Second-order probability, weight of evidence & accuracy

Rafal Urbaniak and Marcello Di Bello

Contents

1 Introduction

A defendant in a criminal case may face multiple items of incriminating evidence whose strength can at least sometimes be assessed using probabilities. For example, consider a murder case in which the police recover trace evidence that matches the defendant. Hair found at the crime scene matches the defendant's hair (call this evidence hair). In addition, the defendant owns a dog whose fur matches the dog fur found in a carpet wrapped around one of the bodies (call this evidence dog).[1] The two matches suggest that the defendant (and the defendant's dog) must be the source of the crime traces (call this hypothesis, source). But how strong is this evidence, really? What are the fact-finders to make of it?

The standard story among legal probabilists goes something like this. To evaluate the strength of the two items of match evidence, we must find the value of the likelihood ratio:

$$\frac{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\mathsf{source})}{\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair}|\neg\mathsf{source})}$$

For simplicity, the numerator can be equated to one. To fill in the denominator, an expert provides the relevant random match probabilities. Suppose the expert testifies that the probability of a random person's hair matching the reference sample is about 0.0253, and the probability of a random dog's hair matching the reference sample happens to be about the same, 0.0256.[2] Presumably, the two matches are independent lines of evidence. In other words, their random match probabilities must be independent of each other conditional on the source hypothesis. Then, to evaluate the overall 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 is a very low number. Two such random matches would be quite a coincidence. Following our advice from Chapter 5, the expert facilitates your understanding of how this low number should be interpreted by showing you how the items of match evidence change the probability of the source hypothesis given a range of possible priors (Figure 1). The posterior of .99 is reached as soon as the prior is higher than 0.061.[3] While perhaps not sufficient for outright belief in the source hypothesis, the evidence seems extremely strong: a minor additional piece of evidence could make the case against the defendant overwhelming.

[1] The hair evidence and the dog fur evidence are stylized after two items of evidence in the notorious 1981 Wayne Williams case (Deadman, 1984b, 1984a).

[2] Probabilities have been slightly but not unrealistically modified to be closer to each other in order to make a conceptual point. The original probabilities were 1/100 for the dog fur, and 29/1148 for Wayne Williams' hair. We modified the actual reported probabilities slightly to emphasize the point that we will elaborate further on: the same first-order probabilities, even when they sound precise, may be affected by second-order uncertainty to different extents.

[3] These calculations assume that the probability of a match if the suspect and the suspect's dog are the sources is one.

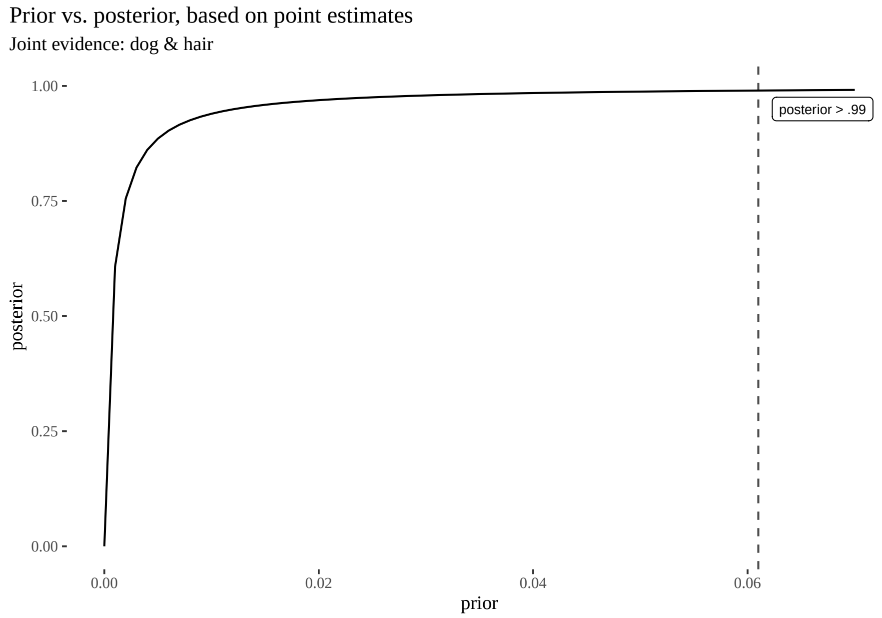

Figure 1: Impact of dog fur and human hair evidence on the prior, point estimates.

Unfortunately, this analysis leaves out something crucial. You reflect on what you have been told and ask the expert: how can you know the random match probabilities with such precision? Shouldn't we also be mindful of the uncertainty that may affect these numbers? The expert agrees, and tells you that in fact the random match probability for the hair evidence is based on 29 matches found in a database of size 1148, while the random match probability for the dog evidence is based on finding two matches in a reference database of size 78.

The expert's answer makes apparent that the precise random match probabilities do not tell the whole story. What to do, then? You ask the expert for guidance: what are reasonable ranges of the random match probabilities? What are the worst-case and best-case scenarios? The expert responds with 99% credible intervals—specifically, starting with uniform priors, the ranges of the random match probabilities are (.015,.037) for hair evidence and (.002, .103) for fur evidence.[4] With this information, you redo your calculations using the upper bounds of the two intervals: $.037$ and $.103$. The rationale for choosing the upper bounds is that these numbers result in random match probabilities that are most favorable to the defendant. Your new calculation yields the following:

$$\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source}) = .037 \times .103 = .003811.$$

This number is around 5.88 times greater than the original estimate. Now the prior probability of the source hypothesis needs to be higher than 0.274 for the posterior probability to be above .99 (Figure 2). So you are no longer convinced that the two items of match evidence are strongly incriminating.

[4]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 XXX.

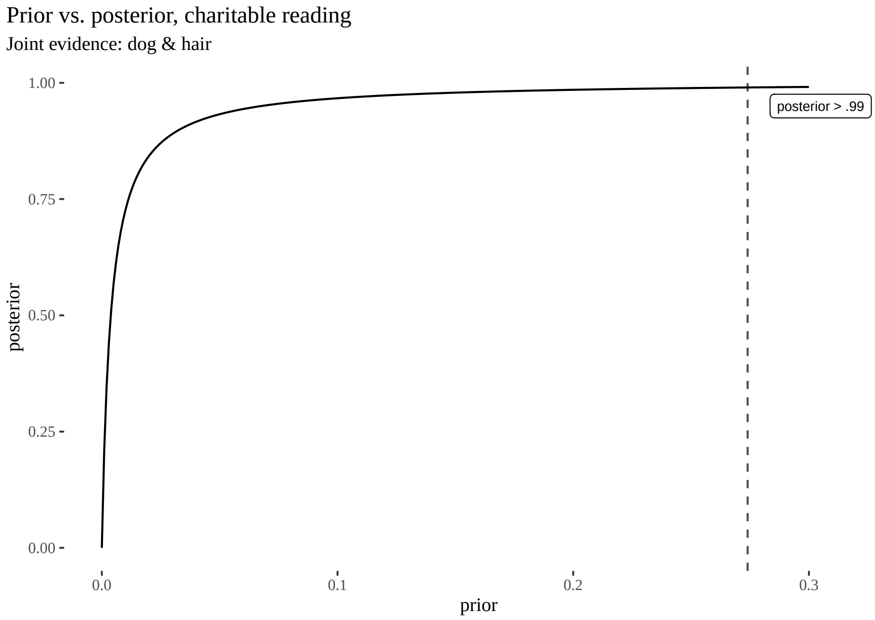

Figure 2: Impact of dog fur and human hair evidence on the prior, charitable reading.

This outcome is puzzling. Are the two items of match evidence strongly incriminating evidence (as you initially thought) or somewhat weaker (as the new calculation suggests)? For one thing, using precise random match probabilities might be too unfavorable toward the defendant. On the other hand, your new assessment of the evidence based on the upper bounds might be too *favorable* toward him. Is there a middle way that avoids overestimating and underestimating the value of the evidence?

To see what this middle path looks like, we should reconsider the calculations you just did. Here you made an important blunder: you assumed that because the worst-case probability for one event is x and the worst-case probability for another independent event is y, the worst-case probability for their conjunction is xy. But this conclusion does not follow so long as the margin of error (or credible interval) is kept fixed.[5] The interval presentation instead of doing us good led us into error.

Even if we are right about the interval that we obtained for the probability of a single event, it is not straightforward to calculate a new interval once a second event is included. We need additional information: the distributions that were used to calculate the intervals for the probabilities of the individual events.[6] So, in the case of the two items of match evidence, 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% credible interval for both items of evidence. As it turns out, given the reported sample sizes, the credible interval for the probability $\mathsf{P}(\mathsf{dog} \wedge \mathsf{hair} | \neg \mathsf{source})$ is $(0.000023, 0.002760)$.[7] Using the upper bound of this interval would then require the prior probability of the source hypothesis to be above .215 for the posterior to be above .99. On this interpretation, the two items of match evidence are still not quite as strong as you initially thought, but stronger than what your second calculation indicated.

Still, the interval approach—even the corrected version just outlined—suffers from a more general problem. Working with intervals might be useful if the underlying distributions are fairly symmetrical. But in our case, they might not be. For instance, Figure 3 depicts beta densities for dog fur and human hair, together with sampling-approximated density for the joint evidence. The distribution for the joint evidence is not symmetric. If you were only informed about the edges of the interval, you would be oblivious to the fact that the most likely value (and the bulk of the distribution, really) does not simply

[5]The intuitive reason is 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.

[6]Also, in principle, we need further information about how these items of evidence are related if we cannot take them to be independent, which we here assume random matches are.

[7]The 99% credible interval (or a 99% margin of error) is not the 99% confidence interval known from classical statistics. There are various reasons not to use these, already discussed in Chapter 3. Another sense, in which we mean it here, is the range to which the true value belongs with posterior probability of 99% given the evidence. Normally we mean highest posterior density intervals, that is the narrowest intervals with this property.

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—the bulk of the density might very well be closer to one of the edges. Therefore, relying on the edges only can lead one to either over-estimate or under-estimate the probabilities at play. This also means that—following our advice to illustrate the impact of evidence on the prior— a better representation of the uncertainty involving the dependence of the posterior on the prior involves multiple possible sampled lines whose density mirrors the density around the probability of the evidence (Figure 4).

