Introduction to Bayesian Inference

jukka.ranta@helsinki.fi (also: @evira.fi)

13.1.2014

Abstract

The course gives an introduction to the theory of bayesian inference and basics of WinBUGS / OpenBUGS for application examples. (Also R software will be modestly used). There are no pre-requirements other than reasonable familiarity with basic differential and integral calculus of functions of one, two and sometimes several variables. But: perhaps most essential is to know probability theory at basic level but reasonably well to be able to do some calculations. Concepts of discrete and continuous random variables, their usual distributions as well as parameterizations of these distributions should be reasonably familiar (not never-heard). In addition to 'usual' distributions such as Normal, Binomial, Multinomial, Poisson, Exponential etc. which you must have more-or-less seen or heard before, we will work with others that can be less familiar to you now: Beta, Dirichlet, Gamma, etc. You won't need to memorize exact mathematical definition of each, but you will need to recognize them, to know what they stand for and to do some short mathematical manipulations with them. For that, you will need to distinguish concepts such as 'random variable' and 'parameters' in each case, and to apply calculus to compute things like conditional probability, marginal probability, expected value, etc. at least in simple cases with given functions. Both with discrete and continuous distributions. For some examples we aim to look for analytical solutions but over time, we increasingly need to pay attention to learning the principle of the method and its underlying working conditions, to let software do computing. It can be an advantage to have knowledge of basic (non-bayesian) statistics, although not necessary. But some understanding of probability theory is needed because that is the main tool that will be used. Also an interest in applied problems is beneficial for making probabilistic model based inference from data, as well as willingness to get hands on command script based software.

Contents

1 Preliminary material and notations

Bayesian inference is very much about probability calculations. The following mathematics can be helpful, both for notations and for calculations. **You can skip this now**, but maybe check later if needed.

Discrete random variable. A variable X that can take on a finite number of possible values.

Continuous random variable. A variable X that can take on an infinite number of possible values.

Events. By an event A we denote some actual event that may (1) either occur at any time in a series of repeatable experiments, e.g. $A =$"a tail occurs in a coin toss", or (2) whose occurrence is unique e.g. $A =$"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). A unique event could also be e.g. $A =$"Eurozone will collapse in the next 5 years". 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". Note that event A could also be described as a binary random variable taking values zero ('false') or one ('true').

A special random variable: indicator variable. This is a function taking values 0 and 1 depending on its argument. For example, using event $A(x)$ whose truth value depends on x:

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

Probability for event A (in bayesian context) denotes the degree of uncertainty we have about the truth value of A. Every probability is conditional on the evidence we have. Hence, if you know well Jack (who is short) and John (who is tall), *you* might have probability $P(A \mid \text{you}) = P(X > Y \mid \text{you}) = 0$, but for an uninformed outsider, it could well be that $P(X > Y \mid \text{outsider}) = 0.5$ (before receiving any more information than just the names). With two events, A and B, the probability that *both* occur (occurrence of both is true) is written $P(A, B)$, or with specific variables: $P(X = x, Y = y)$ - if X & Y are discrete variables, or $P(X \in S_1, Y \in S_2)$ - where X & Y are any variables that may belong to sets S_i.

Probability axioms. Mathematically, probability is a *measure* that takes values between zero and one. For any event A, the probability is $P(A) \geq 0$. Also, if S is the entire sample space, then $P(S) = 1$, for example: if $B =$ "A or not A", then $P(B) = 1$. Finally, if A_i are mutually exclusive events, then $P(A_1 \text{ or } A_2 \text{ or } \ldots) = \sum_{i=1}^{\infty} P(A_i)$. Other laws can be derived from these.

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 positive function does not integrate to one but to some other constant C, $(C < \infty)$, 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, as will be shown later! The *support* of a density is the set of x values for which $\pi(x) > 0$.

Empirical distribution: the plain distribution of data. Can be visualized as a histogram.

Notations: I (often) choose the notation $\pi()$ for probability density function instead of the often used $p()$. Then we can use p to denote e.g. prevalence or proportion. More exactly, a density function should be specific to some variables, e.g. $\pi_X()$. And to be precise, also conditional on some stated evidence, e.g. $\pi_X(x \mid y)$. Shorter notations are used when the rest is (or should be) clear from the context.

Proportional to, "$\propto$" . Probability distributions are often handled without stating explicitly the normalizing constant. For example, if $\pi(p)$ is a beta distribution, we could write $\pi(p) \propto p^{a-1}(1-p)^{b-1}$, thus omitting - in this notation - the constant $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ which does not depend on p.

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 one-to-one 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 $Y = y$

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

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 inference: the Bayes theorem 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 $X \in S \subset \mathbb{R}^n$ where S is the entire sample space

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

For an indicator variable we obtain a very useful result:

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

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

Variance of a random variable:

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

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:

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

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 and marginal density

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

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

Marginal probability 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!

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

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

Special functions

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

Beta-function: $\text{beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$, $t \in [0,1]$, $\alpha > 0, \beta > 0$.

Prior probability. The probability (or density) that depends only on our previous background information I, before having the new data. This can be specified either to reflect existing relevant information, or to reflect the lack of it. "$P(\cdot \mid I)$". Often we just omit the I in the notation for simplicity.

Posterior probability. The probability (or density) that depends on the evidence obtained from the new data, in addition to the prior evidence. "$P(\cdot \mid \text{data}, I)$". Often we just omit the I in the notation for simplicity.

Bayes theorem, Bayes formula, Bayes rule. Based on probability calculus, the solution for computing posterior probability, or probability density: $\pi(\theta \mid x) = \pi(x \mid \theta)\pi(\theta)/C$, where the normalizing constant is $C = \int_\Theta \pi(x \mid \theta)\pi(\theta)d\theta$, with integral over the entire space Θ of the unknown θ. Typically, x denotes the whole set of observed data (therefore is fixed), and θ denotes the unobservable quantity (or quantities) of interest, e.g. parameters to be inferred from the data. See also *conditional probability.*

Bayesian inference. The process of updating one's prior probabilities to posterior probabilities in the light of new data. Also called bayesian learning, probabilistic inference, bayesian statistics, etc. This provides a formal method for starting with some initial description of uncertainty (e.g. about a quantity to be estimated) which is to be gradually reduced when more and more data are obtained. Although probabilities are subjective (and initially, prior uncertainty may be large), the process of updating is firmly based on probability theory and gives a unifying and universally applicable principle for statistical inference from data.

Likelihood inference. Based on analyzing the **likelihood function** $L(\theta; x)$ which is a function of the parameter when the observed data x are fixed. It is the same as the probability (or probability density) of data x, conditional on parameters $\pi(x \mid \theta)$, but interpreted as a function of θ for some fixed data. Maximum likelihood estimate $\hat{\theta}$ is the value that gives the highest probability (or probability density) for the data that were observed. Likelihood function is needed also for computing a posterior distribution of θ in Bayes theorem.

2 Introduction: $\propto$

Who was 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/.

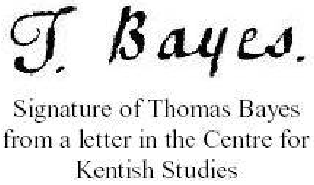

T. Bayes.
Signature of Thomas Bayes
from a letter in the Centre for
Kentish Studies

Figure 1: T. Bayes.

In the background section of bayesian history, the concept of bayesian probability was already briefly introduced as a degree of uncertainty. In our notations of probability, we could thus explicitly 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 on some state of information 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, nor that we would be free from the logical rules of probability calculus. 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 almost 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 information 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 not very meaningful without stating the conditional information and the underlying assumptions. Even a marginal distribution is still conditional on something. (Consider 2D density function $\pi(x, y \mid I)$. The marginal density of x is $\pi(x \mid I) = \int \pi(x, y \mid I)\mathbf{d}y$). There is no such thing as a completely unconditional probability.

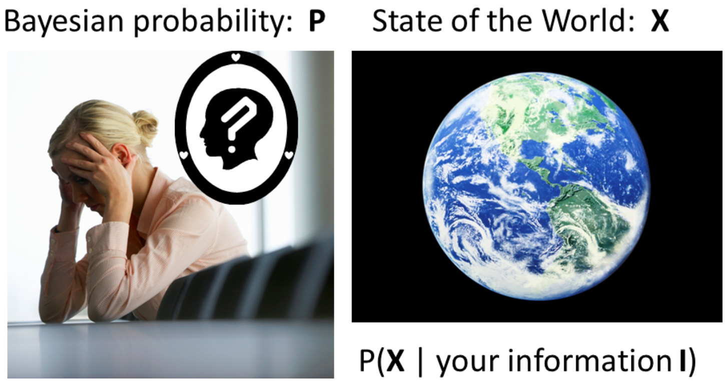

Bayesian probability: **P**
State of the World: **X**
P(**X** | your information **I**)

Figure 2: Probability is in the head of the observer.

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).

An example: in a bag you have M 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 this 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. After all, Dennis Lindley warned that the main danger with (bayesian) methods is that they are used too automatically. 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. Also, probability models are always 'wrong' because they are simplifications that can only include a limited number of features which we can handle.

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:

$$r \in \left\{ \frac{0}{M}, \frac{1}{M}, \frac{2}{M}, \frac{3}{M}, \dots, \frac{M}{M} \right\}.$$

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:

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

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

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

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" [**?**], 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

Recall the basic elements of probability theory. Let E and F denote two events. In general, these can also be logical propositions which are either true or false just like an event either 'occurs' or 'does not occur'. The probability measure P is a mapping from the space of events to the interval $[0, 1]$. Firstly, for any event E we have

$$0 \leq P(E) \leq 1.$$

This also gives the probability of the 'negation' or 'complement event' E^c = 'not E': $P(E^c) = 1 - P(E)$.

Secondly, if E is a sure event (or a proposition known to be true, according to our background knowledge), then we would have

$$P(E) = 1.$$

For example, with the bag of red and white balls, a sure event would be E = 'the ball is red or white'. Thirdly, for any two events E and F we have the joint probability which is *symmetric*

$$P(E \cap F) = P(E \mid F)P(F) = P(F \mid E)P(E) = P(F \cap E),$$

where $P(E \mid F)$ denotes the conditional probability of E given that F is true. For example, if E = 'the bag has i red balls' and F = 'the picked ball is red' then, according to the previously introduced (epistemic and aleatoric) probabilities:

$$P(E \cap F) = P(F \mid E)P(E) = \frac{i}{M} \times \frac{1}{M+1}.$$

In the special case, some events E and F are said to be independent if $P(E \cap F) = P(E)P(F)$ which also means that $P(E \mid F) = P(E)$ so that the probability of E is not influenced by knowing whether F is true or not (is occurred or not). The law of total probability states:

$$P(E) = P(E \cap F) + P(E \cap F^c) = P(E \mid F)P(F) + P(E \mid F^c)P(F^c),$$

which more generally, for mutually disjoint events F_i, is written

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E \mid F_i)P(F_i).$$

Also, more generally the joint probability is

$$P(E_1 \cap \ldots \cap E_n) = P(E_1 \mid E_2 \cap \ldots \cap E_n)P(E_2 \cap \ldots \cap E_n)$$

$$= P(E_1 \mid E_2 \cap \ldots \cap E_n)P(E_2 \mid E_3 \cap \ldots \cap E_n)P(E_3 \cap \ldots \cap E_n)$$

$$= P(E_1 \mid E_2 \cap \ldots \cap E_n)P(E_2 \mid E_3 \cap \ldots \cap E_n) \ldots P(E_{n-1} \mid E_n)P(E_n).$$

In the special case, where event E_i only depends on the event E_{i+1}, then this can be greatly simplified to

$$P(E_1 \mid E_2)P(E_2 \mid E_3) \ldots P(E_{n-1} \mid E_n)P(E_n) = \prod_{i=1}^{n-1} P(E_i \mid E_{i+1})P(E_n).$$

This technique is much exploited in complicated multivariate models where the joint distribution can still be handled by finding useful ways to break it down to some conditional probabilities. In the end of the line, there will be one or more probabilities that are not conditional on other events. In the above expression: $P(E_n)$. These would be called prior probabilities. For example, the above epistemic probability P('there are i red balls in the bag') $= 1/(M+1)$ is a probability which is not conditional on other things, except our initial background knowledge. Note that the product rule is symmetric and allows several different ways to write conditional probabilities.

But let us return to the question: so how exactly the probabilities are updated?

First, we must declare what our prior probability is - to have something to update. To continue the example above, this was already written 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 the joint probability $P(X, r)$, we have:

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

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 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 continuous quantities) 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 theorem is often written in a proportional form:

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

which means that the probability (or density) of r given X, i.e. $P(r \mid X)$, is equal to $P(X \mid r)P(r)$ multiplied by a constant. This 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. It is important to note that both of these are necessary elements for probabilistic inference and hence for all probabilistic learning. Also note that the Bayes formula is not an axiom in itself, but merely a logical consequence of the laws of probability where the product rule also provides Bayes formula.

