2 Introduction: $\propto$

Who is Bayes? Reverend Thomas Bayes (1702-1761). Posthumous publication by Richard Price:

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 330-418. (Reprinted with biographical note by G.A. Barnard in *Biometrika* **45**, 293-315, 1958).

See also:
http://en.wikipedia.org/wiki/Thomas_Bayes
http://www.bayesian.org/.

Signature of Thomas Bayes
from a letter in the Centre for
Kentish Studies

Figure 1: T. Bayes.

In the preliminaries, the concept of (bayesian) probability was already briefly introduced as a degree of uncertainty. More formally, we could write that *every* probability is only a *conditional* probability, that depends on the background information I the observer has. Hence, it is always the case that the probabilities are of the form

$$P(A \mid I).$$

Although, for the convenience of shorter notations, we usually write $P(A)$, bearing in mind that it really is always conditional to some I. It therefore follows that two observers with different background information I_1 and I_2 have two different probabilities concerning the same event

$$P(A \mid I_1) \neq P(A \mid I_2).$$

For this reason, the bayesian definition of probability is said to be *subjective* as opposed to 'objective'. But subjective does not mean that "anything goes" or that the analysis is based on arbitrariness. The fully bayesian viewpoint is that there is no such thing as "pure objectivity". What we can do, is strive for logical coherence of our inferential process, when judging under uncertainty. When the probabilities of two persons disagree, it is because they had different background information. Remember: before you make a bet on a horse, be sure that your opponent does not know better about that horse, or else you're sure to lose! In a sense, bayesian analysis aims to be transparent because it encourages to write explicitly conditional probabilities. Many disagreements typically occur when two experts argue about $P(A)$ as an "objective property" of a phenomenon when, in fact, they should more explicitly argue about $P(A \mid I)$, for some relevant I. In bayesian context, there is no "true probability", but the probabilities obey rules of logic that ensure that the inference is internally coherent. This does not prevent bad conclusions if your background information happens to be seriously misguided. Always explicitly define (as accurately as possible) what your relevant background information is (and find out what

it is for somebody else who is looking at the same problem). Therefore, conditional probability is a really important concept that is repeatedly used in all bayesian work. Actually, a probability is meaningless without stating the conditional information. Even a marginal distribution is still conditional to something: $\pi(x \mid I) = \int \pi(x, y \mid I)\mathbf{d}y$. There is no such thing as a completely unconditional probability.

Another important feature, or consequence, is that the probabilities are updated when new information arrives. They are not constants. Instead, they change when we learn more about the question being assessed (as they should change for learning to take place).

Consider this simple example: in a bag you have N balls that can be white or red, but you don't know how many are red. Initially, you might have a vague idea that perhaps half are red. But after you blindly pick one ball at a time, and always get a red ball, you gradually become more convinced that a larger proportion of them were red. In bayesian context, a scientific inquiry is a process of learning in which we update our previous state of knowledge. Probability theory, particularly the famous Bayes theorem, provides the necessary recipe for the quantitative task. This does not mean that the calculations are always easy, even though the general recipe is straightforward. Hard problems are hard problems, but many problems that may seem cumbersome at first, can be surprisingly easy to analyze with bayesian approach, particularly if only a numerical result is required. However, Bayes does not provide a "click-the-button" analysis that could be blindly applied. But perhaps we should not go for "click-the-button" statistical analysis too easily anyway. With bayesian probabilistic modelling we are free to think as big and complicated problems we want, without resorting to the first available "standard software approach" that does not exactly address our questions and whose assumptions are not exactly even valid in the problem we are trying to solve. But that does not come completely free of charge. Posterior distributions seldom take the form of a standard distribution. Therefore, their calculation typically requires MCMC methods, or some other numerical techniques. And they can be computationally intensive.

2.1 Probability as measure of uncertainty

> *It is unanimously agreed that statistics depends somehow on probability.*
> *But, as to what probability is and how it is connected with statistics,*
> *there has seldom been such complete disagreement and*
> *breakdown of communication since the Tower of Babel. (L J Savage 1972)*

In Bayesian interpretation, probability is the measure of uncertainty about any logical statement, whether that is a statement about the outcome of a repeatable experiment or not. Therefore, 'randomness', as far as it is described by probability, refers to uncertainty. It does not mean that some variable is said to be 'truly random'. Instead, the variable is random to us, as long as we are uncertain about its value. Sometimes, we can reduce our uncertainty by observations so that finally all uncertainties vanish, but more often we will remain more or less uncertain. There are different types of uncertainties, sometimes described as *aleatory* and *epistemic*. Consider again the simple example of drawing red and white balls from a bag. Firstly, we are uncertain about the exact number of red and white balls before any ball was picked. This could be our epistemic uncertainty about the contents of the bag. Assume that we know the total number of balls M. We can then think of all possible proportions (r) of red balls:

$$r \in \{\frac{0}{M}, \frac{1}{M}, \frac{2}{M}, \frac{3}{M}, \ldots, \frac{M}{M}\}.$$

