smfd16.tex
Day 16. 10.12.2014.

In hypothesis testing, such prior knowledge by the experimenter (scientific, manufacturing etc.) is tacitly assumed, because we need it to be able to formulate H_0 and H_1 sensibly. But we may not be willing to enter the 'accept or reject' framework of hypothesis testing [which some statisticians believe is inappropriate and damaging]: how then can we use prior knowledge? In the estimation framework also, we may know a lot about θ before sampling [as in the rope example above]: indeed, if we do *not* have some prior knowledge of the situation to be studied, we would in practice not have enough prior interest in it to be willing to invest the time, trouble and money to study it statistically.

Bayesian statistics addresses this by giving a framework where
1. The statistician knows something before sampling: he has some *prior knowledge*.
2. He then draws a sample, and analyses the *data* to extract some relevant information.
3. He then *updates his prior information* with his *data (or sample) information*, to obtain *posterior information*

(prior: before (sampling); posterior: after (sampling)).

This verbal description of the Bayesian approach is attractive, because it resembles how we learn. Life involves (indeed, largely consists of) a constant, ongoing process of acquiring new information and using it to update our previous ('prior') information/beliefs/attitudes/policies.

To implement the Bayesian approach, we need some mathematics. The formulae below derive from the work of the English clergyman
Thomas BAYES (1702-1761): *An essay towards solving a problem in the doctrine of chances* (1763, posth.).
Recall that if A, B are events of positive probability,

$$P(A) > 0, \qquad P(B) > 0,$$

the *conditional probability* of A *given* (or *knowing*) B is

$$P(A|B) := P(A \cap B)/P(B).$$

Symmetrically,

$$P(B|A) := P(B \cap A)/P(A) = P(A \cap B)/P(A).$$

Combining,

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) :$$

$$P(B|A) = P(A|B)P(B)/P(A) \qquad \text{(Bayes' formula, or Bayes' theorem).}$$

Interpretation.

1. Think of A as a 'cause', B as an 'effect'. We naturally first think of P(effect B|cause A). We can use Bayes' formula to get from this to P(cause A|effect B) (think of B as an effect we can see, A as an effect we can't see).

2. Suppose we are interested in event B. We begin with an initial, *prior* probability $P(B)$ for its occurrence. This represents how probable we initially consider B to be [this depends on us: we will have to estimate $P(B)$!]. Suppose we then observe that event A occurs. This gives us new information, which affects how probable we should now consider B to be, *after* observing A [or, to use the technical term, *a posteriori*]. Bayes' theorem tells us how to do this updating: we multiply by the ratio $P(A|B)/P(A)$:

$$P(B|A) = P(B).P(A|B)/P(A) :$$

posterior probability of B = prior probability of B $\times$ updating ratio.

We first observe some extreme cases.

Independence. If A, B are independent, $P(A \cap B) = P(A).P(B)$, so

$$P(B|A) = P(A \cap B)/P(A) = P(A).P(B)/P(A) = P(B),$$

and similarly $P(A|B) = P(A)$: updating ratio = 1, posterior probability = prior probability – conditioning on something independent has no effect.

Inclusion.

1. $A \subset B$: here, $P(A \cap B) = P(A)$, $P(A|B) = P(A \cap B)/P(B) = P(A)/P(B)$;

updating ratio $P(A|B)/P(A) = 1/P(B)$, posterior probability = 1.

2. $B \subset A$: $P(A \cap B) = P(B), P(A|B) = P(A \cap B)/P(B) = P(B)/P(B) = 1$; updating ratio $P(A|B)/P(A) = 1/P(A)$, posterior probability = $P(B)/P(A)$.

Partitions. B partitions Ω into two disjoint events B; A is the disjoint union of $A \cap B$ and $A \cap B^c$, so

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Similarly, if $\Omega = \cup_1^n B_i$ with B_i disjoint, $A = \cup_1^n (A \cap B_i)$, disjoint. So by finite additivity,

$$P(A) = \Sigma_{r=1}^n P(A \cap B_r) = \Sigma_{r=1}^n P(A|B_r)P(B_r) \quad \text{(Formula of total probability),}$$

using the definition of conditional probability again.

