LECTURE 7

Example 15. A drug company would like to introduce a drug to reduce acid indigestion. It is desirable to estimate θ, the proportion of the market share that this drug will capture. The company interviews n people and Y of them say that they will buy the drug.

Solution: In the non-Bayesian analysis $\theta \in [0,1]$ and $Y \sim Bin(n, \theta)$.

We know that $\widehat{\theta} = Y/n$ is a very good estimator of θ. It is unbiased, consistent and minimum variance unbiased. Moreover, it is also the maximum likelihood estimator (MLE), and thus asymptotically normal.

A Bayesian may look at the past performance of new drugs of this type. If in the past new drugs tend to capture a proportion between say .05 and .15 of the market, and if all values in between are assumed equally likely, then $\theta \sim Unif(.05, .15)$.

Thus, the prior distribution is given by

$$p(\theta) = \begin{cases} 1/(0.15 - 0.05) = 10, & \text{if} \quad 0.05 \le \theta \le 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

and the likelihood function by

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

The posterior distribution is

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)\, p(y|\theta)\, d\theta} = \begin{cases} \frac{\theta^y (1-\theta)^{n-y}}{\int_{0.05}^{0.15} \theta^y \, (1-\theta)^{n-y} \, d\theta} & \text{if} \quad \theta \in (0.05, 0.15) \\ 0, & \text{otherwise.} \end{cases}$$

§14. ONE-PARAMETER MODELS

A one-parameter model is a class of sampling distributions that is indexed by a single unknown parameter. In this section we discuss Bayesian inference for two one-parameter models: the binomial model and the Poisson model. In addition to being useful statistical tools, these models also provide a simple environment within which we can learn the basics of Bayesian data analysis, including conjugate prior distributions and predictive distributions.

14.1 The binomial model

Happiness data. Each female of age 65 or over in the 1998 General Social Survey was asked whether or not they were generally happy. Let $Y_i = 1$ if respondent i reported being generally happy, and let $Y_i = 0$ otherwise. Since $n = 129$ individuals, then our joint beliefs about Y_1,...,Y_{129} are well approximated by the model that, conditional on θ, the Y_is are i.i.d. binary random variables with expectation $\theta = \sum_{i=1}^{129} Y_i/129$.

The last item says that the probability for any potential outcome $\{y_1, ..., y_{129}\}$, conditional on θ, is given by

$$P(y_1, ..., y_{129}|\theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}.$$

What remains to be specified is our prior distribution.

A uniform prior distribution. The parameter is some unknown number between 0 and 1. Suppose our prior information is such that all subintervals of $[0, 1]$ having the same length also have the same probability. Symbolically,

$$P(a \leq \theta \leq b) = P(a + c \leq \theta \leq b + c) \quad \text{for} \quad 0 \leq a < b < b + c \leq 1.$$

This condition implies that our density for θ must be the uniform density:

$$p(\theta) = 1 \quad for \quad all \quad \theta \in [0, 1].$$

For this prior distribution and the above sampling model, Bayes rule gives

$$p(\theta|y_1, ..., y_{129}) = \frac{p(y_1, ..., y_{129}|\theta)\, p(\theta)}{p(y_1, ..., y_{129})} =$$

$$= p(y_1, ..., y_{129}|\theta) \cdot \frac{1}{p(y_1, ..., y_{129})} \propto p(y_1, ..., y_{129}|\theta).$$

The last line says that in this particular case $p(\theta|y_1, ..., y_{129})$ and $p(y_1, ..., y_{129}|\theta)$ are proportional to each other as functions of θ. This is because the posterior distribution is equal to $p(y_1, ..., y_{129}|\theta)$ divided by something that does not depend on θ. This means that these two functions of θ have the same shape, but not necessarily the same scale.

Data and posterior distribution.

129 individuals surveyed;

118 individuals report being generally happy (91%);

11 individuals do not report being generally happy (9%).

