Higher Order Legal Probabilism

An expert testifies that the defendant's genetic profile matches the sample from the crime scene. The match supports the hypothesis that the defendant is the source of the crime sample. But how sure should we be of the truth of this hypothesis? At least two sources of uncertainty should be taken into account: coincidental matches and laboratory errors.[1] They are discussed detail in Chapter XYZ, but a summary is provided below (Section 1).

Coincidental matches and laboratory errors do not exhaust all the sources of uncertainty that are relevant for evaluating match evidence. The risk of a concidental match can be quantified with a number known as 'random match probability'. But this match probability can itself be subject to uncertainty. The same point applies to laboratory errors. If an expert asserts that the laboratory makes false positive identifications only in 0.1% of the cases, this percentage can itself be subject to uncertainty. In what follows, we discuss more systematically these additional sources of uncertainty, what we call *higher-order uncertainty* (Section 2).

How can higher-order uncertainty be combined with sources of first-order uncertainty, for example, coincidental matches and laboratory errors, in evaluating match evidence? Forensic scientists have put forward different proposals for how to model higher-order uncertainty, but these are at best preliminary and disagreement persists (Section 3, Section 4, and Section 5). To make progress, we provide a framework to reason about higher-order uncertainty for the evaluation of match evidence (Section 6). We also show that higher-order uncertainty is not a problem confined to match evidence, but generalizes to other forms of evidence (Section 7). Finally, we apply the formal framework to complex bodies of evidence (Section 8).

1 Coincidental matches and laboratory errors

This section provides a racap about match probabilities and laboratory errors. An expert reports that the defendant matches the crime sample. Besides reporting about the match, the expert also reports the probability of a coincidental match, the probability that a random

[1] Or consider a more complicated case. The expert reports that the defendant matches the crime sample, but the circumstances of the case make it clear that there were two contributors, not just one. Again, the match supports the hypothesis that the defendant was one of them, but what is the likelihood of this hypothesis given the match?

person, unrelated to the crime, would have a matching genotype. Say this probability is 1 in 23 billion, an extremely unlikely event. If the defendant is a match and it is extremely unlikely that a random person would match—so the reasoning goes—the match must be strong evidence that the defendant is the source of the crime sample.

Instead of a match probability, a forensic expert may report a likelihood ratio. In the simplest case, the likelihood ratio compares the probability of seeing a match under the hypothesis that the defendant is the source (usually set to one, the numerator), and the probability of seeing a match under the hypothesis that the defendant is *not* the source (usually set to the match probability, the denominator). In this case, the likelihood ratio equals one over the match probability. If the match probability p is a low number, the likelihood ratio $1/p$ would be a high number. This is as it should be: if the match probability is low, the match must be strong evidece that the defendant is the source.[2]

A randomly selected individual could match, because they happen to have the same genetic profile as the profile associated with the crime scene sample. This type of uncertainty is captured by the match probability. But a randomly selected individual who does not have a matching profile could also match because the *reported* match is a false positive identification.

Think of an inferential chain in two steps. From a match being reported by the laboratory, one infers that there is a true match, namely the defendant and the crime scene sample do actually match. This inference can go wrong if the laboratory makes a false positive identification. The second inferential step goes from the premise of a true match to the conclusion that the matching person, say the defendant, is actually the source of the crime scene traces. This inference can go wrong if the defendant just happens to share—by coincidence—the same genetic profile as that associated with the crime scene. In the language of Bayesian networks, these inferential steps can be represented with three nodes:

$$S \rightarrow M_t \rightarrow M_r,$$

where M_t stands for true match, M_r for reported match, and S for source hypothesis. So, a reported match M_r can be incorrect in two cases: either when the defendant is not the source but happens to genetically match the crime sample ($M_t \wedge \neg S$); or when the defendant is not the source and does not match the crime sample ($\neg M_t \wedge \neg S$).[3]

[2]Consider now the more complex case: the defendant matches the genetic material recovered at the crime scene, but there were two contributors. The question is whether the defendant is one of them. Here, only using the match probability would give an incomplete assessment of the value of the match. The likelihood ratio helps to model this scenario by comparing the following two probabilities: the probability of the match assuming that the defendant and an unknown person are the source (numerator) and the probability of the match assuming that two unknown people are the source (denominator). Evett (1987) shows (see his paper for details) that the likelihood ratio should be $1/2p$ where p is the defendant's genotype probability. Interestingly, the likelihood ratio does not depend on the genotype probability associated with the second person. **REFERENCE:** Evett (1987) On Meaningful Questions: A Two-Trace Transfer Problem, *Journal of the Forensic Science Society*, 27(6): 375–381.

