Recent methodological advances in Bayes factors for use in forensic analysis and reporting

Presentation to the 11th International Conference on Forensic Inference and Statistics in Lund, Skåne County, Sweden

Dan Spitzner, June 14, 2023

1. Greeting. Good afternoon, and thank you for attending this session. It is an honor to be presenting at this conference. Today I will be discussing two forensics-related articles that I recently published in statistics journals. The topic of the first article is evidence reporting using a specific strategy that I call pool reduction. The topic of the second article is robust Bayes factors.

2. Pool reduction: debates. I will spend most of my time discussing the first article, since the second article is more technical in nature.

I start by outlining some of the relevant debates in forensics around evidence reporting. One such debate focuses on two misunderstandings of probability, one known as the *prosecutor's fallacy*, which is most closely connected to the pool-reduction strategy; the other is known as the and *defense fallacy*. The prosecutor's fallacy "confuses the probability of finding the evidence on an innocent person with the probability that a person on whom the evidence is found is innocent." We will explore this idea more closely in a moment. Conversely, the defense fallacy is committed when the defense "uses the probability value for an evidence match... to argue that for a *large enough* population... their client is only *one of many people* that could be guilty, and is thus innocent due to reasonable doubt." Each fallacy insufficiently acknowledges the relevance of prior probabilities in inferential thinking.

As a hypothetical example of the prosecutor's fallacy, consider the following example offered in a 2018 paper by Bill Thompson. Suppose a DNA profile, which is part of a forensic investigation, is such that it "would be found in only one person in 1 million in the general population." A relevant question is whether "one-in-a-million" is rare enough to conclude that the victim is the source of the material. Consider instead, however, that "in a nation as large as the United States there are likely to be over 300 people who share the one-in-a-million DNA profile." Thus, even when DNA evidence alone drastically reduces a large pool of potential sources, hundreds of plausible sources remain. To reduce the pool further—down to the victim—requires other evidence.

As an example of the prosecutor's fallacy playing out in the courtroom, consider the *Doheny* case, which was argued before the England & Wales Court of Appeals in 1996. This case involved a semen stain found on the clothing of a woman who had been raped. As part of their instructions to the jury, the judge conveyed that the conclusions of forensic analysis imply "that there are probably only four or five white males in the United Kingdom from whom that semen could have come," and that "the defendant is one of them." The jurors' task is to decide, based on "all the evidence," whether "it was the defendant who left that stain or whether it is possible that it was one of the other small group of men who share the same DNA characteristics." The *Doheny* example offers not only an articulation of the prosecutor's fallacy, but an example of a court offering guidance in the same terms that underlie the pool-reduction strategy.

A basic outline of the strategy is as follows: state the extent to which forensic material reduces an initial pool of plausible sources to a smaller pool, but not to a particular source.

The latter goal to reduce a pool to a particular source reflects a theory of *individualization*, where "a person or thing is specifically distinguished from all other persons or things of the same kind." Since its popularization by Paul Kirk in 1963, individualization has been severely questioned by the forensic science community, and many doubt that it rests on a sound basis. Pool reduction is closer instead to the related concept of *classification*, wherein "we set out with a goal to individualize," but "fail to narrow the source item to a category of one."

The main objective of my efforts on this topic is to firm-up pool-reduction's statistical foundation. My proposal for doing so is to establish that pool-reduction is a non-standard interpretation of a Bayes factor.

Let us now turn to some experimental results on evidence-reporting. The table presented in the slide lists a number of available strategies for presentation of evidence to a jury. The *random match probability* is a "numerical estimate of the probability that a 'match' or 'nonexclusion' would occur by coincidence." The *likelihood ratio* is the ratio of the probability the observed forensic data is obtained under the same-source hypothesis to that under the different-source hypothesis. The *likelihood ratio verbal equivalent* is where the likelihood ratio is computed, but its result is conveyed using words that describes strength-of-evidence, such as "strong," or "moderately strong." The *random match equivalent* is where a likelihood ratio result is conveyed as an equivalent frequency, as if a random match probability.