The probability of these data for a given value of θ is

$$p(y_1, ..., y_{129}|\theta) = \theta^{118}(1-\theta)^{11}.$$

Our result above about proportionality says that the posterior distribution $p(\theta|y_1, ..., y_{129})$ will have the same shape as this function, and so we know that the true value of θ is very likely to be near 0.91. However, we will often want to be more precise than this, and we will need to know the scale of $p(\theta|y_1, ..., y_n)$ as well as the shape. From Bayes rule, we have

$$p(\theta|y_1, ..., y_{129}) = \theta^{118}(1-\theta)^{11} \cdot \frac{p(\theta)}{p(y_1, ..., y_{129})} = \theta^{118}(1-\theta)^{11} \cdot \frac{1}{p(y_1, ..., y_{129})}.$$

It turns out that we can calculate the scale or normalizing constant $\frac{1}{p(y_1,...,y_{129})}$ using the following result from calculus:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

(the value of the gamma function $\Gamma(x)$ for any number $x > 0$ can be looked up in a table, or with *R* using the *gamma*() function). How does the calculus result help us compute $p(\theta|y_1, ..., y_{129})$? Lets recall what we know about $p(\theta|y_1, ..., y_{129})$:

(a) $\int_0^1 p(\theta|y_1, ..., y_{129})\, d\theta = 1$, since all probability distributions integrate or sum to 1;

(b) $p(\theta|y_1, ..., y_{129}) = \frac{\theta^{118}\,(1-\theta)^{11}}{p(y_1,...,y_{129})}$, from Bayes rule.

Therefore,

$$1 = \int_0^1 p(\theta|y_1, ..., y_{129})\, d\theta \qquad \text{using (a)}$$

$$1 = \int_0^1 \frac{\theta^{118}\,(1-\theta)^{11}}{p(y_1, ..., y_{129})}\, d\theta \qquad \text{using (b)}$$

$$1 = \frac{1}{p(y_1, ..., y_{129})} \int_0^1 \theta^{118}\,(1-\theta)^{11}\, d\theta$$

$$1 = \frac{1}{p(y_1, ..., y_{129})} \frac{\Gamma(119)\,\Gamma(12)}{\Gamma(131)} \qquad \text{using the calculus result, and so}$$

$$p(y_1, ..., y_{129}) = \frac{\Gamma(119)\,\Gamma(12)}{\Gamma(131)}.$$

You should convince yourself that this result holds for any sequence $y_1, ..., y_{129}$ that contains 118 ones and 11 zeros. Putting everything together, we have

$$p(\theta | y_1, ..., y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\,\Gamma(12)} \theta^{118} \, (1-\theta)^{11} =,$$

which we will write as

$$= \frac{\Gamma(131)}{\Gamma(119)\,\Gamma(12)} \theta^{119-1} \, (1-\theta)^{12-1}.$$

This density for θ is called a beta distribution with parameters $a = 119$ and $b = 12$, which can be calculated, plotted and sampled from in R using the function $dbeta()$.

The beta distribution. An uncertain quantity θ, known to be between 0 and 1, has a $beta(a, b)$ distribution if

$$f(\theta) = beta(\theta, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \theta^{a-1} \, (1-\theta)^{b-1} \qquad \text{for} \qquad 0 \le \theta \le 1.$$

For such a random variable,

$$mode[\theta] = \frac{a-1}{(a-1)+(b-1)} \qquad \text{if} \qquad a > 1 \quad \text{and} \quad b > 1;$$

$$E[\theta] = \frac{a}{a+b};$$

$$Var[\theta] = \frac{ab}{(a+b+1)(a+b)^2} = \frac{E[\theta] \cdot E[1-\theta]}{a+b+1}.$$

For our data on happiness in which we observed $(Y_1, ..., Y_{129}) = (y_1, ..., y_{129})$ with $\sum\limits_{i=1}^{129} y_i = 118$,

$$mode[\theta | y_1, ..., y_{129}] = 0.915;$$

$$E[\theta | y_1, ..., y_{129}] = 0.908;$$

$$sd[\theta | y_1, ..., y_{129}] = 0.025.$$

14.1.1 Inference for exchangeable binary data.

Posterior inference under a uniform prior.

