Marcello's notes on Weight

9/1/2022

1 Purpose of this document

This contain some thoughts about weight, to be merged with the main chapter/paper on weight. I am keeping my notes here to avoid conflicts in updating the main file.

2 Structure and contributions of the chapter/paper

Here is a possible structure that the chapter/paper on weight might take:

(1) Motivate the distinction between balance and weight. Examine different conceptions of weight: quantity, completeness, resilience, informativeness (=Rafal's account), what else (precision, specificity, intervals). Provide a characterization of each, show connections between the different conceptions.

(2) Illustrate how the different conceptions of weight play a role in trial decision-making broadly construed (not just standards of proof, but also questions of admissibility, retrial, appeallate decision, etc.) Give at least one illustration each.

R: challenge: mix applicability in sson, without abandoning care about formal details

(3) Show why weighty evidence (under one or more of the conceptions of weight above) improves accuracy. Note that accuracy here is understood as the distance of one's credence (or assigned probability) from the true value. Brier score is the most common measure of accuracy. Accuracy in this sense should be distinguished from accuracy in decision-making (say in terms of rates of false positive and false negative decisions). A separate question is whether decision-making informed by weighty evidence promotes the accuracy of decisions or not.

Additional topics (not to be covered in this chapter/paper):

(4) Discuss theories of trial-decision making that make sense of the different conception of weight. Engage with existing theories: Nance model, Kaye model, Dahlman model. Are we proposing a different model of weight-based trial decision-making or we are improving on existing models?

(5) Show why trial decision-making that takes weight into account promote accuracy (here accuracy is the accuracy of decision, not accuracy as distance from true value like a Brier score).

3 (1) Different conceptions

3.1 Quantity

How much evidence you have. This satisfies monotonicity and is mostly what Kaynes had in might.

Hard to make it formally precise.

- Attempts to count items of evidence are not promising. They are purely syntactical. They cannot justify why, intuitively, several shaky testimonies should be *less weighty* than fewer solid testimonies.

R; use the example in the criticism of counting

- One account is comparative. Take a body of evidence B and add an extra item of evidence, obtaining B+. Then, clearly B+ has more quantity of evidence than B. Comparison can only be made when one body of evidence is a subset of the other.

R: In particular, can't help understand which way to branch if evidence is costly and we need to choose or prioritize

3.1.1 Absolute value of log of LR

Another idea is to take directional measures of the value of evidence, say the likelihood ratio, and make them adirectional. See Nance's book "The Burdens of Proof" section 3.5. The quantification of Keynesian Weight. So the weight of a body of evidence E relative to a claim H would be

$$|\ln(LR_H(E))|,$$

where $LR_H(E) = \frac{P(E|H)}{P(E|\neg H)}$. Nance comments (p. 135):

> This version also strips the evidence of its directionality by treating all natural log likelihoods for evidence as positively signed. And because $ln(1/x) = -ln(x)$ two items of equivalent weight according to the preceding criterion will retain that equivalence under the present criterion. (In terms of the preceding example, $|ln(1/3)| = |ln(3)| = 1.61$.)

3.1.2 Nance's problem of decomposition

Nance does not believe (see again section 3.5) that a quantification of weight in this sense (quantity of evidence) is workable. One key problem is decomposition:

> if one tries to compute the weight of a mass of evidence by decomposing it into smaller pieces and adding up the weights of the pieces, the resulting sum will depend on how the decomposition is performed, and the sum of the weights of the pieces will, in general, not equal the weight of the total mass. (p. 153)

This is interesting, need to compare it to Good's, which is additive, and I don't even know what Nance means by decomposition and why he thinks it should be unique, does he give an example anywhere?

For the point about **decomposition**, see footnote p. 135, page 153. Take two items of evidence such as $LR(E1) = 2$ and $LR(E2) = 0.1$, but $LR(E2|E1) = 0.5$. The weight of the conjunction $E1 \wedge E2$ is zero, since $|ln(LR(E1 \wedge E2))| = |ln(LR(E1)LR(E2|E1))| = |ln(2.0 * 0.5)| = |ln(1)| = 0$. Instead, the weight of $E1$ is $|ln(LR(E1)) = |ln(2.0)|$ and the weight of $E2$ is $|ln(LR(E2|E1)) = |ln(0.5)|$. Adding these two weights does not give rise to zero.

This suggest this *decomposition principle* (also called additvity, see Rafal's discussion of Good):

$$W(A \wedge B) = W(A) + W(B|A),$$

where $W(B|A)$ is the weight of b given A. Presumably the weight of B alone need not be the same as the weight of B given A.

The following *commutative principle* should also hold:

$$W(A \wedge B) = W(A) + W(B|A) = W(B) + (A|B) = W(B \wedge A)$$

Question: Does Rafal's informativeness conception satisfy these principles?

3.1.3 Nance: weight of overall evidence available versus future evidence

However, despite the problem of decomposition (or additivity) Nance does not think we need a measure of total quantity of the evidence currently available. He thinks that the only question relevant for legal decision-making is whether more evidence can be acquired at reasonable costs. So he focuses on the weight of evidence that can be acquired at the margins rather than on the weight of evidence already acquired. For him, the weight of evidence that can be acquired at the margins can be computed using the absolute value of the log of the likelihood, as defined above. No decomposition problem arises here since we are evaluating one single items of evidence one by one. He writes:

> After all, the question of weight comes up only in connection with the issue of whether and how, at the margin, to augment the evidence on hand. Consequently, it is not important how one determines the unique sequence that decomposes the evidence that one already has. With respect to potential evidence, one also does not need to decompose any con- templated item of evidence into smaller "chunks" at the acquisition stage. The circumstances of its acquisition will determine the items of evidence acquired. Thus one only needs to compare at the margin what various items of evidence might contribute (p. 155, book).

