Introduction to Bayesian Inference

jukka.ranta@helsinki.fi

17.1.2012

Abstract

The course gives a practical introduction to bayesian inference and basics of WinBUGS / OpenBUGS. (Also R software will be modestly used). There are no pre-requirements other than reasonable familiarity with basic differential and integral calculus and functions (mostly taught at high school advanced courses already), and probability theory at basic level. Concepts of discrete and continuous random variables, their distributions and parameterizations of basic distributions should be reasonably familiar, as well as basic laws of probability theory. It can be an advantage to have knowledge of basic (non-bayesian) statistics, although not necessary.

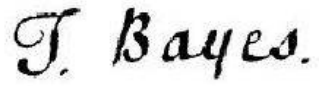
Contents

1 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/.



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 to 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 to 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.

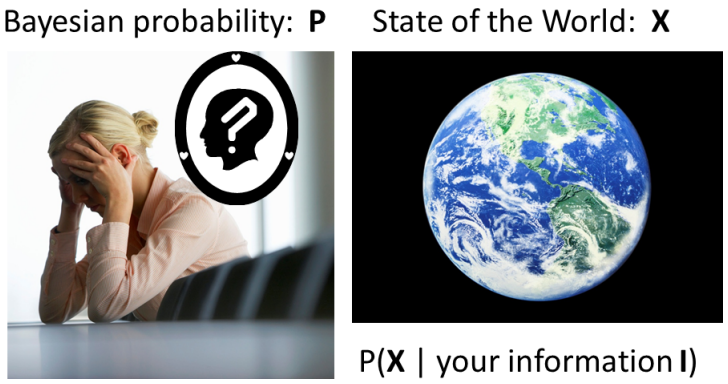


Figure 2: Probability is in the head of the observer.

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.

1.1 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:

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

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:

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

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

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

Note that the selection of a ball was 'randomized' or 'blindfolded' only as far as we could know about it. It may not be 'truly random'. We could always think of someone more informed than us, who knows better the positions of the balls and the movements of the hand that picks the ball. There would not be aleatory uncertainty for him. Someone who knows exactly the initial conditions and how the ball is to be picked also knows the result without any uncertainty. This effect is exploited in magic tricks. But it shows that also aleatory uncertainty is actually a form of our uncertainty, arising from incomplete knowledge. The outcome of every 'random experiment' is predictable *if* we only knew the *exact* initial conditions. E.T. Jaynes has discussed the "physics of random experiments" in his book "Probability theory, the logic of science" [6], discussing also quantum mechanics. For the purpose of quantifying our uncertainty, it remains open whether there really is 'true randomness' out there, or whether everything is thoroughly deterministic (or even something else?). We do not need to assume either way, because we describe and update our uncertainties based on what we *can* know.

1.2 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 to 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 to 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 formula 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. (Notation '$\propto$' means 'proportional to', or 'up to a constant term', ('vakiokerrointa vaille'), so that in this particular example all terms that are not functions of r are left out). 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 $2/(M(M+1)), 4/(M(M+1)), 6/(M(M+1)), ...$

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:

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

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

where the posterior after the 1st observation was:

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

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

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

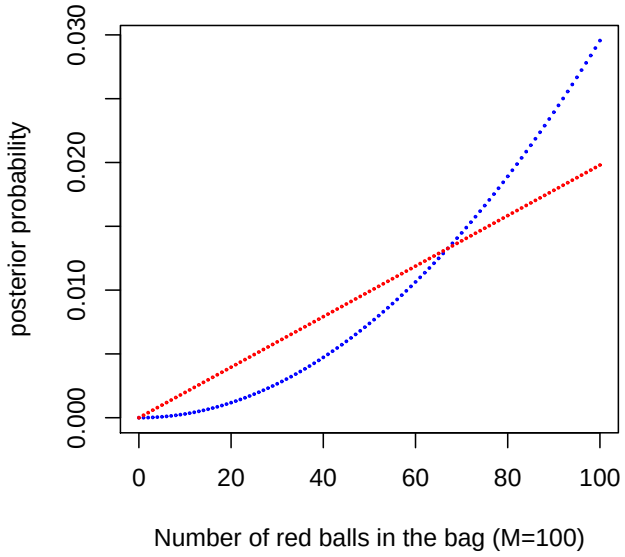


Figure 3: Posterior probabilities for the number of red balls among M in a bag, if one ball is drawn and it is red (red dots), and if two balls are drawn and both are red (blue dots).

1.3 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 *non-informative* prior. It is commonly applied when there is no knowledge indicating unequal probabilities.

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

Broadly, these two approaches are sometimes called as *objective* bayesian [2] and *subjective* bayesian [3] approach. If the data are very informative about the quantity being estimated, then an uninformative prior is a quick and easy choice. Actually, if the data are extremely informative, then nearly any prior would lead to the same posterior probability. But if the data are poor, then the posterior will be heavily influenced by the prior and it is more important to think how the prior was chosen and how sensitive the result is to different priors. Also, there can be really important expert knowledge (that is not part of the observed data already). That 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.

1.3.1 Simple elicitation of informative prior probability

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

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

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

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

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

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

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

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

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

Another way to approach subjective probability is by using *odds*. When making bets (at some 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. The number of questionnaires 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 for universal noninformative priors can be sufficient (and free of elicitation problems!). However, the quest for a truly universal method for a noninformative 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 some transformation 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' [4]: **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).

1.3.2 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.

1.4 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 limited frequency is always unknown to us, because we cannot repeat any experiment truly infinitely. (Compare with bayes: all probabilities are known!).

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

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.

1.5 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 densities:

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

For bayesian inference too, 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 'recognizing'). 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 distributions. Generally, they don't have to be so, and we could choose any other prior distribution, but the resulting posterior would not be among any 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 seen how 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 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. In bayesian analyzes, the posterior distribution is always the primary result and not some selected point values because they would not convey the same information about the uncertainty.

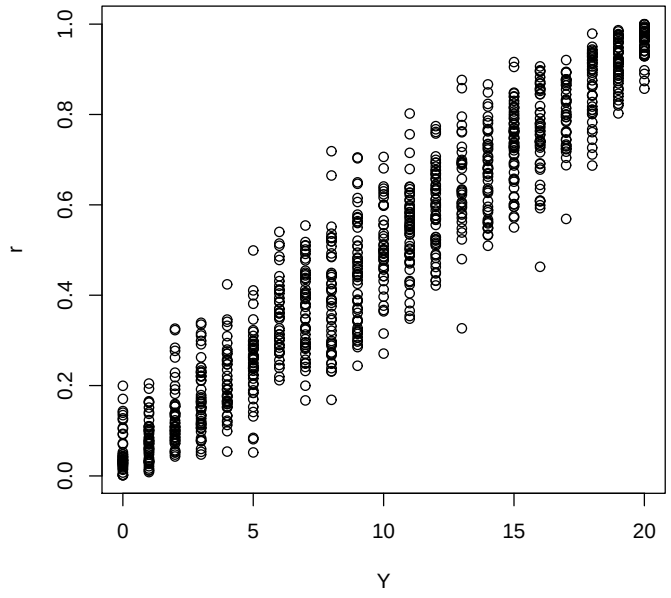


Figure 4: Simulated values from the joint distribution of $\pi(r, Y) = \pi(Y \mid r, N = 20)\pi(r)$ with uniform density $\pi(r)$. In R code: `r <- runif(1000)`, `Y <- rbinom(1000,20,r)`, `plot(Y,r)`. For any fixed r we have Binomial$(20, r)$ for Y, i.e. $\pi(Y \mid r)$. For any fixed Y we have Beta$(Y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

1.5.1 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:

$$\alpha = -m(mm - m + ss)/(ss) \ , \ \beta = (mm - m + ss)(m - 1)/(ss)$$

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.

1.5.2 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. This seems to be a problem because one could say that if I'm completely uncertain about r, I should be similarly uncertain about r^2 - if that happens to be of interest too. The original Bayes-Laplace prior $r \sim \mathrm{U}(0,1)$ would not imply a uniform prior for r^2, and vice versa. (The density of $q = r^2$ would be $\pi(q) = 0.5q^{-0.5}$, by using the transformation of variables rule, if $\pi(r) = \mathrm{U}(0,1)$).

The priors can be interpreted as being equivalent to some amount of 'prior data'. The uniform prior Beta$(1,1)$=U(0,1) corresponds to having 2 prior experiments, one of which was a 'red ball' and the other 'white ball'. The Jeffreys' prior equals to having only one prior experiment in which

one ball was 'drawn' and it was 'half red', 'half white'. In this sense, Haldane's prior corresponds to having no prior data at all, but the prior is actually concentrated at two points: zero and one. Moreover, with Beta$(0,0)$ prior the posterior is not defined if the observed data happens to be either 0 or N under a Binomial(N,r) model.

The Jeffreys' prior is based on the principle that an uninformative prior should be such that it does not depend on which parameter transformation is used: it should be the same for all transformations. 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.

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

(This equality is borrowed, without proof, from classical texts where the Fisher information is more used, and these two equivalent forms are 'well known' parlance).

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. This means that if we calculate the prior $\pi(\psi)$ for some transformation ψ of the original parameter r, we get, using the variable transformation rule:

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

and if the original prior $\pi(r)$ is chosen as Jeffreys, then $\pi(\psi)$ is proportional to

$$\propto \sqrt{E\Big[\Big(\frac{\mathbf{d}\log L}{\mathbf{d}r}\Big)^2\Big]\Big(\frac{\mathbf{d}r}{\mathbf{d}\psi}\Big)^2} = \sqrt{E\Big[\Big(\frac{\mathbf{d}\log L}{\mathbf{d}r}\frac{\mathbf{d}r}{\mathbf{d}\psi}\Big)^2\Big]} = \sqrt{E\Big(\frac{\mathbf{d}\log L}{\mathbf{d}\psi}\Big)^2} = \sqrt{J(\psi)}$$

where L denotes the likelihood $\pi(\text{data} \mid \text{parameter})$.

So, the prior of the transformed parameter $\pi(\psi) \propto \sqrt{J(\psi)}$, if the prior of the original parameter $\pi(r) \propto \sqrt{J(r)}$.

