A short introduction to Bayesian statistics, part II
Math 218, Mathematical Statistics
D Joyce, Spring 2016

Bayes' pool table example. The process we just completed is what Thomas Bayes (1702–1761) did. He illustrated the problem with balls on a table. I'll paraphrase his illustration using the terminology we developed above.

Suppose a ball W is placed on a pool table so that "there shall be the same probability that it rests upon any equal part of the plane [rectangle] as another." We'll suppose the length of the table is 1 and that the distance of W from one end, call it the left end, is p. We don't know where W is placed, so our prior distribution on p is uniform, that is, the density function for p is uniform on the interval $[0, 1]$.

Next suppose another ball O is repeatedly randomly placed on the table n times, and in k of these placements ball O is closer to the left end than ball W is. Suppose all we know is that of the n times O was placed, k times was placed closer to the left end. Given that outcome what is the posterior distribution for p? We just worked out the answer. The prior distribution for p was uniform on $[0, 1]$. Therefore, the posterior distribution is $\text{BETA}(k + 1, n + 1 - k)$.

The conjugate prior family for Bernouli distributions. There are, of course, applications where the prior distribution on p should be uniform on $[0, 1]$. But sometimes you have other information that suggests p isn't uniformly distributed on $[0, 1]$ but has some other distribution. Fortunately, the whole family of beta distributions works well here. That is, if the prior distribution is any beta distribution $\text{BETA}(\alpha_0, \beta_0)$, and we observe k successes and l failures, then the posterior distribution is $\text{BETA}(\alpha_0 + k, \beta_0 + l)$. Here's why. The prior distribution $f(p)$ was proportional to $p^{\alpha-1}(1-p)^{\beta-1}$, and the posterior distribution $f(p|\mathbf{x})$ is proportional to $p^k(1-p)^l$ times the prior, therefore

$$
\begin{aligned}
f(p|\mathbf{x}) &\propto p^k(1-p)^l p^{\alpha-1}(1-p)^{\beta-1} \\
&= p^{\alpha+k-1}(1-p)^{\beta+l-1}
\end{aligned}
$$

Since, when the prior distribution is a beta distribution, then the posterior one is also a beta distribution, we say the family of beta distributions is a *conjugate prior* family for p, the parameter in the Bernoulli process. The first parameter, α, is increased by the number of successes while the second, β, is increased by the number of failures.

Selecting the prior distribution. Here's one way we can incorporate knowledge about p by selecting a particular beta distribution. Typically, our knowledge about a parameter can be summarized in terms of a mean μ and variance σ^2 for its distribution. Since the family of beta distributions is parametrized by two variables, these two values μ and σ^2 should determine exactly one beta distribution. Now, a $\text{BETA}(\alpha, \beta)$ distribution has mean

$$
\mu = \frac{\alpha}{\alpha + \beta}
$$

and variance

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

so we can solve these two equations for the parameters α and β of the prior distribution in terms of μ and σ^2. We find

$$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$

$$\beta = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)$$

For instance, suppose that you think p is about 0.5 with a standard deviation of 0.1 (so that p lies in the interval $[0.3, 0.7]$ with probability close to 95% if p is close to normally distributed). Then $\mu = 0.5$ and $\sigma = 0.1$. It works out that α and β both equal 12. That's the same amount of information that you would get from 24 Bernoulli trials where half of them turn out successes. In other words, if you start with the prior distribution $\text{BETA}(0, 0)$, run those 24 trials and get 12 successes and 12 failures, then your posterior distribution would be $\text{BETA}(12, 12)$ which has a mean $\mu = 0.5$ and standard deviation $\sigma = 0.1$.

The figure below shows the density function for $\text{BETA}(12, 12)$. It's very close to a normal distribution with $\mu = 0.5$ and $\sigma = 0.1$.
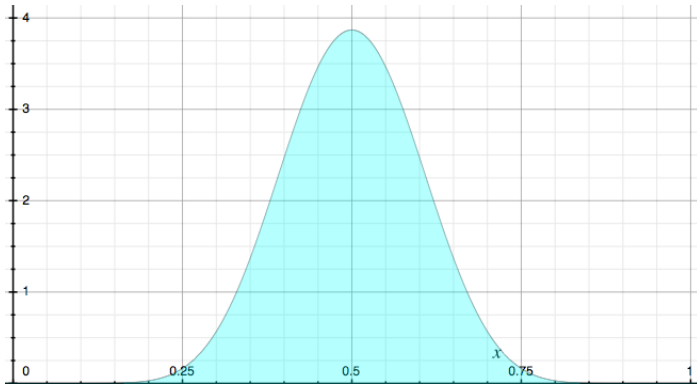
4 Point estimators and probability intervals.