Our epistemic uncertainty could be quantified by assigning a probability for each of these values. If we have no reason to suspect any particular arrangement, this initial uncertainty could be described as a discrete uniform distribution:

$$P(r = \frac{i}{M}) = \frac{1}{M+1} \quad \forall i = 0, 1, \ldots, M.$$

When a ball is picked, we need to consider how this procedure works and does it somehow select more easily red balls than white ones. The outcome must depend on the actual contents of the bag or else the experiment would be meaningless. Also, the selection of a ball is 'randomized' as far as we can control the procedure. Hence, we can have aleatory uncertainty about the color of the resulting ball. This could be described, conditionally (given the unknown true proportion) as

$$P(X = \text{red} \mid r = \frac{i}{M}) = \frac{i}{M}.$$

Note that the selection of a ball was 'randomized' or 'blindfolded' only as far as we could know about it. It may not be 'truly random'. We could always think of someone more informed than us, who knows better the positions of the balls and the movements of the hand that picks the ball. There would not be aleatory uncertainty for him. Someone who knows exactly the initial conditions and how the ball is to be picked also knows the result without any uncertainty. This effect is exploited in magic tricks. But it shows that also aleatory uncertainty is actually a form of our uncertainty, arising from incomplete knowledge. The outcome of every 'random experiment' is predictable *if* we only knew the *exact* initial conditions. E.T. Jaynes has discussed the "physics of random experiments" in his book "Probability theory, the logic of science" [3], discussing also quantum mechanics. For the purpose of quantifying our uncertainty, it remains open whether there really is 'true randomness' out there, or whether everything is thoroughly deterministic (or even something else?). We do not need to assume either way, because we describe and update our uncertainties based on what we *can* know.

2.2 From prior probability to posterior

So how exactly the probabilities are updated? First, we must declare what our prior probability is. To continue the example above, this was done already there: $P(r) = 1/(M+1)$. Then, we must declare the conditional probability of the observable outcome, given the true proportion (r) of red balls. This too was stated already: $P(X = \text{red} \mid r) = r$. We are here dealing with two quantities r and X, both of which are uncertain before observations. (Total number of balls M was assumed known). According to probability theory, due to symmetry of $P(X, r)$:

$$P(X, r) = P(X \mid r)P(r) = P(r \mid X)P(X) = P(r, X).$$

Our prior probability about r is expressed as $P(r)$, and our posterior probability as $P(r \mid X)$, after observing the outcome X. We can now solve the posterior probability:

$$P(r \mid X) = \frac{P(X \mid r)P(r)}{P(X)}.$$

This is known as the Bayes's formula. The idea was first used by Thomas Bayes, 1763, in the form of a specific example problem concerning billiard balls. However, it gives the general recipe for updating prior probabilities into posterior probabilities. But the actual calculation can be laborious. It should be noted that this is a probability (or probability density) for the unknown quantity (here r). It is a conditional probability, given the observed quantity (here X) which is no longer random after it has been observed. The denominator $P(X)$ is constant with respect to r, and has the role of a normalizing constant. Ignoring the normalizing constant, the Bayes's formula is often written in a proportional form:

$$P(r \mid X) \propto P(X \mid r)P(r),$$

which means that $P(r \mid X)$ is proportional to $P(X \mid r)P(r)$. The normalizing constant can be written as:

$$P(X) = \sum_i P(X \mid r_i)P(r_i) \qquad \text{or} \qquad \int_R P(X \mid r)P(r)\mathbf{d}r,$$

depending on whether r is discrete or continuous. Therefore, the solution is completely determined when $P(r)$ and $P(X \mid r)$ are determined mathematically.

For this particular example problem, we can try to calculate the posterior:

$$P(r = i/M \mid X = \text{red}) \propto \underbrace{\frac{i}{M}}_{P(X=\text{red}|r=i/M)} \times \underbrace{\frac{1}{M+1}}_{P(r=i/M)}.$$

The normalizing constant is thus

$$C = \sum_{i=0}^{M} \frac{i}{M}\frac{1}{M+1} = \frac{1+2+\ldots+M}{M(M+1)} = \frac{M(1+M)/2}{M(M+1)} = 1/2.$$

Therefore, the posterior probability is:

$$P(r = i/M \mid X = \text{red}) = \frac{2i}{M(M+1)}.$$

What does it tell us? Firstly, the probability that there were no red balls ($i = 0$) in the bag is zero, obviously because we just observed one. Secondly, it is most probable (probability $2/(M+1)$) that all balls are red ($i = M$) because, so far, the ball that we observed was indeed red, not white, and our prior probability was even for all possible proportions. Thirdly, the probability for all other proportions ($0 < i < M$) is between these extremes, taking values $2/(M(M+1)), 4/(M(M+1)), 6/(M(M+1)), \ldots$.

