Outline

The Bernoulli Model

We often like to think of observables y as having a model indexed by a parameter θ.

The density

$$p(y \mid \theta)$$

describes what kind of y we get given a value of the "latent" parameter θ.

If we put a prior $p(\theta)$ on θ and then observe y we have

Bayes Theorem:

$$p(\theta \mid y) \propto f(y \mid \theta)\, p(\theta).$$

Note:

Given we observe y, $f(y \mid \theta)$ is a function of θ.

This is the *likelihood function*.

Given y,

$$L(\theta) \equiv f(y \mid \theta).$$

Bayes theorem is often written:

$$p(\theta \mid y) \propto L(\theta)\, p(\theta).$$

Let's see how these ideas play out in the simple case where "$\theta = p$", that is, we are inferring about the unknown probability p that something happens.

The Bernoulli Likelihood

Suppose on each independent trial, the probability of a "success" is p.

Let y denote the number of successes given n trials.

$$p(y \mid p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y \, (1 - p)^{n-y}.$$

Thus,

$$L(p) \propto p^y \, (1 - p)^{n-y}.$$

Using a Discrete Prior

How do we choose the form of our prior?

We want two things:

- The ability to flexibly express beliefs about p.
- The ability to "compute" the posterior.

Since $p(p)$ looks bad, let's call our prior $g(p)$.

A very simple approach is to use a discrete distribution for p.

That is, we say that p must be one of the values

$$p \in \{p_1, p_2, \ldots, p_m\}.$$

And we let,

$$P(p = p_i) = g(p_i).$$

Then

$$p(p = p_i \mid y) \propto L(p_i)\, g(p_i).$$

$$p(p = p_i \mid y) = \frac{L(p_i)\, g(p_i)}{\sum_{j=1}^{m} L(p_j)\, g(p_j)}.$$

Using a Beta Prior

Certainly, the most commonly used prior is the Beta prior.

Recall the *Gamma function*:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}\, e^{-y}\, dy.$$

$$\Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1).$$

$$\Gamma(n) = (n - 1)!.$$

$\Gamma(1) = 1.$ $\Gamma(.5) = \sqrt{\pi}.$

The Beta distribution

$$X \sim Beta(\alpha, \beta)$$

then,

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}, \;\; x \in (0,1).$$

$$E(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^{\infty} x^{\alpha}\,(1-x)^{\beta-1}\,dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\frac{\Gamma(\alpha+1)\,\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

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

We use a Beta prior for p, $p \sim Beta(\alpha, \beta)$.

Then,

$$g(p \mid y) \propto L(p)g(p) = p^{y}\,(1-p)^{n-y}\;p^{\alpha-1}\,(1-p)^{\beta-1}.$$

$$\propto p^{\alpha+y-1}\,(1-p)^{\beta+n-y-1}.$$

So,

$$p \mid y \sim Beta(\alpha + y, \beta + n - y).$$

Conjugate Priors:

We started with a Beta prior.

We stared and the likelihood $\times$ prior and saw that it was also of the Beta form.

When the prior and likelihood are in the same parametric family we say the prior is *conjugate*.

Models of the exponential family form have conjugate priors.

Discrete Approximation

If our prior is discrete computation of the posterior is straightforward given calculation of the $L(p)$ is so easy.

If we use the conjugate Beta prior, which is continuous, then computation of the posterior is also straightforward.

Suppose we use an arbitrary continuous prior with density $g(p)$. How do we "compute" the posterior?

Perhaps the most obvious thing is to do the required integrals numerically:

$$g(p \mid y) = \frac{L(p)\, g(p)}{\int L(p)\, g(p)\, dp}.$$

$$P(p \in A \mid y) = \frac{\int_A L(p)\, g(p)}{\int L(p)\, g(p)\, dp}.$$

A nice simple way that works generally is to discretize the prior.

We approximate the continuous prior by a discrete one, and then do all the computations as in the discrete case.

