Weight Chapter Outline

9/1/2022

Contents

1 Introduction

M's comment: Nice exmaple and discussion. The only thing I would change (besides being clearer on certain stuff, see comments below) is make clear that we are after three strategies: precise probabilism (use precise probabilities for rando match probabilities), interval approach (senstivity analysys), and then higher-order approach. I think that sequence of moves should emerge from the discussion as it foreshadoes what is to come.)

Consider two different items of match evidence:[1] The suspects dog's fur matches the dog fur found in a carpet wrapped around one of the bodies (dog). A hair found on one of the victims matches that of the suspect (hair). What are the fact-finders to make of this evidence? To start with, some probabilistic evaluation thereof should be useful.

[1]These are stylized after two items of evidence in the notorious Wayne Williams case. Probabilities have been slightly but not unrealistically shifted to be closer to each other to make a conceptual point. The original probabilities were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair.

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

Accordingly, an expert testifies that the probability of a random person's hair matching the reference sample is about 0.0253, and it so happens that so is the probability of a random dog's hair matching the reference sample, 0.0256.[2] You assume that the probabilities of matches if the suspect (respectively, the suspect's dog) is the source is one, and that these probabilities of a match are independent of each other conditional on either truth value of the source hypothesis (source and ¬source). Then, to evaluate the total impact of the evidence on the source hypothesis you calculate:

$$\begin{aligned}\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\neg\mathsf{source}) &= \mathsf{P}(\mathsf{dog}|\neg\mathsf{source}) \times \mathsf{P}(\mathsf{hair}|\neg\mathsf{source}) \\ &= 0.0252613 \times 0.025641 = 6.4772626 \times 10^{-4}\end{aligned}$$

This seems like a low number. To get a better grip on how this should be interpreted, the expert shows you how the posterior depends on the prior, given this evidence (Figure 1. The posterior of .99 is reached as soon as your prior is higher than 0.061.

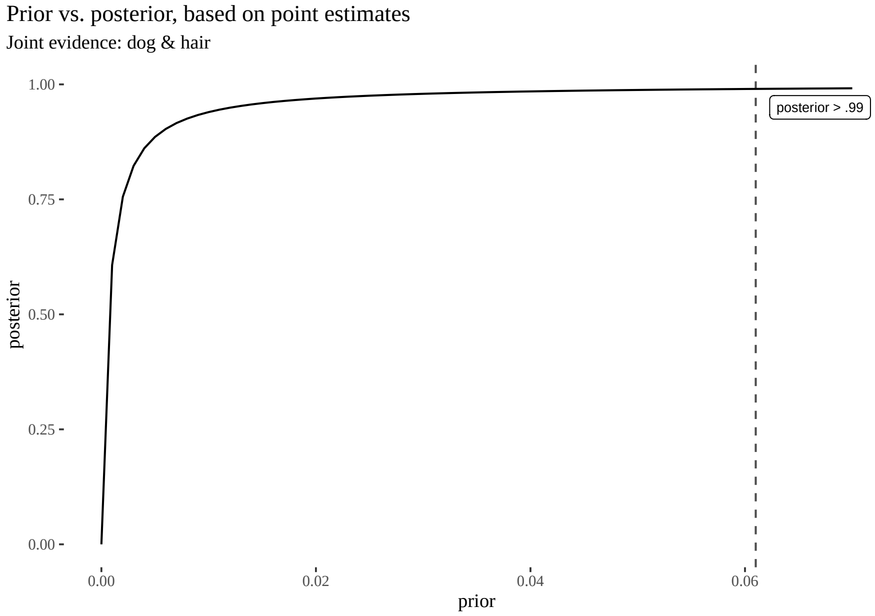

Figure 1: Impact of dog fur and human hair evidence on the prior, point estimates.

While perhaps not sufficient for accepting the source hypothesis, the evidence seems pretty solid: a minor additional piece of evidence could tip the scale. But then, you reflect on what you have been told and ask the expect: *wait, but how do you know these exact point probabilities? There must be some aleatory uncertainties around these estimates, and we should pay attention to these!* The expert agrees, and tells you that in fact the hair evidence estimate is based on 29 matches found in a database of size 1148, and the dog evidence estimate was based on finding two matches in a reference class of size 78.

M: conviction might not be the right word here; R: better now?

Well, that means the point estimates did not tell us the whole story, you think. What to do next? You might try to factor what you have been just informally told into your evaluation, but unless you have some training in probability, you might have hard time doing this correctly. So instead, you push the expert further: *well, with a 99% margin of errors, what are the ranges for these estimates, what are the worst-case and best-case scenarios?* The expert gives you their estimate of credible intervals.[3] He says: *if we start with uniform priors, the intervals are are (.015,.037) for hair and (.002, .103) for fur.*

M: This paragrph is too complicated. Can't we just say: suppose we use these intervals (and then specific how they were arrived at in a footnote). R: no, we can't. The fact that intervals result from margins of errors is important, as keeping the error margins plays an important role later on. I tried to simplify the paragraph though, moving the stuff about confidence intervals into a footnote.

[2] We modified the actual reported probabilities slightly to emphasize the point that we will elaborate further on: the same first-order probabilities, even if they sound precise, might result from various items of evidence connected to various levels of second-order uncertainty.

[3] Roughly, the 99% credible interval is the narrowest interval to which the expert thinks the true parameter belongs with probability .99. For a discussion of what credible intervals are, how they differ from confidence intervals, and why confidence intervals should not be used, see the discussion in Chapter XXXX.

With good intentions, you calculate the estimate that is the most charitable to the suspect. $\mathsf{P}_{char}(\mathsf{dog} \wedge \mathsf{hair}|\neg \mathsf{source}) = .037 * .103 = .003811$. This number is around 5.88 times greater than the original estimate! You ask what the impact of evidence on the prior would be given this scenario, and the answer is that now the prior needs to be higher than 0.274 for the posterior to be above .99 (Figure 2). You are not convinced that the evidence is fairly strong anymore.

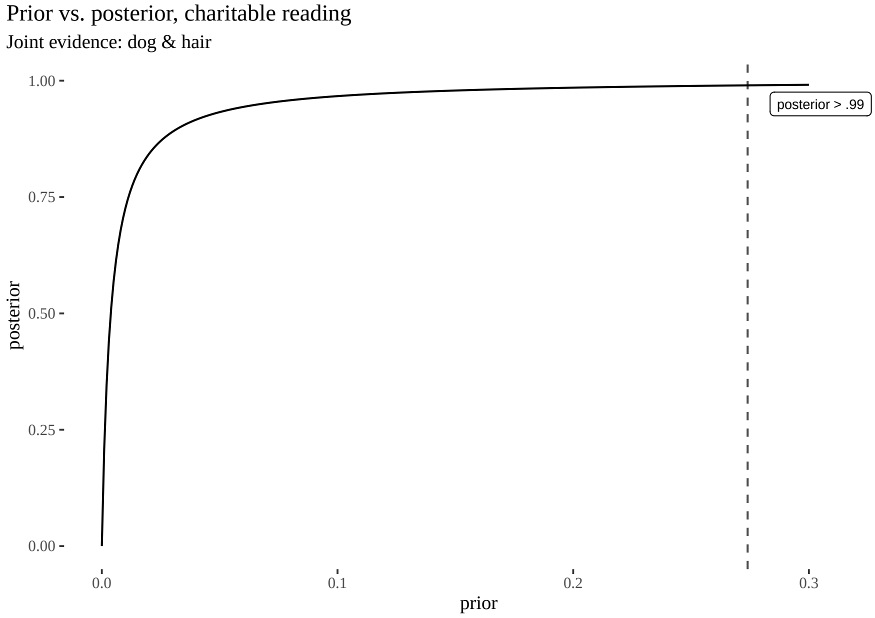

Figure 2: Impact of dog fur and human hair evidence on the prior, charitable reading.

But you made an important blunder. Just because the worst-case probability estimate for one event is x and the worst-case probability estimate for another independent event is y, it does not follow that the worst-case probability estimate for their conjunction is xy, if the margin of error is kept fixed. The intuitive reason is quite simple: just because the probability of an extreme (or larger absolute) value x for one variable is .01, and so it is for the value y of another independent variable, it does not follow that the probability that those two independent variables take values x and y simultaneously is the same. This probability is actually much smaller.

In fact, if you knew what distributions the expert used (it should have been beta distributions in this context), you could work your way back and calculate the .99 highest posterior density interval for the conjunction, which is $(0.000023, 0.002760)$. The proper charitable reading would then require the prior to be above .215 for the posterior to be above .99. Still not enough to convict, but at least now we worked out the consequences of the aleatory uncertainties involved provided the margin of error is fixed. Is this good enough?

Well, it seems the interval presentation instead of doing us good led us into error — the general phenomenon is that intervals do **not** contain enough information to reliably reason about such things as margins of error. Even if we are happy with the interval that we obtained, we will not be able to correctly obtain a new interval once a new item of evidence is included. That is, unless we proceed through the densities.

Another problem is that looking at intervals might be useful if the underlying distributions are fairly symmetrical. But in our case, they might not be. For instance, Figure 3 illustrates are the beta densities for dog fur and human hair, together with sampling-approximated density for the joint evidence. Crucially, the distribution is not symmetric, and so switching the margin of error moves the right edge of the interval much faster towards lower values. If you were only informed about the edges of the interval, you would be oblivious to such phenomena and the fact that the most likely value does **not** simply lie in the middle between the edges of the interval. Just because the parameter lies in an interval with some posterior probability, it does not mean that the ranges near the edges of the interval are equally likely—it still might be the case that the bulk of the density is closer to one of the edges, and therefore relying on the edges only might lead us to either overestimate or underestimate the risk at play.

M: What are reliability and margin of error?

M: This point seems crucial. How to combine intervals? Might need more explanation.R: not sure, what else would you like me to explain here?

M: So the point is that the intervals might be the same but the densities might be different. How do we get these densities though? R: More: the point is, the densities might not be symmetric, in which case intervals might be misleading. How do we

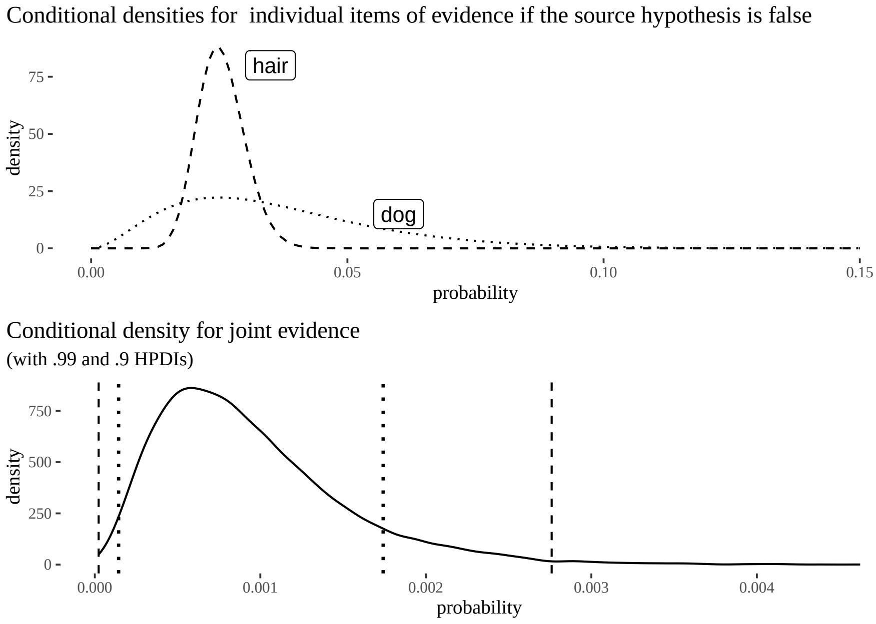

Figure 3: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

This means that a better representation of the uncertainty involving the dependence of the posterior on the prior involves multiple possible lines whose density mirrors the density around the probability of the evidence (Figure 4).

M: Interesting, but from the last graph, the prior needed for .99 posterior has to be around 0.1, though we are given a distribution. So what what are we to make of this?

R: GET BACK TO THIS

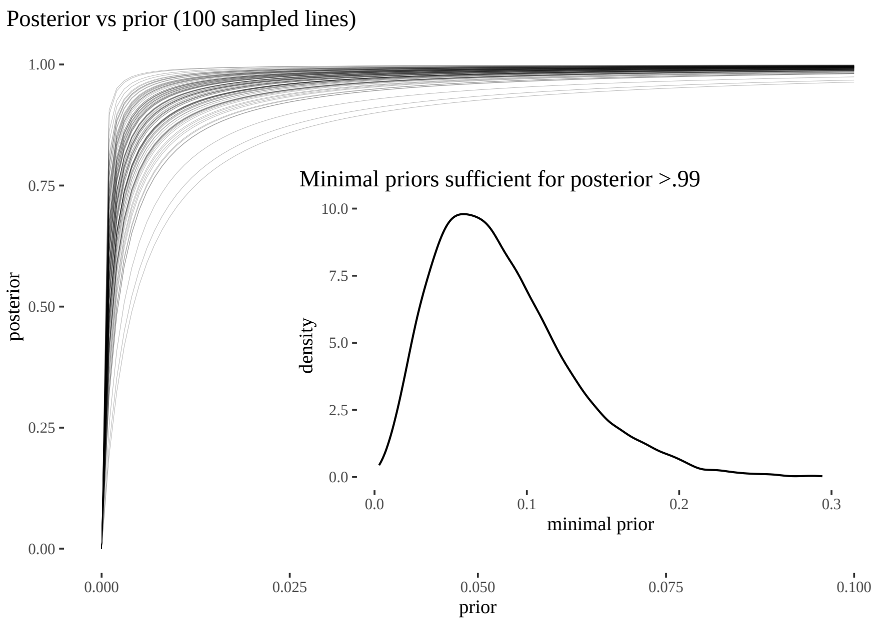

Figure 4: 100 lines illustrating the uncertainty about the dependence of the posterior on the prior given aleatory uncertainty about the evidence, with the distribution of the minimal priors required for the posterior to be above .99.