Calculate the Jeffreys' prior for the binomial proportion r. To begin, we have the following from the binomial model:

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

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

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

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

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

The Jeffreys' prior for binomial proportion r is thus

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

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

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

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

$$\pi(\psi \mid X) = \pi(\, r(\psi) \mid X \,) \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid \;\propto\; \pi(X \mid r(\psi)) \underbrace{\pi(r(\psi))}_{\text{Jeffreys'}} \mid \frac{\mathbf{d}r}{\mathbf{d}\psi} \mid$$

$$\propto \underbrace{\psi^{2X}(1-\psi^2)^{(N-X)}}_{\propto \text{Bin}(N,\psi^2)} \times (\psi^2)^{-1/2}(1-\psi^2)^{-1/2} \times 2\psi.$$

(2). Compute **directly the posterior** $\pi(\psi \mid X) \propto \pi(X \mid \psi)\pi(\psi)$ using ***Jeffreys' prior for*** ψ. In this case, $\log \pi(X \mid \psi) = \text{constant} + 2X \log(\psi) + (N-X)\log(1-\psi^2)$, 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:

$$\pi(\psi \mid X) \propto \pi(X \mid \psi) \underbrace{\pi(\psi)}_{\text{Jeffreys'}}$$

$$= \underbrace{\psi^{2X}(1-\psi^2)^{(N-X)}}_{\propto \text{Bin}(N,\psi^2)} \times (1-\psi^2)^{-1/2}.$$

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 Beta(α, β), then the posterior will be Beta$(X+\alpha, N-X+\beta)$ and the posterior mode is then $(X+\alpha-1)/(\alpha+\beta+N-2)$, and posterior mean is $(X+\alpha)/(\alpha+\beta+N)$. The posterior mode becomes X/N when the Bayes-Laplace prior is used. The posterior mean becomes X/N when the Haldane's prior is used. 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: $\text{argmax}_{r\in[0,1]} P(X \mid N, r)$.

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

1.5.3 Unknown N

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

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

Now the posterior is:

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

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

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

and the normalizing constant is

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

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

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

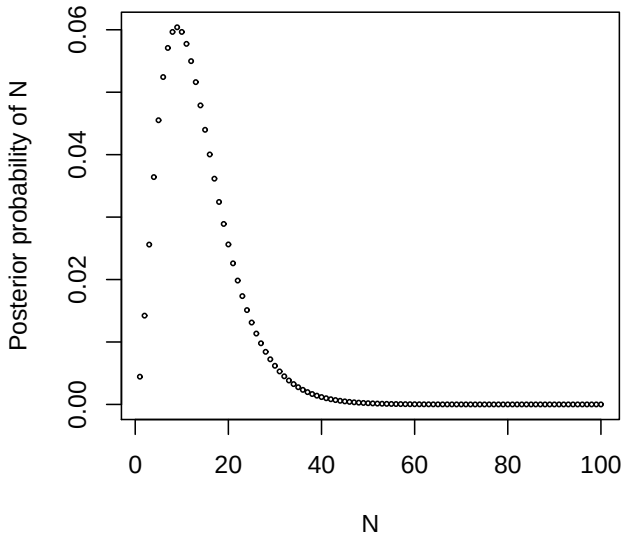


Figure 5: Posterior probability for N, given that $X = 1, r = 0.2$ with uniform prior over $0, 1, 2, \ldots, M = 100$.

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

In all cases, **bayesian model is the full joint distribution**. If we have the binomial model for data, $P(X \mid N, p) = \text{Bin}(N, p)$, the bayesian model is $\pi(X, p)$ as shown in a Figure before. If we also treat N as an uncertain quantity, the full model is $\pi(N, X, p)$. Depending on whatever becomes observed, the bayesian learner will compute a conditional distribution from the full model, by conditioning to the observed variables.

2 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 formula (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: 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 and mode

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

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

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 models only.

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")
```

2.1 Choosing an estimate on the basis of loss functions

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. *The chosen loss function determines which point estimate is best.* (By 'point estimate' we mean one of the possible point values for summarizing the posterior

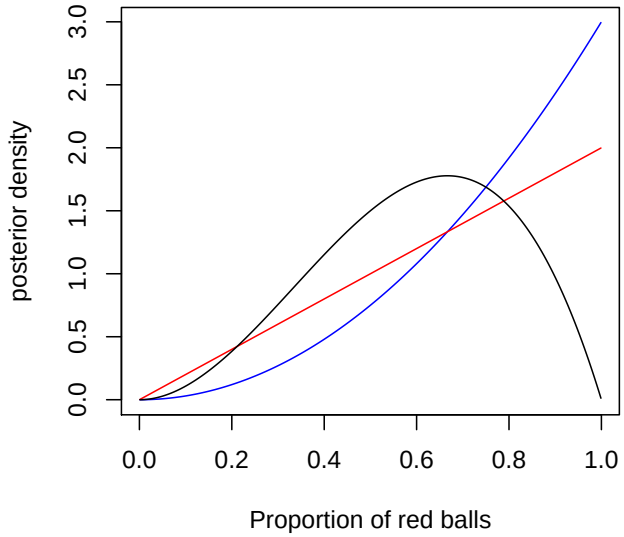


Figure 6: Posterior probability density for the proportion of red balls in an infinitely large bag of infinitely many balls, if one ball is drawn and it is red (red line), and if two balls are drawn and both are red (blue line), and if three balls are drawn and one is white (black line).

distribution). For example, if we estimate some unknown parameter p 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. With the quadratic loss function, we get

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

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

$$-2(\delta_x - E(p \mid x)) \underbrace{\int (p - E(p \mid x))\pi(p \mid x)\mathbf{d}p}_{=0}$$

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

which is minimized when $\delta_x = E(p \mid x)$.

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 $1_{\{p=\delta_x\}}(\delta_x)$.

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.

2.2 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 (6), 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.

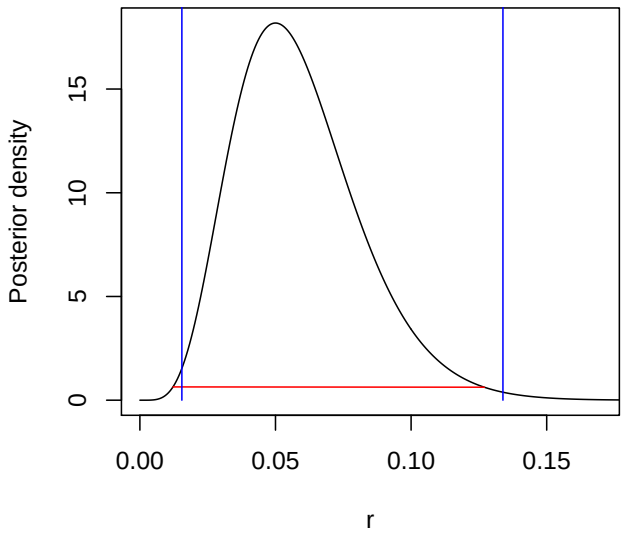


Figure 7: Comparison of HPD credible interval and simple credible interval from Beta(5+1,100-5+1) density. Red line shows 99% HPD interval. The length of 99% HPD CI is 0.1148 compared to 0.1184 of the simple 99% CI.

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')
}
```

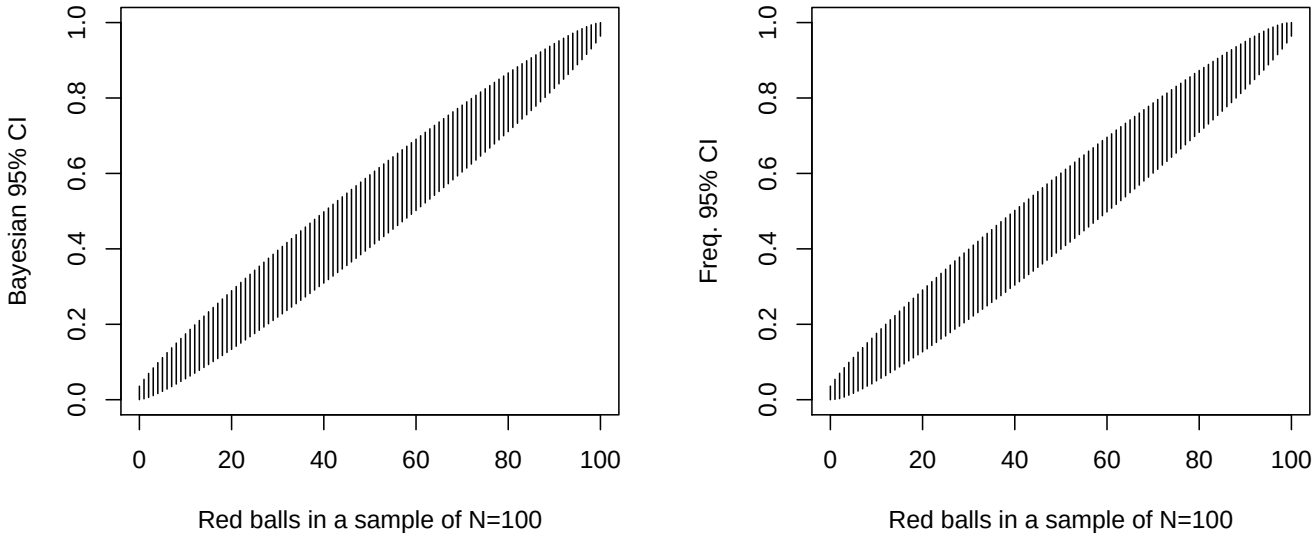


Figure 8: Bayesian Credible Intervals and frequentist Confidence Intervals.

The figure (8) 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 [11], 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).

2.2.1 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.

3 Predictions

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 $X = \{X_1, \ldots, X_n\}$.

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

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 to 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).*

3.1 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 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.