In a study published in 2015, Thompson and Newman examine the first three strategies in an experiment designed to assess perceptions of statistics used to present forensic evidence. Their study finds a complex effect of presentation format on participants' sensitivity to strength of evidence. Comparing contexts involving DNA and shoeprint evidence, the authors generally observe the expected effect in the DNA context, but not in the shoeprint context, unless the evidence is presented in the RMP format. As explanation, the authors speculate that a statement made in the RMP format is perceived as "more scientific, or at least more discriminating," while one made in the LR or VE format may seem like "a conclusion without evidence." They furthermore speculate that these differences do not arise in the DNA context "because DNA is already perceived as highly scientific."

The random match equivalent is not investigated in Thompson and Newman's experiment, but is put forward in another article by Thompson with the suggestion that "although the RME could be criticized as artificial" it would convey evidence "more effectively than likelihood ratios." That is, it, too, may be perceived as more discriminating. I suggest that the pool-reduction strategy would do the same, and would be less subject to criticisms of artificiality than a random match equivalent because of its connection to Bayes factors.

In addition to considering the pool-reduction strategy specifically within the forensic context, the present discussion also resonates with extensive debate about statistical reporting that produced a 2016 statement by the *American Statistical Association* (ASA) on p-values, a 2019 special issue in *The American Statistician* on related issues such as statistical significance and reproducibility, and a 2021 statement by the ASA clarifying the 2016 statement. A substantial portion of that discussion ponders alternative concepts to statistical significance and issues related to the communication of results, all of which motivates a secondary aim of my efforts, which is to strengthen pool reduction as a strategy for forensic evidence-reporting by also establishing it as a strategy for evidence-reporting in general.

3. Pool reduction: proposal. To begin our careful exploration of these ideas, it will benefit discussion to adopt a simplified context of blood-type analysis. Let us consider a toy example, along with the table of hypothetical blood-type frequencies that appears on the slide. The example is as follows:

> *The remains of a murder victim is found in a forest. A suspect is found to have blood stains on his shirt. How strong is the evidence that the blood on the suspect's shirt is that of the victim?*

Suppose the blood of the victim and the blood on the suspect's shirt are both of the same type. We can see from the tabulated blood-type percentages that if the shared blood-type is AB, which is somewhat rare, this should raise some eyebrows. The evidence for a shared source is at least worth noting, or at any rate it is stronger then if the shared blood-type is O, which is relatively common. Suppose the shared blood-type *is* AB. Reflecting the 6.2% prevalence of this blood type, the report produced by the pool-reduction strategy might read as follows:

> *Based on a comparison of forensic material, a pool of 1000 plausible sources would be reduced to 62, leaving 61 that remain to possibly exclude by other evidence.*

I will now cover some of the very basic mathematics that are involved in making a connection to Bayes factors. Our setup assumes an initial pool of one thousand plausible sources, which includes the victim. In my notation, I use an asterisk to single out that particular source. The assertion of the *prosecutor* is that the victim is the source; it corresponds to a subset of just one member. The assertion of the *defense* is that the victim is not among the plausible sources; its subset has 999 members. We will need to consider one additional assertion, which I call the assertion of *jurisprudence*. This assertion is that one or the other of the assertions of the prosecutor and defense are correct. Its subset is all 1000 plausible sources.

The relevant Bayes factor compares the assertion of the prosecutor *versus* the assertion of jurisprudence. To set up a Bayesian analysis, equal-probability sampling of an assertion's subset is assumed, as is the assumption that the blood-type of a plausible source is measured without error. In the case where the shared blood-type is AB, the Bayes factor works out be the ratio of 1000 to 62, which reflects the statement communicated by pool reduction of an initial pool of 1000 plausible sources being reduced to 62. I have written a formula for the Bayes factor that is likely unfamiliar to you, and requires further explanation. This is the formula that connects pool reduction to Bayes factors.

