Bayesian Approach to statistics

The Bayesian Paradigm can be seen in some ways as an extra step in the modelling world just as parametric modelling is. We have seen how we could use probabilistic models to infer about some unknown aspect either by confidence intervals or by hypothesis testing.

The motivation for any statistical analyses is that some ``target population'' is not well understood- some aspects of it are unknown or unsure.

The idea in this paradigm is to say that any uncertainty can be modelled in a probabilistic way.

It is true that there are very rarely situations when one doesn't know anything at all, asked to measure the table, you won't want to use a ``pied de coulisse"(callipers) or a 100 yard measuring ribbon.

The probability model that we build can be quite approximate, it reflects one's beliefs and any prior experience we may have, it is described as personal or subjective.

Why ? Because it is different from person to person, examples that are easy to understand are about horse betting, the stock exchange...

So when the uncertainty about the model can be boiled down to a parameter ϑ the Bayesian statistician treats θ as if it were a random variable Θ whose distribution describes that uncertainty.

Elliciting a whole distribution may seem a challenge, in fact it's done by successive events of the type $\Theta \leq \theta$, and does NOT have to be very precise.

A subjective/personal probability is going to be subject to modification upon acquisition of further information supplied by experimanetal data.

Suppose a distribution with density $g(\theta)$ describes one's present uncertainties about some probability model with density $f(x|\theta)$.

Those uncertainties will change with the acquisition of data obtained by doing the experiment modelled by *f*.

Bayes theorem is essential in updating :

$$P(H|data) = \frac{P(data|H)P(H)}{P(data)}$$

The probability of H given the data is called the posterior probability of H, it is posterior to the data. The unconditional probability of *H* : *P*(*H*) is the prior probability of H.

For given data *P*(*data*|*H*) is the likelihood of H.

For given data we often write :

$$P(H|data) \propto P(data|H)P(H)$$

The posterior is proportional to the likelihood time the prior.

If it helps (some people have a better understanding of odds):

$$\frac{P(H|data)}{P(H^c|data)} \propto \frac{P(data|H)}{P(data|H^c)} \frac{P(H)}{P(H^c)}$$

Posterior odds = Likelihood ratio $\times$ Prior odds.

Now these formulae were written as if the rv were discrete for continuous random variables the behaviour is identical replacing probabilities with densities:

Represent the data by a random variable Y:

$$h(\theta) \propto L(\theta)g(\theta)$$

$L(\theta)$ proportional to the probability density of *Y* given θ. In fact we can consider we are studying the joint distribution of two random variables Θ and *Y*.

The marginal distribution of Y is not exhibited, it is the proportionality factor. It can be written :

$$m(y) = \int f(y|\theta)g(\theta)d\theta$$

: One does NOT have to worry too much about the prior because as soon as the data comes in it is `swamped' in the following sense: Two people with divergent prior opinions but reasonably open-minded will be forced into arbitrarily close agreement about future observations by a sufficient amount of data. We will see an example of this later on.

About Priors

``Gentle'' priors reflect some agreed-upon weakness in the available information, for instance before any instruments went to the moon no one had any precise idea about the answer to the question: ``How deep is the dust''. The initial belief was overthrown as soon as any data came back.

When advance information is available the Bayesian method provides a routine way for updating uncertainty when new information comes in.

There are several steps to building a prior:

Calibrating degrees of belief

Suppose I wanted to discover ``Your'' probability that average adult male emperor penguins weigh more than 50 lbs? We will go through comparison experiments:

1. Would you rather bet on getting one green chip out of 1 R 1G or bet on A true?

   Suppose you prefer the latter.

2. Would you rather bet on getting a green chip out of 3G and 1 R ?

....etc... This allows for statements that enable us to bound probabilites.

Another type of thought experiment could be used to build $P[\Theta \leq \theta]$ for an increasing sequence of θ's.

This is not usually how priors are built though because it seems quite an exhaustive process to build up a whole density prior, instead we are going to use families of priors who have easy updating processes with regards to the specific likelihoods at hand.

Beta priors for the Binomial parameter

Example of the Binomial and the Beta.

A little history: From Bayes 1763: A white billiard ball is rolled along a line and we look at where it stops, scale the table from 0 to 1. We suppose that it has a uniform probability of falling anywhere on the line. It stops at a point p. A red billiard ball is then rolled n times under the same uniform assumption. X then denotes the number of times R goes further than W went.

Given X what inference can we make about p ?

Let's say this again in our terminology: We are looking for the posterior distribution of *p* given X. The prior distribution of p is Uniform(0,1)=Beta(1,1).

$X \sim \mathcal{B}(n,p)$

$$P(X=x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$P(a < p < b and X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

$$P(X=x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$$

$$P(a < p < b | X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

Which is the integral of

You are given a set (here taken as finite) and a probability density $p(x), (p(x) \geq 0, \sum p(x) = 1)$. Also given is a set *A* in . The problem is to compute or approximate *p*(*A*). We consider one of the three basic problems used throughout by Feller - the Birthday Problem, the Coupon Collector's problem and the Matching Problem. These will be considered from a Bayesian standpoint.