Such expressions are often used for the denominator in Bayes' formula:

$$P(B_r|A) = P(B_r)P(A|B_r)/P(A) = P(B_r)P(A|B_r)/\Sigma_k P(B_k)P(A|B_k).$$

3. Prior and posterior densities.

Suppose now we are studying a parameter θ. Suppose we have data x [x may be a single number, i.e. a scalar, or a vector $x = (x_1, \cdots, x_n)$ of numbers; we shall simply write x in both cases]. Recall that x is an observed value of a random variable, X say. In the *density case*, this random variable has a (probability) *density* (function), $f(x)$ say, a non-negative function that integrates to 1:

$$f(x) \geq 0, \qquad \int f(x)dx = 1$$

(here and below, integrals with limits unspecified are over everything).
Interpretation. $P(X \in A) = \int_A f(x)dx$ for measurable sets $A \subset \mathbb{R}$. For instance, if $A = (-\infty, x]$,

$$F(x) := P(X \in (-\infty, x]) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy \quad \forall x \in \mathbb{R};$$

as x varies, $F(x)$ gives the (probability) distribution (function) of X.]
In brief: the density $f(x)$ describes the *uncertainty* in the data x.

The distinctive feature of Bayesian statistics is that it treats *parameters* θ in the same way as *data* x. Our initial (prior) uncertainty about θ should also be described by a density $f(\theta)$:

$$f(\theta) \geq 0, \qquad \int_{-\infty}^{\infty} f(\theta)d\theta = 1, \qquad P(\theta \in A) = \int_A f(\theta)d\theta \qquad \forall A \subset \mathbb{R},$$

where the probability on the left is a *prior probability*. The analogue for densities of Bayes' formula $P(B|A) = P(B)P(A|B)/P(A)$ now becomes

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x). \qquad (*)$$

The density on the left is the *posterior density* of θ *given* the data x; it describes our uncertainty about θ knowing x. Now densities integrate to 1: $\int f(\theta|x)d\theta = 1$, so $\int [f(\theta)f(x|\theta)/f(x)]d\theta = 1$:

$$\int f(\theta)f(x|\theta)d\theta = f(x).$$

Combining,

$$f(\theta|x) = f(\theta)f(x|\theta)/\int f(\theta)f(x|\theta)d\theta.$$

In the *discrete case*, θ and/or x may take discrete values $\theta_1, \theta_2, \cdots, x_1, x_2, \cdots$ only, with probabilities $f(\theta_1), f(\theta_2), \cdots, f(x_1), f(x_2), \cdots$. The above formulae still apply, but with *integrals replaced by sums*:

$$P(X \in A) = \Sigma_{x\in A} f(x), \qquad P(\theta \in B) = \Sigma_{\theta\in B} f(\theta),$$

$$f(x) = \Sigma_\theta f(\theta f(x|\theta), \quad f(\theta|x) = f(\theta)f(x|\theta)/\Sigma_\theta f(\theta)f(x|\theta).$$

In the formula $f(\theta|x) = f(\theta)f(x|\theta)/f(x)$, it is θ, the parameter under study, which is the main focus of interest. Consequently, the denominator $f(x)$ – whose role is simply to ensure that the posterior density $f(\theta|x)$ integrates to 1 (i.e., really is a density) – can be omitted (or understood from context). This replaces the *equation* above by a *proportionality statement*:

$$f(\theta|x) \propto f(\theta)f(x|\theta)$$

(here $\propto$, read as 'is proportional to', relates to the variability in θ, which is where the action is). Now $f(x|\theta)$ can be viewed in two ways:
(i) for fixed θ as a function of x. It is then the density of x when θ is the true parameter value,
(ii) for fixed/known/given data values x as a function of θ. It is then called the *likelihood* of θ (Fisher), familiar from IS, Ch. I, Ch. II, etc.

The formula above now reads, in words:

posterior $\propto$ prior $\times$ likelihood.

This is the essence of Bayesian statistics. It shows how Bayes' theorem may be used to *update* the *prior* information on θ before sampling by using the information in the *data* x – which is contained in the *likelihood* factor $f(x|\theta)$ by which one multiplies – to give the *posterior* information on θ after sampling. Thus posterior information combines two sources: prior information and data/sample/likelihood information.

4. Examples.

Example 1. Bernoulli trials with Beta prior ([O'H], Ex. 1.4, p.5).