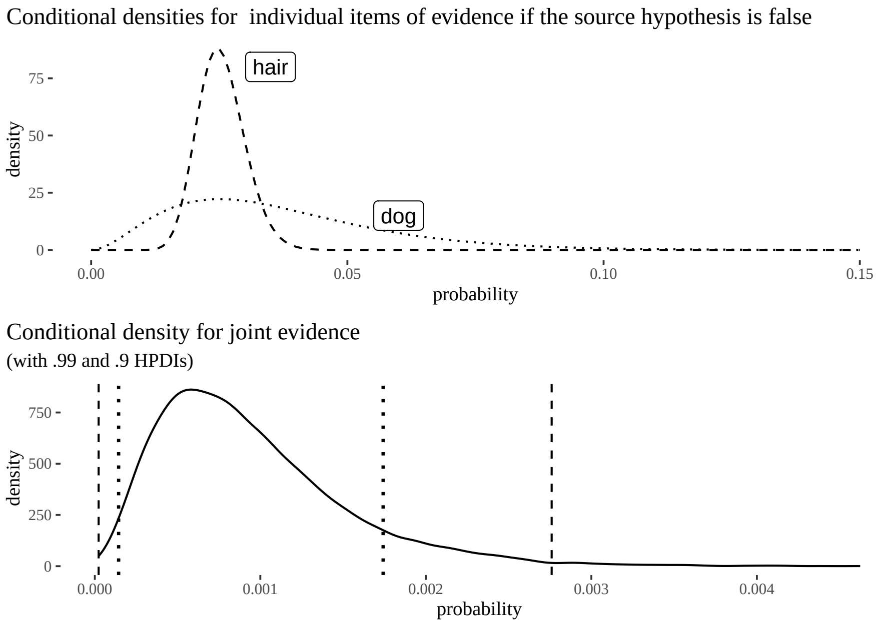

Figure 3: Beta densities for individual items of evidence and the resulting joint density with .99 and .9 highest posterior density intervals, assuming the sample sizes as discussed and independence, with uniform priors.

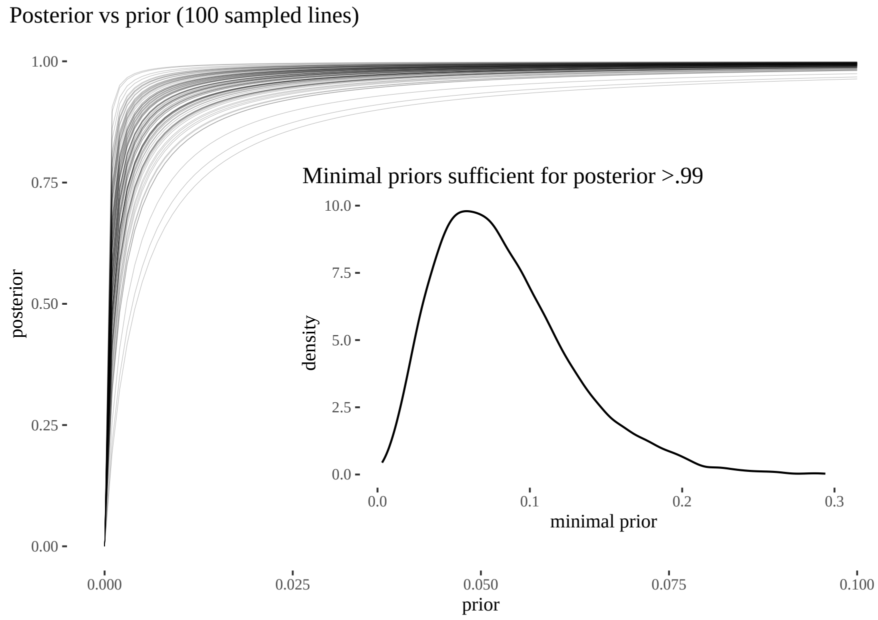

Figure 4: 100 lines illustrating the uncertainty about the dependence of the posterior on the prior given aleatory uncertainty about the evidence, with the distribution of the minimal priors required for the posterior to be above .99.

This, then, is the working hypothesis of this chapter: whenever density estimates for the probabilities of interest are available (and they should be available for match evidence whose reliability has been properly studied), those densities should be reported for assessing the strength of the evidence. This approach avoids hiding actual aleatory uncertainties under the carpet. It also allows for a balanced assessment of the evidence, whereas using point estimates or intervals may exaggerate or underestimate the value of the evidence.

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 a higher-order approach whenever possible. Second, having gained this perspective, we revisit 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.

2 Three Probabilisms

The introduction outlined three probabilistic approaches that one might take for assessing the value of the evidence presented at trial. The first approach uses precise probabilities; the second uses intervals; the third uses distributions over probabilities. By relying on an example featuring two items of match evidence, we suggested that the third approach is preferable. This section buttresses this claim by providing principled, philosophical reasons in favor of the third approach.

The three approaches we considered correspond (roughly) to three ways in which probabilities can be deployed to model a rational agent's fallible and evidence-based beliefs about the world. The first approach, known in the philosophical literature as precise probabilism, posits that an agent's credal state is modeled by a single, precise probability measure. The second approach, known as imprecise probabilism, replaces precise probabilities by sets of probability measures. The third approach, what we call 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 also 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 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 fallible beliefs reflect the evidence we have (or have not) obtained. A couple of stylized examples should make the point clear. (For the sake of simplicity, we will use examples featuring coins, but coin biases can be thought of as random match probabilities in the forensic context.)

> **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, based on overwhelming evidence, that the coin is fair.

On precise probabilism, both scenarios are represented by assigning a probability of .5 to the outcome *heads*. If you are completely ignorant, the principle of insufficient evidence suggests that you assign .5 to both outcomes. Similarly, if you know for sure the coin is fair, assigning .5 seems the best way to quantify the uncertainty about the outcome. The agent's evidence in the two scenario is quite different, but precise probabilities cannot capture this difference.

> **Learning from ignorance** You toss a coin with unknown bias. You toss it 10 times and observe *heads* 5 times. Suppose you toss it further and observe 50 *heads* in 100 tosses.

Since the coin initially had unknown bias, you should presumably assign a probability of .5 to both outcomes. After the 10 tosses, you end up again with an estimate of .5. You must have learned something, but whatever that is, it is not modeled by precise probabilities. When you toss the coin 100 times and observe 50 heads, you learn something. But your precise probability assessment will again be .5.

These examples suggest that precise probabilism is not appropriately responsive to evidence. It ends up assigning the same probability in situations in which one's evidence is quite different. For instance, when no evidence is available about the coin's bias, when there is little evidence that the coin is fair (say, after only 10 tosses), and when there is strong evidence that the coin is fair (say, after 100 tosses). The general problem is, precise probability captures the value around which your uncertainty should be centered, but fails to capture how centered it should be given the evidence.[8]

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 is to be represented by means of a *set of*

[8] 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 $r(X \cap Y) = 5/18 \neq r(X)r(Y) = 1/4$.

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 with the evidence. For instance, if an agent knows that the coin is fair, their credal state would be captured by the singleton set $\{\mathsf{P}\}$, 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.[9]

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:

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

The hope is that, if we start with a range of probabilities that is not extremely wide, point-wise learning will behave appropriately. 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 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, if you observed at least one tail and one heads, you can exclude the measures assigning 0 or 1 to *heads*. But what else have you learned? If you are to update your representor set 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 anything if you start from complete ignorance.[10]

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.[11] But no matter what we think about belief inertia, other problems plague imprecise probabilism. Two 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

[9] For the development of imprecise probabilism, see Keynes (1921); Levi (1974); Gärdenfors & Sahlin (1982); Kaplan (1968); Joyce (2005); Fraassen (2006); Sturgeon (2008); Walley (1991). S. Bradley (2019) is a good source of further references. Imprecise probabilism shares some similarities with what we might call **interval probabilism** (Kyburg, 1961; Kyburg Jr & Teng, 2001). 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. As we have already noted, intervals do not contain probabilistic information sufficient to guide reasoning with multiple items of evidence. So we focus on IP, which is the more promising approach.

[10] 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.

[11] 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, as well as a principled account of why one should use a separate special update rule when starting with complete ignorance.

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 or assigning probabilities to chance hypotheses, in which case it is no longer clear whether the object-level imprecision performs any valuable task.[12]

Second, besides descriptive inadequacy, an even deeper, foundational problem exists for imprecise probabilism. This problem arises when we attempt to measure the accuracy of a representor set of probability measures. Workable *scoring rules* exists for measuring the accuracy of a single, precise credence function, such as the Brier score. These rules measure the distance between one's credence function (or probability measure) and the actual value. A requirement of scoring rules is that they be *proper*: any agent will score their own credence function to be more accurate than every other credence function. After all, if an agent thought a different credence was more accurate, they should switch to it. Proper scoring rules are then used to formulate accuracy-based arguments for precise probabilism. These arguments show (roughly) that, 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. Can the same be done for imprecise probabilism? It seems not. Impossibility theorems demonstrate that no proper scoring rules are available for representor sets. 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 certain plausible formal constraints, it will never strictly recommend an imprecise stance, as for any imprecise stance there will be a precise one with at least the same accuracy.

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 (S. Bradley, 2012, p. 157):

> . . . 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'.

Note, however, that such bunching up cannot be modeled by imprecise probabilism. Joyce (2005), in a paper defending imprecise probabilism, in fact uses a density over chance hypotheses to account for the notion of evidential weight. 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. 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 affecting precise ad imprecise probabilism.

The key idea of the higher-order approach we propose 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.[13] From this perspective, when an agent is asked about their credal stance towards X, they can refuse to summarize it

[12] 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.

[13] Bradley admits this much (S. Bradley, 2012, p. 90), and so does Konek (Konek, 2013, p. 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.