3.2 Prediction for binomial experiment

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:

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

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:*

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

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 (but perhaps 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)
```

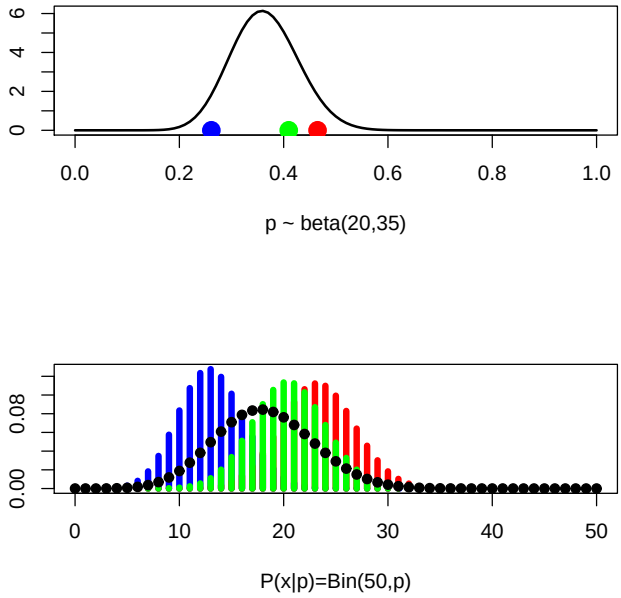


Figure 9: Upper frame: density of p (Beta(20,35)) and three randomly sampled values (red,blue,green). Lower frame: three conditional distributions for X (Bin(50,p)) with different sampled values of p (red,blue,green) corresponding to the upper frame. Integrating over all possible p according to the density 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 ($P(X \mid X_{\text{obs}})$).

3.2.1 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**.

3.3 Example: mixture priors

Returning to the mixture priors mentioned before: the prior predictive distribution from an individual component distribution is

$$\pi_i(x) = \int_\Theta \pi_i(\theta)\pi(x \mid \theta)\, \mathbf{d}\theta,$$

and the prior predictive distribution from the whole mixture is

$$\pi(x) = \int_\Theta \pi(\theta)\pi(x \mid \theta)\, \mathbf{d}\theta = \sum_{i=1}^{k} \alpha_i \pi_i(x).$$

Now, the posterior distribution can be written as:

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

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

which is seen as a weighted average of component specific posterior distributions, with weights calculated from the predictive distributions as shown.

4 Hypotheses

Hypotheses are usually formulated for some parameter θ so that the null hypothesis is written

$$H_0 : \theta \in \Theta_0$$

against an alternative hypothesis

$$H_1 : \theta \in \Theta_1$$

If both sets have positive probabilities, e.g. if they are intervals and θ is a continuous parameter $\in S \subset \mathbb{R}$, then the Bayesian approach is to compute a posterior probability, with data X:

$$\pi(H_0 \mid X)$$

and the posterior probability for the alternative is $1 - \pi(H_0 \mid X)$. The posterior probability summarizes the current evidence. What remains is to evaluate the results and decide how large (small) probability is large (small) enough. The posterior probability is not accepting or rejecting a hypothesis, it simply provides a numerical value for its plausibility. We need to think what level of plausibility is enough, if we had to take some action. Ultimately, this would call for a loss function in order to choose the decision that minimizes the expected loss with respect to the posterior distribution.

Note: frequentist hypothesis testing with p-values gives the probability of more extreme Y than the one observed, given null hypothesis: $P(Y$ more extreme than $Yobs \mid H_0)$. The null hypothesis may thus be rejected (or not), if a more extreme observation than what we had would seem too improbable. The frequentist hypothesis testing does not give probability of the hypothesis. Harold Jeffreys (1939), commented: "an hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. On the face of it, the evidence might more reasonably be taken as evidence for the hypothesis, not against it. The same applies to all the current significance tests based on P-values". The same arguments tend to get repeated back and forth.

For example, hypotheses related to linear models $E(Y_i) = \alpha_0 + \alpha_1 X_i$ could e.g. focus on the slope parameter for inspecting trend. $H_0 : \alpha_1 < 0$ and $H_1 : \alpha_1 \geq 0$. Hence, the posterior distribution $P(\alpha_1 < 0 \mid Y, X)$ directly assesses the probability of this hypothesis, $P(H_0 \mid Y, X)$. Actually, in this case, the hypothesis would not be about a true physical state of the world in the sense that 'a slope' does not exist: it is only a parameter of our model and we could have written a different model with different parameters. Hence, some hypotheses may have more direct interpretation as 'state of the world' than others.

Also, posterior odds can be written: for hypothesis H_0 against H_1 we have

$$\frac{\pi(H_0 \mid X)}{\pi(H_1 \mid X)}.$$

Whenever this is > 1, it shows support for H_0. *Bayes factor* is computed as a ratio of prior and posterior odds:

$$BF = \frac{\pi(H_0 \mid X)/\pi(H_1 \mid X)}{\pi(H_0)/\pi(H_1)} = \frac{\pi(H_0 \mid X)}{\pi(H_1 \mid X)} \frac{\pi(H_1)}{\pi(H_0)}.$$

In other words:

$$\textbf{Posterior odds } = \textbf{Prior odds} \times \textbf{Bayes factor.}$$

The Bayes factor measures how much the data changes the prior odds. If the factor is bigger than one, the data gave some support for the hypothesis H_0. Bayes factor provides a scale of evidence in favor of one hypothesis against another. (But the scale is from zero to infinity, which is not as 'neat' as probability scale).

If we have a point hypothesis (=simple hypothesis) where $H_0 : \theta = \theta_0$ against another point hypothesis $H_1 : \theta = \theta_1$, with some specific values of θ_0 and θ_1 (e.g. 3.5 against 5.7) then we must have positive probability for both, and the posterior odds is

$$\underbrace{\frac{\pi(\theta = \theta_0 \mid X)}{\pi(\theta = \theta_1 \mid X)}}_{\text{Posterior odds.}} = \overbrace{\underbrace{\frac{\pi(\theta = \theta_0)\times}{\pi(\theta = \theta_1)\times}}_{\text{Prior odds.}} \underbrace{\frac{\pi(X \mid \theta = \theta_0)}{\pi(X \mid \theta = \theta_1)}}_{\text{Likelihood ratio.}}}^{\text{Constant } \pi(X) \text{ cancels out}}$$

so that the likelihood ratio is the Bayes factor and this does not depend on the prior. In the expression above, it was sufficient to write the two posterior probabilities (on the left) 'proportional to' (i.e. just prior times likelihood, on the right) because the normalizing constant $\pi(X)$ cancels out from the ratio. From the equation, likelihood ratio could also be written as

$$\frac{\pi(X \mid \theta = \theta_0)}{\pi(X \mid \theta = \theta_1)} = \frac{\pi(\theta = \theta_0 \mid X)}{\pi(\theta = \theta_1 \mid X)} \frac{\pi(\theta = \theta_1)}{\pi(\theta = \theta_0)} = BF.$$

With a composite hypothesis where $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ we have a probability density of θ so that the probability for any single value of θ is zero. There is positive probability only for the set Θ_0, (and Θ_1). In a *one sided hypothesis test*: $H_0 : \theta < \theta_0$ against $H_1 : \theta \geq \theta_0$. The Bayes factor is then more complicated, and depends on prior:

$$BF = \underbrace{\frac{\int_{\Theta_0} \pi(\theta \mid X)\mathbf{d}\theta \int_{\Theta_1} \pi(\theta)\mathbf{d}\theta}{\int_{\Theta_1} \pi(\theta \mid X)\mathbf{d}\theta \int_{\Theta_0} \pi(\theta)\mathbf{d}\theta}}_{\text{post.odds}\times\text{prior odds}^{-1}} = \frac{\int_{\Theta_0} \pi(X \mid \theta)\pi(\theta)\mathbf{d}\theta}{\int_{\Theta_1} \pi(X \mid \theta)\pi(\theta)\mathbf{d}\theta} \frac{\int_{\Theta_1} \pi(\theta)\mathbf{d}\theta}{\int_{\Theta_0} \pi(\theta)\mathbf{d}\theta}$$

For a continuous parameter θ with a density function, *a two sided hypothesis* $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ does not make sense unless we place positive prior probability for the point θ_0 so that $q_0 = P(\theta = \theta_0) > 0$. With probability $q_1 = 1 - q_0 = P(\theta \neq \theta_0) > 0$. The prior is thus a mixture

$$\pi(\theta) = 1_{\{\theta=\theta_0\}}(\theta)q_0 + \pi_1(\theta)q_1$$

where a density function $\pi_1(\theta)$ is applied under H_1. This could be seen as a model choice problem where we have two competing models, M_0 and M_1 with prior probabilities for the models: q_0 and $q_1 = 1 - q_0$. Then:

$$\underbrace{\frac{P(\text{model} = M_0 \mid X)}{P(\text{model} = M_1 \mid X)}}_{\text{Post.odds}} = \frac{q_0}{q_1} \underbrace{\frac{\int \pi_0(X \mid \theta_0)\pi_0(\theta_0)\mathbf{d}\theta_0}{\int \pi_1(X \mid \theta_1)\pi_1(\theta_1)\mathbf{d}\theta_1}}_{BF}$$

where the probability of data X is either based on model M_0 as $\pi_0(X \mid \theta_0)$ or on model M_1 as $\pi_1(X \mid \theta_1)$. Parameters θ_0 and θ_1 in each model could have different dimensions. One model could be a three-parameter model, and the other a one-parameter model...

4.1 Example: evidence for population prevalence

If the hypothesis with a binomial model $\mathrm{Bin}(N, r)$ is that the large population prevalence $r < 0.5$, then the prior probability of that hypothesis is

$$P(H_0) = P(r < 0.5) = \int_0^{0.5} \pi(r)\mathbf{d}r = 0.5 \qquad \text{(from U(0,1)-prior)}$$

but the posterior probability, with data $Y = 2, N = 3$, would be

$$P(H_0 \mid Y, N) = P(r < 0.5 \mid Y, N) = \int_0^{0.5} \mathrm{Beta}(r \mid Y + 1, N - Y + 1)\mathbf{d}r$$

which is the cumulative probability of the beta-density at $r = 0.5$. The approximate value (0.3125) is obtained in R by typing `pbeta(0.5,Y+1,N-Y+1)`. The posterior probability became smaller than the prior probability.

We may also compute posterior odds to compare the difference. The prior odds for the hypothesis were

$$\frac{P(r < 0.5)}{P(r \geq 0.5)} = 1$$

but the posterior odds are only about half of that

$$\frac{P(r < 0.5 \mid Y, N)}{P(r \geq 0.5 \mid Y, N)} = \frac{0.3125}{0.6875} = 0.4545.$$

Therefore, posterior odds became smaller than prior odds, i.e. there was some evidence against the hypothesis. Jeffreys (1961) suggested that a Bayes factor bigger than 10 means strong evidence for the hypothesis, whereas a Bayes factor smaller than 1/10 means strong evidence against it.

Hypotheses could also involve comparisons of two quantities. For example, we could study two different bags, each with a different proportion of red balls, r_1 and r_2, and we get some observations from both, (Y_1, N_1) and (Y_2, N_2). The hypothesis could then be e.g. $H_0 : r_1 < r_2$. What is the prior and the posterior probability of the hypothesis? To study this, we can create a new variable: $s = r_1 - r_2$, so that $H_0 : s < 0$. But now the distribution of s is a convolution of two independent distributions and generally it may be difficult to compute, at least analytically. With iterative sampling methods, such posterior probabilities are routinely computed for applied problems.

4.2 Example: analysis of birth data

Example from Gelman [5]: the proportion of female births in Germany is 0.485. In a study of a rare condition of pregnancy it was observed that in 980 of such births, 437 were female. That's 0.4459184, which is a little lower than expected. How much evidence this gives for the claim that the proportion of female births in such conditions is lower than in the large population? Assuming uniform prior probability for the female proportion r, the posterior density becomes

$$\pi(r \mid X = 437, N = 980) = \mathrm{Beta}(438, 544).$$

The posterior mean of r is 0.446, and the posterior standard deviation 0.016. The median is 0.446, (`qbeta(0.5,438,544)`). The probability $P(r < 0.485)$ is

$$P(r < 0.485 \mid X, N) = \texttt{pbeta(0.485,438,544)} = 0.992826$$

which seems quite high. The posterior odds is 0.992826/(1-0.992826) = 138.4, and the prior odds $0.485/(1 - 0.485) = 0.94$, giving a Bayes factor of about 147.

This result was obtained when the prior was uniform. This was somewhat against our actual prior knowledge. We can check how much difference does it make if the prior would be more concentrated around population mean 0.485.

$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	posterior median	95%posterior interval
0.5	2	0.446	[0.415, 0.477]
0.485	2	0.446	[0.415, 0.477]
0.485	5	0.446	[0.415, 0.477]
0.485	10	0.446	[0.415, 0.477]
0.485	20	0.447	[0.416, 0.478]
0.485	100	0.450	[0.420, 0.479]
0.485	200	0.453	[0.424, 0.481]

The prior mean is outside the 95% interval in all of these. In the last case, the prior sample size equals already 200, and the prior is Beta(97, 103). From this we get prior odds of about 1.00. The posterior odds are about 78, so the Bayes factor is about 78. This is still large, but lower than 147. The choice of prior has an effect.