The above calculation may be simple but it demonstrates how prior probability actually is updated to a posterior probability. We might continue the experiment by drawing more balls and update the posterior again and again. But we then need to specify how the additional draws are actually done. If we take out each ball we are exhausting the bag and eventually we will be completely sure about its contents. This type of experiment leads to hypergeometric distribution for the total number of red balls (k) in a given number (K) of draws ($K < M$). But assume that we replace the ball in the bag after every draw. Then, the conditional probability for obtaining a red ball remains the same for each

draw (assuming a thorough lottery mixing of balls), but our prior probability will change according to the observation history. If the first ball was red, our current state of knowledge is summarized by the posterior we just calculated. It is no longer the uniform discrete distribution we started with. The obtained posterior becomes our new prior in the face of the next experiment. (Unless we deliberately want to forget what information we just learned). Assume then that the second draw also results to a red ball. What is the posterior for proportion r now? The current prior is:

$$P(r = i/M) = \frac{2i}{M(M+1)},$$

So, the new posterior will be

$$P(r = i/M \mid 2^{\text{nd}}X = \text{red}) \propto \frac{i}{M}\frac{2i}{M(M+1)} = \frac{2i^2}{M^2(M+1)},$$

and its normalizing constant is

$$C = \frac{2}{M^2(M+1)}\sum_{i=0}^{M} i^2 = \frac{2}{M^2(M+1)}\frac{M(M+1)(2M+1)}{6} = \frac{2M+1}{3M}.$$

Hence, the posterior probability:

$$P(r = i/M \mid 2^{\text{nd}}X) = \frac{2i^2}{M^2(M+1)} \times \frac{3M}{2M+1} = \frac{6i^2}{M(M+1)(2M+1)}.$$

This is the result after two red balls (assuming replacement) and we see that the posterior probability is now higher for the event that all balls are red. The same result would have been obtained if we had used the original prior but calculated the probability for two successive red balls (assuming replacement). It does not matter if we really update the prior step-by-step after each observation or if we update it once by using all the data simultaneously. This is formally expressed as:

$$P(r \mid X_1, X_2) = \frac{P(X_1, X_2 \mid r)P(r)}{P(X_1, X_2)} = \frac{P(X_2 \mid X_1, r)P(X_1 \mid r)P(r)}{P(X_1, X_2)} = \frac{P(X_2 \mid r)P(X_1 \mid r)P(r)}{P(X_1, X_2)}$$

$$= \frac{P(X_2 \mid r)P(r \mid X_1)P(X_1)}{P(X_2 \mid X_1)P(X_1)} = \frac{P(X_2 \mid r)P(r \mid X_1)}{P(X_2 \mid X_1)} \propto P(X_2 \mid r)P(r \mid X_1),$$

where the posterior after the 1st observation was:

$$P(r \mid X_1) = \frac{P(X_1 \mid r)P(r)}{P(X_1)}.$$

What probability laws were used in this? Why were they valid?

In short:

$$P(r \mid X_1, X_2) \propto P(X_1, X_2 \mid r)P(r) = P(X_1 \mid r)P(X_2 \mid r)P(r) \propto P(r \mid X_1)P(X_2 \mid r)$$

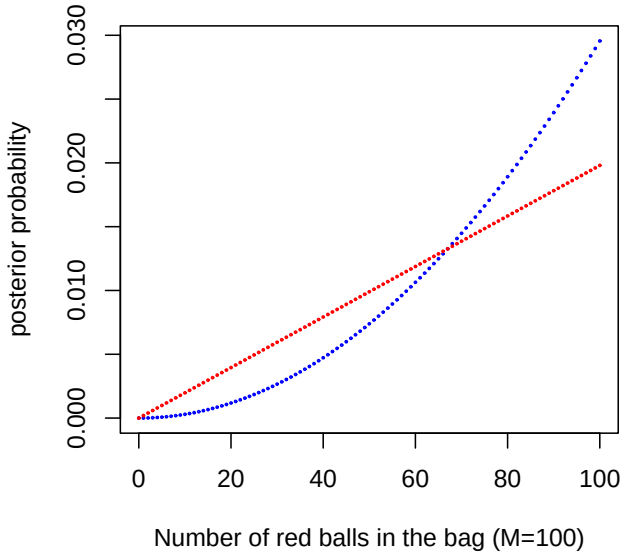

Figure 2: Posterior probabilities for the number of red balls among M in a bag, if one ball is drawn and it is red (red dots), and if two balls are drawn and both are red (blue dots).