[3]More formally, the (reported) match probability $P(M_r|\neg S)$ can be broken down, as follows:

Without taking laboratory errors into account, the reported match probability is simply equated to the genotype probability, the probability that a randomly selected person would have the matching genotype. In fact, match probability and genotype probability are sometimes used interchangeably. But the two concepts are distinct and they come apart most clearly after taking laboratory errors into account. The reported match probability equals a more complicated formula that contains the genotype probability as one of its inputs:

> TP probability × genotype probability + FP probability × (1-genotype probability)

It is helpful to plug in some numbers. Suppose the genotyope probability is an extremely small number, and the laboratory true positive (TP) probability is 100%. Then, the reported match probability will mostly depend on the laboratory false positive (FP) probability. If the latter is 0.1%, the reported match probability will also be 0.1% (plus an extremely small number). It does not matter how low the genotype probability could be. So, by taking laboratory errors into account, the reported match probability will usually be much less impressive.[4]

2 Higher-order uncertainty

The expert asserts that the genotype probability is in 1 in 23 billion. What is the basis for this claim? Could this number be wrong? Below we identify different sources of higher-order uncertainty for genotype probabilities: model assumptions, sample (un)representativeness, sample size and sampling variability. Questions of higher-order uncertainty also arise for laboratory error rates. But to keep this section more focused, we discuss them at the end of the chapter.

The role of model assumptions is apparent from how genotype probabilities are computed. The number 1 in 23 billion comes from multiplying the genotype probabilities of however many loci there are in a genetic profile. CODIS currently requires that the match occur at twenty loci. The multiplication of the single-locus genotype probabilities is an application of

---

$$P(M_r|\neg S) = P(M_r|M_t \wedge \neg S) \times P(M_t|\neg S) + P(M_r|\neg M_t \wedge \neg S) \times P(\neg M_t|\neg S).$$

Using the probabilistic independencies in the Bayesian network, the equality simplifies to this:

$$P(M_r|\neg S) = P(M_r|M_t) \times P(M_t|\neg S) + P(M_r|\neg M_t) \times P(\neg M_t|\neg S)$$

[4]The uncertainty associated with laboratory errors can be modeled by the reported match probability, as illustrated above, or using the likelihood ratio. As Steele and Colyvan (2023) note, this uncertainty 'bears directly on the first-order probabilities' (p. 1940). They call the uncertainty due to laboratory errors, meta-uncertainty. Since it can be modelled directly with first-order probability, we prefer to label it first-order uncertainty. We discuss meta-uncertainty (or more generally, higher-order uncertainty) in the next section. (**REFERENCE:** Steele and Colyvan (2023), Meta Uncertainty and the Proof Paradoxes, *Philosophical Studies*, 180, pp. 1927-1950). Other forms of uncertainity—for example, model uncertainty or uncertainty about the representativeness of the data—are not like that. They cannot be modeled by the first-order probability. We discuss them in the next section.

the product rule—that the probability of multiple events is the product of their probabilities—and is allowed by the assumption of probabilistic independence between single-locus genotypes. In turn, each single-locus genotype probability results from multiplying the probabilities of the alleles occuring at the locus. These calculations rest on another assumption, Hardy–Weinberg equilibrium, which states that allele probabilities in a population remain constant from a generation to the next absent specific evolutionary pressures.[5] In the early days of DNA evidence, the assumption of probabilistic independence between single-locus genotypes was fiercely debated. This assumption today is more well-established, although disagreements can still arise about more subtle assumptions.[6]