If $Y_1, ..., Y_n|\theta$ are i.i.d. binary(θ), we showed that

$$p(\theta|y_1, ..., y_n) = \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \cdot p(\theta)}{p(y_1, ..., y_n)}.$$

If we compare the relative probabilities of any two θ-values, say θ_a and θ_b, we see that

$$\frac{p(\theta_a|y_1, ..., y_n)}{p(\theta_b|y_1, ..., y_n)} = \frac{\theta_a^{\sum y_i}(1-\theta_a)^{n-\sum y_i} \cdot p(\theta_a)/p(y_1, ..., y_n)}{\theta_b^{\sum y_i}(1-\theta_b)^{n-\sum y_i} \cdot p(\theta_b)/p(y_1, ..., y_n)} =$$

$$= \left(\frac{\theta_a}{\theta_b}\right)^{\sum y_i} \left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum y_i} \frac{p(\theta_a)}{p(\theta_b)}.$$

This shows that the probability density at θ_a relative to that at θ_b depends on $y_1, ..., y_n$ only through $\sum\limits_{i=1}^{n} y_i$. From this, you can show that

$$P(\theta \in A|Y_1 = y_1, ..., Y_n = y_n) = P\left(\theta \in A|\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} y_i\right).$$

We interpret this as meaning that $\sum\limits_{i=1}^{n} Y_i$ contains all the information about θ available from the data, and we say that $\sum\limits_{i=1}^{n} Y_i$ is a sufficient statistic for θ and $p(y_1, ..., y_n|\theta)$. The word sufficient is used, because it is sufficient to know $\sum Y_i$ in order to make inference about θ. In this case, where $Y_1, ..., Y_n|\theta$ are i.i.d. binary (θ) random variables, the sufficient statistic $Y = \sum\limits_{i=1}^{n} Y_i$ has a binomial distribution with parameters (n, θ).

The Binomial distribution. A random variable $Y \in \{0, 1, ..., n\}$ has a binomial (n, θ) distribution if

$$P(Y = y|\theta) = dbinom(y, n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \qquad y \in \{0, 1, ..., n\}.$$

For a binomial(n, θ) random variable,

$$E[Y|\theta] = n\,\theta;$$

$$Var[Y|\theta] = n\theta(1-\theta).$$

Posterior inference under a uniform prior distribution

Having observed $Y = y$ our task is to obtain the posterior distribution of θ:

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)} = \frac{\binom{n}{y}\,\theta^y(1-\theta)^{n-y}\,p(\theta)}{p(y)} = c(y)\,\theta^y\,(1-\theta)^{n-y}\,p(\theta)$$

where $c(y)$ is a function of y and not of θ. For the uniform distribution with $p(\theta) = 1$, we can find out what $c(y)$ is using our calculus trick:

$$1 = \int_0^1 c(y)\,\theta^y(1-\theta)^{n-y}\,d\theta$$

$$1 = c(y)\int_0^1 \theta^y(1-\theta)^{n-y}\,d\theta$$

$$1 = c(y)\,\frac{\Gamma(y+1)\,\Gamma(n-y+1)}{\Gamma(n+2)}.$$

The normalizing constant $c(y)$ is therefore equal to $\frac{\Gamma(n+2)}{\Gamma(y+1)\,\Gamma(n-y+1)}$, and we have

$$p(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\,\Gamma(n-y+1)}\theta^y\,(1-\theta)^{n-y} =$$

$$\frac{\Gamma(n+2)}{\Gamma(y+1)\,\Gamma(n-y+1)}\theta^{(y+1)-1}\,(1-\theta)^{(n-y+1)-1} = beta(y+1, n-y+1).$$

Recall the happiness example, where we observed that $Y \equiv \sum Y_i = 118$:

$$n = 129, \quad Y \equiv \sum Y_i = 118 \quad \Rightarrow \quad \theta\,|\,\{Y = 118\} \sim beta(119, 12).$$

This confirms the sufficiency result for this model and prior distribution, by showing that if $\sum y_i = y = 118$,

$$p(\theta\,|y_1, \ldots, y_n) = p(\theta|y) = beta(119, 12).$$

In other words, the information contained in $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ is the same as the information contained in $\{Y = y\}$, where $Y = \sum Y_i$ and $y = \sum y_i$.

Posterior distributions under beta prior distributions.

The uniform prior distribution has $p(\theta) = 1$ for all $\theta \in [0, 1]$. This distribution can be thought of as a beta prior distribution with parameters $a = 1$, $b = 1$:

$$p(\theta) = \frac{\Gamma(2)}{\Gamma(1)\,\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{1 \times 1}1 \times 1 = 1.$$

Note that $\Gamma(x+1) = x! = x \times (x-1) \times \cdots \times 1$ if x is a positive integer, and $\Gamma(1) = 1$ by convention.

