1 Bayesian inference

1.1 Bayes theorem

We begin by refreshing our memories on the subject of conditional probability. We define the probability of A given B has occurred, written $P[A|B]$ as

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

While conditional probabilities can have interesting philosophical implications they also allow one to do calculations. Thus

$$P[A] = P[A|B]P[B] + P[A|B^c]P[B^c]$$

or more generally if $\cup_{i=1}^{n} B_i = \Omega$ then

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i]$$

We also have Bayes Theorem

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]}$$

this implies that

$$P[B|A] \propto P[A|B]P[B] \tag{1}$$

A slight generalization of the Bayes theorem is by considering the events $\cup_{i=1}^{n} B_i = \Omega$,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}, \quad i = 1, \ldots, k.$$

Example 1.1. *In a population with high risk to be affected from HIV we try an new diagnostic method. The 10% percent of the population is believed to be affected by HIV. From the results of the test, the test is positive for 90% of the people that they are really affected by HIV and negative for 85% of the people that the are not affected by HIV. Which are the probabilities to obtain false negative and false positive results?*

Suppose that A is the event of someone to be affected by HIV and B the event that the test is positive. Then $P(A) = 0.1, P(B|A) = 0.9$ *and* $P(B^c|A^c) = 0.85$.

$$P(\textit{the test is false positive}) = P(A^c|B) = \frac{P(B|A^c)P(A^c)}{P(B)} =$$

$$\frac{(1 - P(B^c|A^c))P(A^c)}{P(B|A^c)P(A^c) + P(B|A)P(A)} = \frac{(1 - 0.85) \times 0.90}{0.15 \times 0.90 + 0.90 \times 0.10} = 0.6$$

Accordingly,

$$P(\textit{the test is false negative}) = P(A|B^c) = \frac{P(B^c|A)P(A)}{P(B^c)} =$$

$$\frac{(1 - P(B|A))P(A)}{P(B^c|A^c)P(A^c) + P(B^c|A)P(A)} = \frac{(1 - 0.9) \times 0.10}{0.85 \times 0.90 + 0.10 \times 0.10} = 0.0129$$

1.1.1 The Continuous Analogue

In the continuous case we usually consider distributions with a joint density. In the two variable case with a density $f(x,y)$ so

$$P[(X,Y) \in C] = \int_C f(x,y)dxdy$$

Here we define the marginals as

$$f_x(x) = \int f(x,y)dy \text{ and } f_y(y) = \int f(x,y)dx$$

The conditional distributions are

$$f(x|y) = \frac{f(x,y)}{f_y(y)} \text{ and } f(y|x) = \frac{f(x,y)}{f_x(x)}$$

Thus

$$f(y|x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x|y)f_y(y)}{f_x(x)}$$

or $f(y|x)$ is proportional to $f(x|y)f_y(y)$

$$f(y|x) \propto f(x|y)f_y(y)$$

1.2 Subjective Inference

Despite the careful derivations of frequentist inference many people would argue that one always starts an experiment with a prior belief and this is then modified by experience. This subjective view is generally known as Bayesian and can be developed to give an alternative approach to inference.

Example 1.2. *Consider three statistical experiments:*

1. *A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In all 3 trials, she distinguished correctly.*
2. *A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all 3 trials, she correctly determined which was poured first.*
3. *A drunkard claims to be able to predict the outcome of a flip of a fair coin. In all 3 trials, he guessed the outcomes correctly.*

Our interest is the the probability of the person answering correctly. The distribution for this case is binomial $B(3,p)$. The maximum likelihood estimator is $\hat{p} = \frac{X}{n} = \frac{3}{3} = 1$. Is this a satisfactory analysis for all the three experiments?

No reasons to doubt the conclusion for Experiment 1 but we have serious doubt for Experiment 3.

Suppose we have some parameter of interest θ. Unlike classical inference we have some subjective, or prior belief about this parameter. To quantify this belief we construct a probability distribution $f(\theta)$ which we call the *prior distribution*.

We perform some experiment with the aim of gaining *more* information about θ and suppose the result of our experiment is some data

$$(x_1, x_2, x_3, \ldots, x_n)^T = \boldsymbol{x}$$

This has a likelihood of the form

$$f(\boldsymbol{x}|\theta)$$

Note we use the above notation since we perform the experiment *given our prior belief about* θ.