2.3 Where do priors come from?

In the original work of Bayes, he considered billiard balls and the position of a 'randomly' thrown ball on a billiard table. The position was assumed known to the experimenter but unknown to the observer. The observer is told about the positions of subsequent balls with respect to the first ball; whether they end up left or right from the first ball. The position of the first ball was to be estimated by the observer. The prior was chosen as uniform distribution across the table, based on physical intuition that the ball could stop at any position 'equally likely'. In the example of red and white balls, we chose a uniform discrete distribution to express our initial uncertainty that any proportion (i/M) of red balls is as likely as any other. Both of these choices are examples of the principle of insufficient reason (or indifference). This gives the simplest *non-informative* prior. It is commonly applied when there is no knowledge indicating unequal probabilities.

An alternative approach would be to choose an *informative* prior. That would be based on careful examination of expert knowledge and *elicitation* of a prior distribution from the expert or group of experts.

Broadly, these two approaches are sometimes called as *objective* bayesian [1] and *subjective* bayesian [2] approach. If the data are very informative about the quantity being estimated, then an uninformative prior is a quick and easy choice. Actually, if the data are extremely informative, then nearly any prior would lead to the same posterior probability. But if the data are poor, then the posterior will be heavily influenced by the prior and it is more important to think how the prior was chosen and how sensitive the result is to different priors. Also, there can be really important expert knowledge (that is not part of the observed data already). That can be a basis for an informative prior, by conducting

a careful elicitation process.

2.3.1 Simple elicitation of prior probability

We would like to obtain your prior probability of $A =$"salmonella is detected from this pig". You are given a choice between these two options:

(1) You'll get 300 EUR if salmonella is detected from this pig.

(2) You'll receive a lottery ticket such that n tickets from a hundred will win 300 EUR.

Which option would you choose? Assume that n is really small number. If you believe (based on your background knowledge about salmonella in pigs) that you then have better chances to win with the first choice, it means that for you

$$\frac{n_{\text{small}}}{100} < P(A \mid I_{\text{your}}).$$

Likewise, assume that n is really large number. Then you would probably go for the lottery ticket, which means that

$$P(A \mid I_{\text{your}}) < \frac{n_{\text{large}}}{100}.$$

By making n_{small} larger and n_{large} smaller, we would eventually find such value, n^*, that you could not make the choice. Both options would then be equally attractive for that n^*. This means that, for you:

$$P(A \mid I_{\text{your}}) = \frac{n^*}{100}.$$

Another way to approach subjective probability is by using *odds*. When making bets (at some stake M) about some event A, the possible rewards are as follows: if event A happens, you will gain ωM, but if it does not happen, you'll lose M. If you strongly believe that A happens, then you would accept the bet for a small ω, but if you strongly believe A does not happen, then ω would have to be large before you would accept the bet. A fair bet is such that

$$P(A)\omega M + (1 - P(A))(-M) = 0$$

from which the probability $P(A)$ can be obtained as

$$P(A) = \frac{1}{1+\omega}.$$

For example, if you consider the odds $\omega = 1/400$ as fair, then $P(A) = 400/401$.

Note: definition of odds above may be used in gambling, but in probability and statistics, odds *for* event A is defined as $P(A)/(1 - P(A))$.

In practice, we often need to consider distributions for continuous quantities or even more complicated multivariate objects. Elicitation of expert's knowledge can then be very laborious and prone to

psychological effects leading to inconsistencies in the expert's stated opinions. Therefore, 'objectivist' techniques for universal noninformative priors can often be sufficient (and free of elicitation problems). However, the quest for a truly universal method for a noninformative prior may be the quest for Holy Grail. There are different approaches, each with some drawbacks. For example, the simplest idea of a uniform distribution for a variable X, does not give a uniform distribution for some transformation of X, for example X^2, or $\log(X)$.

There are no unknown probabilities in a Bayesian analysis,
only unknown - and therefore random - quantities for which you have a probability
based on your background information (O'Hagan 1995).

Question from the audience:
"But of course, a mere machine can't really `think`*, can it?"*
John von Neumann replied:
"You insist that there is something a machine cannot do.
If you will tell me precisely what it is that a machine cannot do,
then I can always make a machine which will do just that!" (Lecture in Princeton, 1948).