Why TF spend so much time writing about weight of evidence then?? How to you even know if the cost of obtaining an item of evidence is worth it if you can't evaluate it's potential impact/weight?

So, for Nance, the question is only whether the weight of the potential evidence is cost-justified, in light of (a) the ratio $W(E)/K(E)$ being above/below a threshold ratio of weight $W(E)$ to cost of

acquisition $K(E)$ and (b) what other potential evidence could be acquired and at what costs. He think there should be a prioritization ordering of acquisition of potential evidence (p. 155, book). See also discussion later under completeness.

3.1.4 Complication regardig future evidence: expected versus actual weight

Future evidence—evidence that one could potentially acquire but one has not acquired—does not have a definite value yet. The test could be positive or negative. You do not know that before doing the test. So the weight of future evidence should be a weighted average—that is, the *expected weight of the evidence*, not the weight of an actual piece of evidence.

This point turns into an objection against Nance. If he is measuring weight (of potential evidence) using the log of the LR, this does not seem correct, as this would only be a measure of actual weight, not expected weight. The latter is more appropriate in the case of potential evidence. But Nance seems to think that the content of potential evidence can be "roughly anticipated":

> The task is complicated by the difficulty of assessing, however approximately, the weight of evidence that has not yet been obtained and so the content of which can only be roughly anticipated. (book, p. 156)

It is not clear whether Nance has any account of expected weight. Perhpas this account is not even needed. Suppose we are considering taking a DNA test. The test could be positive or negative. But, presumably, the weight of the rest (for either result) should be the same since we are taking the absolute of the log of the LR. The same applies to a potential eyewitness testimony. And so on. The key here seems to be that weight is unidirectional and thus it does not mater what the actual value of the evidence is. Whether the test is positive or negative, it will have the same weight. This allows to bypass the problem of expected weight of evidence.

Question: Is it the case, under Rafal's account of weight, that the weight of a DNA test, positive or negative—or the weight of whatever other evidence that can take a positive or negative value—is the same in either direction?

3.2 Completeness

3.2.1 Kaye: List of all items of evidence: script-narrative conception

Presumably there is a list of *all* items of evidence one would expect in a case, and thus completeness measures the gap between that list and the evidence actually presented.

This list can be drawn in a number of different ways:

(a) using a *script* (=this is the type of case in which one would normally expect to see these kinds of evidence);
(b) using a case-specific *shared narrative* (=the following facts are well-established, and given those facts, we would normally expect to see these items of evidence);
(c) using a *partisan narrative* (=the prosecutor offers this narrative from which we would normally expect to see these items of evidence); or
(d) some combination of these.

These options need not be exclusive We can view them as describing a natural progression and refinement. At first, the script identifies a generic complete list of items of evidence. Then, this generic list of complete items of evidence is further refined once we learn more specific details about the case at hand and settle on a detailed story about what happened.

- The **narrative-center conception of completeness** can be found in David Kaye 1986 piece "Do We Need a Calculus of Weight to Understand Proof Beyond a Reasonable Doubt?"

For him a story S makes a prediction about what the items of evidence that should exist and should not exist. Difference between these prediction and the evidence actually available form gaps, call them G. These gaps are part of the evidence also. They describe facts about what evidence was not found or not presented. So, broadly speaking, the evidence in a case included both E (evidence presented) and G (gaps, evidence not presented). For Kaye, the probability to be assessed is not $P(S|E)$, but rather, $P(S|E \wedge G)$.

Often $P(S|E) > P(S|E \wedge G)$, depending on the reason why the evidence is missing.

Yeah, I've read it. My worry: this only works if you build in the identification and gaps brute force. Nothing in this overly simplistic machinery tells you how to identify them and what their impact should be. We need to do better.

3.2.2 Nance: reasonable completeness - cost benefit analysis of completeness

- Nance has a notion of completeness in mind when he talks about reasonable completeness. See his 1998 piece "Evidential Completeness and the Burden of Proof." He seems to have in mind a **cost-benefit analysis of completeness** more than a reasonable list of complete evidence.

Evidence is, for him (pp. 627-628, article), reasonably complete (relative to a claim H), when

(a) it is not missing any relevant evidence that it would be cost justified (=reasonable) to obtain, and

(b) it is missing a relevant item of evidence that—though currently impossible to obtain because of a fault (negligence, intentional destruction, etc.) of one of the parties—could have been obtained with no unreasonable costs.

- Because of condition (b), the same body of evidence could count as complete or incomplete deepening on the circumstances. For example, if DNA evidence is missing in a rape case, but there is a good justification for why it is missing (say, the lab samples were destroyed in an accident), the body of evidence missing DNA evidence could very well count as reasonably complete. But if there is no good justification for why the DNA evidence is missing, the same body of evidence could count as incomplete.
- Focusing on (a), Nance does not speak of a list of *reasonably complete* evidence. Nance in his book seems to think that evidence is complete if no more evidence can be acquired in a cost justified manner. He writes:

  > the decision whether Keynesian weight is adequate to proceed to the determination about the underlying hypothesis or claim does not depend on any such quantitative measure of either the weight of the evidence on hand or its degree of completeness. It depends, instead, on what further weight can be obtained and at what costs. (p. 156, book)

He illustrates this point by proving a list of possible further items of evidence, each associated with a weight to costs ratio (where the weight is presumably understood as quantity of evidence). He writes:

> This creates a prioritization for the acquisition of additional evidence as long as the ratio remains above some point of insufficient return on investigative cost. (p. 155, book)

So, it seems, once the return on investigative cost is below a threshold ratio, there is no point in searching for and presenting further evidence, and thus the body of evidence available must be considered complete.

