WinBUGS/OpenBUGS with applications

jukka.ranta@evira.fi

January 2011

Abstract

The course gives an introduction to WinBUGS/OpenBUGS as a tool for bayesian computations in applied problems using examples from e.g. food safety risk assessments, microbiology, diagnostics, veterinary science, epidemiology and examples from the course book. Pre-requirements are basic knowledge of the principles of bayesian statistics and familiarity with basic probability calculus including elementary probability (and multivariate) distributions and their parameters. The course (5 cu) consists of 25 hours with lectures and practicals, an exam and a written homework to be completed and returned in the end. In the homework, a short applied problem is described, a bayesian model defined, a WinBUGS code implemented and results discussed. Suitable example problems are given in the course material, but you can also suggest your own. The total grade is a combination of the exam (30 p) and the written homework (10 p). Preliminary course outline: intro to bayesian computation and MCMC, WinBUGS software, model definition language and DAGs, analysis of the output, application examples, scripts, extensions.

1 Textbooks

The number of Bayesian textbooks has grown enormously and you can find books for many specialized topics. Many of them include material about WinBUGS or OpenBUGS. For this course, a useful starting point can be Ioannis Ntzoufras: Bayesian Modeling using WinBUGS. Many examples are from this book.

`http://stat-athens.aueb.gr/~jbn/winbugs_book/`

Also a recent book is Marc Kéry: Introduction to WinBUGS for Ecologists.

`http://137.227.242.23/pubanalysis/kerybook/`

Moreover, there are three books by Peter Congdon about Bayesian modeling with WinBUGS, containing many examples. For Bayesian theory and computation see Gelman et al: Bayesian Data Analysis. This also contains instructions for using WinBUGS with or without R. An introduction to Bayesian theory 'for scientists and engineers' can be found e.g. in Sivia: Data Analysis, a Bayesian tutorial. Also, the book ET Jaynes: Probability Theory, the Logic of Science provides some interesting background and history.

The material for this course is a blend of these sources, WinBUGS manuals and some example problems from separate published papers and my previous projects and other miscellaneous materials. The introductory part is meant to be a brief reiteration of bayesian theory so that we can soon proceed to the BUGS and applications. Therefore, not all of the introductory part will be discussed in detail over the lectures, but you may use the text as a reference material if needed, depending on how much you already know about bayesian theory.

2 Preliminaries

Bayesian inference is so thoroughly based on calculating probabilities, that we might be tempted to say that it is nothing else than probability calculus put in practice. Every bayesian analysis involves definitions and calculations of probabilities or probability densities. They are the essential molding mass of probabilistic modeling. Hence, we could start with a quick review of some common concepts and notations (which are loosely used here). Note: these are mainly simply mathematical tools of probability calculus without (necessarily) any interpretation of probability beyond its mathematical definition! The same material can be found more extensively written in text books on elementary probability theory.

Random event, random variable. By an event A we can denote some actual event that may occur in a series of repeatable experiments, e.g. A ="a tail occurs in a coin toss", or A ="a measurement X is larger than 46". In the latter case, the event would be a statement regarding a specific (random) variable X - the value of the measurement. The value of X determines whether event A happened (is true) or not (is false). Generally, the 'event' is a logical proposition that is either true or false, and can also describe an unrepeatable state of affairs, e.g. A ="Jack is taller than John". (Using random variable notation: if X ="height of Jack", Y ="height of John", then A ="X>Y").

Probability for event A (in bayesian context) denotes the degree of uncertainty we have about the truth value of A. Hence, if you know Jack (short) and John (long) well, you might have $P(X > Y) = 0$, but for an uninformed outsider, it could well be that $P(X > Y) = 0.5$ (before receiving any more information than the names). Mathematically, probability is a *measure* that takes values between zero and one. With two events, A and B, the probability that *both* occur (is true) is written $P(A, B)$, or with specific variables: $P(X = x, Y = y)$, or $P(X \in S_1, Y \in S_2)$, where upper case letter denotes the random variable, and lower case letter denotes a specific value of it.

Probability distribution. For a discrete variable X, taking values in the set $\{x_1, x_2, \ldots\}$, this is the numerative collection of point probabilities $P(X = x_i) = P_i \geq 0$ so that $\sum P_i = 1$. Likewise, for a continuous variable X, taking values x in some set $S \subset \mathbb{R}^n$, this is the *probability density function* $\pi(x) \geq 0$ so that $\int_S \pi(x)\mathbf{d}x = 1$. Note that probability *density* is not the same as probability, because $P(x) = 0$ for all x, but the density is $\pi(x) \geq 0$. If a function does not integrate to one but to some other constant C, $(-\infty < C < \infty)$, it can always be *normalized* to make a **proper** probability distribution. If it integrates to infinity, it is said to be an **improper** probability distribution. Surprisingly, these can sometimes be used too! The *support* of a density is the set of x values for which $\pi(x) > 0$.

Cumulative probability distribution function: $F(x) = P(X \leq x)$, where X can be either discrete or continuous. Note: $F(-\infty) = 0$, and $F(\infty) = 1$. It may sometimes be useful to calculate things like: $P(a < X \leq b) = F(b) - F(a)$, or $P(X > c) = 1 - F(c)$.

Transformation of variable. If $\pi(x)$ is a probability density, and $y = g(x)$ is a continuous smooth function of x, $(x = g^{-1}(y))$, then the probability density of y is $\pi(g^{-1}(y)) \mid \frac{dx}{dy} \mid$. (Note that the support of this new density is usually different from the original).

Conditional probability for events A and B, and conditional probability density for values of $X = x$ and $Y = y$

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \text{ and } \pi(x \mid y) = \frac{\pi(x, y)}{\pi(y)}$$

Product rule. Due to symmetry of $P(A, B)$ and $\pi(x, y)$ we have: $P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$, and $\pi(x, y) = \pi(x \mid y)\pi(y) = \pi(y \mid x)\pi(x)$. The product rule leads to the most important equation in bayesian modelling: the bayes formula itself.

Sum rule. $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually distinct, i.e. $P(A, B) = 0$. Otherwise, more generally, $P(A \text{ or } B) = P(A) + P(B) - P(A, B)$.

Expected value of a random variable $X \in S \subset \mathbb{R}^n$

$$E(X) = \sum_i x_i P(X = x_i) \qquad \text{or} \qquad \int_S x\pi(x)\mathbf{d}x,$$

where P denotes probability and π denotes probability density.

Variance of a random variable:

$$V(X) = E(X - E(X))^2 = \sum_i (x_i - E(X))^2 P(X = x_i) \qquad \text{or} \qquad \int_S (x - E(X))^2 \pi(x)\mathbf{d}x$$

Variance can also be written in this form:

$$V(X) = E(X^2) - (E(X))^2.$$

Independence and conditional independence. Variables X and Y are said to be independent if $P(X,Y) = P(X)P(Y)$. They are said to be conditionally independent, given Z, if $P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. (Likewise with probability densities $\pi(X,Y \mid Z)$).

If variables X & Y are independent, then the following equations hold:

$$E(XY) = E(X)E(Y) \text{ and } V(X+Y) = V(X) + V(Y).$$

The following equations hold for any random variables, whether they are independent or not:

$$E(X+Y) = E(X) + E(Y) \text{ and } V(cX) = c^2V(X) \text{ and } E(cX) = cE(X),$$

where c is a constant.