On this next slide I have written three equivalent formulas for a Bayes factor, the last of which copies from the previous slide. The first formula writes the Bayes factor as a likelihood ratio, which in the Bayesian context is a ratio of *integrated* likelihood functions. The second formula is a widely known expression that reflects the operations of Bayesian updating. Here, the Bayes factor is the factor by which prior odds is multiplied to obtain posterior odds. The third formula, which is our main interest, writes the Bayes factor as the ratio of the *relative* initial pool-size, to the *relative* pool-size after having examined the data. I will talk about relative size *versus* absolute size in just a moment. To reach this interpretation it is necessary to treat an inverse-probability as a pool size, as is reflective of equal-probability sampling; for example, a probability of one-in-a-thousand reflects by inversion equal-probability sampling from a pool of size one thousand. Note, in addition,

this formula makes use of a parameter value that is common to each assertion, which in the forensic example is the victim.

To understand the importance of distinguishing *relative* from *absolute* pool-size, let us consider a second toy example:

> *Two blood stains are found at a crime scene where it is clear there has been a struggle. Were two people injured in the struggle, or just one?*

For this example, the relevant pools are much larger than that those of the previous example, since they are composed of *pairs* of plausible sources, some of which are pairs of duplicate sources. The pairs with duplicate sources comprise the subset associated with the assertion that one person was injured in the struggle. Should the two stains share the blood-type AB, the evidence report might read as follows:

> *For each plausible source of one stain, forensic evidence reduces the pool of 1000 sources of the other stain to 62.*

Observe that this is a *relative* reduction in pool-size, and it reflects the Bayes factor computed for this example. The *absolute* reduction in pool-size is one thousand *squared* to sixty-two *squared*.

Several additional comments are in order. A key issue is that because the mathematics of pool reduction expresses only the *reduction* of uncertainty, additional context is needed to specify the size of the initial pool. In Thompson's DNA example, the initial pool-size is 300 million, the approximate size of a concrete population. Reference to a relevant concrete population is the ideal situation, provided one can be identified. When one cannot, a potentially suitable alternative may be to offer several *versions* of the pool-reduction scenario, each in terms of a distinct population and initial pool-size. For instance, one version would state, "the initial pool of 100,000 county residents would be reduced to 6,200 plausible sources;" another would refer to residents in the five-county area, and state a reduction of 500,000 plausible sources to 31,000.

In our toy examples, the reported initial pool-size is one thousand, which is based on a convention. When the identification of a concrete population is out of reach, the use of conventions may still be a useful option to pursue. The number one thousand for initial pool-size is offered as a tidy option that is consistent with percentage values carried to one decimal point. Variations for other decimal lengths are readily formulated. For situations of drastic reductions, an alternative convention is to set the initial pool-size such that the final pool-size is exactly two. This can avoid references to a fraction of a source. For example, if the Bayes factor is 625, one need not puzzle over the meaning of reducing an initial pool of 1000 to a fractional 1.6 sources. Instead the initial pool-size would be set to 1250, which is reduced to a final pool of size of two. The numerical convention of two and not one as the final pool-size would remind the reader that, even in the face of strong evidence, the decision-making task is not fully informed by examination of empirical data alone.

As for the extension of pool reduction to an evidence-reporting strategy for general use, I will not cover details, but will instead list a few of the more interesting aspects that require careful consideration. Extension to unequal-probability sampling is rather straightforward mathematically, but requires additional interpretation, to cover such situations as, for example, when the plausibility of a source depends on its distance from a crime scene. Extension to the comparisons of non-nested assertions is possible, but more complicated, especially as

regards to conventions that would be used to report evidence. We do learn from such considerations, though, that comparing the assertion of the prosecutor to that of jurisprudence, and not to that of the defense—as is traditional in formulations of this problem—, slightly benefits the defense. This may seem counterintuitive since the victim is removed from the defense's subset. Other aspects that I consider are the possibility of reporting evidence in terms of a pool *expansion*, rather than a reduction, and conventions that might be employed to convey strength-of-evidence through descriptive graphics.