3.2.3 Difficulties with Nance's resonable completeness

- **Rafal's comment**: Ok, I don't get this fully. So fine, reasonable completeness is supposed to be a necessary condition for making a decision, whatever the decision will be. But clearly we need to be able to distinguish cases where no further evidence can feasibly be obtained, the posterior probability is centered around a high value but uncertainty is still high, and so no conviction should be made (see some examples from my higher-order probability paper/Girona talk). So we still need some explanation of when such feasibility-restricted collection of evidence is sufficient, and some quantitative take on this, right? Am I missing something? It's not sufficient either, if the evidence is overwhelming you can make a decision, even if there is some evidence you could easily obtained, but what you have is already resilient.

- Reasonable completeness isn't sufficient (for making a decision, one way or another): one might have gathered all the evidence that is practically available, but the evidence could still have very low weight, be quite shaky, etc. (e.g. the distribution associated with the evidence is spread out). So reasonable completeness does not seem a good proxy for sufficient weight as a condition for making a decision.
- Reasonable completeness isn't necessary (for making a decision): body of evidence could be reasonably incomplete and yet potential further evidence could make no difference to the probability assessment, if the evidence is very resilient. So a body of evidence could be reasonably incomplete and still have very good weight and be good enough for a decision.
- UPSHOT: reasonable completeness is a good heuristic and might be the best we have as a proxy for weight in some cases, but doe not capture conceptually what we mean by weight.

3.2.4 Open questions

- An open question about completeness is, how do we measure the gap between the complete list of evidence and the actual evidence available? Counting the number of items of evidence missing does not seem a promising approach. It is merely syntactic. We seem to run into the same problems as with measuring quantity.
- **Rafal's response**: Ok, how about the complete list is an idiom-insipred-narration-developed wishlist. Then on one hand you need to pay attention to what you're missing, what the potential impact of it would be (my expected weight seems to do the job), and on the other, how telling the evidence that you already have you are. But also, you need to be able to distinguish various reason for incompleteness —- not only cost-related, because it's a mixed bag. If the evidence has been purpsefully destroyed, the impact is different than if it has been destroyed by negligence, and thinking about this in terms of costs is useless, as both are cost-wise degenerate cases (you can't obtain these items of evidence at any cost)
- Connection with quantity: is a body of evidence that has more quantity also more evidentially complete?

I think this question is too vague, I don't get it.

3.3 Resilience

The evidence currently available might support a certain claim to some degree, say the LR or the posterior probability (or some other measures of balance) tips strongly in favor of the claim. But, would the balance (posterior probability, LR) change in light of new evidence that might be presented? If there would be no change (within some set boundaries perhaps), the evidence counts as resilient. If there would be change, the evidence counts as not resilient.

- Skyrms' approach in his 1977 piece "Resiliency, Propensities, and Causal Necessity:

  Def. Resiliency for Conditional Probabilities: The resiliency of $P(H|E) = a$ is 1 - the maximum $|Pr(H|E\&E') - a|$ (so-called wiggle), for any E', where E' is another statement in the language that entails neither $E \rightarrow H$ now $E \rightarrow \neg H$.
- Need to restrict the resilience test to a *reasonable set* of possible further items of evidence, otherwise no body of evidence could ever count as resilient. This problem was identified by Skyrms already in his 1977 piece:

  Resiliency over the whole language may be a requirement of unrealistic stringency. There is no unique answer as to which sublanguage resiliency must be evaluated over, for lawlikeness. Rather, we must again say that the larger the sublanguage over which we have high instantial resiliency, the more lawlike the statistical law. At one end of the scale we have statements like "the probability of death within a year given that one is an American male of age 65 = d," which is extremely sensitive to auxiliary information, and whose resiliency is limited indeed. At the other end we have laws of radioactive decay, which have been tested under an enormous variety of circumstances and whose resiliency extends over a language of impressive scope. (Skyrm , Resiliency, Propensities, and Causal Necessity, p. 708).

Right, and it is not insane to think that this list of potential pieces of further evidence that the decision should be ideally resilient to comes from the wishlist of complete evidence, no?

3.4 Informativeness (Rafal's account)

The notion of weight/informativeness (in Rafal's formal account) is related but not identical to quantity, completeness, resilience. Weight refers to:

(a) the weight of a distribution (=how informative the distribution is relative to the uniform distribution, which is by definition uninformative and thus has zero weight);

(b) the weight of evidence, called wDelta (=the difference between the weight of the prior distribution and the weight of the posterior distribution).

Questions: How does this model of weight look in trial proceedings? How can it be applied? How would this conception of weight enter into a theory of decision-making? Do we identify a weight threshold, like a probability threshold?

Question: How does this conception relate to Joyce measure of weight for the case of chance hypotheses. Let $P(X)\sum_x P(Ch(X) = x)x$, where X is a statement such as "the coin comes up heads" and $Ch(X) = x$ stands for "The chance that the coin come sup heads is x." Weight is defined as (see Joyce, How Probabilities Reflect Evidence, p. 166):

$$w(X,E) = \sum_x |P(Ch(X) = x)(x - P(X))^2 - P(Ch(X) = x|E)(x - Pr(X|E))^2|$$

All valid questions. I propose we get started with going over the example in the Girona talk to identify which questions are unanswered, which need more work, and what the weak spots are.

3.5 What else?

There could be other notions that might be good to discuss, such as: *specificity* (how specific is the evidence), *intervals and imprecision*, etc. Need to state connection to other conceptions of weight.

Like I said, I did think about this, this is briefly discussed in the HOP paper, but I agree this needs to be more extensively discussed in the chapter.

4 DNA Evidence Evaluation

We usually assess DNA matches using likelihood ratios (LRs). But there may be uncertainty about the random match probability (RMP) used to estimate the LR. See the case by Dhalman of Missing Finger below or multiple reference classes available to estimate RMP. What to do about this uncertainty? Below I identify a few key challenges that a theory of weight might help to address.

