Modified Upper and Lower Probabilities based on Imprecise Likelihoods

Nic Wilson
School of Computing and Mathematical Sciences,
Oxford Brookes University, UK
nic.wilson@brookes.ac.uk

Abstract

The large sample behaviour is examined for upper and lower probabilities generated by precise priors (on a finite set) and imprecise likelihoods, for equally independently distributed variables taking values in a finite set. It can easily happen that the posterior upper probability of θ will tend to 1 in situations where the relative frequencies are tending to a measure far removed from the set of conditional measures associated with θ. A number of different modifications to upper and lower probability are made in such a way that the posterior probabilities will, with probability one, become or tend to 0 in such situations; these all involve rejecting measures which are extremely implausible given the data, and hence considering more restricted sets of likelihoods.

Keywords. Upper and Lower Probability, Imprecise Likelihoods, Large sample behaviour, Bernouilli Trials.

1 Introduction

Imagine the following scenario: A sequence X of N bits is randomly generated either by mechanism θ_1 or by mechanism θ_2, but we don't know which. However, we do know that a fair coin was tossed to decide which mechanism is used, and that with either θ_i, each bit is generated independently of all the others, and with a fixed chance p_i of being 0. In the case of mechanism θ_1, p_1 is known to be precisely 0.5; for θ_2, it is only known that the chance of picking a zero for each trial is in the interval $[0.1, 0.6]$. On the basis of the sequence of bits we would like to infer which is more likely to have been the mechanism, θ_1 or θ_2.

Now, suppose that N is large and that the fraction of zeroes in the sequence is close to 0.2. It does not seem credible that the first mechanism could have generated this data, since we know that the relative frequency would almost certainly then have been close to 0.5. More precisely we can consider the probability that θ_1 generates data such that the proportion of zeroes is more than 0.3 from its chance 0.5, i.e., the probability, given that θ_1 is the correct model, that the relative frequency will not be in the interval $[0.2, 0.8]$. Even for N as small as 100 this is much less than 10^{-8}. Therefore, since θ_1 is not credible, and the data is entirely consistent with θ_2, it seems that the data must almost certainly have been generated by θ_2.

However, this is not suggested by the upper and lower probabilities conditional on X. For $N = 100$, $\underline{\Pr}(\theta_2|X)$, the lower probability of θ_2 conditional on X, is less than 10^{-6}, and so the upper probability $\overline{\Pr}(\theta_1|X)$ is greater than $1 - 10^{-6}$. The reason for this is that there is a possible value for p_2, in particular, 0.6, which is even (in fact much) less likely to have generated data with a relative frequency of about 0.2. But we are then comparing two incredible possibilities, when there are alternatives that are entirely consistent with the data, i.e., θ_2 when p_2 is around 0.2.

The purpose of this paper is to look at ways of modifying upper and lower probability for this type of situation, so that they have behaviour more consistent with the intuitive argument given above. The approach taken can be roughly summarised as follows: We assume that the observed data will not be implausibly far from the true distribution; then we consider upper and lower probabilities given this assumption.

For convenience, we assume an objective view of probability in this paper, though this work can be applied more widely. A crucial assumption that we make is that the data are being generated by an unknown, but fixed chance distribution. Clearly there are many situations in which this assumption is untenable (see e.g., the discussion in Walley [5] and Walley and Fine [4]); however, where it does apply, it can give much stronger inferences.

Section 2 describes the mathematical setup, and standard upper and lower probabilities generated from imprecise likelihoods. Section 3 motivates and defines the modifications for upper and probabilities, while leaving the precise definition of plausibility open, except for its desirable properties. Section 4 gives two natural ways of defining when a distribution is plausible or not given the data.

2 Upper and Lower Probabilities from Imprecise Likelihoods

Suppose F is an unknown chance distribution over some finite set V. A sequence X of N values is generated, each value being generated independently with F. For $k = 1, \ldots, N$, let $X(k)$ be the kth element of the sequence. Define R to be the observed distribution, so that, for $x \in V$, $R(x)$ is the proportion of the N values in the sequence which are x, i.e., $\frac{1}{N}|\{k : X(k) = x\}|$. The sequence can be thought of as a random variable, and when viewed as such it will be written $\mathbf{X}$ and its associated relative frequencies as $\mathbf{R}$. The chance of particular sequence X being generated is $\prod_{k=1}^{N} F(X(k))$ which can be written as $\prod_{x \in V} F(x)^{NR(x)} = \exp(N\mathcal{Q}(F, R))$, where for probability functions R and S on V, $\mathcal{Q}(S, R)$ is defined to be $\sum_{x \in V} R(x) \log_e S(x)$, and for $r = 0$, $r \log_e r$ is taken to be 0.

2.1 The Objective Bayesian Case: Precise Likelihoods and Priors

Let Θ be a set of models, which will be assumed to be finite. To each $\theta \in \Theta$ is associated a measure F_θ on the subsets of V; for each trial $F_\theta(x)$ is the chance that x occurs, conditional on θ being the correct model. Θ is assumed to represent a mutually exclusive and exhaustive set of possibilities; the exhaustivity implies that $F = F_\theta$ for some $\theta \in \Theta$.

We also are given a prior distribution ρ over Θ. To simplify interpretation and exposition, we assume that an element of Θ was picked randomly with ρ; this enables us to unproblematically speak of *the* probability of events, as they are objective chances, even when we are considering imprecise priors and/or likelihoods. However the methods developed in this paper still apply for subjective probabilities. It is also assumed that, for all $\theta \in \Theta$, $\rho(\theta) \neq 0$. Using Bayes' Rule, the posterior probability of θ given observations X, $\Pr(\theta|X)$, is proportional to $\rho(\theta)F_\theta(X)$, which equals, by the discussion above, $\rho(\theta)\exp(N\mathcal{Q}(F_\theta, R))$. Hence

$$\Pr(\theta|X) = \frac{\rho(\theta)\exp(N\mathcal{Q}(F_\theta, R))}{\sum_{\tau \in \Theta} \rho(\tau)\exp(N\mathcal{Q}(F_\tau, R))}.$$