Consider now the second source of higher-order uncertainty: sample unpresentativeness. The computation of the single-locus genotype probabilities starts with allele probabilities. But where do these probabilities come from? The allele *frequencies* that are observed in a database usually replace the allele probabilities in the calculations. To find out these frequencies, one simply counts how often the alleles of interest occur in a database.[7] One concern here is sample representativeness. A sample to be representative should be a random selection from the relevant population. DNA databases are not random samples. They are built from blood banks, paternity testing databases, data collected at hospitals, etc. These are unlikely to be random samples. The good thing is that the alleles used in DNA typing do not have any specific function, and thus DNA databases are equivalent to random samples. (**But what about racial composition?**)[8]

[5]If different alleles i and j occur at the locus, the match probability of the single-locus genotype is given by the formula $2p_ip_j$, where p_i and p_j are the probabilities of each allele. If two copies of the same allele are at the locus, the formula to use is p^2, where p is the probability of the allele. So, to calculate the match probsbility, or more specifically, 20-loci genotype probability, the model usually takes this form: $kp_i^1p_j^1 \times kp_i^2p_j^2 \times \cdots \times kp_i^{20}p_j^{20}$ where k equals 2 if the alleles at the locus are different and 1 otherwise.

[6]In a recent case, two different softwares, STRMix and TrueAllele, reported different likelihood ratios for the same item of match DNA evidence. One software reported a likelihood ratio of 24 while the other reported a likelihood ratio of over a million. The discrepancy can be attributed to different model assumptions. As Thompson (2023) has documentaed, the differences between the two softwares depended on 'subtle differences in modeling parameters and methods' (p. 1). **REFERENCE:** William C. Thompson (2023), Uncertainty in Probabilistic Genotyping of Law Template Dna: A Case Study Comparing STRMix and TrueAllele, *Journal of Forensic Sciences*.

[7]One might wonder, why not directly count the multi-locus genotype frequencies in the database instead of the allele frequencies? This would be simpler and do away with the multiplication of the allele probabilities and the single-locus genotype probabilities. The problem is that genotypes are exceedingly rare. Precisely this fact makes DNA evidence so powerful. So, databases will hardly ever contain most multi-locus genotypes: most genotypes will have a frequency of zero. Alleles, instead, are more common and thus their frequencies can be counted. There are complications, of course. For example, rare alleles may also occur zero times in a database. It would be incorrect to use a zero frequency for the probability of a rare allele. So the parameters of the model must be adjusted in some cases.

[8]There is good empirical evidence for that. As the National Research Council explains: "The saving point is that the features in which we are interested are believed theoretically and found empirically to be essentially uncorrelated with the means by which samples are chosen. Comparison of estimated profile frequencies from different data sets shows relative insensitivity to the source of the data, as we document later in the chapter." (Chapter 5) **REFRENCE:** National Research Council (US) Committee on DNA Forensic Science: An Update. The Evaluation of Forensic DNA Evidence. Washington (DC): National Academies Press (US);

As sources of meta-uncertainty, model assumptions and sample (un)representativeness are often set aside. Typical assumptions such as Hardy–Weinberg equilibrium are well-tested and DNA databases are considered representative of the population.[9] This conclusion might leave the impression that meta-uncertainty in the evaluation of DNA matches is likely to be a marginal issue. But this impression would be mistaken.

Any reported match probability relies on a database of a certain size. If a database of a different size had been used, the random match probability would have likely been different. For example, studies have shown that for databases of 500 people, certain alleles can have an observed frequency of zero. Instead, for databases of 1,000 individuals, the same allele can have a frequency greater than zero.[10] These discrepancies are not due to lack of representativeness of the data or contested model assumptions. And, in fact, the variability is not confined to databases of different size. If another database of the same size had been used, the random match probability would have likely been different, merely as a result of sampling variability.

The effects of sample size and sampling variability are not technical quibbles. A good defense lawyer can bring up these complications during cross-examination rather directly:

> As a DNA expert, you asserted that my client is a match and that the chance of a random match is extremely low, 1 in 23 million, a very unlikely, almost impossible event. At face value, the match seems strong evidence against my client. But how strong is the evidence really? What if you used a different database? Would that number be higher, say in 1 in 1,000? Could it be 1 in 10?