Yeh, imprecision is going to be a big thing in the chapter, as this is one of the modern strategies to make sense of Keynes. Not sure what to do with the notion of specificity, as this smells of Thompson and the blue bus problem, I don't think the notion has ever been successfully expicated. To be discussed.

4.1 Taroni Sjerps dispute: Multiple levels of uncertainty

There is a dispute in the literature between single value reporting (Biederman, Taroni, Bozza, Colin Aitken, Charles Berger, Duncan Taylor, Tacha Hicks, Chirstopher Champond) and critics of single values reporting (Marjan Sjerps, Kristy Martire, Gary Edmonf, Geoffrey Morrison).

See 2016 LPR "Uncertainty and LR: to integrate or not to integrate, that's the question" by M. J. Sjerps, I. Alberink, A. Bolck, R. D. Stoel, P. Vergeer, J. H. van Zanten. This is a response to a piece by the Taroni group.

The crux of the matter is the RMP: calculating the probability γ that an individual drawn at random has a specific DNA profile. The probability γ is "the population limiting frequency θ of the observed genetic profile in the population" (p. 25). Even though Taroni does like the language of "estimating" a probability, Sjerps points out that θ (to which γ is identified) is indeed estimated and has a true value.

How should one report the probability γ or RMP?

> (Option 1: Taroni) "This probability can be obtained by integrating over all possible values of θ, weighting with its density according to probability calculus, and obtain the (posterior) expected value of θ, $E(\theta)$" (p. 25)
>
> (Option 2: Sjerps) "θ is a population limiting frequency that has an unknown 'true' value, and there is uncertainty attached to it in the form of a known probability density .... Hence, we can estimate θ by e.g. its mean $E(\theta)$ and report this to the legal justice system. However, 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

Sjerps motivates her approach with this example:

> Suppose that we have a very partial DNA profile, and compare three experts:
> (1) Expert 1 has profiled all individuals in the population, he is certain that θ is 10%.

(2) Expert 2 assumes, based on the very limited number of loci, a $Be(1,9)$ prior and profiles a thousand individuals: one hundred of them have the profile.

(3) Expert 3 also assumes a $Be(1,9)$ prior and profiles 10 individuals: one of them has the profile.

Now when all experts go to court, they will all report exactly the same when they follow Taroni's advice: report only $E(\theta)$. In this example, they will all report a probability of 10

Seems clear that Sjerps is right Maybe I'll ask Alex Biederman what's wrong with that!

The general question here is whether things like the likelihood ratio—and only that—can capture all the different levels of uncertainty involved in the assessment of a DNA match or whether a more sophisticated approach, say using first-order probability *plus* weight, is more appropriate. It seems that the Taroni/Biederman group would be opposed to this latter approach.

Another useful paper on this debate is J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (2002) 29–37.

Rafal's comment: If weight was something external; if it actually falls out from otherwise probabilistic phenomena standard in Bayesian data analysis, maybe not. But it will be good to discuss this potential source of disagreement: hey, you might think this is crazy and external, but actually it's natural once you agree the HOP approach makes sense and is in line with the usual methods of Bayesian data analysis. }

4.2 DNA evidence and the brother problem

There is uncertainty about whether the defendant has a brother or a twin. If the defendant has a brother or a twin, the uncertainty about the RMP increases. The RMP might have to be lower than it actually is, or alternatively, the greater uncertainty can be captured by a greater spread of the distribution. This again seems to fit with Rafal's approach.

Here too it'd be great to have an example precise enough for an implementation to be possible. Do you recall seeing one in the literature?

Some papers about this issue:

- Relatedness and DNA: are we taking it seriously enough? John Buckleton,Christopher M. Triggs.

  It is necessary to consider by what criterion we may decide whether or not to include the estimated match probability for a sibling in addition to that for an unrelated person. Clearly, including an additional number makes an already complex evidential statement more so. There would need to be a good reason to include an additional number in a statement. It would seem reasonable to include this number if it was "important" for the decision making process. (p. 117)

- Evett, Evaluating DNA profiles in a case where the defence is "It was my brother." This paper gives a fairly detailed and reliastic scenaruio that may occur in practise:

  In this case the defendant has a number of brothers. There is no particular item of evidence to implicate any one of them, neither have they been eliminated from suspicion. No samples have been taken from any of them. The proposal that a brother left the stain is put forward by defence counsel as a potential explanation for the scientist to consider. This is the scenario which is most likely to occur. It will be most helpful to work it through in fairly general terms so the numerical values used in the present scenario are replaced for the time being with symbols. (p. 12)

Here we are going to have one prosecution hypothesis (C: the defendant is the source) and two defense hypotheses ($\overline{C}_1$ an unrelated persons is the source and $\overline{C}_2$ a brother is the source). Evett notes:

> In general, however, the expert is going to be in a difficult position. In the simpler kind of case where there are only two alternatives, the likelihood ratio gives a measure of evidence strength which is quite independent of the prior odds. In the present kind of situation, however, the import of the scientific evidence cannot be detached neatly from the other evidence. Like it or not, to assess his evidence it is necessary for the scientist to gain some idea of how the Court is thinking in relation to $\overline{C}_1$ and $\overline{C}_2$. The difficulties of doing this should not be underestimated.

So the problme here is that since the denominator of the likelihood has two hypotheses, then their prior probabilities do matter and affect the value of the likelihood ratio. But we might uncertain about what these prior probabilities should, or we have a range of values, or we thin some are more plausible then others. Any any rate. . . .**question**: To resolve this problem, wouldn't it be a good idea to take into a

second order probability about the possible priors of the alternative hypotheses? Rafal's approach to weight seems to fit well here, no?