4. Robust Bayes factors: debates. I switch gears now to the second of my recently-published articles, whose topic is robust Bayes factors. It is inspired by two papers that bookend a forty-year timespan of forensics discussion. The first is the seminal article by Dennis Lindley, published in 1977, which introduced Bayes factors to the forensics community. In that article, Lindley describes the source attribution problem mathematically as two-sample testing for equal means under Gaussian samples and Gaussian priors. Forty years since the publication of Lindley's article, and after extensive discussion of his ideas, Steven Lund and Hari Iyer argue against the normative use of Bayes factors in forensic reporting, based in part on their oversensitivity to the prior distribution. Moreover, Lund and Iyer, in alignment with the perspectives of other forensic methodologists, stress the relevance of nonparametric representations of background information. In addition, their arguments make frequent reference to subjective Bayesian foundations. As it turns out, Lund and Iyer's foundational arguments have been effectively refuted by others. Nevertheless, their argumentation remains motivational for deeper exploration of Bayesian robustness in forensics problems, particularly in the context of subjective Bayesian perspectives.

An article by Brunero Liseo, put forward to the statistics community, provides a helpful summary of debate around robust Bayes factors. He describes three broad approaches: *sensitivity analysis*, where the result of a Bayes factor is examined across an entire class of prior distributions; he also describes two versions of *intrinsically robust procedures*, one of which is a default prior, which is a prior that is analytically derived for the problem—examples are the Jeffreys-rule prior and Berger and Pericchi's intrinsic priors—; the other class of intrinsically robust procedures is asymptotic approximations, whose focus is the elements of a posterior formula that are stable within asymptotic analysis—the Schwarz criterion, or BIC, is a well known example; the third approach is *robustifying procedures*, which modify or reinterpret techniques to reduce their sensitivity to certain modeling choices. Anthony O'Hagan's fractional Bayes factor is an example, as is the robust Bayes factor I present to you today.

5. Robust Bayes factors: proposal. The objective of my efforts on this topic is to demonstrate a robustifying procedure for Lindley's problem that is consistent with subjective Bayesian perspectives and accommodating to flexible parametric priors.

What I mean by a flexible parametric prior is one that reflects the breadth of nonparametric representations in terms of the variety of shapes that can be captured. It takes the form of a mixture density, to be applied to each mean of Lindley's two-sample problem, and could possibly depend on nuisance parameters. The class of mixtures with which I work includes component-densities that could be symmetric, taking the form of a Gaussian density, or skewed, being drawn from the location-scale family generated by the chi-square distributions.

For example, suppose we consider the canonical refractive-index-of-glass dataset, which has been examined by an array of authors. Its histogram is copied three times onto the slide. Two large symmetric modes are observed at lower index values, and, directly to the right, a pattern is observed that could be interpreted as in Panel A, whose fitted probability density depicts a skewed mode that is mostly disconnected with the symmetric mode to its left, or as in Panel B, where a single right-skewed mode blends with the symmetric mode, or as in Panel C, which displays a series of smaller, dampening symmetric modes.

The robust Bayes factor that I have been exploring makes novel use of default priors. Specifically, it applies the analyst's favorite default prior toward the purpose of calibrating a Bayes factor. In broad terms, the two steps in achieving the calibration are as follows. First, the default prior is used to set up the equation shown in the slide wherein the corresponding Bayes factor is set to one. The equation is to be solved for the data, and its solution is to be interpreted as *neutral* in the sense that the balance of evidence that it provides favors neither the equal- nor unequal-means assertion. The default prior is then discarded from the formulation, while in the second step the Bayes factor on neutral data that is formulated from the elicited prior is used to calibrate the Bayes factor on the observed data. I call the resulting statistic a neutral-data comparison.