What should the expert respond? The literature contains a variety of different suggestions. We now consider three approaches: intervals, single number probabilities and distributions.

1996. 5, Statistical Issues.

[9]The question of which is the relevant population remains, though. Genotype probabilities can differ significantly across races, even though geographical subgroups within the same race do not show significant differences. If the race of the perpetrator is unknown and the suspect is, say, white, what is the relevant population whose allele frequencies should be used to calculate genotype probabilities? Suppose that, for a particular matching profile, the genotype probability is 1 in 25,000 in the Hispanis population, 1 in 2,500,000 for the Black population and 1 in 10,000,000 for the White population. Which genotype probability should the expert use in testifying in court? Which one is the relevant popoulation? The answer: all three are relevant, so long as the perpetrator could be an individual of any of these three races. **REFERENCE:** David Kaye (2004), Logical Relevance: Problems with the Reference Population and DNA Mixtures in People v. Pizarro David Kaye 3 Law, Probability and Risk 211.**Mention Charles Berger, model uncertainity, e.g. when the competing hypthesis has to do with a brother or close relative rather than an unrelated person. Not sure this is about meta-uncertainity though, but simply choice of competing hypotheses**

[10]**REFRENCE:** Hyun-Chul Park, Eu-Ree Ahn, and Sang-Cheul Shin, Comparative analysis of allele variation using allele frequencies according to sample size in Korean population, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8478755/

3 Intervals

Instead of a single number—a single match probability or likelihood ratio—the expert can present a plausible *interval* with a lowest and highest value.[11] This idea should be familiar from statistics. Statisticians routinely estimate the value of population parameters, such as population proportions. A sample is used to estimate these parameters. Say we are estimating the population allele proportion. We collect data and arrive at a sample proportion. We then attach a measure of uncertainty to the sample proportion as an estimate for the population proportion, for example, an interval of plausible values. That estimate is more or less uncertain—the interval is larger or smaller—depending on the size of the representative sample.[12] The interval more spread out, the smaller the sample.[13]

There is a complication here, though. Estimating the proportion of a population parameter using confidence intervals is standard statistical practice. There is nothig controversial about that. But, a trial expert is not just offerring an estimate of the *genotype proportion*. Their task is to make an inference about the probability that the defendant is the source of the traces at the crime scene. And, in order to do that, the expert should assess the *genotype probability*, the likelihood that a random person would have the matching genotype.[14] Confidence intervals are typically predicated of estimates of population parameters (such as proportions), but not of probabilities.[15] To apply confidence intervals to genotype probabilities or likelihood ratios would certainly bring us beyod standard statistical practice.

Proponents of confidence intervals face a technical hurdle, as Berger and Slooten (2016) note:

> We invite those that propose to report … an interval to demonstrate how one should update one's prior odds into posterior odds, based on that interval, for the purpose of decision making.

[11] In the same spirit, the expert can also present a *distribution* of values of the match probability or likelihood ratio, with some values more likely than others; more on this later.

[12] For example, suppose an expert asserts they are 95% confident that the true genotype proportion lies between 1 in 12 million (lower bound) and 1 in 8 million (higher bound). This means that, in 95% of cases, the true genotype proportion will lie within that interval. **Formulation of CI, not completely correct. RU Revise.**

[13] How large should a database (our sample) be so that an observed allele frequency can be relied on? Chakraborti (1992) calculated one needs to sample at least 300 individuals to have 95% confidence that the observed allele frequency correspond to the allele proportion in the population. (**Not sure this is completely correct. Need to rephrase.**) This holds only for alleles with at least a 1% proportion ()a sample of 300 individuals will ensure that all common alleles (alleles with frequency greater than 1%) will be represented in a sample with at least 95% confidence" (p. 152). **REFRENCE:** Chakraborti (1992) Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing, *Human Biology*, 64(2), pp. 141-159). The proportion of rare alleles should be estimated using larger databases.

[14] Even if the expert was just providing an estimate of the genotype proportion, the fact-finders will then have to translate that estimate into a probability.