N.B. Actually, (by Cox, advocated by Jaynes), Bayesian inference can be founded as extended logic, when some minimal requirements of consistency are met. The usual interpretation of events as subsets is not necessary then. For example, the general sum rule is often explained by using Venn diagrams where 'events' A and B are drawn as overlapping circles and where $P(A \cup B)$ represents the area under at least one of the circles. Hence, the overlapping area needs to be subtracted in the general formula. A special case is $P(A \cup B) = P(A) + P(B)$ when the sets are not overlapping, i.e. the corresponding events are said to be independent. However, we can also think of A and B as any logical propositions, e.g. $A =$ 'it rains tomorrow' and $B =$ 'it is cloudy tomorrow'. Then, instead of knowing exactly the truth value (zero/one) of these propositions, we have uncertainty P about them, and $P \in [0, 1]$. In such Bayesian theory, we aim to an objective formulation of priors, so that it might be used by a 'rational robot' rather than by a subjective individual with subjective prior information. However, the ultimate objectivity of priors remains a controversial issue.

For this particular example problem, we can now 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 and shake the bag for mixing. 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 is now:

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

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 after each draw). 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:

$$\begin{aligned} P(r \mid X_1, X_2) &= \frac{P(X_1, X_2 \mid r)P(r)}{P(X_1, X_2)} = \frac{P(X_2 \mid X_1, r)P(X_1 \mid r)P(r)}{P(X_1, X_2)} = \frac{P(X_2 \mid r)P(X_1 \mid r)P(r)}{P(X_1, X_2)} \\ &= \frac{P(X_2 \mid r)P(r \mid X_1)P(X_1)}{P(X_2 \mid X_1)P(X_1)} = \frac{P(X_2 \mid r)P(r \mid X_1)}{P(X_2 \mid X_1)} \propto P(X_2 \mid r)P(r \mid X_1), \end{aligned}$$

where the posterior after the 1st observation was:

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

It would also make no difference if both draws were already made by someone and then the results were only revealed to us later in reverse order.

What probability laws were used in this? Why were they valid?
In short:

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

This is an example which is often generalized to make Bayesian inference from a set of observations, $X_1, \ldots, X_n$, when these can be modeled as conditionally independent variables, given the parameter of interest r. Then we can conveniently write the probability of the *complete data set* (also known as 'full likelihood') as

$$P(X_1, \ldots, X_n \mid r) = \prod_{i=1}^{n} P(X_i \mid r)$$

With this, the posterior $P(r \mid X_1, \ldots, X_n)$ would be of the form

$$\propto P(r) \prod_{i=1}^{n} P(X_i \mid r).$$

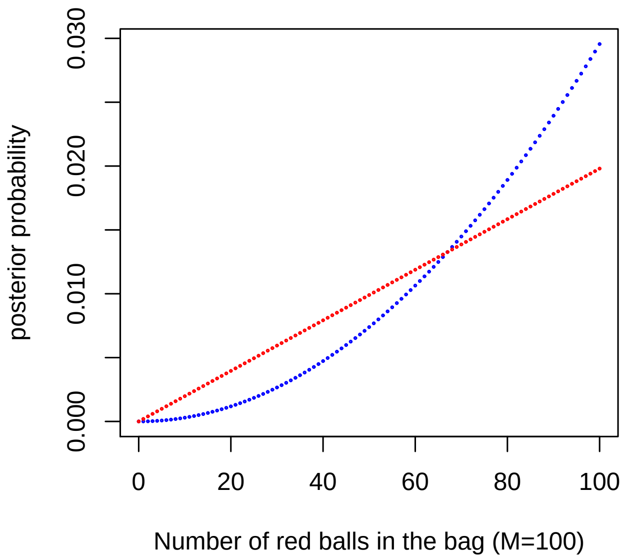


Figure 3: Posterior probabilities for the number of red balls among M in a bag, if one ball is drawn and it is red (red dots), and if two balls are drawn and both are red (blue dots).

2.3 Where do priors come from?

In the original work of Bayes, he considered (something like) 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 *uninformative* 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 [**?**] and *subjective* bayesian [**?**] 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 knowledge can be used as a basis for an informative prior, by conducting a careful elicitation process. The bayesian history shows many examples where the 'sample data' has not been the only source of important information for tackling a problem of inference.

Expert knowledge elicitation play an important role in areas where we clearly can make use of accumulated knowledge of experts, but where the amount of exact data is small. The data may be nonexisting perhaps because the problem is so urgent that there is no time to collect data, or it may be too expensive, or technologically or logistically unfeasible to get it. As one example, there is a recent document: Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment, European Food Safety Authority EFSA.
(http://www.efsa.europa.eu/en/consultationsclosed/call/130813.pdf).

2.3.1 Elicitation of prior probability via referencing standard random experiments

Consider eliciting the probability of an event A: "Eurozone will collapse in the next 5 years". This is a unique event for which there are no repeatable experiments. A financial expert may have some useful background knowledge about this and we could try to elicit the probability of the event based on that knowledge. The subjective probability can be compared with a more familiar 'random experiment'. For example, think of a grid of 100cm × 100cm with grid cells of size 1 cm^2. There are 10,000 cells in the grid. Tossing a pin 'randomly' on the grid implies that the probability for the pin landing on a specific cell is 1/10,000. Taking this as a reference, would the expert consider the probability $P(A)$ to be equal, or smaller, or larger than that? The question can be further refined by using other familiar examples of 'standard' random events for which the probabilities are easily calculated. An expert needs to think the event of interest in comparison to specified events in well-known standard play examples. After some discussion and thinking, a rough number could be elicited for $P(A)$, describing the expert's probability - actually $P(A \mid I)$, to be precise. (I denotes the background knowledge of this expert).

2.3.2 Elicitation of prior probability via gambling

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 monetary stake R) about some event A, the possible rewards are as follows: if event A happens, you will gain ωR, but if it does not happen, you'll lose R. 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

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

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 *prior 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. Some typical effects are, for example:

Representativeness heuristics (edustavuusharha)

This concerns elicitation of conditional probabilities such as 'What is the probability that a person of type A is of type B?' or 'What is the probability that a condition A leads to condition B in a system?'.

For example: 'Mr A is mean, pedant and introvert. Which of the following is his probable profession: B_1 salesman, B_2 journalist, B_3 doctor, B_4 accountant?'

Here we should quantify the conditional probability $P(B_i \mid A)$. Typical psychological error is to make a stereotypic association between A and B_i, based on perceived similarity. For example, by thinking that the personalities of accountants match this description. What is neglected is the proportionality of different professions in the population. The association is based on similarity, and similarity is symmetric. However, the conditional probabilities are generally not symmetric. The representativeness heuristic leads to violations of the Bayes formula, because it will assume $P(A \mid B) = P(B \mid A)$ instead of the correct formula. If A and B are perceived to be similar, then the answer we get will be a 'high probability', and if A and B are perceived to be very different, we typically get a 'low probability'. Hence, 'accountant' is typically given the highest probability $P(B_4 \mid A)$ than the other options. If B is not similar to A, the probability that B originates from A is judged to be low.

Availability heuristics (saavutettavuusharha)

This effect is due to thinking that familiar events occur more frequently than less familiar events. Likewise, events that we can easily imagine feel like more frequent than events that are hard to imagine. Also, events that have just recently happened, or events that received lots of publicity (like bad accidents), seem more probable compared to others. It is also difficult to assess correctly probabilities of very rare events, which hardly ever have been observed. Probabilities of place crash deaths can be overestimated compared to car crash deaths, if a recent plane crash is widely reported in media.

Anchoring (ankkurointiharha)

Experts can think of some special source of information, or it may be written in the questionnaire for them. It may happen that the expert then becomes anchored to this value. Even though the expert may try to shift his opinion away from this initial value during the elicitation, the shift may not be sufficient. The resulting answer tends to be anchored to the initial value. For example, when asking the unknown percentage: 'Is it less or over 10%?' compared to 'Is it less or over 80%?'. An arbitrary reference point is given in the question, and the answers tend to be closer to that.

Read more: Garthwaite PH, Kadane JB, O'Hagan A: Statistical methods for eliciting probability distributions. JASA (2005), Vol 100, (470), 680-700.

Also: Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR: Bayesian methods in health technology assessment: a review. Health Technology Assessment 2000, Vol 4, (38). chapter 3. (http://www.ncchta.org).

It can be laborious to avoid all psychological fallacies. Therefore, elicitation of informative prior probabilities is not necessarily easy. Moreover, probability of a complicated event is always more difficult to assess than the probability of its subevents. For example, the probability of failure of a machine could be assessed by eliciting the failure probabilities of its components, and describing how they are combined. Assessing each of the components should be easier to the experts than assessing the complete machine directly.

For events with continuous variables, *probability density functions* need to be elicited. This could be done with several techniques, e.g. eliciting the median and some percentiles. The number of questions for the expert(s) can then become large. However, in many problems we can rely on the data itself and the prior can be safely left vague. We then would like to have minimal information in the prior. In such case, 'objectivist' techniques with default uninformative priors can be sufficient (and free of elicitation problems!). However, the quest for a truly universal method for a uninformative prior may be the quest for the 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 all transformations of X, for example X^2, or $\log(X)$. It seems that we can only be uninformative in some aspects of the problem. To see how the transformation of variable affects the probability density, recall the following:

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 $\pi(g^{-1}(y)) \mid \frac{dx(y)}{dy} \mid$. (Note that the support of this new density is usually different from the original).

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).

Quote from the book of 'Bayesian Ideas and Data Analysis' [?]: **there is no *true* prior, only priors that adequately reflect uncertainty and information**.

...after all, the aim is to update the probabilities with new data. We don't intend to stick with the prior. But if that is our main, or only, information, we should be careful that it represents what we want it to represent. (As Lindley said: the danger is to use it in a too automatical fashion).

2.3.3 Combining expert opinions

For simplicity, assume that we take a simple parametric density function to represent the opinion of a single expert. This could be obtained by asking e.g. the median value from the expert, and then another value representing the upper 90% limit, or something similar. These can be used for solving the parameters for a simple density which then *approximates* the expert's opinion. As a result, we then have one density elicited from each expert. Two basic approaches of combining are the sum and the product of densities. For the sum we take

$$\pi(\theta) = \sum w_i \pi_i(\theta)$$

where each of the n experts has similar weight $w_i = 1/n$. The result is automatically a probability density, because it is a mixture of proper probability densities. Alternatively, for the product we take

$$\pi(\theta) = \prod \pi_i(\theta)^{w_i}/C$$

where we need to normalize the product because it does not lead to a proper density otherwise. A special case is obtained by setting $w_i = 1/n$, which corresponds to having the combination as the geometric mean of individual distributions. Whereas the sum will preserve all diverging opinions with equal weights, the product will emphasize the area of mutual certainty, so that whenever a single expert places a zero probability for some region, $\theta \in S$, this will also remain zero probability in the combined opinion, no matter how many other experts would think otherwise. This could work well if the experts are absolutely sure about 'impossible events'. But if the opinions of the experts are not overlapping, we have a contradiction.

2.4 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 limit 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 shuffled 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 which in a way describes the symmetry of our subjective probabilities (the observer's uncertainty about the events).

These other definitions share the underlying idea that probability is a purely objective 'true' property of the natural phenomenon we study - just like the mass of a physical object which has a specific value regardless of our state of knowledge. This is in contrast to the bayesian view that the probability is in the head of the observer, and thus must be changing when we get new information from observations.

2.5 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, e.g. $0, 1, 1, 0, 1$ can be written as

$$(1-r) \times r \times r \times (1-r) \times r = r^3(1-r)^2$$

which is the same as for another sequence of $1, 1, 1, 0, 0$. Generally:

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

where r is the proportion of red balls in the bag. It is apparent that only the sum of red (or white) balls matters for the probability of the sequence, not their order of appearance in the sequence. When making classical statistical inference about r, based on this conditional probability model of the X_is given r, the above expression is seen as a function of (the unknown) r, for a given data $X_1, \ldots, X_N$. The function is called *likelihood function*, and the sum is said to be *sufficient statistic*, (tyhjentävä tunnusluku) [1]. In classical statistics a sufficient statistic contains all the information in the sample needed to compute an estimate for a parameter. In this example: $\hat{r} = \sum_i^N X_i/N$. If we only observe the sum $Y = \sum X_i$, but not the exact sequence, then

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

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 for a bayesian statistical inference we must specify a prior *density* for this.

About notations: usually, probability is denoted as P whereas a probability density is written with a different symbol. In some cases we need to write a multivariate distribution where some of its variables are continuous and some discrete. To avoid switching symbols, below π is used loosely to denote all distributions, so that the reader should guess from the context if it means a probability mass function or probability density. Then, symbol P can be reserved to denote probabilities.

Analogous choice to the previously used discrete uniform distribution would be uniform probability density (as in the original example of reverend Bayes):

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

[1] $T(X)$ is sufficient for r if $P(X)$ can be written in the form $h(X)g(r, T(X))$

This uniform prior is a special case of a Beta(α, β)-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 density functions π for real values $r \in [0, 1]$ instead of probabilities P for $r \in 0, 1/N, \ldots, 1$.

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

For bayesian inference, like in likelihood based inference, the result is the same if we have observed the exact sequence of X_i's or if we just observe the sum Y. For a given Y, the posterior density is still of the same form, regardless of the sequence. From the functional form above - taken as a density for r, and knowing that this is indeed a probability density (the remaining terms, whatever they are, must be the normalizing constant) - the posterior density of r is *recognized* to be a Beta-density, with parameters $Y + \alpha$ and $N - Y + \beta$. (You can also calculate this exactly, without simply 'recognizing' the form of the density function by comparison to beta-density). The expected value of r from the posterior density is

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

which can also be written as a weighted average:

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

where $w = (\alpha + \beta)/(\alpha + \beta + N)$. The parameters of the prior can thus be chosen so that they represent some imaginary data Y_0, N_0, corresponding to $(\alpha, \beta) = (Y_0, N_0 - Y_0)$.