Therefore, we saw that

$$\text{if} \quad \theta \sim beta(1,1) \quad (\text{uniform}), Y \sim \text{binomial}(n,\theta), \text{then } \{\theta\,|Y=y\} \sim beta(1+y, 1+n-y),$$

and so to get the posterior distribution when our prior distribution is $beta(a=1,\, b=1)$, we can simply add the number of 1s to the a parameter and the number of 0s to the b parameter.

Does this result hold for arbitrary beta priors? Lets find out: Suppose $\theta \sim beta(a,b)$ and $Y\,|\theta \sim binomial(n,\theta)$. Having observed $Y=y$,

$$p(\theta|y) = \frac{p(\theta)\,p(y|\theta)}{p(y)} = \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^{y}(1-\theta)^{n-y} =$$

$$= c(n,y,a,b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1} =$$

$$= beta(\theta, a+y, b+n-y).$$

It is important to understand the last two lines above: The second to last line says that $p(\theta|y)$ is, as a function of θ, proportional to $\theta^{a+y-1} \times (1-\theta)^{b+n-y-1}$. This means that it has the same shape as the beta density $beta(\theta, a+y, b+n-y)$. But we also know that $p(\theta|y)$ and the beta density must both integrate to 1, and therefore they also share the same scale.

These two things together mean that $p(\theta|y)$ and the beta density are in fact the same function. Throughout my lectures we will use this trick to identify posterior distributions: We will recognize that the posterior distribution is proportional to a known probability density, and therefore must equal that density.

CONJUGACY.

We have shown that a beta prior distribution and a binomial sampling model lead to a beta posterior distribution. To reflect this, we say that the class of beta priors is ***conjugate*** for the binomial sampling model.

Definition 5 (Conjugate). A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}.$$

Conjugate priors make posterior calculations easy, but might not actually represent our prior information. However, mixtures of conjugate prior distributions are very flexible and are computationally tractable.

Combining information

If $\theta|\{Y=y\} \sim beta(a+y, b+n-y)$, then

$$E[\theta|y] = \frac{a+y}{a+b+n}, \quad mode[\theta|y] = \frac{a+y-1}{a+b+n-2}, \quad Var[\theta|y] = \frac{E[\theta|y]\,E[1-\theta|y]}{a+b+n+1}.$$

The posterior expectation $E[\theta|y]$ is easily recognized as a combination of prior and data information:

$$E[\theta|y] = \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} =$$

$$= \frac{a+b}{a+b+n} \times prior \quad expectation + \frac{n}{a+b+n} \times data \quad average.$$

For this model and prior distribution, the posterior expectation (also known as the posterior mean) is a weighted average of the prior expectation and the sample average, with weights proportional to $a+b$ and n respectively. This leads to the interpretation of a and b as prior data:

$$a \approx prior \quad number \quad of \quad 1s,$$

$$b \approx prior \quad number \quad of \quad 0s,$$

$$a+b \sim prior \quad sample \quad size.$$

If our sample size n is larger than our prior sample size $a+b$, then it seems reasonable that a majority of our information about θ should be coming from the data as opposed to the prior distribution. This is indeed the case: For example, if $n >> a+b$, then

$$\frac{a+b}{a+b+n} \approx 0, \quad E[\theta|y] \approx \frac{y}{n}, \quad Var[\theta|y] \approx \frac{1}{n}\frac{y}{n}(1-\frac{y}{n}).$$

PREDICTION.

An important feature of Bayesian inference is the existence of a predictive distribution for new observations. Reverting for the moment to our notation for binary data, let $y_1, \ldots, y_n$ be the outcomes from a sample of n binary random variables, and let $\tilde{Y} \in \{0,1\}$ be an additional outcome from the same population that has yet to be observed. The predictive distribution of $\tilde{Y}$ is the conditional distribution of $\tilde{Y}$ given $\{Y_1 = y_1, \ldots, Y_n = y_n\}$. For conditionally i.i.d. binary variables this distribution can be derived from the distribution of $\tilde{Y}$ given θ and the posterior distribution of θ:

$$P(\tilde{Y}=1|y_1,...,y_n) = \int P(\tilde{Y}=1,\theta|y_1,...,y_n)d\theta =$$

$$= \int P(\tilde{Y} = 1|\theta, y_1, \ldots, y_n)\, p(\theta|y_1, ..., y_n)\, d\theta = \int \theta\, p(\theta|y_1, ..., y_n)\, d\theta =$$

$$= E[\theta|y_1, ..., y_n] = \frac{a + \sum_{i=1}^n y_i}{a + b + n}$$

$$P(\tilde{Y} = 0|y_1, ..., y_n) = 1 - E[\theta|y_1, ..., y_n] = \frac{b + \sum_{i=1}^n (1 - y_i)}{a + b + n}.$$

You should note two important things about the predictive distribution:

1. The predictive distribution does not depend on any unknown quantities. If it did, we would not be able to use it to make predictions.