[15] As Berger and Slooten (2016) point out: "limited data and resulting sampling variation relates to the precision of e.g. our estimate of the real and existing proportion of a feature in the population, but not to the precision of an LR [likelihood ratio]" (p. 389) (**REFERENCE:** Charles Berger and Klaas Slooten (2016), The LR does not exist, *Science and Justice*, 56(5), pp. 388-391)

We will refer to this as *Berger-Slooten challenge*. It is difficult to answer. The update can be performed by means of Bayes' theorem via a select value of the likelihood ratio, assuming the likelihood ratio lies within an interval of plasible values. The question, then, is this: if a likelihood ratio comes with an interval, which value of the ratio should be used in the update? A few options are: use the extreme value, either upper or lower bound; use the midvalue; or use the value closer to the neutral point (which would be one for the likelihood ratio). Unfortunately, each of these options lead to incoherent posterior probabilities, biased probabilities in favor or against one of the parties, or overvaluing the evidence.[16] Another option is to perform point-by-point updates via Bayes's theorem for all the values within the interval. This process will result in an interval of posterior probabilities. This strategy resembles what is known in formal epistemology as *imprecise probabilism*. This theory has significant shortcomings. **NEED TO EXPLAIN THIS MORE**

Using just a single number for the genotype probability, not an interval, would eliminate this technical hurdle. At the same time, genotype proportions and genotype probabilities are closely related. If intervals are routinely used to model uncertainty about estimates of proportions, why not use them to model uncertainty about genotype probabilities?

As a preliminary point, a genotype proportion can be viewed in at least two different ways. First, the genotype proportion is the *limiting frequency* of the observed genotype: if we were to sample from a population, the observed relative frequency of the genotype would approximate in the limit the genotype population proportion (Sjerps et al, 2016). Alternatively, think of this quantity as the genotype *objective probability*. Denote the population proportion, limiting frequency or objective genotype probability with parameter θ. The other quantity, the genotype probability, is the probability p that an individual picked at random would have the matching genotype. On the simplest model, the genotype probability p is nothing other than the expected value of the genotype proportion θ based on the available data, that is, $E(\theta)$. This expectation can be obtained by updating θ's prior distribution with sample frequency data and obtain θ's posterior distribution. Integrating over the possible values of θ yields $E(\theta)$. **EXPLAIN THIS MORE PRECISELY**.

So the two notions—genotype proportion and genotype probability—are closely related. If uncertainty applies to one, why not to the other? Suppose an expert testifies about p as the expected value of θ, but not about θ itself (**CITE TARONI HERE**). In this case, important information woud be lost becaue the differences in sample size would be overlooked. This information loss is apparent by comparing these three cases. Case one: expert has profiled everybody in the population and is certain that the genotype proportion is 10%. Case two: expert profiled 1,000 individuals and observed that 100 of them have the genotype. Case three: expert profiled 10 individuals and observed that 1 has the genotype. An expert who only reports about the genotype probability p would make no difference between these three cases and report a genotype probability of 10% in all three. But the three experts are in much different epistemic positions. Clearly, the first expert is the most trustworthy and the third

[16] See **REFERENCE:** Ommen, Saunders, Neumann (2016), An argument against presenting interval quantifications as a surrogate for the value of evidence, *Science and Justice*, 56(5), pp. 383-387.

expert is the least trustworthy at identifying the true population proportion. (Example from **REFERENCE:** Sjerps et al. (2016), Uncertainty and LR: to integrate or not to integrate, that's the question, *Law, Probability and Risk* (2016) 15, pp. 23–29)

The upshot here is that, besides reporting about the genotype probability p (or the relevant likelihood ratio), the expert should convey the uncertainty due to the differences in sample size. The expert could do this: first, inform the fact-finders about the genotype probability p (or use a likelihood ratio); and second, inform the fact-finders about the uncertainty in the estimate of the genotype proportion θ, for example, using intervals.[17] This hybrid approach tracks the difference between larger and smaller samples without conflating probabilities and proportions. In this format of presentation, the expert's estimate of the genotype proportion is uncertain, but not the likelihood ratio or the genotype probability.