We now use Bayes theorem to modify our belief, that is to derive a new distribution for θ. If θ is continuous

$$f(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)f(\theta)}{\int f(\boldsymbol{x}|\theta)f(\theta)d\theta},$$

and if θ is discrete,

$$f(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)f(\theta)}{\sum f(\boldsymbol{x}|\theta_j)f(\theta_j)}.$$

Since in the denominator we integrate or sum with respect to θ the denominator is a function of $\boldsymbol{x}$. Therefore for given data $\boldsymbol{x}$, the denominator is constant, the so-called constant of proportionality. Accordingly an alternative way of presenting the Bayes theorem is

$$f(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)f(\theta)$$

where $f(\theta|\boldsymbol{x})$ is the posterior distribution, that is the distribution of θ after our view has been modified by experiment. This is just an analog of Bayes theorem as discussed above.

$$p[A|B] = p[B|A]\,p[A]\,/\,p[B]$$

The basic idea:

1. Treat θ as a random variable.
2. The prior $f(\theta)$ reflects the beliefs on the true value of θ before taking any observations.
3. The goal of Bayesian inference is to update the beliefs from prior $f(\theta)$ to posterior $f(\theta|\boldsymbol{x})$, after taking observations from a random sample $(x_1, x_2, x_3, \ldots, x_N)^T = \boldsymbol{x}$.

Example 1.3. *A coin is tossed 70 times, the number of heads is 34. The probability of being a head is some unknown value* θ*. We have some prior belief about* θ*, i.e.,* $E(\theta) = 0.4$ *and* $Var(\theta) = 0.02$ *which we need to quantify. One way of doing this is to use a probability distribution to quantify our belief. This distribution* $f(\theta)$ *is the prior distribution is assumed to contain our prior beliefs about the parameter* θ*. A possible distribution for* θ *is the Beta distribution with density*

$$f(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a,b)} \text{ for } 0 \le \theta \le 1. \tag{2}$$

Since the mean and variance of the Beta distribution for the prior distribution are known i.e.,

$$E(\theta) = \frac{a}{a+b} = 0.4 \qquad Var(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = 0.02.$$

We can easily derive the parameters of the Beta distribution a, b by solving the above system of equations. So the resulted parameters are,

$$a = \frac{(1-m)m^2}{u} - m \quad and \quad q = \frac{(1-m)^2 m}{u} - (1-m),$$

where $m = E(\theta)$ and $u = Var(\theta)$. For our example the parameters are $a = 4.4$ and $b = 6.6$. When we conduct the experiment of tossing coin n times we have the probability, (likelihood when x the number of heads is known)

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Using Bayes theorem we have

$$f(\theta|x) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x} f(\theta)$$

so if we had chosen

$$f(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a,b)}$$

we would have

$$f(\theta|x) \propto \theta^{(a+x-1)}(1-\theta)^{(N-x+b-1)}$$

We know that $f(\theta|x)$ is a distribution function and since that

$$\theta|x \sim Beta(a+x, b+n-x).$$

The constant of proportionality is obtained easily since we know that the distribution must integrate to 1. This latter function is the posterior distribution of the parameter θ given the data - it is our belief given the data.

So if we start with a prior which is

$$\theta \sim Beta(4.4, 6.6)$$

and we toss a coin 70 times and observe 34 heads then

$$\theta|x \sim Beta(38.4, 42.6).$$

We will study now how the data change our prior beliefs by comparing the expected values for the prior and posterior distributions:

$$E(\theta) = \frac{a}{a+b} \qquad E(\theta|x) = \frac{a+x}{a+b+n}$$

In our example they are:

$$E(\theta) = 0.4 \qquad E(\theta|x) = 0.474.$$

So after taking observations from a random sample the prior estimate increases from 0.4 to 0.474. Note that using the standard maximum likelihood estimation the MLE based on the data is $\frac{x}{n} = 0.486$. *Summing up the posterior is a combination of the data and the prior belief.*

Example 1.4. *Drinks dispensed by a vending machine may overflow. Suppose the prior distribution of p the proportion that overflow is*

p	*0.05*	*0.10*	*0.15*
$f(p)$	*0.30*	*0.50*	*0.20*

If two of the next nine drinks dispensed overflow find the posterior distribution of p .
Assuming independence the likelihood given *the experiment is*

$$f(x|p) = \binom{9}{2} p^2(1-p)^7$$

Thus the posterior is

$$f(p|x) = \frac{\binom{9}{2} p^2(1-p)^7 \times f(p)}{\sum_{j=1}^{3} \binom{9}{2} p_j^2(1-p_j)^7 \times f(p_j)}$$