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. 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 of the well known standard distributions. Yet, it could still be computed by using numerical methods in the absence of analytical solution.

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 which somehow involve this parameter. When the prior was chosen as uniform density, the posterior density actually equals to the likelihood function which simply would be normalized to represent a proper probability density of r, for a given data Y. In figure (4) you see the joint distribution of Y, r based on uniform prior for r. From this you can get either $\pi(r \mid Y)$ or $\pi(Y \mid r)$.

In classical statistics, a popular estimate of the parameter is the maximum likelihood estimate, which is the parameter value that gives highest probability to the data. In the special case of uniform prior, the maximum likelihood estimate coincides with the value that has maximum posterior density. Note that for a bayesian, r is viewed as random (because unknown), but in non-bayesian statistics r would be thought as fixed.

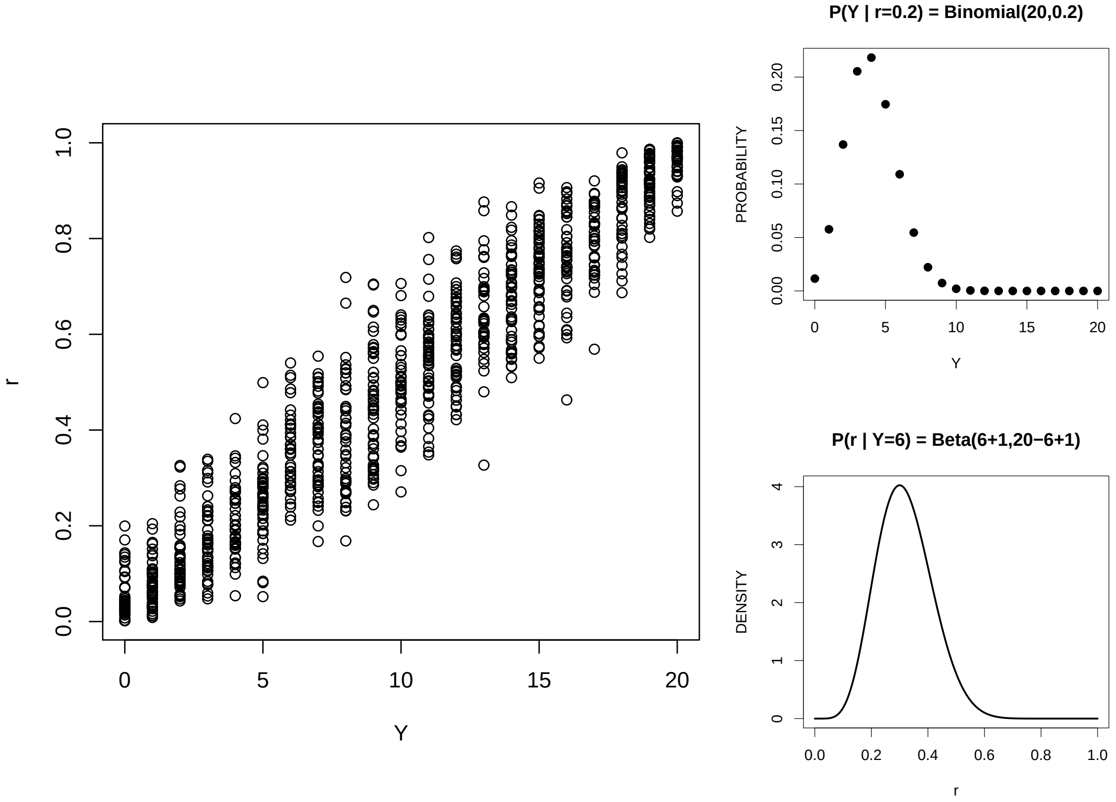


Figure 4: Simulated values from the joint distribution of $\pi(r, Y) = \pi(Y \mid r, N = 20)\pi(r)$ with uniform density $\pi(r) = U(0, 1)$. In R code: `r <- runif(1000)`, `Y <- rbinom(1000,20,r)`, `plot(Y,r)`. Note: for any fixed r we have Binomial$(20, r)$ for Y, i.e. $\pi(Y \mid r)$. For any fixed Y we have Beta$(Y + 1, 20 - Y + 1)$ for r, i.e. $\pi(r \mid Y)$. Examples of such conditional distributions on the right.

2.5.1 Informative priors for unknown proportion

Depending on what the prior information is, there can be different ways to formulate the prior as a density over $[0, 1]$ to reflect such prior information. the simplest case is to have a previous similar binomial experiment from which we have data Y_0, N_0 which can be directly translated to a Beta-density with parameters $Y_0, N_0 - Y_0$. Then we are assuming that the old sample and the forthcoming sample could be combined as one sample. Another source of information could be to ask from experts, or search from literature, what is the most plausible value of prevalence and call it m. Then, we should quantify also the width of the distribution by determining e.g. the standard deviation and call it s. If these can be reasonably quantified, we can then solve parameters for Beta-density:

$$\alpha = -m(m^2 - m + s^2)/(s^2) \ \ , \ \ \beta = (m^2 - m + s^2)(m - 1)/(s^2)$$

These follow from the expressions for mean (m) and variance (s^2) for beta distribution. It may be difficult to get an opinion about s, so we could work around it by formulating the problem differently. First start by asking for the most plausible value m. This could be taken to represent the mean as above, or perhaps more accurately the mode. By looking up the formula of the mode for Beta-distribution, we write

$$m = \frac{\alpha - 1}{\alpha + \beta - 2}$$

so that the Beta-prior is then Beta$((1 + (\beta - 2)m)/(1 - m), \beta)$. Next we determine what is a value for which the expert is 95% sure that the actual value is below. Call this value u. We then have prior probability $P(r < u) = 0.95$. By using the Beta-density which now only depends on β, we look for such value of β that we get $P(r < u \mid \beta) = 0.95$. Finally, we have solved the prior Beta(α, β). Solving the last step requires numerical techniques, e.g. using R to find percentiles. In a bayesian model, when computing posteriors can also require numerical techniques, one does not necessarily want to solve the prior numerically. Then, approximations based on normal distributions can be used to find analytical solution for the prior parameters.

In all cases, if the final prior density is Beta, we can also study what amount of prior data this would equal to. Note that Beta densities cannot represent bimodal or more complicated prior densities. However, these are rare in practice. But such prior might be obtained when combining the priors of a group of experts, as a group opinion. Then, the prior density could be expressed as a mixture of Beta-distributions.

Generally, a mixture prior distribution is a mixture of densities each specified by some parameter β_i:

$$\pi(\theta) = \sum_{i=1}^{k} \alpha_i \pi(\theta \mid \beta_i) = \sum_{i=1}^{k} \alpha_i \pi_i(\theta).$$

The weights α_i are the mixing weights of the component distributions ($\sum \alpha_i = 1$). Denote the model for data x as $\pi(x \mid \theta)$. The posterior distribution is then

$$\pi(\theta \mid x) = \frac{\sum_{i=1}^{k} \alpha_i \pi_i(\theta) \pi(x \mid \theta)}{\pi(x)}$$

which unfortunately is no longer recognized as a standard distribution, but this could be handled with numerical methods, e.g. in BUGS.

2.5.2 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 $r \sim$ 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

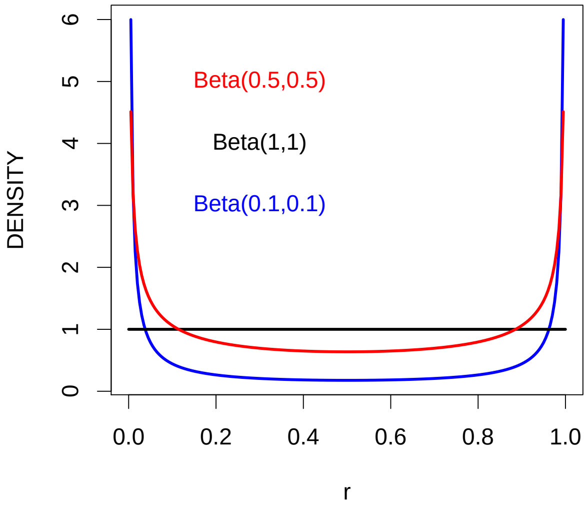


Figure 5: Beta-distributions which might be used as uninformative priors. "Beta(0,0)" is not a proper probability density. In practice, this Haldane's prior is often implemented approximately as Beta(0.001,0.001). The priors Beta(α,β) contain "prior data" which is comparable to sample size of $\alpha+\beta$.

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).

More about Jeffreys' prior

What does all this mean for Bayesian inference about some transformation of r? We do Bayesian inference by computing posterior distribution, so let's see this for the binomial example. 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:

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

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

(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!

However, Jeffreys' prior violates so called *likelihood principle* which states that whenever the likelihood function is (proportionally) the same, the inferences should be the same too. For example, the binomial model (for a sample result with fixed N) and the negative binomial model (for the number of samples N needed before fixed number of successes X is obtained) produce (proportionally) the same likelihood function for the success probability r. Therefore any differences in posterior must be due to different priors. In this example, Jeffreys' prior leads to two different prior distributions depending on which of the two models is used in the calculations. The difference is because in the first case, the expected value in the Fisher information is taken of derivatives of log-likelihood in which the random variable is X (given r, N), but in the second case the random variable is N (given r, X). Jeffreys' prior can also lead to improper prior distributions which cannot be normalized to proper probability distributions (which should integrate to one).

Note also that if the prior of r is $\mathrm{Beta}(\alpha, \beta)$, then the posterior will be $\mathrm{Beta}(X+\alpha, N-X+\beta)$ 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. Note that the fraction X/N is also 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: $\mathrm{argmax}_{r\in[0,1]} P(X \mid N, r)$.

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 BUGS, 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.

2.5.3 Unknown N

The usual application of binomial model $\mathrm{Bin}(N, r)$ 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 $0, 1, 2, \dots$ 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:

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

Now the posterior is:

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

and the normalizing constant is

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

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 or a text book. 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))
}
```

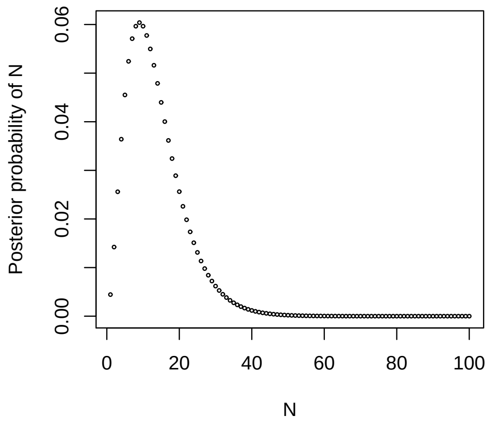
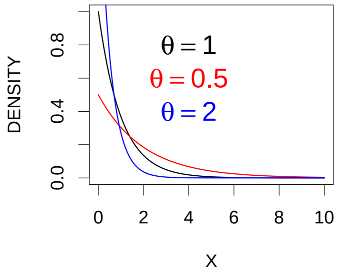


Figure 6: Posterior probability for N, given that $X = 1, r = 0.2$ with uniform prior over $0, 1, 2, \ldots, M = 100$.

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. Then we would not 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).

2.6 Exponential model

As an example of one-parameter inference from data which consist of continuous variables, consider e.g. modeling waiting times (times to next event), or concentrations, or other *positive valued* variables X. First, assume only a single observation $x \in \mathbb{R}^+$ is given as data, for which the conditional distribution is exponential:



$$\pi(x \mid \theta) = \theta \exp(-\theta x)$$

Note that in some versions of exponential distributions it is parameterized by θ as here, but sometimes by $\phi = 1/\theta$, so you need to check what parametrization is used. Using the parametrization here, the (conditional) mean of X is $E(X \mid \theta) = 1/\theta$. For this model a conjugate prior of θ is Gamma(α, β)-distribution, so that the posterior $\pi(\theta \mid x)$ can be solved exactly as Gamma$(\alpha + 1, \beta + x)$. Then, posterior mean is the mean of this gamma-distribution (check from the list of standard distributions):

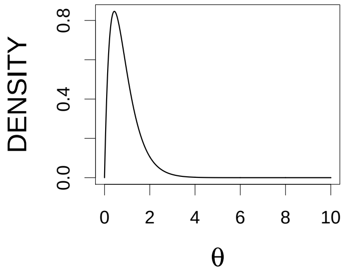
$$E(\theta \mid x, \alpha, \beta) = \frac{\alpha + 1}{\beta + x}$$

With a set of observations $x_1, \ldots, x_n$ (mean$= \bar{x} = \sum_{i=1}^n x_i/n$) we get

$$\pi(x_1, \ldots, x_n \mid \theta) = \theta^n \exp(-n\bar{x}\theta)$$

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 α prior observations $x_1^0, \ldots, x_\alpha^0$ for which the sum $\sum x_i^0$ equals to β. In the posterior distribution, these are updated by size of data n (number of observations) and the sum of observations $\sum x_i$, respectively.

This shows how the effect or the prior is comparable to the evidence from data and with $\alpha \approx 0, \beta \approx 0$ the posterior would be Gamma$(n, n\bar{x})$ with mean $1/\bar{x}$. This is comparable to a classical estimate based on the observed mean, when the theoretical mean is substituted by that, solving from: $\bar{x} \approx E(X \mid \theta) = 1/\theta$. With only two observations, with mean $\bar{x} = 2.3$, (e.g. average life time of two light bulbs) the posterior density of θ would be:



$$\pi(\theta \mid \bar{x} = 2.3) = \text{Gamma}(1, 2.3)$$

Note, the uninformative prior used here was "Gamma(0, 0)" which is no longer a proper probability density at all. Yet, the posterior is proper if $\bar{x} > 0$ which, in practice, is the case for actual observations.

2.7 Poisson model

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 a 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: recognized to be 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.

Informative prior would need to be elicited from some useful knowledge, by e.g. specifying the most probable value of λ and some upper limit (e.g. 95% percentile), and solving parameters of gamma-prior from this. An **uninformative prior** would obviously have 'small' values α, β. In the limit, these could be set to zero, so that the posterior then only depends on data. However, the prior is then not proper density and it would not be possible to get a prior predictive distribution. Also, e.g. with the single observation X, it could happen that $X = 0$, in which case the posterior would be Gamma$(0, 1)$ - not proper. (With improper priors, always check if posterior distribution is proper). Another approach would be to choose a uniform prior over a wide range U$(0, L)$ where $L \to \infty$ leads to improper prior. However, the posterior can still be solved and it is a proper distribution.

3 Approximating posterior density with normal density

With enough data, the posterior distributions eventually become very similar to normal distributions. If we are confident about this, posterior distributions could be approximated by finding out the posterior mean and variance, and then using normal distribution as an approximation:

$$\mathrm{N}\Big(E(\theta \mid X), V(\theta \mid X)\Big)$$

in place of the exact posterior density. Moreover, Posterior density can be approximated focusing on the mode as:

$$\pi(\theta \mid X) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1}),$$

where $\hat{\theta}$ is the posterior mode and $I(\theta)$ is so called *observed information*

$$I(\theta) = -\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(X \mid \theta).$$

3.1 Basis: Taylor series expansion

The approximation is based on Taylor series expansion of $\log \pi(\theta \mid X)$ centered at the posterior mode, $\hat{\theta}$. From Bayes theorem, the log-posterior is

$$\log \pi(\theta \mid X) = \log \pi(X \mid \theta) + \log \pi(\theta) + c$$

where the prior remains fixed so that the log-likelihood term will dominate when sample size increases. Thus we can approximate the log-posterior by using Taylor series expansion for the log-likelihood term. For a scalar valued θ the Taylor series expansion is

$$\log \pi(\theta \mid X) \approx \log \pi(X \mid \hat{\theta}) + \underbrace{[\frac{\mathbf{d}}{\mathbf{d}\theta} \log \pi(X \mid \theta)]_{\theta=\hat{\theta}}}_{=0} \frac{(\theta - \hat{\theta})}{1!} + [\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(X \mid \theta)]_{\theta=\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2!} + \cdots,$$

where the first derivative at posterior mode $\hat{\theta}$ is zero. When θ is near the mode, the higher order terms are small compared to the first terms. As a function of θ, the first term in the expression is constant whereas the 2nd order term is proportional to the logarithm of a normal density, which provides the approximation. For a vector valued θ, the Taylor series would be

$$\log \pi(\theta \mid X) \approx \log \pi(X \mid \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \Big[\frac{\mathbf{d}^2}{\mathbf{d}\theta^2} \log \pi(X \mid \theta)\Big]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots.$$

This normal approximation (modal approximation) can be a useful benchmark and it gives a quick approximation of the posterior density. For final results, more accurate computations are usually needed. Even so, the first rough estimates can be obtained from the approximation, if only as realistic starting values for more complicated calculations. Also, this makes a close connection between likelihood inference and bayesian inference: with large data the results become identical and maximum likelihood estimate would equal posterior mode. Note that here sample size $n \to \infty$, but number of parameters does not change.

4 Monte Carlo method

A posterior probability distribution $\pi(\theta \mid X)$ captures all the information we have about the unknown parameter θ, after selecting the prior distribution $\pi(\theta)$ and the conditional distribution of data $\pi(X \mid \theta)$, and after observing what the data X was. *All results follow from the posterior distribution.* Usually, for the results we need to *integrate* over the posterior density. This can be simple to obtain from tabulated values of cumulative probability distribution (or from statistical functions on computer), when the posterior is among well known standard distributions - *as is the case with conjugate distributions.* Generally, this convenience is not available. Monte Carlo method is based on approximating the distribution by a sufficiently large sample from it. All we need to do is to find a way to draw random samples from the distribution. Our target distribution will naturally be the posterior distribution.

Assume we have been able to draw a random sample $\theta_1, \ldots, \theta_n$ from posterior distribution $\pi(\theta \mid X)$, generated by some software. Then, when $n \to \infty$

- $E(\theta \mid X) \approx \frac{1}{n}\sum_{i=1}^{n}\theta_i$.

Want to calculate posterior mean $E(\theta \mid X)$?

⇒ compute the sample mean $\frac{1}{n}\sum_{i=1}^{n}\theta_i$.

- $E(g(\theta) \mid X) \approx \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$.

Want to calculate posterior mean of a transformation: $E(g(\theta) \mid X)$?

⇒ compute the sample mean $\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$.

- $P(\theta \in S \mid X) = E(1_{\{\theta \in S\}}(\theta)) \approx \frac{1}{n}\sum_{i=1}^{n} 1_{\{\theta \in S\}}(\theta)$

Want to calculate posterior probability $P(\theta \in S \mid X)$?

⇒ compute the sample mean of the indicator variable $\frac{1}{n}\sum_{i=1}^{n} 1_{\{\theta \in S\}}(\theta)$.

Many standard distributions are available in statistical software as R, and we could also use those in WinBUGS/OpenBUGS. These could be used for simulating samples from given distributions, e.g. from posterior distribution that was found to be among standard distributions. This is always possible when using conjugate priors. Just plug in the appropriate parameter values for the standard distribution. With direct Monte Carlo method you can simulate any quantities of interest, and get approximate posterior means, medians, modes, and CIs.

For the conjugate models, you need to learn about analytical solutions of posterior distributions. Example for binomial and exponential models was shown above. Other conjugate priors exist e.g. for multinomial, poisson, gamma, and normal models.

So, for example with binomial model $\text{Bin}(N, r)$ with observed data x 'successes' out of N trials, and estimating the unknown r:

(1) if prior is $r \sim \text{Beta}(\alpha, \beta)$
(2) if data model is $x \sim \text{Binomial}(n, r)$
(3) then posterior is $r \sim \text{Beta}(x + \alpha, n - x + \beta)$
(4) use any software available to sample from this posterior distribution, with these given parameters.
(5) monitor any quantity of interest, $g(r)$, and see the Monte Carlo sample histogram.

Price: for each problem, you need to solve the posterior probability density. This is unfortunately very restrictive, and the solutions for any more realistic problems become increasingly challenging. However, Monte Carlo method is far more general and versatile tool. Some examples are briefly explained below.

4.1 Some versions of Monte Carlo

For some problems, you may try to apply these with e.g. R. General Monte Carlo methods give you computational freedom from the available set of standard distributions in a software. Basically, you would only need a random number generator to produce random samples from Uniform(0,1), to generate random samples from any distribution you need.

4.1.1 Inverting cumulative distribution function

If the target density, say $\pi(\theta)$ has a cumulative distribution function

$$F(\theta') = P(\theta < \theta') = \int_{-\infty}^{\infty} \pi(\theta) d\theta = u(\theta')$$

which can be inverted for solving $\theta' = F^{-1}(u)$, we can then generate u ~Uniform(0,1) and calculate $\theta' = F^{-1}(u)$. These are distributed as the target density. Multivariate distributions would need to be handled in a sequence of univariate distributions, taking advantage of the product rule: $\pi(x, y, z) = \pi(x \mid y, z)\pi(y \mid z)\pi(z)$, etc.

4.1.2 Rejection sampling

Assume target density is $\pi(\theta)$. Choose instrumental density $g(\theta)$, and some constant M so that $\pi(\theta)/(Mg(\theta)) \leq 1$. This instrumental density should be easy to sample and it should have the same support as $\pi(\theta)$. Then: (1) generate random value θ from density g. (2) Accept this with probability $\pi(\theta)/(Mg(\theta))$. Repeat until enough large sample was obtained.

4.1.3 Importance sampling

Again, target density is some $\pi(\theta)$. Choose instrumental density $g(\theta)$ that is easy to sample. After you have obtained a sample $\theta_1, \ldots, \theta_K$ from g, use a weighted sample for final results, with weights $\pi(\theta_k)/g(\theta_k)$. For example, to calculate mean $E_\pi(\theta)$, use

$$E_\pi(\theta) = \int \theta \pi(\theta) d\theta = \int [\theta \frac{\pi(\theta)}{g(\theta)}] g(\theta) d\theta = E_g(\theta \frac{\pi(\theta)}{g(\theta)}) \approx \frac{1}{K} \sum_{k=1}^{K} \theta_k \frac{\pi(\theta_k)}{g(\theta_k)}$$

4.1.4 Approximate Bayesian Computation (ABC)

This 'brute force' method could even be applied to compute posterior distributions when it is not possible to write the likelihood function in analytically closed form. For example, the likelihood may result from a separate simulation model that can only simulate observable values X conditionally on some underlying parameter(s) θ. In its simplest form, the method is actually exact when X is a discrete variable. It is based on the following rejection method applied for the joint distribution of (θ, X): (1) Sample θ from prior distribution $\pi(\theta)$, then (2) sample X from the conditional distribution $\pi(X \mid \theta)$, and (3) repeat until a large sample is obtained. Finally, (4) select those samples where the X matched the observed value X_{obs} and analyze the corresponding θ in that sample.

In this way, we have effectively simulated from the conditional distribution (posterior) $\pi(\theta \mid X = X_{\text{obs}})$. However, the method can waste a lot of simulations because most of the time it will be the case that $X \neq X_{\text{obs}}$. The resulting subsample should still be large enough to make reliable approximation for the posterior distribution. Moreover, if X is not discrete, the sampled X will practically never match the observed value. In that case, some error tolerance is needed so that we accept samples where $X \in [X_{\text{obs}} - \epsilon, X_{\text{obs}} + \epsilon]$. This is why it is Approximate Bayesian Computation, ABC.

4.1.5 Markov chain Monte Carlo sampling (MCMC)

A much more general alternative to direct Monte Carlo sampling is Markov Chain Monte Carlo (MCMC). It is based on iterative approach where we start with some *initial values* θ_0, then sample next value conditional on that: $f(\theta_1 \mid \theta_0)$, and continue sampling from $f(\theta_i \mid \theta_{i-1})$, $i = 1, 2, 3, \ldots$, where i denotes the ith sample of the parameter, the ith iteration step. The *transition distribution* f of the Markov chain is chosen so that, in the limit, the Markov chain converges to a stationary distribution, and this stationary distribution is the same as the distribution we want to draw samples from. A target distribution in Bayesian applications is naturally the posterior distribution. Hence, for each posterior distribution we are interested in, it is possible to construct a Markov chain sampler that will eventually draw random samples from it. There can be many different Markov chain samplers that will lead to the same target distribution but some are more efficient than others. These are implemented in BUGS.

Note: the consequent samples are no longer independent and identically distributed as they are with direct Monte Carlo sampling where the next sampled value did not depend on the previously sampled value. Nevertheless, with sufficiently large number of iterations, we get approximately correct sample. (But one usually needs to discard a burn-in period).

4.1.6 Gibbs sampling

A special case of MCMC sampling is Gibbs sampling. This is sometimes called 'alternating' (vuorotteleva) sampling, because there we sample one of the unknown parameters at a time, from a distribution that is conditional on the current values of all other parameters and data. This is based on solving the 'full conditionals' from the joint distribution. For example with 2D-parameters θ_1, θ_2, we have the joint distribution as $\pi(\theta_1, \theta_2 \mid X)$. We can look at this in the 'proportional to' form, given by Bayes formula: $\pi(X \mid \theta_1, \theta_2)\pi(\theta_1, \theta_2)$. When you write down what these functions are in this product, look for an expression that is proportional to a familiar distribution for θ_1, given θ_2 and X. This should appear when you re-write the joint posterior probability of θ_1, θ_2 in the form $\pi(\theta_1 \mid \theta_2, X)\pi(\theta_2 \mid X)$. Then do the same to find a distribution for θ_2, given θ_1 and X. This should appear when re-writing the joint posterior as $\pi(\theta_2 \mid \theta_1, X)\pi(\theta_1 \mid X)$. If these two conditional distributions can be identified from the expression, you can use them to sample $\theta_1 \sim \pi(\theta_1 \mid \theta_2, X)$ and $\theta_2 \sim \pi(\theta_2 \mid \theta_1, X)$ sequentially: first θ_1, then θ_2, then θ_1, then θ_2...

For example: think again the binomial model, but let both X and p be unknown, and set $N = 20$ fixed. We try to compute the joint distribution of X and p, given $N = 20$, describing jointly the prior distribution of p and prior predictive for X. These are not independent because $\pi(X, p) \neq \pi(X)\pi(p)$, see also the figure shown in the section for binomial model. Here, we have N as fixed number in all calculations, so for simplicity it might be dropped from all notations; it is an underlying assumption here. The joint distribution of p and X is (according to product rule) written in two possible ways

$$\pi(p, X \mid N) = \quad \underbrace{\pi(p \mid X, N)}_{\text{Beta}(X+1,N-X+1)} \quad \underbrace{\pi(X \mid N)}_{*} = \underbrace{\pi(X \mid p, N)}_{\text{Bin}(N,p)} \underbrace{\pi(p \mid N)}_{**}$$

When you look at these two alternative expressions you can find that (1) if keeping X fixed in the first expression you have a distribution function for p proportional to Beta$(X+1, N-X+1)$. Likewise, (2) if keeping p fixed in the second expression, you have a distribution function for X, proportional to Binomial(N, p). These are the full conditional distributions for p and X. MCMC sampler is given in R below. This joint distribution is the same as earlier when introducing binomial model and when simulating the joint distribution from the prior. There, the prior of p was uniform U$(0, 1)$ which is also the **marginal distribution of p**. This leads to the **marginal distribution of** X to be discrete uniform $\pi(X) = 1/(N + 1)$. From the joint distribution, these two marginal distributions can be written as ('**' and '*' above) $\pi(p) = \sum_X \pi(X, p)$ and $\pi(X) = \int_0^1 \pi(X, p)\mathbf{d}p$.