But this proposal is not without problems. How should the fact-finders rationally update their beliefs based on the genotype probability *together with* the interval associated with the expert's estimate about the genotype proportion? How are the two pieces of information to be combined in updating one's beliefs? Probabilities should be updated in the standard way using Bayes's theorem, but it is unclear how the interval should be integrated with probabilities. Berger-Slooten challenge stands. And even if we did away with confidence interval altoghether and simply presented the facf-findere with information about the sample size, this would not eliminate the problem. (**THIS SEEMS TARONI VIEW**) How are the fact-finders supposed to pur toghether the match probabilkity and the additional information about sample size?

4 Single number probabilities

If intervals are not going to work, another route suggests itself: revert back to single number probabilities. Recall the lawyer's challenge to the expert. What would have happeded if a different database had been used? The resulting match probability could have been different, perhaps much greater. There is no denying that, and intervals aim to capture that variability. But why is this a relevant question to start with? Probabilities should reflect the state of uncertainty about a proposition in light of the evidence available, not what would have happened had one considered other evidence. As Brenner (2010) puts it,

> It is not relevant what our probability estimate would be if we had different population data than we have … any more than makes sense to speculate about alleles in loci that have not been tested. (**REFERENCE:** Brenner 2010, The Fundamental Problem of Forensic Mathematics)

This is a familiar point often made by Bayesians. Probabilities reflect the uncertainty associated with the current evidence, not hypothetical evidence.

[17] Sjerps et al. (2016) write: "we can estimate θ by e.g. its mean $E(\theta)$ … as with any estimate we think it is also important to provide a measure of uncertainty in the form of an interval, e.g. a 95% credible interval" (p. 25).

Consider an analogy. You toss a coin 1,000 times and obtain roughly half heads and half tails. This evidence favors the hypothesis that the coin is fair. Still, had you tossed the coin 5,000 times—or tossed it again 1,000 times—the frequency of heads and tails could have been different. You could have obtained mostly heads, or mostly tails. Still, these unrealized possibilities are irrelevant to the question of what your current evidence says.[18]

Using a single number probability will eliminate the techinical complications with intervals in updating prior beliefs in light of new evidence. Berger-Slooten challenge would not be an issue. At the same time, single number probabilities risk obliterating differences in sample size and sampling variability. These are not differences that pertain to hypothetical evidence, but the current evidence. As seen earlier, a larger sample—compared to a smaller sample—provides stronger evidence for assessing the match probability even though $E(\theta)$ can be the same in both cases. Can this difference be modeled using single number probabilities?

Following Berger and Slooten (2016), let X be a random variable for the population genotype proportion. If we know for sure the population proportion, the likelihood ratio is simply $\frac{p}{p^2} = \frac{1}{p}$, where p is the known population proportion. If the population proportion is not known for sure, the likelihood ratio can be formulated, as follows (derivation on p. 390 of their paper):

$$LR = \frac{1}{E[X] + \frac{VAR[X]}{E[X]}}$$

(**REFERENCE:** Charles Berger and Klaas Slooten (2016), The LR does not exist, *Science and Justice*, 56(5), pp. 388-391)

The greater the variance, with fixed $E[X]$, the lower the likelihood ratio. So, the likelihood ratio decreases as the variance increases, say, due to a smaller database. The likelihood ratio is greatest when the variance is zero, so it simply equals $1/p$.

The upshot is that the likelihood ratio—as Berger and Slooten (2016) aptly put it—is 'not a reflection of the rarity of the characteristic, but of *our knowledge* of that rarity' (p. 390).

OBJECTION: IS THIS A BAD IDEA? EXPLAIN. IF THE EXPECTATION of THETA IS THE SAME FOR SMALL AND LARGE SAMPLES BUT THE MATCH PROBABILITY IS DIFERENT (BECUAUSE OF TEH EFFECTS OF VARIANCE), THEN IS THE MATCH PROBABILITY BASDED ON VARIANCE STILL COHERENT? HOW CAN SOME CLAIM THAT THE MATCH

[18] Another example can solidify the same point. A witness testifies against the defendant. They claim they saw the defendant near the crime scene at the relevant time. This is no doubt incriminating evidence. Now, had a different witness been questioned at trial, they could have testified differently. They could have provided exculpatory information. But so what? It does not follow that in evaluating the testimony of *this* witness, the fact-finders should consider what another witness could have said. It is the evidence that was *actually* presented at trial that should be assessed, not hypothetical evidence that was not presented.}

PRBABILITY MUST BE THE SAME WHILE THEY CLAIM IT IS NOT THE SAME? WHICH ONE IS IT?