Conditional expected value $E(X \mid Y)$. This is obtained from the previous formulation of $E(X)$ by substituting the distribution of X by the conditional distribution of X. The marginal expected value can be written as $E(X) = E(E(X \mid Y))$, where the outer expected value is taken with respect to Y.

Conditional variance $V(X \mid Y)$. This is similar to conditional expected value. But now we have: $V(X) = E(V(X \mid Y)) + V(E(X \mid Y))$.

Marginal probability:

$$P(X) = \sum_i P(X, y_i) \quad \text{if } X \in \{x_1, \ldots\} \text{ and } Y \in \{y_1, \ldots\} \text{ discrete.}$$

$$\pi(x) = \int_{S_y} \pi(x,y)\mathbf{d}y \quad \text{if } X \text{ and } Y \text{continuous.}$$

Marginal probability is computed similarly from multivariate models; by 'integrating out' the other variables. The marginal probability can also be computed for a k-dimensional vector variable that is part of a n-dimensional larger vector, $(n > k)$, for which the *joint distribution* is $P(X_1, \ldots, X_n)$. Using marginal distributions is an essential practical method for computing and visualizing results from multidimensional joint distributions. This will be used in nearly all practical bayesian applications!

A special random variable: indicator variable

$$I_{\{A(x)\}}(x) = \begin{cases} 1 & \text{if } A(x) \text{ is true} \\ 0 & \text{if } A(x) \text{ is false} \end{cases}$$

For an indicator variable we obtain:

$$E(I_{\{A(x)\}}) = 1P(A(x) \text{ is true}) + 0P(A(x) \text{ is false}) = P(A(x) \text{ is true})$$

The indicator variable is sometimes convenient in mathematical manipulations. Moreover, it will later provide us a simple tool for calculating many probabilities in WinBUGS by taking the average of a suitable indicator variable.

Completion of squares. This is a mathematical routine that is often used in solving posterior densities with Gaussian (normal) models. A square that needs to be completed is typically of the form $(a-b)^2 = a^2 - 2ab + b^2$. An incomplete square is thus completed by adding and subtracting one of the missing terms, e.g.:

$$a^2 - 2ab = a^2 - 2ab + b^2 - b^2 = (a-b)^2 - b^2.$$

In matrix algebra, if a and b are vectors (of size $n \times 1$):

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

and $a^T = (a_1, \ldots, a_n)$ and $b^T = (b_1, \ldots, b_n)$ are transposes of the vectors (of size $1 \times n$), the square (a scalar) is:

$$(a-b)^T(a-b) = a^Ta - a^Tb - b^Ta + b^Tb = a^Ta - 2a^Tb + b^Tb$$

Special functions:

Gamma-function, some useful properties: $\Gamma(N+1) = N!$, and $\Gamma(N+1) = N\Gamma(N)$ for integers N.

Beta-function: $\text{beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

2.1 Exercises

1. Microbiological sampling is repeated three times under same conditions. Assume the probability to catch salmonella in each is p. What is the probability $P(A)$ for the event $A =$"at least one of the three samples contains salmonella"? Assume that the sample of three is not analyzed as three separate tests but as a single pooled sample, and the detection probability of the pooled sample is q given that at least one of the three subsamples contained salmonella. What is the probability for the event $B =$"negative result for the pooled test", and $\bar{B} =$"positive result for the pooled test". Show that they add up to one.

2. Assume that influenza types T_1, T_2 and T_3 each will spread into a population, reaching prevalence $p_{T_1} = 0.05, p_{T_2} = 0.1$ and $p_{T_3} = 0.3$. Assume that influenza specific mortalities are $P(\dagger \mid T_1) = 0.005$, $P(\dagger \mid T_2) = 0.0001$, and $P(\dagger \mid T_2) = 0.00001$. For simplicity, assume also that an individual can only acquire one of the diseases or none. What is the probability of death ($P(\dagger)$) for an arbitrary citizen?

3. To continue the previous exercise, assume that a disease of type T_i will cause costs C_i per each case of illness. The expected values of the costs are $E(C_1) =$ EUR 100, $E(C_2) =$ EUR 200, $E(C_4) =$ EUR 400. What is the expected cost for an arbitrary citizen? Use the conditional expected value in calculations. (Hint: indicator variable may help).

4. Assume that the joint probability for X (rain='yes'/'no'=1/0) and Y (windy='yes'/'no'=1/0) is given in the table below. (Chosen numbers are purely fictional).

	Y=0	Y=1
X=0	0.3	0.4
X=1	0.1	0.2

What is the conditional probability that it rains if it is windy? What is the conditional probability that it is windy if it rains? What is the marginal probability of rain, and the marginal probability of wind? Are rain and wind independent?

5. Extend the previous model of X, Y with a new variable Z (cloudiness='High'/'Low') so that X and Y have the following conditional probabilities (again fictional), given Z:

	Z=0		Z=1	
	Y=0	Y=1	Y=0	Y=1
X=0	0.45	0.45	0.15	0.35
X=1	0.05	0.05	0.15	0.35

Are X and Y conditionally independent, given Z? Verify that with $P(Z = 1) = 0.5$, this extended model gives the previous table of marginal probabilities $P(X, Y)$.

6. If X and Y are independent, show that $P(X \mid Y) = P(X)$.

7. Using product rule, factorize the joint density $\pi(x, y, z)$ into a product of three parts.

8. Broiler flocks can be either infected or not infected with probability p. Define an indicator variable for each flock $i = 1, \ldots, n$:

$$I_i = \begin{cases} 1 & \text{if flock } i \text{ is infected} \\ 0 & \text{otherwise.} \end{cases}$$

Assume that the flock size has some distribution with mean μ and the flocks are independent.

Show that the expected number of broilers in infected flocks is $n\mu p$.

9. Show that the function $\pi(x) = -\ln(x)$ is a probability density over the interval $[0, 1]$. (You need this detail: $\lim_{x\to 0^+} x\ln(x) = 0$).

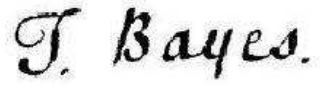
10. X has uniform probability density over $[0, 1]$, ($\pi(x) = 1$ if $x \in [0, 1]$ and zero elsewhere). What is the probability density of $Y = \ln(X)$? Show that it integrates to one.

3 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.

3.1 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" [6], 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.

3.2 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:

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

3.3 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

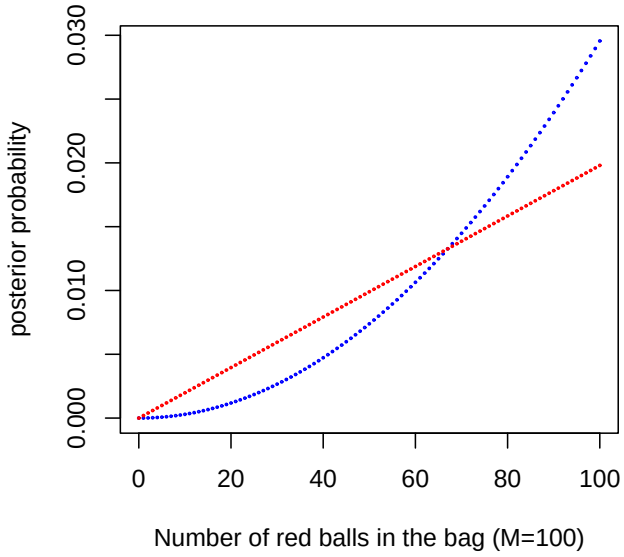


Figure 2: Posterior probabilities for the number of red balls among M in a bag, if one ball is drawn and it is red (red dots), and if two balls are drawn and both are red (blue dots).