In order to go further we need to extend what we did before for the binomial and its Conjugate Prior to the multinomial and the the Dirichlet Prior. This is a probability distribution on the *n* simplex

$$\Delta_n = \{\tilde{p} = (p_1, \cdots, p_n),\ p_1 + \cdots + p_n = 1,\ p_i \geq 0\ \}$$

It is a *n*-dimensional version of the beta density. Wilks (1962) is a standard reference for Dirichlet computations.

The Dirichlet has a parameter vector: $\tilde{a} = (a_1, \ldots, a_n)$. Throughout we write $A = a_1 + \cdots + a_n$.

With respect to Lebesgue measure on Δ_n normalised to have total mass 1 the Dirichlet has density:

$$D_{\tilde{a}}(\tilde{x}) = \frac{\Gamma(A)}{\prod \Gamma(a_i)} x_1^{a_1-1} x_2^{a_2-1} \cdots x_n^{a_n-1}$$

The uniform distribution on Δ_n results from choosing all $a_i = 1$. The multinomial distribution corresponding to *k* balls dropped into *n* boxes with fixed probability $(p_1, \cdots, p_n)$ (with the ith box containing k_i balls) is

$$\binom{k}{k_1 \ldots k_n} p_1^{k_1} \cdots p_n^{k_n}$$

If this is averaged with respect to $D_{\tilde{a}}$ one gets the marginal (or Dirichlet/ Multinomial):

$$P(k_1, \ldots, k_n) = P() = \frac{(a_1)_{(k_1)} (a_2)_{(k_2)} \cdots (a_n)_{(k_n)}}{A_{(k)}} \text{ where } \ m_{(j)} \stackrel{\text{def}}{=} m(m+1) \cdots (m + (j -$$

From a more practical point of view there are two simple procedures worth recalling here:

- To pick $\tilde{p}$ from a Dirichlet prior; just pick $X_1, X_2, \ldots, X_n$ independant from gamma densities

$$\frac{e^{-x} x^{a_i - 1}}{\Gamma(a_i)} \text{ and set } p_i = \frac{X_i}{X_1 + \cdots X_n}, 1 \leq i \leq N$$

- To generate sequential samples from the marginal distribution use **Polya's Urn**: Consider an urn containing α_i balls of color *i* (actually fractions are allowed).

  Each time, choose a color *i* with probability proportional to the number of balls of that color in the urn. If *i* is drawn, replace it along with another ball of the same color.

The Dirichlet is a convenient prior because the posterior for $\bar{p}$ having observed $(k_1, \cdots, k_n)$ is Dirichlet with probability $(\alpha_1 + k_1, \cdots, \alpha_n + k_n)$. Zabell (1982) gives a nice account of W.E. Johnson's characterization of the Dirichlet: it is the only prior that predicts outcomes linearly in the past. One frequently used speical case is the symmetric Dirichlet when all $\alpha_i = c > 0$. We denote this prior as D_c.

Monty Hall example of use of using prior information

The best Monty Hall experiment to try

The Let's Make a Deal

As a motivating example behind the discussion of probability, an applet has been developed which allows students to investigate the Let's Make a Deal Paradox. This paradox is related to a popular television show in the 1970's. In the show, a contestant was given a choice of three doors of which one contained a prize. The other two doors contained gag gifts like a chicken or a donkey. After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?

The intuition of most students tells them that each of the doors, the chosen door and the unchosen door, are equally likely to contain the prize so that there is a 50-50 chance of winning with either selection. This, however, is not the case. The probability of winning by using the switching technique is 2/3 while the odds of winning by not switching is 1/3. The easiest way to explain this to students is as follows. The probability of picking the wrong door in the initial stage of the game is 2/3. If the contestant picks the wrong door initially, the host must reveal the remaining empty door in the second stage of the game. Thus, if the contestant switches after picking the wrong door initially, the contestant will win the prize. The probability of winning by switching then reduces to the probability of picking the wrong door in the initial stage which is clearly 2/3.

Despite a very clear explanation of this paradox, most students have a difficulty understanding the problem. It is very difficult to conquer the strong intuition which most students have in this case. As a challenge to students who don't believe the explanation, an instructor may ask the students to actually play the game a number of times by switching and by not switching and to keep track of the relative frequency of wins with each strategy. An applet has developed which allows students to repeatedly play the game and keep track of the results. The applet is given below.

Within the applet, the computer plays the role of the host. Upon loading the applet, students are asked to pick a door by clicking the mouse on the proper region. After the initial pick, the computer reveals an empty door by displaying a picture of a donkey behind one of the numbers. The students then have the option of staying with their initial selection or switching to the remaining door. The computer keeps track of the number of times the game is played with each strategy and the number of times the game is won with each strategy. This information is displayed at the bottom of the applet after each game. Using empirical techniques, the student is then able to see the strategy with the highest probability of winning as the relative frequencies converge to the true probabilities. An instructor may use this game to motivate the ideas of repeated trials as a means for investigating a random phenomena.