4.3 DNA evidence assessment and multiple reference classes

There could be multiple random match probabilities available (to due multiple reference classes) but we do not know which should be applied. Simply taking the random match probability that is most favorable to the defendant would not be too one-sided. Dahlman Swedish missing finger case (see below in later sections) falls into this type of cases.

OK, so purely combinatorial RMP probabilities have been criticized for being unrealistic, either because they don't take into account lab errors, or because they do not easily entail differences in frequencies between races (we need to read up on the details here if we are to use this example, but I remember an example with multiple possible classes in an early version of the SEP entry). Need to think about this.

It is not clear whether the problem of multiple reference classes can be addressed via weight or needs something else, but it is a pressing problem for DNA evidence assessment (or any other evidence), as pointed in Allen and Pardo 2007 piece "The Problematic Value of Mathematical Models of Evidence":

> Any attempt to mathematically model the value of evidence, however, must somehow try to isolate an item of evidence's probability for establishing a particular conclusion. Generating these probabilities will, in turn, involve isolating characteristics about the evidence, the event, and the relationship among those characteristics. This relationship may be established either by objectively known base rates or through subjective assessments. In either case, the modeled values arise through abstracting from the specific evidence and event under discussion and placing various aspects of each within particular classes, with varying frequencies, propensities, or subjective probabilities instantiated by the various characteristics on which one has chosen to attend (p. 114)

In our SEP entry, we have a good discussion of this problem applied to DNA evidence evaluation (citing Allen and Pardo 2007 piece). But we do not offer a solution, only pose the problem:

> It is tempting to dismiss this challenge by noting that expert witnesses work with multiple reference classes and entertain plausible ranges of values (Nance 2007). In fact, relying on multiple reference classes is customary in the assessment of DNA evidence. In Darling v. State, 808 So. 2d 145 (Fla. 2002), for example, a Polish woman living in Orlando was sexually assaulted and killed. The DNA expert testified about multiple random match probabilities, relying on frequencies about African-Americans, Caucasians and Southeastern Hispanics from the Miami area. Since the perpetrator could have belonged to any of these ethnic groups, the groups considered by the expert were all relevant under different scenarios of what could have happened.
>
> Unlike expert witnesses, appellate courts often prefer that only one reference class be considered. In another case, Michael Pizarro, who matched the crime traces, was convicted of raping and suffocating his 13-year-old-half-sister (People v. Pizarro, 110 Cal.App.4th 530, 2003). The FBI analyst testified at trial that the likelihood of finding another unrelated Hispanic individual with the same genetic profile was approximately 1/250,000. Since the race of the perpetrator was not known, Pizarro appealed arguing that the DNA evidence was inadmissible. The appellate court sided with Pizarro and objected to the presentation of frequency estimates for the Hispanic population as well as frequencies for any other racial or ethnic groups. The court wrote:
>
> "It does not matter how many Hispanics, Caucasians, Blacks, or Native Americans resemble the perpetrator if the perpetrator is actually Asian"
>
> The uneasiness that appellate courts display when expert witnesses testify about multiple references classes is understandable. Perhaps, the reference class most favorable to the defendant should be selected, giving the accused the benefit of the doubt. This might be appropriate in some cases. But suppose the random match probability associated with a DNA match is 1 in 10 for people in group A, while it is 1 in 100 million for people in group B. Always going for the reference class that is most favorable to the defendant will in some cases weaken the incriminating force of DNA matches more than necessary.

4.4 DNA evidence and levels of propositions

One controversial issue is how to use DNA match evidence to decide about activity level propositions rather than mere source level propositions. Say, if the defendant is the source of the crime scene DNA, how likely is the defendant to have touched, pushed, stabbed the victim, or whatever? Data about transfer seems less solid then data about random match probabilities. A 2016 article by Biederman and others ("Evaluation of Forensic DNA Traces When Propositions of Interest Relate to Activities: Analysis and Discussion of Recurrent Concerns" *https://doi.org/10.3389/fgene.2016.00215*) says:

> So, when a scientist is faced with assigning a probability for finding trace material given the proposition of handling an object by a person of interest (e.g., the activity of discharging a firearm), we do see no harm in referring to studies that have focused on rates of transfer not exactly the same in the alleged circumstances of the case. Although some features of the individual case at hand may differ, nothing will prevent the scientist from also judging that some additional case-tailored experiments should be conducted in order to extend their knowledge and understanding, but case backlogs and limited resources may render this difficult.

use as motivation for HOP, analogy to Bayesian statistical reasoning with weak prior information.

The paucity of data available or the fact that the data available does not match perfectly the circumstances of the specific case at hand suggest that the probability assignment should be less weighty – say the spread of the distribution should be greater. This fits with Rafal's account of weight.

Question: Can Rafal's account of weight also apply to the assessment of source level propositions about which data are usually available and less controversial? There can be circumstances in which Rafal's weight might still be a useful notion. For example:

Neat, it'd be great to have a worked-out example with an implementation. Can we find a precise enough example to work with in the literature?

5 Legal cases and illustrations

5.1 Example: Salem Trial

Where does this come from, has this been discussed in the literature? Oh wait, is that the one I brought up? :)

Aggravated murder case. Victim was stabbed to death in her house. Defendant is convicted. He appeals. Oregon App Ct first grants reconsideration by post-conviction court. Then, post-conviction court rejects defendant's arguments. Defendats appeals again. Oregon App Ct disagrees with post-conviction court and agrees with defendant. See Jesse Lee Johnson v. Jeff Premo 2021 Oregon Appellate case.

Link to decision: *https://law.justia.com/cases/oregon/court-of-appeals/2021/a159635.html*

Evidence against defendant:

- fingerprint match in victim's home
- cigarette butt in victim's home genetic match
- statement by informant that defendant said "offed the bitch to rob her"
- jewelry found with defendant matched victim's
- blood boot prints in victim's house match defedant's boot prints

Exculpating evidence:

- negative genetic match between weapon and defendant
- negative genetic match between blood in victim's home and defendant
- defendant's boots did not test positive for blood

Missing evidence (not presented at trial, fault of counsel and police):

- testimony by women living across the street. She saw white man enter victim's house first, then loud noises and screams followed and finally he run away. Later she saw black man enter the house and then leave. Defendant is black.