Revise this structure description once the chapter is done

This is the gist of our chapter: whenever honest density estimates are available (and they should be available for match evidence evaluation methods whose reliability has been properly studied), it is those densities that should be reported and used in further reasoning. This avoids hiding actual aleatory uncertainties under the carpet, and allows for more correct reasoning where interval-based representation might either lead one astray or leave one oblivious to important probabilistic considerations.

The rest of this chapter expands on this idea in a few dimensions. First, it places it in the context of philosophical discussions surrounding a proper probabilistic representation of uncertainty. The main alternatives on the market are precise probabilism and imprecise probabilism. We argue that both options are problematic and should be superseded by the second-order representation whenever possible. Second, having gained this perspective, we visit a recent discussion in the forensic science literature, where a prominent proposal is that the experts, even if they use densities, should integrate and present only point estimates to the fact-finders. We disagree. Third, we explain how the approach can be used in more complex situations in which multiple items of evidence and multiple propositions interact—and the idea is that such complexities can be handled by sampling from distributions and approximating densities using multiple Bayesian Networks in the calculations. Last but not least, we turn to the notion of weight of evidence. Having distinguished quite a few notions in the vicinity, we explain how the framework we propose allows for a more successful explication and implementation of the notion of weight of evidence than the ones currently available on the market.

M's comment: On the structure of the chapter. So basically, the strcture of the chapter would be: (1) higher order approach in general, (2) illustration 1 (first application): DNA evidence debate Taroni, etc., (2) illustration 2 (second application): combing items of evidence, complex cases, etc. and (3) illustration 3 (theory): modeling weight.

M's comment: We might concede that the higher order approach outlined in the introduction might not always be pratically useful, or its applicabiolity might be limited. Not sure. Need to discuss. Perhaps once the discussion about DNA evidence is in place, this will become more clear. One question might be: what kind of information do we need to get the higher order approach going? Regardless of its applicability—we can say, we want to leave this decision to practionares in the end—the higher order approach has important theorical significance because it allows to model notions such as weight.

2 Three probabilisms

This section outlines three version of probabilism: precise, imprecise and higher-order. Precise probabilism, as the name suggests, posits that an agent' credal state is modeled by a single, precise probability measure. Imprecise probabilism replaces precise probabilities by sets of probability measures, while higher-order probabilism relies on distributions over parameter values. There are good reasons to abandon precise probabilism and endorse higher-order probabilism. Imprecise probabilism is a step in the right direction, but as we will see, it suffers from too many difficulties of its own.

2.1 Precise Probabilism

Precise probabilism (PP) holds that a rational agent's uncertainty about a hypothesis H is to be represented as a single, precise probability measure. This is an elegant and simple theory. But representing our uncertainty about a proposition in terms of a single, precise probability runs into a number of difficulties. Precise probabilism fails to capture an important dimension of how our uncertainty connects with the evidence we have or have not obtained. Consider the following simple examples (we will be using examples with coins for a bit, but the points are general and hold for all sampling based frequency estimation methods, including those of random match probability for various pieces of forensic evidence):

> **No evidence v. fair coin** You are about to toss a coin, but have no evidence whatsoever about its bias. You are completely ignorant. Compare this to the situation in which you know the coin is fair.

R: isolated scenario from PP, check.

Sticking to PP you follow the principle of insufficient evidence and assign probability .5 to the outcome being *heads* in both cases.

> **Learning from ignorance** You start tossing the coin with unknown bias, toss it ten times and observe *heads* five times. You started with your bias estimate at .5, but you also end with your bias estimate at .5. Clearly, you have learned something, but whatever that is, it is not captured in the representation recommended by PP. Suppose you toss further, observing

50 *heads* in 100 tosses. Again, your epistemic situation has changed, but it is hard to see how this shift could be represented on PP.

Clearly, precise probabilism has difficulties modeling such situations.[4] The examples suggest that precise probabilism is not appropriately responsive to evidence. It ends up assigning a probability of .5 to situations in which one's evidence is quite different: when no evidence is available about the coin's bias; when there is little evidence that the coin is fair (say, after only 2 draws); when there is strong evidence that the coin is fair (say, after 1000 draws).[5]

M: Should we add the bean example in main text? Sample size is different which is key in the earlier examples about match evidence? Why put it in a footnote? R: meh, I thought it's too repetitive. But I added a bit about sample size in the coin example. Check.

2.2 Imprecise Probabilism

What if we give up the assumption that probability assignments should be precise? **Imprecise probabilism** (IP) holds that an agent's credal stance towards a hypothesis H is to be represented by means of a *set of probability measures*, typically called a representor $\mathbb{P}$, rather than a single measure P. The representor should include all and only those probability measures which are compatible (in a sense to be specified) with the evidence. For instance, if an agent knows that the coin is fair, their credal state would be captured by the singleton set {P}, where P is a probability measure which assigns .5 to *heads*. If, on the other hand, the agent knows nothing about the coin's bias, their credal state would rather be represented as the set of all probabilistic measures, as none of them is excluded by the available evidence. Note that the set of probability measures does not represent admissible options that the agent could legitimately pick from. Rather, the agent's credal state is essentially imprecise and should be represented by means of the entire set of probability measures.[6]

Imprecise probabilism, at least *prima facie*, offers a straightforward picture of learning from evidence, that is a natural extension of the classical Bayesian approach. When faced with new evidence E between time t_0 and t_1, the representor set should be updated point-wise, running the standard Bayesian updating on each probability measure in the representor:[7]

REF Kyburg. Probability and the Logic of Rational Belief. Wesleyan University Press, Middletown Connecticut, 1961 and H. E. Kyburg and C. M. Teng. Uncertain Inference. Cambridge University Press, Cambridge, 2001.

$$\mathbb{P}_{t_1} = \{\mathsf{P}_{t_1} | \exists\, \mathsf{P}_{t_0} \in \mathbb{P}_{t_0}\ \forall H\ \left[\mathsf{P}_{t_1}(H) = \mathsf{P}_{t_0}(H|E)\right]\}.$$

The hope is at least that if we start with a range of probabilities that is not extremely wide, point-wise learning will behave appropriately.[8] For instance, if we start with a prior probability of *heads* equal to .4 or .6, then those measure should be updated to something closer to .5 once we learn that a given coin has already been tossed ten times with the observed number of heads equal 5 (call this evidence E). This would mean that if the initial range of values was $[.4, .6]$ the posterior range of values should be more narrow. But even this seemingly straightforward piece of reasoning is hard to model if we want to avoid using densities. For to calculate $\mathsf{P}(\mathsf{heads}|E)$ we need to calculate $\mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads})$ and divide it by $\mathsf{P}(E) = \mathsf{P}(E|\mathsf{heads})\mathsf{P}(\mathsf{heads}) + \mathsf{P}(E) = \mathsf{P}(E|\neg\mathsf{heads})\mathsf{P}(\neg\mathsf{heads})$. The tricky part is obtaining the

[4] Examples of this sort date back to C. S. Peirce, who in his 1872 manuscript 'The Fixation of Belief' (W3 295) comments: "when we have drawn a thousand times, if about half [of the beans] have been white, we have great confidence in this result … a confidence which would be entirely wanting if, instead of sampling the bag by 1000 drawings, we had done so by only two.' ' Similar remarks can be found in Peirce's 1878 *Probability of Induction*. There, he also proposes to represent uncertainty by at least two numbers, the first depending on the inferred probability, and the second measuring the amount of knowledge obtained; as the latter, Peirce proposed to use some dispersion-related measure of error (but then suggested that an error of that estimate should also be estimated and so, so that ideally more numbers representing errors would be needed).

[5] Precise probabilism suffers from other difficulties. For example, it has problems with formulating a sensible method of probabilistic opinion aggregation Stewart & Quintana (2018). A seemingly intuitive constraint is that if every member agrees that X and Y are probabilistically independent, the aggregated credence should respect this. But this is hard to achieve if we stick to PP (Dietrich & List, 2016). For instance, a *prima facie* obvious method of linear pooling does not respect this. Consider probabilistic measures p and q such that $p(X) = p(Y) = p(X|Y) = 1/3$ and $q(X) = q(Y) = q(X|Y) = 2/3$. On both measures, taken separately, X and Y are independent. Now take the average, $r = p/2 + q/2$. Then $r(X \cap Y) = 5/18 \neq r(X)r(Y) = 1/4$.

[6] For the development of imprecise probabilism, see (Fraassen, 2006; Gärdenfors & Sahlin, 1982; Joyce, 2005; Kaplan, 1968; Keynes, 1921; Levi, 1974; Sturgeon, 2008; Walley, 1991), (S. Bradley, 2019) is a good source of further references.

[7] Imprecise probabilism shares some similarities with what we might call **interval probabilism** due to [KYBURG 1961]. On interval probabilism, precise probabilities are replaced by intervals of probabilities. On imprecise probabilism, instead, precise probabilities are replaced by sets of probabilities. This makes imprecise probabilism more general, since the probabilities of a proposition in the representor set do not have to form a closed interval. Moreover, learning on Kyburg's approach is somewhat idiosyncratic and is strongly connected to reference classes and selection and reshaping rules for intervals. See [PEDDEN] for a introduction. As we have already signaled, as intervals do not contain probabilistic information sufficient to guide reasoning with multiple propositions and items of evidence, we keep our focus on IP, which is the more promising candidate method.

[8] The hope is also that IP offers a feasible aggregation method (Elkin & Wheeler, 2018; Stewart & Quintana, 2018): just put all representors together in one set, and voil'a! However, this is a very conservative method which quickly leads to extremely few points of agreement, and we are not aware of any successful practical deployment of this method.

conditional probabilities $\mathsf{P}(E|\mathsf{heads})$ and $\mathsf{P}(E|\neg\mathsf{heads})$ in a principle manner without explicitly going second-order, estimating the parameter value and using beta distributions.

The situation is even more difficult if we start with complete lack of knowledge, as imprecise probabilism runs into the problem of **belief inertia** (Levi, 1980). Say you start tossing a coin knowing nothing about its bias. The range of possibilities is $[0,1]$. After a few tosses, once you observed at least one tail and at least one heads, you have excluded the measures assigning 0 or 1 to *heads*. But what else have you learned? Let's charitably agree that each particular measure from your initial representor gets updated to one that is closer to .5, but also now each value in your original interval can be obtained by updating some *other* measure in your original representor on the evidence, and the picture does not change no matter how many observations you have made. For instance, some measure that initially assigned .4 to heads might now assign .45 to heads, but now a measure that assigned .37 to heads has been updated to one that assigns .4 to heads. Thus, if you are to update your representor point-wise, you will end up with the same representor set. Consequently, the edges of your resulting interval will remain the same. In the end, it is not clear how you are supposed to learn that the proportion of beans is such and such.[9]

Some downplay the problem of belief inertia. They insist that vacuous priors should not be used and that imprecise probabilism gives the right results when the priors are non-vacuous. After all, if you started with knowing truly nothing, then perhaps it is right to conclude that you will never learn anything. Another strategy is to say that, in a state of complete ignorance, a special updating rule should be deployed.[10] But no matter what we think about belief inertia, other problems plague imprecise probabilism. Three more problems are particularly pressing.

One problem is that imprecise probabilism fails to capture intuitions we have about evidence and uncertainty in a number of scenarios. Consider this example:

> **Even v. uneven bias:** You have two coins and you know, for sure, that the probability of getting heads is .4, if you toss one coin, and .6, if you toss the other coin. But you do not know which is which. You pick one of the two at random and toss it. Contrast this with an uneven case. You have four coins and you know that three of them have bias .4 and one of them has bias .6. You pick a coin at random and plan to toss it. You should be three times more confident that the probability of getting heads is .4. rather than .6.

The first situation can be easily represented by imprecise probabilism. The representor would contain two probability measures, one that assigns .4. and the other that assigns .6 to the hypothesis 'this coin lands heads'. But imprecise probabilism cannot represent the second situation, at least not without moving to higher-order probabilities, in which case it is no longer clear whether the object-level imprecision performs any valuable task.[11]

Second, besides descriptive inadequacy, an even deeper, foundational problem exists for imprecise probabilism. This problem affects imprecise probabilism, but not precise probabilism. It arises when we reflect on the notion of the accuracy of imprecise credal states. A variety of workable **scoring rules** for measuring the accuracy of a single credence function, such as the Brier score, are available. One key feature that some key candidates have is that they are *proper*: any agent will score her own credence function to be more accurate than every other credence function. After all, if an agent thought a different

[9] Here's another example from (Rinard, 2013). Either all the marbles in the urn are green (H_1), or exactly one tenth of the marbles are green (H_2). Your initial credence $[0,1]$ in each. Then you learn that a marble drawn at random from the urn is green (E). After conditionalizing each function in your representor on this evidence, you end up with the the same spread of values for H_1 that you had before learning E, and no matter how many marbles are sampled from the urn and found to be green.

[10] (Elkin, 2017) suggests the rule of *credal set replacement* that recommends that upon receiving evidence the agent should drop measures rendered implausible, and add all non-extreme plausible probability measures. This however, is tricky: one needs a separate account of what makes a distribution plausible or not. Elkin admits that he has no solution to this: "But how do we determine what the set of plausible probability measures is relative to E? There is no precise rule that I am aware of for determining such set at this moment, but I might say that the set can sometimes be determined fairly easily" [p. 83] He goes on to a trivial example of learning that the coin is fair and dropping extreme probabilities. This is far from a general account. One also needs a principled account of why one should use a separate special update rule when starting with complete ignorance.