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 [2] and *subjective* bayesian [3] 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. A free software to assist in expert elicitation can be found from

`http://www.tonyohagan.co.uk/shelf/`

Something to think about: was Bayes a bayesian? We can only speculate. Perhaps not in the sense of modern bayesian analysis which was rigorously developed much later, both for subjective and objective bayesian approaches. Also, it was Laplace who more extensively developed and actually put the probability theory into action for many inferential problems, whereas Bayes only wrote the first example. Therefore, some of us advocate the notion of 'probabilistic inference' instead of 'bayesian inference'. This would highlight the fact that the approach is more about

using probability theory 'as the logic of science', than about Bayes. The only example by Bayes was about billiard balls, and the uniform prior was maybe based on the (frequentist?) idea that it could represent the 'physical' and 'true' distribution of the position of the first ball on the table. Interpreting probability as our subjective uncertainty, this does not need to be the assumption. It could well be that someone just chooses to put the first ball at some position, without any real 'physical random experiment'. The question is then: what is our uncertainty about the position, if we do not know at all how the ball was placed? Could we use the uniform distribution to describe that? For us, as far as we could know, any position would be just as 'possible' as any other. The prior then is really an expression of subjective uncertainty. Even if we would be told that there was a physical experiment of 'pushing the ball randomly', remember what E.T. Jaynes writes about the 'physics of random experiments'. It would still be a question of how accurately we know the initial conditions. If these were exactly known, there would be no 'true randomness' left.

The dilemma is still alive and occasionally seen in some published papers where the authors may accept only priors which aim to describe the 'true distribution' before subsequent data are obtained. They would accept the uniform prior only if it was based on 'knowing' that there was a real experiment that would result into more or less uniform distribution of the 'first ball' if repeated many times. Yet, even so, there is no randomness *after* the ball has stopped, but the position could still be uncertain to us. In situations where the 'first ball' is just placed in an unknown way, some would not accept any prior distribution if 'we have no way to tell what the true distribution then is'. But this reasoning reduces the applicability to frequentist problems! If we consider more extended set of problems like Laplace, e.g. in celestial mechanics, then we need to quantify uncertainty e.g. about the mass of Jupiter. Or if we want an estimate of the population in the geographical area of Helsinki in the year 1330, or the winner of next presidential election. There is more or less uncertainty, but no repeated experiment involved.

Figure 3: Laplace in Versailles castle.

3.3.1 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:

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 (assumed) symmetry of the coin. Likewise, probability of Ace of Spades is 1/52 due to (assumed) symmetry of the cards, etc. 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. However, symmetry argument is very closely related to the concept of exchangeability in bayesian inference, but in terms of symmetry with respect to available information on the finite sequence of random variables $X_1, \ldots, X_n$, rather than symmetry of the physical object.

Frequentist definition: probability of event A is the limiting frequency of occurrences of A in a series of repeated experiments under same conditions. But this limited frequency is always unknown to us, because we cannot repeat any experiment truly infinitely. (Compare with bayes: all probabilities are known!). Moreover, the requirement of 'same conditions' is not exactly defined for any practical problem. However, beliefs about the limiting proportion take the form of a subjective prior distribution as a *consequence* of assuming infinitely exchangeable sequence of random variables $X_1, X_2, \ldots$ in de Finetti's representation theorem which is foundational in the subjective bayesian theory.

3.4 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.

3.4.1 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.

3.4.2 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:

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

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))
}
```

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).

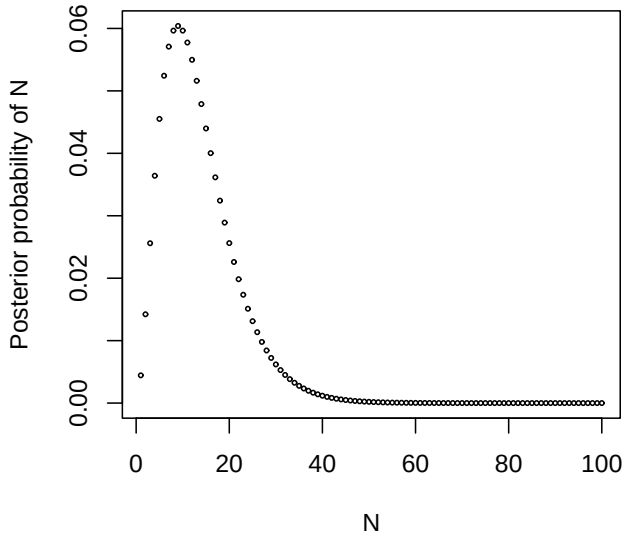


Figure 4: Posterior probability for N, given that $X = 1, r = 0.2$ with uniform prior over $0, 1, 2, \ldots, M = 100$.

3.5 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': $\propto$.

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

3.6 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:

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

Bayesian inference consists of the first four steps.

3.6.1 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.

3.7 Exercises

1. Explain the role of uncertainty (epistemic uncertainty) and variability (aleatory uncertainty) in the context of Bayesian inference by using some example.

2. In the example of eliciting your subjective probability of event A you are offered to choose between a (1) lottery ticket and a (2) possible reward if event A occurs. What is the expected value E_i of your gain if you choose option i? Solve $P(A)$ from the equation $E_1 = E_2$.

3. Grandfather goes shopping in four shops. In each shop, he can forget his umbrella with probability 0.25, and remembers it with probability 0.75. After he has gone through all four shops, what is the probability that the umbrella was forgotten in the third shop? What is the probability that it was forgotten in the third shop, given that he really did forget it in one of the shops?

4. In Bayesian inference, the goal is to infer (using Bayes's formula) about some quantity X based on observations about another quantity Y. If the joint distribution $P(X, Y)$ is known, show that either $P(Y \mid X)$ or $P(X \mid Y)$ could be derived from this. The fact that we can make inference about Y based on observed X, or vice versa, does not mean that Y is caused by X, or X caused by Y. Find examples.

5. Continue Exercise 1.2. Assume there is some genotype in humans that can protect from serious infections, and the population prevalence of this type is $g = 0.03$. If a person has the genotype and is infected, the conditional probability of survival is one, regardless of the virus type. If a person does not have the genotype, the conditional probabilities only depend on the influenza type as given in the exercise. Assume a patient survived infection. What is the probability that he had this advantageous genotype?

6. Approximately 1/125 of all births are fraternal twins and 1/300 identical twins. Elvis Presley had a twin brother who died at birth. What is the probability that Elvis was an identical twin?

7. Show that we really get the same posterior probability formula as above when using $P(r = i/N) = 1/(N + 1)$ as the prior, and $P(1^{\text{st}}X = \text{red}, 2^{\text{nd}}X = \text{red} \mid r = i/N)$ as the model for observed data.

8. You strongly suspect your car must have either electricity problems, or some other problems, but not both. If the car has electricity problems, the key's remote control button won't open the door with probability $P(\text{not open} \mid \text{e-problem}) = 0.1$. Otherwise, the probability is $P(\text{not open} \mid \text{other problem}) = 0.2$. Then, $P(\text{e-problem}) + P(\text{other problem}) = 1$, and $P(\text{no problem}) = 0$. Now, you push the key-button and nothing happens. What is the probability that the car really has e-problems? Discuss how this depends on your initial beliefs about the existence of e-problems versus other possible problems. (Plot the posterior as a function of prior). In this model, have you included the possibility that you might have wrong keys?

9. When dad bakes gingerbread, they get burned with probability 0.5. When mom bakes, they get burned with probability 0.9. Gingerbread is made every month. Every other month dad is cooking, and mom is cooking for other months. Their son is visiting at a random day and served with unburned gingerbread. What is the probability for him, that dad was the cook? What if he knows dad is cooking for one month per year only? What if he knows the calendar for cooking shifts? Which is more decisive: the prior or the data?

4 Beyond binomial models

4.1 Poisson-distribution

Poisson-distribution is one of the most commonly used models in e.g. reliability research and epidemiology. It is used for describing number of 'rare events'. Poisson distribution can be derived as a limiting case of binomial distribution $\text{Bin}(N_k, r_k)$ when $N_k \to \infty$ and $r_k \to 0$ so that the product $N_k r_k \to \lambda$, when $k \to \infty$. Then, the (Poisson) distribution of a single observation $X \in \{0, 1, 2, 3, \ldots\}$ is

$$P(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}.$$

The Poisson distribution also emerges from Poisson process (a special case of stochastic process) with constant intensity λ. If, e.g. accidents occur with constant intensity λ per time unit, then the expected number of accidents in a time unit is λ and the number of them (per time unit) follows Poisson distribution with parameter λ, which is both the mean and the variance of Poisson distribution. Due to additivity of Poisson variables, if $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$, then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. Likewise, the number of events during time T has Poisson distribution $\text{Poisson}(\lambda T)$. In a Poisson process with constant intensity λ, the waiting time until next event is exponentially distributed with mean $1/\lambda$, regardless of the past history, (if λ given).

As a conjugate distribution, the prior of λ is $\text{Gamma}(\alpha, \beta)$-density

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which leads to the posterior:

$$\pi(\lambda \mid X) \propto \frac{\lambda^X}{X!} e^{-\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

which is, up to a normalizing constant, the same as

$$\lambda^{x+\alpha-1} e^{-(1+\beta)\lambda}.$$

In other words: $\text{Gamma}(X + \alpha, 1 + \beta)$-density. The posterior mean is thus

$$E(\lambda \mid X, \alpha, \beta) = \frac{X + \alpha}{1 + \beta} = \frac{1}{1 + \beta} X + \frac{\beta}{1 + \beta} \frac{\alpha}{\beta}$$

which is a weighted average of prior mean α/β and X. If we have a series of observations $X_1, \ldots, X_n$, an analogous result can be derived.

4.1.1 Example: Asthma mortality

Epidemiological Example from Gelman [4]: Poisson model parameterized in terms of rate and exposure:

$$X_i \sim \text{Poisson}(E_i \theta)$$

where X_i is the number of e.g. disease cases in a group with exposure E_i and θ is the unknown parameter of interest, the 'underlying rate'. The probability of the data $X = (X_1, \ldots, X_N)$ is

$$\pi(X \mid \theta) \propto \theta^{\sum_{i=1}^{N} X_i} \exp((-\sum_{i=1}^{N} X_i)\theta)$$

With the conjugate prior Gamma(α, β), the posterior is

$$\pi(\theta \mid X) = \text{Gamma}(\alpha + \sum_{i=1}^{N} X_i, \beta + \sum_{i=1}^{N} E_i)$$

Assume there were $X = 3$ deaths due to asthma in a city during a year, out of a population of 200000. Hence the crude estimate per 100000 per year would be 1.5 cases. The model for the observed count could be

$$X \sim \text{Poisson}(2\theta) = \text{Poisson}(E\theta)$$

where θ represents the 'underlying mortality rate' per 100000 per year, and E 'exposure'. To compute the posterior $\pi(\theta \mid X)$, we choose a conjugate prior $\pi(\theta) = \text{Gamma}(\alpha, \beta)$ by choosing (α, β) so that the prior represents reasonable background information. According to literature, the typical asthma mortality rate in Western countries would be around 0.6 per 100000. It is also known that values above 1.5 are rare. Hence, Gamma$(3, 5)$ prior has mean 0.6, standard deviation 0.35, and this prior also has $P(\theta < 1.44) = 97.5\%$. All this seems to fulfill both prior specifications. (The prior parameters can be chosen by trial and error). The posterior distribution is then Gamma$(6, 7)$, which has mean 0.86. That is substantial shrinkage towards prior distribution.

(Note: in spatial analysis, the prior can be chosen to represent local dependencies so that if a geographically small area with small population can only provide weak data, the estimate becomes influenced by its neighbors, i.e. 'borrow strength'. This is used for bayesian smoothing in spatial disease mapping. See GeoBUGS manual. Similar prior construction based on 'neighbors' can be used for one-dimensional smoothing as well, e.g. in temporal models).

4.2 Exponential distribution

Assume a single observation $X \in \mathbb{R}^+$ (typical example: waiting times, time of next event) for which the conditional distribution is exponential:

$$\pi(X \mid \theta) = \theta \exp(-X\theta).$$

As a conjugate prior of θ, we choose Gamma(α, β), so that the posterior $\pi(\theta \mid X)$ becomes Gamma$(\alpha + 1, \beta + X)$. The posterior mean is

$$E(\theta \mid X, \alpha, \beta) = \frac{\alpha + 1}{\beta + X}$$

With a set of observations $X_1, \ldots, X_n$ (mean $\bar{X} = \sum_{i=1}^{n} X_i/n$) we get

$$\pi(X \mid \theta) = \theta^n \exp(-n\bar{X}\theta)$$

which leads to the posterior Gamma$(\alpha+n, \beta+n\bar{X})$, so that the Gamma$(\alpha, \beta)$ prior can be thought as equivalent of $\alpha - 1$ prior observations $X_1^0, \ldots, X_{\alpha-1}^0$ for which the sum $\sum X_i^0$ equals to β.

4.2.1 Censored data