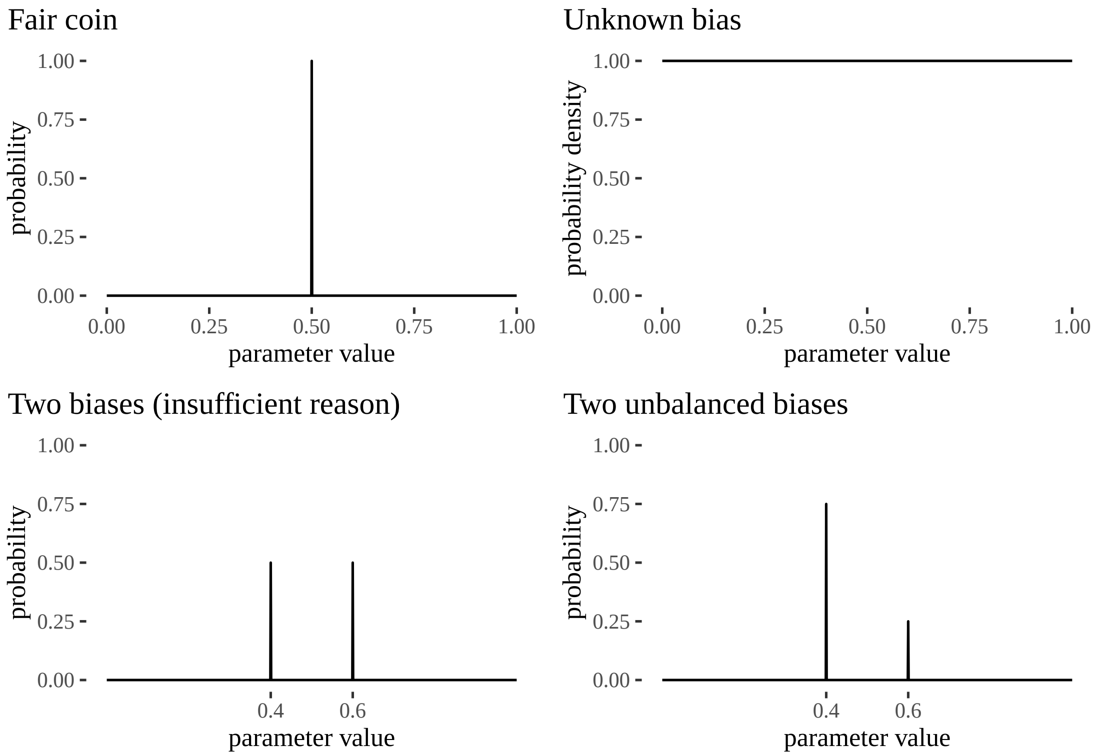

Figure 5: Examples of higher-order distributions for scenarios brought up in the literature.

in terms of a point value $\mathsf{P}(X)$. They can instead express their credal stance in terms of a probability (density) distribution f_x treating $\mathsf{P}(X)$ as a random variable. To be sure, an agent's credal state toward X could sometimes be usefully represented by the expectation

$$\int_0^1 xf(x)\,dx$$

as the precise, object-level credence in X, where f is the probability density over possible object-level probability values. But this need not always be the case. If the probability density f is not sufficiently concentrated around a single value, a one-point summary might fail to do justice to the nuances of the agent's credal state.[14] For example, consider again the scenario in which the agent knows that the bias of the coin is either .4 or .6 but the former is three times more likely. The agent might refuse to summarize their credal state with the expectation $\mathsf{P}(X) = .75 \times .4 + .25 \times .6 = .45$.

The higher-order approach can easily model all the challenging scenarios we discussed so far in the manner illustrated in Figure 5. In particular, the scenario in which the two biases of the coin are not equally likely—which imprecise probabilism cannot model—can be easily modeled within high-order probabilism by assigning different probabilities to the two biases.

Besides its flexibility in modelling uncertainty, higher-order probabilism does not fall prey to belief inertia. Consider a situation in which you have no idea about the bias of a coin. So you start with a uniform density over $[0,1]$ as your prior. By using binomial probabilities as likelihoods, 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. The posterior distribution will become more centered around the parameter estimate as the observations come in. Figure 6 shows—starting with a uniform prior distribution— how the posterior distribution changes after successive observations of heads, heads again, and then tails.[15]

[14] This approach lines up with common practice in Bayesian statistics, where the primary role of uncertainty representation is assigned to the whole distribution. Summaries such as the mean, mode standard deviation, mean absolute deviation, or highest posterior density intervals are only succinct ways for representing the uncertainty of a given scenario. Whether the expectation should be used in betting behavior is a separate problem. Here we focus on epistemic issues.

[15] More generally, learning about frequencies, assuming independence and constant probability for all the observations, is modeled the Bayes way. You start with some prior density p over the parameter values. If you start with complete lack of information,

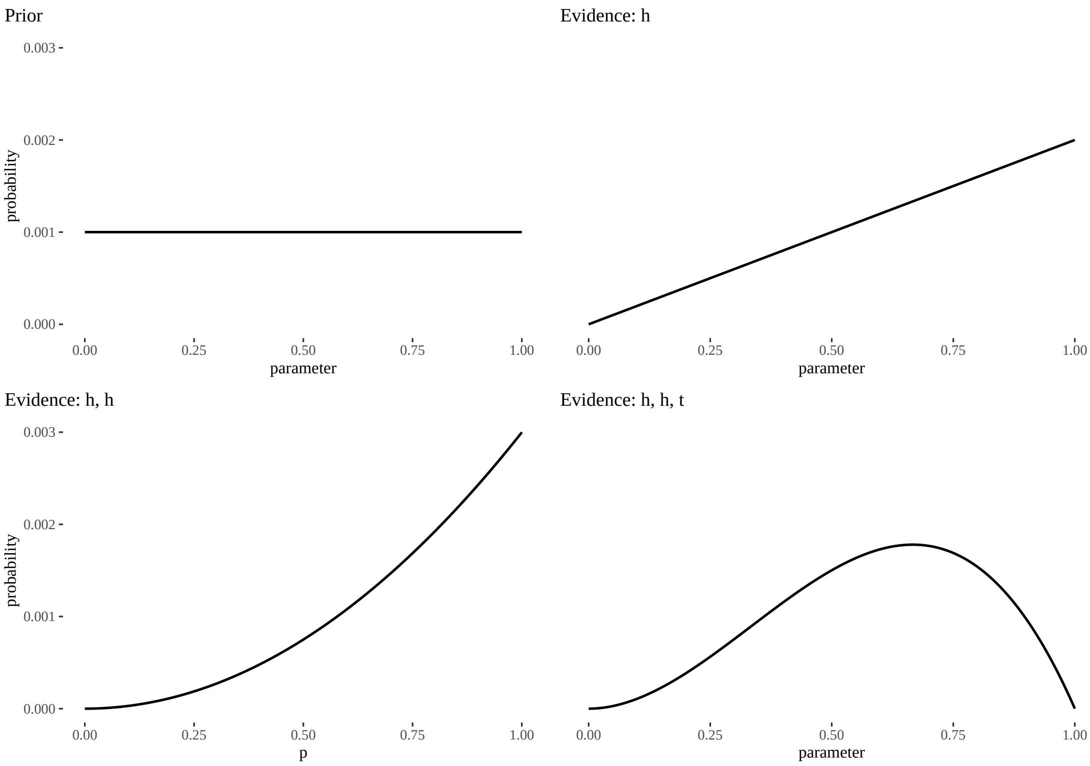

Figure 6: As observations of heads, heads and tails come in, extreme parameter values drop out of the picture and the posterior is shaped by the evidence.

A further advantage of high-order probabilism over imprecise probabilism is that the prospects for accuracy-based arguments are not foreclosed. This is a significant shortcoming of imprecise probabilism, especially because such arguments exist for precise probabilism. One can show that there exist proper scoring rules for higher-order probabilism. These rules can then be used to formulate accuracy-based arguments. Another interesting feature of the framework is that the point made by Schoenfield against imprecise probabilism does not apply: there are cases in which accuracy considerations recommend an imprecise stance (that is, a multi-modal distribution) over a precise one ((Urbaniak, 2022 manuscript).

All in all, higher-order probabilism outperforms both precise and imprecise probabilism, at the descriptive as well as the normative level. From a descriptive standpoint, higher-order probabilism can easily modeled a variety of scenarios that cannot be adequately modeled by the other versions of probabilism. From a normative standpoint, accuracy maximization may sometimes recommend that a rational agent represent their credal state with a distribution over probability values rather than a precise probability measure.

---

p should be uniform. Then, you observe the data D which is 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:

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

3 Objections to the higher-order approach

This section addresses a number of conceptual difficulties that may arise in using higher-order probabilities, with focus on those brought up by prominent legal evidence scholars. In discussing these conceptual issues, we will formulate an accuracy-based argument that using higher-order probabilities is actually preferable to precise probabilities.

Our main polemic target is a discussion initiated by Taroni, Bozza, Biedermann, & Aitken (2015), who argue extensively that trial experts should only report point estimates. Their point of departure is a reflection on match evidence. Say an expert reports at trial that the sample from the crime scene matches the defendant. The significance of this match should be evaluated in light of the population frequency θ of the matching profile. This frequency, however, cannot be known for sure and must instead be estimated. The expert will estimate the true parameter θ by means of a probabilistic distribution $p(\theta)$ over its possible values. For example, 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.

Nothing so far should be controversial. However, the question arises of how the expert should report their own uncertainty about θ. To fix the notation, let the prosecution hypothesis H_p be that the suspect is the source of the trace, and the defense hypothesis H_d that another person, unrelated to the suspect, is the source. For simplicity, assume that if H_p holds, the laboratory will surely report a match M, so that $\mathsf{P}(M|H_p) = 1$. The likelihood ratio, then, reduces to $1/\mathsf{P}(M|H_d)$. Taroni et al. (2015) claim that the probability of the match evidence given the defense hypothesis should be calculated as follows:

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

In case of a DNA match, they recommend that the expert report the expected value of the $\mathtt{beta}$ distribution, which reduces to $\alpha+s+1/\alpha+\beta+n+1$. They claim that this number satisfactorily expresses the posterior uncertainty about θ. For them, it is this probability alone that should be used in the denominator in the calculation and reporting of the likelihood ratio.