MANY ARGUE THE FOLLOWING: IF YOU ARE BETTING ON THE MATCH, YOUR BETTING QUOTIENT SHOULD FOLLOW EXPECTATION(THETA), WHETHER THE SAMPLE IS BIG OR SMALL. SAME WITH BETTING ON COINS. IF THIS IS RIGHT A PROBABILITY THAT CHANGES BY SAMPLE VARIABILITY WOULD BE INCOHERENT. IS THIS CORRECT?

POSSIBLE DILEMMA: IF SINGLE NUMBER PROBABILITY TAKE INTO ACCOUNT SAMPLE VARIABILITY VIA VAR(X), THEN THEY BECOME INCOHERENT (SEE ARGUMENT ABOVE). IF SINGLE NUMBER PROBABILITIES DO NOT TAKE INTO ACCOUNT SAMPLE VARIABILITY, THEY ARE COHERENT BUT FAIL TO MODEL IMPORTANT UNCERTAINTY-RELEVANT INFORMATION SUCH AS SAMPLE VARIABILITY.

5 Distributions

Using intervals raises a number of problems that have not been satisfactorily addressed in the literature. Berger-Slooten challenge remains unaddressed. Reverting to single number probabilities is not without difficulties either. We consider now a third opption: distributions over genotype probabilities or likelihood ratios.

At the technical level, prior probabilities can be updated using a distribution of likelihood ratios and obtaining as a result a distribution of posterior probabilities. (**REFERENCE:** Morrison and Enzinger (2016), What should a forensic practitioner's likelihood ratio be?*Science and Justice*, 56(5), pp. 374-379). So, Berger-Slooten challenge, at least for the case of one single item of evidence, can be addressed. As we will show, the formal machinery of Bayesian networks can be used to extent the theory beyond just updating beliefs with single items of evidence. It is possible to update across multiple items of evidence using distributions over probabilities or likelihood ratios (Section 8).

It is not not clear how these distribution should be interpreted, however.

Some believe that likelihood ratios have a true value. So our estimate of them can be more or less precise. Those who subscribe to this view treat likelihood ratios just like population parameters to be estimated. Likelihood ratio are seen as a function of the genotype population proportion. So if it makes sense to to be uncertain about the population proportion, it must also make sense to be uncertain about the likelihood ratio (which is a function of the population proportion). The use of distributions is quite natural in this setting. Distributions convey the degree of uncertainty, based on our evidence, about the true value of the likelihood ratio (Morrison and Enzinger, 2016).

Others are hesistant to treat likelihood ratios as parameters. They are ratios of probabilities, and probabilities have no true value. Probabilities reflect our uncertainty about propositions of interest. Distributions can be seen as a measure of the robustness of one's assessment of the likelihood ratio, but not of one's uncertainty about the likelihood ratio (Taylor, Hicks and Champod, 2016).

Recall again the lawyers's challenge to the expert. What if a database of a different size had been used? Or what if a different database of the same size had been used? Could that number have been different? How different? Running a simulation with different databases of different sizes allows us to plot a distribution of possible likelihood ratios (or genotype probabilities). (**EXPLAIN HOW THIS WORK AND SHOW SIMULATION**) If the resulting distribution is sharply concentrated around 100, the 100 likelihood ratio should be considered more robust than if the distribution were spread out. Robustness, then, is a measure of how one's current probability assessment is likely to change in light of further evidence that one might come accross. The more robust a probability, the less likely that it would change in light of other information.

On this interpretation, a distribution of likelihood ratios is not a measure of how precise a given likelihood ratio is. A proposed likelihood ratio is not more precise if the spread of the distribution is narrow and the likelihood ratio sits at the center of the distribition, compred to a case in which the distribuition is more spread out. As Taylor, Hicks and Champod (2016), who favor this approach, underscore:

> there is no "true" value for a LR so this distribution does not show the uncertainty of the LR, but how robust (or sensitive) it is depending on the data used. Each data point on the distribution only represents a LR computed under the conditioning of different data.