2.2 Imprecise Likelihood and Precise Priors

Here we relax the assumption that the probability measures F_θ are known; they are still assumed to exist, i.e., each model θ is taken to generate values in V with chance distribution F_θ, but we don't generally know the values precisely; instead it is assumed that all we know are constraints for each $x \in V$, $F_\theta(x) \in [L_\theta(x), U_\theta(x)]$, where $L_\theta(x)$ and $U_\theta(x)$ are known numerical values with $L_\theta(x) \leq U_\theta(x)$ and $\sum_{x \in V} L_\theta(x) \leq 1 \leq \sum_{x \in V} U_\theta(x)$. To simplify the results in this paper, it will be assumed that for all θ and all x, $L_\theta(x) > 0$ and $U_\theta(x) < 1$. The constraints give rise, for each $\theta \in \Theta$, to a family $\mathcal{P}_\theta$ of strictly positive distributions P_θ defined by $P_\theta \in \mathcal{P}_\theta$ iff for all $x \in V$, $P_\theta(x) \in [L_\theta(x), U_\theta(x)]$. Write $\mathcal{P}$ for $\bigcup_{\theta \in \Theta} \mathcal{P}_\theta$.

It is natural to ask, for given data X, what are the (optimal) upper and lower bounds that the constraints imply for $\Pr(\theta|X)$, i.e., the upper and lower probabilities conditional on X. Using the Bayesian case above, the lower probability, $\underline{\Pr}(\theta|X)$, is equal to the infimum, over all possible values of $F_\theta \in \mathcal{P}_\theta$, (for each $\theta \in \Theta$), of $\Pr(\theta|X)$. This equals

$$\frac{\rho(\theta)\underline{P}_\theta(X)}{\rho(\theta)\underline{P}_\theta(X) + \sum_{\tau \in \Theta - \{\theta\}} \rho(\tau)\overline{P}_\tau(X)}$$

where $\underline{P}_\theta(X) = \min_{P \in \mathcal{P}_\theta} P(X)$ and $\overline{P}_\tau(X) = \max_{P \in \mathcal{P}_\tau} P(X)$. Similarly, the upper probability, $\overline{\Pr}(\theta|X)$ can be expressed as

$$\frac{\rho(\theta)\overline{P}_\theta(X)}{\rho(\theta)\overline{P}_\theta(X) + \sum_{\tau \in \Theta - \{\theta\}} \rho(\tau)\underline{P}_\tau(X).}$$

$\underline{P}_\theta(X)$ can be written as $\exp(N\underline{\mathcal{Q}}_\theta(R))$, where $\underline{\mathcal{Q}}_\theta(R) = \min\{\mathcal{Q}(P_\theta, R) : P_\theta \in \mathcal{P}_\theta\}$, and $\overline{P}_\tau(X)$ as $\exp(N\overline{\mathcal{Q}}_\tau(R))$, where $\overline{\mathcal{Q}}_\tau(R) = \max\{\mathcal{Q}(P_\tau, R) : P_\tau \in \mathcal{P}_\tau\}$.

Therefore, in order to compute these upper and lower bounds, we just need to compute $\underline{\mathcal{Q}}$ and $\overline{\mathcal{Q}}$, the minimum and maximum entropy relative to R, within sets $\mathcal{P}_\theta$.

2.3 Imprecise Priors and Imprecise Likelihoods

More generally we will have imprecise priors as well as likelihoods. This has been considered, for example, in [5, 2]. In all the methods described here, just the priors and the upper and lower likelihoods are important, so adaption of the standard methods should be straight-forward.

2.4 Behaviour of Upper and Lower Probabilities for Large N

As can be seen from section 2.2, the upper and lower probabilities are determined by the prior and the differences $N(\overline{\mathcal{Q}}_\tau(R) - \underline{\mathcal{Q}}_\theta(R))$ for various $\theta \neq \tau \in \Theta$. It can be shown using the strong law of large numbers (e.g., [3]) and continuity of $\mathcal{Q}$ that if $\overline{\mathcal{Q}}_\tau(F) - \underline{\mathcal{Q}}_\theta(F) > 0$ then with probability one, $N(\overline{\mathcal{Q}}_\tau(R) - \underline{\mathcal{Q}}_\theta(R))$ will tend to ∞ as N tends to infinity; similarly, if $\overline{\mathcal{Q}}_\tau(F) - \underline{\mathcal{Q}}_\theta(F) < 0$ then with probability one, $N(\overline{\mathcal{Q}}_\tau(R) - \underline{\mathcal{Q}}_\theta(R))$ will tend to $-\infty$.

Now, suppose that θ is such that $\mathcal{P}_\theta$ is not the singleton set $\{F\}$ and that there exists some $\tau \neq \theta$ with $\mathcal{P}_\tau \ni F$. Then $\overline{\mathcal{Q}}_\tau(F) = \mathcal{Q}(F, F) > \underline{\mathcal{Q}}_\theta(F)$, and so, with probability one, as N tends to infinity, $\underline{\Pr}(\theta|X)$ tends to 0.

In particular this holds for any $\mathcal{P}_\theta$ not containing F, because, by exhaustivity of Θ, $F = F_\tau$ for some $\tau \in \Theta$, so $\mathcal{P}_\tau \ni F$ and $\tau \neq \theta$. This is what we would hope for. However, we do not always have, with probability one, that $\overline{\Pr}(\theta|X)$ tends to 0. Define function d_F on $\mathcal{P}$ by $d_F(P) = \mathcal{Q}(F, F) - \mathcal{Q}(P, F)$ (known as the Kullback-Liebler divergence between F and P). d_F is always non-negative, and $d_F(P) = 0$ if and only if $P = F$. $d_F(P)$ can be considered as a measure of how far P is from F. Now, since $\mathcal{P}_\theta$ is topologically closed, there exists an element, say P^*_θ, which is closest to F in this sense, i.e., is such that $\mathcal{Q}(P^*_\theta, F) = \overline{\mathcal{Q}}_\theta(F)$. If for each $\tau \neq \theta$ there's an element $P_\tau \in \mathcal{P}_\tau$ with P_τ further from F than P^*_θ, i.e., $d_F(P_\tau) > d_F(P^*_\theta)$, then, with probability one $\overline{\Pr}(\theta|X)$ tends to 1 (since for all τ, $\overline{\mathcal{Q}}_\theta(F) - \underline{\mathcal{Q}}_\tau(F) > 0$). With probability one $\overline{\Pr}(\theta|X)$ tends to 0 if for some τ, for all $P_\tau \in \mathcal{P}_\tau$ and all $P_\theta \in \mathcal{P}_\theta$, $d_F(P_\tau) < d_F(P_\theta)$, i.e., for some τ, every element of $\mathcal{P}_\tau$ is closer to F than any element of $\mathcal{P}_\theta$.