```
 n<-20; p <- numeric(); x<- numeric()

 p[1] <- 0.5  # initial values
 x[1] <- 10
 for(i in 2:1000){
 p[i] <- rbeta(1,x[i-1]+1,n-x[i-1]+1)
 x[i] <- rbinom(1,n,p[i])
 }
 plot(x,p)
```

From this we could actually compute also e.g. the posterior probability $\pi(p \mid X = 7, N = 20)$ by collecting all those iterations where we had $X = 7$. This is intuitive because we produced the joint distribution of X, p, and then we can conditionalize on a specific value of X and see the distribution of p at that value of X. In principle, all posterior distributions might be simulated in this way: by sampling the joint distribution of the parameter and all possible data sets, and then collect those samples where the data coincides with our actually observed data X. This would often be very inefficient sampler, though. In most cases in practice, it is best to keep your data fixed to what they were, and only sample the quantities that are uncertain. (Posterior distribution is defined for those).

```
# R code for drawing the Gibbs sample:
 n<-20; p <- numeric(); x<- numeric()
p[1] <- 0.5; x[1] <- 5  # initial values
plot(x[1],p[1],xlim=c(0,20),ylim=c(0,1),
ylab="p",xlab="x",main="Gibbs sampling")
for(i in 2:250){
p[i] <- rbeta(1,x[i-1]+1,n-x[i-1]+1)
```

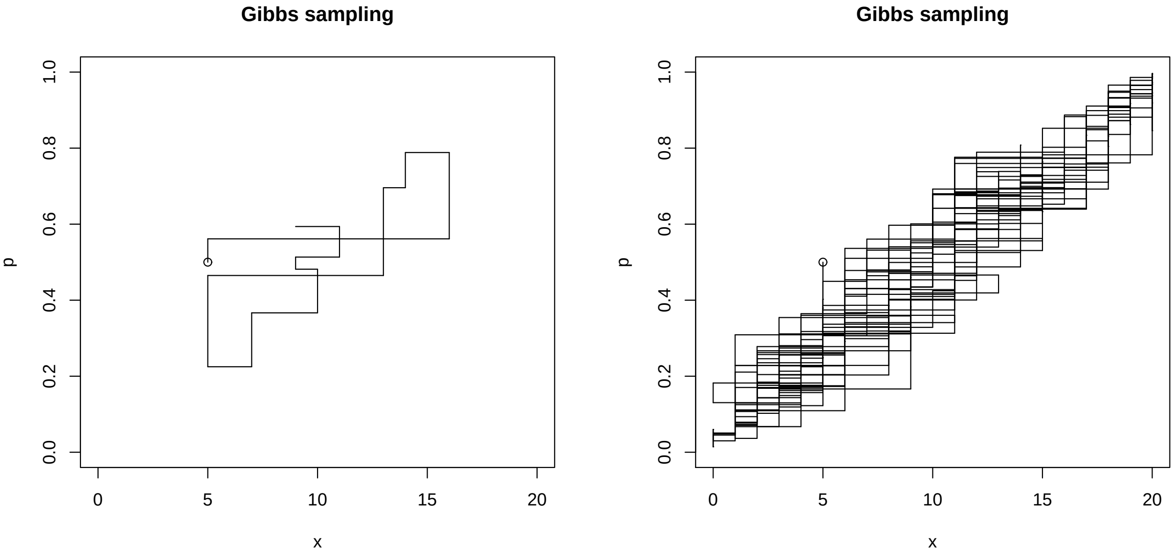


Figure 7: Sample path of Gibbs sampling with $\pi(X, p \mid N)$

```
points(c(x[i-1],x[i-1]),c(p[i-1],p[i]),'l')
x[i] <- rbinom(1,n,p[i])
points(c(x[i-1],x[i]),c(p[i],p[i]),'l')
}
```

Example: Gibbs sampling for a simple 2D normal density:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \Big)$$

Recall that the 2D normal density function (mean zero, unit variance) is

$$\pi(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \Big).$$

It can be written in the form $\pi(x \mid y)\pi(y)$ or $\pi(y \mid x)\pi(x)$ since the marginal, and conditional, densities can be solved from the joint density:

$$\pi(x) = \int_{-\infty}^{\infty} \pi(x, y)\mathbf{d}y = N(0, 1)$$

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x, y)\mathbf{d}x = N(0, 1)$$

and

$$\pi(y \mid x) = \frac{\pi(x, y)}{\pi(x)}$$

$$= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \Big)}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(\rho x - y)^2) \quad = \quad \mathrm{N}(\rho x, 1-\rho^2)$$

and similarly: $\pi(x \mid y) = \mathrm{N}(\rho y, 1-\rho^2)$.

4.1.7 Metropolis-Hastings algorithm

More general sampling algorithm is Metropolis-Hastings method, where within each iteration, the next sampled value is *proposed* from a proposal distribution. Then it is either rejected or accepted by a probability given by the Metropolis-Hastings ratio

$$R = \min\Big(\frac{\pi(\theta^* \mid \text{data})Q(\theta_{i-1} \mid \theta^*)}{\pi(\theta_{i-1} \mid \text{data})Q(\theta^* \mid \theta_{i-1})}, 1\Big),$$

where Q is the proposal distribution, x^* is the proposed new value and x_{i-1} is the current value from the previous iteration step. If θ^* is rejected, the previously sampled value θ_{i-1} is taken as the next value. The important innovation in this MH-ratio is that - again - the normalizing constant of the posterior is not needed. It cancels out from the ratio. **Therefore, we only need to be able to calculate the posterior probabilities in the 'proportional to' form**. Gibbs sampler is a special case of Metropolis-Hastings, where the acceptance probability becomes one, because the full conditional distributions are used.

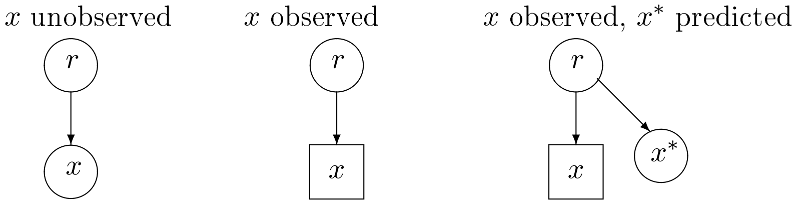
Note: with all MCMC methods, the initial value can be chosen anywhere in the parameter space. The MCMC method is based on a constructed Markov chain **that will only converge** to the correct target distribution. But it can be slow to converge. Generally, due to arbitrary initial values that may be far off the target distribution, it is common practice to run a *burn in* period. Only the sample collected after that will be used. The length of the burn in period needs to be judged separately in every case. There is no guarantee that a specific number of iterations is always enough. There are diagnostic tests for bad convergence (one is available in BUGS), but these can only detect some possible problems. They cannot prove that the result is correct. Also, large autocorrelation between consecutive MCMC iterations is indicative of slowly mixing MCMC chain. In Metropolis-Hastings algorithm, acceptance rate of the proposed values is an indicator of possible problems: acceptance rate should not be near zero, nor near one. If it is zero, none of the proposed values are accepted and the sampler is stuck with the current value. If it is nearly one, every proposal is accepted and this can happen e.g. if the proposal distribution is very narrowly centered around the previous value so that the proposed values are nearly as good as the previous, and the chain is not moving fast enough to more remote areas of parameter space. It would make a move almost every time, but too small moves to cover the whole parameter space efficiently. A Gibbs sampler can also run into problems if the posterior distribution covers the parameter space in such way that it is very difficult to move around within the space of positive posterior density by sampling one coordinate at a time. For example, a 2D posterior distribution that is mostly diagonally aligned, or multimodal (in which case the sampler might produce a sample around one peak of the distribution only, never entering the other peaks).

5 Graphical modeling and BUGS

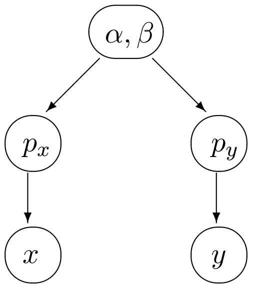
Bayesian models can be shown graphically as Directed Acyclic Graphs (DAG). Every Bayesian model defines a joint probability model, e.g. $\pi(X,\theta)$ which can also consist of several variables

$\pi(X, Y, Z, \theta, \lambda, \phi)$. Some of the variables represent observable data e.g. X, Y, Z and some represent parameters or unobservable quantities to be estimated e.g. θ, λ, ϕ. As always, the task is to compute π(parameters | data) based on Bayes theorem. The DAG is also how models can be thought and constructed in different versions of BUGS software. The joint posterior density is always fully specified when all the Child-Parent conditional distributions are defined in a DAG, including the prior distributions and founder nodes which have no parents. Therefore, WinBUGS is a **declarative** language, as opposed to **procedural** programming languages. (**This is important to remember**). The order in which code lines are written in the file does not matter.

Some Directed Acyclic Graphs (DAG) which could represent e.g. the Binomial model:



Also, there could be two binomial models each with own data (x, y) for estimating two population prevalences (p_x, p_y) with common prior distribution given by parameters α, β which also could be estimated (to describe variation between populations).



5.1 WinBUGS/OpenBUGS

WinBUGS = **B**ayesian inference **U**sing **G**ibbs **S**ampling

WinBUGS is a computer program designed for Monte Carlo simulation of posterior distributions, by using Markov chain Monte Carlo methods (MCMC). Its interface is fairly easy to use, and it can also be called from programs such as R. WinBUGS is free and can be found on the website:

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

OpenBUGS program is the version of BUGS that is further developed, found at: http://www.openbugs.net/

Given a likelihood and prior distribution, the aim of both WinBUGS and OpenBUGS is to sample model parameters (and other unknown quantities) from their posterior distribution. After the parameters have been sampled for many iterations, parameter estimates can be obtained and inferences can be made by using the Monte Carlo generated sample as approximation of the posterior distribution.

For a given application project, three files are used:

1. A program file containing the model specification.
2. A data file containing the data in a specific (slightly strange) format.
3. A file containing starting values ('initials') for model parameters (optional).

File 3 is optional because WinBUGS/OpenBUGS can generate its own starting values. There is no guarantee that the generated starting values are good starting values, though. All three files can be written in one if manually choosing the model code, data and inits by clicking-and-pointing.

5.1.1 Quick guide for getting started

1. Open an existing model code or an empty window from `file -> open`, or `file -> new`.

2. In case of empty window, write the model code, for example the binomial:

```
model{
x ~ dbin(p,N)
p ~ dbeta(a,b)
}
```

3. Check the syntax: click `model -> specification...` from menu. This opens a new window, where you should click `check model`. If the code was syntactically correct, you should get a message about it in the bottom of the main BUGS window.

4. Then you need to load data and other constants that need to be specified. In this case e.g.