This gives

$$f(0.05|x) \propto 0.05^2(0.95)^7 \times 0.3$$
$$f(0.10|x) \propto 0.10^2(0.90)^7 \times 0.5$$
$$f(0.05|x) \propto 0.15^2(0.85)^7 \times 0.2$$

or pulling all the results together

p	*0.05*	*0.10*	*0.15*
$f(p)$	*0.30*	*0.50*	*0.20*
$f(p\|x)$	*0.12*	*0.55*	*0.33*

1.3 Choice of prior

The Prior Distributions are the most critical and most criticized point of Bayesian analysis:

- The prior distribution is the key to Bayesian inference.
- In practice, it seldom occurs that the available prior information is precise enough to lead to an exact determination of the prior distribution.
- There is no such thing as the prior distribution.
- The prior is a tool summarizing available information as well as uncertainty related with this information.
- Ungrounded prior distributions produce unjustified posterior inference.

1.4 Conjugate class

The other difficulty is the integration required to obtain the posterior density, i.e., the calculation of the constant of the proportionality. We can overcome this to some extent by choosing a prior in sufficiently clever way. It is not very restrictive to choose a prior from a conjugate class. Conjugate priors are specific parametric families with analytical properties.

Definition 1.1. *A family $\mathscr{F}$ of probability distributions on θ is conjugate for a likelihood function $f(x|\theta)$ if, for every $f \in \mathscr{F}$, the posterior distribution $f(\theta|x)$ also belongs to $\mathscr{F}$.*

Switching from prior to posterior distribution is reduced to an updating of the corresponding parameters, see Example 1.3 where the Beta distribution for binomial likelihood as you can see is a conjugate prior. The justification of the use of conjugate priors is mainly their tractability and simplicity along with the preservation of the structure of $f(\theta|x)$.

The following table represent the most usual conjugate priors.

$f(x\|\theta)$	$f(\theta)$	$f(\theta\|x)$
$x \sim B(n,\theta)$	$Beta(a,b)$	$Beta(a+x, b+n-x)$
$x_1,\ldots,x_n \sim P(\theta)$	$Gamma(a,1/b)$	$Gamma\big(a+\sum x_i, 1/(b+n)\big)$
$x_1,\ldots,x_n \sim Gamma(k,1/\theta)$ k is known	$Gamma(a,1/b)$	$Gamma\big(a+nk, 1/(b+\sum x_i)\big)$
$x_1,\ldots,x_n \sim Geom(\theta)$	$Beta(a,b)$	$Beta(a+n, b+\sum x_i - n)$
$x \sim NB(r,\theta)$	$Beta(a,b)$	$Beta(a+r, b+x-r)$
$x_1,\ldots,x_n \sim N(\theta,1/\tau)$ $\tau isknown$	$N(b,1/c)$	$N(\frac{cb+n\tau\bar{x}}{c+n\tau}, \frac{1}{c+n\tau})$

Table 1: List of conjugate priors.

Example 1.5. *Suppose a random sample $\boldsymbol{x} \sim Gamma(k,1/\theta)$, where k is known and $\theta \sim Gamma(a,1/b)$. Show that*

$$\theta|\boldsymbol{x} \sim Gamma\big(a+nk, 1/(b+\sum x_i)\big).$$

The likelihood is,

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{\theta^k}{\Gamma(k)} x_i^{k-1} \exp(-\theta x_i) \\ &= \prod_{i=1}^{n} \frac{x_i^{k-1}}{\Gamma(k)} \theta^k \exp(-\theta x_i) \\ &\propto \prod_{i=1}^{n} \theta^k \exp(-\theta x_i) \\ &= \theta^{nk} \exp(-\theta \sum x_i) \end{aligned}$$

So

$$\begin{aligned} f(\theta|\boldsymbol{x}) &\propto f(\theta) \times f(\boldsymbol{x}|\theta) \\ &\propto \theta^{a-1} \exp(-b\theta) \times \theta^{nk} \exp(-\theta \sum x_i) \\ &\propto \theta^{a+nk-1} \exp\left\{-(b+\sum x_i)\theta\right\} \end{aligned}$$

Therefore

$$\theta|\boldsymbol{x} \sim Gamma\left(a+nk, 1/(b+\sum x_i)\right).$$

Example 1.6. *Suppose a random sample $\boldsymbol{x} \sim P(\theta)$ of size n and $\theta \sim Gamma(a, 1/b)$. Show that*

$$\theta|\boldsymbol{x} \sim Gamma\left(a+\sum x_i, 1/(b+n)\right).$$

The likelihood is,

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{\exp(-\theta)\theta^{x_i}}{x_i!} \\ &\propto \prod_{i=1}^{n} \exp(-\theta)\theta^{x_i} \\ &= \exp(-n\theta)\theta^{\sum x_i} \end{aligned}$$