The strong law of large numbers implies that, if $\mathcal{P}_\tau \not\ni F$, it will almost certainly be the case that, for all sufficiently large N, the relative frequencies will be extremely close to F, so to some element in some $\mathcal{P}_\tau$, but will not be close (relatively speaking) to any element in $\mathcal{P}_\theta$. As N increases we can be increasingly confident that θ is not the true element of Θ. However we very easily have situations where $\overline{\Pr}(\theta|X)$ tends to 1. It thus seems desirable to develop versions of upper and lower probability that are less cautious, and are consistent with the argument above.

3 Sets of Plausible Measures

In this section an abstract approach is taken to implausibility of probability measures given the data; we define three modifications of upper and lower probability, and give properties which are sufficient to obtain the essential property we want, that is, if $\mathcal{P}_\theta$ does not contain the true measure F, then, with probability one, the upper probability of θ given the data should tend to zero. It is also shown here that the concrete definitions of implausibility defined in section 4 satisfy these properties under fairly weak conditions.

3.1 The General Framework

First some notation will be defined. Recall that N is the number of trials that will be performed; and let $\mathcal{X}$ be the set of all possible sequences X. Each $P \in \mathcal{P}$ is extended to $\mathcal{X}$ in the obvious way, by $P(X) = \prod_{k=1}^{N} P(X(k))$; as observed above, $\log P(X)$ can be written as $N\mathcal{Q}(P, R_X)$ where R_X is the relative frequency probability measure summarising the sequence X. If $\mathcal{N}$ is another set of probability measures then define $\mathcal{N}^*$ to be $\{X \in \mathcal{X} : R_X \in \mathcal{N}\}$.

For each N and probability measure $P \in \mathcal{P}$ we assume a set of probability measures $\mathcal{N}_{P,N}$ (usually abbreviated to $\mathcal{N}_P$), which contains P; we look in detail at two ways of defining these in section 4. It is intended that $\mathcal{N}_P$ contains all the relative frequency functions R that might plausibly have been generated by P after the N trials. It is therefore assumed that the sets $\mathcal{N}_P$ have been defined in such a way as to make $\Pr(\mathbf{R}^P \in \mathcal{N}_P)$ very close to 1, where $\mathbf{R}^P$ is the relative frequency random variable, given that P is the true measure F; hence $\Pr(\mathbf{R}^P \in \mathcal{N}_P)$ equals $P(\mathcal{N}_P^*)$.

Let C be the (random) event that $\mathcal{N}_F \ni \mathbf{R}$. Since, by exhaustivity of Θ, $F \in \bigcup_{\theta\in\Theta} \mathcal{P}_\theta = \mathcal{P}$, $\Pr(C)$ is very close to 1. (More precise bounds for $\Pr(C)$ can often be achieved by observing that it is the ρ-weighted average over Θ of $\Pr(\mathbf{R}^{F_\theta} \in \mathcal{N}_{F_\theta})$; each of these terms is known to be close to 1.) C is a well-defined event—F, though unknown, exists, $\mathcal{N}_F$ is defined before the sequence of trials takes place, and it is an objective fact after the sequence X has been generated whether or not R_X is in $\mathcal{N}_F$.

The upper and lower probabilities we define will be conditional on this C (as well as X). The first approach we consider does precisely that; the second and third take further steps. If the event C occurs then it must be the case that $F \in \bigcup_{\theta\in\Theta} \mathcal{P}'_\theta$, where $\mathcal{P}'_\theta = \mathcal{P}_\theta \cap \mathcal{N}^R$, and $\mathcal{N}^R = \{P : \mathcal{N}_P \ni R\}$. The second approach computes upper and lower probabilities over the reduced sets $\mathcal{P}'_\theta$ instead of the sets $\mathcal{P}_\theta$; the third modifies the second by not allowing these reduced sets to become empty.

Assuming C

Before describing in detail the modified upper and lower probabilities, the question will be addressed as to whether knowing the upper and lower probabilities conditional on C has much value. At first sight it may seem as if C is being assumed just because it has very high probability; of course this would not be reasonable; for example, the chance of any given person winning the lottery is very small, but if we assumed, for each person, that they wouldn't win, we'd be assuming that no-one would win. It is important here that C is fixed before the trials take place, and that there is one C (as opposed to a very large number) for each experiment.

It might be argued that what is important is a high value of $\Pr(C|X)$, the probability conditional on the sequence X that actually occurs (rather than a high value of $\Pr(C)$) since we're assuming C in the context X. If the likelihoods and priors were precise, we'd of course almost certainly have a very high probability $\Pr(C|X)$ because $\Pr(C)$ is very high, but in the type of situation we're interested in, where the bounds on upper and lower probabilities of elements of Θ given X are disappointly imprecise, then we'll tend to get the same for $\Pr(C|X)$. (If it did turn out that the upper probability for C given X was not close to 1, we shouldn't assume C; but this is very unlikely to happen.) In this situation, at least if $\Pr(\neg C)$ is small enough, it can be considered important in itself in justifying the assumption of C. For example, if $\Pr(\neg C)$ were less than, say, 10^{-8}, one could reason as follows:- *Even if I were to observe such a sequence of trials every hour of my life it would be unlikely that $\neg C$ would occur even once. Therefore it seems safe to assume that the event C will occur for this experiment.* Essentially, if $\Pr(\neg C)$ is small enough, I'm not in practice going to be wrong if I assume C.

In any case it is not at all unusual in practice to be making assumptions, and considering probabilities given these; for example, in many settings it is not certain that the set of possibilities is exhaustive but it can be useful to assume it.

3.2 Upper and Lower Plausible Probabilities

In this section modifications of upper and lower probability for this scenario are defined, and the properties are explored. Here only the upper and lower probabilities of singletons sets $\{\theta\}$ conditional on X will be considered, but the definitions can be extended easily to arbitrary subsets of Θ, and to upper and lower expectations of gambles on Θ.