[11] Other scenarios can be constructed in which imprecise probabilism fails to capture distinctive intuitions about evidence and uncertainty; see, for example, (Rinard, 2013). Suppose you know of two urns, GREEN and MYSTERY. You are certain GREEN contains only green marbles, but have no information about MYSTERY. A marble will be drawn at random from each. You should be certain that the marble drawn from GREEN will be green (G), and you should be more confident about this than about the proposition that the marble from MYSTERY will be green (M). In line with how lack of information is to be represented on IP, for each $r \in [0,1]$ your representor contains a P with $\mathsf{P}(M) = r$. But then, it also contains one with $\mathsf{P}(M) = 1$. This means that it is not the case that for any probability measure P in your representor, $\mathsf{P}(G) > \mathsf{P}(M)$, that is, it is not the case that RA is more confident of G than of M. This is highly counter-intuitive.

credence is more accurate, they should switch to it. The availability of such scoring rules underlies an array of accuracy-oriented arguments for precise probabilism (roughly, if your precise credence follows the axioms of probability theory, no other credence is going to be more accurate than yours whatever the facts are). When we turn to imprecise probabilism, there are impossibility results to the effect that no proper scoring rules are available for representors. So, as many have noted, the prospects for an accuracy-based argument for imprecise probabilism look dim (Campbell-Moore, 2020; Mayo-Wilson & Wheeler, 2016; Schoenfield, 2017; Seidenfeld, Schervish, & Kadane, 2012). Moreover, as shown by (Schoenfield, 2017), if an accuracy measure satisfies some basic formal constraints, it will never strictly recommend an imprecise stance, as for any imprecise stance there will be a precise one with the same accuracy.

Third, on IP, much is made of the notion of representors containing probability measures compatible with evidence—the idea is that thanks to this feature, imprecise credal stances are evidence-responsive in a way precise probabilistic stances are not. But how exactly does the evidence exclude probability measures? This is not a mathematical question: mathematically (S. Bradley, 2012), evidential constraints are fairly easy to model, as they can take the form of the *evidence of chances* $\{\mathsf{P}(X) = x\}$ or $\mathsf{P}(X) \in [x, y]$, or be *structural constraints* such as "X and Y are independent" or "X is more likely than Y." While it is clear that these constraints are something that an agent can come to accept if offered such information by an expert to which the agent completely defers, it is not trivial to explain how non-testimonial evidence can result in such constraints. Most of the examples in the literature start with the assumption that the agent is told by a believable source that the chances are such-and-such, or that the experimental set-up is such that the agent knows that such and such structural constraint is satisfied. But outside of such ideal circumstances what observations exactly would need to be made to come to accept such constraints remains unclear.[12]

Bradley suggests that "statistical evidence might inform [evidential] constraints [. . . and that evidence] of causes might inform structural constraints" [125-126]. This, however, is far cry from a clear account of how exactly this should proceed. Now, one suggestion might be that once a statistical significance threshold is selected, a given set of observations with a selection of background modeling assumptions yields a credible interval. But this is to admit that to reach such constraints we already have to start with a second-order approach, and drop information about the densities, focusing only on the intervals obtained with fixed margins of errors. But as we already illustrated, if you have the information about densities to start with, there is no clear advantage to going imprecise instead, and there are multiple problems associated with this move. Moreover, such moves require a choice of an error margin, which is extra-epistemic,[13] and it is not clear what advantage there is to dropping information contained in second-order probabilities based on extra-epistemic considerations of this sort.

2.3 Higher order probabilism

There is, however, a view in the neighborhood that fares better: a second-order perspective. In fact, some of the comments by the proponents of imprecise probabilism tend to go in this direction. For instance, Seamus Bradley compares the measures in a representor to committee members, each voting on a particular issue, say the true chance or bias of a coin. As they acquire more evidence, the committee members will often converge on a specific chance hypothesis. He writes:

Maybe add reference to Joyce here as well?

> . . . the committee members are "bunching up". Whatever measure you put over the set of probability functions—whatever "second order probability" you use—the "mass" of this measure gets more and more concentrated around the true chance hypothesis' [BRADLEY p. 157]

[12] And the question is urging: even if you were lucky enough to run into an expert that you completely trust that provides you with a constraint like this, how exactly did the expert come to learn the constraint? The chain of testimonial evidence has to end somewhere! Admittedly, there are straightforward degenerate cases: if you see the outcome of a coin toss to be heads, you reject the measure with $\mathsf{P}(H) = 0$, and similarly for tails. Another class of cases might arise if you are randomly drawing objects from a finite set where the real frequencies are already known, because this finite set has been inspected. But such extreme cases aside, what else? Mere consistency constraint wouldn't get the agent very far in the game of excluding probability measures, as way too many probability measures are strictly speaking still consistent with the observations for evidence to result in epistemic progress.

[13] In forensic evidence evaluation even scholars who disagree about the value of going higher-order agree that interval reporting is problematic, as the choice of a limit or uncertainly level is rather arbitrary [Taroni, Bozza, Biedermann, & Aitken (2015);Sjerps2015Uncertainty].

Note, however, that such bunching up cannot be modeled by imprecise probabilism.[14]

Similarly, Joyce (2005), in a paper defending imprecise probabilism in fact uses a density over chance hypotheses to account for the notion of evidential weight and conceptualizes the weight of evidence as an increase of concentration of smaller subsets of chance hypotheses, without any reference to representors in his explication of the notion of weight (we will get back to his explication when we discuss the notion of weight of evidence).

The idea that one should use higher-order probabilities has also been suggested by critics of imprecise probabilism. For example, Carr (2020) argues that sometimes evidence requires uncertainty about what credences to have. On Carr's approach, one should use vague credences, assigning various weights to probabilities—agent's credence in propositions about either what credences the evidence supports, or about objective chances. Carr, however, does not articulate this suggestion more fully, does not develop it formally, and does not explain how her approach would fare against the difficulties pestering precise ad imprecise probabilism.

Our goal now is to develop a higher-order approach that can handle the problems that imprecise probabilism runs into. The key idea is that uncertainty is not a single-dimensional thing to be mapped on a single one-dimensional scale such as a real line. It is the whole shape of the whole distribution over parameter values that should be taken under consideration. [15] From this perspective, sometimes, when an agent is asked about her credal stance towards X, they can refuse to summarize it in terms of a point value $\mathsf{P}(X)$, instead expressing it in terms of a probability (density) distribution f_x treating $\mathsf{P}(X)$ as a random variable. Coming back to an example we already argued imprecise probabilism cannot handle, when the agent knows that the real chance is either .4 or .6 but the former is three times more likely, she might refuse to summarize her credal state by saying that $\mathsf{P}(H) = .75 \times .4 + .25 \times .6 = .45$.[16] This approach in fact lines up with common practice in Bayesian statistics, where the primary role of uncertainty representation is assigned to the whole distribution, and summaries such as the mean, mode standard deviation, mean absolute deviation, or highest posterior density intervals are only summary ways of representing the uncertainty involved in a given study, to be used mostly due to practical restrictions.

From this perspective, the scenarios we discussed—some of which imprecise probabilism has hard time distinguishing—can be easily represented in the manner illustrated in Figure 5.

[14] Bradley seems to be aware of that, which would explain the use of scare quotes: when he talks about the option of using second-order probabilities in decision theory, he insists that 'there is no justification for saying that there is more of your representor here or there.' ~[p.~195]

[15] Bradley admits this much [90], and so does Konek [59]. For instance, Konek disagrees with: (1) X is more probable than Y just in case $p(X) > p(Y)$, (2) D positively supports H if $p_D(H) > p(H)$, or (3) A is preferable to B just in case the expected utility of A w.r.t. p is larger than that of B.

[16] More generally, on this perspective, the agent might deny that $\int_0^1 x f(x)\, dx$ is their object-level credence in X, if f is the probability density over possible object-level probability values and f is not sufficiently concentrated around a single value for such a one-point summary to do the justice to the complexity of the agent's credal state. Whether this expectation should be used in betting behavior is a separate problem, here we focus on epistemic issues.

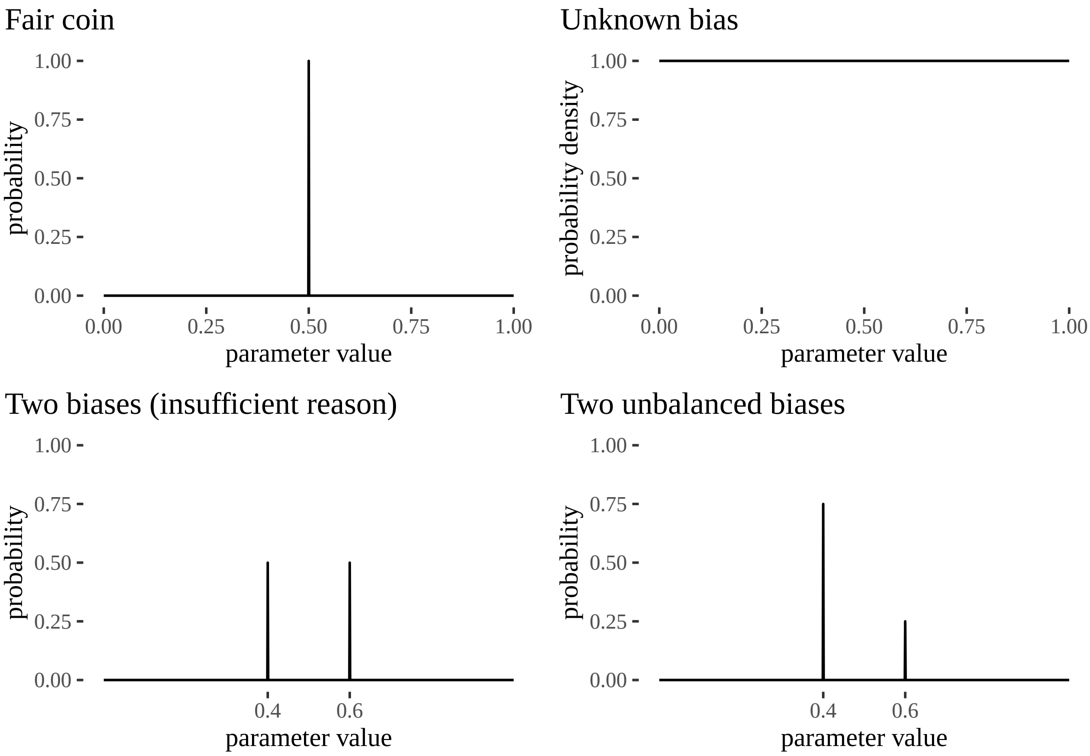

Figure 5: Examples of RA's distributions responding to various types of evidence for typical cases brought up in the literature.

How is learning about frequencies modeled on this approach, assuming independence and constant frequency/probability for all the observations? The Bayes way. You start with some prior density p over the parameter values. For instance, if you start with complete lack of information, p might be uniform. Then you observe the data D which is basically the number of successes s in a certain number of observations n. For each particular possible value θ of the parameter, the probability of D conditional on θ follows the binomial distribution. The probability of D is obtained by integration. That is:

$$
\begin{aligned}
p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)} \\
&= \frac{\theta^s(1-\theta)^{(n-s)}p(\theta)}{\int \theta'^s(1-\theta')^{(n-s)}p(\theta')\,d\theta'}
\end{aligned}
$$

For instance, belief inertia does not arise. If you just start with a uniform density over $[0,1]$ as your prior, use binomial probability as likelihood, observing any non-zero number of heads will exclude 0 and observing any non-zero number of tails will exclude 1 from the basis of the posterior, and the posterior distribution becomes more centered around the parameter estimate as the observations come in. Let's see an example with a grid approximation ($n = 1k$) and coin tossing (grid approximation allows us also to talk about probabilities rather than densities). Our prior is uniform, and then, in subsequent steps, we observe heads, another heads, and then tails. This is what happens with the posterior as we go (Figure 6).

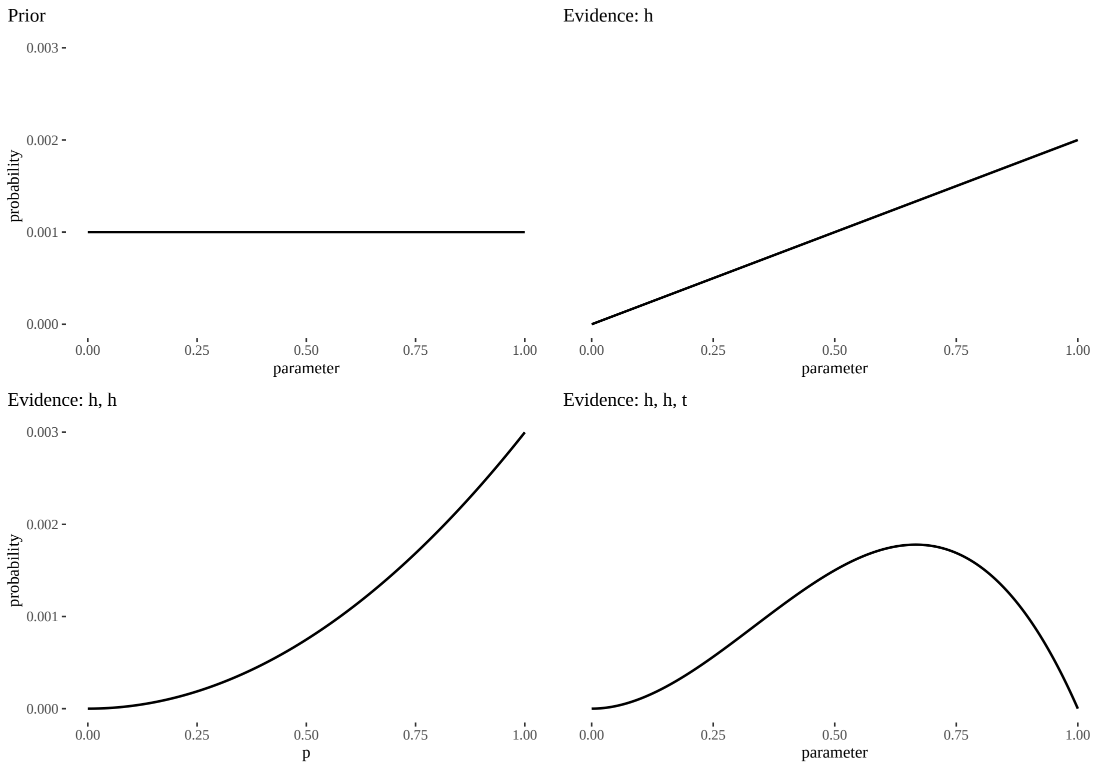

Figure 6: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

3 Examples of applications

3.1 Adding carpet evidence in the Wayne Williams case

3.2 Higher-order probabilities and Bayesian networks

The reader might be worried: how can we handle the computational complexity that comes with moving to higher-order probabilities? The answer is, as long as we have decent ways of either basing densities on sensible priors and data, or eliciting densities from experts (CITE UNCERTAIN JUDMENTS), implementation is not computationally unfeasible, as we can approximate densities using sampling. To illustrate, let us start with a simplified BN developed by CITE FENTON to illustrate how conviction was unjustified in the Clark case (Figure 7). to illustrate a point about the notorious Sally Clark case (Figure 7).[17] The arrows depict relationships of influence between variables. Amurder and Bmurder are binary nodes corresponding to whether Sally Clark's sons, call them A and B, were murdered. These influence whether signs of disease (Adisease and Bdisease) and bruising (Abruising and Bbruising) were present. Also, since son A died first, whether A was murdered casts some light on the probability of son B being murdered.