Examining all the particulars is difficult as they are infinite in number.
(Wikipedia: Sextus Empiricus, Outlines Of Pyrrhonism.
Trans. R.G. Bury, Harvard University Press, Cambridge, Massachusetts, 1933, p. 283).

Other definitions of probability:

Frequentist definition: probability of event A is the limiting frequency of occurrences of A in a series of repeated experiments. But this limited frequency is always unknown to us, because we cannot repeat any experiment truly infinitely. (Compare with bayes: all probabilities are known!).

Classical definition: this is familiar from most school books. Based on symmetry of 'elementary events'. For example, in coin tossing 'Heads' and 'Tails' are equally possible because of the symmetry of the coin. Likewise, probability of Ace of Spades is 1/52 due to symmetry of the cards. But symmetry arguments can be difficult to find for more complicated events which cannot be easily broken down into elementary events. Furthermore, even if the coin is perfectly symmetric, the result depends on how the coin is tossed. But symmetry argument is very closely related to the concept of exchangeability in bayesian inference.

2.4 Binomial model

In the example of red and white balls, we described bayesian inference when only two balls were drawn and both happened to be red. In general, if N balls are drawn (with replacement) from a bag with M balls, we can observe a sequence of red and white balls. If we define

$$X_i = \begin{cases} 1 & \text{if the } i\text{th ball is red} \\ 0 & \text{if the } i\text{th ball is white} \end{cases}$$

then, the (conditional) probability for a specific sequence can be written as

$$P(X_1, \ldots, X_N \mid r) = r^{\sum X_i}(1-r)^{(N-\sum X_i)}$$

where r is the proportion of red balls in the bag. If we only observe the sum $Y = \sum X_i$, but not the exact sequence, then

$$P(Y \mid r) = \binom{N}{Y} r^Y (1-r)^{N-Y}$$

which is the binomial distribution with parameters r and N. Individual draws are said to be Bernoulli experiments, corresponding to binomial distribution with parameters r and $N = 1$. So far, the proportion r has been considered as discrete valued. But if the number of balls in the bag is very large, we can think of the limiting value

$$\lim_{M \to \infty} \frac{R(M)}{M} = r$$

where $R(M)$ is the number of red balls among M balls. The object of inference is now a continuous valued parameter $r \in [0, 1]$ and we must specify a prior *density* for this. Analogous choice to the previous discrete uniform distribution would be uniform probability density:

$$\pi(r) = 1 \; \forall r \in [0, 1] \text{ and } 0 \; \forall r \notin [0, 1]$$

This uniform prior is a special case of a beta-density, obtained by setting $\alpha = \beta = 1$ (Bayes-Laplace uniform prior):

$$\pi(r) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha - 1}(1-r)^{\beta - 1}.$$

The posterior distribution of r is then obtained again by applying Bayes's formula, but now with probability densities:

$$\pi(r \mid Y) \propto r^{(Y + \alpha - 1)}(1-r)^{(N - Y + \beta - 1)}$$

The result is the same if we have observed the exact sequence of X_i's or if we just observe the sum Y. This shows that for inferring r, it is sufficient to know the sum of red balls. The posterior density of r is recognized to be a beta-density, with parameters $Y + \alpha$ and $N - Y + \beta$. The expected value of r from the posterior density is

$$E(r \mid X, N, \alpha, \beta) = \frac{\alpha + X}{\alpha + \beta + N},$$

which can also be written as a weighted average:

$$w \frac{\alpha}{\alpha + \beta} + (1 - w) \frac{X}{N},$$

where $w = (\alpha + \beta)/(\alpha + \beta + N)$. The parameters of the prior can thus be chosen so that they represent some imaginary data X_0, N_0, corresponding to $(\alpha, \beta) = (X_0, N_0 - X_0)$.

In this example, the posterior density could actually be solved so that the solution is among standard probability densities. This was possible because the binomial distribution of the data, and the beta-density prior are conjugate distributions. Generally, they don't have to be so, and we could choose any other prior distribution, but the resulting posterior would not be among any standard distributions. Yet, it could still be computed numerically.

So, now we have learned to obtain a posterior density for the unknown proportion r. It can be summarized in various ways, but it can also be made to work for us as a tool for many kind of scientific questions.

2.4.1 Uninformative priors for unknown proportion

If an uninformative prior is required for binomial proportion r, there are actually several choices. They are all uninformative, but in different ways.