Sjerps et al. (2015) disagree. In reporting a single value, the expert would refrain from providing the fact-finders with relevant information that can make a difference in the evaluation of the evidence. There is a difference between (a) an expert who is certain θ is $.1$; (b) an expert whose best estimate of θ is $.1$ based on a thousand of observations; and (c) an expert whose best estimate of θ is again $.1$ but based on only ten observations. As our earlier discussion of precise probabilism makes clear, a simple point estimate (or precise probability) would fail to capture these differences.[16]

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.

[16]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 this is unsatisfactory. Depriving the fact-finders of 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 is wildly optimistic, to say the least.

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 (yes, we need three hands to juggle the concepts here), 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 $\mathtt{beta}(1+\mathtt{successes}, 1+\mathtt{samplesize}-\mathtt{successes})$ 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 7.

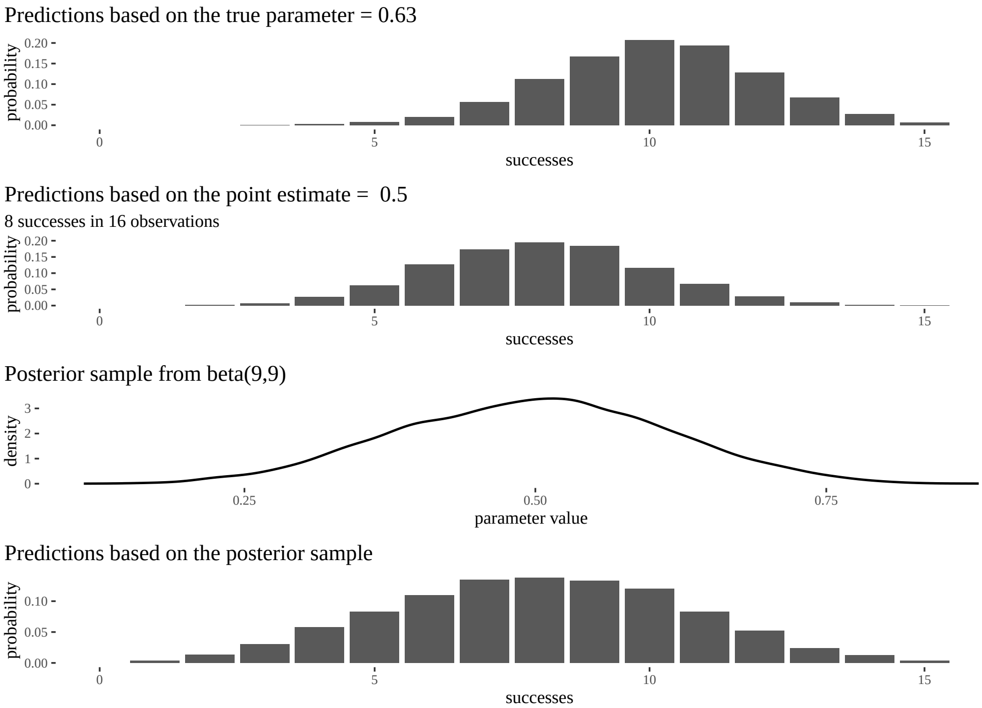

Figure 7: Real probability mass, probability mass calculated using a point estimate, sampling distribution from the posterior, and the posterior predictive distribution based on this sampling distribution.

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 8)—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 over point-based predictions. That is, accuracy-wise, point-estimate-based PMFs are systematically worse than posterior predictive PMFs.

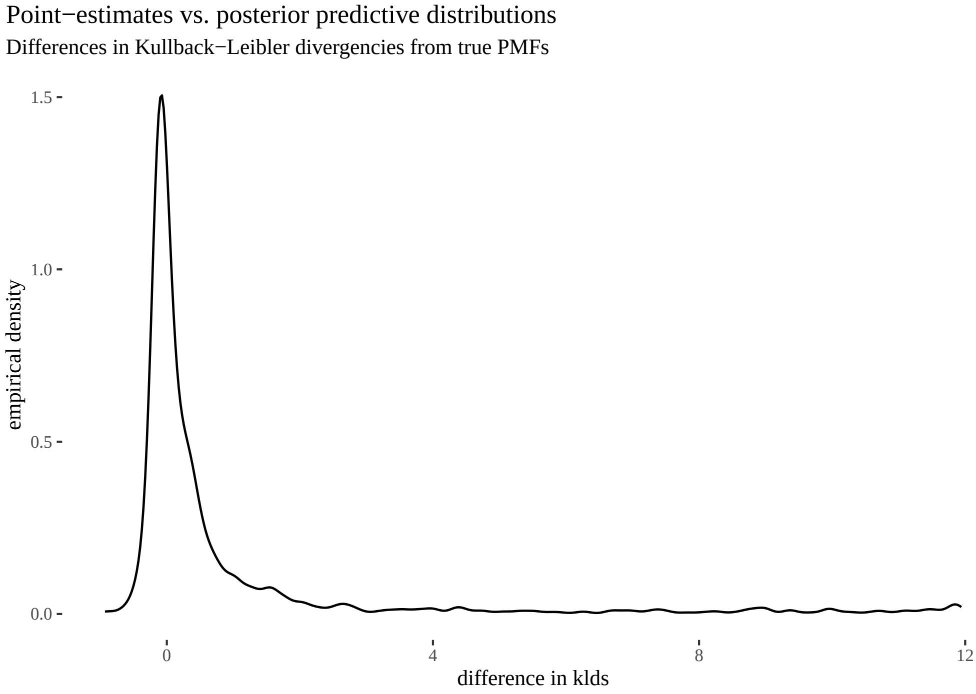

Figure 8: Differences in Kullback-Leibler divergencies from the true distributions, comparing the distributions obtained using point estimates and posterior predictive distributions. Positive values indicate the point-estimate-based PMF was further from the true distribution that its competitor.

Accuracy considerations aside, we can also engage with the more philosophical arguments on the market. 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.

As the betting behavior interpretation is actually not as obvious or uncontroversial as Taroni et al. (2015) suggest,[17] 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. So is estimation of one's uncertainty pointless?

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 $\mathtt{beta}$ prior with parameters α, β. Say they observe s matches and the database size is n. They reach a $\mathtt{beta}(\alpha+s+1, \beta+n-s)$ 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),[18] they want to

[17] See philosophical textbooks on formal epistemology such as (D. Bradley, 2015) or (Titelbaum, 2020).
[18] Otherwise they need to think harder and dampen their convictions.

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 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.

Another worry raised by Taroni et al. (2015) is that first-order probabilities 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 we can get get closer to 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.

Taroni et al. (2015) insist that the situation is even worse for likelihood ratios, as 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.

Taroni et al. (2015) argue also that once we allow second-order probability we run into the threat of infinite regress. But do we? 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.

4 Examples of applications

4.1 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 (Thompson, Taroni, & Aitken, 2003, p. 3):

> 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.

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:

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

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 11 (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 is what happens between these edges.

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 9.

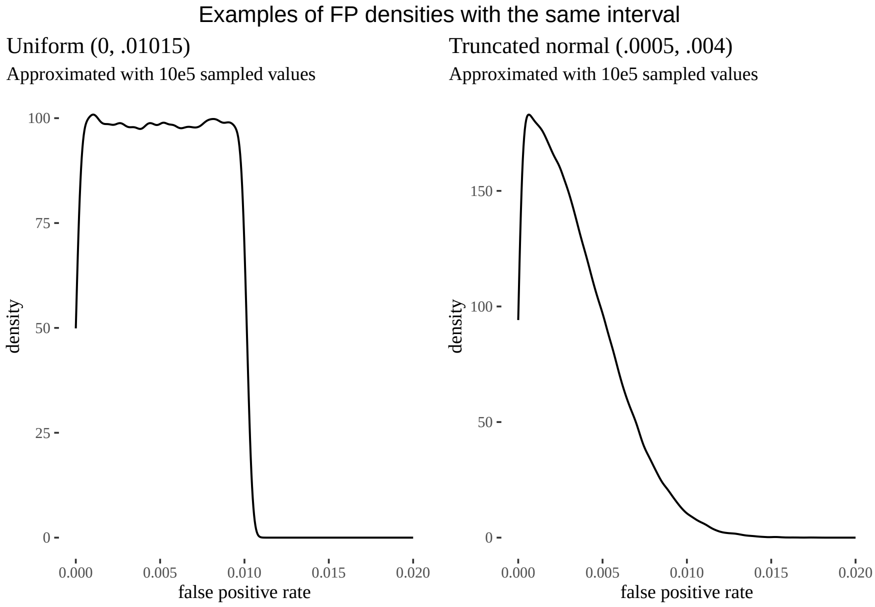

Figure 9: Two examples of assumptions about the false positive rates, both having pretty much the same 99% highest density intervals. (Top) all error rates are equally likely, (Bottom) the most likely values are closer to 0, but also some high values while unlikely are possible.

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 10.

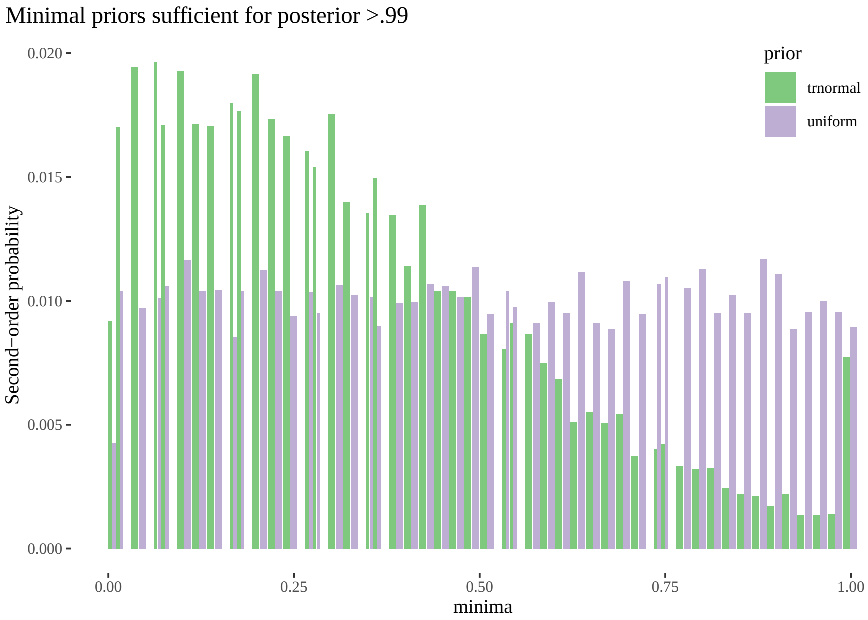

Figure 10: The distribution of minimal priors sufficient for the posterior being above .99 on the two distributions of false positive rates. Note that the truncated normal distribution has its bulk towards the left, but at the same time has higher ratio of evens in which this posterior is never reached.