In addition to r, an interesting quantity is the sex ratio $z = (1 - r)/r$ and some research questions could be framed about it. Distribution of z could be found using the transformation of variables technique. In practice, it is easier to produce it by simulation techniques.

4.3 Example: winning Monty Hall

Monty Hall problem is a famous game in which you are first offered a choice over 3 boxes, one of which contains a prize and others are empty. Once you have made your initial choice, you are not yet allowed to open your box. Instead, one of the other boxes is shown to be empty by the game master who knows exactly what was placed in each box. You are then asked to make your final choice: do you keep your initially chosen box, or do you change for the other unopened box? The hypothesis under judgement is that A='the prize is in your box already' or B='the prize is in the other box'.

Initially, the probability to make a correct choice is $P(A) = 1/3$, hence $P(B) = 2/3$. We then need to define the conditional probabilities for the data that you'll be shown. Given that the prize is already in your box, the probability that an empty box is revealed to you is surely one: $P(\text{Monty shows empty'} \mid A) = 1$. But since Monty knows exactly the contents of all boxes, there will always be at least one empty box for him that he can reveal. So: $P(\text{'Monty shows empty'} \mid B) = 1$. Now we get $P(B \mid \text{'Monty shows empty'})$

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{2}{3}}{\frac{2}{3} + \frac{1}{3}} = \frac{2}{3}.$$

The posterior odds for Monty having the price is 2, compared to prior odds of 2, so that the revealing of an empty box is not giving information, other than simply reducing the number of

Monty's boxes by one empty box - which he always can do. But let's change the rules! Assume then that Monty is allowed to choose randomly (blindfolded) which one of his boxes he opens. (This could result into Monty's error of showing the price). Now we still have P(Monty shows empty' | A) = 1, but if the prize is in the other boxes, then P('Monty shows empty' | B) = 1/2. This will change the result:

$$= \frac{P(\text{'Monty shows empty'} \mid B)P(B)}{P(\text{'Monty shows empty'} \mid B)P(B) + P(\text{'Monty shows empty'} \mid A)P(A)} = \frac{\frac{1}{2}\frac{2}{3}}{\frac{1}{2}\frac{2}{3} + \frac{1}{3}} = \frac{1}{2}.$$

Posterior odds for Monty winning is then 1, compared to prior odds of 2, so the revealing of empty box makes the odds for Monty having the price smaller by a factor of 2. (Bayes factor 0.5). We really need to know how the game is played!

4.4 Fair coin or not

In 200 tosses of a coin, 115 were heads, 85 tails. The null hypothesis is to assume a fair coin: $H_0 : p = 0.5$. Alternatively, it is something else, $H_1 : p \neq 0.5$ in which case we apply a uniform prior distribution $U(0, 1)$. This can be seen as a model choice problem, where each hypothesis corresponds to a different model, M_0 and M_1. The Bayes factor is the posterior odds divided by prior odds

$$BF = \frac{P(M_0 \mid X)/P(M_1 \mid X)}{P(M_0)/P(M_1)} = \frac{P(M_0)P(X \mid M_0)/(P(M_1)P(X \mid M_0))}{P(M_0)/P(M_1)} = \frac{P(X \mid M_0)}{P(X \mid M_1)}$$

similarly to the case of simple (point) hypothesis (the point hypothesis is now the model). Now the probability of the data X under each model needs to be computed, to get:

$$BF = \frac{P(X \mid p = 0.5)}{\int_0^1 P(X \mid p)\pi(p \mid M_1)\mathbf{d}p}$$

The probability of data X under H_0 (model M_0) and under H_1 (model M_1) is

$$P(115 \mid M_0) = \binom{200}{115} 0.5^{200} = 0.005956$$

$$P(115 \mid M_1) = \int_0^1 \binom{200}{115} p^{115}(1-p)^{85}\mathbf{d}p = \frac{1}{201} = 0.004975$$

This gives $BF = 1.197$.

5 Other models

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

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

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

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 $\text{Gamma}(0, 1)$ - not proper. (With improper priors, always check if posterior distribution is proper).

Posterior predictive distribution can be analytically solved when the prior is conjugate (Gamma) distribution. The idea is the same as before (as always: follow the probability calculus):

$$P(X^{\text{next}} \mid X) = \int_0^\infty \text{Poisson}(X^{\text{next}} \mid \lambda)\text{Gamma}(\lambda \mid X)\mathbf{d}\lambda$$

where $\text{Gamma}(\lambda \mid X)$ is the posterior distribution of λ, based on earlier data X. By doing this integration we get Negative Binomial distribution. The mean and variance of NegBin distribution can be found as before with BetaBinom distribution, by using the law of total expectation and total variance:

$$E(X) = E(\underbrace{E(X \mid \lambda)}_{\text{from Poisson}}) = \underbrace{E(\lambda)}_{\text{from Gamma}} = \alpha/\beta$$

and

$$V(X) = E(\underbrace{V(X \mid \lambda)}_{\text{from Poisson}}) + V(\underbrace{E(X \mid \lambda)}_{\text{from Poisson}}) = \underbrace{E(\lambda) + V(\lambda)}_{\text{from Gamma}} = \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2}$$

In the posterior predictive distribution, the parameters α, β would correspond to the parameters of the posterior distribution for λ, shown above for a single observation X. Again, if we think of the Gamma distribution with just some parameters α, β, this can be used for **modeling overdispersed data** when the Poisson model as such is not adequate. (Poisson has mean $= \lambda =$ variance, which can be restrictive. NegBin distribution would depend on the two Gamma parameters α, β which allows more flexibility). In general, similar effects can be obtained by modeling the intensity λ of a Poisson process as a random parameter or some function over time, so that we have a *nonhomogeneous Poisson process* with intensity $\lambda(t)$. This is further specified depending on the 'second order' model structure ('=model of parameters') so that the intensity at one time point could depend on other intensities at other time points and other parameters - but then we have a more complex model, and the posterior distribution might not have a closed form anymore.

5.1.1 Example: asthma mortality

Epidemiological Example from Gelman [5]: Poisson model parameterized in terms of rate and exposure:

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

where X_i is the number of e.g. disease cases in a group with exposure E_i and θ is the unknown (common) parameter of interest, the 'underlying rate'. The 'exposure' E_i could be the person years in the ith group 'under risk' and X_i is then the observed disease cases as an outcome of that. The rate would then be interpreted as the underlying disease incidence per person per year.

(For rare diseases or 'once-in-a-lifetime' events, we do not expect more than one occurrence per person. Especially, if we model incidence of death. More exactly, such incidence for a person should be modeled as θ before the event, and zero after the event has occurred. Otherwise, the Poisson process with intensity θ assumes that any number of events is possible. In a large population with small number of cases X, it does not matter much if the diseased are subtracted after their onset of disease from the total exposure group or not. See more: *survival models* below).

In epidemiological applications, it is this underlying incidence θ we aim to estimate. A simple point estimate would be $\hat{\theta} = X_i/E_i$. For example, if the group i is the population of a town with 50000 inhabitants, and if the population does not vary significantly during a year, then $E_i \approx 50000$. If some disease count is $X_i = 30$ for some year, we get point estimate $30/50000$, or

'60 per 100000'. Assuming all individuals are 'exchangeable', then only the person years at risk matters. Similarly, if we estimate the failure rate of light bulbs: only the total time in use would matter, not the number of light bulbs. Two light bulbs in use for one year, then one burned, would give the same estimate (1 per 2×1 years $= 0.5$ per year) as one light bulb in use for two years, then burned.

The conditional probability of the data from N different groups (cities, etc.) $X = (X_1, \ldots, X_N)$ each with different exposures E_i is

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

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

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

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

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

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

Estimating several θ_i instead of one common θ:
The same idea is exploited further e.g. in spatial epidemiology, where we wish to estimate disease incidences θ_i in different geographical areas, instead of assuming there is a common incidence θ everywhere. The exposure E_i can be very small in some geographical areas due to low population. The risk estimates based on local data would be very unstable because - by chance - there could be ± 1 case, which already could cause a high/low point estimate. Therefore, results from small population groups are analyzed with respect to expected results based on the larger population. The former gives the likelihood part, the latter the prior $\rightarrow$ the result is a shrinkage towards the prior.

Examples are sometimes seen in daily news, reporting 'exceptional rise of crime rate in a small area'. With many small areas with small populations, most of them would show observed incidence of zero, but by chance some would have one or two cases which would give observed incidence much higher than the national incidence. This should not be interpreted as if the local risk is *really* that much bigger. Likewise, a small university department might be unfairly reported to score badly if for one year there happens to be no doctoral dissertations. Similarly, salmonella subtypes detected annually in some species could show so low counts that most types have zero frequency whereas few types would have only 1 or 2 counts. Does it mean that the other types are

absolutely nonexistent - just because they have not been detected this year? Or, in the cracking of the Enigma code: having not yet observed some bits of code does not mean that their probability is absolutely zero. A sensible prior distribution would put some positive probability for events not yet seen (if they logically can exist). Otherwise, the model would exclude their possibility completely and could not update those probabilities.

In spatial models in health geography, the disease rates (estimated Poisson intensities) are therefore **smoothed** over the map, so that any local estimate is taken as a compromise between the local data and its neighborhood (=prior). If the local data are heavy, the final estimate will be determined by that, but if the local data are scarce, the estimate is more influenced by neighborhood. This is also called borrowing strength. The same is often done with time series data, to get smoothed temporal rates. In other applications, we might have a stratification of a population so that the population counts in some strata are too low, hence the estimates would be smoothed towards other data. Smoothing can be either towards neighborhood mean or towards global mean.

Different alternatives would be built in the posterior of the local disease rates:

$$\pi(\theta_1, \ldots, \theta_N \mid X_i, \ldots, X_N) \propto \prod_{i=1}^{N} \pi(X_i \mid \theta_i)\pi(\theta_i \mid \text{something})$$

by choosing the definition of prior $\pi(\theta_i \mid \text{something})$ as needed:

either 1: the prior of each disease rate θ_i (of $i = 1, \ldots, N$ subpopulations) is based on background information about larger population. (With this prior, we expect disease rates in different geographic areas to be as the national rate is - or as the rate in surrounding area - or whatever reference group thought to be relevantly informative). Here, the prior is given as a fixed distribution.