Bayes-Laplace prior: Beta(1,1)

Jeffreys' prior: Beta(1/2,1/2)

Haldane's (improper) prior: Beta(0,0)

The Bayes-Laplace prior reflects the idea of 'insufficient reason', which says that unless there is specific reason to assign unequal probabilities, they should be equal for all possible values of r. But the problem is that the uniform prior is not uniform for all transformations. If, instead of r, we were interested in r^2, the prior $r \sim \mathrm{U}(0,1)$ would not imply a uniform prior for r^2, and vice versa. The uniform prior Beta(1, 1)=U(0,1) corresponds to having 2 prior experiments, one of which was a 'red ball' and the other 'white ball'. The Jeffreys' prior equals to having only one prior experiment in which one ball was 'drawn' and it was 'half red', 'half white'. In this sense, Haldane's prior corresponds to having no prior data at all, but the prior is actually concentrated at two points: zero and one. Moreover, with Beta$(0, 0)$ prior the posterior is not defined if the observed data happens to be either 0 or N under a Binomial(N, r) model. The Jeffreys' prior is based on the principle that an uninformative prior should be such that the posterior remains the same regardless of the parameter transformation used. For single parameters, the Jeffreys' prior is sometimes used but for multiparameter problems the results are more controversial, and a hierarchical modeling approach is more common. Generally, for some single parameter, r, the Jeffreys' prior is chosen so that

$$\pi(r) \propto [J(r)]^{1/2},$$

where $J(r)$ is so called *Fisher information* for r.

$$J(r) = E\Big[\Big(\frac{\mathbf{d} \log \pi(X \mid r)}{\mathbf{d}r}\Big)^2 \mid r\Big] = -E\Big[\frac{\mathbf{d}^2 \log \pi(X \mid r)}{\mathbf{d}r^2} \mid r\Big].$$

It can be shown that for a transformation $\psi = h(r)$, with $r = h^{-1}(\psi)$, the following equation can be obtained:

$$J(\psi)^{1/2} = J(r)^{1/2} \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

and the Jeffreys' prior is defined as proportional to $J(\cdot)^{1/2}$ which makes it invariant under transformation. Let's see by example what this means.

For a binomial model we have:

$$\log \pi(X \mid r) = \text{constant} + X \log(r) + (N - X) \log(1 - r)$$

$$\frac{\mathbf{d} \log \pi(X \mid r)}{\mathbf{d}r} = \frac{X}{r} - \frac{N - X}{1 - r}$$

$$\frac{\mathbf{d}^2 \log \pi(X \mid r)}{\mathbf{d}r^2} = \frac{-X}{r^2} - \frac{N - X}{(1 - r)^2},$$

and taking the negative of expected value, $-E(\cdot \mid r)$, gives

$$J(r) = -\Big(\frac{-rN}{r^2} - \frac{N - rN}{(1 - r)^2}\Big) = \frac{N}{r(1 - r)}.$$

The Jeffreys' prior for binomial proportion r is thus

$$\pi(r) \propto [J(r)]^{1/2} \propto r^{-1/2}(1 - r)^{-1/2}$$

which is Beta(1/2,1/2).

What does all this mean for some transformation of r? For example $\psi(r) = \sqrt{r}$, with inverse transform $r(\psi) = \psi^2$, and $\mid \mathbf{d}r/\mathbf{d}\psi \mid = 2\psi$. If we want the posterior density of ψ, we can obtain it in two ways:

(1). Compute the posterior density $\pi(r \mid X) \propto \pi(X \mid r)\pi(r)$ using Jeffreys' prior for r, and then use transformation of variables to get the posterior density of ψ:

$$\pi(\psi \mid X) = \pi(\, r(\psi) \mid X \,) \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid \;\propto\; \pi(X \mid r(\psi))\, \pi(r(\psi)) \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

$$\propto\; \psi^{2X}(1 - \psi^2)^{(N-X)} \times (\psi^2)^{-1/2}(1 - \psi^2)^{-1/2} \times 2\psi.$$

(2). Compute directly the posterior $\pi(\psi \mid X) \propto \pi(X \mid \psi)\pi(\psi)$ using Jeffreys' prior for ψ. In this case, $\log \pi(X \mid \psi) = c + 2X \log(\psi) + (N - X) \log(1 - \psi^2)$, and after some calculations we get $J(\psi) = 4N/(1 - \psi^2)$. Therefore, Jeffreys' prior for ψ is

$$\pi(\psi) \propto [J(\psi)]^{1/2} = \frac{2\sqrt{N}}{\sqrt{1 - \psi^2}} \propto (1 - \psi^2)^{-1/2}.$$

Using this prior, we calculate the posterior of ψ directly:

$$\pi(\psi \mid X) \propto \pi(X \mid \psi)\pi(\psi)$$

$$= \psi^{2X}(1 - \psi^2)^{(N-X)} \times (1 - \psi^2)^{-1/2}.$$

By comparing (1) and (2), either way, the posterior of ψ is the same!

