Introduction to Bayesian Inference

jukka.ranta@helsinki.fi

17.1.2012

Abstract

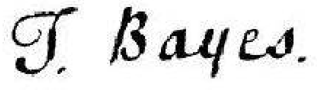
The course gives a practical introduction to bayesian inference and basics of WinBUGS / OpenBUGS. (Also R software will be modestly used). There are no pre-requirements other than reasonable familiarity with basic differential and integral calculus and functions (mostly taught at high school advanced 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.

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.

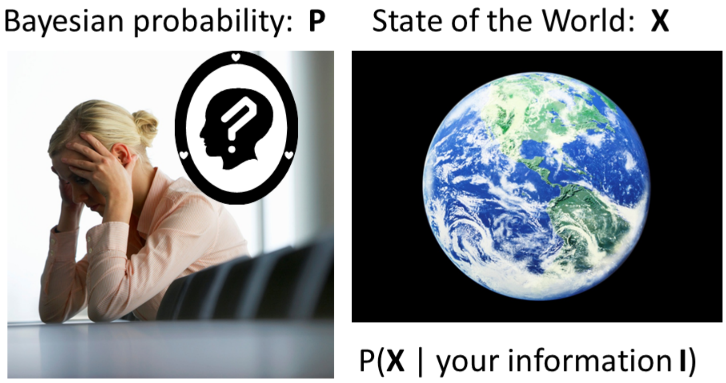

Bayesian probability: **P**
State of the World: **X**
P(**X** | your information **I**)

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

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

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

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. 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

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

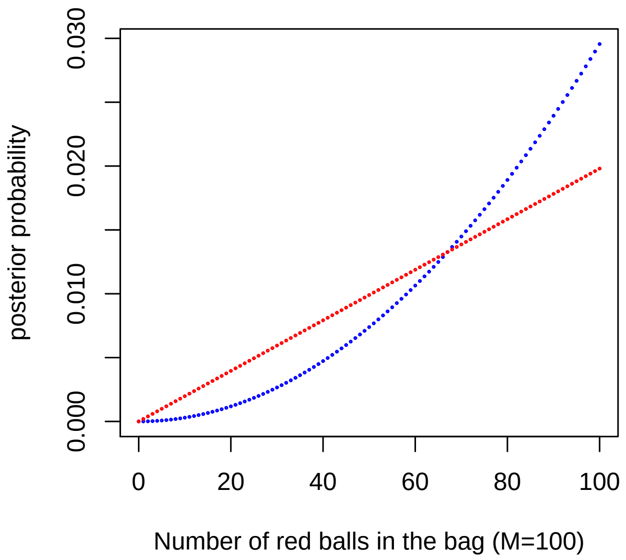


Figure 3: Posterior probabilities for the number of red balls among M in a bag, if one ball is drawn and it is red (red dots), and if two balls are drawn and both are red (blue dots).

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 Exercises

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. 515Ű533. 2006.