```
 list(x=3,N=8,a=1,b=1)
```

This data listing could be in its own window (saved in a separate file), or it could be written below the model code in the same window. If in its own window, you can click that to make it 'active', then click `load data` button. If written in the same window below code lines, then select the word 'list' by using mouse, then click `load data`. Otherwise, BUGS tries to read the model as data which leads to error. It should be reading in the data list. If all goes well, you find the message 'data loaded' in the bottom of the main window.

5. Click `compile`. If all goes well, you get the message 'model compiled' in the bottom of the main window.

6. Generate of choose initial values for the MCMC sampler. You can let BUGS generate random initial values by clicking ĝen inits. It generates them from priors. BUT: if your priors were 'uninformative' and therefore wide, the values could be far in the tails of the posterior distribution. This can lead to numerical errors when BUGS tries to continue from that. To set your own initial values, you need to write them in the same manner as the data, possibly in a different file too. For example:

```
list(p=0.4)
```

If all goes well, you get message 'model initialized' in the bottom of the main window.

7. At this step, you should have checked the syntax, loaded data, compiled the model, and loaded initial values. You no longer need 'Specification Tool' and you can close it.

8. Time to run a 'burn-in period' (usually if the initial values are poor, it takes some iterations before the generated sample represents properly the target distribution). Click `Model -> Update` from the menu to get 'Update Tool'-window. Click `update`. It runs 1000 first iterations. Click more if you like. Let this window stay open, and click `Inference -> Samples` from the menu. It opens 'Sample Monitor Tool' window. Write the name of the parameter you want to sample in the dialog box after 'node'. In this case p. Then click `set`. Move back to Update Tool and run again more samples. After that, move back to Sample Monitor Tool. Select from the 'node' dialog box the name of the parameter you want. The names should appear by clicking the small arrow in the box. Now most of the buttons of the Sample Monitor Tool should become available and you can click for example `density` to see the density plot for this parameter. This is the posterior distribution for it (or marginal posterior, if multi-parameter model). Or click `stats`.

Good luck!

5.1.2 Type of nodes in BUGS

Nodes in the graph are of three types.

1. **Constants** are fixed by the design of the study: they are always founder nodes (i.e. do not have parents), and are (often) denoted as rectangles in the graph. They must be specified in a data file.

2. **Stochastic nodes** are variables that are given a conditional distribution, and are denoted as ellipses in the graph; they may be parents or children (or both). Stochastic nodes may be observed in which case they are data, or may be unobserved and hence be parameters, which may be unknown quantities underlying a model, observations on an individual case that are unobserved say due to censoring, or simply missing data. They are coded with $\sim$ before the specified conditional distribution. (e.g. `x` $\sim$ `dnorm(mu,tau)`).

3. **Deterministic nodes** are logical functions of other nodes. Note that they are not allowed to be given data values. Data should always be given to a stochastic node as an observed value for that. Therefore, we cannot specify a structure where e.g. $X \sim \mathrm{N}(\mu, \tau)$ and $Y \sim \mathrm{N}(\mu, \tau)$ and $Z \leftarrow (X+Y)/2$, and then assign Z some observed data value. This would lead to an error message indicating multiple definition of Z. Instead, we need to define Z as a stochastic node $Z \sim \mathrm{N}(\mu, 2\tau)$, to be able to assign data value for it. This has been causing some headache when trying to define distributions implicitly. It is not possible. A conditional distribution needs to be chosen for every stochastic node in the graph. Deterministic nodes are coded with $\leftarrow$. (e.g. `x <- log(z*z)/2 + u`).

Stochastic quantities can be specified as data by giving them observed values in a data file, in which values for constants are also given.

5.1.3 Doodle BUGS: create models graphically

Models can be defined in either by writing the corresponding BUGS code, or by drawing the graphical model description, DAG, using doodle-BUGS. Once the 'doodle' is drawn, which defines the DAG, the corresponding BUGS code is automatically generated from it. But the opposite is not possible: a picture of a DAG cannot be generated from BUGS code, you need to draw it manually with something else. But the BUGS language is much more versatile than doodle-BUGS, so it is best to learn to write BUGS codes, and do drawing of the corresponding DAGs with some other tools.

As an illustration, we could construct a simple model for medical diagnosis, using doodle-BUGS (in OpenBUGS).

Assume a patient can have one or more of the following symptoms: 'sore throat', 'fever', 'white spots in throat'. The possible causes for these symptoms are assumed to be 'cold' and 'angina' or both, or neither of the two. Note that all of these are binary variables that can only take values '0' and '1'. The following probability model for the symptoms is assumed, conditional on the causes:

		cold,angina		
	(0,0)	(0,1)	(1,0)	(1,1)
sore_throat	0.05	0.8	0.1	0.95
fever	0.01	0.7	0.1	0.85
white_spots	0	1	0	1

This table specifies the probabilities for all observable symptoms (=data), given the unknown causes. For Bayesian inference, prior probabilities for the causes are needed. Assuming no prior knowledge in favor of any cause, the prior probabilities could be

cold	angina
0.5	0.5

These numbers were arbitrary, but in more realistic example they would be parameters that could be estimated from prevalence data. If we just assume the numbers, then the only unknown quantities (after observing the symptoms) are the binary variables representing the causes. We can use Bayes theorem to compute the posterior probability for each combination of possible causes, given any observed set of symptoms. (Try this with paper and pencil). In doodle-BUGS, we construct a DAG corresponding to the above tables. Click *Doodle* from menu, then *New* to open a new window. Click inside the window to create a node. This could be named as 'cold' and its density chosen as *dbern* (for Bernoulli distribution). Setting *proportion* to 0.5 specifies the prior probability. This corresponds to setting the one parameter of Bernoulli distribution (which is 'proportion') to 0.5. Then create similarly a node for 'angina'. Next create Bernoulli-variables likewise for 'sore_throat', 'fever' and 'white_spots'. Finally, these observables need conditional probabilities depending on each cause. In some software for *Bayesian networks* this might be created by drawing arrows from the 'cause' nodes to the 'effect' nodes, and giving the probability tables, but in BUGS we need to specify explicitly what the parameters are for each defined distribution. Since the distributions here were Bernoulli-distributions, each of them requires one parameter ('proportion' in doodle syntax). The parameter for Bernoulli-distributed (binary) variable sore_throat can be written down as 'p1':

```
p1<- cold*(1-angina)*0.1+(1-cold)*angina*0.8+cold*angina*0.95+(1-cold)*(1-angina)*0.05
```

P1 needs to be created explicitly as a node in the graph. This makes a deterministic node, since the value of 'p1' is deterministically calculated from the values of 'cold' and 'angina'. To create the node, click within the window to create node. Then select its type *deterministic* and write the expression shown after the arrow sign above as *value* for the node. Then create a deterministic arrow from nodes 'angina' and 'cold' to node 'p1' by clicking 'p1' and pressing ctrl while clicking 'angina'. (Similarly for creating arrow from 'cold' to 'p1'). As a result, there should be node 'p1' which has arrows coming from 'angina' and 'cold', which determine the value for it. Likewise, create nodes 'p2' and 'p3' for parameters of the other two Bernoulli variables 'fever' and 'white_spots', according to the above table. After all these parameters are specified as deterministic nodes, they should be connected to the stochastic nodes 'sore_throat', 'fever' and 'white_spots' by arrows. These nodes need to be first created, their types selected as *stochastic* and densities chosen as *dbern* and proportions defined as 'p1' or 'p2' or 'p3', accordingly. Effectively, the graphical construction defines a complete DAG of a Bayesian model. In BUGS language, the model is written as

```
model{
angina ~ dbern(0.5)
cold ~ dbern(0.5)
fever ~ dbern(p2)
sore_throat ~ dbern(p1)
white_spots ~ dbern(p3)
p1<-cold*(1-angina)*0.1+(1-cold)*angina*0.8+cold*angina*0.95+(1-cold)*(1-angina)*0.05
p2<-cold*(1-angina)*0.1+(1-cold)*angina*0.7+cold*angina*0.85+(1-cold)*(1-angina)*0.01
p3<-angina
}
```

Click *Model*, then *Pretty print* to see the BUGS code. To calculate posterior probabilities for different causes, specify the symptoms as data and run the model. Each of the symptoms can take any of the values 0, 1 or 'NA'. The data specification could be e.g.

```
list(sore_throat=1,fever=0,white_spots=NA)
```

Explore how different symptoms will change the diagnosis. You could also try to change the underlying conditional probabilities to see their influence on diagnosis. For Bayesian networks, see e.g. the book (which contains basically the above example):

F V Jensen: Bayesian Networks and Decision Graphs. Springer-Verlag. 2001.

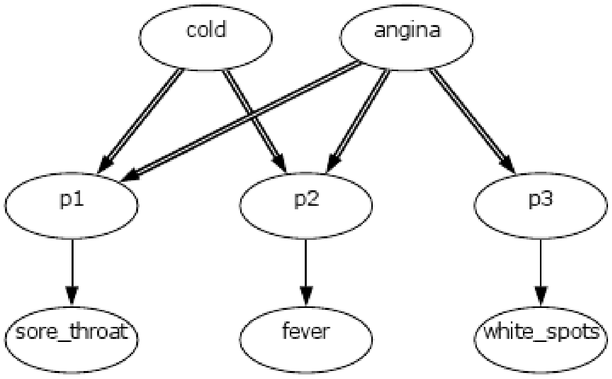


Figure 8: DAG of the diagnosis model from Doodle-BUGS

6 Summarizing the posterior distribution

Often, the posterior distribution is presented graphically, possibly with the analytical mathematical expression of the density (if it could be solved) or as given in the Bayes theorem (prior times likelihood). A graphical display is very informative, but sometimes we need simple summaries. In non-bayesian statistics we often deal with 'estimators', which are functions of the data and therefore 'random', conditionally to some hypothetical parameter values. The calculated values of such estimators are then taken as estimates of the (nonrandom) unknown parameters. But in Bayesian statistics, the parameters are random (i.e. uncertain), described by the posterior distribution. Therefore, the usual ways to summarize a probability distribution are directly applicable. Typically central values: mean, mode, or median. Also the width of the distribution is important, since it represents how uncertain we are. Therefore, variance, or standard deviation can be reported. For standard densities, these are easily calculated. For less common distributions, they may be easily available numerically in various software. Also, percentiles of the distribution can be informative. Very often, *credible intervals* (or regions for higher dimensional parameters) are reported.

The binomial model of red balls led to the posterior of the unknown proportion in the form of a beta-density. Since the expected value of a Beta(α,β)-density is $\alpha/(\alpha+\beta)$, and the mode is $(\alpha-1)/(\alpha+\beta-2)$ it is easy to summarize the posterior density by reporting the mean or mode

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

$$\mathrm{Mod}(r \mid \alpha,\beta,N,X) = \frac{X+\alpha-1}{N+\alpha+\beta-2}.$$

Note that mode is well defined only when both parameters of the beta-distribution are larger than one. Median would be yet more tricky due to lack of exact solution. As noted, the posterior mean can also be written as a weighted average:

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

showing how the prior and the data contribute to the estimate. This is a useful way to summarize the relative importance of both sources of information. But the simple analytical expression is limited to conjugate modeling only.

With numerical and graphical approaches we can draw this posterior density in each situation by simply plotting the beta-density. But then we need a software, such as R. For example, using the commands

```
X <- 2; N <- 20
p <- seq(0,1,by=0.01)
plot(p,dbeta(p,X+1,N-X+1),type="l")
```

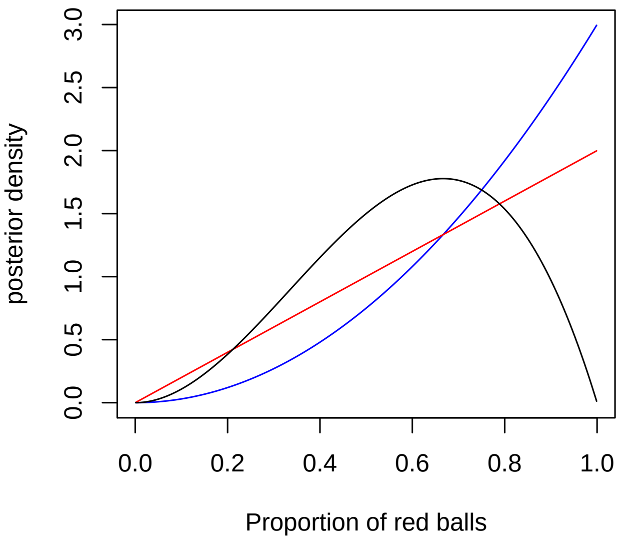


Figure 9: Posterior probability density for the proportion of red balls in an infinitely large bag of infinitely many balls, if one ball is drawn and it is red (red line), and if two balls are drawn and both are red (blue line), and if three balls are drawn and one is white (black line).

6.1 What to report: posterior mean, median or mode?

Finally, should we summarize a posterior distribution by its mean, median, mode or something else? Eventually, this should depend on the context and the purpose of the analysis. This could be mathematically tackled by a *loss function*. This function should define how much the error 'costs' when making a decision. For example, if we estimate the unknown proportion p in a population by choosing a point estimate δ_x that depends somehow on our data x, and if we define a quadratic loss

$$L(p, \delta_x) = (p - \delta_x)^2$$

then the Bayes risk

$$\int \int L(p, \delta_x) \pi(p \mid x) \mathbf{d} x \mathbf{d} p$$

should be minimized. This will be minimized by minimizing the *posterior loss*

$$E(L(p,\delta_x) \mid x) = \int L(p,\delta_x)\pi(p \mid x)\mathbf{d}p$$

for each x. In this example with the quadratic loss, we minimize

$$\int (p-\delta_x)^2\pi(p \mid x)\mathbf{d}p = \int (p^2 - 2p\delta_x + \delta_x^2)\pi(p \mid x)\mathbf{d}p = E(p^2 \mid x) - 2\delta_x E(p \mid x) + \delta_x^2.$$

The minimizer is found by derivation with respect to δ_x, and solving

$$-2E(p \mid x) + 2\delta_x = 0$$

which gives $\delta_x = E(p \mid x)$, the posterior mean. Similarly, we can think of some function of the parameter $h(p)$, so that the posterior mean $E(h(p) \mid x)$ is again the choice which will minimize the posterior loss with quadratic loss function $(h(p) - \delta_x)^2$.

Posterior median will minimize the loss with absolute error $L(p,\delta_x) =\mid p - \delta_x \mid$, and posterior mode will minimize the loss with 'all-or-nothing' error $1_{\{p\neq\delta_x\}}(\delta_x)$.

In general, a full Bayesian analysis would indeed consist of a decision problem for a real life application, where the decisions and consequences have specific losses. Then we need to choose a decision which minimizes the posterior loss. Thinking of practical computation, it is easy to see that this can be hard. Firstly, the posterior density $\pi(p \mid x)$ is generally not available in closed form. Secondly, the calculation of $E(h(p) \mid x)$ is generally difficult and not possible analytically. Also, other loss functions can be even more difficult.

6.2 Credible Intervals (CI)

Mode shows where the distribution is mostly concentrated, but it does not convey information about how uncertain we are. This is always the problem with point summaries (as with point estimates in non-Bayesian statistics). Hence, variance of a distribution could be reported in addition. However, we are often required to report a region, or interval, to describe the uncertainty. From a posterior distribution we can immediately obtain intervals that contain a specific probability. The interval is usually defined so that the point summary is somewhere in the middle, but not necessarily exactly in the middle. Any interval $[a,b]$ for which

$$\int_a^b \pi(r \mid \text{data})\mathbf{d}r = Q$$

is said to be a $Q \times 100\%$ *Credible Interval.* This is usually constructed simply by taking $Q/2$ off from both ends of the distribution. But this is not necessarily the shortest possible interval. The shortest Credible Interval is called Highest Posterior Density Interval (HPD-interval). The simple Credible Interval is computationally easier to obtain. For standard distributions, it can be calculated by using tabulated (or computerized) quantiles. For example, to compute the 95% CI for the posterior of r, shown as black line in Figure (9), in R-software:

```
> qbeta(c(0.025,0.975),2+1,3-2+1)
[1] 0.1941204 0.9324140
```

And to calculate all 95% Credible Intervals of r for all possible outcomes $x \in [0,N]$:

```
N<-100; y<-0:N
lower<-qbeta(0.025,y+1,N-y+1);
upper<-qbeta(0.975,y+1,N-y+1);
plot(c(y[1],y[1]),c(lower[1],upper[1]),'l',
xlab='Red balls in a sample of N=100',
ylab='Bayesian 95% CI',
xlim=c(0,100),ylim=c(0,1));
for(i in 2:length(y)){
points(c(y[i],y[i]),c(lower[i],upper[i]),'l');
}
```

In comparison, the corresponding HPD interval of r would contain the same probability (e.g. 0.95), but we would need to find such interval that $\pi(r^* \mid X, N) > \pi(r \mid X, N)$ when r^* and r are any values within and outside the interval, respectively.

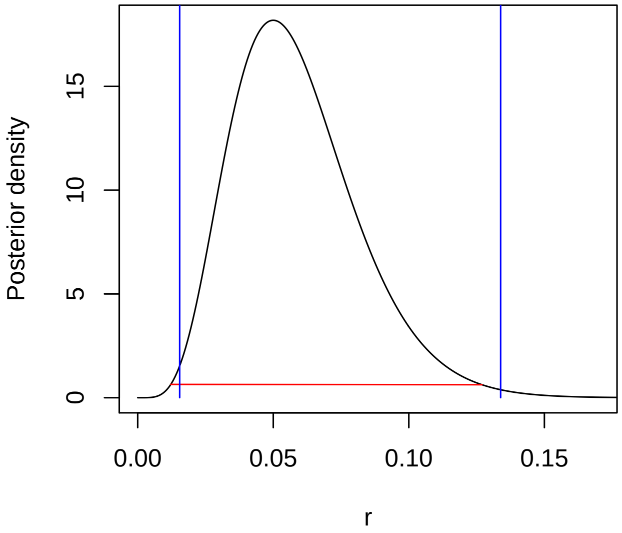


Figure 10: Comparison of HPD credible interval and simple credible interval from Beta(5+1,100-5+1) density. Red line shows 99% HPD interval. The length of 99% HPD CI is 0.1148 compared to 0.1184 of the simple 99% CI.

As a non-bayesian alternative, the exact frequentist 95% Confidence Interval (Clopper-Pearson interval) would be the set

$$\{r : P(Y \leq Y^{obs} \mid N, r) \geq 0.025\} \cap \{r : P(Y \geq Y^{obs} \mid N, r) \geq 0.025\}$$

which could be calculated for every outcome $y \in [0, N]$ as:

```
N<-100; y<-0:N
p<-seq(0,1,by=0.001);
```

```
I<-(1-pbinom(y[1]-1,N,p)>0.025)&(pbinom(y[1],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
plot(c(y[1],y[1]),c(lower,upper),’l’,
xlab=’Red balls in a sample of N=100’,
ylab=’Freq. 95% CI’,xlim=c(0,N),ylim=c(0,1));
for(i in 2:length(y)){
I<-(1-pbinom(y[i]-1,N,p)>0.025)&(pbinom(y[i],N,p)>0.025);
lower<-min(p[I*(1:length(p))]);
upper<-max(p[I*(1:length(p))]);
points(c(y[i],y[i]),c(lower,upper),’l’)
}
```

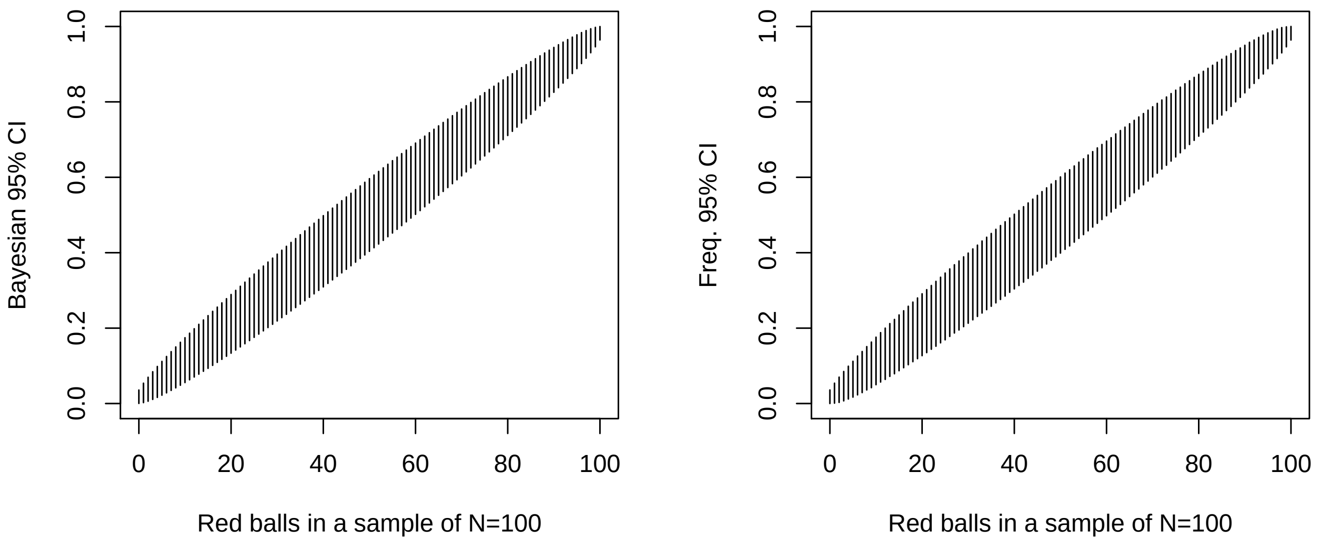


Figure 11: Bayesian Credible Intervals and frequentist Confidence Intervals.

The figure (11) looks very similar in both frequentist and bayesian calculations. Note, however, the difference of interpretation. In the bayesian approach, the unknown proportion r has distribution. In the frequentist approach, r is fixed unknown constant, and the *interval* is random, and it *would* cover the true unknown value of r in 95% of the cases if the experiment was repeated, but it says nothing about the probability that r belongs to this interval for any given sample Y that actually occurred. (See [**?**], page 453).

The bayesian CI was solved by finding the integration limits for the posterior, such that the required probability is achieved between $[a, b]$. In general, the HPD-CI can be a set of distinct intervals if the posterior density happens to be multimodal. Numerical techniques for solving the CI's would require that we can calculate the posterior density function accurately (which was possible above).

6.2.1 The more data, the narrower CI can be expected

Obviously, the resulting width of a CI depends on the amount of information we had. When the amount of data increases, we can expect the posterior to become more peaked, and hence the CI more narrow. On average, this is guaranteed because the prior variance of r can be written as

$$V(r) = E(V(r \mid X)) + V(E(r \mid X))$$

which shows that the posterior variance $V(r \mid X)$ is *expected* to be smaller than the prior variance. We can study the expected width of the CI with different sample sizes N and choose the value of N that gives the required expected width.

7 Predictive distributions

While posterior density summarizes our current uncertainty about an unknown quantity, predictions of future experiments and events could sometimes be even more interesting. (Some have even argued that it is the ultimate purpose of modeling). The posterior density is the basis for this too. What we get is a **posterior predictive distribution**. Assume a parametric model $\pi(X_i \mid \theta)$ for each variable in the sequence $X_1, X_2, \ldots$, so that the variables X_i are conditionally independent of each other, given the parameter θ. The goal is to predict next $X^{\text{next}} = X_{n+1}$, given the previous observed values $X = \{X_1, \ldots, X_n\}$.

$$\pi(X^{\text{next}} \mid X) = \int \pi(X^{\text{next}}, \theta \mid X)\mathbf{d}\theta = \int \underbrace{\pi(X^{\text{next}} \mid \theta, X)}_{=\pi(X^{\text{next}}|\theta)} \pi(\theta \mid X)\mathbf{d}\theta$$

This is the solution to the practical problem: having some probability model $\pi(X \mid \theta)$, how to compute a prediction, based on our observations X, **without knowing** the underlying value of θ? It is easy to calculate $\pi(X \mid \theta)$ and generate random values for X, when we assume some specific value for θ. In practice, this is unknown in every real application. Therefore, we use probability distribution to describe our uncertainty about θ. But the data informs us about probable values of θ. Hence, the posterior distribution is used, and the prediction distributions $\pi(X^{\text{next}} \mid \theta)$ are weighted by this posterior distribution.

Before having data, we just had the prior. From this, we can similarly compute the **prior predictive distribution**:

$$\pi(X^{\text{next}}) = \int \pi(X^{\text{next}}, \theta)\mathbf{d}\theta = \int \pi(X^{\text{next}} \mid \theta)\pi(\theta)\mathbf{d}\theta$$

In these notations, we could have written that they are conditional on the prior information, so that the prior predictive distribution actually is $\pi(X^{\text{next}} \mid I)$. This corresponds to our prior beliefs about the *observable* variables X. The parameter θ can be seen as purely a technical device, which provides a way to write this. This parameter may or may not have a close interpretation as a physical condition. Our focus is on assigning our probabilities to the actually observable quantities X. Parameter θ may have no interest in its own right.

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).

7.1 Exchangeability

Consider a sequence of binary variables 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 bayesian modeling. 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).

In all this, 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. Hence, parameter r is just a mathematical device that allows us to update our probabilities concerning the X_i.

Similarly, exchangeability works for other sequences of variables, not just binary variables. Whenever our beliefs about the observable variables X_i are exchangeable, it follows that there must exist a parametric model $\pi(X \mid \theta)$ and a distribution $\pi(\theta)$ so that our probability of $X_1, \ldots, X_n$ can be expressed as

$$\pi(X_1, \ldots, X_n) = \int_\Theta \prod_i^n \pi(X_i \mid \theta)\pi(\theta)\mathbf{d}\theta$$

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

Therefore, the X_i are not independent of each other, **only conditionally independent**, given θ. This means that we can learn from the observed X_i to predict other X_j that are not yet observed.

Quoting J Bernardo: *It is important to realise that if the observations are conditionally independent, - as it is implicitly assumed when they are considered to be a random sample from some model - , then they are necessarily exchangeable. The representation theorem, - a pure probability theory result - proves that if observations are judged to be exchangeable, then they must indeed be a random sample from some model and there must exist a prior probability distribution over the parameter of the model, hence requiring a Bayesian approach. Note however that the representation theorem is an existence theorem: it generally does not specify the model, and it never specifies the required prior distribution. An additional effort is necessary to assess a prior distribution for the parameter of the model.*

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.

7.2 Prediction with binomial model

For example, assume that the experiment of drawing balls is to be continued after the first three balls were picked. We should then predict the color of the next ball. Our model tells us that, conditionally on r, the probability of red ball in the next draw is simply r (according to a parametric model and de Finetti). But the true value of r was unknown (and will remain unknown, representing an infinite population). In such parametric model, we could use our current estimate for the parameter, but a fixed point estimate does not account for the fact that we are still uncertain about the parameter. The posterior predictive probability for the next ball to be red is:

$$P(\text{red} \mid Y, N) = \int_0^1 \underbrace{P(\text{red} \mid r)}_{=r} \times \underbrace{P(r \mid Y, N)}_{\text{Beta(Y+1,N-Y+1)}} \quad \mathbf{d}r \quad = \; E(r \mid Y, N) = \frac{Y+\alpha}{N+\alpha+\beta}$$

which is the same as the posterior mean of parameter r.

Next: consider an experiment where N new balls are to be picked, X of them will be red, so $X \sim \text{Bin}(N, r)$, and our current uncertainty about r is represented by beta-distribution $\text{Beta}(\alpha, \beta)$ (which could be the posterior of r, based on some earlier data). What is the predictive distribution of X in this new experiment?

$$P(X \mid N, \alpha, \beta) = \int_0^1 \underbrace{P(X \mid N, r)}_{\text{Bin}(N,r)} \underbrace{\pi(r \mid \alpha, \beta)}_{\text{Beta}(\alpha,\beta)} \mathbf{d}r$$

$$= \int_0^1 \frac{\Gamma(N+1)}{\Gamma(X+1)\Gamma(N-X+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{X+\alpha-1}(1-r)^{N-X+\beta-1} \, \mathbf{d}r$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 r^{X+\alpha-1}(1-p)^{N-X+\beta-1} \mathbf{d}r$$

Then, write: $A = X + \alpha$, $B = N - X + \beta$, so that

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_0^1 \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} r^{A-1}(1-r)^{B-1} \mathbf{d}r}_{=1} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \frac{\Gamma(N+1)\Gamma(\alpha+\beta)}{\Gamma(X+1)\Gamma(N-X+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

$$= \binom{N}{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)}$$

which can also be written using so called *beta-functions:*

$$\binom{N}{X}\frac{\mathrm{beta}(A,B)}{\mathrm{beta}(\alpha,\beta)}$$

This distribution of X is said to be *beta-binomial* distribution. It is sometimes used e.g. in food safety microbial risk assessments to describe e.g. the number of contaminated servings X among N servings, under uncertainty about the true fraction, r, of contaminated servings in a large (infinite) population. In risk assessment literature, the conditional distribution of X (binomial distribution) is often called as the variability distribution of X, and the distribution of r (beta distribution) as the uncertainty distribution of r. Hence, it is often said in RA-literature that ’variability and uncertainty are separated’. In bayesian context, both distributions are expressions of uncertainty (’epistemic uncertainty’ and ’aleatoric uncertainty’), and the resulting beta-binomial distribution reflects both types of uncertainties. The result can be either prior predictive distribution (in which case α, β represent parameters of a prior (Beta-) distribution), or posterior predictive distribution (in which case α, β represent parameters of a posterior (Beta-) distribution). Beta-binomial distribution can be used to account for **overdispersion in binomial models**: the distribution has two parameters, α, β, in place of the single parameter r of the binomial distribution.

By using the two general (often useful) probability laws for total expectation and total variance:

$$E(X) = E(E(X \mid Z))$$

and

$$V(X) = E(V(X \mid Z)) + V(E(X \mid Z)),$$

the mean of beta-binomial can be found from

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

Similarly, its variance can be found from

$$V(X) = E(V(X \mid r, N)) + V(E(X \mid r, N)) = \frac{N\alpha\beta(\alpha+\beta+N)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