or 2: the prior can be further generalized to allow prior parameters to be unknown, so that they are estimated too as part of a posterior distribution. This is accomplished by letting the prior to be e.g. $\pi(\theta_i \mid \mu)$ with a further prior $\pi(\mu)$ which could represent global mean, around which local means θ_i are distributed (θ_i are conditionally independent, given μ). In turn, local observed disease counts X_i would be conditionally independent, given θ_i. This makes a hierarchical model where the smoothing depends on the unknown parameters in $\pi(\theta_i \mid \mu)$. A 'tight' prior makes shrinkage towards the global μ (which itself is uncertain) whereas a 'loose' prior lets each θ_i to be close to the observed incidence X_i/E_i.

or 3: same as before, but the prior of θ_i could depend on other θ_j, so the prior could e.g. specify θ_i to be distributed around a local mean of other θ_j, $j \neq i$ in the neighborhood of the ith geographical area. Or in temporal models: conditionally on the θ_{i-1} of the previous time interval.

In any case, the structure of smoothing becomes specified by the choice of prior.
(•) In the first case, a prior for each parameter is set independently of other parameters and it is a fixed distribution: $\pi(\theta_i)$. This model cannot borrow strength from other data because the prior is a fixed distribution and cannot change. Every θ_i is estimated separately from the others.
(•) In the latter cases, prior of θ_i depends on a parameter(s) that is also estimated, and this common parameter will get estimated from information about all other θ_j too, which finally are informed by their specific data X_i. Every θ_i is estimated with influence from other θ_j as well, via the common parameter(s) which need 'hyper prior': a prior for the parameters of a prior.

Large models could have several layers of priors, describing the structural assumptions of conditional dependency we make.

5.2 Exponential distribution

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

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

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

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

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

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

which leads to the posterior Gamma$(\alpha + n, \beta + n\bar{X})$, so that the Gamma$(\alpha, \beta)$ prior can be thought as equivalent of 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 data sum $\sum X_i$, respectively.

5.2.1 Survival models

Consider a non-homogeneous Poisson process with intensity (hazard) function $\lambda(t)$. The definition of hazard is the limit of a conditional probability

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(t < X \leq t + \delta t \mid t < X)}{\delta t}$$

In other words:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$$

where $f(t)$ is the probability density function of the event time $X \in [0, \infty)$ and $S(t) = 1 - F(t)$ is the *survival function*, (F is the cumulative probability function).

Solving the differential equation $\lambda(t) = -S'(t)/S(t) = -\,\mathbf{d}\log(S(t))/\mathbf{d}t$ gives

$$S(t) = e^{-\int_0^t \lambda(\tau)\mathbf{d}\tau}.$$

If the intensity is just constant $\lambda(t) = \lambda$, this results to the cumulative probability function of exponential distribution

$$S(t) = e^{-\lambda t}$$

so that the event time X is exponentially distributed, conditionally on λ. An observed data of several event times (life times, failure times, etc.) gives the likelihood function

$$\prod_{i=1}^N e^{-\lambda X_i} \lambda^N = \lambda^N e^{-N\bar{X}\lambda}$$

which can be seen as the product of probabilities to (1) 'survive' up to time X_i and (2) to 'die' at time X_i, for all data points $X_1, \ldots, X_N$ representing the event times for individuals. *Each individual contributes to the likelihood only for the length of time he/she survived.* If all these event times are observed, and with Gamma-prior for λ, bayesian inference for the intensity is straightforward, leading to a Gamma-posterior. Typically, some of the event times are not observed. This is called censoring.

5.2.2 Censored data

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

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

Applying the Gamma(α, β)-prior, the posterior is then Gamma$(\alpha+k, \beta+\sum_{i=1}^{k} X_i+(n-k)T)$.

More generally, we may know that for some individuals the event occurred before some given time, or between two given times. In each case, this information should be included by writing the corresponding conditional probability. (This is sometimes called as the 'full likelihood'). For example, if some events are only known to have been before time T_1 and some are known to be after time T_2, and for the rest we know the exact time, then **the full likelihood** would be of this form

$$P(X \mid \lambda) = \prod_{i \in E_1} \underbrace{F(T_1 \mid \lambda)}_{P(X_i<T_1|\lambda)} \times \prod_{i \in E_2} \underbrace{S(T_2 \mid \lambda)}_{P(X_i>T_2|\lambda)} \times \prod_{i \in E_3} \underbrace{\lambda \exp(-\lambda X_i)}_{P(X_i|\lambda)}.$$

and this could still be expanded by interval censored data, by incorporating probabilities of the type $P(L_1 < X_i < L_2 \mid \lambda)$. All these censored observations give likelihood contributions involving integrations of the density function of X_i.

Whatever the expression of the full likelihood, the principle is generally the same: to compute posterior distribution, conditionally to 'full data' (=censored and exact event times):

$$\pi(\lambda \mid \text{full data}) \propto \pi(\lambda) P(\text{full data} \mid \lambda)$$

which might take a form that does not reduce to a known density function in closed form!

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

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

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

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

Censoring can happen also with other variables than event times. In microbiology, bacterial concentrations are measured and some (or many) measurements can be censored. As an example of what a heavily censored data could look like, the following represents observed bacterial concentrations in samples. There is only one exactly observed measurement. All others are somehow censored:

Number of samples	Concentration (CFU/g (Colony Forming Units per gram))
54	<0.04
2	<100
26	0.04-10
1	15
8	0.04 -100
2	>100
1	<1
1	>1
7	0.04 -1
1	1-100

Reference: P Busschaert, AH Geeraerd, M Uyttendaele, JF Van Impe. Estimating distributions out of qualitatite and (semi)quantitative microbial contamination data for use in risk assessment. International Journal of Food Microbiology. 138 (2010), 260-269.

6 Approximating posterior density

Posterior distributions could be approximated by finding out the posterior mean and variance, and then using normal distribution

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

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*

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

The approximation is based on Taylor series expansion of $\log \pi(\theta \mid X)$ centered at the posterior mode, $\hat{\theta}$. For a scalar valued θ this is

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

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

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

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.

7 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.

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

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

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

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

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

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

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

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

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

Again, prior parameters $\alpha_1, \ldots, \alpha_k$ can be interpreted to represent 'prior data' so that the 'prior sample size' is $\sum \alpha_i$. A usual uninformative prior choice is $\text{Dir}(1, \ldots, 1)$, which is the generalization of $\text{Beta}(1, 1)$. The posterior means can be written as weighted mean of prior and data proportions

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

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

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:

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

7.2 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 $\mathbf{N}(\mu, \sigma^2)$ can be casually written as $\mathbf{N}(\mu, \tau)$ 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'.

7.2.1 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. [9] 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: $\mu \sim \mathrm{N}(\bar{X}, \sigma^2/N)$

7.2.2 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 $\pi(\sigma^2) \propto 1/\sigma^2$.

7.2.3 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: $\text{N}(\mu_0, c\sigma^2)$. 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, [5]. As a result, the marginal posterior is found to be a t-distribution so that

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

8 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 condition al 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 was 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 drawn a sample $\theta_1, \ldots, \theta_n$ generated from the posterior distribution $\pi(\theta \mid X)$. Then,

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

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

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

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

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

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

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

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

$\Rightarrow$ 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 the technical material about analytical solutions of posterior distributions given earlier for binomial, multinomial, poisson, gamma, and normal models.

So, for e.g. binomial model $\text{Bin}(N, r)$ with observed data x, and 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.
(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.

8.1 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 to 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.

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.

Note also: with all these Monte Carlo methods in Bayesian applications, the target distribution is the posterior distribution. It is not about simulating the biological random process. It is about simulating values according to the posterior distribution, i.e. describing our uncertainty distribution, given the observations we had from the biological process. An observation can be missing or predicted, in which case it becomes one of the unknown parameters among others, to be simulated by MCMC from the joint posterior distribution of all unknowns.

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 to 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

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

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 $\text{Beta}(X+1, N-X+1)$. Likewise, (2) if keeping p fixed in the second expression, you have a distribution function for X, proportional to $\text{Binomial}(N, p)$. 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 $\text{U}(0, 1)$ 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 condition to 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 be very inefficient and clumsy sampler, though. (But sometimes seen in some applications!). In practice, it is best to keep your data fixed to what it was, 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)
points(c(x[i-1],x[i-1]),c(p[i-1],p[i]),'l')
```