Here θ represents the probability of a head on tossing a biased coin. On the basis of prior information, θ is assumed to have a prior density proportional to $\theta^{p-1}(1-\theta)^{q-1}$ $(0 \le \theta \le 1)$ for $p, q > 0$:

$$f(\theta) \propto \theta^{p-1}(1-\theta)^{q-1} \qquad (0 \le \theta \le 1).$$

Writing

$$B(p,q) := \int_0^1 \theta^{p-1}(1-\theta)^{q-1} d\theta$$

(the *Beta function)*,

$$f(\theta) = \theta^{p-1}(1-\theta)^{q-1}/B(p,q).$$

[We quote the *Eulerian integral* for the Beta function: for

$$\Gamma(p) := \int_0^\infty e^{-x}x^{p-1}dx \quad (p>0), \quad B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q) \quad (p,q>0).]$$

Note that, as p, q vary, the shape of $f(\theta)$ varies – e.g, the graph is u-shaped if $0 < p, q < 1$, n-shaped if $p, q > 1$. Here p, q are called *hyperparameters* - they are parameters describing the parameter θ.

Suppose now we toss the biased coin n times (independently), observing x heads. Then x is our data. It has a discrete distribution, the binomial $B(n,\theta)$, described by

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad (x = 0, 1, \cdots, n).$$

We apply Bayes' theorem to update our prior information on θ – our prior values of p, q – by our data x. Now

$$f(x) = \int f(\theta)f(x|\theta)d\theta = \int \frac{\theta^{p-1}(1-\theta)^{q-1}}{B(p,q)}.\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta$$

$$= \binom{n}{x}.\frac{1}{B(p,q)}.\int_0^1 \theta^{p+x-1}(1-\theta)^{q+n-x-1}d\theta = \binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}.$$

So Bayes' theorem gives

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x) = \binom{n}{x}.\frac{1}{B(p,q)}.\theta^{p+x-1}(1-\theta)^{q+n-x-1}/\binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}$$

or

$$f(\theta|x) = \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x, q+n-x)}.$$

The posterior density of θ is thus another Beta density, $B(p+x, q+n-x)$. Summarising:

• prior $B(p,q)$ is updated by data x heads in n tosses to posterior $B(p+x, q+n-x)$.

Graphs. To graph the three functions of θ – prior, likelihood and posterior – first find their maxima.

Likelihood: $f(x|\theta)$ has a maximum where $\log f(x|\theta)$ has a maximum, i.e. where

$x \log\theta + (n-x)\log(1-\theta)$ has a maximum, i.e. where

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 : \qquad x - x\theta = n\theta - x\theta : \qquad \theta = x/n.$$

Prior: similarly, $f(\theta)$ has a maximum where $\log f(\theta)$ does, i.e. where

$$\frac{p-1}{\theta} - \frac{q-1}{1-\theta} = 0 : \quad p - p\theta - 1 + \theta = q\theta - \theta : \quad \theta = (p-1)/(p+q-2).$$

Example 2. Normal family with normal prior ([O'H], Ex. 1.5 p.7). Suppose x is the sample mean of a sample of n independent readings from a normal distribution $N(\theta, \sigma^2)$, with σ known and θ the parameter of interest. So x is $N(\theta, \sigma^2/n)$:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}.\sigma/\sqrt{n}} \exp\{-\frac{1}{2}(x-\theta)^2/\frac{\sigma^2}{n}\}.$$

Suppose that on the basis of past experience [prior knowledge] the prior distribution of θ is taken to be $N(\mu, \tau^2)$:

$$f(\theta) = \frac{1}{\sqrt{2}\pi\tau} \exp\{-\frac{1}{2}(\theta-\mu)^2/\tau^2\}.$$

Now $f(x) = \int f(\theta) f(x|\theta) d\theta$:

$$f(\theta) f(x|\theta) = \frac{1}{2\pi.\tau\sigma/\sqrt{n}}. \exp\{-\frac{1}{2}\Big[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2/n}\Big]\}.$$

The RHS has the functional form of a bivariate normal distribution (IV.2 D7, [BF] 1.5). So to evaluate the θ-integration, we need to *complete the square*

(cf. solving quadratic equations!). First,

$$(x-\theta)^2 = [(x-\mu)-(\theta-\mu)]^2 = (x-\mu)^2 - 2(x-\mu)(\theta-\mu) + (\theta-\mu)^2.$$