For probability measure $P \in \mathcal{P}$ define $P(X \wedge C)$ to be the probability of $X \wedge C$ given that $F = P$, which equals $P(X)$ if $\mathcal{N}_P \ni R_X$ (i.e., $P \in \mathcal{N}^{R_X}$) and 0 otherwise.

$\Pr_C(\theta|X)$, i.e., $\Pr(\theta|X \wedge C)$, can be written as

$$\frac{\rho(\theta)F_\theta(X \wedge C)}{\sum_{\tau\in\Theta}\rho(\tau)F_\tau(X \wedge C)},$$

as long as for some τ, $F_\tau(X \wedge C) \neq 0$, i.e., $F_\tau \in \mathcal{N}^{R_X}$ (since $F_\tau(X) \neq 0$).

We're interested in upper and lower probabilties of elements of Θ conditional on the data, X, and C. In order for these to be well-defined we need that there's some possibility that the probability of $X \wedge C$ is non-zero. This is the case if and only if there exists some $\theta \in \Theta$ and $P \in \mathcal{P}_\theta$ with $\mathcal{N}_P \ni R_X$ (since by hypothesis, $P(X) \neq 0$), which happens if and only if $\bigcup_\theta \mathcal{P}'_{\theta,N} \neq \emptyset$, where $\mathcal{P}'_{\theta,N}$ (usually abbreviated to $\mathcal{P}'_\theta$) is defined to be $\mathcal{P}_\theta \cap \mathcal{N}^{R_X}$. If we define $\mathcal{P}'$ to be $\mathcal{P} \cap \mathcal{N}^{R_X}$ then the condition we need is that $\mathcal{P}' \neq \emptyset$. Since $\mathcal{N}^{R_X}$ is intended to be all P from which R_X could plausibly arise, $\mathcal{P}'_\theta$ can be thought of as all the plausible values for F in $\mathcal{P}_\theta$, and $\mathcal{P}'$ as all plausible values for F, given the data X.

Three different modifications of upper and lower probability will be considered. The first one is to consider the upper and lower probabilities conditional on C; this amounts to considering the supremum and infimum of $\Pr_C(\theta|X)$ over all possible F_τ in $\mathcal{P}_\tau$, for $\tau \in \Theta$, subject to the condition that some F_τ is in $\mathcal{P}'_\tau$, i.e., that at least one F_τ is plausibly F.

The second ($\overline{\Pr}'_C$ and $\underline{\Pr}'_C$) is much bolder; it is also defined as the supremum and infimum of $\Pr_C(\theta|X)$, but only over choices F_τ in $\mathcal{P}'_\tau$; i.e., for each $\tau \in \Theta$, we only consider plausible F_τ.

In the first and second methods, if, as N increases, $\mathcal{P}'_\theta$ becomes empty, θ at that N suddenly makes no contribution, giving rise to a jump in the value of the upper/lower probability. This is most significant when we are considering the upper conditional probability of θ, as it will often happen that, for some increment of N, the upper probability will jump from a substantial value to 0. The third modifies the second, by not allowing the family of measures corresponding to θ to become empty, giving rise to a gradual fading of the influence of θ as N increases. $\overline{\Pr}''_C(\theta|X)$ is defined to be the supremum of $\Pr_C(\theta|X)$ over sets $\mathcal{P}''_{\tau,N}$ (abbreviated to $\mathcal{P}''_\tau$) where $\mathcal{P}''_\tau$ equals $\mathcal{P}'_\tau$ if the latter is non-empty; otherwise it equals the singleton set consisting of the (unique) element of $\mathcal{P}_\tau$ with maximal value of $P_\tau(X)$.

Define for all $\theta \in \Theta$ $\underline{P}'_\theta(X) = \inf_{P\in\mathcal{P}'_\theta} P(X)$, $\overline{P}'_\theta(X) = \sup_{P\in\mathcal{P}'_\theta} P(X)$, and similarly $\underline{P}''_\theta(X)$ and $\overline{P}''_\theta(X)$, with

$\mathcal{P}''_\theta$ replacing $\mathcal{P}'_\theta$ in the equations. Therefore when $\mathcal{P}'_\theta \neq \emptyset$, $\underline{P}''_\theta(X) = \underline{P}'_\theta(X)$ and $\overline{P}''_\theta(X) = \overline{P}'_\theta(X)$. If $\mathcal{P}'_\theta = \emptyset$, $\underline{P}''_\theta(X) = \overline{P}''_\theta(X) = \overline{P}_\theta(X)$. Let $\Theta' = \{\theta : \mathcal{P}'_\theta \neq \emptyset\}$, which can be thought of as the set of all the models that are still plausible. Defining Θ'' analogously gives $\Theta'' = \Theta$. Each of $\underline{\Pr}_C(\theta|X)$, $\underline{\Pr}'_C(\theta|X)$ and $\underline{\Pr}''_C(\theta|X)$ can be written in the form $A/(A+B)$. If $\theta \in \Theta'$, let

$$a' = \rho(\theta)\underline{P}'_\theta(X) \quad \text{and} \quad b' = \sum_{\tau \in \Theta' - \{\theta\}} \rho(\tau)\overline{P}'_\tau(X),$$

and define a'' and b'' similarly. $\underline{\Pr}''_C(\theta|X)$ then equals $a''/(a''+b'')$. Note that when a' is defined (i.e., when $\mathcal{P}'_\theta \neq \emptyset$) it equals a'', and that b'' is b' with (possibly) some additional terms. When $\underline{\Pr}'_C(\theta|X)$ is well-defined (i.e., when $\mathcal{P}' \neq \emptyset$), it equals 0 if $\mathcal{P}'_\theta = \emptyset$, and otherwise equals $a'/(a'+b')$. Similarly, $\underline{\Pr}_C(\theta|X)$ is only defined when $\mathcal{P}' \neq \emptyset$; when it is defined, it equals 1 if $\Theta' = \{\theta\}$ (i.e., if θ is the only plausible model); otherwise it equals 0 if $\mathcal{P}_\theta \not\subseteq \mathcal{N}^{R_X}$, and otherwise equals $a'/(a'+b')$.