```
R-code for producing the figure:
par(mfcol=c(2,1))
plot(seq(0,1,by=0.01),
dbeta(seq(0,1,by=0.01),20,35),’l’,lwd=2,xlab="p ~ beta(20,35)",ylab="")
p <- rbeta(3,20,35)
points(p[1],0,cex=2,col="blue",pch=16)
points(p[2],0,cex=2,col="red",pch=16)
points(p[3],0,cex=2,col="green",pch=16)
x <- 0:50
plot(x,dbinom(x,50,p[1]),’h’,lwd=5,col="blue",ylab="",xlab="P(x|p)=Bin(50,p)")
points(x,dbinom(x,50,p[2]),’h’,lwd=5,col="red")
points(x,dbinom(x,50,p[3]),’h’,lwd=5,col="green")
N <- 50; a <- 20; b <- 35; A <-x+a; B <- N-x+b
pr <- (gamma(N+1)*gamma(a+b)/(gamma(x+1)*gamma(N-x+1)*gamma(a)*gamma(b)))*
      gamma(A)*gamma(B)/gamma(A+B)
points(x,pr,pch=16)
```

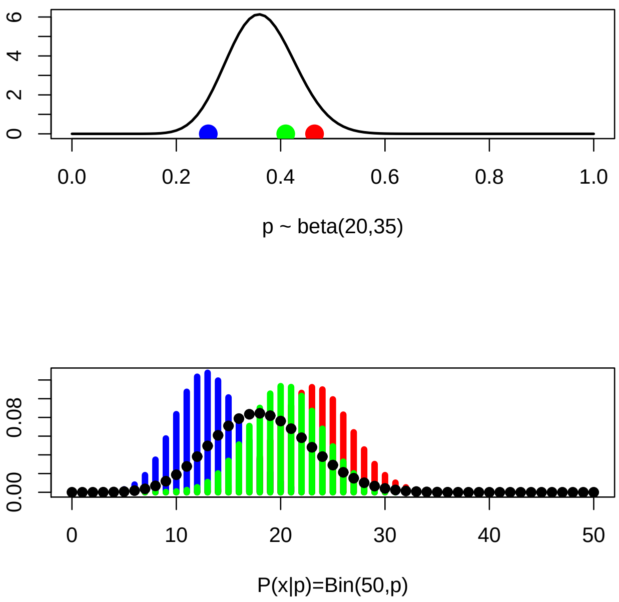


Figure 12: Upper frame: density of p (Beta(20,35)) and three randomly sampled values (red,blue,green). Lower frame: three conditional distributions for X (Bin(50,p)) with different sampled values of p (red,blue,green) corresponding to the upper frame. Integrating over all possible p according to the density of p, gives beta-binomial distribution for X (black dots). If the density of p was a posterior density based on earlier observed X_{obs}, then this gives the posterior predictive distribution of next X ($P(X \mid X_{\text{obs}})$).

7.2.1 Overdispersion not possible for Bernoulli variables

As a side step, consider a situation in which we pick N new balls, but assuming that each of the balls is picked from a different population (e.g. different bags) so that for each draw we have Bernoulli-distribution with different parameter r_i. (Bin$(1, r_i)$). Our uncertainty about all r_i is assumed to be described as some distribution $\pi(r_i)$, (which could be Beta(α, β)). What is the distribution of X?

$$P(X \mid N) = \int_0^1 \dots \int_0^1 P(X \mid r_1, \dots, r_N) P(r_1, \dots, r_N) \mathbf{d}r_1 \dots \mathbf{d}r_N$$

$$= \int_0^1 \dots \int_0^1 \binom{N}{X} \prod_{i=1}^{X} r_{k_i} \prod_{i=N-X}^{N} (1 - r_{k_i}) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1} (1 - r_{k_i})^{\beta-1} \mathbf{d}r_{k_1} \dots \mathbf{d}r_{k_N}$$

Here, $k_1, \dots, k_N$ is some permutation of the indices i. After re-arranging the terms in this expression, we get:

$$\binom{N}{X} \int_0^1 \dots \int_0^1 \prod_{i=1}^{X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha+1-1} (1 - r_{k_i})^{\beta-1} \prod_{i=N-X}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r_{k_i}^{\alpha-1} (1 - r_{k_i})^{\beta+1-1} \mathbf{d}r_{k_1} \dots \mathbf{d}r_{k_N}$$

and by integrating over each r_i one by one, we get:

$$= \binom{N}{X} E(r_i)^X E(1-r_i)^{N-X} = \text{Bin}\Big(N, \frac{\alpha}{\alpha+\beta}\Big)$$

This is a distribution that depends on N and the expected value of r_i, so the prior distribution of r_i affects the result via its expected value only. **It is not possible to model overdispersion for Bernoulli variables**.

8 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.

8.1 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 $\mathrm{Dir}(\alpha) = \mathrm{Dirichlet}(\alpha_1,\ldots,\alpha_k)$:

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

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 $\mathrm{Dir}(1,\ldots,1)$, which is the generalization of $\mathrm{Beta}(1,1)$. 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 $r \sim \mathrm{Dir}(\alpha)$, then the marginal distribution of each r_j is $\mathrm{Beta}(\alpha_j, \sum_i \alpha_i - \alpha_j)$, 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 $\frac{\alpha_j}{A}(1-\frac{\alpha_j}{A})/(A+1)$.

The marginal posterior distribution (here solved as Beta) allows to make probability statements of any single parameter, while accounting for the uncertainty in all parameters.

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

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

8.2 Normal model

As an example of a 2-dimensional problem, consider normal model for X, $\pi(X) = \mathrm{N}(\mu, \sigma^2)$, with unknown parameters μ and σ. Sometimes, another parametrization is used by defining *precision* $\tau = 1/\sigma^2$ instead of variance. (This is used in WinBUGS and OpenBUGS). **Note: notations get easily mixed! Below $\mathrm{N}(\mu, \sigma^2)$ can be casually written as $\mathrm{N}(\mu, \tau)$ which should not be understood as if τ was in the place of variance. Remember:** $\tau = 1/\sigma^2$.

Before the 2-dimensional problem, take a look at the one-dimensional problems where one of the parameters is assumed to be 'known'.

8.2.1 Unknown mean, known variance

Assume that variance σ^2 is known, but mean μ unknown. We would like to estimate the mean. Consider first a single observation X_i only. The conditional density of the observation is

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

where $\tau = 1/\sigma^2$ is the *precision*. As always, before calculating posterior of μ, we need to choose the prior. Assume that, for all practical purposes it is acceptable to consider the whole set $\mathbb{R}$ of real numbers as the range of possible values. It is possible 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 X_i, 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 e.g. [?] BSM p. 62). The posterior mean can be written as weighted average

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

where the weight is $w = \tau_0/(\tau_0 + \tau)$.

Next, assume the data has several values $X_1, \ldots, X_N$. The probability of the whole data set can be written using the average $\bar{X} = \sum X_i/N$ (which is the sufficient statistic):

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

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 τ_0 approaches zero, the prior density becomes flat and approaches zero everywhere. To describe a 'distribution' that is flat everywhere, we define an improper uniform density, $\pi(\mu) \propto 1$. 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 *statistic*, 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. The roles of $\bar{X}$ and μ are reversed in the two paradigms:

frequentist says: $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/N)$ bayesian says: $\mu \sim \mathrm{N}(\bar{X}, \sigma^2/N)$

8.2.2 Unknown variance, known mean

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:

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

$$= (\sigma^2)^{-N/2} \exp(-\frac{N}{2\sigma^2} s_0^2) = \tau^{N/2} \exp(-\frac{N\tau}{2} s_0^2)$$

where we have used the notation:

$$s_0^2 = \frac{1}{N} \sum_i^N (X_i - \mu)^2.$$

Since τ is unknown we must choose a prior for it. Alternatively, we could work out using $\sigma^2 = 1/\tau$, but let's use τ, because that is actually the parametrization used in WinBUGS and OpenBUGS. A conjugate choice would be Gamma(α, β) -distribution The posterior is then proportional to

$$\tau^{N/2} \exp(-\frac{N\tau}{2} s_0^2) \times \tau^{\alpha-1} \exp(-\beta\tau) = \tau^{N/2+\alpha-1} \exp(-(\frac{N}{2} s_0^2 + \beta)\tau)$$

Which can be recognized as Gamma$(N/2 + \alpha, \frac{N}{2} s_0^2 + \beta)$. An uninformative Gamma-prior is again obtained by setting α, β 'nearly zero'. So, in the limit the posterior would be Gamma$(\frac{N}{2}, \frac{N}{2} s_0^2)$ which has mean $1/s_0^2$. Setting $\alpha = \beta = 0$ in the Gamma-prior density gives $\pi(\tau) \propto \tau^{-1}$. By making the transformation $\theta = 1/\tau$, we get $\pi(\theta) \propto \theta \mid \frac{\mathbf{d}\theta^{-1}}{\mathbf{d}\theta} \mid = 1/\theta$. Hence, the corresponding improper prior for $\sigma^2 = 1/\tau$ is $\pi(\sigma^2) \propto 1/\sigma^2$.

8.2.3 Unknown mean and unknown variance

The previous solutions provided conditional distributions $\pi(\mu \mid \tau, \text{data})$ and $\pi(\tau \mid \mu, \text{data})$. These are called *full conditional distributions* which are obtained from the joint posterior density, based on the data and the two (independent) priors $\pi(\mu)$ and $\pi(\tau)$. These could be used for drawing random samples of μ and τ (one after another) from these full conditionals, which finally produces samples from the joint posterior distribution. (Gibbs sampling).

(1) Conjugate prior for the 2D-problem can be formulated as

$$\pi(\mu,\tau) = \pi(\mu \mid \tau)\pi(\tau) \quad \text{or } \pi(\mu,\sigma^2) = \pi(\mu \mid \sigma^2)\pi(\sigma^2)$$

In this case, joint posterior density can still be solved as a known distribution. 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. For practical data analysis purposes, this 2D-prior specification is slightly problematic because it requires to specify the prior distribution of μ conditionally on τ. This can be difficult to get e.g. from expert opinions, or any judgements of the application context. It seems more natural to specify priors for μ and τ separately. This leads to independent priors:

(2) Independent priors can be chosen as

$$\pi(\mu,\tau) = \pi(\mu)\pi(\tau) \quad \text{or } \pi(\mu,\sigma^2) = \pi(\mu)\pi(\sigma^2)$$

In this case, it is not possible to choose the distributions $\pi(\mu)$ and $\pi(\tau)$ so that the joint posterior density could be solved in any familiar form. In this case, we are forced to numerical calculations instead of analytical solutions.

(3) Finally, improper prior would be $\pi(\mu,\tau) \propto 1/\tau$, or $\pi(\mu,\sigma^2) \propto 1/\sigma^2$.

A *computationally useful* result is always to find out the full conditional distributions of τ and μ (or, in general, with any multidimensional bayesian inference). A numerical method called Gibbs sampler can be constructed from these. This provides a way to draw samples from the joint posterior distribution of τ and μ. With a large enough sample, we can calculate everything we need from the posterior, as a Monte Carlo approximation.

Solution with improper priors:

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: $\pi(\sigma^2 \mid X) = \text{Scaled Inv-}\chi^2(n-1, s^2)$ or $\pi(\tau \mid X) = \text{Gamma}(\frac{n-1}{2}, \frac{n-1}{2}s^2)$.

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, [**?**]. As a result, the marginal posterior is found to be a t-distribution so that

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