In survival analysis and reliability applications, it is common that the 'failure times' (times of death, infections, illness, etc.) are exactly known for only some individuals. For others, the time can be censored, which means that we only know that the event has not happened before some known time point. (This is also information!). Often, the censoring time can be the ending time of the follow-up period, or ending time of the study, T. The probability for such event is the *survival probability*: $P(X_i > T \mid \theta) = 1 - P(X_i < T \mid \theta) = 1 - F(T \mid \theta) = \exp(-\theta T) = S(T \mid \theta)$. The conditional probability of the whole data is then of the form

$$P(X \mid \theta) = \prod_{i=1}^{k} \theta \exp(-\theta X_i) \times S(T \mid \theta)^{n-k} = \theta^k \exp(-\theta[\sum_{i=1}^{k} X_i + (n-k)T]).$$

The posterior is then Gamma$(\alpha+k, \beta+\sum_{i=1}^{k} X_i+(n-k)T)$. More generally, we may know that for some individuals the event occurred before some given time, or between two given times. In each case, this information should be included by writing the corresponding conditional probability. (This is sometimes called as the 'full likelihood'). For example, if some events are only known to have been before time T_1 and some are known to be after time T_2, and for the rest we know the exact time, then the full likelihood would be of this form

$$P(X \mid \theta) = \prod_{i \in E_1} F(T_1 \mid \theta) \times \prod_{i \in E_2} S(T_2 \mid \theta) \times \prod_{i \in E_3} \theta \exp(-\theta X_i).$$

Note: by using the cumulative probability function F, probability expressions for all different situations of censoring might be written.

Note: when the event time is known, the conditional probability of this observation is $P(X_i \mid \theta) = \theta \exp(-\theta X_i)$, but when the censoring time is known, the observation can be interpreted as a Bernoulli variable (indicator variable!) that was one:

$$Y_i = \begin{cases} 0 & \text{if } X_i < T \\ 1 & \text{if } X_i > T \end{cases}$$

so that $P(Y_i = 1 \mid \theta) = S(T \mid \theta)$.

4.3 Normal-distribution

The normal, or Gaussian, distribution is the most widely used model and has connections to many other models and their asymptotic approximations. For an example of bayesian inference, consider that a measurement, e.g. temperature, is measured from N items, resulting to $X_1, \ldots, X_N$ as the observed temperatures. These are assumed to be normally distributed with mean μ and variance σ^2, representing the infinite population from which the items are drawn. We then have two unknown parameters in our model. Consider first estimating one of them, assuming the other as 'known', and finally estimating both.

4.3.1 Estimating the mean

Assume that variance σ^2 is known, but mean μ unknown. We would like to estimate the mean, representing the average temperature in an infinite population of items. Consider first a single observation. The conditional density is

$$\pi(X_i \mid \mu, \sigma) = \mathrm{N}(X_i \mid \mu, \sigma^2) = \mathrm{N}(X_i \mid \mu, \tau) \propto \exp(-0.5\tau(X_i - \mu)^2).$$

where $\tau = 1/\sigma^2$ is the *precision*. (Gaussian model is parameterized using precision in WinBUGS and often in bayesian notation. Be careful to note which notation is used!!). Before calculating posterior of μ, we need to choose the prior. As we know from physics, there is absolute minimum temperature, but for this example we assume that our measurements are well beyond absolute minimum. Therefore, for all practical purposes it is acceptable to consider the whole set $\mathbb{R}$ of real numbers as the range of possible measurement values. It is convenient to use a conjugate prior density, $\mathrm{N}(\mu_0, \tau_0)$:

$$\pi(\mu) \propto \exp(-0.5\tau_0(\mu - \mu_0)^2).$$

With the single measurement, the posterior density would be of the form

$$\pi(\mu \mid X_i, \tau, \mu_0, \tau_0) \propto \exp(-0.5(\tau_0(\mu - \mu_0)^2 + \tau(X_i - \mu)^2)),$$

and this is the same as

$$\mathrm{N}\Big(\frac{n_0\mu_0 + X_i}{n_0 + 1}, \frac{\sigma^2}{n_0 + 1}\Big),$$

where $n_0 = \tau_0/\tau$ can be interpreted as *a priori* sample size. The normal density is obtained from the bayes formula by using the technique of completing a square. (See [9] BSM p. 62). The posterior mean can be written as

$$w\mu_0 + (1 - w)X_i,$$

where the weight is $w = \tau_0/(\tau_0 + \tau)$. The probability of the whole data set can be written using the average $\bar{X} = \sum X_i/N$:

$$\pi(\bar{X} \mid \mu, \sigma) = \mathrm{N}(\bar{X} \mid \mu, \sigma^2/N) = \mathrm{N}(\bar{X} \mid \mu, N\tau).$$

By using bayes formula, this leads to the posterior

$$\mathrm{N}\Big(\frac{n_0\mu_0 + \bar{X}}{n_0 + 1}, \frac{\sigma^2/N}{n_0 + 1}\Big),$$

with $n_0 = \tau_0/(N\tau)$. The posterior mean and variance can also be written in this form:

$$E(\mu \mid X) = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{X}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}} \qquad V(\mu \mid X) = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}.$$

Improper prior. When the prior precision approaches zero, the prior density becomes flat and improper density, $\pi(\mu) \propto 1$, but the posterior density still exists, becoming $\mathrm{N}(\bar{X}, \sigma^2/N)$. The posterior mean then equals sample mean, and posterior variance equals the variance of the sample average. This is a perfect mirror image of the non-bayesian approach where a sampling distribution is derived for a *statistics*, such as sample mean, whereas the unknown population mean μ is considered constant. In bayesian inference μ is unknown, therefore random, but the data $\bar{X}$ is known, therefore constant:

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/N) \qquad \mu \sim \mathrm{N}(\bar{X}, \sigma^2/N)$$

4.3.2 Estimating the variance

It is next assumed that the mean μ is known, and we would like to estimate the unknown variance σ^2, (or precision τ). It is not sensible to estimate variance unless there are several (at least more than one) observations. Therefore, we assume that we have some number of observations $X = X_1, \ldots, X_N$. We can start again with the conditional density of all observations:

$$\pi(X \mid \mu, \sigma) \propto \sigma^{-N} \exp(-\frac{1}{2\sigma^2} \sum_i^N (X_i - \mu)^2).$$

$$= (\sigma^2)^{-N/2} \exp(-\frac{N}{2\sigma^2}\nu)$$

where we have used the notation:

$$\nu = \frac{1}{N} \sum_i^N (X_i - \mu)^2.$$

Since τ is unknown we must choose a prior for it. Following the presentation in Gelman et al [4], a convenient choice for the prior $\pi(\sigma^2)$ is a Scaled Inverse χ^2 distribution. The density function is:

$$\pi(\sigma^2 \mid \nu_0, \sigma_0^2) = \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \sigma_0^{\nu_0} (\sigma^2)^{-(\nu_0/2+1)} \exp(-\nu_0\sigma_0^2/2\sigma^2)$$

It has two parameters, ν_0, σ_0^2, and it has some connections to other densities, which provide alternative ways in constructing the prior:

$$\sigma^2 \sim \text{Scaled Inv-}\chi^2(\nu_0, \sigma_0^2) = \text{Inv-}\Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}) \iff \tau = \frac{1}{\sigma^2} \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$

$$\theta \sim \chi^2_{\nu_0} \iff \frac{\nu_0\sigma_0^2}{\theta} \sim \text{Scaled Inv-}\chi^2(\nu_0, \sigma_0^2)$$

Using the Scaled Inverse-χ^2 prior, the posterior is of the form:

$$\pi(\sigma^2 \mid X) \propto \pi(\sigma^2)\pi(X \mid \sigma^2)$$

$$\propto (\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{\nu_0\sigma_0^2}{2\sigma^2}) \times (\sigma^2)^{-N/2} \exp(-\frac{N\nu}{2\sigma^2})$$

$$= (\sigma^2)^{-((\nu_0+N)/2+1)} \exp(-\frac{\nu_0\sigma_0^2 + N\nu}{2\sigma^2})$$

Which is the Scaled Inverse-χ^2 density:

$$\pi(\sigma^2 \mid X, \mu) = \text{Scaled Inv-}\chi^2(\nu_0 + N, \frac{\nu_0\sigma_0^2 + N\nu}{\nu_0 + N}).$$

Note that the prior can be thought of as ν_0 observations with average squared SD of σ_0^2.

Improper prior. When the ν_0 parameter of the prior is set to zero, we obtain an improper prior

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

which does not integrate to one. Nevertheless, the posterior density still exists, and it is Scaled Inv-$\chi^2(N, \nu)$ where $\nu = \frac{1}{N}\sum(X_i - \mu)^2$. The prior is equivalent to the uniform improper prior $\pi(\log(\sigma)) \propto 1$.

4.4 Multiparameter models

In nearly all inference problems there is more than one unknown quantity. Often, only one of them is of interest and the others are *nuisance parameters*. Assume there are two unknown parameters θ_1, θ_2 (both can be vectors) and some set of data X. The posterior density is

$$\pi(\theta_1, \theta_2 \mid X) \propto \pi(X \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2),$$

and the marginal density of θ_1 is

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1, \theta_2 \mid X)\mathbf{d}\theta_2,$$

which can also be calculated as

$$\pi(\theta_1 \mid X) = \int \pi(\theta_1 \mid \theta_2, X)\pi(\theta_2 \mid X)\mathbf{d}\theta_2.$$

This integral is usually not computed directly, but it shows an important structure that is used when hierarchical models are constructed, and also when MCMC algorithms are implemented.

Note: the unknown parameters θ can be 'unknown model parameters', or missing data variables, or variables to be predicted, or unobservable latent (hidden) variables. They are all simply unknown, and in bayesian inference they are all treated as unknown quantities, so that we aim to compute the posterior:

$$P(\text{'all unknowns'} \mid \text{'all known things'})$$

Note: it is difficult to visualize a posterior density for three or more unknown quantities. Therefore, we often plot one-dimensional marginal distributions, or two-dimensional marginal distributions for selected quantities of interest. This is always based on the full posterior density that can be multidimensional.