So

$$\begin{aligned} f(\theta|\boldsymbol{x}) &\propto f(\theta) \times f(\boldsymbol{x}|\theta) \\ &\propto \theta^{a-1} \exp(-b\theta) \times \exp(-n\theta)\theta^{\sum x_i} \\ &\propto \theta^{a+\sum x_i-1} \exp\{-(b+n)\theta\} \end{aligned}$$

Therefore

$$\theta|\boldsymbol{x} \sim Gamma\left(a+\sum x_i, 1/(b+n)\right).$$

1.4.1 Existence of conjugate priors

Under the assumption that the conjugate priors do not contradict with our prior beliefs and given that a such family of distribution exists the computations could be simplified. Although in which cases a family of conjugate priors can be obtained?

The only case that the conjugate priors can be derived easily is for models of the exponential family of distributions.

Definition 1.2. *The family of distributions:*

$$f(x|\theta) = h(x)g(\theta)\exp\big(t(x)c(\theta)\big)$$

is called an exponential family for the functions h, g, t, c such as

$$\int f(x|\theta)dx = g(\theta)\int h(x)\exp\big(t(x)c(\theta)\big)dx = 1.$$

The exponential family includes the exponential, the Poisson, the Gamma with one parameter, the binomial and the normal distribution with known variance.

Suppose the prior distribution $f(\theta)$, then:

$$\begin{aligned} f(\theta|\boldsymbol{x}) &\propto f(\theta)\times f(\boldsymbol{x}|\theta) \\ &= f(\theta)\times\prod_{i=1}^{n} h(x_i)g(\theta)\exp\big(t(x_i)c(\theta)\big) \\ &= f(\theta)\times\big(\prod_{i=1}^{n} h(x_i)\big)g(\theta)^n\exp\big(\sum t(x_i)c(\theta)\big) \\ &\propto f(\theta)g(\theta)^n\exp\big(\sum t(x_i)c(\theta)\big) \end{aligned}$$

So if we choose

$$f(\theta)\propto g(\theta)^d\exp\big(bc(\theta)\big) \tag{3}$$

then the following posterior distribution will be obtained:

$$\begin{aligned} f(\theta|\boldsymbol{x}) &\propto g(\theta)^{n+d}\exp\big((\sum t(x_i)+b)c(\theta)\big) \\ &= g(\theta)^{\tilde{d}}\exp\big(\tilde{b}c(\theta)\big), \end{aligned}$$

which belongs to the same family of distributions with the prior distribution with adjusted parameters. All the examples of conjugate priors that we have seen were obtained in this way.

Example 1.7. *Consider the binomial distribution. The probability mass function is given by,*

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\big(\frac{\theta}{1-\theta}\big)^x \\ &= \binom{n}{x}(1-\theta)^n\exp\Big(\log\Big(\big(\frac{\theta}{1-\theta}\big)^x\Big)\Big) \\ &= \binom{n}{x}(1-\theta)^n\exp\Big(x\log\big(\frac{\theta}{1-\theta}\big)\Big). \end{aligned}$$

For Definition 1.2,

$$\begin{aligned} h(x) &= \binom{n}{x} \\ g(\theta) &= (1-\theta)^n \\ t(x) &= x \\ c(\theta) &= \log\big(\frac{\theta}{1-\theta}\big) \end{aligned}$$

Therefore we can construct a conjugate prior, see Eq. 3 of the form,

$$\begin{aligned} f(\theta) &\propto g(\theta)^d\exp\big(bc(\theta)\big) \\ &= \big((1-\theta)^n\big)^d\exp\Big(b\log\big(\frac{\theta}{1-\theta}\big)\Big) \\ &= (1-\theta)^{nd-b}\theta^b, \end{aligned}$$

which is member of Beta distributions.

1.5 Noninformative priors

What if all we know is that we know "nothing"? In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$.

Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.