[17] R. v. Clark (EWCA Crim 54, 2000) is a classic example of how the lack of probabilistic independence between events can be easily overlooked. Sally Clark's first son died in 1996 soon after birth, and her second son died in similar circumstances a few years later in 1998. At trial, the paediatrician Roy Meadow testified that the probability that a child from such a family would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Meadow calculated that therefore the probability of both children dying of SIDS was approximately 1 in 73 million. Sally Clark was convicted of murdering her infant sons (the conviction was ultimately reversed on appeal). The calculation illegitimately assumes independence, as the environmental or genetic factors may predispose a family to SIDS. The winning appeal was based on new evidence: signs of a potentially lethal disease—contrary to what was assumed in the original case—were found in one of the bodies.

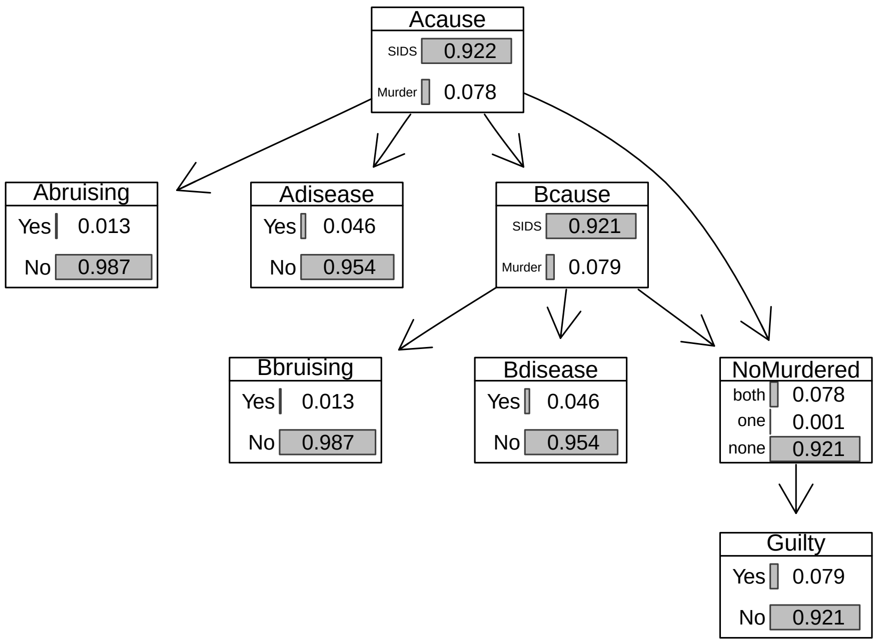

Figure 7: The BN developed by FENTON ET AL., with marginal prior probabilities.

The point to be illustrated was that with a sensible choice of probabilities for the conditional probability tables in the BN, conviction was not justified at any of the major stages (Figure 8).

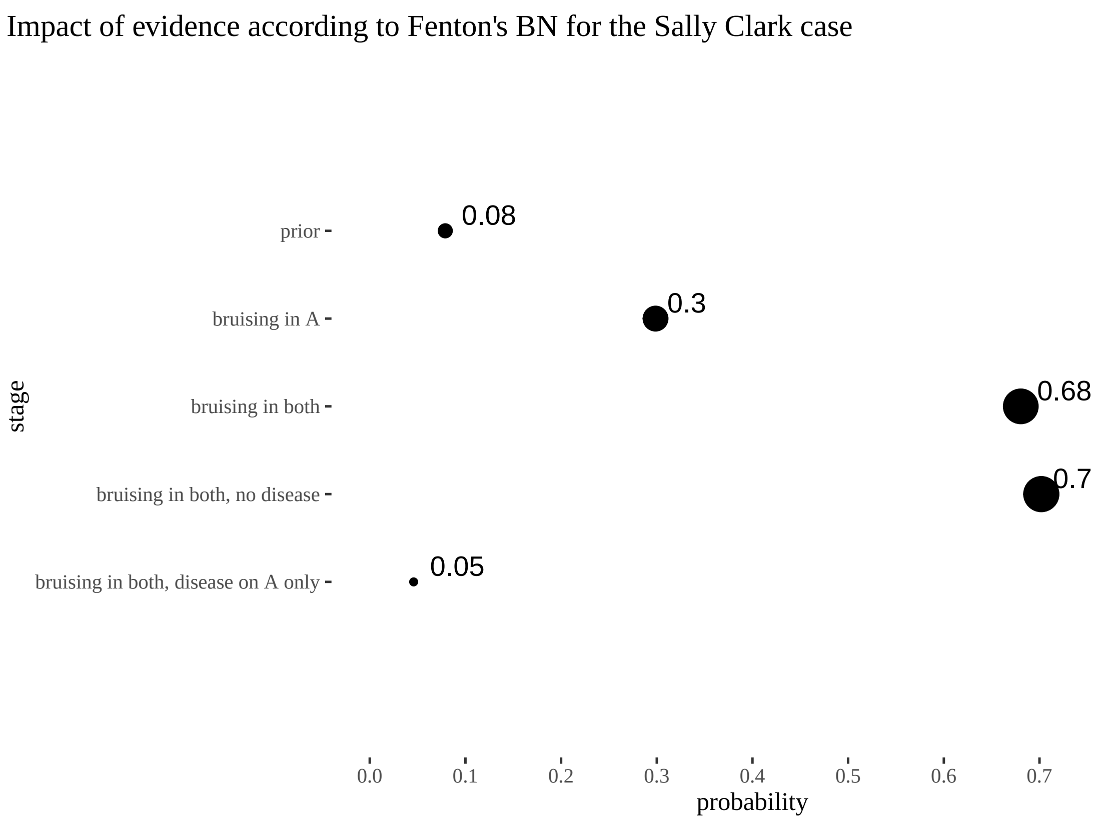

Figure 8: The prior and posterior probabilities for Fenton's Sally Clark BN.

One reason the reader might worry is that the choice of the probabilities is fairly specific, and it is not obvious where such precise values should come from. We have already discussed how frequency and probability estimates usually come at least with some aleatory uncertainty around them that cannot be represented by first-order probabilities. The usual response REFS FOR SENSITIVITY ANALYSIS is that a range of such selections should be tested, perhaps with special focus on extreme but still plausible values. We have already discussed how much care is needed on such approach as it to some extent ignores the shape of the underlying distributions. Crucially, on the sensitivity approach different probability measures (or point estimates) are not distinguished in terms of their plausibility, and so this plausibility is not accounted for in the analysis. Moreover, if in the sensitivity analysis the further decision is guided by the results for the extreme measures, they might play an undeservedly strong role. The best case scenario for my way home from the office is that I find a suitcase with a lot of money inside. The worst case scenario is that I get run over by a bus. What is the quality of the guidance that the consideration of the worst case and the best case scenario provide me with? Limited.

Some of these concerns are at least dampened when we deploy the higher order probabilities in the BN. The general method is as follows. Each particular node in a precise BN has a probability table determined by a finite list of numbers. If it's a root node, its probability table is determined by one number, if it's a node with one parent, its table is determined by two numbers etc. Now, suppose that instead of precise numbers we have densities over parameter values for those determining numbers. Densities of interests can then be approximated by (1) sampling parameter values from the specified distributions, (2) plugging them into the construction of the BN, and (3) evaluating the probability of interest in that precise BN. The list of the probabilities thus obtained will approximate the density fo interest. In what follows we will work with sample sizes of 10k. For instance, your conditional probabilities might look as illustrated in Figure 9. One of them is based on a truncated normal distribution to emphasize that the framework give us much freedom in the specification of distributions.

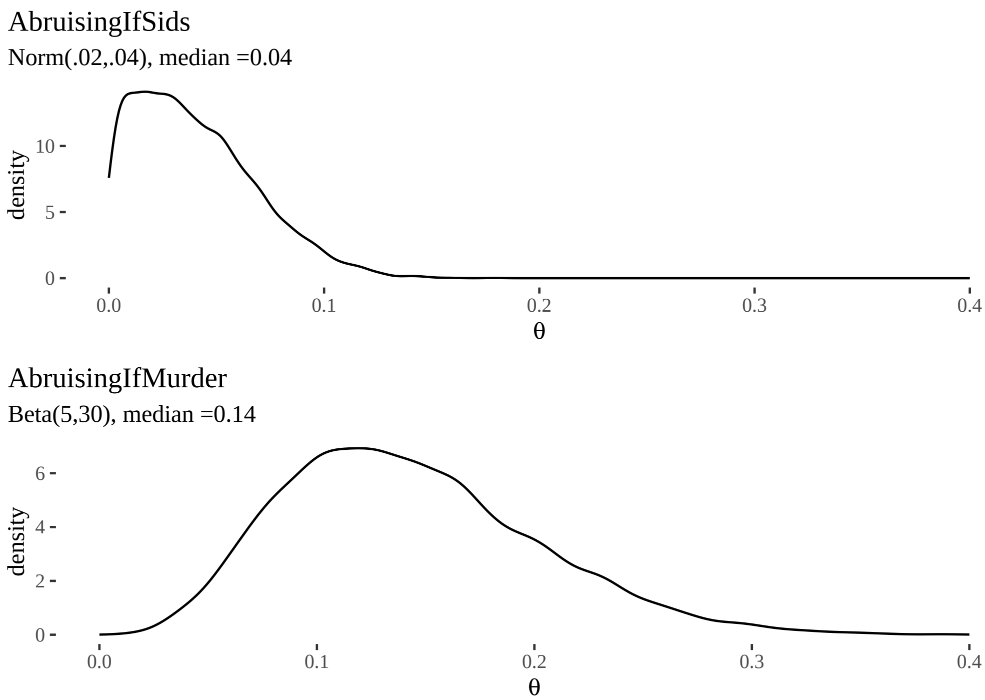

Figure 9: Example of approximated uncertainties about conditional probabilities in the Sally Clark case.

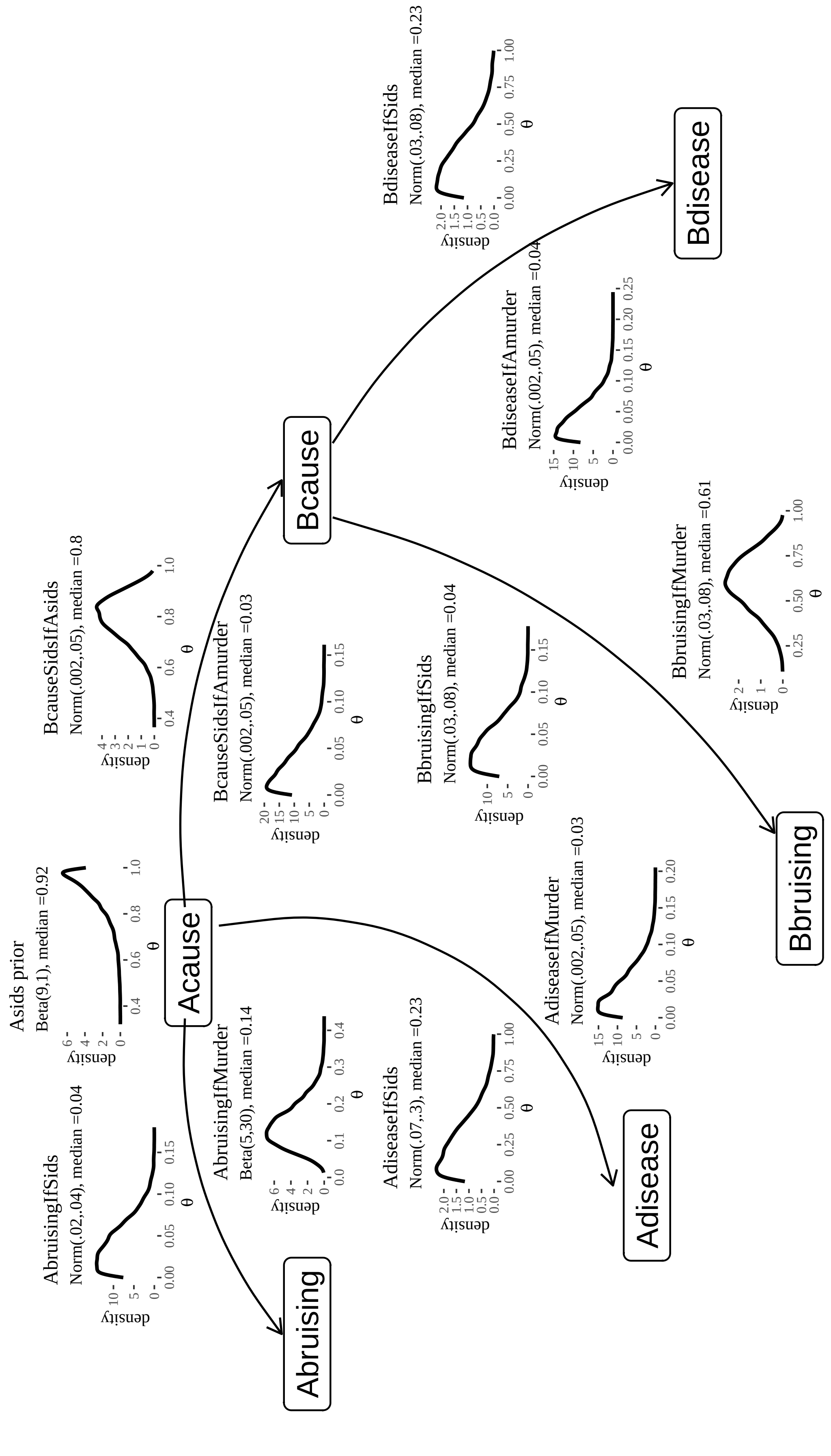

Figure 10: Example of a HOP approach tor the Sally Clark Case approximated by sampling probabilities and constructing 10k BNs.