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 11.

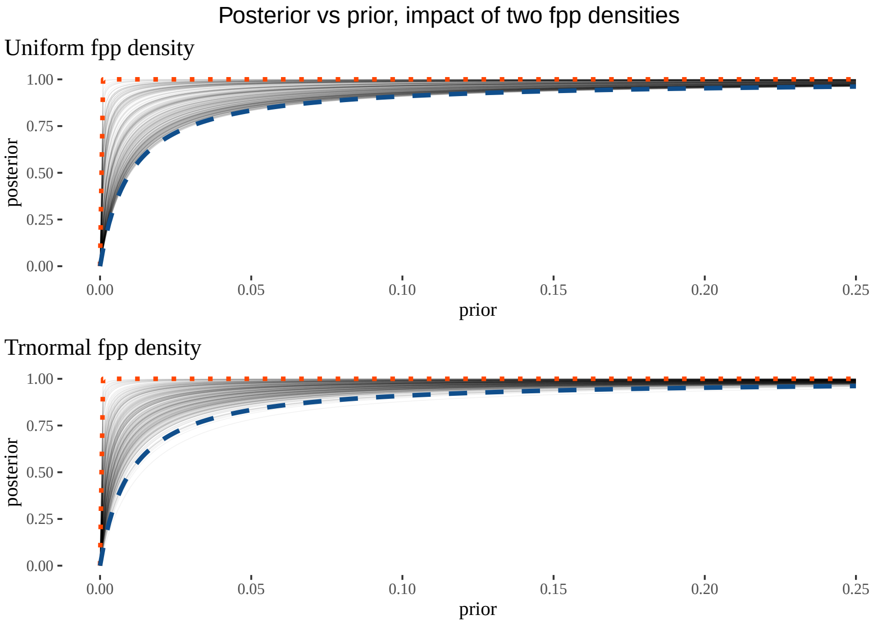

Figure 11: Impact of prior on the posterior assumign two different densitites for false positive rates. Note how both the "pristine" error-free point estimate (orange) and the charitable version (blue) are quite far from where the bulks of the distributions in fact are. Note also how the trnormal density 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.

4.2 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 (O'Hagan et al., 2006), implementation is not computationally unfeasible, as we can approximate densities using sampling. To illustrate, let us start with a simplified BN developed by Fenton & Neil (2018) to illustrate how conviction was unjustified in the Clark case (Figure 12). to illustrate a point about the notorious Sally Clark case (Figure 12).[19] 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

[19] 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.

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.

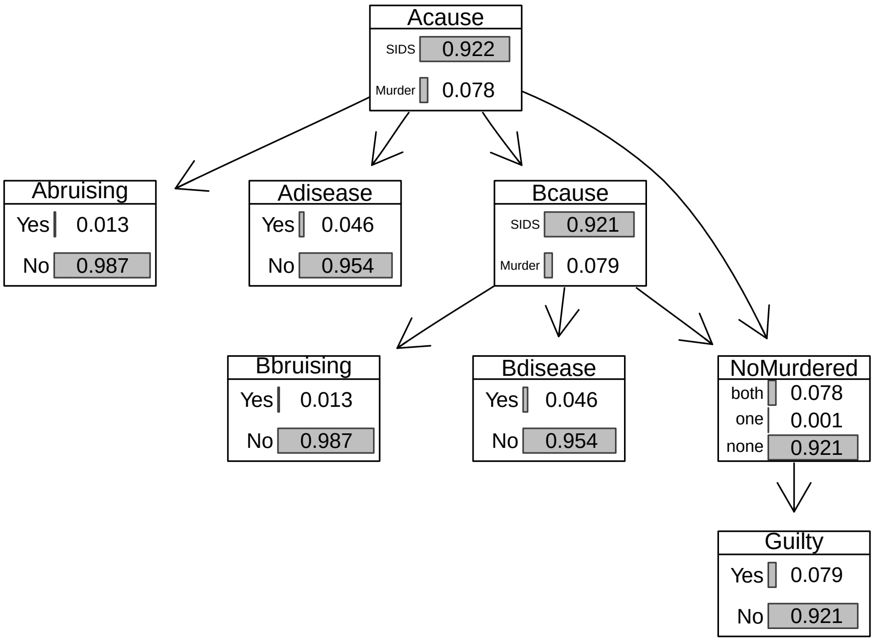

Figure 12: The BN developed by FENTON ET AL., with marginal prior probabilities.

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 13).

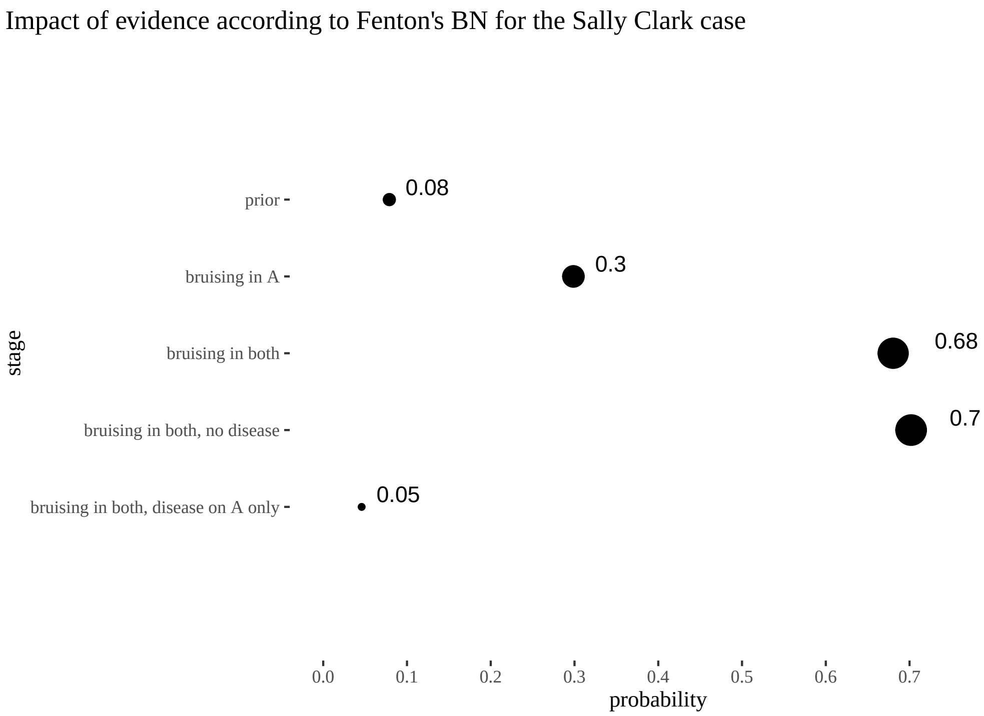

Figure 13: The prior and posterior probabilities for Fenton's Sally Clark BN.

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 14. One of them is based on a truncated normal distribution to emphasize that the framework give us much freedom in the specification of distributions.

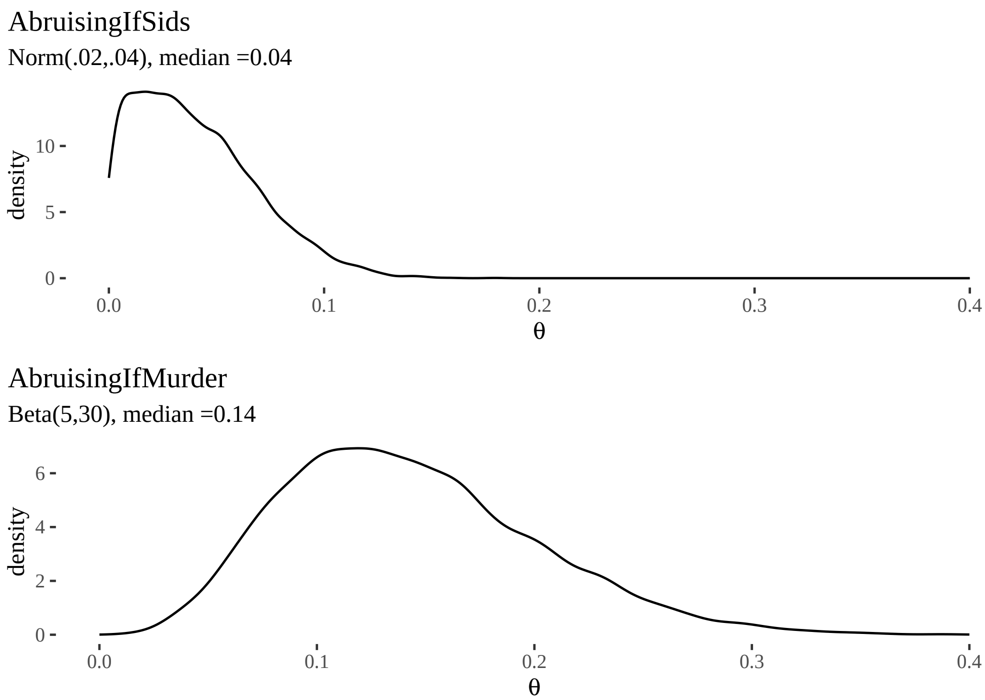

Figure 14: Example of approximated uncertainties about conditional probabilities in the Sally Clark case.