1.5.1 The Jeffreys' prior

One perspective on defining the noninformative prior distribution is the invariance to '1-1' transformations of the parameters. This leads to the Jeffreys' prior which is based on Fisher information

$$I(\theta) = -E\Big(\frac{d^2 \log f(x|\theta)}{d\theta^2}\Big) = E\Big\{\Big(\frac{d \log f(x|\theta)}{d\theta}\Big)^2\Big\}.$$

Definition 1.3. *The prior distribution of Jeffreys is defined as*

$$f_0(\theta) \propto |I(\theta)|^{1/2}.$$

Proposition 1.1. *The prior distribution of Jeffreys is invariant under transformations of the parameters i.e., if* $\phi = g(\theta)$ *then*

$$f_0(\theta) = f_0(\phi)\Big|\frac{\partial \phi}{\partial \theta}\Big|.$$

Example 1.8. *Suppose that* $(x_1, x_2, x_3, \ldots, x_n)^T = \boldsymbol{x} \sim N(\theta, \sigma^2)$*, where* σ^2 *is known. Then*

$$f(\boldsymbol{x}|\theta) \propto \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right)$$

and

$$\log f(\boldsymbol{x}|\theta) \propto -\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}.$$

Furthermore:

$$\begin{aligned} I(\theta) &= -E\Big(\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\Big) \\ &= E\Big(\frac{n}{\sigma^2}\Big) \\ &= \frac{n}{\sigma^2}. \end{aligned}$$

Therefore $f_0(\theta) \propto 1$.

Example 1.9. *Suppose that* $x|\theta \sim B(n, \theta)$*. Then*

$$f(x|\theta) \propto \theta^x (1-\theta)^{n-x}$$

and

$$\log f(\boldsymbol{x}|\theta) \propto x \log\theta + (n-x)\log(1-\theta).$$

Furthermore:

$$\begin{aligned} I(\theta) &= -E\Big(\frac{d^2 \log f(\boldsymbol{x}|\theta)}{d\theta^2}\Big) \\ &= -E\Big(\frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\Big) \\ &= \frac{n\theta}{\theta^2} - \frac{n-n\theta}{(1-\theta)^2} \quad \textit{because} \quad E(x) = n\theta \\ &= n\theta^{-1}(1-\theta)^{-1}. \end{aligned}$$

Therefore $f_0(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ *and thus*

$$\theta \sim Beta(1/2, 1/2).$$

1.6 Predictive distribution

With a prior distribution $f(\theta)$ the Bayes theorem leads to the posterior distribution $f(\theta|\boldsymbol{x})$. Suppose we consider taking a further observation y. We know the likelihood so $f(y|\theta)$ is known. The predictive distribution of y given x is

$$f(y|x) = \int f(y|\theta) f(\theta|\boldsymbol{x}) d\theta.$$

Example 1.10. *20 windows in a high rise office block broke in the first year of occupancy of the building. The question was how many of these were due to a specific defect D. If they were caused by D then the manufacturer of the windows will replace them, otherwise they will not. Only in 4 of the 20 widows was glass available for analysis; to this end we will assume a sample of 4 in 20. So all in our sample of 4 were found to have broken because of D.*

In the subsequent legal wrangle a glass expert claimed that the distribution of θ *the probability a window suffers from D is*

$$f(\theta) = \frac{\theta^{-3/4}(1-\theta)^{-1/4}}{B(1/4, 3/4)} \qquad 0 \le \theta \le 1 \quad or \quad \theta \sim Beta(1/4, 3/4).$$

For a sample of 4 with 4 with defect D the likelihood is $f(x|\theta) = \binom{4}{4}\theta^4(1-\theta)^{4-4} = \theta^4$ *so the posterior is*

$$f(\theta|x) = \frac{\theta^{13/4}(1-\theta)^{-1/4}}{B(1/4+4, 3/4+4-4)} \qquad 0 \le \theta \le 1 \quad or \quad \theta \sim Beta(17/4, 3/4).$$

This is a conjugate prior, see for the general case Table 3. The distribution of Y *the number of defectives is (given* θ*)*

$$f(y|\theta) = \binom{16}{y}\theta^y(1-\theta)^{16-y}$$

The predictive distribution is then

$$\begin{aligned} f(y|x) &= \int f(y|\theta) f(\theta|\boldsymbol{x}) d\theta \\ &= \frac{\binom{16}{y}}{B(17/4,3/4)} \int_0^1 \theta^{y+13/4} (1-\theta)^{63/4-y} d\theta \\ &= \binom{16}{y} \frac{B(y+13/4+1, 63/4-y+1)}{B(17/4,3/4)} \\ &= \binom{16}{y} \frac{B(y+17/4, 67/4-y)}{B(17/4,3/4)} \end{aligned}$$