Trial:

- Defendant is convicted

Post conviction court:

- counsel was at fault for not presenting neighbor testimony
- still, no prejudice occurred since the testimony corroborates other evidence which was presented

Appellate court:

- counsel was at fault, agrees with post-conviction coirt on this point
- but, contra post-conviction court, prejudice did occur since the testimony of the neighbor can be the basis of a solid defense case (=white man killed woman before defendant came in: black man came in afterwards)

Comment: Note that both resilience and completeness play role here:

- (*completeness test*) first ask whether evidence is missing. Note that police have a justification for why the evidence is missing, but it does not seem a very good one. So the evidence is reasonably incomplete.
- (*resilience test*) second assess whether missing evidence could have changed the verdict. In legal jargon, this is the question of prejudice.

This is a great example for various reasons. One: completeness wrt. to list again seems to be a heuristics: coming up with an ideal list to be used (i) to guide evidence collection and interpretation and to (ii) estimate resiliency wrt. to potential impact of the missing items (although, to what extend the witness' testimony could be predicted and its impact evaluated?). To discuss

5.2 Example: Tin Box Case

Missing fingerprint evidence in a murder case due to police oversight. This is from Dahlman's paper circulated during the Girona workshop. Full quotation below:

> In 2005 a man walks into a Swedish police station and says that he wants to turn himself in. His name is AA and he says that he has just killed an elderly woman who lives by herself in an apartment nearby. The police rush to the apartment and find the woman stabbed to death. In his confession AA explains to the police that he had heard that the woman kept a huge amount of cash in a tin box, and had knocked on her door and tricked her to let him in by pretending to work for the local church. He says that his plan was to distract the woman and quickly grab some money from the tin box, but she caught him in the act, and he panicked and stabbed her. At the police station AA pulls out a switch blade knife from his pocket and puts in on the table. The knife is smeared in blood, and is sent to forensics, who quickly confirm that the blood belongs to the victim. The autopsy report is consistent with AAs confession. The angle of the stab wound suggests that the perpetrator is above medium height, which is somewhat odd since AA is shorter than medium, but can be explained if AA held the knife high. AA is prosecuted for murder. At the trial, AAs defense attorney says that he suspects that his client is giving a false confession to cover for someone else. AA has no criminal record, but he has two sons who both have previous convictions for burglary and assault. Both sons are above medium height, and are known to carry switch blade knives. AA insists that he did it. He claims that his sons have nothing to do with the murder, and gives the court a detailed and vivid story of how he committed it, that fits perfectly with the crime scene. During the trial the court learns that the tin box was found open at the crime scene, but was never examined for fingerprints or DNA. Apparently, the police did not consider this necessary, since AA had confessed and the victim's blood had been found on his knife. When the defense attorney tried to have the tin box examined for fingerprints or DNA, it was too late. The box had been wiped clean from the victim's blood, which had removed all potential traces from the perpetrator. The defense attorney argues that the police investigators committed a huge blunder when they missed to look for forensic traces on the tin box, since the results of this investigation could have produced evidence favorable to the defendant. If fingerprints or DNA from one of AAs sons had been found on the box, AAs confession would have been falsified. AA is acquitted. The court says in its verdict that the police should have examined the tin box for fingerprints or DNA, and explains that the evidence against AA would have been sufficient for a conviction if the tin box had been properly examined and this had not produced any evidence against the prosecution's case, but since this examination is now missing from the investigation, the evidence against AA is not sufficiently robust for the standard of proof in criminal cases.

Question: do we want to spend time and effort developing a BN HOP treatment here? Worry: this might be too complex to clearly illustrate the basic points. Counter: so perhaps later in a chapter as a proof of concept?

I don't think I've seen this paper, would you mind forwarding it to me? Also, a good example of how difficult a practical resilience evaluation is, and how prior weight estimation would help

Comment: Here again completeness and resilience are key.

- Given what we know about what happened, fingerprint evidence from the tin box is missing ("the tin box was found open at the crime scene, but was never examined for fingerprints or DNA"). As in the Salem case, police have a justification ("police did not consider this necessary, since AA had confessed"), but it does not seem a very good one. So the evidence is reasonably incomplete.

Wait, what happened with the sons??

- Next, we have a resilience test, suggesting that fingerprint evidence could weaken the case against

Note, both cases failed not because the evidence was intentionally destroyed, but due to negligence.

Do we also have a neat example of evidence being intentionally destroyed and this actually having an impact

the defendant ("the results of this investigation could have produced evidence favorable to the defendant").

5.3 Example: Missing Fingers

Reference class used to compute random match probability of missing finger is too generic. A difference reference class might yield a different random match probability. This is also from Dahlman's paper:

> In 2013 a beheading video is spread on the internet. The video is made with a smartphone in Syria and shows in graphic detail how a British journalist is decapitated by ISIS. The face of the executor is masked but his hands are visible and two fingers are partly missing on his right hand. A couple of months later the Swedish police receives an anonymous tip from a woman who has seen the beheading video on the internet and says that she recognizes the hand. She believes that the executor is BB, a man of Syrian origin living in Sweden. The police investigate BB and find that he made two trips from Sweden to Syria in 2013 to support ISIS in its cause. BB admits that he has participated in ISIS activities in Syria, but denies that he is the executioner in the beheading video, and claims that he has never killed anyone. A forensic image analyst compares the hand in the video with BBs hand, and report that they match. The missing fingers are severed in the same places. To assess the probability of a random match, the forensic analyst consults reference data on the prevalence of missing fingers. Searching a data base with 20 000 pictures of hands collected from the general Swedish population the forensic analyst finds 20 hands (1 in 1000) with severed fingers. At closer scrutiny, two of them (1 in 10 000) are severed in the same way as the hand in the beheading video, and match it just as well as BBs hand. The forensic analyst therefore concludes that the probability of a random match is approximately 1 in 10 000. The two matching hands in the reference data base belong to men who died before 2013 and can therefore be ruled out as suspects. BB is prosecuted for murdering the British journalist. The case for the prosecution is based on BBs affiliation with ISIS and the expert testimony of the forensic image analyst. BBs defense attorney argues that the random match probability assigned by the forensic expert is too small, since it is based on the prevalence of missing fingers in the general Swedish population and it is reasonable to assume that such injuries are more common among men that are affiliated with ISIS. In the cross-examination of the forensic expert, the defense attorney asks if it is possible that the gathering of further reference data about people affiliated with ISIS could have shown that missing fingers are considerably more frequent in this reference class, for example that 1 in 1000 rather than 1 in 10 000 are disfigured in this way. The forensic expert replies that this possibility cannot be ruled out. BB is acquitted. The court says in the verdict that the prosecution should have backed their case with better reference data. The court explains that a random match probability of 1 in 10 000 would have been sufficient for proof beyond reasonable doubt, given the other circumstances of the case, if this probability had been robust, but in the absence of more reference data on people affiliated with ISIS it is not sufficiently robust for the standard of proof in criminal cases.

Whoa, great example to talk about why HOP would do much better at gauging uncertainty here!!

Comments:

- Completeness and resilience still play a role here, but not so clearly as in the previous two examples.

resilience, reference class: two different ways to talk about it

- In addition to completeness and resilience, this example best illustrates Rafal's model of weight/informativeness of evidence, since the evidence here is quantitative. The distribution of the random match probability could be more or less spread out depending on the data relied upon.

Nah, I'd say, two things, poor choice of reference class, and despite low frequency high uncertainty based on sample size. But I still need to think about this.

- The notion of specificity seems applicable as well, since the random match probability can be more or less specific to an individual, deepening on the specificity of the reference class used.

Ah, yeah, you already thought about the point I made. :)

5.4 Example: Howard - No. 18-CF-157 District of Columbia Court of Appeals

Link to case: *https://www.casemine.com/judgement/us/5fbb6f1d4653d07a51f93921*

Maybe, but then I still don't think I have a good enough understanding of the notion and perhaps adding it to the mix in this chapter might be too much. To be discussed.

Unlawful possession of firearm. Police stop defendant's vehicle and find illegal firearm. Police does not retain all the items found in a backpack in the car. Defendant request missing evidence instructions at trial. Request is denied. Defendant appeals. Appeal court rejects because the items in the backpack did not seem at all relevant in understanding what happened.

Comments:

- The evidence was in some broad sense incomplete (items in backpack were not presented as evidence), but since they were clearly irrelevant or would have made no difference to the case, there is no need to request missing evidence instructions.
- Here again e see an appeal to completeness as well as to resilience (missing evidence would have made no difference to the outcome since the missing evidence seems utterly irrelevant)
- This case opens up a discussion about various legal rules, such as missing evidence and spoliation

Ok, interesting, was there a deeper story explaining how there was no resilience and how the content of the backpack would explain something or undermine the decision? Otherwise a nice short example of a positive resiliency check, to be used in the chapater.

5.5 Example: Missing evidence instrcutions

Below are some examples, not intended to exhaust the complexity of the legal doctrine on these matters.

> In order for a party to be entitled to a missing witness instruction, the court must first determine that the requesting party has satisfied two conditions: 1) that the witness be "peculiarly available" to the party against whom the inference is sought to be made, and 2) that the witness' testimony would be likely to elucidate the transaction at issue. See Graves v. United States, 150 U.S. 118, 121, 14 S.Ct. 40, 41, 37 L.Ed. 1021 (1893); Thomas v. United States, 447 A.2d 52, 57 (D.C. 1982); Cooper v. United States, 415 A.2d 528, 533 (D.C. 1980).
>
> Miles v. United States, 483 A.2d 649 (D.C. 1984)
>
> There are two prerequisites to the giving of a missing witness instruction. First, the witness' testimony must be likely to elucidate the transaction at issue. Second, the absent witness must be peculiarly available to the party against whom the adverse inference is sought to be drawn. Miles v. United States, 483 A.2d 649, 657 (D.C. 1984) (citing cases).
>
> Hinnant v. United States, 520 A.2d 292 (D.C. 1987)
>
> Hinnant v. United States, 520 A.2d 292, 295 (D.C. 1987). The party seeking a missing evidence instruction must make a twofold showing. First, the evidence "must be likely to elucidate the transaction at issue"; second, it "must be peculiarly available to the party against whom the adverse inference is sought to be drawn. Id. at 294.
>
> Moreover, we have "recognized several dangers inherent in the use of a missing [evidence] instruction," Dent v. United States, 404 A.2d 165, 171 (D.C. 1979), since it "represents a radical departure" from the principle that the jury should decide the case by evaluating the evidence before it. See Thomas v. United States, 447 A.2d 52, 58 (D.C. 1982).
>
> Tyer v. United States, 912 A.2d 1150 (D.C. 2006)

Ok, but are there any clear cases of the instruction being given and of the instruction being denied.