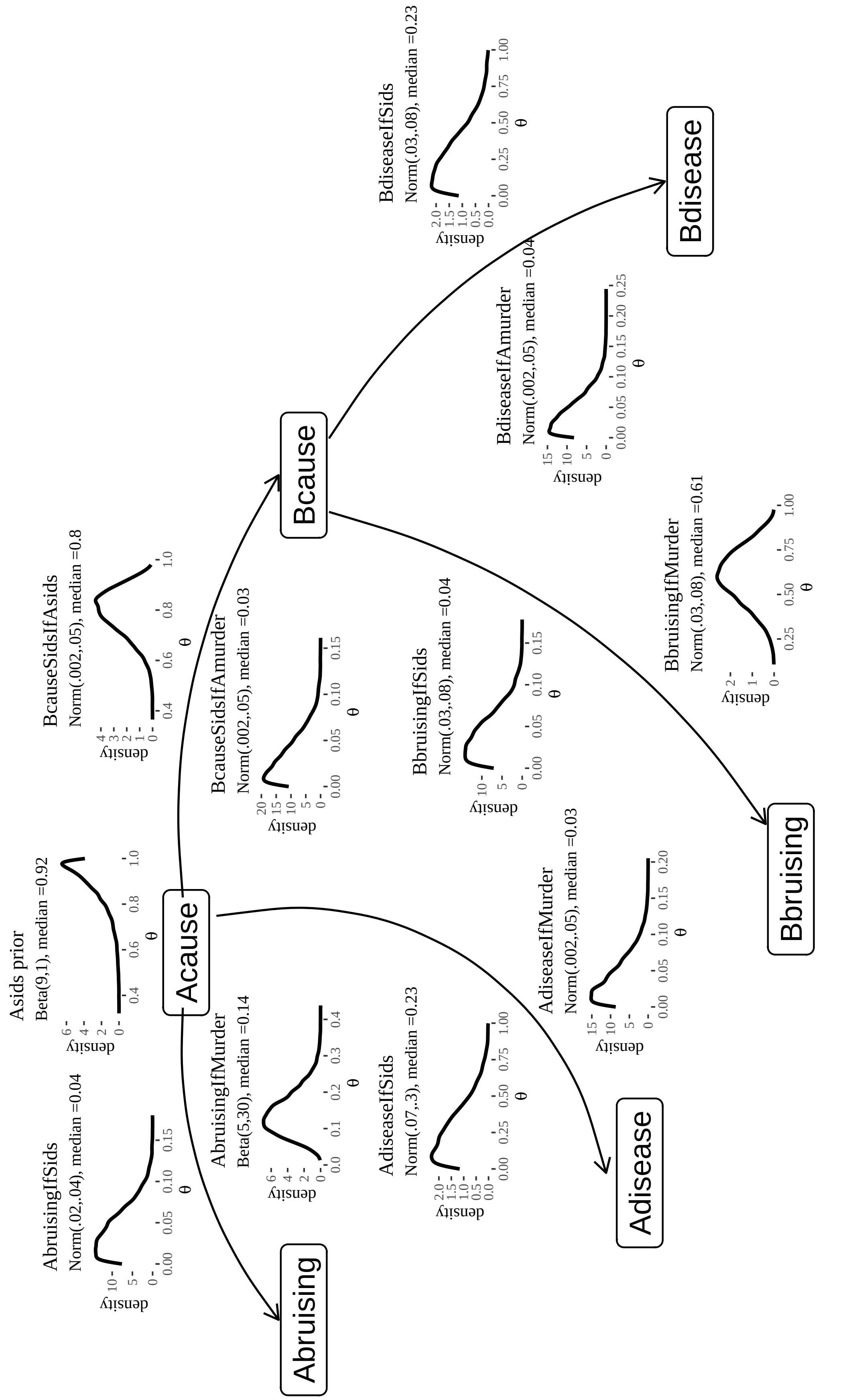

Figure 15: Example of a HOP approach tor the Sally Clark Case approximated by sampling probabilities and constructing 10k BNs.

Using these we can investigate the impact of incoming evidence as it arrives (Figure 16). 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.

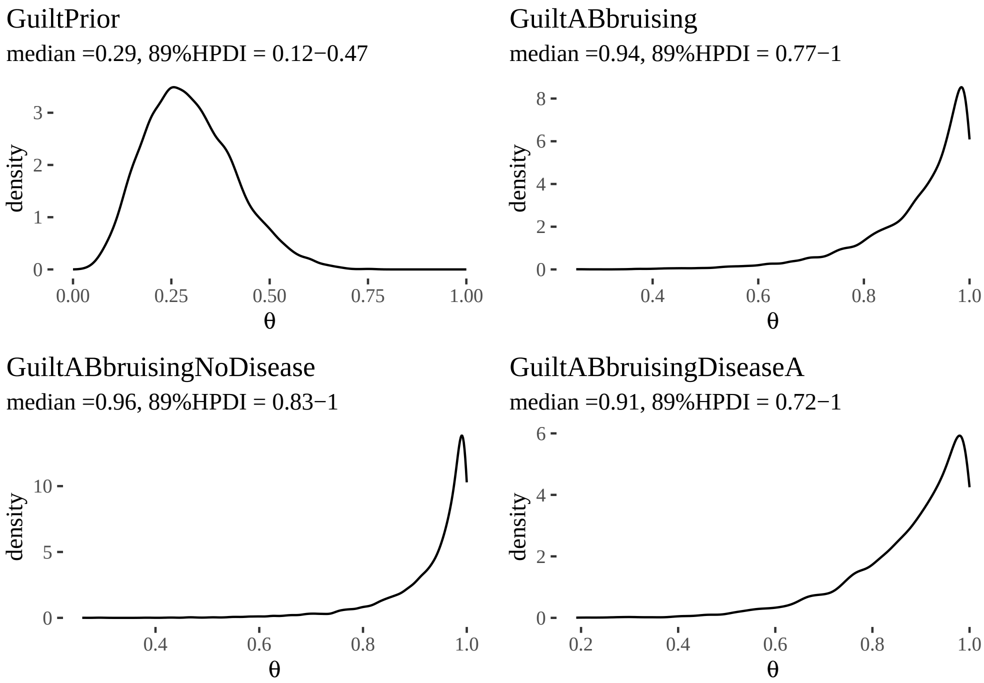

Figure 16: Impact of incoming evidence in the Sally Clark case.

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 17).

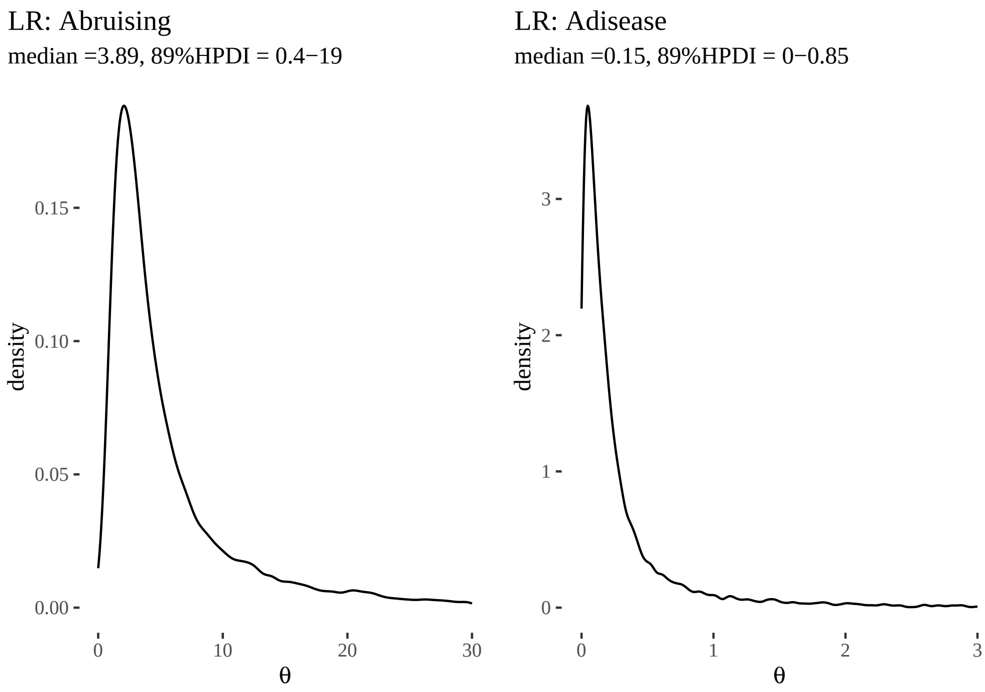

Figure 17: Likelihood ratios forbruising and signs of disease in child A in the Sally Clark case.

5 Weight of evidence

We now illustrate one of the theoretical payoffs of higher-order probabilism. It allows to develop an elegant theory of the weight of evidence that outperforms the existing proposals. We will start with an informal sketch of the idea of the weight of evidence, as opposed to the balance of the evidence. We will then explore a few attempts at modeling this idea, first based on precise probabilism and then based on imprecise probabilism. Finally, we will show how higher-order probabilism can offer a better theory.

5.1 Motivating examples

In the 1872 manuscript *The Fixation of Belief* (W3 295), C. S. Peirce makes the following observation about sampling from a bag of beans that are either black or white:

> When we have drawn a thousand times, if about half 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.

In both cases, our best assessment of the probability that the next draw will be a black bean is .5, but how sure we should be of that assessment is quite different depending on whether it is based on two or one thousands draws. In other words, the *weight* of the evidence seems much greater after drawing a thousand beans and finding out that half are black, compared to drawing just two beans and again finding out that half are black. The weight of the evidence is different in the two cases, but its *balance*—understood here as the empirical proportion of black-to-white beans—is the same.

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. For the latter, Peirce proposes to use some dispersion-related measure of error (but then suggests that an error of that estimate should also be estimated and so on, so that ideally more numbers representing errors would be needed).