Note also that if the prior of r is Beta(α, β), then the posterior will be Beta$(X + \alpha, N - X + \beta)$ and the posterior mode is then $(X + \alpha - 1)/(\alpha + \beta + N - 2)$, and posterior mean is $(X + \alpha)/(\alpha + \beta + N)$.

The posterior mode becomes X/N when the Bayes-Laplace prior is used. The posterior mean becomes X/N when the Haldane's prior is used. The fraction X/N is the *maximum likelihood estimator* for r in *likelihood inference*. I.e., it is the value of $r \in [0,1]$ that gives the highest probability for the data, X, that was observed: $\text{argmax}_{r\in[0,1]} P(X \mid N, r)$.

Warning: improper priors may lead to improper posteriors. Therefore, it may be advisable to use proper priors also when aiming at an uninformative prior. Later, when using WinBUGS, it is possible to explore what happens when the prior parameters are tuned towards a nearly improper distribution. Numerical difficulties may sometimes happen even if the prior is just proper, e.g. if the parameters of beta-density are nearly zero. Sensitivity analysis is always recommended to check how sensitive the posterior results are to the choice of prior.

2.4.2 Unknown N

The usual application of binomial model $\text{Bin}(N, r)$ involves inference about unknown r with known N. In general, any quantity could be unknown, so let's see how to make inference about N, assuming that r is known. We then would know the true proportion of red balls in a 'large' bag, and someone has done the sampling of N balls but he does not tell us what the sample size N was. Instead, we are only told how many red balls (X) there were. Again, we first have to specify a prior for N. But N could be any integer value $0, 1, 2, \ldots$ and there is no way to know how large it could be. It seems difficult to assign an uninformative probability distribution. But let's start with a simple choice that assumes some very large maximum value M, so that the prior is uniform from 0 to M:

$$P(N = i) = \frac{1}{M+1} \forall i \in \{0, 1, \ldots, M\}$$

Now the posterior is:

$$P(N \mid X, r) \propto P(X \mid N, r)P(N) = \frac{N!}{X!(N-X)!} r^X (1-r)^{N-X} \frac{1}{M+1}$$

$$\propto \frac{N!}{(N-X)!}(1-r)^N$$

$$= N(N-1)\ldots(N-X+1)(1-r)^N$$

and the normalizing constant is

$$\sum_{i=X}^{M} i(i-1)\ldots(i-X+1)(1-r)^i$$

This posterior distribution is not among the well known standard distributions. But it is a distribution. We just cannot find this distribution in a common statistical software. If our tools only allow to operate with a limited number of well known distributions, then we could not handle this. Therefore, it is good to have a software that allows some self-made programming in this kind of situations, e.g. in R: try the following, but be careful to use correct values: $X \leq N \leq M$.

```
p0 <- function(X,N,r){
s <- log(N)
for(i in 1:X-1){
```

```
s <- s+log(N-i)
}
s<-s+N*log(1-r)
exp(s)
}
postn <- function(X,N,M,r){
p0(X,N,r)/sum(p0(X,X:M,r))
}
```

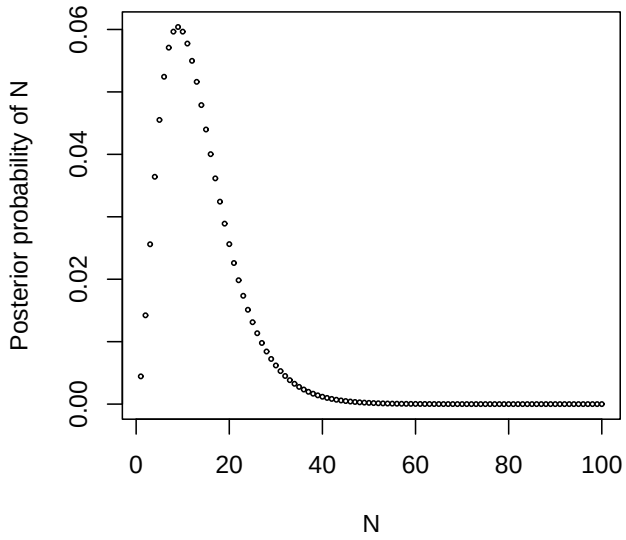

Figure 3: Posterior probability for N, given that $X = 1, r = 0.2$ with uniform prior over $0, 1, 2, \ldots, M = 100$.