4.4.1 Multinomial model, unknown $r_1, \ldots, r_k$

Binomial model can be generalized to multinomial model by considering outcomes of several types instead of two types. For example, in a large bag there are balls of k different colours. The proportions of these are $r = r_1, \ldots, r_k$. A sample of N balls is drawn, and we observe the number of balls of each colour $X_1, \ldots, X_k$. The goal is now to solve the posterior density:

$$\pi(r_1, \ldots, r_k \mid X_1, \ldots, X_k).$$

Note that the unknown proportions have to sum to one: $\sum r_i = 1$. The conditional distribution of the data is now

$$P(X_1, \ldots, X_k \mid r_1, \ldots, r_k, N) = \binom{N}{X_1, \ldots, X_k} r_1^{X_1} \times \cdots \times r_k^{X_k}.$$

The conjugate prior density is $\text{Dir}(\alpha) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$:

$$\pi(r_1, \ldots, r_k) = \frac{\Gamma(\alpha_1, \ldots, \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} r_1^{\alpha_1 - 1} \times \cdots \times r_k^{\alpha_k - 1},$$

so that the posterior density will also be Dirichlet, with parameters $(\alpha_1 + X_1, \ldots, \alpha_k + X_k)$:

$$\propto r_1^{\alpha_1+X_1-1} \times \cdots \times r_k^{\alpha_k+X_k-1}.$$

Again, prior parameters $\alpha_1, \ldots, \alpha_k$ can be interpreted to represent 'prior data' so that the 'prior sample size' is $\sum \alpha_i$. A usual uninformative prior choice is $\text{Dir}(1, \ldots, 1)$, which is the generalization of $\text{Beta}(1,1)$. Note: this prior is equal to having k prior data points. If we want a prior to weight as just one data point, then we should define $\alpha_i = 1/k$. This is analogous with the $\text{Beta}(1/2, 1/2)$-prior for binomial problems, as opposed to $\text{Beta}(1,1) = \text{U}(0,1)$. The posterior means can be written as weighted mean of prior and data proportions

$$E(r_i \mid X, \alpha) = \frac{\alpha_i + X_i}{\sum(\alpha_i + X_i)} = \frac{\sum \alpha_i}{\sum(X_i + \alpha_i)} \frac{\alpha_i}{\sum \alpha_i} + \frac{\sum X_i}{\sum(X_i + \alpha_i)} \frac{X_i}{\sum X_i}$$

Note also that if $r \sim \text{Dir}(\alpha)$, then the marginal distribution of each r_j is $\text{Beta}(\alpha_j, \sum_i \alpha_i - \alpha_j)$, with variance $\alpha_j(\sum_i \alpha_i - \alpha_j)/((\sum_i \alpha_i)^2(\sum \alpha_i + 1))$. To simplify notations, write $A = \sum_i \alpha_i$. Then the marginal variance may be written as $\frac{\alpha_j}{A}(1 - \frac{\alpha_j}{A})/(A+1)$.

If dirichlet distribution is not found in a software, the following result can be useful:

$$Z_i \sim \text{Gamma}(\alpha_i, 1) \quad \Rightarrow \quad (\frac{Z_1}{\sum Z_i}, \ldots, \frac{Z_k}{\sum Z_i}) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

Congdon, BSM, p. 38, shows a possible method for constructing an informative prior based on 'expert opinion'. It starts by picking up two estimates for the parameters. Denote them as $p_1, \ldots, p_k$ and $q_1, \ldots, q_k$. Their differences are $d_i = p_i - q_i$ and means are $\eta_i = (p_i + q_i)/2$. The expected value of the sum of squared differences can be written as

$$\begin{aligned}
&E(\textstyle\sum(p_i - q_i)^2) = \sum E(p_i^2) - 2\sum E(p_i)E(q_i) + \sum E(q_i^2) \\
&= \textstyle\sum(2E(p_i^2) - 2E(p_i)^2) \\
&= 2\textstyle\sum V(p_i) \\
&= 2\textstyle\sum \eta_i(1 - \eta_i)/(A+1) \\
&= 2\textstyle\sum(\eta_i - \eta_i^2)/(A+1) \\
&= 2(1 - \textstyle\sum \eta_i^2)/(A+1).
\end{aligned}$$

This expected value and the prior 'observed' sum of squared differences of the two estimates are marked as equal. Then, using the 'observed' prior estimate η_i, we can solve the prior sample size A. The parameters of the Dir-prior are finally obtained as $(A\eta_1, \ldots, A\eta_k)$.

4.4.2 Normal model, unknown μ and σ^2

The goal is to solve the posterior (joint) density $\pi(\mu, \sigma^2 \mid X)$, i.e. both parameters are unknown. The prior density is assumed **improper** and uninformative so that

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This prior is the same as an improper uniform prior

$$\pi(\mu, \log(\sigma)) \propto 1.$$

First, there's some preliminary math that will be needed when solving the posterior density.

$$\sum_i^n (X_i - \mu)^2 = \sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Proof:

$$\sum_i^n (X_i - \mu)^2 = \sum_i^n (X_i^2 - 2X_i\mu + \mu^2)$$

$$= \sum_i^n (X_i^2 - 2X_i\mu + \mu^2 - \bar{X}^2 + \bar{X}^2 - 2X_i\bar{X} + 2X_i\bar{X})$$

$$= \sum_i^n (X_i - \bar{X})^2 + \sum_i^n (\mu^2 - 2X_i\mu - \bar{X}^2 + 2X_i\bar{X})$$

$$= \sum_i^n (X_i - \bar{X})^2 + n(\mu^2 - 2\bar{X}\mu - \bar{X}^2 + 2\bar{X}\bar{X}) = \sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Then, using this 'trick', the posterior density can be solved as

$$\pi(\mu, \sigma \mid X) \propto \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} \sum_i^n (X_i - \mu)^2)$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} [\sum_i^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{X} - \mu)^2]),$$

where $s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$.

The posterior density is finally solved by using factorization:

$$\pi(\mu, \sigma^2 \mid X) = \pi(\mu \mid \sigma^2, X) \pi(\sigma^2 \mid X).$$

We already know from earlier results that $\pi(\mu \mid \sigma^2, X) = \mathrm{N}(\bar{X}, \sigma^2/n)$. Therefore, we only need to find out what the marginal density $\pi(\sigma^2 \mid X)$ is. This can be calculated from the joint density by integrating over μ:

$$\pi(\sigma^2 \mid X) \propto \int_{-\infty}^{\infty} \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{X} - \mu)^2]) \mathbf{d}\mu$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} (n-1)s^2) \times \int_{-\infty}^{\infty} \exp(-\frac{n}{2\sigma^2} (\bar{X} - \mu)^2) \mathbf{d}\mu$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2} (n-1)s^2) \times \sqrt{2\pi\sigma^2/n}$$

$$\propto (\sigma^2)^{-(n+1)/2} \exp(-\frac{(n-1)s^2}{2\sigma^2}).$$

In other words: $\pi(\sigma^2 \mid X) = \text{Scaled Inv-}\chi^2(n-1, s^2)$.

Compare this with the earlier result where μ was assumed to be known.

The full joint density can thus be computed as a product of two known densities $\pi(\sigma^2 \mid X)$ and $\pi(\mu \mid \sigma^2, X)$. This is also convenient for Monte Carlo implementations, because we can then simulate both unknown parameters from these known distributions. This example happens to be such that it is also possible to solve the marginal posterior density of the mean $\pi(\mu \mid X)$. This follows from calculating the integral:

$$\pi(\mu \mid X) = \int_0^\infty \pi(\mu, \sigma^2 \mid X) \mathbf{d}\sigma^2.$$

The details are given in Gelman et al, [4]. As a result, the marginal posterior is found to be a t-distribution so that

$$\pi\Big(\frac{\mu - \bar{X}}{s/\sqrt{n}} \mid X\Big) = t_{n-1}.$$

Note also that a common choice is to use normal-inverse gamma prior for (μ, σ^2) so that an inverse gamma prior is applied for σ^2 and a conditional normal density for μ: $\mathrm{N}(\mu_0, c\sigma^2)$. In other words, the prior for (μ, τ) is then normal-gamma, with density

$$\pi(\mu, \tau) = (2\pi c)^{-0.5} \tau^{-0.5} \exp(-\frac{\tau}{2c}(\mu - \mu_0)^2) \times \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau)$$

The resulting posterior for (μ, σ^2) is then normal-inverse gamma.

4.5 Comment

The above posterior distributions were obtained using conjugate priors. Conjugate priors are convenient, because (1) the posterior density is among well known standard densities (exact solution exists), (2) the prior can be thought of as some amount of 'prior data'. On the other hand, conjugate priors may not be flexible enough to represent more complicated prior information. But if non-conjugate priors are used, then the posterior does not take the form of any standard distribution and we must use numerical methods for all computations. Conjugate priors can only be used for a limited number of problems. They can also be useful as a first approach. Some examples are given in Table (1):

Data distribution	Prior
Binomial(n, p), n known	$p \sim$ Beta
Multinomial$(n, p_1, \ldots, p_k)$, n known	$p_1, \ldots, p_k \sim$ Dirichlet
Poisson(λ)	$\lambda \sim$ Gamma
$\mathrm{N}(\mu, \sigma^2)$, σ known	$\mu \sim$ N
$\mathrm{N}(\mu, \sigma^2)$, μ known	$\frac{1}{\sigma^2} \sim$ Gamma
$\mathrm{MN}((\mu_1, \ldots, \mu_k), \Sigma)$, μ known	$\Sigma^{-1} \sim$ Wishart
Gamma(α, β), α known	$\beta \sim$ Gamma
Beta(α, β), β known	$\alpha \sim$ Gamma
Exp(θ)	$\theta \sim$ Gamma

Table 1: Some conjugate models.

When conjugate priors are used, and the posterior solved in the form of some well known distribution, it is usually easy to draw whatever summaries from this. For example, 95% intervals, point estimates (mean, mode, median, variance), or cumulative probabilities etc., simply by using any familiar statistical software where these distributions can be found. This makes computations very fast, simple and reliable. Despite of simplifying limitations, this approach can be sufficient for basic problems, or as a useful benchmark for more complicated models. (Even if *simulation* is used as the tool for computing approximate results, the conjugate models are still the easiest, because simple Monte Carlo sampling is possible for them).

4.6 Exercises

1. Assume the number of infections X_i (per 100 000) in age groups $i = 1, \ldots, n$ are reported and $X_i \sim$ Poisson(λ). What is the posterior distribution of the mean incidence λ? Assume Gamma(α, β) prior. Interpret the prior as 'prior data'.

2. Observed bacterial counts (cfu /10 grams) in 17 samples were as follows:

```
X <- c(0, 0, 0, 0, 5, 3, 0, 0, 70, 0, 0, 0, 8, 0, 0, 3, 0)
```

Assume Poisson model $X_i \sim$ Poisson(λ). Use Gamma(α, β) prior for λ. What could be a suitable choice for prior parameters? Compute the posterior mean, mode and standard deviation of λ.

3. The observed life times were

```
X=c(1.54, 0.70, 1.23, 0.82, 0.99, 1.33, 0.38, 0.99, 1.97, 1.10, 0.40)
```

and there were 4 censored observations at time $T = 2$. Assume $X_i \sim \text{Exp}(\theta)$ and prior $\theta \sim$ Gamma$(2, 1)$. Compute posterior mean $E(\theta \mid \text{data})$.

4. Assume $X_i \sim \text{N}(\mu, \sigma^2)$. Derive the posterior of the unknown mean μ, assuming known variance σ^2 and observations $X_1, \ldots, X_N$. Prior is $\text{N}(\mu_0, \sigma_0^2)$. Hint: completion of a square.

5. Using the table of distributions, check that the mean of Y is the same as the mean of Y'

$$\alpha = \nu/2, \beta = vs^2/2 \qquad X \sim \Gamma(\alpha, \beta) \qquad Y = 1/X$$

$$Z \sim \chi^2(\nu) \qquad Y' = \nu s^2/Z$$

6. Using R, plot the posterior of σ^2 assuming that the prior information equals to having 3 observations with $\sigma_0^2 = 10$, and the data consists of 20 observations with $\frac{1}{20}\sum_{i=1}^{20}(X_i - \mu)^2 = 1.5$. Explore the posterior with different amount of data.

7. Janne Ahonen and Jakub Janda shared the Four Hills Tournament (Vierschanzentournee, Keski-Euroopan mäkiviikot) championship in 2006. Both scored a total of 1081.5 points from four competitions. Before the tournament, both took part in four other competitions. Their scores from all eight competitions were