Peirce did not use the expression the weight of evidence (and, in fact, he used the phrase to refer to the balance of evidence, W3 294) (Kasser, 2016). However, his remarks anticipated what came to be called weight of evidence by Keynes in his 1921 *A Treatise on Probability*:

> As the relevant evidence at our disposal increases, the magnitude of the probability of the

> argument may either increase or decrease, according as the new knowledge strengthens the unfavourable or the favourable evidence; but something seems to have increased in either case—we have a more substantial basis upon which to rest our conclusion. I express this by saying that an accession of new evidence increases the weight of an argument. (p. 71)

The key point is the same (Levi, 2011): since the balance of probability alone cannot characterize all important aspects of evidential appraisal, another dimension—the weight of an argument—must be deployed to quantify uncertainty.

Keynes entertained the possibility of measuring weight of evidence in terms of the variance of the posterior distribution of a certain parameter, but was quite attached to the idea that weight should increase with new information, even if the dispersion may increase with new evidence [TP 80-82]. So he proposed only a very rough sketch of a positive proposal. Moreover, he was unclear how a measure of weight could be part of decision-making. He was ultimately skeptical about the practical significance of the notion [TP 83].

5.2 Monotonicity?

It is instructive to examine more closely Keynes' claim that the weight of evidence, unlike its balance, always increases as more evidence is taken into account. While balance can oscillate one direction or the other, weight would seem to always increase. We can state this requirement as follows:

(Monotonicity) If E is relevant to X given K, where K is background knowledge, $V(X|K \wedge E) > V(X|K)$, where V is the weight of evidence.

Monotonicity is corroborated by Peirce's example about drawing from a bag of beans. As the sample size increases and the relative proportion of black-to-white beans remains constant, the weight of the evidence increases. That is:

(Weak increase) In urn-like cases, the evidential weight obtained by a larger sample is greater, if the relative frequencies in the samples remain the same.

This formulation can be strengthened by dropping the assumption of equal relative frequencies:

(Strong increase) In urn-like cases, the evidential weight obtained by a larger sample is higher.

We think there are good reasons to reject Monotonicity and Strong Increase, but we agree with Weak Increase. Monotonicity and Strong Increase are consistent with a certain conception of the weight of evidence, what we might call *quantity* of evidence. After all, there is no doubt that, as more evidence is taken into account, the quantity of the evidence must increase. But the weight of evidence need not be identified with its quantity alone. As more evidence is taken into account, the new evidence may speak less clearly in favor or against a hypothesis. For consider this example:

> **A (possibly) rigged lottery:** Initially, you think the lottery is fair. You have no reason to doubt that. So, you calculate precisely the probability that a certain ticket number will be drawn. Then, rumors begin to surface that the lottery is rigged and that only numbers that satisfy a complicated equation will be drawn. You now have more relevant evidence at your disposal, but that evidence is more confusing and muddled than before.

Arguably, this is a scenario in which the quantity of evidence has increased, but the weight of evidence has not. If this is right, weight and quantity can come apart.

So, we seek a theory of evidential weight that can model two intuitions. The first is that, as the sample size increases, the weight of the evidence must also increase (under certain conditions, along the lines of Weak Increase). The second intuition is that that, even when the quantity of evidence increases, the weight of evidence might not (as illustrated by the example of the rigged lottery).

5.3 Weight and precise probabilism

An obvious place to look for a theory of the weight of evidence is within precise probabilism. But, we will argue, our best bet is to look beyond it.

5.3.1 Absolute value of the likelihood ratio

Weight of evidence can be modeled by generalizing the likelihood ratio as a measure of the value of evidence. The likelihood ratio is a *directional* measure: the evidence may be favorable or unfavorable to a hypothesis (compared to another). By contrast, Keynes' remarks suggests that weight is a non-directional measure. It appears to always increase no matter the balance (though we will ultimately reject this claim). To strip the likelihood ratio of its directionality, it is enough to take the absolute value of the natural log (Nance, 2016, sec. 3.5). The weight of evidence E relative to the pair of hypotheses H, H' would be

$$|\ln(LR_{H,H'}(E))|,$$

where $LR_{H,H'}(E) = \frac{P(E|H)}{P(E|H')}$. By the properties of logarithms, $ln(1/x) = -ln(x)$, so two items of evidence of equivalent strength—but opposite directionality—would have the same weight. So, for example, $|ln(1/3)| = |ln(3)| = 1.61$.

This account is simple and elegant, but faces two difficulties. The first is technical in nature. As many have pointed out, there is a problem with decomposition. Consider two items of evidence that, taken together, have a likelihood ratio of one, say one has a likelihood ratio of 1/3 and the other of 3. Assuming they are probabilistically independent given a hypothesis of interest, their combined likelihood ratio results from multiplying the individual likelihood ratios. Thus, their combined weight would be zero since $\ln(LR_H(E1 \wedge E_2)) = \ln(1) = 0$. However, by adding the weights one by one, the combined weight would be different from zero, since $|ln(1/3)| + |ln(3)| = 3.22$. So, do the two items of evidence have zero weight or not? Depending on how evidence is decomposed, it appears to have different weights.

There might be a technical fix to the problem of decomposition. But another, conceptually deeper problem exists. Whether the log of the likelihood has a positive or negative sign, its absolute value is always positive. Therefore, as more evidence is added, the weight of evidence would always increase. This shows that this account agrees with Monotonicity and Strong Increase, and in this sense, it only captures the notion of quantity of evidence. The problem, then, is that it cannot capture the intuition underlying the scenario of the rigged lottery in which, even when quantity increases, weight does not.

5.3.2 Completeness and resilience

Another strategy, within precise probabilism, is to model notions that are related, albeit not identical, to weight. The literature has suggested that precise probabilism contains the conceptual resources to model two related ideas: resilience (or stability) and completeness. Roughly speaking, resilience tracks how additional evidence may change the current probability assessment, while completeness tracks how the fact that some evidence is missing may change the current probability assessment.

Skyrms (1977) offers the following account of resilience: a precise probability assessment of H based on some evidence E, say $P(H|E)$, is resilient (or stable) whenever it does not wiggle too much relative to a reasonable set of additional items of evidence E' that can be conditioned on. In other words, $P(H|E)$ is resilient enough if and only if the absolute difference $|P(H|E) - P(H|E \wedge E')|$ (for any element E' in a reasonable set of additional evidence) is within a certain acceptable limit. For an account of completeness, Kaye's sketches the following proposal: instead of merely assessing $P(H|E)$, CITE KAYE, which one?
one should assess $P(H|E \wedge M)$, where M is the known fact that some evidence is missing. Whenever the evidence available is missing information that one would reasonably expect to see in a case (in a sense to be specified), this fact must be taken into account in assessing the probability of the hypothesis of interest.[20]

A good account of resilience and completeness should address a number of questions. For example, in assessing resilience, which items of additional evidence should be included in the reasonable set?[21]

[20] Interestingly, the missing evidence is not a higher-order fact about the currently available evidence, but itself a fact that counts as evidence together with other facts. This, we think is a weakness in the theory, because much of the heavy lifting has to be done by assessing the conditional probabilities involving missing evidence conditions and these do not fall out of the theory, but have to be plugged in by hand on a case-by-case basis.

[21] The set must be restricted somehow, or else no probability assessment would ever be resilient. Another problem is that resilience consists in evaluating a probability based on evidence not yet available. The evidence could take different values—for example, in a criminal case, the additional evidence could be exculpatory or incriminating. How can additional evidence whose value is unknown affect one's assessment of the probability of a hypothesis? Another question is, which items of evidence should be included is the reasonable set?

Or, in assessing completeness, when should an item of evidence count as missing?[22] But, it is clear that resilience and completeness both play a role in trial proceedings. On appeal, a defendant might argue that, had additional evidence be taken into account, the verdict should have been different. Questions of missing evidence are also routinely brought up in litigation. And the rules of evidence often deal with questions of missing evidence and offer a variety of remedies, from jury instructions to sanctions.

So, a good theory of how evidence is assessed at trial should include an account of completeness and resilience. A question remains, however. Are resilience and completeness the theory of weight we are looking for? This is far from clear. We think that weight, completeness and resilience are conceptually distinct notions. And, they are likely to play distinct roles in legal fact-finding. Or perhaps completeness and resilience, together with precise first-order probabilities, are all we need to assess evidence at trial, and the notion of weight is redundant. These are difficult questions that require a close examination of each notion. We focus here on weight and leave the discussion of how weight compares to completeness and resilience to another time.

5.4 Weight and imprecise probabilism

A theory of weight can also be formulated using the conceptual resources of imprecise probabilism. The idea is this: a body of evidence is weightier whenever the representor set of probability measures compatible with the evidence is in some sensible sense smaller. Consider a case of complete ignorance about the bias of a coin. The representor set will contain *any* probability measure. This corresponds to complete lack of evidence and null weight, as expected. At the other extreme, consider a case in which the fairness of the coin is known for sure. The supporting evidence here would have maximal weight and the representor set would only contain the precise probability measure that assigns .5 to the two outcomes. All other intermediate cases would fall somewhere in between.[23]

CITE WEATHERSON ON WEIGHT AND IMPRECISE PROBABILISM

This account accommodates the intuition underlying the rigged lottery example. As more evidenced is accumulated, the representor set can become larger and include more probability measures than before. Thus, weight of evidence can decrease even when the quantity of evidence increases. For this reason, a theory of weight based on imprecise probabilism is promising. The problem, however, is that this theory inherits the difficulties of imprecise probabilism, which we have already mentioned. For example, recall that imprecise probabilism cannot model a situation in which an epistemic agent believes that the coin could have bias .8 or .4, but thinks that one bias is more likely than the other. Arguably, evidence compatible with the coin having two equally likely biases should possess different weight than evidence compatible with the coin having two biases one more likely than the other. The latter situation is closer to full weight—the situation in which the exact bias of the coin is known for sure. But a theory of weight based on imprecise probabilism is ill-equipped to accommodate these nuances.

5.5 Weight and higher order probabilism

We are finally ready to present our own account of the weight of evidence. This accout has two distinctive features. First, it is based on higher-order probabilism. Second, it is information-theoretic. To develop this account, we will begin with a very short introduction to Shannon's theory of information.

5.5.1 Entropy of a distribution

Let X be a random variable and P a probability distribution over its values. Shannon's measure of

I think you crippled the explanation too much, as now there is no connection to expected surprise, more needs to be said here. If not today, at some later point.

