values for θ is the problem we focus on. We will therefore pay little attention from now on to the eventual goal of calculating probabilities or expectations, although that is likely to be the motivation for trying to estimate parameters like θ. For example, for someone investigating the life time of lightbulbs, it is not the parameter of the exponential distribution itself that is of interest, but some more readily interpretable measure of the life time, such as the median or mean life of a lightbulb, or the probability that the lifetime exceeds 500 hours.

What information do we have for infering values of θ? We study the case where we have an observation, that is, a realization of this random variable. For example, if the experiment is tossing a coin, and the model is a binomial one with parameters $n = 1$ and unknown probability p, the realization will be a one or a zero. More typically, the experiment could be tossing a coin n (say n =100) times, and X would be a vector of zeros and ones, with joint distribution

$$f_X(x_1, x_2, \ldots, x_N) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

Here we could also look at the distribution of the sum $y = \sum x_i$, which has a binomial distribution with parameters $n = 100$ and p, directly, but the point is that often we have

more than one realization of the random variable, which is clearly going to make our inferences more precise.

We look at two types of questions: (a) (point) estimation, where the goal is to find a single number that is the "best" guess for the value of θ, and (b) testing, where we wish to investigate whether a particular value of θ is consistent with the evidence from the observations, and relatedly, interval estimation. An example of the latter, if we have a single observation from a Bernoulli trial with probability p, and the value of the observation is 1, we know for sure that the probability of success is greater than zero. We cannot rule out other values of p, and even with a very large number of observations we can never rule out for sure values of p strictly between zero and one, but we can make probabilistic statements about the plausibility of such values. Specifically, if, with very many observations, e.g., millions of observations, the fraction of ones is roughly $1/2$, it is extremely unlikely that we would see something like that if in fact the true value of p is equal to 0.01. We will make precise the meaning of extremely unlikely.

First we look at point estimation. There are two approaches to estimation, right and wrong, also known as Bayesian and classical inference. Most of the remainder of the course will be devoted to classical inference.

First we look at Bayesian inference. Suppose we are interested in the probability that a coin comes up heads. We plan an experiment to toss the coin once and it comes up heads. What are our beliefs regarding the long run probability that the coin comes up heads? Go back to the point in time before we tossed the coin. At that point you may have some ideas about the probability that the coin comes up heads. Let's call this probability P. It could be that this probability is less than 0.1, but that may be unlikely. It could be that this probability is greater than 0.9, but again this may be unlikely. Let us organize our beliefs regarding this probability P into a probability density function, $f_P(p)$. To keep things simple, let's suppose that $f_P(p) = 1$, for $0 < p < 1$, a uniform distribution on the unit interval. This is obviously not very realistic, but we'll deal with generalizations later.

So, now we have:

$f_P(p)$,

the marginal distribution of P, and

$$f_{X|P}(x|p),$$

the conditional distribution of X given P. Therefore we can calculate the joint distribution:

$$f_{XP}(x,p) = f_{X|P}(x|p) \cdot f_P(p),$$

and the conditional distribution of P given X:

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

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.

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

Now let us look at a more realistic distribution for our beliefs prior to the experiment. Suppose our beliefs are described by the following 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, choosen to make the distribution conform to our prior beliefs. Before we choose $\alpha = \beta = 1$ so we got the uniform distribution. Now we have a much more flexible distribution. 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 think that our prior beliefs about p are represented by a distribution with mean 1/4 and variance 1/100. Then there is a Beta distribution corresponding to that, namely 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 realistical might be a mean of 1/2 and a variance or 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 our beliefs of the plausible values of P upwards: the conditional mean, $(\alpha + 1)/(\alpha + \beta + 1) = (71/16 + 1)/(71/16 + 213/16 + 1) = 0.29$, 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, that is, choosen by 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 found the constant that makes this expression integrate out to one as a function of the parameter. At that stage it is often 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

$$f_{\mu|X}(\mu|x) \propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot 100}\mu^2\right)$$

$$= \exp -\frac{1}{2}\Big(x^2 - 2x\mu + \mu^2 + \mu^2/100\Big)$$

$$= \exp\Big(-\frac{1}{2(100/101)}(\mu - (100/101)x)^2\Big).$$

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