We again choose a grid of points:

$$\{p_1, p_2, \ldots, p_m\}.$$

And then our approximate prior is

$$P(p = p_i) = \frac{g(p_i)}{\sum_{i=1}^{m} g(p_i)},$$

where g is the continuous density.

We choose the grid $\{p_i\}$ to be evenly spaced and cover the support of the posterior and compute as in the discrete case.

Prediction

Suppose we are trying to predict the outcome of the next trial?

Let $\tilde{y}$ be 1 if the next trial is a success and 0 if it is a failure.

$$\tilde{y} \sim Bern(p).$$

$$P(\tilde{y} = 1 \mid y) = \int P(\tilde{y} = 1 \mid p, y)\, g(p \mid y)\, dp$$

$$= \int P(\tilde{y} = 1 \mid p)\, g(p \mid y)\, dp$$

$$= \int p\, g(p \mid y)\, dp$$

$$= E(p \mid y).$$

For the Beta prior we have

$$E(p \mid y) = \frac{\alpha + y}{\alpha + \beta + n}.$$

Note that if α and β are "large" then the prior dominates and the posterior mean is close to the prior mean of $\frac{\alpha}{\alpha+\beta}$.

If n is "large" then the posterior is mean is $\frac{y}{n}$ which is the sample fraction of successes.

For the discrete methods we can just sum:

$$E(p \mid y) = \sum_{i=1}^{m} p_i \, P(p = p_i \mid y).$$

Change of Variable

Suppose we want inference for the odds:

$$o = p/(1-p).$$

More generally, we may want inference for some function of the parameter $\theta = f(p)$.

In the discrete case it is easy to figure out the distribution of θ.

Let's assume that $f(p)$ has the inverse function $h(\theta)$.

$$\theta = f(p), \quad p = h(\theta).$$

If our grid for p is $\{p_i\}$ then possible values for θ are $\{\theta_i = f(p_i)\}$.

$$P(\theta = \theta_i) = P(p = h(\theta_i)) = g(h(\theta_i)).$$

where g is the probability mass function of p.

We can think of the function f as just "relabeling" the possible outcomes.

In the continous case we can do a change of variable using the Jabobian..

Let $g(p)$ be the density of p and $\theta = f(p)$.
Let $h(\theta) = p$ be the inverse of f.

$$p(\theta) = g(h(\theta)) \, |\frac{dh(\theta)}{d\theta}|.$$

or,

$$p(\theta) = g(p(\theta)) \, |\frac{dp}{d\theta}|$$

Example:

Let Z be a random variable.

$y = f(z) = \mu + \sigma z.$

Then

$$z = h(y) = \frac{y - \mu}{\sigma}, \quad \frac{dz}{dy} = \frac{1}{\sigma}.$$

So,

$$p_Y(y) = p_Z(\frac{y - \mu}{\sigma}) \, \frac{1}{\sigma}.$$

To get back to our Bernoulli p problem, if we use the Beta prior then our posterior is Beta.

To get the marginal posterior of $o = f(p) = p/(1-p)$, we have

$$h(o) = p = \frac{o}{1+o}, \quad \frac{dp}{do} = \frac{1}{(1+o)^2}.$$

We can now easily compute the prior and posterior densities of o.

$$p(o) = g(p(o))\,\frac{1}{(1+o)^2}.$$

where g is the prior or posterior of p and hence a Beta density.

Monte Carlo

Suppose we want $P(o > 1)$?

Suppose we want $E(o^2)$?

Well, we could figure these out, but it is very easy to do everything by Monte Carlo.

- Get iid draws p_j, $j = 1, 2, \ldots, N$. from the distribution of p.
- Then $o_j = f(p_j)$ are iid draws from the distribution of o.

For distributions like the Beta, a lot of work has gone into coming up with nice algorithms for getting iid draws from the computer.

For N large, $P(o > 1)$ is just the fraction of $o_j > 1$.

$E(o^2)$ is the average of the o_j^2 values.

The density of o is like the histogram of the o_j values.

We will be studying *Markov Chain Monte Carlo* which is a very general Monte Carlo technique which works great in Bayesian problems.