The estimation of unknown proportion r is a common application in many applied areas, e.g. epidemiology. Applications with unknown N are rare because usually we know the sample size. In some situations this information may be missing. For example, if only positive results are reported in some reporting system, omitting negative results. We would not then know what the sample size was. It would also be difficult to estimate r, because all standard approaches assume N is known. In bayesian inference, unknown N just adds one more source of uncertainty to the problem (which then becomes described by a two-dimensional distribution).

2.5 Comment: $\propto$

The Bayes formula gives the general recipe for solving posterior distributions, and we usually need to focus only on the part that is a function of the unknown variable, the rest becomes the normalizing constant. We can derive all the estimates and even draw random samples from the density without knowing what the constant is (as long as we know the density itself really is a probability density).

The constant can be ignored in many calculations. Therefore, one of the most often used mathematical symbol in bayesian calculations is 'proportional to': $\propto$.

$$\boxed{\pi(\theta \mid X) \propto \pi(X \mid \theta)\pi(\theta)}$$

2.6 Elements of full bayesian analysis

In this course, we mainly focus on bayesian inference, which is only one part of a full baysian analysis aiming at decisions. All the elements would be:

1. An unknown quantity of interest, θ, which can be something more complicated than a scalar, e.g. it could be vector or matrix, etc.

2. The prior distribution of θ: $\pi(\theta)$.

3. The conditional distribution, or data generating model, of observations $\pi(X \mid \theta)$, sometimes also called as the likelihood function of θ.

4. The posterior of θ: $\pi(\theta \mid X)$, obtained from the Bayes' formula.

5. A set of actions $\{a_1, a_2, \ldots, a_n\}$, from which the decision maker has to choose. E.g. using a specific vaccine (a_1) or not using it (a_2).

6. A loss function $L(\theta, a_i)$ that depends both on the action chosen, a_i, and the unknown quantity θ. Usually, actions are taken after observing some data X, so that the action can be some function of data: $a_i(X)$.

7. Choosing the decision $a_i(X)$ that minimizes the expected loss:

$$E_{a_i(X)}(L \mid X) = \int_\Theta L(\theta, a_i(X))\ \pi(\theta \mid X)\ \mathbf{d}\theta.$$

Bayesian inference consists of the first four steps.

2.6.1 Example: HIV testing using confirmatory 2nd test

Assume that there are two possible strategies for testing patients (Example from D Draper):

$R1$: use ELISA test, at a cost of $c_1 = 20$; if positive, diagnose HIV+, but if negative, diagnose HIV-.

$R2$: same as $R1$, except that if ELISA gives positive result, use Western Blot to get 2nd result (cost $c_2 = 100$). If the 2nd test is positive, diagnose HIV+, if negative, diagnose HIV-.

With $R1$, the probabilities of different outcomes are

Probability	True HIV status	ELISA status	Cost
0.0095	+	+	c_1
0.0005	+	−	$c_1 + L1$
0.0198	−	+	$c_1 + L2$
0.9702	−	−	c_1

Here, $L1$ is the false negative cost of diagnosing HIV- when the patient really is HIV+, and $L2$ is the false positive cost.

The expected cost under $R1$ is then

$$E_{R1}(\text{cost}) = c_1 + 0.0005L1 + 0.0198L2$$

The corresponding table with $R2$ is

Probability	True HIV status	ELISA status	W B status	Cost
0.00945	+	+	+	$c_1 + c_2$
0.00005	+	+	−	$c_1 + c_2 + L1$
0.00004	+	−	+	$c_1 + L1$
0.00046	+	−	−	$c_1 + L1$
0.0001	−	+	+	$c_1 + c_2 + L2$
0.0197	−	+	−	$c_1 + c_2$
0.00095	−	−	+	c_1
0.96925	−	−	−	c_1

The expected cost under $R2$ is then

$$E_{R2}(\text{cost}) = c_1 + 0.0293c_2 + 0.00055L1 + 0.0001L2$$

Decision $R2$ should be preferred to decision $R1$ if $E_{R2}(\text{cost}) < E_{R1}(\text{cost})$, that is, when

$$0.0197L2 - 0.00005L1 - 0.0293c_2 > 0$$

Assume a modest value $L2 = 1000$. Then the advantage of $R2$ is quite small even for a huge value of $L1 = 100000$. But in this simple example, the probabilities of each outcome were assumed to be pre-assigned. So we did not need to do bayesian inference to calculate them.

References

[1] Berger J: The Case for Objective Bayesian Analysis. Bayesian Analysis, 2006, Vol 1, 3, 385-402.

[2] Goldstein M: Subjective Bayesian Analysis: Principles and Practice. Bayesian Analysis, 2006, Vol 1, 3, 403-420.

[3] Jaynes E T: Probability theory: the logic of science. Cambridge university press. 2003.