For the upper probabilities, define, when $\theta \in \Theta'$, $y' = \rho(\theta)\overline{P}'_\theta(X)$ and, for any θ, $y'' = \rho(\theta)\overline{P}''_\theta(X)$,

$$z' = \sum_{\tau \in \Theta' - \{\theta\}} \rho(\tau)\underline{P}'_\tau(X), \quad z^* = \sum_{\tau \in \Theta^* - \{\theta\}} \rho(\tau)\underline{P}'_\tau(X),$$

where $\Theta^* = \{\tau \in \Theta : \mathcal{P}_\tau \subseteq \mathcal{N}^{R_X}\}$. Define z'' analogously to the definition of z'.

Then $\overline{\Pr}''_C(\theta|X)$ equals $y''/(y''+z'')$. Assuming $\mathcal{P}' \neq \emptyset$, we have $\overline{\Pr}_C(\theta|X) = \overline{\Pr}'_C(\theta|X) = 0$ if $\mathcal{P}'_\theta = \emptyset$; otherwise, $\overline{\Pr}_C(\theta|X) = y'/(y'+z^*)$ and $\overline{\Pr}'_C(\theta|X) = y'/(y'+z')$.

$\underline{P}'_\theta(X)$, for example, can be written as $\exp(N\underline{\mathcal{Q}}'_\theta(R))$, where $\underline{\mathcal{Q}}'_\theta(R) = \inf\{\mathcal{Q}(P_\theta, R) : P_\theta \in \mathcal{P}'_\theta\}$. Similarly $\overline{P}'_\theta(X)$ can be written as $\exp(N\overline{\mathcal{Q}}'_\theta(R))$ with $\overline{\mathcal{Q}}'_\theta(R) = \sup\{\mathcal{Q}(P_\theta, R) : P_\theta \in \mathcal{P}'_\theta\}$.

By dividing through the fractions by the numerator it can be seen that, as for standard upper and lower probability (see section 2.4), the behaviour for large N of, for example, $\overline{\Pr}'_C(\theta|X)$ and $\underline{\Pr}'_C(\theta|X)$ is essentially determined by the differences $N(\overline{\mathcal{Q}}'_\tau(R) - \underline{\mathcal{Q}}'_\theta(R))$ for various $\theta \neq \tau \in \Theta$.

To ensure the desired behaviour we need to make some conditions on the sets $\mathcal{N}_P$.

(i) For any possible observed distribution R (i.e., any probability measure on the subsets of V taking only rational values), and any $\delta > 0$, there exists M such that[1] for all $N \geq M$, $d_R(P) < \delta$ for all $P \in \mathcal{N}^{R,N}$; and

(ii) For any possible F (i.e., any element of $\mathcal{P}$) with probability one, there exists M' such that for all $N \geq M'$, $\mathcal{N}^{\mathbf{R},N} \ni F$.

(i) is roughly saying that for any R, $\mathcal{N}^{R,N}$ shrinks with N, but doesn't say anything about how fast. (ii) implies that with probability one for all sufficiently large N, $\mathcal{N}_{F,N} \ni \mathbf{R}$, i.e., that C will occur. Hence (ii) implies that $\Pr(C)$ tends to 1 as N tends to infinity.

Suppose $\mathcal{P}_\theta \not\ni F$. Then, it can be shown using the strong law of large numbers that (i) implies that there exists $c > 0$ such that with probability one there exists M'' such that for all $N \geq M''$,

$$\inf_{P \in \mathcal{N}^{R,N}} \mathcal{Q}(P, \mathbf{R}) - \sup_{P \in \mathcal{P}_\theta} \mathcal{Q}(P, \mathbf{R}) > c.$$

An immediate consequence of this is that when $\mathcal{P}_\theta \not\ni F$, with probability one for all sufficiently large N, $\mathcal{P}'_{\theta,N}$ is empty, and hence $\overline{\Pr}_C(\theta|X) = \overline{\Pr}'_C(\theta|X) = 0$, as long as $\mathcal{P}' \neq \emptyset$. But by exhaustivity there exists $\tau \neq \theta$ with $\mathcal{P}_\tau \ni F$, and so by (ii), we indeed have that with probability one, for all sufficiently large N, $\mathcal{P}' \neq \emptyset$, and therefore $\overline{\Pr}_C(\theta|X) = \overline{\Pr}'_C(\theta|X) = 0$. A similar argument using the last displayed inequality implies that with probability one $\overline{\Pr}''_C(\theta|X)$ tends to 0 (in fact, this doesn't even need property (ii)).

Verifying Properties (i) and (ii)

The first definition of plausible distributions (section 4.1) lets $\mathcal{N}_{P,N}$ be the set of all R such that for all $x \in V$, $|R(x) - P(x)| \leq J_{P(x),N}$ where $J_{P(x),N}$ is chosen so that $\Pr(|\mathbf{R}^P(x) - P(x)| > J_{P(x),N})$ is just less than ϵ. For property (ii) we require that with probability one, for all sufficiently large N, and all $x \in V$, $|\mathbf{R}^F(x) - F(x)| \leq J_{F(x),N}$. Since (see below) $N\mathbf{R}^F(x)$ has a binomial distribution, by the law of the iterated logarithm [3] sufficient for this is that for some $\lambda > 1$, for all sufficiently large N, $K_{F(x),N} > \sqrt{2\lambda \log\log N}$, where $K_{p,N} = J_{p,N}\sqrt{N}/\sqrt{p(1-p)}$. This means that we just need ϵ to decrease very slowly with N. Using the fact that a re-scaled binomial tends to a normal distribution (the DeMoivre-Laplace limit theorem and its extensions [3]), and bounds on the tail of a normal distribution, it can be shown that sufficient is that for some $\lambda > 1$, for all sufficiently large N, $\epsilon < (\log N)^{-\lambda}$.

To satisfy property (i) it is sufficient that $J_{p,N}$ tends to 0 uniformly over p as N tends to infinity. Using Chebyshev's Inequality it can be shown that a (crude) sufficient condition for this is that $N\epsilon$ tends to infinity as N tends to infinity. Therefore as long as we have for large N that ϵ satisfies these constraints then the two properties are satisfied.

The second definition is of the form: $\mathcal{N}_{P,N}$ is the set

[1] Note that any other sensible distance can be used instead of d, e.g., Euclidean distance, without changing the property.

of all R such that $Nd_R(P) \leq \mu_P + K\sigma_P$. Within $\mathcal{P}$, μ_P and σ_P can be bounded above by quantities not dependent on P and N. So property (i) will be satisfied as long as K/N tends to 0 as N tends to infinity.

For the property (ii) to hold we need with probability one, for all sufficiently large N, $Nd_{\mathbf{R}}(F) \leq \mu_F + K\sigma_F$. For $x \in V$, let $\delta(x) = (R(x) - F(x))/F(x)$. By the law of the iterated logarithm, for any $\lambda > 1$, with probability one, for all sufficiently large N, for all x,

$$(\delta(x))^2 < \frac{1 - F(x)}{NF(x)} \times 2\lambda \log\log N.$$

$d_{\mathbf{R}}(F)$ can be written as $\sum_x F(x)(1 + \delta(x)) \log(1 + \delta(x))$. Almost certainly for all sufficiently large N we have $|\delta(x)| < 1$. Therefore the last expression can be expanded using the power series form for $\log(1+\delta(x))$ and simplified using the fact that $\sum_x F(x)\delta(x) = 0$, to give

$$d_{\mathbf{R}}(F) = \sum_{x \in V} F(x) \sum_{i=2}^{\infty} \frac{(-1)^i\, \delta^i(x)}{i(i-1)}.$$

The most significant term here is $\frac{1}{2}\sum_x F(x)\delta^2(x)$. Using the above constraint on $\delta(x)$ and the expressions for μ_F and σ_F in section 4.2, it can be seen that K need only increase extremely slowly with N for property (ii) to be satisfied, i.e., that for all sufficiently large N, $Nd_{\mathbf{R}}(F) \leq \mu_F + K\sigma_F$. (This is because μ_F and σ_F are bounded above zero as functions of N, and for large enough N, $Nd_{\mathbf{R}}(F)$ is at worst only very slowly increasing in N.) For example it is sufficient that for some $t > 0$ (any will do), for all sufficiently large N, $K > (\log N)^t$.

Hence for both definitions of plausible probabilities, for a wide range of the key parameters[2] (ϵ in the first case and K in the second), we have when $\mathcal{P}_\theta \not\ni F$, with probability one for all sufficiently large N, $\overline{\Pr}_C(\theta|X) = \overline{\Pr}'_C(\theta|X) = 0$, and as N tends to infinity, $\overline{\Pr}''_C(\theta|X)$ tends to 0.

4 Eliminating Implausible Measures

In this section two different definitions of implausibility are given. As discussed in the last section, for each θ we replace $\mathcal{P}_\theta$ by a smaller set of distributions $\mathcal{P}'_\theta$ by eliminating those distributions which are incredibly unlikely to have given rise to the observed distribution R. We then can calculate the new posterior upper and lower probabilities, based on the sets $\mathcal{P}'_\theta$, rather than the sets $\mathcal{P}_\theta$.

[2] In practice one would probably use ϵ and K varying very slowly with N.

The first definition is pointwise: P is rejected if for some x, $R(x)$ is incredibly unlikely to have arisen from $P(x)$. The second uses a measure of how far the observed distribution R is from P, rejecting P if R is not close to it.

4.1 Implausibility I

Suppose P is the true probability measure F. Then the distribution $\mathbf{R}^P$ we observe after N trials is a random variable, depending randomly just on P and N. In fact, for each x, $N\mathbf{R}^P(x)$ has a binomial distribution: the chance that $N\mathbf{R}^P(x) = k$ is equal to $b(k; N, P(x))$, where the binomial probability $b(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$. This is because $N\mathbf{R}^P(x) = \sum_{k=1}^{N} I_{\mathbf{X}(k)=x}$, and so is the total number of successes of a sequence of N Bernouilli trials, where $I_{\mathbf{X}(k)=x}$ is the indicator function for the kth trial being a success, equalling 1 if $\mathbf{X}(k) = x$, and 0 otherwise. Hence $\mathbf{E}[\mathbf{R}^P(x)] = P(x)$ and $\mathbf{Var}[\mathbf{R}^P(x)] = \frac{1}{N}P(x)(1 - P(x))$.

We want to reject P such that for some x, the observed value $R(x)$ is incredibly far from $P(x)$. To do this a very small probability ϵ is picked; a very cautious agent might use e.g., $\epsilon = 10^{-10}$ or smaller. Define $J_{P(x)}$ to be the minimum of all real values J such that $\Pr(|\mathbf{R}^P(x) - P(x)| > J) < \epsilon$. (Note that $J_{P(x)}$ depends also on N and ϵ; however, for the moment, only its dependence on $P(x)$ concerns us, and N and ϵ can be considered fixed.) We then have $J_p = \frac{1}{N}\min\{B : T(B; N, p) < \epsilon\}$, where $T(B; N, p)$, the tails probability[3] of the binomial distribution, is defined to be

$$\sum_{k\,:\,|k-Np|>B} b(k; N, p).$$

The idea is to reject probability function P if for some $x \in V$, $|R(x) - P(x)| > J_{P(x)}$, as these are implausible given the data. So we define $\mathcal{N}_P$ to be the set of R such that for all x, $|R(x) - P(x)| \leq J_{P(x)}$.

Abbreviate $P(x)$ to p, and $R(x)$ to r. In practice $J_{P(x)}$ may well have to be approximated. In view of the normal approximations to the binomial, it is natural to express J_p in terms of numbers of standard deviations from the mean. J_p corresponds to K_p standard deviations of $\mathbf{R}^P(x)$ from the mean, where $K_p = J_p\sqrt{N}/\sqrt{p(1-p)}$. K_p is then the number of standard deviations from the mean that $\mathbf{R}^P(x)$ has to be for the tails probability to be less than ϵ.

[3] Both tails of the binomial are being considered together. An alternative is to deal with each tail separately, arranging each tail to have probability just below $\epsilon/2$.

p is not being rejected iff

$$|r-p| \leq K_p \sqrt{\frac{p(1-p)}{N}},$$

i.e., iff $(r-p)^2 \leq \frac{K_p^2}{N} p(1-p)$. This occurs if and only if p is between the two numbers

$$\frac{1}{1+c_p}\left(r+\frac{c_p}{2} \pm \sqrt{c_p}\sqrt{r(1-r)+c_p/4}\right)$$

where c_p is defined to be K_p^2/N. When the normal makes a good approximation of the binomial (N needs to be sufficiently large, and ϵ large enough to make K_p^6/N small), K_p hardly varies at all with p (and N —it essentially just depends on ϵ), and so it is natural to approximate by writing $K_p = K$ and $c_p = c$. Given ϵ, K can then be looked up in a table of the normal distribution.

In other cases, if N is sufficiently large, and $P(x)$ is not too close to 0 or 1, a necessary condition for the above equation to be satisfied is that p must be close to r. K_p will then be close to K_r, so the acceptable $P(x)$ will be approximately those between the two numbers

$$\frac{1}{1+c_r}\left(r+\frac{c_r}{2} \pm \sqrt{c_r}\sqrt{r(1-r)+c_r/4}\right).$$

Since these bounds do not depend on p, we have constraints similar in form to the original constraints, and so various modified upper and lower probabilities can be computed as before, by maximising and minimising entropy relative to the observed distribution.

4.2 Implausibility II

Although the method described in the last section seems quite a natural way of rejecting distributions, it is perhaps a little inelegant testing each $P(x)$ separately. An alternative is to define a function $D(P,R)$ which is a measure of how far R is from P, and reject P if this is unusually large. A natural choice is $D(P,R) = N d_R(P)$ which equals

$$N\big(\mathcal{Q}(R,R)-\mathcal{Q}(P,R)\big) = N\sum_{x\in V} R(x)\log\frac{R(x)}{P(x)}$$

D is non-negative, and $D(P,R)=0$ if and only if $P = R$. This seems a particularly appropriate measure as it gives precisely the degree to which the data support P, normalised by the maximum degree of support that the data gives to any distribution. The closer P is to the observed distribution, the more plausible it is that R could have arisen from P. As before we want to reject any distribution P for which it is incredible that R could be a random sample from it.

Recall F is the true distribution and random variable $\mathbf{R}^F$ is the observed distribution after the N trials. $D(F,\mathbf{R}^F)$ can then be considered as a random variable. It can be shown using combinatorial arguments that $\mathbf{E}[D(F,\mathbf{R}^F)]$ equals

$$\frac{|V|-1}{2}+\frac{1}{12N}\left(\sum_{x\in V}\frac{1}{F(x)}-1\right)+O\left(\frac{1}{N^2}\right).$$

and

$$\mathbf{Var}[D(F,\mathbf{R}^F)] = \frac{|V|-1}{2}+O\left(\frac{1}{N}\right).$$

Abbreviate the expectation to μ_F and the variance to σ_F^2. The idea is then to choose a number K of standard deviations we require from the mean, and to reject P such that

$$D(P,R)-\mu_P > K\sigma_P.$$

P is thus being judged implausible if, on the assumption that $P = F$, the true distribution, R is more than K standard deviations further from P than expected—since then it is not plausible that it arose from P.

So the acceptable P all satisfy

$$\mathcal{Q}(P,R) \geq \mathcal{Q}(R,R)-\frac{\mu_P}{N}-\frac{K\sigma_P}{N}.$$

$\mathcal{N}^R$ is therefore defined to be the set of all P satisfying the above inequality.

Therefore acceptable P all satisfy

$$\mathcal{Q}(P,R) \geq \mathcal{Q}(R,R)-\frac{|V|-1}{2N}-\frac{K}{N}\sqrt{\frac{|V|-1}{2}}+U_P,$$

for some U_P which is $O(\frac{1}{N^2})$. The upper values of $\mathcal{Q}$ are not affected by this additional constraint, so $\overline{\mathcal{Q}}'_\theta(R) = \overline{\mathcal{Q}}_\theta(R)$ which can be computed easily. We can approximate $\underline{\mathcal{Q}}'_\theta(R)$, with an error at worst $O(1/N^2)$ as

$$\mathcal{Q}(R,R)-\frac{|V|-1}{2N}-\frac{K}{N}\sqrt{\frac{|V|-1}{2}}$$

unless this is less than $\mathcal{Q}(P,R)$ for all $P\in\mathcal{P}_\theta$, when $\underline{\mathcal{Q}}'_\theta(R) = \underline{\mathcal{Q}}_\theta(R)$, the standard case.

If more accurate values for $\underline{\mathcal{Q}}'_\theta(R)$ are required, one can compute e.g., the $O(1/N)$ term for the variance. It can be seen that for even moderately large N, for the above inequalities for $\mathcal{Q}(P,R)$ to be satisfied, for each x, $P(x)$ and $R(x)$ must be close, and then μ_P and σ_P can be approximated, respectively, by μ_R and σ_R, and $\underline{\mathcal{Q}}'_\theta(R)$ more closely approximated by the maximum of $\underline{\mathcal{Q}}_\theta(R)$ and $\mathcal{Q}(R,R)-\frac{\mu_R}{N}-\frac{K\sigma_R}{N}$.

One can also consider the random variable $D(P, \mathbf{R}^F)$ for $P \neq F$. It can be written as $D(F, \mathbf{R}^F) + NZ$ where $Z = \big(\mathcal{Q}(F, \mathbf{R}^F) - \mathcal{Q}(P, \mathbf{R}^F)\big)$. It can be shown that $\mathbf{E}[Z]$ and $N\mathbf{Var}[Z]$ are both positive and independent of N. Therefore $D(P, \mathbf{R}^F)$ will almost certainly be large for large N. Roughly speaking, for large N, unless for all x, $|P(x) - F(x)|$ is at most $O(1/\sqrt{N})$ then $D(P, \mathbf{R}^F)$ will almost certainly be large.

4.3 Example Continued

Consider the example again, but with upper and lower plausible probabilities. The definition of implausibility from 4.1 will be used, but for $|V| = 2$, the two definitions are strongly related. Using $\epsilon = 10^{-8}$, and $\mathcal{N}_{P,N} = \{R : |R(0) - P(0)| \leq J_{P(0),N}\}$, where $J_{P(0),N}$ is chosen so that $P(\mathcal{N}_{P,N})$ is just greater than $1 - \epsilon$, and $P(\mathcal{N}_{P,N})$ is the probability, conditional on P being F the true measure, of an observation in $\mathcal{N}_{P,N}$ being generated. Given the relative frequencies R, probability measure P is then deemed implausible if and only if $\mathcal{N}_{P,N} \not\ni R$. Given that as N increases, $R(0)$ stays around 0.2, $P = (0.6, 0.4)$ (i.e., with $P(0) = 0.6$, $P(1) = 0.4$) will get rejected around $N = 50$; around $N = 90$, $(0.5, 0.5)$, and hence θ_1, will get rejected.

For $N = 10$, $\underline{\mathrm{Pr}}'_C(\theta_2|X)$ is about 0.2; it then decreases exponentially with N until at around $N = 50$ it's approximately 10^{-3}. Up to this point the values are the same as the standard lower probability (section 2.2). Then, as elements of $\mathcal{P}_{\theta_2}$ get rejected, $\underline{\mathrm{Pr}}'_C(\theta_2|X)$ rapidly increases (without a jump); at $N = 75$ it becomes about 0.1. Just before $(0.5, 0.5)$ gets rejected, $\underline{\mathrm{Pr}}'_C(\theta_2|X)$ moves towards 0.5, then at the rejection point (around $N = 90$), $\underline{\mathrm{Pr}}'_C(\theta_2|X)$ jumps to 1.

$\underline{\mathrm{Pr}}_C(\theta_2|X)$ behaves similarly up to about $N = 50$. Then some elements of $\mathcal{P}_{\theta_2}$ become incompatible with C, so the lower probability $\underline{\mathrm{Pr}}_C(\theta_2|X)$ suddenly becomes zero. It stays at zero until $(0.5, 0.5)$ gets rejected, when it jumps to 1. In contrast, $\underline{\mathrm{Pr}}''_C(\theta_2|X)$ behaves identically to $\underline{\mathrm{Pr}}'_C(\theta_2|X)$ until $(0.5, 0.5)$ is rejected; it reaches about 0.5 around this point, but instead of jumping to 1 it more gradually increases and tends to 1.

5 Discussion

Although all three modified definitions of upper posterior probability of θ achieved the goal of becoming or tending to zero in the situation where $\mathcal{P}_\theta$ doesn't contain the true distribution F, they perhaps do so with varying degrees of success. The apparently most well-founded of the three, $\overline{\mathrm{Pr}}_C(\theta|X)$ will very often jump in an extreme fashion; in the example, with $\theta = \theta_1$, it jumps at one point up to 1, and stays there until it jumps down to 0. This restricts its usefulness, as we're not getting a clear idea of the way things are tending.

$\overline{\mathrm{Pr}}'_C(\theta|X)$ will jump to 0 if θ becomes implausible; it also exhibits typically very much smaller 'discontinuities' when various τ become implausible, and so drop out of the denominator. The definition of this upper and lower probability seems in some sense natural, but it could do with a stronger justification. Smoother behaviour, as exhibited by $\overline{\mathrm{Pr}}''_C(\theta|X)$, is clearly desirable; however, its modification from $\overline{\mathrm{Pr}}'_C$ also needs justification. It would be worth exploring other possible methods of smoothing; in particular it would be interesting to see if the non-standard methods of combining imprecise priors and likelihoods in Cano et al. [1] could be usefully applied or adapted for this type of situation.

Another example of arguably counter-intuitive behaviour of upper and lower probability is illustrated by the following example: $\Theta = \{\theta, \tau\}$ where P_θ is just the singleton set consisting of the correct model F. P_τ also contains F, but as an interior point, i.e., for all $x \in V$, $L_\tau(x) < F(x) < U_\tau(x)$. The prior is uniform, $\rho(\theta) = \rho(\tau) = 0.5$. One might imagine that for large N the lower probability of θ given X would perhaps tend to 0.5. However it will almost certainly not tend to a limit. With probability one, for all sufficiently large N, the observed distribution R will be in $\mathcal{P}_\tau$. Then $N(\overline{\mathcal{Q}}_\tau(R) - \underline{\mathcal{Q}}_\theta(R))$ equals $N(\mathcal{Q}(R, R) - \mathcal{Q}(F, R))$, which equals $D(F, R)$. By the results in 4.2, its expected value tends to $\frac{1}{2}(|V| - 1)$, and the variance similarly. So, if $D(F, R)$ is close to its expected value, the posterior lower probability of θ will be about $\exp(-\frac{|V|-1}{2})$, so will be extremely small unless V is small.

The method in 4.2 might be adapted to deal with this problem, by rejecting P such that the observed value $D(P, R)$ is much smaller than we would expect—i.e., rejecting the P which are *too close* to the observed distribution R. However, rejecting these, which give rise to unusually high likelihoods seems rather different to rejecting the furthest distributions which give rise to unusually low values of likelihood. In any case, it just reduces the problem rather than eliminating it, so extensions of the idea, as well as other possible approaches should be explored.

Possible Generalisation

The scenario described here involves a set of models, where for each model θ there is associated a (non-convex) family of probability measures on $\mathcal{X}$ (the set of all possible sequences X). An element X is observed and we want to learn about which models θ are more or less likely given X. The method in this paper involves rejecting some of these probability measures in various ways, to attempt to obtain more intuitive upper and lower probabilities; in particular in 4.2, it is (roughly speaking) the measures with the lowest likelihoods which are rejected. The set $\mathcal{X}$ has a very special structure, as do the measures considered on $\mathcal{X}$ (their independence properties). It would be worth exploring if this type of approach, based on notions of which measures are implausible given the data, could usefully be applied more widely, where $\mathcal{X}$ is replaced by a set with a different structure, and with other types of constraints on probability measures.

Acknowledgements

I am grateful to Fabio Cozman for discussions beginning at ISIPTA'99 from which this work germinated. Thanks also to various friends and relations for giving me their intuitions about the example, and to the reviewers for their thought-provoking and insightful comments.

References

[1] J. E. Cano, S. Moral, J. F. Verdegay-López, Combination of Upper and Lower Probabilities, *Proceedings of Uncertainty in Artificial Intelligence (UAI91)*, ed. P. B. D'Ambrosio, Smets, P. Bonissone, Morgan Kaufmann Publishers, 61–68, 1991.

[2] F. G. Cozman, Computing Posterior Upper Expectations, *ISIPTA'99*, 131–140, 1999.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications.* Third edition, John Wiley and Sons, New York, London, 1968.

[4] P. Walley, and T. L. Fine, Towards a Frequentist Theory of Upper and Lower Probability *The Annals of Statistics* Vol. 10, No.3, 741–761, 1982.

[5] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.