We write for convenience

$$c := \frac{1}{\tau^2} + \frac{1}{\sigma^2/n}.$$

Then

$$\begin{aligned} f(\theta)f(x|\theta) &= const.\exp\{-\frac{1}{2}\Big[c(\theta-\mu)^2 - \frac{2}{\sigma^2/n}(\theta-\mu)(x-\mu) + \text{function of } x\Big]\} \\ &= const.\exp\{-\frac{1}{2}c\Big[(\theta-\mu)^2 - \frac{2(\theta-\mu)(x-\mu)}{c\sigma^2/n} + \text{function of } x\Big]\} \\ &= const.\exp\{-\frac{1}{2}c\Big(\theta-\mu-\frac{x-\mu}{c\sigma^2/n}\Big)^2 + \text{function of } x\}. \end{aligned}$$

Then from (*), to get the posterior density $f(\theta|x)$ we have to take the product $f(\theta)f(x|\theta)$ above, and divide by $f(x)$ – a function of x *only* (θ has been integrated out to get it). So: the posterior density $f(\theta|x)$ is itself of the form above, as a function of θ (with a different constant and a different function of x – but these do not matter, as our interest is in θ).

We can now recognise the posterior $f(\theta|x)$ – it is *normal*. We can read off:
(i) its mean, $\mu + (x-\mu)/(c\sigma^2/n)$,
(ii) its variance, $1/c$. Thus the *posterior precision* is c. But from the definition of c, this is the sum of $1/\tau^2$, the *prior precision*, and $1/(\sigma^2/n)$, the *data precision*. By (i), the mean is

$$\mu[1 - \frac{\text{data precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}],$$

or

$$\mu[\frac{\text{prior precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}].$$

This is a *weighted average* of the prior mean μ and the data value x (the sample mean of the n readings), *weighted according to their precisions*. So:
(a) the form, mean and variance (or precision) of the posterior density are intuitive, statistically meaningful and easy to interpret,
(b) the conclusions above show clearly how the Bayesian procedure synthesises prior and data information to give a compromise,

(c) the family of normal distributions is closed in the above example: a normal prior and normal data give a normal posterior. This is an example of *conjugate priors*, to which we return later.
Note. The example above on the normal distribution makes another important point: often θ will be a vector parameter, $\theta = (\theta_1, \cdots, \theta_p)$ – as with, e.g., the normal distribution $N(\mu, \sigma^2)$. For simplicity, the variance σ^2 in the above was taken known. But in general, we will not know σ^2. Instead, we should include it in the Bayesian analysis, representing our uncertainty about it in the prior density. We then arrive at a posterior density $f(\theta|x)$ for the vector parameter $\theta = (\theta_1, \cdots, \theta_p)$. If our interest is in, say, θ_1, we want the posterior density of θ_1, $f(\theta_1|x)$. We get this just as in classical statistics we get a marginal density out of a joint density – by *integrating out the unwanted variables*.

In the normal example above, Ex. 2, we could impose a prior density on σ without assuming it known. This can be done ([O'H], Ex. 1.6 p.8, Lee [L], §2.12), but there is no obvious natural choice, so we shall not do so here.

Example 3. The Dirichlet distribution ([O'H], Ex. 1.7 p.10, §10.2-6). Consider the density in $\theta = (\theta_1, \cdots, \theta_k)$ on the region

$$\theta_1, \cdots, \theta_k \geq 0, \qquad \theta_1 + \cdots + \theta_k = 1$$

(a *simplex* in k dimensions), with density

$$f(\theta) \propto \Pi_{i=1}^k {\theta_i}^{a_i - 1}$$

for constants a_i. We quote that the constant of proportionality is

$$\Gamma(a_1 + \cdots + a_k)/\Gamma(a_1) \cdots \Gamma(a_k),$$

by *Dirichlet's integral*, an extension of the Eulerian integral for the gamma function (see [O'H] 10.4, or, say, 12.5 of
WHITTAKER, E. T. & WATSON, G. N.: *Modern analysis*, 4th ed., 1927/1963, CUP).
Thus the *Dirichlet density* $D(a_1, \cdots, a_k)$ with *parameters* $\theta_1, \cdots, \theta_k$ is

$$f(\theta) := \frac{\Gamma(a_1 + \cdots + a_k)}{\Gamma(a_1) \cdots \Gamma(a_k)} . {\theta_1}^{a_1 - 1} \cdots {\theta_k}^{a_k - 1}.$$