2. The predictive distribution depends on our observed data. In this distribution, $\tilde{Y}$ is not independent of $Y_1, ..., Y_n$. This is because observing $Y_1, ..., Y_n$ gives information about θ, which in turn gives information about $\tilde{Y}$. It would be bad if $\tilde{Y}$ were independent of $Y_1, ..., Y_n$ - it would mean that we could never infer anything about the unsampled population from the sample cases.

Example 16. The uniform prior distribution, or *beta*$(1, 1)$ prior, can be thought of as equivalent to the information in a prior data set consisting of a single "1" and a single "0". Under this prior distribution,

$$P(\tilde{Y} = 1|Y = y) = E[\theta|Y = y] = \frac{2}{2+n}\frac{1}{2} + \frac{n}{2+n}\frac{y}{n},$$

$$mode(\theta|Y = y) = \frac{y}{n},$$

where $Y = \sum_{i=1}^n Y_i$. Does the discrepancy between these two posterior summaries of our information make sense? Consider the case in which $Y = 0$, for which $mode(\theta|\, Y = 0) = 0$ but $P(\tilde{Y} = 1|Y = 0) = 1/(2 + n)$.

Example 17. Consider again the problem of Example 12, in which the probability of pile failure at a load of 300 tons is of concern; this time, however assume that the probability p is a continuous random variable. If there is no (prior) factual information on p, a uniform prior distribution may be assumed (known as diffuse prior), namely,

$$f(p) = \begin{cases} 0 & \text{if} \quad p \notin (0, 1) \\ 1 & \text{if} \quad 0 \le p \le 1 \end{cases}.$$

a) On the basis of the single test, the likelihood function is simply the probability of the event $A =$ capacity of test pile less than 300 tons, which is simply p. Therefore, likelihood function is proportional to p:

$$P(X_1 = 0/p) = k\,p,$$

and the constant k is equal to

$$k = \left[\int_0^1 pdp\right]^{-1} = 2$$

Hence posterior distribution of p is

$$P(p/X_1 = 0) = \begin{cases} 0 & \text{if} \quad p \notin (0,1) \\ 2p & \text{if} \quad 0 \le p \le 1 \end{cases},$$

that is beta distribution with parameters $a = 2$ and $b = 1$:

$$\frac{\Gamma(3)}{\Gamma(2)\,\Gamma(1)}\,p = 2\,p.$$

Thus the Bayesian estimator of p is

$$p^* = E(p/X_1 = 0) = \frac{a}{a+b} = \frac{2}{3}.$$

b) If a sequence of n piles were tested, out of which r piles failed at loads less that the maximum test load, then the likelihood function is the probability of observing r failures among the n piles tested. If the failure probability of each pile is p, and statistical independence is assumed between piles, the likelihood function would be

$$P(\sum_{i=1}^{n} X_i/p) = \binom{n}{r} p^r (1-p)^{n-r}.$$

Then with the uniform prior, the posterior distribution of p becomes

$$f(p/\sum_{i=1}^{n} X_i) = k\,\binom{n}{r} p^r (1-p)^{n-r}, p \in [0,1],$$

where

$$k = \left[\int_0^1 \binom{n}{r} p^r (1-p)^{n-r}\right]^{-1}.$$

In our case we have beta distribution with parameters $a = 1 + r$ and $b = 1 + n - r$. Therefore, the Bayesian estimator of p is

$$p^* = E(p/\sum_{i=1}^{n} X_i = r) = \frac{r+1}{r+1+1+n-r} = \frac{r+1}{n+2}.$$

From this result, we may observe that as a number of tests n increases (with the ratio r/n remaining constant), the Bayesian estimate for p approaches that of the classical estimate; that is

$$\frac{r+1}{n+2} \to \frac{r}{n}$$

for large n.