(**REFERENCE:** Taylor, Hicks and Champod (2016), Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios, *Science and Justice*, 56(5), pp. 402-410)

6 The formal framework

Let us take stock. We have looked at three ways to model meta-uncertainty for evaluating match evidence: using intervals, single number probabilities and distribitions. From a technical point of view, intervals seem ill-suited. They can give rise to incoherent probabilities and other problems. Distributions of likelihood ratios are more promising. They can address Bergen-Sloten chellenge. Alternatively, the variance due to sample variability can be embedded in the likelihoodratio itself. Greater variance will usually mean a lower likelihood ratio. But this single number probability approach also runs in technical difficulties. **EXPLAIN**

We will start by presenting the formal framework. The discussion will be limited to technical aspects. Later we will offer a few thoughts on how the formal framework can be interpreted. The debate in the literatyre is partly conceptual and partly technical. The point we want to demonstrate here is mostly technical: by using distribution of likelihood ratios and distributions of probabilities, prior probabilities can be updated into posterior probabilities. This can be done coherently. The update process can become complicated in dealing with more than one item of evidence. So we will go over a few examples involving two items of evidence, converging and diverging.

6.1 Combining items of evidence

This subsection shows that various combinations of items of evidence exists, not all can be handled with intuitively:

- two items of evidence (equal strength), conflicting, LR=3 for one and LR=1/3 for the other. Overll effect: null. Intuitive.
- two items of evifdence (equal strength), converging, LR=3 for both. Combined LR is 6. Intuitive.
- two items of evidence (equal average strength, but one is less higher-order uncertain than the other), say LR=3 for both LR's but higher-order is lower for one LR. How to combine them? No so intuitive. (This is the like the example in the paper, dog fur etc.)
- two items of evidence, (different average strength), one LR=3 and the other LR=1/2. Conflicting. Seems like overall strength is still LR=3/2, just by multiplication. Is this right? But what if LR=3 is more higher-order uncertain than LR=1/2? Say 1/2 is very concentrated (very stable, very certain) and instead LR=3 very spread out (very unstable, very uncertain). Then, could LR=1/2 ultimately win out LR=3? If so, the combination of the two LR's would less then one (say close 1/2), not greater than one (not close to 2/3). Which one is it? These are radically different assessments. Not so intuitive.

Question: Suppose we assess LR1 of E1 and LR2 of E2, along with respective higher-order uncertainities. Procedure 1: We integrate out and the LR as point averages. Then, we combine LR's and get a new overall LR. Procedure 2: What if instead we do this? We Take LR1 and LR2 and keep their respective higher-uncertainties. Next, we combine them LR's and their higher-order uncertainities. Then, we integrate out and obtain a combined overall LR. Is this the same as by the other procedure?

7 Generalizing

There can be meta-uncertainty about numerical information used to assess the value of evidence, in case of genetic matches, statistical evidence and eyewitness evidence. So meta-uncertainty is not a minor issue, but a general one.

7.1 Labopratory Error rates

ADD EXAMPLE ABOUT UNCERTAINITY OF THE LABORATORY ERROR RATES ALSO DEPEDENT ON SAMPLE VARIABILITY AND SAMPLE REPRESENTATIVENESS, NOT SURE ANY MODEL ASSUMPTIONS ARE INVOLVED HERE

IT WOULD BE INTERESTING TO HAVE AN EXAMPLE IN WHICH WE SHOW HOW TO EVALUATE MATCH EVIDENCE BY TAKING INTO ACCOUNT THESE FOUR SOURCES OF UNCERTAINTY: 1. RANDOM MATCHES PROBABILITY, 2. LABORATORY ERROR PROBABILITY, 3. SAMPLING VARIABILITY FOR RANDOM MATCH PROBABILITY, 4. SAMPLING VARIABILITY FOR LABORATORY ERROR PROBABILITY.

7.2 Statistics in the Shonubi

7.3 Eyewitness Evidence

8 Higher-order Bayesian Networks