[22] Admittedly, there are cases in which it is clear that evidence is missing, say because it was destroyed, lost or could not be retrieved or collected for one reason or another. But evidence could be missing in a more speculative sense. For example, if it is routine to see breathalyzer evidence in drunk driving cases, what should one make of the fact that such evidence is missing in a given case? The more general worry here is that evidence is always incomplete. Peirce's example of drawing from a bag of beans makes this clear. In principle, more draws could be made and more evidence could be gathered. Another difficulty concerns the impact that the missing evidence should have on the other evidence. For example, in a case of tax fraud, suppose the recording of phone conversation was deleted by accident. There is no dispute about that. Now, should this fact increase or decrease the probability that the defendant committed fraud?

[23] More formally, let x_1 and x_2 the two extreme probability assignments in the representor set compatible with the available evidence E. Then, the weight of E would be 1-$x_1 - x_2$. As expected, if there is only one probability measure, the weight would be one.

information, $H(X)$, reads:

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

Consider the simple case in which X can take two values—outcome 1 and outcome 0—whose probabilities are p and $1-p$. Figure (**REFER TO PLOT of H(X)**) shows $H(X)$ as a function of p. As is apparent from the graph, $H(X)$ is greatest when the two outcomes have an equal probability of .5. The more the probabilities deviate from .5, the more $H(X)$ approaches zero. WHAT PLOT?

To make sense of this, $H(X)$ can be thought as the *entropy* (i.e. the lack of information) contained in the distribution associated with random variable X. When the two outcomes have equal probability, entropy is greatest. When they have different probabilities, one outcome will be more probable than the other. The more probable one of the outcomes (and thus the less likely the other), the lower the entropy. Intuitively, $H(X)$ captures the idea that entropy is greatest when the indecision about which outcome will occur is maximal, and the entropy decreases when such indecision decreases.

If $H(X)$ is a measure of entropy—that is, a measure of lack of information—why call it a measure of information? $H(X)$ is also a measure of information in the following sense: it describes the *expected amount of information* one would receive upon learning the actual value of X. After all, the higher the entropy, the less informative a distribution, the more you expect to learn upon finding out the actual value of X. Conversely, the lower the entropy of the distribution, the more informative the distribution, the less one expects to learn upon finding out the actual value of X.[24]

5.5.2 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 than 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. How should we measure the informativeness (or weight) of a distribution?

Since informativeness is the opposite of entropy, it is tempting to take the short route and define $W(P)$ as $1-H(P)$. But, for reasons that will be become clear (**WHEN?**), the weight (or informativeness) of a distribution is more aptly modeled by comparing it to the least informative distribution, the uniform distribution which expresses complete uncertainty. The idea is this: the more informative a distribution, the more it departs from the uniform distribution, the more weight it has, on a scale from 0 to 1. If the drop from uncertainty is complete, the entropy drops to zero, and thus the weight should be 1; if the drop is null, the entropy remains the same, and thus the weight should be zero; if the drop is half, the weight should be .5; and so on for other intermediate cases. This pattern can be captured by the following definition of informativeness (or weight) of a distribution:

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

where P is the probability distribution of interest and $\mathsf{uniform}$ is the baseline uniform distribution.[25]

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

[24] 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.

[25] 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:

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

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.

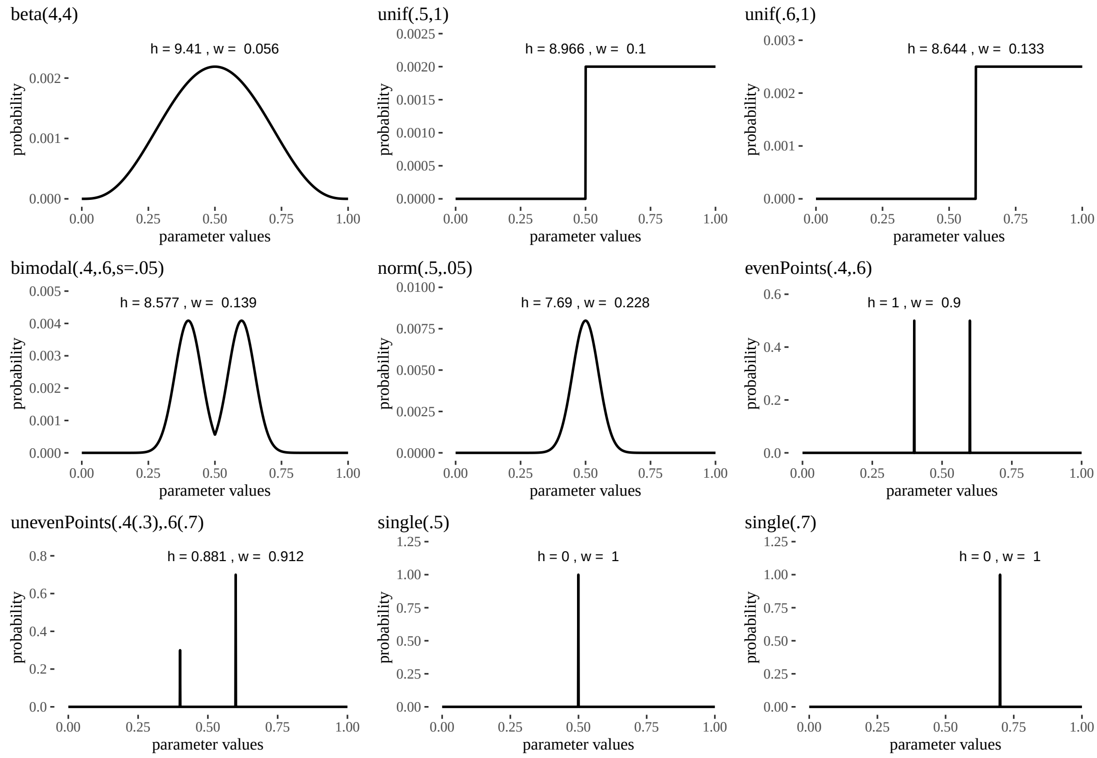

Figure 18: Examples of various distributions with their entropies and weights, ordered by weights. (1) beta(4,4), (2) uniform starting from .5 to 1, (3), uniform strating from .6 to 1, (4) two normal distributions centered around .4 and .6 with standard deviation .05, glued at .5. (5) normal centered 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.

5.5.3 Weight of evidence

Suppose a distribution P depicts what an agent thinks, at some point in time, about the probability of the possible outcomes a random variable H. In this sense, the informativeness (or weight) of a distribution $W(P)$ measures how informed an agent is about H. But it measures the information level of an agent only relative to the state of full uncertainty represented by the uniform distribution. There are two things missing from this account: first, not every agent starts with a uniform prior; second, how informed an agent is must depend on the evidence available to that agent. What we need, then, is an account of how informed agents are *on the basis of the evidence* they have, or in other words, an account of the weight (or informativeness) of the evidence they have.

If the agent starts with a uniform prior over the values of a random variable of interest, $W(P)$ would be a good enough approximation of how informed the evidence made them. In general, however, how much more information is obtained is context-dependent. The weight of evidence, then, must depend on what the agent already knows. Here is a general recipe. In a given context, consider the prior distribution P_0 for the target hypothesis H given what the agent already knows. Then, the agent updates by a body of evidence E. Call this posterior distribution P_E, where the updating is done by standard Bayesian conditionalization. Take the difference between the weight of the prior distribution, $W(P_0)$, and the weight of the posterior distribution, $W(P_E)$. The difference between the two–$W(P_E) - W(P_0)$ or more succinctly ΔW—masures the impact that evidence E has on the information level of the evidence. The difference ΔW, then, is our proposed measure of the weight of the evidence.[26]

More precisely, the calculation follows 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, say the likelihood of the evidence under different possible outcomes.
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.

For some reason you dropped Good, so my comment won't make much sense

Notice that the notion of weight, in principle, is a separate package from our proposal to go second-order. In principle, even if you just consider two hypotheses and plain old likelihoods and likelihood ratios, you could deploy the notion of weight. But this would bring us back to Good's notion of weight, which does not capture the intuitions about weight of evidence that we wanted to capture.

Here comes an explanation about first-order

5.6 Limits of our contribution

I think going through a BN example with Sally Clark with weights is important here, perhaps also with expected values and so on, if you have the time to copy and clean up those bits!

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),

[26] If you prefer to think that weight of evidence should be always positive as a result of adding evidence, you might prefer the absolute value of the difference. We, however, prefer to keep track of whether the evidence makes the agent more or less informed about an issue.

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.

5.7 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).[27] 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?

[27] 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).

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 decision making* (PhD thesis). London School of Economics; Political Science (University of London).

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

Deadman, H. A. (1984a). Fiber evidence and the wayne williams trial (conclusion). *FBI L. Enforcement Bull.*, *53*, 10–19.

Deadman, H. A. (1984b). Fiber evidence and the wayne williams trial (part i). *FBI L. Enforcement Bull.*, *53*, 12–20.

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

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with bayesian networks*. Chapman; Hall.

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.

Kasser, J. (2016). Two conceptions of weight of evidence in peirce's illustrations of the logic of science. *Erkenntnis*, *81*(3), 629–648.

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

Konek, J. (2013). *New foundations for imprecise bayesianism* (PhD thesis). University of Michigan.

Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.

Kyburg Jr, H. E., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press.

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.

Levi, I. (2011). The weight of argument. In *Fundamental uncertainty* (pp. 39–58). Springer.

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

Nance, D. A. (2016). *The burdens of proof: Discriminatory power, weight of evidence, and tenacity of belief*. Cambridge University Press.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., … Rakow,

T. (2006). *Uncertain judgements: Eliciting experts' probabilities*.
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
Skyrms, B. (1977). Resiliency, propensity, and causal necessity. *Journal of Philosophy*, *74*(11), 704–713.
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
Thompson, W. C., Taroni, F., & Aitken, C. G. G. (2003). How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Science*, *48*(1), 47–54.
Titelbaum, M. G. (2020). *Fundamentals of bayesian epistemology*.
Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman; Hall London.