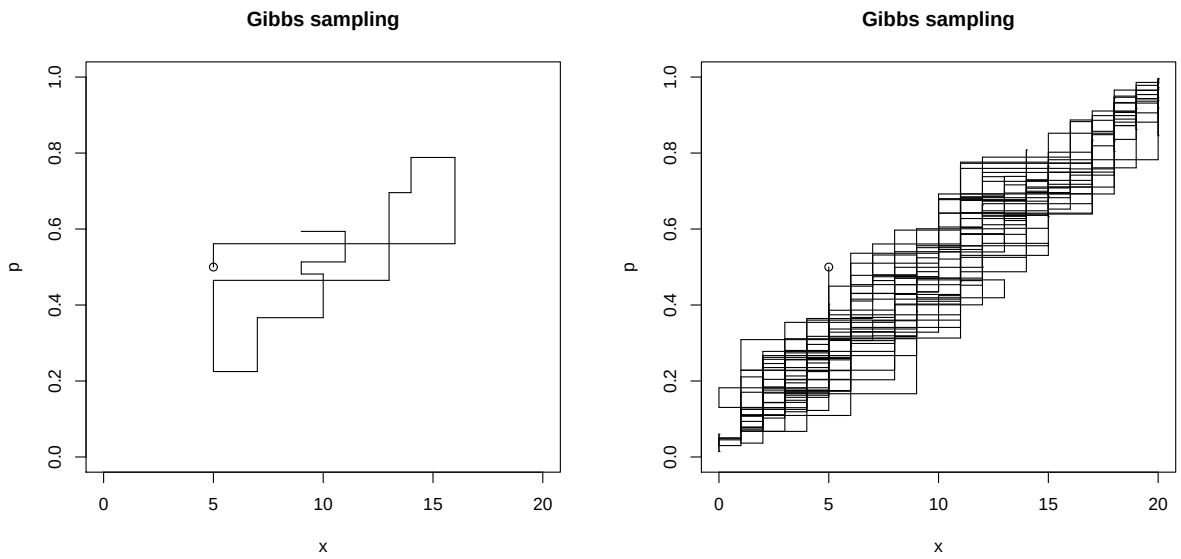


Figure 10: Sample path of Gibbs sampling with $\pi(X, p \mid N)$

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

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 Markov Chain **that will only converge** to the correct target distribution. But the Markov chain 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).

8.2 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. 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 that is further developed, found at: http://www.openbugs.info/w/

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 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 by click-and-point the model code, data and inits.

Advice for new users:

1. Step through the simple worked example in the tutorial.
2. Try other examples provided with this release
(see Examples Volume 1 and 2, also Vol 3 in OpenBUGS)
3. Edit the BUGS language to fit an example of your own.

It is easiest to take existing code for a simple model and modify that for your purpose. It as been recommended that 'users should already be aware of the background to bayesian Markov chain Monte Carlo methods'. That's why it was included in the introduction part.

The current Metropolis MCMC algorithm is based on a symmetric normal proposal distribution, whose standard deviation is tuned over the first 4000 iterations in order to get an acceptance rate of between 20% and 40%. All summary statistics for the model will ignore information from this adapting phase. In OpenBUGS, the samplers have been further developed and this process is expected to continue. WinBUGS will no longer be updated. Hence, version 1.4.3 will be the last of WinBUGS.

Strong recommendation: the first step in any analysis should be the **construction of a directed graphical model**. Briefly, this represents all quantities as nodes in a **Directed Acyclic Graph (DAG)**, in which arrows run into nodes from their direct influences (parents). The model represents the assumption that, given its parent nodes pa[v], each node v is independent of all other nodes in the graph except descendants of v, where descendant has the obvious definition. This visualization of the model is very useful for presenting the model in a glance.

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 denoted as rectangles in the graph. They must be specified in a data file.

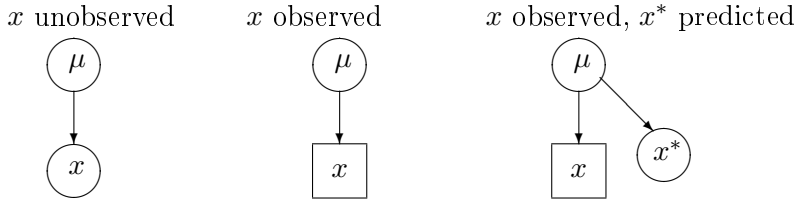
2. **Stochastic nodes** are variables that are given a conditional distribution, and are denoted as ellipses in the graph; they may be parents or children (or both). Stochastic nodes may be observed in which case they are data, or may be unobserved and hence be parameters, which may be unknown quantities 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 ~ 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 values in a data file, in which values for constants are also given.

Example: $x \sim \mathrm{N}(\mu, 1)$ with prior $\mu \sim \mathrm{N}(0, 10^4)$. The DAG would run from stochastic node μ (parent of x) to stochastic node x (child of μ). The latter would be assigned some data value, which leaves only μ as an unknown parameter for which the posterior will be computed, given data value for x. A deterministic node could be added by defining e.g. $logmu \leftarrow \log(\mu)$, for monitoring the MCMC samples for the log of μ, if its posterior happens to be of any interest. The constant value of 1 for the variance in the conditional model of x could be given as a fixed parameter τ which then needs to be given also as data, $\tau = 1$. This would be a founder node in the DAG since there would be no parents for it. To summarize: these specifications together are needed for constructing the MCMC sampler (inside WinBUGS/OpenBUGS) for computing the posterior $\pi(\mu \mid x)$. If x is not given a data value, it too remains an unknown stochastic node. Effectively, it would then be simulated from the prior predictive distribution, whereas μ would be simulated from its prior only. After observing x, the posterior predictive distribution $\pi(x^* \mid x)$ for a new, x^*, observation can be computed simply by adding x^* as an unobserved node in the graph, with the same conditional distribution as there was for x, given μ.

Some Directed Acyclic Graphs (DAG):



8.3 Steps of installing WinBUGS/OpenBUGS

Go to the website (either Win- or Open-) and follow instructions - it usually works. Because the development of WinBUGS is not to be continued, its compatibility with other new software in the future is not sure. New updates will appear for OpenBUGS. If installing WinBUGS, you should get version 1.4 which is then upgraded to 1.4.3 by installing a patch as instructed. Also, a keyfile was required for getting the fully functional version. This keyfile used to expire at the end of the year, and new keyfiles were mailed to registered users. However, finally an 'immortal' key was given for all users, now from the Website. In OpenBUGS these steps are not involved. You just install the latest version of OpenBUGS.

Installation in Windows machines has usually been straightforward. Some difficulties(?) may occur with Linux/Mac, but it is possible to run WinBUGS from Mac and OpenBUGS from Linux (and also WinBUGS via emulators). For detailed instructions with each platform, you should carefully read the installation instructions provided in the websites. For example, they say:

Note: There appears to be a problem with installing WinBUGS and/or various patches in Windows Vista. Vista doesn't seem to like anyone overwriting files in the "C:\Program Files" directory (regardless of permissions). Hence we recommend that WinBUGS be installed elsewhere, e.g. "C:\".

If all else fails (for example with a 64-bit machine), you can download a zipped version of the whole file structure and unzip it into Program Files or wherever you want it. WinBUGS makes no changes to the Registry.

I have also installed WinBUGS on a memory stick. Seems to be running!

8.4 Steps of running WinBUGS/OpenBUGS models

Assume you have an existing BUGS model code in a file ('.odc' or '.txt'). Assume also that the data list is included in the same file (a list written below the model code).

1. Open the file in WinBUGS/OpenBUGS
2. Open `Model > Specification...`
3. Check the model's syntax
4. Load data
5. Compile model
6. Set initial values (from a preset list, or let the software generate them)
7. Run the model `Model > Update...`
8. After convergence, set parameters of interest for monitoring. `Inference > Samples...`
9. Run again the model to get sufficiently large sample
10. See the output graphically (`history, density`), and summary statistics (`stats`)

8.5 Structure of the model

The syntax and form of a model in WinBUGS follows (nearly) directly from the structure of the required densities in the Bayes formula. In OpenBUGS, the structure of the language is the same, with some added new functions or distributions which gradually may evolve. Other features have also emerged in OpenBUGS (see the Webpage for details), including the possibility to get your

model code printed with LaTeX commands. However the core of the model specification language is still the same. Understanding of the product rule and bayes formula, as well as other basic theorems of probability calculus is as essential as understanding grammatical rules and structure of sentences in natural languages. We need to specify a conditional distribution of data, and a prior. These can consist of several conditional distributions. The whole structure is convenient to draw as a Directed Acyclic Graph (DAG) which you can frequently find in BUGS examples. (There are some conventions to draw different arrows for stochastic dependencies and deterministic dependencies, etc.). This makes a visual expression of the **logical structure** for a complete specification of a joint probability model. It is this logical structure we need to code for WinBUGS/OpenBUGS.

The joint posterior density is always fully specified when all these necessary parts are defined. Therefore, WinBUGS is a **declarative** language, as opposed to **procedural** programming languages. (**This is important to remember**). In a procedural language the following code could be valid:

```
X := 1;
Y := 1;
Z := X+Y;
```

but the following would not compute procedurally:

```
Z := X+Y;
X := 1;
Y := 1;
```

In WinBUGS/OpenBUGS, the order of these statements would not matter, because only the logical structure is defined which can be *written out* in any order, as long as all the quantities are defined somewhere and their combination defines a valid model. That is: *all the conditional distributions and priors needed in the Bayes formula!*

Therefore, in WB, we define a chain of conditional distributions that was obtained from the product rule when writing the Bayes formula. Each variable $v \in V$ in the set of all variables in the model can be a 'child node' that is conditionally dependent on its 'parent nodes'. And these 'parent nodes' can be 'child nodes' of other nodes, etc. By applying the product rule in the Bayes formula, a joint distribution of all variables V is broken into a product of conditional distributions:

$$\pi(V) = \prod_{v \in V} \pi(v \mid \text{parents}\{v\}),$$

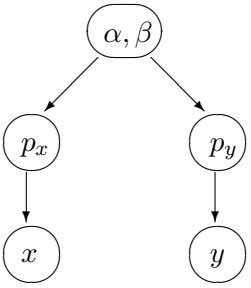
and the last variables in this chain have no further parents, i.e. their conditional distribution does not depend on any further variables - these have only the prior distribution. The whole structure specifies a bayesian model. A simple example (assuming data x, y, n, m) could be:

```
model{
x ~ dbin(px,n)
y ~ dbin(py,m)
px ~ dbeta(a,b)
py ~ dbeta(a,b)
a ~ dexp(1)
b ~ dexp(1)
}
```

or a linear model (assuming data x, y):

```
model{
for(i in 1:N){
 y[i] ~ dnorm(mu[i],tau)
 # note that: tau = 1/sigma^2
 mu[i] <- alpha + beta * (x[i]-mean(x[]))
 }
alpha ~ dnorm(0,0.0001); beta ~ dnorm(0,0.0001)
tau ~ dgamma(0.001,0.001)
}
list(N=5,x=c(1,2,3,4,5),y=c(1,3,3,3,5))
```

Comment lines can be written, starting with '#', and sufficient commenting is indeed recommended! In the binomial model above the DAG would be:



A DAG is really the visualized skeleton that gives the necessary structure for a valid bayesian model as well as for a valid WB model code. (The exception is that in WB we can also define Gibbs sampling algorithm directly using 'full conditionals', but this is not a typical way of using it).