Point Estimators. The posterior distribution $f(\theta|\mathbf{x})$ gives a whole distribution for the parameter θ. A point estimator for θ is supposed to be a single number, a best guess for θ, in some sense of the word "best".

With a whole distribution for θ, there are a lot of choices for the best guess. One that's often used is just the mean of the distribution, and, in this case, that's the mean of the posterior distribution $f(\theta|\mathbf{x})$, so it's $\mu_{\theta|\mathbf{x}} = E(\theta|\mathbf{x})$. That's what we'll use, and we'll call it the *Bayesian point estimator*. Other choices for estimators are the median of the distribution and the mode of the distribution.

For the Bernoulli distributions studied above, the prior distribution is any beta distribution $\text{BETA}(\alpha_0, \beta_0)$, often chosen to be $\text{BETA}(1, 1)$, the uniform distribution on $[0, 1]$. After observing n trials, of which k are successes and l failures, then the posterior distribution is $\text{BETA}(\alpha_0 + k, \beta_0 + l)$. The prior point estimator was

$$\mu_p = E(p) = \frac{\alpha_0}{\alpha_0 + \beta_0},$$

and the posterior estimator is

$$\mu_{p|\mathbf{x}} = E(p|\mathbf{x}) = \frac{\alpha_0 + k}{\alpha_0 + \beta_0 + k + l}.$$

If we take the prior to be $\text{BETA}(1, 1)$, that makes the prior estimator $\frac{1}{2}$ and the posterior estimator $\frac{k+1}{k+l+2}$.

This doesn't agree with the maximum likelihood estimator for that data, which is $\frac{k}{k+l}$, but it's close. The maximum likelihood estimator is actually the mode of the distribution, the maximum value of the posterior distribution $\text{BETA}(\alpha_0 + 1, \beta_0 + 1)$. That beta distribution has a density proportional to $p^k(1-p)^l$. A little calculus shows that the maximum occurs when $p = \frac{k}{k+l}$.

There are good arguments for concluding $\frac{k+1}{k+l+2}$ is a better estimator for p than $\frac{k}{k+l}$ is, but it's easy to change our prior to make the Bayesian estimate equal to $\frac{k}{k+l}$. Just make the prior $\text{BETA}(0, 0)$, a sort of know-nothing prior. Of course, that is not a valid distribution, but an improper distribution. Indeed, $\text{BETA}(\alpha, \beta)$ is only a probability distribution if both $\alpha \geq 1$ and $\beta \geq 1$. Still, asserting that the prior is $\text{BETA}(0, 0)$ can be taken to be a formal statement

so that when the data $\mathbf{x}$ comes in, with at least one success and one failure, the resulting posterior distribution becomes BETA(k, l).

Interval estimators. In classical statistics, we have confidence intervals. When we say, for example, that a 90% confidence interval for μ is $[0.3, 0.6]$, we aren't saying that the probablility that μ lies in $[0.3, 0.6]$ is 0.9. Instead, we're saying that in the long run, when we use the 90% confidence levels, we'll be right about 90% of the time.

In Bayesian statistics, we can actually have probability intervals, and we can get them from the posterior density function. For example, suppose our prior on p is uniform, and we perform $n = 2$ trials, have $k = 1$ success and $l = 1$ failure. Then the posterior is BETA$(2, 2)$ which has the density function

$$f_{p|\mathbf{x}} = 6p(1 - p,)$$

and its integral is the c.d.f

$$F_{p|\mathbf{x}} = 3p^2 - 2p^3.$$

We can find a 90% probablility interval for p if we remove from the interval $[0, 1]$ two ends each with probability 0.05. To find the left interval, we're looking for a value a so that $F_{a|\mathbf{x}} = 0.05$, which you can find from a table of beta distributions, or solve $3a^2 - 2a^3 = 0.05$. I graphed it to find $a = 0.135$. So, we'll remove the 5% interval $[0, 0.135]$ from the left end. Likewise, we'll remove the 5% interval $[0.865, 1]$ from the right end. That leaves the 90% interval $[0.135, 0.0.865]$.

You could also remove 10% from one end and nothing from the other to get a 90% probability interval.

Math 218 Home Page at