```
ahonen <-c(299.7, 255.2, 281.7, 238.0,   270.9, 262.2, 255.4, 293.0)
janda <-c(238.7, 285.6, 287.1, 252.2,   262.6, 264.7, 263.2, 291.0)
```

Assuming a normal model $\text{N}(\mu_i, \sigma_i^2)$ for both jumpers ($i = 1, 2$) and the uninformative prior $\pi(\mu_i, \sigma_i^2) \propto 1/\sigma_i^2$, a posterior density can be obtained. If both mean and variance are unknown, and the prior is uninformative, can we have realistic results of (μ_1, μ_2) with these data only? Note: a 95% posterior interval of μ_i is obtained from t_{n-1} distribution as $\bar{X}_i \pm 1.997 s_i/\sqrt{n}$. Compare the two jumpers using such posterior intervals.

8. Assume a multinomial model $X \sim \text{Multi}(N, p)$ with a fairly small number, say $N = 4$. The data could then be e.g. $x = (0, 1, 0, 0, 2, 0, 0, 1)$. Using Dirichlet prior, solve the posterior density of p and discuss what would be an uninformative Dirichlet prior for p. How your prior will affect the result if the number of categories (dimension of p and X) is large, and yet N is small? What if we had prior information about the (unequal) proportions but if we do not wish that prior to dominate the result overwhelmingly?

5 Exchangeability and conditional independence briefly

The assumption behind the binomial distribution was that the individual experiments (balls drawn) were independent events, given the true proportion r. This means that

$$P(X_1, \ldots, X_N \mid r) = P(X_1 \mid r) \times \ldots \times P(X_N \mid r)$$

so that the exact order of the results X_i does not matter because the resulting probability will be the same as long as the sum $\sum X_i$ is the same. Hence, the binomial distribution becomes defined for the sum of 'successful events' in a series of N trials. This is a special case, in which the true proportion r encapsulates all 'essential' background knowledge. In order to calculate the probability, we need to know (or assume) a value for r. But in practice, we cannot know what r is. In a more general approach, we can study probabilities of the form

$$P(X_1, \ldots, X_N \mid I)$$

where I denotes all our background information (which we *do* have!) for a given problem, when assigning our (subjective) probability for some sequence of observations X_i. If our probability is such that it remains the same regardless of the ordering of the sequence,

$$P(X_1, \ldots, X_N \mid I) = P(X_{s_1}, \ldots, X_{s_N} \mid I)$$

for all permutations s of the indexes, then the sequence of X_i is said to be (finitely) *exchangeable*. This is an important concept in subjective bayesian modeling theory. An important result (by Bruno de Finetti, 1906-1985, http://www.brunodefinetti.it/) follows from the assumption of *infinite* exchangeability. It can be shown that then the probability can be written in the form

$$P(X_1, \ldots, X_N \mid I) = \int_0^1 \prod_i^N r^{X_i}(1-r)^{1-X_i}\pi(r)\mathbf{d}r$$

The interpretation of parameter r is that $r = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i$. It can also be interpreted as marginal probability of a single event, $r = P(X_i = 1)$.

Interpretation of de Finetti's theorem of subjective probability:

(**I**) Parameter r can be thought *as if* it was the proportion of successful events in an infinite sequence, or the probability of an individual event.
(**II**) Parameter r has to be considered as a random quantity with probability density $\pi(r)$.
(**III**) Conditionally, given r, the variables X_i are independent and equally distributed, as Bernoulli(r).

Note that parameter r emerges only as a mathematical device when the subjective probability concerning the X_i is such that it obeys exchangeability. We are still assigning probabilities for the observable events X_i. The density $\pi(r)$ is not a 'probability of probability'. We have just written our probability of the sequence X_i as a mathematical expression that directly follows from the exchangeability assumption. In fact, parameter r and its prior provide a device that allows us to update our probabilities concerning the X_i.

With the predictive approach parameters diminish in importance,
especially those that have no physical meaning.
From the Bayesian viewpoint, such parameters can be regarded as
just place holders for a particular kind of uncertainty
on your way to making good predictions. (Draper 1997, Lindley 1972).

The conditional probability $P(X_i \mid r)$ provides an important tool for parametric modeling in which we simplify our background knowledge I into a one or few parameters. This is the problem of model choice that is always a subjective choice (in all modeling, not just bayesian). The prior density $\pi(r)$ is an important tool in bayesian analysis, where the whole model is not just of the form $P(X \mid r)$, but it is the joint model $P(X, r)$ of both the observable part X *and* the unobservable part r.

Therefore, the X_i are not independent of each other, only *conditionally independent*, given r. This means that we can *learn* from the observed X_i *to predict* other X_j that are not yet observed. For example, the **prior predictive distribution** of X_1 is

$$P(X_1) = \int_0^1 P(X_1 \mid r)\pi(r)\mathbf{d}r$$

and after we have observed X_1, the **posterior predictive distribution** of X_2 is

$$P(X_2 \mid X_1) = \int_0^1 P(X_2 \mid r)\pi(r \mid X_1)\mathbf{d}r$$

where $\pi(r \mid X_1)$ is the posterior distribution of parameter r, which may be of interest in itself only if r can be taken to closely represent some real quantity of interest, e.g. percentage in a concrete but large population.

D V Lindley reports that Bruno de Finetti was especially fond of the aphorism:
Probability does not exist
which conveys his idea that probability is an expression of the observer's view of the world and as such it has no existence of its own.

Reported by D V Lindley, de Finetti insisted that "random variables" should more appropriately be called "random quantities", for "What varies?" Furthermore, coherently with his view of probabilistic thinking as a tool to deal with uncertainty in life, he thought that it should be taught to children at an early age.

Figure 5: Bruno de Finetti (1906-1985).

References

[1] Ntzoufras I: Bayesian Modeling Using WinBUGS. Wiley 2009.

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

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

[4] Gelman A, Carlin J B, Stern H S, Rubin D B: Bayesian data analysis, 2nd edition. Chapman & Hall/CRC. 2004.

[5] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515Ű533. 2006.

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

[7] Sivia D S: Data Analysis, a Bayesian tutorial, 2nd edition. Oxford university press. 2006.

[8] Robert C P, Casella G: Monte Carlo Statistical Methods. Springer 1999.

[9] Congdon P: Bayesian Statistical Modelling. John Wiley & Sons, Ltd. 2001.

[10] Congdon P: Applied Bayesian Modelling. John Wiley & Sons, Ltd. 2003.

[11] Bernardo J M, Smith A F M: Bayesian Theory. John Wiley & Sons, Ltd. 2000.