Since cycles are not allowed in a DAG, then how to define models where some variable has some feedback into itself? For example, the size of a population drives population growth which again determines the population size. The question is: what is the probability model for this? It is a model of a stochastic process. The variables need to be indexed with respect to time, so that the conditional distribution of X_{t+1} depends on X_t, and this can be written as a DAG without cycles. Alternatively, (but this can be more complicated), we could try to solve the conditional probability distribution for the whole set of values $X_1, \ldots, X_t$, given some other parameters of the model. But if the X variables are unknown, their simulation might require block updating which is not possible in WinBUGS which is based on single site updating algorithms (unless there are extensions available). In General, Gibbs sampling theory allows block updating if we can just solve what the full conditional density for a block (vector of parameters) is.

doodle BUGS

Models can be defined in WB either by writing the corresponding WB code, or by drawing the DAG using doodle-BUGS. Once the 'doodle' is defined, the corresponding WB code is automatically generated. But the opposite is not possible: a picture of a DAG cannot be generated from winBUGS code, you need to draw it elsewhere. But the WB language is much more versatile than doodle-BUGS, so it is best to learn to write WB codes, and do drawing of DAGs elsewhere.

8.6 Logical expressions

Long expressions can sometimes get too long for WinBUGS to compile. You would then get the error message: "logical expression too complex". This can be avoided only by using additional variables:

```
a <- g + t + g*u + 7*pow(w,s)
b <- r + sqrt(h) - inprod(z[],zz[]) + e/p
c <- a+b
```

instead of writing the whole expression as a one-liner. Some useful logical functions in WinBUGS are (there are more in OpenBUGS):

```
abs(e)
equals(e1,e2)
step(e)
exp(e)
log(e)
inprod(v1,v2)
inverse(v)
max(e1,e2)
min(e1,e2)
ranked(v,s)
mean(v)
sum(v)
sd(v)
phi(e)
pow(e1,e2)
sqrt(e)
```

Note: there is no function for computing a product. (Except, in OpenBUGS there is). But the summation can be used for doing this, by taking logarithms: $ab = \exp(\log(ab)) = \exp(\log(a) + \log(b))$.

Note: these logical functions calculate a deterministic value that must be assigned to some variable, "$\leftarrow$", and those variables must not have an assigned value as data, nor initial value. That would lead to 'multiple definition' errors.

difficulty of IF-THEN programming

Sometimes we wish to have IF-THEN -structures in model code, but these are not part of BUGS syntax in the same way as in procedural languages. Remember, BUGS is a declarative language. Therefore, if we need something like this:

$$\begin{aligned} \text{if } y = 1 \quad & \text{then } x \sim N(\mu_1, 1) \\ & \text{else } x \sim N(\mu_2, 1), \end{aligned}$$

it has to be coded, for example, as:

```
x[1] ~ dnorm(mu[1],1)
x[2] ~ dnorm(mu[2],1)
z <- equals(y,1)*x[1] + (1-equals(y,1))*x[2]
```

But we could not use the above when z is given as observed data value. (Data should always be assigned to a stochastic node, defined by $\sim$. A logical node $\leftarrow$ within the code and a definition in the data listing for the same variable would cause multiple definition error). Therefore, if we want to compute posterior distribution of anything conditional to observed z, we should write something like

```
z ~ dnorm(par,1)
par <- mu[1]*equals(y,1)+mu[2]*(1-equals(y,1))
```

or using nested indexing, if y is binary variable:

```
z ~ dnorm(mu[ind],1)
ind <- y+1
```

Variable y could be indicator variable of the alternatives, for which we set Bernoulli(θ)-model with unknown parameter θ. This would correspond to probability of z being from the first model, $\mathrm{N}(\mu_1, 1)$. Also `step`-function could be used. This would make a mixture model for z:

$$\pi(z \mid \theta, \mu_1, \mu_2) = \theta \mathrm{N}(\mu_1, 1) + (1 - \theta)\mathrm{N}(\mu_2, 1)$$

where we need to choose prior distributions for all parameters θ, μ_1, μ_2. Moreover, multiple logical choices could be implemented by using `categorical`-distribution:

```
a ~ dcat(p[])
z ~ dnorm(mu[a],1)
```

Another possible difficulty

Sometimes you may need to compute an expression that is undefined for some values, e.g. $1/X$. Now, if X has e.g. Poisson distribution model $X \sim \mathrm{Poisson}(\theta)$, and θ is given as constant or random, it can happen that for some iterations $X = 0$. Maybe it is not sensible to define such variables that may lead to this problem anyway. But it could happen, for example if we have random N in a binomial model $X \sim \mathrm{Bin}(N, p)$ with $N \sim \mathrm{Poisson}(\lambda)$ and then try to define `Y <- X/N`, which could happen to be $0/0$.

Computing $1/X$ with the possibility of $X = 0$ would lead to runtime error. How to avoid this? If we had a procedural language, we could use IF-THEN structure by first calculating the value of X and only then choose what to calculate (or not) after knowing what X value was. But in declarative language we need to express a definition using only logically structured expressions. For example:

```
Y <- equals(X,0)*(-9) + (1-equals(X,0))*(1/X))
```

It would seem that this solves the problem by setting an arbitrary value of -9 whenever $X = 0$, and only calculate $Y = 1/X$ when $X \neq 0$. But WinBUGS produces an error, because it calculates also the case $1/0$ even though it would be multiplied by zero (which still would be undefined $0/0$). This does not work! The only hope is to replace X by $X + \epsilon$ with a very small value for ϵ to ensure that WinBUGS can compute $1/(X + \epsilon)$ for all values of X. But this introduces small error when $X > 0$ (which may be insignificant). Alternatively, we really need to think over the modeling, and redefine a new model which does not involve even the possibility of $1/0$.

8.7 Data structures

Anything that is not an unknown (random) quantity in the model, has to be fixed value, i.e. given as data or constant. Data are listed separately from the model code, for example:

```
list(x=4,
     y=c(3.5,7.2,9.1),
     z=structure(
       .Data=c(7,3,5,1,8,2),
       .Dim=c(2,3)))
```

which defines a scalar x, vector y and a matrix z of size 2×3. Data matrices can also be defined in this form:

```
 z[,1] z[,2] z[,3]
 7     3     5
 1     8     2
 END
```

so that first index of z needs to be left empty, and there must be an empty line after `END`. You can avoid much trouble if you always check carefully that you have indexed your data correctly. There are no useful tools for checking data inconsistencies within WinBUGS. When data variables have been defined and assigned, there should be a conditional distribution for them in the code. For example, if variable y is given as data, then we might have a model directly for it:

```
y ~ dnorm(mu,tau)
```

Alternatively, we might be interested in modeling some transformation of this variable, which could be done as:

```
yy <- log(y)
yy ~ dnorm(mu,tau)
```

Of course, we might have calculated the transformed y already beforehand, and then use that as data. Note that the previous use of transformations within the code is actually against WB syntax which prohibits multiple definitions of the same node. This is the exception to the rule. Otherwise, you get error messages: 'multiple definition of yy'. An error would be caused also if variable y was a vector with values given in the data, and some values would be missing (`'NA'`). The missing values would then be stochastic nodes and the above construction would lead to error.

9 Some BUGS models

Uncertainty with diagnostic testing

A diagnostic test has sensitivity $q_1 = P(\text{test} + \mid \text{true disease})$, and specificity $q_2 = P(\text{test} - \mid \text{truly no disease})$. Developers of the test have tested individuals that were first confirmed to be healthy or diseased. These data can be used independently to compute posterior density of both parameters. Having this information, we can simultaneously compute posterior distribution of

population prevalence from a sample whose testing results are observed. Priors for the three parameters are Uniform$(0,1)$. Run this model in OpenBUGS. Conditional distributions of data are:

$$X_1 \sim \text{Binom}(N_1, q_1) \quad , N_1 = 50, X_1 = 45$$

$$X_2 \sim \text{Binom}(N_2, q_2) \quad , N_2 = 30, X_2 = 28$$

$$Y \sim \text{Binom}(M, pq_1 + (1-p)(1-q_2)) \quad , M = 100, Y = 10$$

```
model{
x[1] ~ dbin(q[1],N[1])
x[2] ~ dbin(q[2],N[2])
y ~  dbin(pr,M);  pr <- p*q[1]+(1-p)*(1-q[2])
q[1] ~ dunif(0,1); q[2] ~ dunif(0,1); p ~ dunif(0,1)
}
list(x=c(45,28),N=c(50,30),M=100,y=10)
```

Comparison of two populations

Problem 1: to study whether the prevalence in one population is smaller than the prevalence in another population, based on a sample from both. The goal is to compute $P(\theta_1 < \theta_2 \mid \text{data})$. This is straightforward to simulate also in R. Assuming uniform prior density for both prevalence parameters, we draw samples from two beta-distributions, and calculate percentage of simulated values satisfying the requirement $\theta_1 < \theta_2$. In BUGS:

```
model{
x[1] ~ dbin(theta[1],N[1]); theta[1] ~ dunif(0,1)
x[2] ~ dbin(theta[2],N[2]); theta[2] ~ dunif(0,1)
# calculate average of T in the simulations:
T <- step(theta[2]-theta[1])
}
list(x=c(3,7),N=c(30,30))
```