I am skipping over quite a number of details that I attend to carefully in my article. First, formulas for the Bayes factor calculated on neutral data are available in certain broad scenarios, which implies that neutral data typically need not actually be calculated. This drastically simplifies the calibration procedure's implementation. In addition, I explore a scheme for dealing with nuisance parameters by conditioning, which induces compatibility with Markov-chain Monte Carlo computational algorithms.

This next slide shows some of the results of a sensitivity analysis applied to the calibrated Bayes factors that result, in comparison with ordinary Bayes factors. The analysis focuses on the sensitivity of these quantities to the scale parameters of the prior, a characteristic that is known to be particularly problematic. The left panel depicts an overlaid sequence of flexible prior densities such that prior scale is initially set to small values within individual mixture components and then gradually increased to the fitted values that are to be used in the final analysis. Depicted are the results of analyses of multiple data sets, across every value of prior scale. Error variance and sample sizes are held fixed. Appearing in the graph are values of weight-of-evidence, a transformation of a Bayes factor, wherein larger values indicate stronger evidence for the unequal-means assertion. Each data set is formulated such that its two means are centered around the peak of one or the other of the two large symmetric modes at the left side of the fitted density. Between these two modes, corresponding values of weight-of-evidence tended to be very nearly the same, and appear in the graph as close-together curve-pairs. Two different values of the distance between means are examined, producing in the graph a general pattern that is roughly copied and shifted upwards with the larger distance.

The crucial pattern that is to be observed is as follows: In the right panel, the trajectories of weight-of-evidence calculated from ordinary Bayes factors persistently become more negative as prior scale increases, illustrating in detail a known pattern of the Bayes factor's sensitivity to prior scale. In contrast, the trajectories calculated from calibrated Bayes factors become increasingly stable as prior scale increases, a pattern that indicates their robustness.

As a final comment, the use of neutral data to calibrate a Bayes factor may alternatively be understood as calibrating prior odds. This follows by rewriting a well known formula for posterior odds as the product of a Bayes factor and prior odds; it is instead written as the product of a neutral-data comparison and what might be called a *draft value* of prior odds. Solving for prior odds shows that that quantity is a refinement of the draft value, obtained by calibration with respect to the Bayes factor calculated on neutral data.

In a subjective Bayes context, this operation may be interpreted in two ways. The first is where the draft value is obtained within elicitation, and then calibrated afterwards, in which case the framework I am describing would be understood as a robustifying procedure, referring to Liseo's three approaches to robustness. I would like to put forward the second interpretation wherein both the draft value *and* refined value are obtained within elicitation, with the connection between them arising from discussion between the analyst and expert aiming to grasp the sensitivity phenomenon. I offer further perspective on this phenomenon in my article. By this interpretation, the refined prior odds is less a calibration of draft odds, but the result of de-biasing within the process of elicitation. Importantly, should this interpretation be adopted, it would induce wholesale consistency of the robust Bayes factor with subjective Bayesian perspectives.

6. Closing. In closing, I have provided a statistical basis for the pool-reduction strategy of evidence reporting, and taken you through a bare-bones look at a robust Bayes factor that tracks closely to subjective Bayesian viewpoints and is amenable to use with complicated priors. The latter may not offer a complete solution to current-day problems in forensic science, but it does offer a promising concept by which to guide methodology toward complex situations.

Note to the reader: To access references to articles made in the above text, please see the two articles that anchor the discussion:

Spitzner, D. J. (2023a). A statistical basis for reporting strength of evidence as pool reduction. The American Statistician, 77:1, 62-71. DOI: 10.1080/00031305.2022.2026478

Spitzner, D. J. (2023b). Calibrated Bayes factors under flexible priors. Statistical Methods & Applications. DOI: 10.1007/s10260-023-00683-4.