Using these we can investigate the impact of incoming evidence as it arrives (Figure 11). We start with the prior density for the Guilt node. Then, we update with the evidence of signs of bruising in both children. Next, we consider what would have happened if also both children showed no sign of potentially lethal disease. Finally, we look at the (simplified) evidential situation at the time of the appeal: signs of bruising in both children, and signs of lethal disease discovered in only the first child. One thing to notice that even in the strongest scenario against Sally Clark (third visualization), while the median of the posterior distribution was above .95, the uncertainty around that median is still too wide to legitimize conviction as the lower limit of the 89% HPDI is at .83. This illustrates the idea that taking the point estimates and running with them might lead to overconfidence, and that paying attention to uncertainties about the estimates can make an important difference to our decisions and their accuracy.

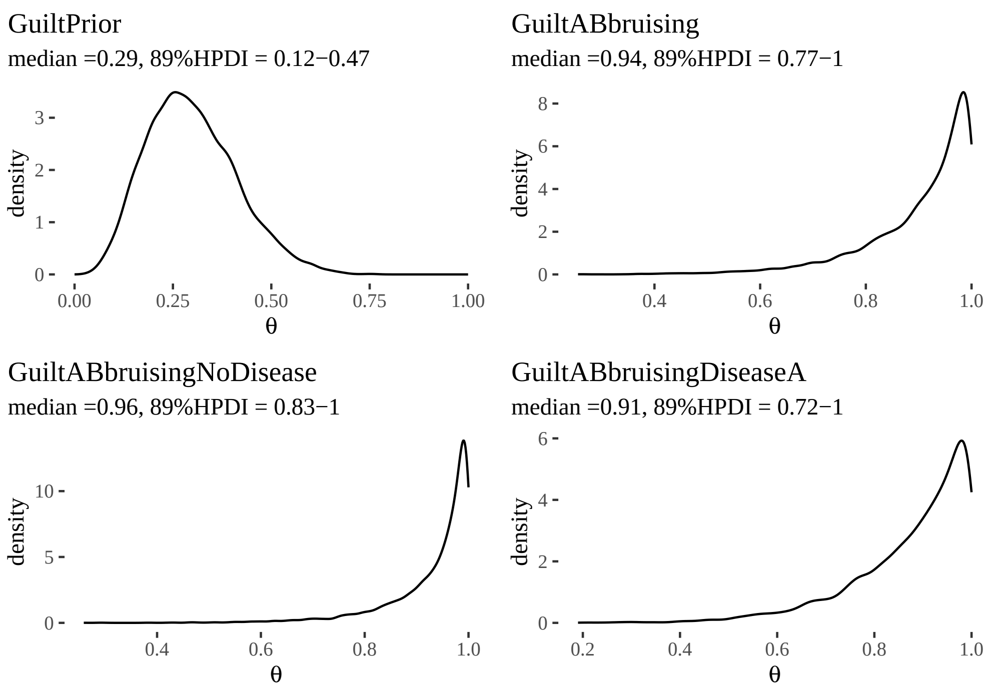

Figure 11: Impact of incoming evidence in the Sally Clark case.

Moreover, if we are interested in likelihood ratios, the same approach can be used: sample from the selected distribution appropriate for the conditional probabilities at hand, then divide the corresponding samples, obtaining a sample of likelihood ratios, approximating the density capturing the recommended uncertainty about the likelihood ratio. For instance, we can use this tool to gauge our uncertainty about the likelihood ratios corresponding to the signs of bruising in son A and the presence of the symptomps of a potentially lethal disease in son A (Figure 12).

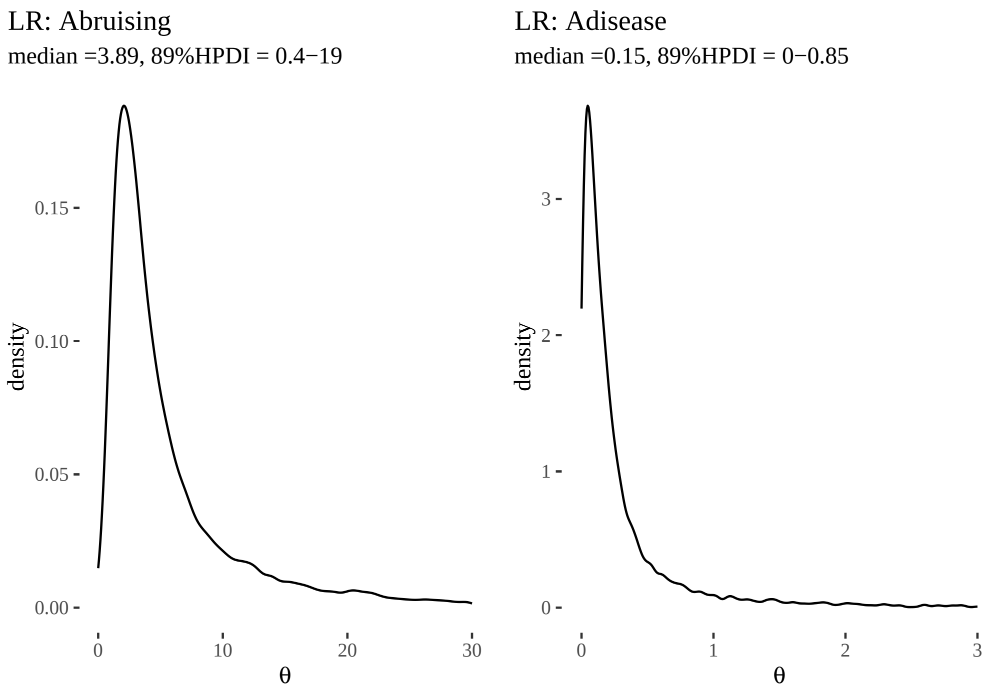

Figure 12: Likelihood ratios forbruising and signs of disease in child A in the Sally Clark case.

Discuss the negation problem?
R: Meh, I never though it was very interesting; but let's talk about this.

3.3 Impact of false positives in DNA identification

Let us get back to the problem we already discussed in Chapter 5: the question of the extent to which the probability of a false positive impacts the value of DNA match evidence. We already argued that the probability of false positives is non-negligible. Here higher-order probability will assist us in thinking through a comment in an important paper on the topic:

> If, as commentators have suggested, the rate of false positives is between 1 in 100 and 1 in 1000, or even less, then one might argue that the jury can safely rule out the prospect that the reported match in their case is due to error and can proceed to consider the probability of a coincidental match. For reasons we will explain more fully below, this argument is fallacious and profoundly misleading . . . As we will explain below, the probability that a reported match occurred due to error in a particular case can be much higher, or lower, than the false positive probability. [@Thomason2003How-the-Probabi, 3]

One option would be to use these interval edges to investigate the consequences of the risk of a false positive in DNA identification. But, we hope to have already convinced the reader, the consequences of doing so might be too skeptical, and it would be much better if we had a sensible distribution to reason with.

Let us use Bayes' theorem again to think about false positives. This time, instead of on likelihood ratios, we focus on posterior probability. The posterior probability of the source hypothesis (S) conditional on the match evidence (E) is:

$$\mathsf{P}(S|E) = \frac{\mathsf{P}(E|S)\mathsf{P}(S)}{\mathsf{P}(E)}$$

$$= \frac{\overbrace{\mathsf{P}(E|S)}^{1}\mathsf{P}(S)}{\underbrace{\mathsf{P}(E|S)}_{1}\mathsf{P}(S) + \underbrace{\mathsf{P}(E|RM)}_{1}\mathsf{P}(RM) + \underbrace{\mathsf{P}(E|FP)}_{1}\mathsf{P}(FP)}$$

$$= \frac{\mathsf{P}(S)}{\mathsf{P}(S) + \mathsf{P}(RM) + \mathsf{P}(FP)}$$

For simplicity we take the false negative rate to be zero, that is we assume $\mathsf{P}(E|S) = 1$. We also assume there are three ways the evidence could arise: the source hypothesis is true, a random match (RM) occurred, or we are dealing with a false positive (FP).

Now, let us start with calculations which ignore false positive risk and take $FP = 0$. Suppose RMP is $10e-9$ (as in some of the examples we discussed in Chapter 5). The relation between priors and the posteriors is illustrated with dashed orange line in Figure 15 (to be explained fully later on), and the evidence seems quite strong: the minimal prior sufficient for the posterior to be above .99 is 0.001.

Next, let us run charitable calculations with the upper edge of the interval offered for FP, 0.01. Then, the posterior of .99 is reached only once the prior is above .99, and the evidence seems rather weak. So which is it? None. Running with the extreme scenario over-appreciates and running with the point estimate under-appreciates the uncertainty involved. What matters between the edges matters.

Consider two scenarios. In both you think that with 99% certainty the false positive rate is between 0.001 and 0.01. On one approach you think that any value between these values (with a slight leeway on top) is equally likely, while on the other you think that it is 50% likely that FP is below .0033. The latter distribution, while being centered closer to zero has another features: it is long-tailed, so at the same time you do not think that FPs above .01 are impossible—you allow for the rare possibility of a false positive being much higher (say, if some specific conditions or circumstances arise, but you have no knowledge of how to identify them, because extensive studies on false positives are only forthcoming), in line with the passage quoted above. These distributions are illustrated in Figure 13.

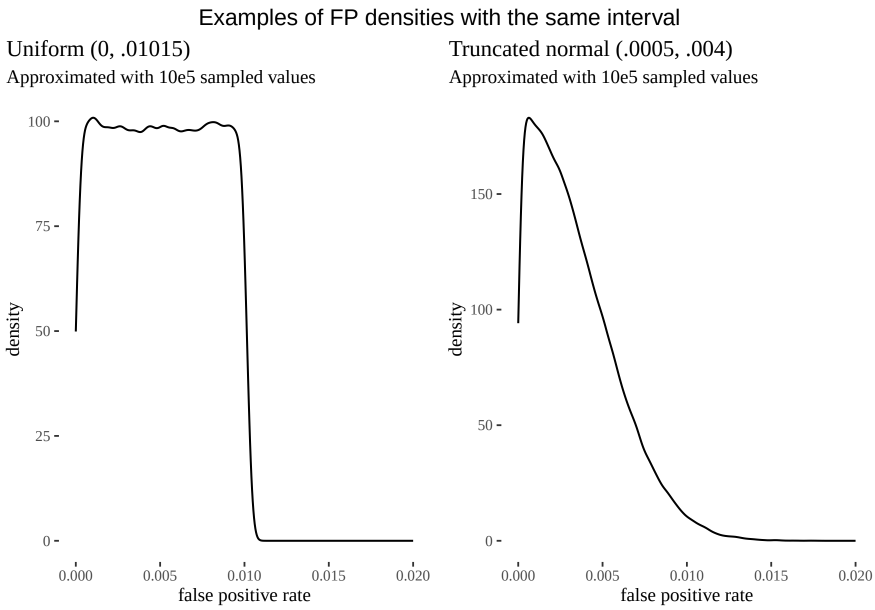

Figure 13: Two examples of assumptions about the false positive rates, both having pretty much the same 99% highest density intervals. (Top) all error rates are equally likely, (Bottom) the most likely values are closer to 0, but also some high values while unlikely are possible.

Let us use sampling from the distributions to investigate what differences arise depending on what one's second-order convictions about error rates are. First, we ask what the distributions of minimal priors sufficient for the posteriors being above .99 for these two options are (if the minimum is 1 this means that no prior results in such a posterior). The answer is illustrated in Figure 14.

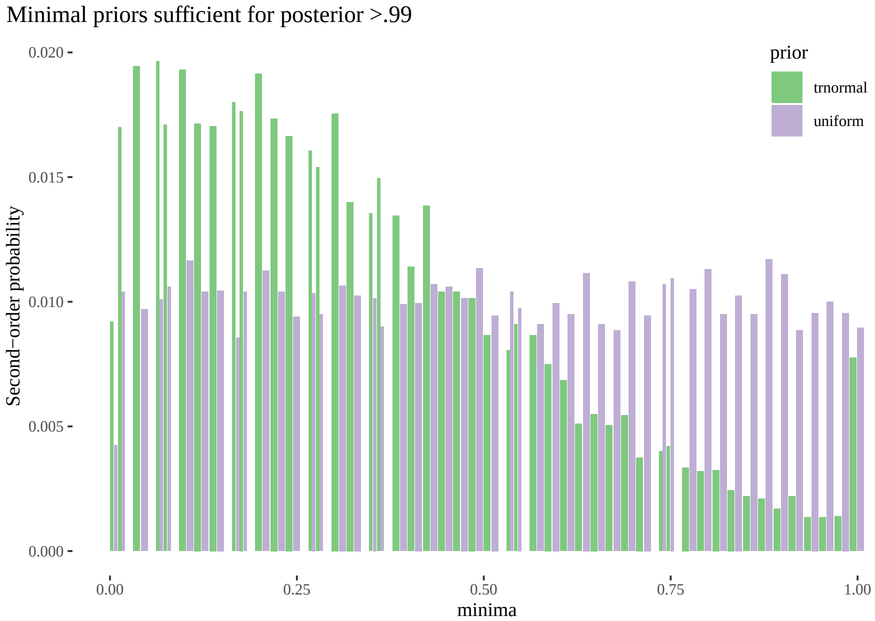

Figure 14: The distribution of minimal priors sufficient for the posterior being above .99 on the two distributions of false positive rates. Note that the truncated normal distribution has its bulk towards the left, but at the same time has higher ratio of evens in which this posterior is never reached.

Notice how the uniform distribution which seriously "thinks" that all false positive rates in the interval are equally likely leads to highly skeptical evaluation of the evidence. On both approaches the evidence is insufficient for conviction, but 95% of the minima on the truncated normal distribution are below .8, whereas the 95th quantile for the minima for the uniform distribution is rather unsurprisingly at .95. That is, the truncated normal gives a more balanced and honest picture that the charitable reading which runs only with one extreme value.

Another perspective on the impact of such differences can be taken by inspecting a large number of lines (in our case, 300) of how the posterior depends on the posterior: their density corresponds to the density of the false positive rates. We illustrate these in Figure 15.

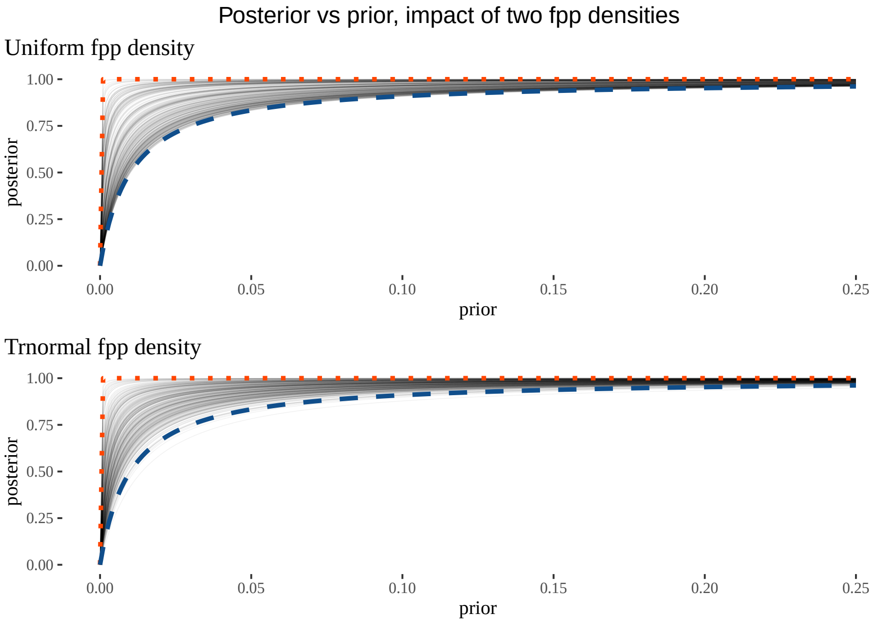

Figure 15: Impact of prior on the posterior assumign two different densitites for false positive rates. Note how both the "pristine" error-free point estimate (orange) and the charitable version (blue) are quite far from where the bulks of the distributions in fact are. Note also how the trnormal density allows for even more charitable cases, which results from it being long-tailed.

Suppose now you indeed are convinced that the distribution over possible false probability values indeed matters for the evaluation of DNA evidence. But where do we take these distributions from, you might ask. The problem is, studies on false positives are very limited and so only give a rough and foggy picture. Ideally, better experiments and studies which would allow for a better justification of the choice of a distribution, should be conducted. However, this does not mean that until then we should stick to using point estimates and interval edges. Once the functional form of the distribution, such as truncated normal or beta, which are known to be pretty standard and reliable in such contexts, are chosen, only a few numbers need to be elicited from experts to be able to construct a density. For instance, once we agree on the truncated normal form, it is enough that the expert says that the 99% interval is as the one we used, and that she believes with more than 50% confidence the false positive rates to be below .033 for the curve to be determined. There is, of course, some idealization involved, and having to rely on such elicitation is not perfect. But it is still better than asking the experts for single point estimates and relying on these.

4 Objections to the higher-order approach

You might have some concerns about the way we propose to use higher-order probabilities and densities in evidence evaluation. Some prominent scholars in field certainly do. Taroni et al. (2015) argue extensively for the experts reporting only point estimates, and their objections deserve attention.

Their point of departure is a reflection on match evidence. The prosecution hypothesis, H_p is that the suspect is the origin of the trace, and the defense hypothesis, H_d, is that another person, unrelated to the suspect is the origin of the trace. The evidence E is that a match has been reported. Typically, if H_p holds, it is assumed the laboratory will report a match, so that $\mathsf{P}(E|H_p) = 1$ and the likelihood ratio reduces to $1/\mathsf{P}(E|H_d)$. This probability can be estimated on the frequentist approach, on which probability is interpreted as the limiting or population frequency θ, and an agent in fact does estimate the true parameter in terms of a probabilistic distribution $p(\theta)$ over its possible values. For instance, in DNA evidence evaluation the allelic relative frequency f of DNA matches found in an available database is an estimate of the true frequency in a given population of interest. If the observations are realizations of independent and identically distributed Bernoulli trails given θ, the expert's uncertainty about θ can be captured as $\mathtt{beta}(\alpha + s + 1, \beta + n - s)$, where s is the number of observed successes, n the number of observations in the database (1 is added to the first shape parameter to include the match with the suspect), and α and β capture the expert's priors.

However, they insist, in front of the fact-finders, the expert is supposed to report their epistemic uncertainty, which is to be understood according to the subjective interpretation. On this interpretation, probabilities express an agent's epistemic attitude towards a proposition, their uncertainty about its truth value. Such probabilities, Taroni et al. (2015) insist, unlike frequencies, are not states of nature, but states of mind associated with individuals. For this reason, they claim, it makes no sense to talk about second-order uncertainty about subjective probabilities, as there is no underlying state of the nature to estimate. Further, they insist this also applies to likelihood ratios:

> ...there is no meaningful state of nature equivalent for the likelihood ratio in its entirety, as it is given by a ratio of two conditional probabilities. [p. 12]

In line with this perspectives, in the elicitation of probabilities Taroni et al. (2015) recommend investigating an agent's betting preferences, and that a proper elicitation of this form will lead to a single number.[18] Moreover, while they claim that it is fine to use distributions to talk about chance, they deny this possibility for personal uncertainty, under the threat of infinite regress:

> One can, in fact, have probabilities for events, or probabilities for propositions, but not probabilities of probabilities, otherwise one would have an infinite regression. [p. 8]

Accordingly, Taroni et al. (2015) insist that given a frequentist estimation of the probability of the evidence given the defense hypothesis, ...

> ...personal beliefs [...] can be computed as:

$$\begin{aligned}\mathsf{P}(E) &= \int_\theta \mathsf{P}(E|\theta)\mathsf{P}(\theta)\, d\theta \\ &= \int_\theta \theta \mathsf{P}(\theta)\, d\theta\end{aligned}$$

In particular, for the DNA match example, they advice that the personal belief is the expected value of the $\mathtt{beta}$ distribution, which reduces to $\alpha+s+1/\alpha+\beta+n+1$. They claim that it satisfactorily expresses the posterior uncertainty about θ, and that it is solely this probability that should be used in the denominator in the calculation and reporting of the likelihood ratio.

This criticism has been countered by Sjerps et al. (2015). One point they raise is that the distinction between states of nature which can be estimated and mental attitudes which are purely subjective and cannot be in any meaningful way estimated is a bit too hasty. Once the background information, the priors, the hypotheses and the evidence are specified, the higher-order uncertainty about θ follows from the probability calculus, and it is by no means arbitrary to what extent various potential values of θ are supported by the evidence.

[18] "Clearly, one can adjust the measure of belief of success in the reference gamble in such a way that one will be indifferent with respect to the truth of the event about which one needs to give one's probability. This understanding is fundamental, as it implies that probability is given by a single number. It may be hard to define, but that does not mean that probability does not exist in an individual's mind. One cannot logically have two different numbers because they would reflect different measures of belief." [p. 7]

Moreover, in reporting only a single value, Sjerps et al. (2015) argue, the expert refrains from providing the fact-finders with other information about θ which is reasonably supported by the data and can have impact on the result of evidence evaluation and incorporation. Crucially, the fact finders are interested in what Sjerps et al. (2015) call the reliability of the evidence, and it depends on the distribution (as we already illustrated in the first section of this chapter). There is a difference between the expert being certain θ is .1, their best estimate being .1 based on a thousand of observations, an their best estimate being .1 based on ten observations. Complete information about the impact of the evidential basis of an estimate on expert's uncertainty about it allows the fact-finders to reasonably go about trying to be on the safe side working on the presumption of innocence.

To be fair, Taroni et al. (2015) at some point suggest that the expert, aside of providing a point estimate, should also informally explain how the estimate was arrived at, and that it would be helpful if the recipients of this information were instructed in "the nature of probability, the importance of an understanding of it and its proper use in dealing with uncertainty" [p. 16]. But we think and our examples illustrate that depriving non-expert fact finders of clearly quantifiable information about aleatory uncertainty related to the parameter of interest, replacing this information with an informal description of what the expert did, while telling them that the nature of probability is important and hoping for the best when it comes to their proper assessment and use of the point estimates is wildly optimistic, to say the least.

Dahlman & Nordgaard (2022) criticize (Taroni et al., 2015) on more philosophical grounds. Replying to the criticism based on probabilities being legitimately assigned only to events, they use the betting interpretation that the authors use themselves and push back, saying that since a probability assessment is a betting preference, a probability assessment is in itself an event: "the formation of a betting preference by a certain person at a certain time" [p. 15]. Replying to the infinite regress objection, they point out that integrating higher order uncertainty does not a priori decrease the first order probability and might as well increase it, so that going higher-order does not in principle have to lead to skepticism. Finally, their reply to the integration recommendation is somewhat case-specific, as on their approach to the weight of evidence they deal with multiple probability measures, and so their integration would be illegitimate.

Before we get into the surprisingly philosophical intricacies of the debate between forensic scientists and legal evidence scholars, let us make a point that should convince you that the higher-order approach is preferable, no matter your philosophical convictions about the nature of probability, propositions or the relation between uncertainty and betting behavior. The point is simple. If you dump free information that you already have in the densities and run with point estimates, your predictions about the world will be less accurate in a very precise quantifiable sense.

First, let us go over a particular example. We randomly draw a true population frequency from the uniform distribution. In our particular case, we obtained 0.632. Then we randomly draw a sample size as a natural number between 10 and 20 (our points holds with larger samples, just the discrepancies get smaller). In our particular case, it is 16. Now we simulate an experiment in which we draw that number of observations from the true distribution. In our particular case this happened to lead to the observation of 8 successes. We use this number to calculate the point estimate of the parameter, which is 0.5. Now we ask about the probability mass function for all possible outcomes of an observation of the same size. On one hand, we have the true probability mass based on the true parameter. On the other, we have the probability mass function based on the point estimate which is basically binomial around the point estimate. On the third hand, we take the uncertainty involved in our estimate of the parameter seriously, and so we first take a sampling distribution of size 1e4 of possible parameter values from the posterior $\mathsf{beta}(1+\mathsf{successes}, 1+\mathsf{samplesize}-\mathsf{successes})$ distribution (we assume uniform prior for the sake of an example). Then we use this sample of parameter values to simulate observations, one simulation for each parameter value in the sample. This results in the so-called posterior predictive distribution, which instead of a point estimate, propagated our uncertainty about the parameter value into our predictions about the outcomes of possible observations. Then we take simulated frequencies as our estimates of probabilities. This distribution is more honest about uncertainty and wider than the one obtained using the point estimate. All these are illustrated in Figure 16.

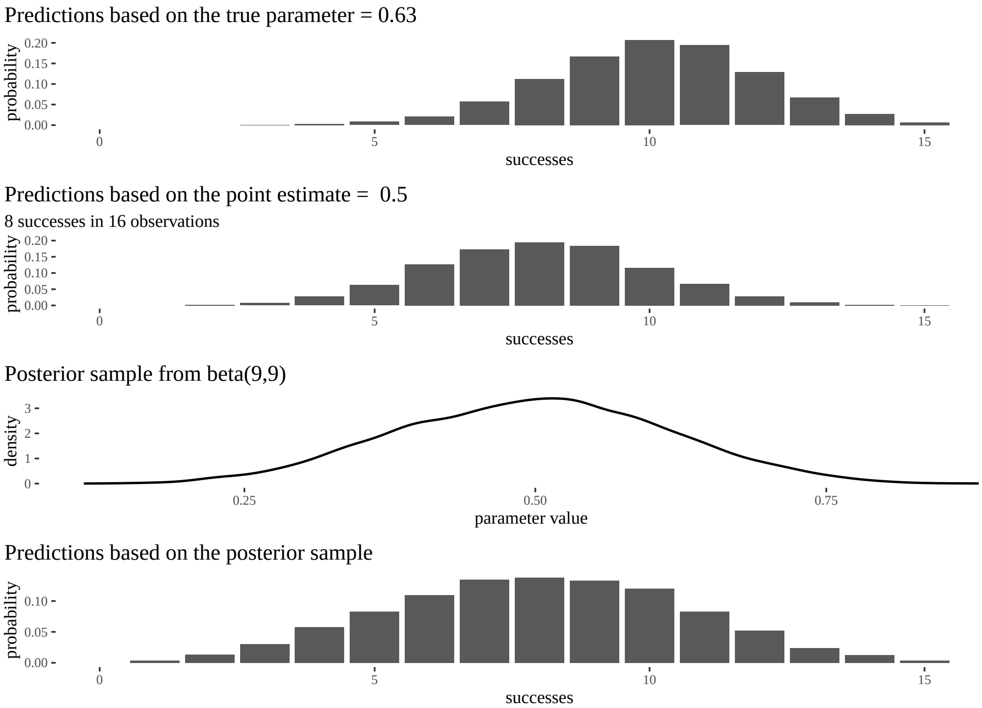

Figure 16: Real probability mass, probability mass calculated using a point estimate, sampling distribution from the posterior, and the posterior predictive distribution based on this sampling distribution.

Notably, the PMF based on a point estimate is further off from the real PMF than the posterior predictive distribution. For instance, if we ask about the probability of the outcome being at least 9 successes, the true answer is 0.7984, the point estimate PMF tells us it is 0.4056, while the posterior predictive distribution gives a somewhat better guess at 0.4277. Interestingly, a similar thing happens when we ask about the probability of the outcome being at most 9 successes. The true answer is 0.3681, the point-estimate-based answer is 0.778, while the posterior predictive distribution yields 0.7051. More generally, we can use Kullback-Leibler divergence to measure how far the point-estimate PMF and the posterior predictive PMF are from the true PMF. In our particular case, the former distance is 0.7905638 and the latter is 0.5681121. That is, the posterior predictive distribution is information-theoretically closer to the true distribution.

Now let us see how this point holds generally. Let's repeat the whole simulation 1000 times, each time with a new true parameter, a new sample size, and a new sample. Every time we construct the PMFs using the methods we described, and measure their KLD distances. Here is the empirical distribution of the results of such a simulation (Figure 17)—we are looking at the differences in KLD divergencies, positive differences mean the point-estimate based distribution was further from the true PMF than the posterior predictive PMF. Notably, the mean difference is 0.865, the median difference is 0.044, and the distribution is asymmetrical, as there are multiple cases of huge differences favoring posterior predictive distributions. That is, accuracy-wise, point-estimate-based PMFs are systematically worse than posterior predictive PMFs.

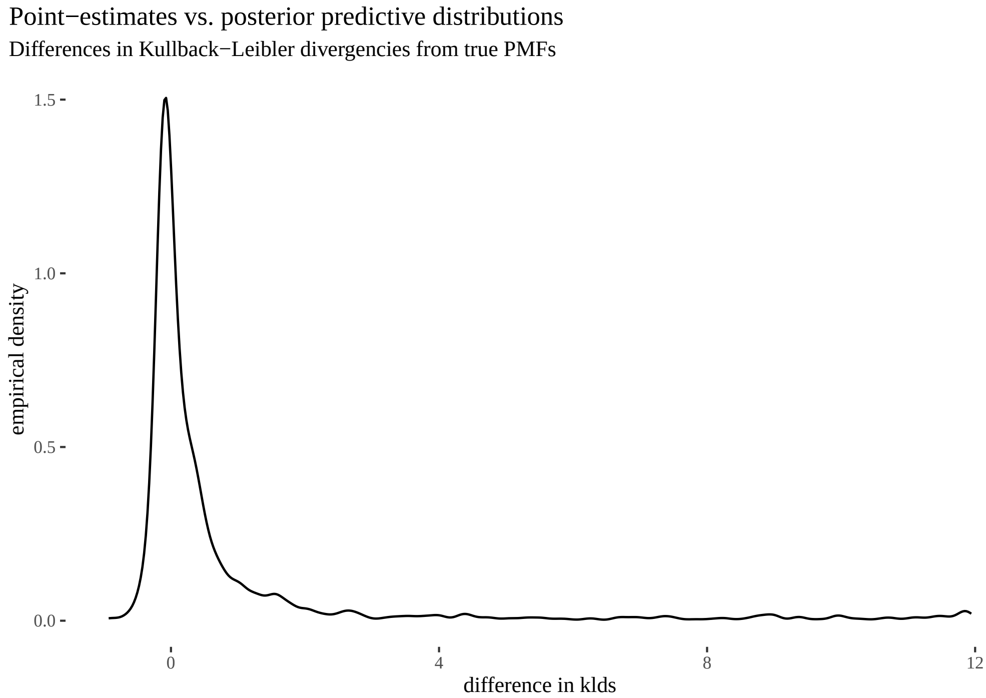

Figure 17: Differences in Kullback-Leibler divergencies from the true distributions, comparing the distributions obtained using point estimates and posterior predictive distributions. Positive values indicate the point-estimate-based PMF was further from the true distribution that its competitor.

If accuracy considerations do not convince you, we can also get more philosophical. As the betting behavior interpretation is actually not as obvious or uncontroversial as Taroni et al. (2015) suggest,[19] we prefer to respond without assuming it, just working with the assumption that an agent's probabilities or densities are to represent or capture their uncertainty.

First, Taroni et al. (2015) seem to argue that since first-order probabilities capture your uncertainty about a proposition of interest, second-order probabilities are supposed to capture your uncertainty about how uncertain you are, and that "estimating" your first-order uncertainties would be a bit silly—after all, they seem to think, you can simply figure out your fair odds in a related bet on the proposition in question, and those would give you your unique first-order uncertainty without any uncertainty about it.

There are more controversial claims that philosophers have argued for than the infallibility of introspection. Those involved with epistemic logic often argue that it is not even the case that if an agent knows (or doesn't know) p, they also automatically know that they (don't) know k. Analogously, it would not be a completely implausible philosophical position to say that we sometimes are uncertain about what we think the fair bets are or what our first-order uncertainties are. This is one possible path of answering this.

But we think there is a better way. Let us think again about our expert, who gathers information about the allelic relative frequency f of DNA matches in an available database, and starts with some defensible beta prior with parameters α, β. Say they observe s matches and the database size is n. They reach a $\mathsf{beta}(\alpha+s+1, \beta+n-s)$ distribution over the possible RMP values. So far, nothing controversial happens—they are estimating the relevant population frequency. Assuming the conditions are pristine (the expert has no modeling uncertainty, does not have consider la ab errors and so on),[20] they want to use this distribution to shape their subjective uncertainty. But uncertainty about what? One obvious proposition is *a match is observed if another person, unrelated to the suspect is the origin of the trace*. And indeed, insofar as only this proposition is being considered, it is yet not clear what second-order uncertainties would be uncertainties about. But the expert also considers a continuum of propositions, each of the form *the true population frequency is* θ for each $\theta \in [0,1]$. A density over θ can be simply

[19] See philosophical textbooks on formal epistemology such as (D. Bradley, 2015) or (Titelbaum, 2020).

[20] Otherwise they need to think harder and dampen their convictions.

seen as capturing the comparative plausibility that the expert assigns to such propositions in light of the evidence, and normalization allows them to calculate their subjective probabilities for θ belonging to various sub-intervals of $[0,1]$. So if you were worried that there were no propositions that the expert could be "second-order" uncertain about, the good news is, there are plenty. In particular, if θ is supposed to be a population frequency, gauging which density captures the extent to which the evidence justifies various estimates of that frequency is pretty much the same as gauging comparative plausibility of the corresponding propositions about the population frequencies. Now, you might complain that this should no longer be called "estimation", but you might also think that the connection is strong enough to justify this terminology. This is a verbal discussion that we will not get into, language users' linguistic behavior will decide this for us anyway.

More generally, evidence justifies various first-order probability assignments to varying extent. If we have no evidence about the bias of a coin, each first-order uncertainty about it is equally (un)-justified (if you like to think in terms of bets: the evidence gives us no reasons to prefer any particular odds as fair). If we know the coin is fair, the evidence clearly selects one preferred value, .5 (and, again, if you like the betting metaphor, one preferred betting odds). But often the evidence is stronger than the former and weaker than the latter case—one clear example of this phenomenon is if we consider propositions about population frequencies in light of the results of observations—in such circumstances, the evidence justifies different values of first-order uncertainty to varying degree, and densities simply capture the extent to which different first-order uncertainties are supported by the evidence. So what is it that you are estimating? The ideal evidential support you would get if you had all the evidence. Why is the estimate off? because your evidence is limited and the best you can do is gauge your uncertainty going where this limited evidence leads you.

What about the objection that these are not "states of nature" and so cannot be estimated? We are not sure about "nature" and why the requirement of being part of nature is important for estimation. We have seen cases of mathematicians using approximate methods to estimate answers to fairly abstract questions not obviously related to "states of nature", whatever these are, and so we think estimation makes sense whenever there are some objective answers that get closer or further from. But if there is some objectivity to what the ideal evidence would support, or to the extent to which the actual evidence supports various competing hypotheses, we can be more or less wrong about such things, and so it is not implausible to say that there is a clear sense in which we can estimate them.

What about the threat of infinite regress? Surely, Taroni et al. (2015) would agree that one can be uncertain about a statistical model. But this can be the case even if this model spits out a point estimate rather than a density. This would suggest that if you think the possibility of putting uncertainty on top of propositions about possible values of a first-order parameter leaves us in an epistemically hopeless situation, you might have hard time explaining why your point estimation is in a better situation. After all, if asking further questions about probabilities up the hierarchy is always justified, we can keep asking about the probability of a point-estimate-spitting model being adequate, the probability of that probability (and the way we have reached it) being adequate and so on. Alternatively, we could concede that while our situation is often epistemically difficult, it is not hopeless. Of course, the mathematical and statistical models we use are exactly those: models, which can be more or less epistemically helpful. And when we decide which models to use, we always face a trade-off between various factors. In particular, second-order estimation is more complex than running with point estimates. But by now we hope to have convinced the reader this complexity is worth the effort. What about more complex models going even higher? Well, if a case is made that there is a workable approach that does that, which is worth the additional complexity, we are all for it! In fact, some of the model selection methods can be thought of this way. But to say that just because in principle more complex models are always possible, we are facing some sort of epistemic infinite regress is too hasty.

Finally, what about the idea that there is no "meaningful state of nature equivalent for the likelihood ratio in its entirety, as it is given by a ratio of two conditional probabilities?" If you think it is meaningful to estimate two conditional probabilities (that is, frequencies in the population), or to compare the relative plausibility of various propositions about them in terms of density, it is equally meaningful to estimate any function of the numbers involved. Otherwise it would also be meaningless to try to estimate the BMI of an average 21 years old male student in the USA just because BMI is a ratio of other quantities. There are various good reasons not to care about BMI too much, but being a ratio of other numbers is not one of them.

5 Weight of evidence

Comment: It is conceptually important to separate the discussion about precise, imprecise, and higher order probabilism (previous sections) from weight (this section). Weight of evidence is one way in which higher order probabilism can be put to use. It can be confusing to run the discussion of higher order probabilism together with weight of evidence. Higher order probabilism can still make perfect sense even if no theory of weight cen be worked out.

5.1 Motivating examples

This section should start with illustrative examples of the weight/balance distinction and why “balance alone” isn’t enough to model the evidential uncertainty relative to a hypothesis of interest. These examples should be chosen carefully. We can use legal and non-legal examples. The driving intuition is given by Keynes with the weight/balance distinction. Some of the examples we saw earlier in talking about imprecise probabilism can be mentioned here again, such as (a1) and (a2), and perhaps also (b) and (c).

Upshot is that uncertainty cannot be captured by balance of evidence alone. There is a further dimension to uncertainty. So we need a theory that can accommodate this further level of uncertainty. This theory is essentially the higher order probabilism introduced before.

5.2 Desiderata

Here we can discuss monotonicity, completeness, strong increase, etc (see current **section 1**). We can list the intuitive properties (based on the example we presented in both philosophy and law) that any theory of weight (and perhaps also of completeness/resilience, but on these notions, see later) should be able to capture. We should try to keep these requirements as simple as possible and leave complications to footnotes.

5.3 Formal characterization of weight

Our account of weight of evidence will be information-theoretic. To develop this account, we will first sketch an account of information, based on Shannon’s theory of information.

5.3.1 Entropy of a distribution

Let X be a random variable and P a probability distribution over its values. Shannon’s measure of information, $H(X)$, reads:

$$H(X) = -\sum \mathsf{P}(x_i)\log_2 \mathsf{P}(x_i)$$

Suppose the random variable X can take two values—outcome 1 and outcome 0—whose probabilities are .5 each. Then, $H(X) = 1$. On the other hand, if the outcomes are not equally likely, $H(X)$ will return a value lower than one. At the extreme, if one outcome is close to certainty (and thus the other is extremely unlikely), $H(X)$ would return a value close to zero.

To make sense of this, $H(X)$ can be thought as a measure of the *entropy* (i.e. the lack of information) contained in the whole distribution associated with random variable X. So, in this sense, $H(X)$ is greatest when the outcomes are regarded as equally probable (maximum entropy), and smallest when one outcome is regarded as certain (lowest entropy). On other hand hand, $H(X)$ is also a measure of information, specifically, the *expected amount of information* one would receive upon learning the actual value of X. After all, the less informative a distribution, the more you expect to learn when you find out the value of X, the higher the entropy.[21]

[21] Since working with continuous distributions is not straightforward, we will be using *grid approximations* of continuous distributions: we will split X into a 1000 bins and use the normalized densities for their centers to obtain their corresponding probabilities. As long as we do not change our level of precision (which would inevitably lead to changes in entropy) in our comparisons, this is not a problem. An additional advantage is that now we do not have to deal with the intricacies of explicit analytic calculations for continuous variables.

5.3.2 Informativeness of a distribution

Since $H(X)$ can be thought as the entropy of a distribution, we will switch to the notation $H(P)$. This notation emphasizes the distribution P rather then the random variable X. The entropy of a distribution is to be contrasted with its informativeness, which we will denote by $W(P)$, the weight of the distribution. To measure the weight (or informativeness) of a distribution, it is best to compare it to the least informative distribution, the uniform distribution. The idea is that the more informative a distribution, as compared to the uniform distribution, the more weight it has, on scale 0 to 1: if the drop from uncertainty is complete, the entropy drops to zero, and we would like the weight to be 1, if the drop is null we would like to be zero, and if the drop is half, we would like to be .5 (and so on for other proportions). This can be achieved by the following definition:

$$\mathsf{w}(\mathsf{P_i}) = 1 - \left(\frac{H(\mathsf{P})}{H(\mathsf{uniform})}\right)$$

where P is probability distribution of interest and baseline uniform distribution.[22]

The behavior of the proposed measure of weight can be illustrated using distributions of various of shapes, displayed in Figure 18). A bimodal normal distribution "glued" from two normal distributions carries less weight than a unimodal normal distribution with the same standard deviation centered around the mean of the two modes. If multiple points have non-zero probability, the weight depends on how uneven the distribution is. If the distribution falls entirely on a single point, the weight is maximal (=1), as expected.

[22] Since we are using grip approximation, P is the discrete probability distribution for a given number of bins n, and uniform is the discrete uniform distribution for the same number of bins. In some contexts it might make sense to measure improvement with respect to a non-uniform prior. In such cases, $H(\mathsf{uniform})$ is to be replaced by $H(\mathsf{prior})$. Note that the entropy of a uniform distribution is pretty straightforward, so we can simplify:

$$\begin{aligned} H(\mathsf{uniform}) &= \sum_{i=1}^{n} 1/n \log_2 \frac{1}{1/n} \\ &= \log_2(n) \\ \mathsf{w}(\mathsf{P_i}) &= 1 - \left(\frac{H(\mathsf{P})}{\log_2(n)}\right) \end{aligned}$$

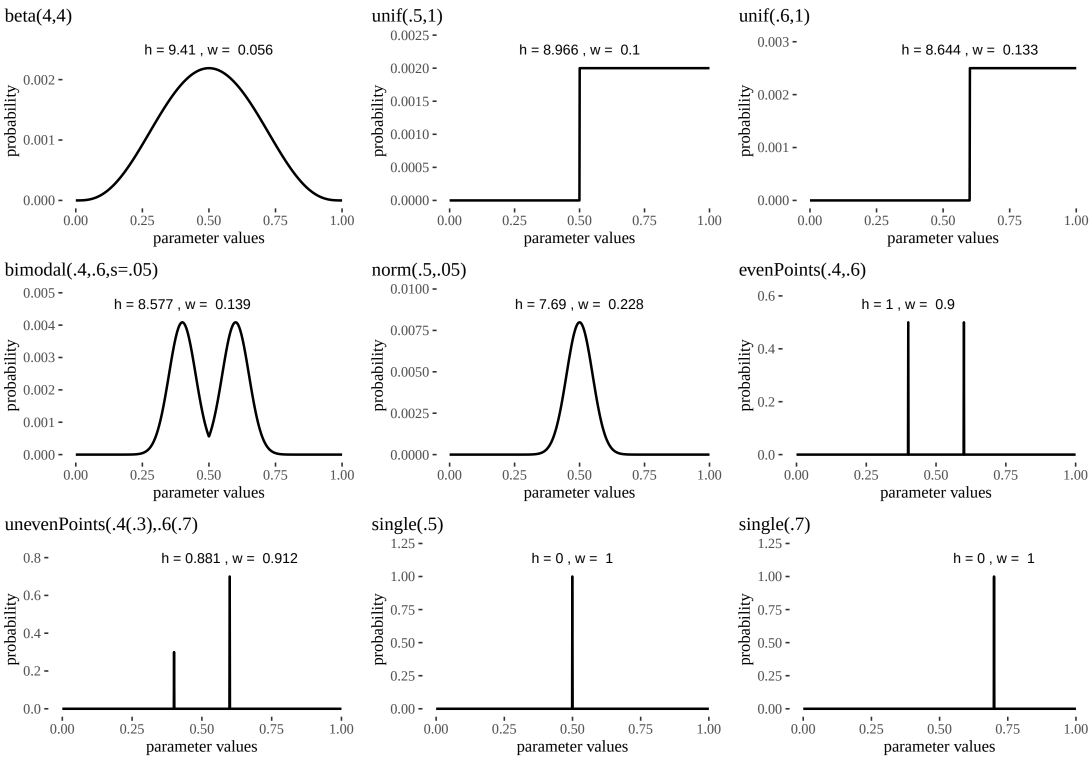

Figure 18: Examples of various distributions with their entropies and weights, ordered by weights. (1) beta(4,4), (2) uniform starting from .5 to 1, (3), uniform strating from .6 to 1, (4) two normal distributions centered around .4 and .6 with standard deviation .05, glued at .5. (5) normal centered around .5 with the same standard deviation, (6) one that assigns .5 to each of .4 and .6, (7) One that assigns .3 to .4 and .7 to .6., (8) one that assigns all weight to .5, and (9) one that assigns all weight to .7.

5.3.3 Weight of evidence

So far we have discussed the weight of a distribution, meant to measure how informed an agent is about an issue. If the agent starts with a uniform prior, this is a good enough approximation of how informed the evidence made them. But in general, how much more information is obtained is context-dependent. We want a prior-relative notion of weight, following the intuition that weight considerations should guide our information gathering also in making us stop collecting further evidence in light of what we already know. But for weight of evidence to have this feature, it has to depend on what we already know.

So here is a general recipe. In a given context, consider your distribution for the target hypothesis H given what you already know. Then update on the evidence. This might increase the weight for H, if the evidence confirms your conviction, or decrease it, if it goes against what the previous evidence tells you. Take the difference between the prior weight and the posterior weight (Δw) as your measure of the weight of evidence in that context. If you prefer to think that weight of evidence should be always positive, you might prefer the absolute value thereof. We, however, prefer to keep track of whether the evidence makes you more or less confused. The calculation goes along the following schema:

1. Start with a prior distribution over the parameter space of interest, and with distributions expressing the agent's uncertainty about other probabilities involved in the calculation of the posterior.
2. Sample from these distributions.
3. For each sample, treat it as a selection of precise probabilities, apply Bayes' theorem to calculate the posterior.
4. The set of the results is the sampling distribution expressing your posterior uncertainty.

Higher order probabilism is then put to use to deliver a theory of weight. What is now in **section 11** ("Weight of a distribution") and **sections 13** and **14** (" Weight of evidence" and "Weights in Bayesian Networks") forms the bulk of the theory.

We should also demonstrate that the proposed theory of weight does meet the intuitive desiderata and can handle the motivating examples. To better appreciset the novelty of the proposal, It might be interesting to raise the following questions:

q1 what does a theory of weight based on precise probabilism look like? (maybe it consists of something like Skyrms' resilience or Kaye's completeness, the problem being that these are not measures of weight, but of something else, more on these later)

q2 what does a theory of weight based on imprecise probabilism look like? (is Joyce's theory essentially an attempt to use imprecise probabilism to construct a theory of weight?)

Well, it's a bit funny as Joyce's weight uses precise chance hypotheses instead of IP, so hard to say

q3 what does a theory of weight based on higher order probabilism look like?

Here we are defending a theory fo weight based on higher order probabilism, but it is interesting to contrast it with a theory of weight based on the other version of legal probabilism. Here we can also show why Joyce's theory of weight does not work (either in the main text or a footnote).

Comment: The current exposition in chapter 11, 13 and 14, however, is complicated—perhaps overly so. The move from "weight of a distribution" to "weight of evidence" is not intuitive and can confuse the reader. Is there a simpler story to be told here? I think so. See below.

Brilliant, I think I can start talking about conditional probabilities to begin with

Suggestion: There seems to be a nice symmetry. Start with precise probabilism. We can use sharp probability theory to offer a theory of the value of the evidence (i.e. likelihood ratio). Actually, I think that the likelihood ratio model the idea of balance of the evidence. What Keynes distinction weight/balance shows is that likelihood ratio are not, by themselves, enough to model the value of the evidence. The straightforward move here seems to just have **higher order likelihood ratios**. Wouldn't higher order likelihood ratio be essentially your formal model of the weight of the evidence? Your measure of weight tracks the difference between (the weight of the) prior distribution (and the weight of the) posterior distribution. But higher order likelihood ratios essentially do the same thing, just like precise likelihood ratios track the difference between prior and posterior. Is this right?

Yup, more or less

Comment: If weight if measured by higher order likelihood ratios, then this can be seen as a generalization of thoughts that many other had – say that the absolute value of the likelihood ratio is a measure of weight (Nance, Glenn Shafer) or that likelihood ratio must be a measure of weight (Good; see current **section 4**). So I think using ""higher order likelihood ratio" could be a more appealing way to sell the idea of weight of evidence since most people are already familiar with likelihood ratios.

5.4 Limits of our contribution

Work by Nance of Dahlman suggests that "weight" should play a role in the standard of proof. We do not take a position on that. Weight could be regulated by legal rules at the level of rules of decision, rule of evidence, admissibility, sanctions at the appellate level. All that matters to us is that, in general, legal decision-making is sensitive to these further levels of uncertainty (quantity, completeness, resilience), but whether this should be codified at the level of the standard of proof or somewhere else, we are not going to take a stance on that.

5.5 Objection

Ronald Allen or Bart Verheij might object as follows. Precise probabilism is bad because we do not always have the numbers we need to plug into the Bayesian network. Imprecise probabilism partly addresses this problem by allowing for a range instead of precise numbers. How does higher order probabilism help address the practical objection that we often we do not have the numbers we need to plug into the Bayesian network?

6 Completeness (and resilience?)

Next the chapter turns to notions related to the weight of evidence, such as completeness (and perhaps resilience as well). See current **sections 5** and **6**.

6.1 Motivating example

Give an example using completeness of evidence (pick one or more court cases). The court case we can use is Porter v. City of San Francisco (see file with Marcello's notes).[23] The jury is given an instruction that a call recording is missing, but no instruction whether the call should be assumed to be favorable or not.

What is the jury supposed to do with this information? If the call could contain information that is favorable or not, shouldn't the jury simply ignore the fact that the call recording is missing (Hamer's claim)? Modelling with Bayesian network might turn out useful. Cite also David Kaye on the issue of completeness. His claim is that when evidence is known to be missing, then this information should simply be added as part of the evidence, which is precisely what the court in Porter does. But again, once we add the fact that the evidence is missing what is the evidentiary significance of that? What is te jury supposed to do with that? Does Pr(H) go up, down or stays the same? Kaye does not say…

6.2 Bayesian network model

Comment I am thinking that incompleteness is modeled by adding an evidence node to a Bayesian network but without setting a precise value for that node, and then see if the updated network yields a different probability than the previous network without the missing evidence node. The missing evidence node could be added in different places and this might changes things. In the Porter case the missing evidence seems to affect the credibility of the other evidence in the case, we would have a network like this: $H \rightarrow E \leftarrow C$, where C is the missing evidence node and E is the available evidence nose. My hunch is that (see also our paper on reverse Bayesianiam and unanticipated possibilities) the addition of this credibility node will affect the probability of the hypothesis (thus proving Hamer wrong).

I think this will depend on how the probability of obtaining new evidence given guilt and given innocence are, I will keep thinking about this, we'll move to this once the earlier bits are done

6.3 Expected weight model

Question: If what I say above in the comment is correct, then a question arises, do we need higher order probabilism to model completeness?

Possible answer: We can use expected weight (see current **section 14**). If the expected weight of an additional item of evidence is null, that would mean that its addition (not matter the value the added evidence would take) cannot change the probability of the hypothesis. If the expected weight is different from zero (pace Hamer who thinks the expected weight is always null), then the evidence can change the probability of the hypothesis.

LR ratio and weight

7 Weight and accuracy

This section addresses the question, why care about weight?

Conclusion

Bradley, D. (2015). *A critical introduction to formal epistemology*. Bloomsbury Publishing.

Bradley, S. (2012). *Scientific uncertainty and DecisionMaking* (PhD thesis). London School of Economics; Political Science.

[23]This is a wrongful death case in which victim was committed to a hospital facility, but escaped and then died under unclear circumstances. So the nurses and other hospital workers—actually, the city of San Francisco—are accused of contributing to this person's death. Need to check exact accusation—this is not a criminal case. A phone call was made to social services shortly after the person disappeared, but its content was erased from hospital records. Court agrees that content of phone call would be helpful to understand what happened and to assess the credibility of hospital's workers ("The Okupnik call is the only contemporaneous record of what information was reported to the SFSD about Nuriddin's disappearance, and could contain facts not otherwise known about her disappearance and CCSF's response. Additionally, the call is relevant to a jury's assessment of Okupnik's credibility"). The court thought that the hospital should have kept records of that call. But court did not think the hospital acted in bad faith or intentionally, so it did NOT issue an "adverse inference instruction" (=the missing evidence was favorable to the party that should have preserved it, but failed to do it).

Bradley, S. (2019). Imprecise Probabilities. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019). https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/; Metaphysics Research Lab, Stanford University.

Campbell-Moore, C. (2020). *Accuracy and imprecise probabilities*.

Carr, J. R. (2020). Imprecise evidence without imprecise credences. *Philosophical Studies*, *177*(9), 2735–2758. https://doi.org/10.1007/s11098-019-01336-7

Dahlman, C., & Nordgaard, A. (2022). *Information economics in the criminal standard of proof*.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability*. Oxford: Oxford University Press.

Elkin, L. (2017). *Imprecise probability in epistemology* (PhD thesis). Ludwig-Maximilians-Universität; Ludwig-Maximilians-Universität München.

Elkin, L., & Wheeler, G. (2018). Resolving peer disagreements through imprecise probabilities. *Noûs*, *52*(2), 260–278. https://doi.org/10.1111/nous.12143

Fraassen, B. C. V. (2006). Vague expectation value loss. *Philosophical Studies*, *127*(3), 483–491. https://doi.org/10.1007/s11098-004-7821-2

Gärdenfors, P., & Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese*, *53*(3), 361–386. https://doi.org/10.1007/bf00486156

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*(1), 153–178.

Kaplan, J. (1968). Decision theory and the fact-finding process. *Stanford Law Review*, *20*(6), 1065–1092.

Keynes, J. M. (1921). *A treatise on probability, 1921*. London: Macmillan.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy*, *71*(13), 391. https://doi.org/10.2307/2025161

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT Press.

Mayo-Wilson, C., & Wheeler, G. (2016). Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, *92*(1), 55–78. https://doi.org/10.1111/phpr.12256

Rinard, S. (2013). Against radical credal imprecision. *Thought: A Journal of Philosophy*, *2*(1), 157–165. https://doi.org/10.1002/tht3.84

Schoenfield, M. (2017). The accuracy and rationality of imprecise credences. *Noûs*, *51*(4), 667–685. https://doi.org/10.1111/nous.12105

Seidenfeld, T., Schervish, M., & Kadane, J. (2012). Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, *53*, 1248–1261. https://doi.org/10.1016/j.ijar.2012.06.018

Sjerps, M. J., Alberink, I., Bolck, A., Stoel, R. D., Vergeer, P., & Zanten, J. H. van. (2015). Uncertainty and LR: to integrate or not to integrate, that's the question. *Law, Probability and Risk*, *15*(1), 23–29. https://doi.org/10.1093/lpr/mgv005

Stewart, R. T., & Quintana, I. O. (2018). Learning and pooling, pooling and learning. *Erkenntnis*, *83*(3), 1–21. https://doi.org/10.1007/s10670-017-9894-2

Sturgeon, S. (2008). Reason and the grain of belief. *Noûs*, *42*(1), 139–165. Retrieved from http://www.jstor.org/stable/25177157

Taroni, F., Bozza, S., Biedermann, A., & Aitken, C. (2015). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, *15*(1), 1–16. https://doi.org/10.1093/lpr/mgv008

Titelbaum, M. G. (2020). *Fundamentals of bayesian epistemology*.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman; Hall London.