Problem 2: similar as before, but now with normally distributed measurements. Here we compare population means

```
model{
# below only one measurement from both populations, but
# this could be expanded in a for-loop, to have x[i], y[i]
x ~ dnorm(theta[1],tau[1]); theta[1] ~ dnorm(0,0.001)
y ~ dnorm(theta[2],tau[2]); theta[2] ~ dnorm(0,0.001)
# calculate average of T in the simulations:
T <- step(theta[2]-theta[1])
}
list(x=3,y=4,tau=c(1,1))
```

Identifiability of parameters: an example

Consider two measurements $X_1 \sim \text{N}(\mu_1, 1)$ and $X_2 \sim \text{N}(\mu_2, 1)$. If we only observe the sum $Y = X_1 + X_2$ what can we infer about μ_1 and μ_2? To write a BUGS model, we have to invent a way to write Y as a stochastic node, so that it has a conditional distribution with some parameters. From

probability theory of normal distributions we know that if $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are conditionally independent, then $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. This is our model. Assuming σ_i^2 are known, we aim to compute posterior distribution of μ_1 and μ_2. Clearly, these are not uniquely identified from data Y. The likelihood function is constant along a line $\mu_1 + \mu_2 = c$, so the individual parameters can take any values that are 'just as likely' combinations. A likelihood inference would conclude without a unique maximum likelihood estimate. Maximization algorithms would be trapped to never ending loops. BUT: bayesian inference could still get a solution, but this depends on priors. The priors must be informative in this case. The solution is sensitive to the choice of prior. **This is the danger with Bayesian methods: it is not always easy to notice identifiability problems because they might be covered up by the choice of prior**. If the problem remains in the posterior distribution, there should be also problems of convergence in MCMC sampling. Try the BUGS model:

```
model{
s ~ dnorm(mu,tau);  mu <- sum(m[1:2]);  tau <- 1/(sum(v[1:2]))
for(i in 1:2){ m[i] ~ dnorm(0,0.001);  # or more informative dnorm(0,1)
               v[i] <- 1/t[i]; t[i] <- 1
             }
s <- 3
}
```

Another identifiability problem: Eyes, WinBUGS Examples Vol2

Assume a set of observations X_i, some of them are from $\mathrm{N}(\theta_1, \sigma_1^2)$ and some are from $\mathrm{N}(\theta_2, \sigma_2^2)$. In this example, the mixture model likelihood becomes:

$$w\pi_1(X \mid \theta_1) + (1-w)\pi_2(X \mid \theta_2) = (1-w)\pi_1(X \mid \theta_2) + w\pi_2(X \mid \theta_1).$$

The likelihood function takes the same value if we switch the 'labels' (indexing) of θ and reverse the weights $(1-w)$ and w. This type of unidentifiability is called '*label switching problem*', or '*aliasing*'. The switching is eliminated by setting suitable constraints, in the implementation below it is ensured by defining θ_2 to be strictly greater than θ_1. Using the latent (unobserved) indicator variables, the model is:

$$\begin{array}{lcl} Z_i & \sim & \text{Bernoulli}(w) \\ X_i \mid Z_i & \sim & \pi(X_i \mid Z_i) \\ \theta_1 & \sim & \text{prior} \\ \theta_2 & \sim & \text{prior} \\ w & \sim & \text{prior} \end{array}$$

and in BUGS code of the Eyes-example:

```
model
    {
        for( i in 1 : N ) {
            y[i] ~ dnorm(mu[i], tau)
            mu[i] <- lambda[T[i]]
            T[i] ~ dcat(P[])
        }
        P[1:2] ~ ddirch(alpha[])
```

```
    theta ~ dnorm(0.0, 1.0E-6)I(0.0, )
    lambda[2] <- lambda[1] + theta
    lambda[1] ~ dnorm(0.0, 1.0E-6)
    tau ~ dgamma(0.001, 0.001) sigma <- 1 / sqrt(tau)
}
```

Linear model: effect of standardization of covariates

Normal linear model is given by

$$y_i \sim \mathrm{N}(\beta X_i, \sigma^2)$$

where the expected value of observation y_i $(i = 1, \ldots, n)$ is a linear sum of the effects of explanatory variables $\beta X_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$, so that X_i is a set of k explanatory variables for individual i. In this model, we have $k+2$ unknown parameters for which we need to specify a prior. Often, uninformative priors are sought. One possibility is to use $\pi(\beta, \sigma^2) \propto 1/\sigma^2$ (improper prior). In the normal linear model with uninformative prior, the posterior density of regression parameters becomes (**conditionally to** σ) the following multivariate normal density:

$$\pi(\beta \mid y, X, \sigma) = N\Big(\underbrace{(X^T X)^{-1} X^T y}_{\text{mean vector}} , \underbrace{(X^T X)^{-1} \sigma^2}_{\text{cov. matrix}} \Big).$$

Typical BUGS-model of this linear model could be e.g.

```
model{
for(i in 1:n){
y[i] ~ dnorm(mu[i],tau)
mu[i] <- beta0 + beta1*x[i]
# mu[i] <- beta0 + beta1*(x[i]-mean(x[])) # standardized covariates
# or with more variables:
# mu[i] <- beta[1] + beta[2]*x[i,1] + beta[3]*x[i,2]
}
for(i in 1:k){
beta[i] ~  dnorm(0,0.001)
}
tau ~ dgamma(0.01,0.01)
# prediction with given value xnew:
ynew ~ dnorm(munew,tau); munew <- beta0 + beta1*xnew
# munew <- beta0 + beta1*(xnew-mean(x[])) # standardized covariates
}
list(y=c(41,52,18.7,55,40,29.2,51,17.6,46.6,57),
x=c(23.9,43.3,36.3,40.6,57,52.5,46.1,142,112.6,23.7),xnew=50)
```

In the above code, a prediction of `ynew` is included for a given covariate value `xnew` given in the data listing. This is the BUGS implementation of posterior predictive distribution $\pi(y_{\text{new}} \mid y_1, \ldots, y_n, x_1, \ldots, x_n, x_{\text{new}})$ which is **basically the same thing as before** in the examples where the predictive distribution could be analytically solved. In this case, the prediction is just based on the posterior distribution of all model parameters, and the assumed linear model with a given value of `xnew`.

Alternatively to the basic model $\beta_0 + \beta_1 x_i$, we can use standardized covariates: $\beta_0 + \beta_1(x_i - \bar{x})$ where $\bar{x} = \frac{1}{n}\sum_i x_i$. It can be useful to standardize the explanatory variables before modeling because this has an effect on the posterior covariance of parameters β. With the data below, compute posterior means in R according to the above multivariate normal density, and compute the covariance matrix (assume $\sigma = 1$) under original x variables, and under standardized variables $x_s = x - \bar{x}$:

```
y <- c(41,52,18.7,55,40,29.2,51,17.6,46.6,57)
x <- c(23.9,43.3,36.3,40.6,57,52.5,46.1,142,112.6,23.7)

X <- matrix(1,10,2)
for(i in 1:10){X[i,2]<-x[i]}
# posterior means for regression parameters beta:
betahat <- ((solve(t(X)%*%X))%*%t(X))%*%y
# covariance matrix of beta, assuming sigma=1:
C <- solve(t(X)%*%X)
```

This gives

$$\mathrm{C} = \begin{bmatrix} 0.346762974 & -4.269256e-03 \\ -0.004269256 & 7.386255e-05 \end{bmatrix}$$

```
# the same with standardized x:
xs <- x-mean(x)

Xs <- matrix(1,10,2)
for(i in 1:10){Xs[i,2]<-xs[i]}
# posterior means for regression parameters beta:
betahats <- ((solve(t(Xs)%*%Xs))%*%t(Xs))%*%y
# covariance matrix of beta, assuming sigma=1:
Cs <- solve(t(Xs)%*%Xs)
```

This gives

$$\mathrm{Cs} = \begin{bmatrix} 1.0000e-01 & -2.099300e-19 \\ -2.0993e-19 & 7.386255e-05 \end{bmatrix}$$

which is nearly a diagonal matrix, so that the correlations are now almost vanished. They should be exactly zero but the computer has limited accuracy. Analytically, we get

$$X_s^T X_s = \begin{bmatrix} n & \sum(X_{i2} - \bar{X}_{,2}) \\ \sum(X_{i2} - \bar{X}_{,2}) & \sum(X_{i,2} - \bar{X}_{,2})^2 \end{bmatrix} = \begin{bmatrix} n & 0 \\ 0 & \sum(X_{i,2} - \bar{X}_{,2})^2 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 13538.66 \end{bmatrix}$$

Whereas computer gives something like

$$X_s^T X_s = \begin{bmatrix} 1.000000e+01 & 2.842171e-14 \\ 2.842171e-14 & 1.353866e+04 \end{bmatrix}$$

Anyhow, we can expect to obtain a posterior distribution of β with minimal correlations. (*This can be very useful in MCMC simulations of β such as Gibbs sampling*). The parameters of the standardized model and the original model are related, because:

$$\mu_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) = \underbrace{\beta_0^* - \beta_1^*\bar{x}}_{\beta_0} + \beta_1^* x_i = \beta_0 + \beta_1^* x_i$$

Hence, with this standardization, β_j for $j \geq 1$ would remain the same, but β_0 of the original model would correspond to $\beta_0^* - \beta_1^* \bar{x}$ in the standardized model.

For curiosity, in BUGS, we could compute the same matrix, but since it is a function of data, it is a constant and cannot be monitored as an uncertain parameter. Its values can be checked using `Info` → `Node info` → `values`.

```
model{
for (i in 1:n) {
y[i] ~ dnorm(mu[i],tau.y)
xx[i] <- (x[i]-mean(x[]))
mu[i] <- a + b*xx[i] # standardized x
#mu[i] <- a + b*x[i]
X[i,1] <- 1
X[i,2] <- xx[i]
for(t in 1:2){XT[t,i] <- X[i,t] }
}
for(i in 1:2){for(j in 1:2){XTX[i,j] <- inprod(XT[i,],X[,j]) }}
XTXinv[1:2,1:2] <- inverse(XTX[,])
... ...
```

References

[1] McGrayne S B: The theorem that would not die. Yale University Press. 2011.

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

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

[4] Christensen R, Johnson W, Branscum A, Hanson E: Bayesian Ideas and Data Analysis. CRC Press. 2011.

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

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

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

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

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

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

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

[12] Gelman A: Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, No 3, pp. 515-533. 2006.