5.6 Example: Spoliation evidence rules

- Here is a very early formulation, back in 1893:

> It was said by Chief Justice Shaw in the case of the Commonwealth v. Webster, 5 Cush. 295, 316: "But when pretty stringent proof of circumstances is produced tending to support the charge, and it is apparent that the accused is so situated that he can offer evidence of all the facts and circumstances as they existed, and show, if such was the truth, that the suspicious circumstances can be accounted for consistently with his innocence, and he fails to offer such proof, the natural conclusion is that the proof, if produced, instead of rebutting, would tend to support the charge." The rule even in criminal cases is that if a party has it peculiarly within his power to produce witnesses whose testimony would elucidate the transaction, the fact that he does not do it creates the presumption that the testimony, if produced, would be unfavorable. 1 Starkie on Evidence, 54; People v. Hovey, 92 N.Y. 554, 559; Mercer v. State, 17 Tex. App. 452[17 Tex.Crim. 452], 467; Gordon v. People, 33 N.Y. 501, 508.
>
> Graves v. United States, 150 U.S. 118, 14 S. Ct. 40, 37 L. Ed. 1021 (1893).

Ok, need to use it when we talk about spoilation having diferent impact than negligence or than cost-consideration.

- The more recent doctrine seems more complicated, more the general idea seems the same, with some qualifications:

  The doctrine of what has been termed spoliation of evidence includes two sub-categories of behavior: the deliberate destruction of evidence and the simple failure to preserve evidence. It is well settled that a party's bad faith destruction of a document relevant to proof of an issue at trial gives rise to a strong inference that production of the document would have been unfavorable to the party responsible for its destruction.

  The prevailing rule is that, to justify the inference, "the circumstances of the [destruction] must manifest bad faith. Mere negligence is not enough, for it does not sustain the inference of consciousness of a weak case."

  When the loss or destruction of evidence is not intentional or reckless, by contrast, the issue is not strictly "spoliation" but rather a failure to preserve evidence. The rule that a fact-finder may draw an inference adverse to a party who fails to preserve relevant evidence within his exclusive control is well established in this jurisdiction ... Like the spoliation rule, it derives from the common sense notion that if the evidence was favorable to the non-producing party's case, it would have taken pains to preserve and come forward with it.

  Battocchi v. Washington Hosp. Center, 581 A.2d 759 (1990)

6 (3) Weight and accuracy (Brier score)

NOTHING TO SAY FOR NOW

7 (4) Weight-sensitive models of trial decision-making

7.1 Nance model

Nance model is complicated, but a simplified version should go something like this:

- First, the judge (not the jury!) should ask whether the evidence is reasonably complete (see definition of reasonable completeness above). If yes, then the jury can assess balance (say using posterior probability). If not, then the judge should make some preemptive decision (direct verdict, dismissal, etc.).
- In criminal cases, evidence that is reasonably incomplete will always end up against the prosecutor, so should result in a diorect verdict against the prosecution.
- In civil cases, matters are more complicated. The party that is at fault for the evidential incompleteness and that has better access to the missing evidence shoud be penalized.

7.2 Two part model: : completeness plus resilience

Nance seems also to subscribe to a two part model, including completeness and resilience.

- First, ask whether the evidence is reasonably complete. If yes, then assess balance. If not, before taking a preemptive decision against one party or the other, assess resilience.
- Second, relative to the reasonably missing evidence, test whether your current evidence is resilient. That is, test the resilience of your evidence (relative to a claim H) only against the evidence that is missing relative to the reasonably complete list. The problem is that you do not know the value of that evidence (e.g. you do not know if a missing DNA test will be positive or negative), so it might sometimes be appropriate to assume the worst case scenario but this will depend case-by-case. If current evidence is resilient, then assess balance (say posterior probability). If it is not resilient (=balance goes below the required standard), then you cannot assess balance and must reconsider (direct verdict, dismissal or more investigation might be necessary).

This two part model is followed by trial courts, as well as post trial courts and post conviction appellate judgments. See earlier cases.

7.3 David Kaye model: gaps

Kaye agrees trial decision-making should be sensitive to gaps in the evidence. Instead of the standard,

$$P(S_p|E) > t,$$

he proposes the following revised threshold model:

$$P(S_p|E \wedge G) > t$$

The idea is to see whether the total evidence presented, E, as well as gaps G in the evidence, support the prosecutor's story S_p to the required threshold probability t.

Interestingly, G is part of the evidence together with evidence proper E. After all, absence of evidence is itself a fact and thus evidence in a broad sense.

Question: How do we come up with G? Any story, if true, induces a list of evidence we would expect. Whatver the difference between that list and the evidence actually presented E is the missing evidence G. This goes back to the assessment of reasonable completeness of the evidence.

Question: How do we assess $P(S_p|E \wedge G)$ and not just $P(S_p|E)$? What kind of evidentiary contribution does G provide?

In general, it seems that $P(S_p|E \wedge \neg G) > P(S_p|E \wedge G)$, or in other words, the presence of gaps should reduce the probability of S_p, other things being equal. So Kaye is advocating for *discounting* by the trier of facts. Nance opposes this idea.

Question: Kayes model follows a two part approach: first, assess completeness, and second, assess how gaps affect balance (say posterior probability). How does Kaye model compare to the two part model consisting of completeness plus resilience? Are they equivalent? What are the differences?

7.4 Dahlman model: information economics

TO BE COPMPLETED

8 (4) Weight and Accuracy in Trial Decision-making

It is instructive to examine each conception of weight (quantity, completeness, resilience, weight/informativeness) and see if trial decision-making is guided by weight (in addition to balance), it can better promote accuracy.

Quantity:

Recall the simple comparative definition of quantity: if B and $B+$ are bodies of evidence and B is a subset of $B+$, then $B+$ has more quantity of evidence than B.

The question about *accuracy* can be put this way. Are decisions based on B+ as opposed to B, in the long run, always more accurate (=they yield fewer false positive and fewer false negatives) *all else being equal*?

Answering this question is by no means obvious and already requires setting a rather complex set of formal definitions. A computer simulation might be a good way to address this question.

Completeness:

Resilience:

Weight/informativeness: