# Value of evidence in the rare type match problem

## Common source versus specific source

Van Dorp, I. N.; Leegwater, A. J.; Alberink, I.; Jongbloed, G.

**DOI**
10.1093/lpr/mgaa002

**Publication date**
2020

**Document Version**
Accepted author manuscript

**Published in**
Law, Probability and Risk

# Value of evidence in the rare type match problem: common source versus specific source

I.N. van Dorp, A.J. Leegwater, I. Alberink, G. Jongbloed

November 11, 2019

## Abstract

In the so-called rare type match problem, the discrete characteristics of a crime stain have not been observed in the set of background material. To assess the strength of evidence, two competing statistical hypotheses need to be considered. The formulation of the hypotheses depends on which identification of source question is of interest (Ommen, 2017). Assuming that the evidence has been generated according to the beta-binomial model, two quantifications of the value of evidence can be found in the literature, but no clear indication is given when to use either of these. When the likelihood ratio is used to quantify the value of evidence, an estimate is needed for the frequency of the discrete characteristics. The central discussion is about whether or not one of the traces needs to be added to the background material when determining this estimate. In this paper it is shown, using fully Bayesian methods, that one of the values of evidence from the literature corresponds to the so-called 'identification of common source' problem and the other to the 'identification of specific source' problem (Ommen, 2017). This means that the question whether or not one of the traces needs to be added to the background material reduces to the question whether a common source or specific source problem is under consideration. The distinction between the two values is especially important for the rare type match problem, since the values of evidence differ most in this situation.

***Keywords*** — Value of evidence, rare type match problem, identification of source problem, beta-binomial model

# 1 Introduction

The proper evaluation of evidence in case of a rare type match is a fundamental problem in forensic statistics (Brenner, 2010). Typically, the characteristics of a crime stain (for instance a DNA profile) are compared with the corresponding characteristics of some material from another source (for example from a suspect). The strength of evidence is assessed through comparison of the evidence given two competing statistical hypotheses, presented against a background of knowledge and experience about the world (Robertson & Vignaux, 1993), which is accomplished by considering some set of relevant background material. When the characteristics of the crime stain have not been observed in the background material, one speaks of a 'rare type match problem' (Cereda, 2017).

The evidence in the rare type match problem is usually represented by the beta-binomial model (Cereda, 2017; Dawid, 2017; Dawid & Mortera, 1996; Taroni, Bozza, Biedermann, & Aitken, 2016; Taroni, Bozza, Biedermann, Garbolino, & Aitken, 2010; Weir, 1996). Two approaches are used in the literature to determine the value of evidence, using either a fully Bayesian procedure (Cereda, 2017; Dawid, 2017) or a likelihood ratio with some 'plug-in' estimate for the parameters (Dawid & Mortera, 1996; Taroni et al., 2016, 2010; Weir, 1996). This plug-in estimate is either determined with (Dawid & Mortera, 1996; Taroni et al., 2016) or without (Taroni et al., 2010; Weir, 1996) adding one of the traces to the background material. The approaches result in two different values of evidence and the difference between the values is largest in case of a rare type match. Currently, there is no clear indication which of the approaches should be used, although it has been argued that the fully Bayesian approach is the only proper method to evaluate the value of evidence (Cereda, 2017).

The evidence evaluation process is primarily focused on gathering information about the source of the recovered evidence. Recent work has shown that there are several possibilities to formulate the competing hypotheses, depending on which identification of source question is of interest (Ommen, 2017; Ommen & Saunders, 2018; Ommen, Saunders, & Neumann, 2017). The main focus is on the so-called identification of a *common source* problem and the identification of a *specific source* problem. In a common source problem, all evidence is assumed to come from unknown sources, whereas the specific source problem states that one of the sources is fixed (Ommen, 2017). In this paper we will show, using fully Bayesian methods, that one of the values of evidence of the rare type match problem found in the literature corresponds to the identification of common

source problem and the other to the identification of specific source problem.

In Section 2 both identification of source problems from (Ommen, 2017) are explained more thoroughly, and the corresponding hypotheses and underlying statistical models are made precise. Under these models, both the likelihood ratio and the Bayes factor for discrete evidence are derived in Section 3. Section 4 covers the explanation of the beta-binomial model and the difference between the value of evidence in the common source and specific source problem is presented, which has the largest impact on the value of evidence in the rare type match problem.

# 2 Forensic identification of source problems

In a commonly used illustration of the evidence evaluation process, one of two competing hypotheses is presented by the prosecution (denoted $H_p$) and the other by the defence (denoted $H_d$). Since the strength of evidence is affected by the choice of hypotheses, a correct formulation is of great importance. The hypotheses are often focused on source-level identification and these problems are therefore usually referred to as *forensic identification of source problems* (Ommen & Saunders, 2018). Although other types of identification of source problems may be encountered in forensic science, we will focus on quantifying the value of evidence in the identification of common source problems and the identification of specific source problems when a rare type trace has been recovered. In the next sections, we will follow the framework from (Ommen, 2017) and apply it to a discrete setup.

## 2.1 Identification of common source

In the identification of a *common source* problem, the question of interest is whether or not two sets of unknown source evidence share the same, but unknown origin (Ommen, 2017). This problem could correspond to a situation where DNA is found at two different crime scenes and the question of interest is whether the DNA comes from the same (unknown) person, i.e. whether the two crimes are related. In the identification of common source problem, the hypotheses are typically stated as follows (Ommen & Saunders, 2018):

$H_p$ : The two sets of unknown source evidence ($e_{u_1}$ and $e_{u_2}$) both originate from the same unknown source.

$H_d$ : The two sets of unknown source evidence ($e_{u_1}$ and $e_{u_2}$) originate from two different unknown

sources.

Following (Ommen, 2017), the available evidence for the common source problem consists of the evidence from the first unknown source $e_{u_1}$, the evidence from the second unknown source $e_{u_2}$ and the background material $e_a$, and is denoted $e = \{e_{u_1}, e_{u_2}, e_a\}$. In order to test the competing hypotheses, statistical models for the evidence need to be specified. Since in forensic identification of source problems the hypotheses do not provide a clear concept of the underlying mathematical models, a set of possible sampling models should be considered from which a selection is to be made (Ommen, 2017). The sampling models indicate how the evidence is assumed to be generated and mainly concern the exchangeability of the observations. The sampling models from (Ommen, 2017) can be reformulated for discrete evidence, where no within-source variation is present. This means that each sample from a source is equal and therefore only the generation of the source is of interest. The sampling models can be formulated as follows:

$M_a$ : The background material $e_a$ is generated by randomly selecting $n_a$ sources from the population of sources.

$M_p$ : The unknown source evidence $e_u = \{e_{u_1}, e_{u_2}\}$ is generated by randomly selecting a single source from the population of sources.

$M_d$ : The unknown source evidence $e_u = \{e_{u_1}, e_{u_2}\}$ is generated by independently randomly selecting two sources from the population of sources.

The prosecution will argue that the unknown source evidence is generated according to sampling model $M_p$ and the background material according to $M_a$, whereas the defence states that the unknown source evidence is generated according to $M_d$ and the background material according to $M_a$ (Ommen, 2017). Thus, both the prosecution and the defence agree on the generation of the background material. Note that the prosecution hypothesis implies that $e_{u_1}$ and $e_{u_2}$ are the same with probability 1, since there is no within-source variation, whereas under the defence model $e_{u_1}$ and $e_{u_2}$ are independent.

Since no within-source variation is present for discrete evidence, the frequently used two-level model (Aitken, Zadora, & Lucy, 2007), which models the within-source distribution in the first level and between-source distribution in the second level, reduces to a 'one-level' model where only the between-source distribution is of interest. The model for discrete evidence can be seen as a special case of the general two-level model considered in (Ommen, 2017) by using a degenerate

distribution for the within-source variation, which has been discussed in (van Dorp, 2018).

Now, we introduce a probabilistic model for the rare type match problem. Having observed $e_{u_1}$ and $e_{u_2}$, we consider the experiment of checking for a match in the available evidence set $e = \{e_{u_1}, e_{u_2}, e_a\}$. In practical discrete evidence evaluation, only the situation when the characteristics of $e_{u_1}$ and $e_{u_2}$ are the same will be considered. This means that in practice a match in the available evidence set with $e_{u_1}$ is equivalent to a match with $e_{u_2}$ and therefore it suffices to consider either one. To be more specific, for each source in the available evidence, we define a random variable indicating the 'level of matching' of $e_{u_1}$ (or $e_{u_2}$, which is exactly the same) with the observed evidence from the source under consideration. Since the source(s) of $e_{u_1}$ and $e_{u_2}$ are unknown, we will also consider the experiment of checking for a match with $e_{u_1}$ and $e_{u_2}$.

Let $Y_i$ denote the random variable corresponding to the matching of the evidence from the $i$th source in the background material $e_a$, for $i = 1, 2, \ldots, n_a$. Moreover, let $Y_{u_1}$ denote the random variable corresponding to the matching of the first unknown source evidence $e_{u_1}$ and let $Y_{u_2}$ denote the random variable corresponding to the matching of the second unknown source evidence $e_{u_2}$. The sampling model $M_a$ then implies that

$$Y_i \overset{\text{iid}}{\sim} G(\cdot|\theta_a), \qquad i = 1, 2, \ldots, n_a,$$

where $G$ denotes the probability distribution of the matching of the population of sources indexed by the parameter $\theta_a$. Under the prosecution model, $e_{u_1}$ and $e_{u_2}$ are generated by the same source and therefore we have

$$Y_{u_1} \sim G(\cdot|\theta_a)$$

and $Y_{u_2}$ is equal to $Y_{u_1}$ with probability 1. Under the defence model, $e_{u_1}$ and $e_{u_2}$ are generated by two different sources and we have

$$Y_{u_1} \sim G(\cdot|\theta_a) \qquad \text{and} \qquad Y_{u_2} \sim G(\cdot|\theta_a) \qquad \text{independently.}$$

In Section 3, the representation given in this section will be used to quantify the value of evidence in the common source problem. Later on we will choose the Bernoulli distribution for $G(\cdot|\theta_a)$ to indicate either a match or no match.

## 2.2 Identification of specific source

The identification of a *specific source* problem focuses on the question whether a single set of unknown source evidence comes from a known, specified source or that it originates from a source with unknown origin (Ommen, 2017). This problem could correspond to a situation where DNA is found at a crime scene and a suspect is identified, and the question of interest is whether the DNA comes from the suspect. In the identification of specific source problem the hypotheses are usually stated as follows (Ommen & Saunders, 2018):

$H_p$ : The unknown source evidence $e_u$ originates from the specific source.

$H_d$ : The unknown source evidence $e_u$ does not originate from the specific source, but from some other unknown source.

Following (Ommen, 2017), the available evidence for the specific source problem consists of the unknown source evidence $e_u$, the specific source evidence $e_s$ and the background material $e_a$, and is denoted $e = \{e_u, e_s, e_a\}$. Again, sampling models can be defined to illustrate how the evidence is generated. For the specific source model, the sampling models from (Ommen, 2017) can be reformulated for discrete evidence as follows:

$M_a$ : The background material $e_a$ is generated by randomly selecting $n_a$ sources from the population of sources.

$M_s$ : The specific source evidence $e_s$ is generated by the known, fixed specific source.

$M_p$ : The unknown source evidence $e_u$ is generated by the known, fixed specific source.

$M_d$ : The unknown source evidence $e_u$ is generated by randomly selecting a single source from the population of sources.

The prosecution will state that the unknown source evidence is generated according to sampling model $M_p$. Contrary, the defence will argue that the unknown source evidence is generated according to sampling model $M_d$. Both the prosecution and the defence agree on the generation of the specific source evidence and the background material, which are assumed to be generated according to sampling models $M_s$ and $M_a$, respectively. (Ommen, 2017) Note that the sampling model $M_a$ is identical to the one in the identification of common source problem. For the specific source problem, the prosecution hypothesis implies that $e_u$ and $e_s$ are the same with probability

1, since there is no within-source variation, whereas under the defence model $e_u$ and $e_s$ are independent.

Again, a probabilistic model for the rare type match problem needs to be defined. Because the specific source is known and fixed, all randomness is removed from the model for discrete evidence, so $e_s$ is also fixed. We consider the experiment of checking for a match with the specific source evidence in the available evidence set $e = \{e_u, e_s, e_a\}$. As in the common source problem, let $Y_i$ denote the random variable corresponding to the matching of the evidence from the $i$th source in the background material $e_a$, for $i = 1, 2, \ldots, n_a$. Let $Y_u$ denote the random variable corresponding to the unknown source evidence $e_u$. Since the specific source is known, i.e. not random, the artificial experiment corresponding to the specific source evidence $e_s$ is not random and results always in a realisation $y_s$ indicating a match. As in the common source problem, the sampling model $M_a$ can be represented by

$$Y_i \overset{\text{iid}}{\sim} G(\cdot|\theta_a), \qquad i = 1, 2, \ldots, n_a,$$

where $G$ denotes the probability distribution of the matching of the population of alternative sources, i.e. other sources than the specific source, indexed by the parameter $\theta_a$. For the prosecution model, we have $Y_u = y_s$ with probability 1 since the unknown source evidence is assumed to come from the fixed specific source, and no probability distribution is involved in the evaluation of evidence. Lastly, under the defence model we have

$$Y_u \sim G(\cdot|\theta_a).$$

In Section 3, this representation will be used to quantify the value of evidence in the specific source problem.

# 3 Quantifying the value of evidence

To decide which hypothesis is most probable after observing all evidence, the posterior odds

$$\frac{P(H_p|e)}{P(H_d|e)}$$

is the most natural and frequently used ratio to consider. A generally accepted method to evaluate forensic evidence is based on Bayes' theorem and splits the posterior odds into

$$\frac{P(H_p|e)}{P(H_d|e)} = \frac{P(e|H_p)}{P(e|H_d)} \cdot \frac{P(H_p)}{P(H_d)},$$

or in words

$$\text{Posterior odds} = \text{Value of evidence} \times \text{Prior odds}.$$

It has been argued that the role of the forensic expert is to determine the value of evidence, whereas the prior odds are beyond his or her scope.

Let $f$ denote the likelihood structure of the evidence, where we follow the terminology coined in (Ommen, 2017). In the statistics community, there are two commonly used approaches to evaluate the value of evidence. In one approach the value of evidence is given by the *likelihood ratio*

$$LR(\theta_a; e) = \frac{f(e|\theta_a, H_p)}{f(e|\theta_a, H_d)},$$

which depends on the parameter $\theta_a$, that is unknown but has a true value. Since the true value of $\theta_a$ is unknown, in practice some estimate of the unknown parameter based on the background material is substituted into the likelihood ratio function. Note that in the likelihood ratio approach, the true value of $\theta_a$ is seen as a fixed quantity and not as a random variable.

On the other hand, a fully Bayesian approach can be taken by constructing the *Bayes factor*

$$BF(e) = \frac{\int f(e|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e|\theta_a, H_d)\ d\Pi(\theta_a|H_d)},$$

where a prior distribution is imposed on $\theta_a$ and the unknown parameter is integrated out of the expression. The prior belief of $\theta_a$ is the same given each hypothesis, so that $\Pi(\theta_a) := \Pi(\theta_a|H_p) = \Pi(\theta_a|H_d)$ (Ommen, 2017). Although in forensic statistics the terms likelihood ratio and Bayes factor are often used interchangeably, we will make a strict distinction between these objects in this paper.

Let $g$ denote the probability mass function corresponding to $G$, as defined in the previous

section. For the common source problem, the likelihood ratio is given by

$$LR_{CS}(\theta_a; e) = \frac{1}{g(y_{u_2}|\theta_a)}$$

and the Bayes factor is

$$BF_{CS}(e) = \frac{\int g(y_{u_1}|\theta_a)\ d\Pi(\theta_a|e_a)}{\int g(y_{u_1}|\theta_a)g(y_{u_2}|\theta_a)\ d\Pi(\theta_a|e_a)}, \tag{1}$$

see Appendix 6.1. Alternatively, in the specific source problem the likelihood ratio is

$$LR_{SS}(\theta_a; e) = \frac{1}{g(y_u|\theta_a)}$$

and the Bayes factor is given by

$$BF_{SS}(e) = \frac{1}{\int g(y_u|\theta_a)\ d\Pi(\theta_a|e_a)}, \tag{2}$$

see Appendix 6.2. These general formulas will be used in the next section to quantify the value of evidence for the beta-binomial model.

# 4 The beta-binomial model

In the rare type match problem, the random variables $Y_i$ corresponding to the background material are usually regarded as the result of a sequence of $n_a$ Bernoulli trials with probability of success $\theta_a$. Here, *success* corresponds to the event of observing the same discrete characteristics as on the crime stain, i.e. a match, and *failure* to the event of observing any other characteristic(s). Hence, in the setup from Section 3, the distribution $G(\cdot|\theta_a)$ is the Bernoulli distribution with parameter $\theta_a$. This means that the total number of matches in the background material can be represented by a binomial model with parameters $n_a$ and $\theta_a$. In forensic statistics, the prior probability distribution of $\theta_a$ is often modelled by the beta distribution (Brenner, 2010; Cereda, 2017; Weir, 1996). This is a conventional choice of prior, because of the known conjugacy with the binomial distribution.

Let $s_a$ denote the total number of matches in the background material, i.e. $s_a = \sum_{i=1}^{n_a} y_i$.

Denote the prior distribution of $\theta_a$ by

$$\Theta_a \sim \text{Beta}(\alpha, \beta), \qquad \alpha, \beta > 0.$$

Updating the prior distribution of $\theta_a$ with the background material and using the conjugacy property of the beta-binomial model results in

$$\Theta_a | e_a \sim \text{Beta}(\alpha + s_a, \beta + n_a - s_a).$$

This property will be used recurrently in this section when quantifying the value of evidence. Note that in the rare type match problem $s_a = 0$.

Throughout this section, it is assumed that $y_{u_1} = 1$ and $y_{u_2} = 1$ in the common source problem, and $y_u = 1$ and $y_s = 1$ in the specific source problem, which corresponds to the situation that the characteristics of all the traces match. Note that this is the only situation that will be considered in practical evidence evaluation: it does not make sense to determine the value of evidence if we can already observe that the discrete characteristics do not match. In the following sections, the values of evidence resulting from the approaches used in the literature are compared with the Bayes factors corresponding to the common source and specific source problem.

## 4.1 Two values of evidence from the literature

Currently, the literature does not distinguish between specific source and common source problems when calculating the value of evidence for the rare type match problem. Usually, the value of evidence is based on the likelihood ratio. It is generally accepted that for the rare type match problem the likelihood ratio is given by $1/\theta_a$, which indeed corresponds to the likelihood ratio for both the common source and specific source problem. In this approach, an estimate is needed for $\theta_a$ which will be plugged in the likelihood ratio to arrive at the value of evidence.

Although different estimates can be used, a frequently used estimator is given in (Taroni et al., 2010; Weir, 1996) and considers the mean of the updated prior distribution (posterior mean) of $\theta_a$ given the background material:

$$\hat{\theta}_a = \text{E}[\Theta_a | e_a] = \frac{\alpha + s_a}{\alpha + \beta + n_a}.$$

Plugging this estimate in the likelihood ratio, the value of evidence is given by

$$\frac{\alpha + \beta + n_a}{\alpha + s_a}. \tag{3}$$

Alternatively, in (Dawid & Mortera, 1996; Taroni et al., 2016) it is argued that the value of evidence is given by

$$\frac{\alpha + \beta + n_a + 1}{\alpha + s_a + 1}, \tag{4}$$

which is obtained by considering the mean of the updated prior distribution of $\theta_a$ with both the background material and one of the traces, i.e.

$$\hat{\theta}_a = \mathrm{E}[\Theta_a | e_a, e_s] = \mathrm{E}[\Theta_a | e_a, e_{u_1}] = \frac{\alpha + s_a + 1}{\alpha + \beta + n_a + 1}.$$

The work from (Cereda, 2017) already provided many insights in the evaluation of evidence under the beta-binomial model. It was explained that the value obtained in (3) corresponds to a likelihood ratio approach with a 'standard' Bayesian plug-in estimate, whereas the value in (4) coincides with a fully Bayesian procedure.

The difference between the two values of evidence given in equations (3) and (4) has resulted in a broad discussion about whether or not one of the traces needs to be added to the database. Currently, it is not clear when each value should be used which leads to inconsistencies in the evidence evaluation process. Since $s_a \leq n_a$ and $\beta > 0$ by definition of the problem, we always have that

$$\frac{\alpha + \beta + n_a + 1}{\alpha + s_a + 1} < \frac{\alpha + \beta + n_a}{\alpha + s_a}.$$

Moreover, the difference between both values is most noticeable in the rare type match problem when $s_a = 0$. In Section 4.3 this difference will be further explored.

## 4.2 Common source and specific source Bayes factor

To determine the Bayes factor for the common source problem, equation (1) can be used. Considering this expression, the numerator of the Bayes factor for the common source problem is given

by

$$\int g(y_{u_1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a = \int \theta_a \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a$$
$$= \frac{\alpha+s_a}{\alpha+\beta+n_a}.$$

The denominator is given by

$$\int g(y_{u_1 1}|\theta_a)g(y_{u_2 1}|\theta_a)\pi(\theta_a|e_a)\ d\theta_a = \int \theta_a^2 \frac{\Gamma(\alpha+\beta+n_a)}{\Gamma(\alpha+s_a)\Gamma(\beta+n_a-s_a)}\theta_a^{\alpha+s_a-1}(1-\theta_a)^{\beta+n_a-s_a-1}\ d\theta_a$$
$$= \frac{(\alpha+s_a)(\alpha+s_a+1)}{(\alpha+\beta+n_a)(\alpha+\beta+n_a+1)},$$

so that the common source Bayes factor becomes

$$BF_{CS}(e) = \frac{\alpha+\beta+n_a+1}{\alpha+s_a+1}. \tag{5}$$

Similarly, using the expression in equation (2) the specific source Bayes factor is found to be

$$BF_{SS}(e) = \frac{1}{\int g(y_u|\theta_a)\pi(\theta_a|e_a)\ d\theta_a} = \frac{\alpha+\beta+n_a}{\alpha+s_a}. \tag{6}$$

This means that the value of evidence in equation (3) corresponds to a specific source problem, whereas the value in equation (4) corresponds to a common source problem. Therefore, the question whether or not one of the traces needs to be added to the database reduces to the question whether a common source or specific source problem is under consideration. If one considers a common source problem, equation (5) should be used. If one considers a specific source problem, equation (6) needs to be used.

### 4.3 Value of evidence in the rare type match problem

To illustrate the influence of adding a trace to the database, both the common source and specific source Bayes factor are evaluated for different values of the hyperparameters $\alpha$ and $\beta$ with a database of size $n_a = 100$. In Figure 1, the Bayes factor is shown for the rare type match problem, i.e. when $s_a = 0$. It is immediately visible that the influence of $\beta$ is limited, whereas the Bayes factor reduces as $\alpha$ increases. Moreover, the Bayes factor of the common source problem leads to far more conservative values than the Bayes factor of the specific source problem.

When $\alpha$ is large, the difference between both values of evidence becomes smaller (see also Figure 2). This observation was already made in (Cereda, 2017) and can be explained by the fact that for $\beta/\alpha \to \infty$ the beta distribution becomes a degenerate distribution placing all mass at $\theta_a = 1$. This means that the probability of observing the characteristics of interest tends to 1, which is of course inappropriate for the rare type match problem.

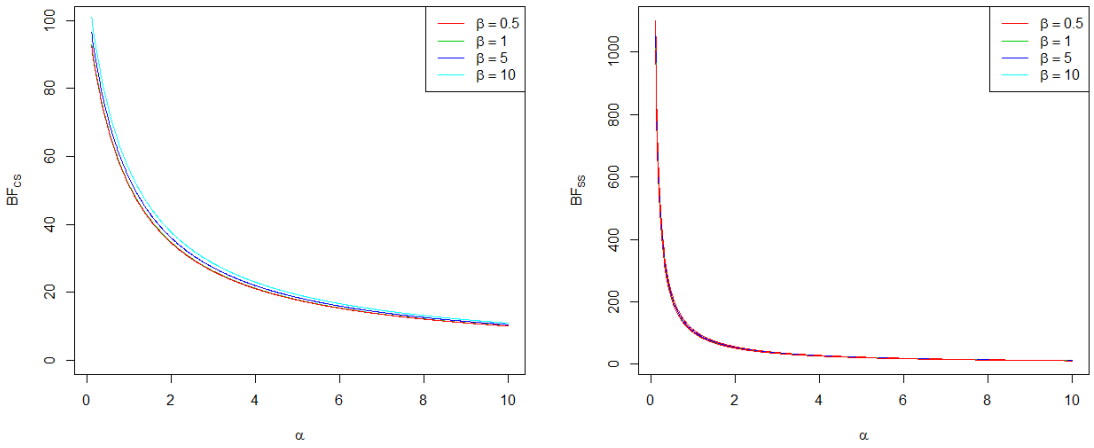


Figure 1: Bayes factor corresponding to the common source problem (left) and the specific source problem (right) in the rare type match problem as function of $\alpha$ for $\beta \in \{0.5, 1, 5, 10\}$ and $n_a = 100$.

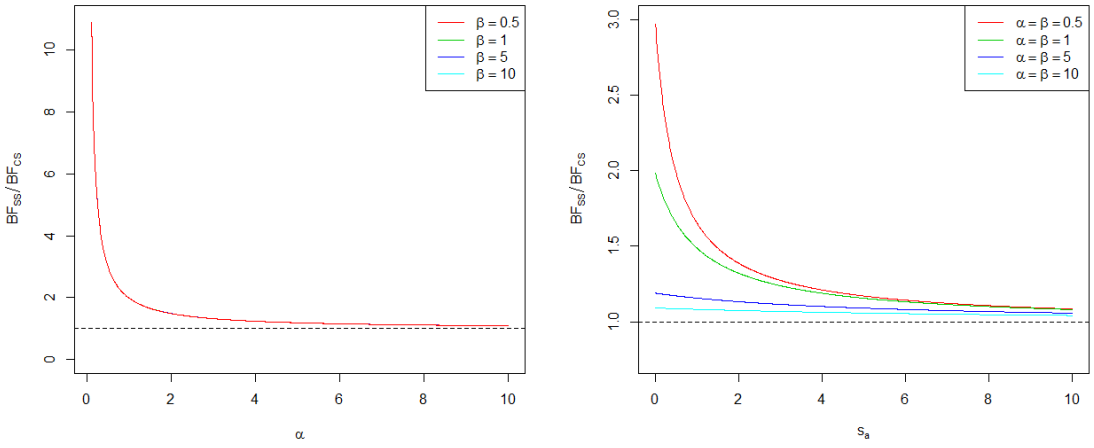


Figure 2: Ratio between $BF_{SS}$ and $BF_{CS}$ as function of $\alpha$ in the rare type match problem (left) and as function of $s_a$ for $\alpha = \beta$ (right), where $\beta \in \{0.5, 1, 5, 10\}$ and $n_a = 100$. The dashed line indicates the value 1.

Figure 2 also shows that the difference between the common source and specific source Bayes factor is most noticeable in the rare type match problem. For other problems, when a reasonable

number of matches is observed in the background material, the difference becomes negligible and both values of evidence would essentially lead to the same conclusions.

# 5 Discussion

The value of evidence in the rare type match problem depends on which identification of source question is considered. In forensic casework, this choice mainly depends on the assumptions a forensic expert makes based on the context of the evidence. The main difference between the common and specific source problem is whether the (first) unknown source evidence is compared to evidence originating from either a fixed or a random source (Ommen, 2017). Of course, one could argue that all evidence is generated from an overall distribution and that the specific evidence under consideration is also a realisation of a random source, which would be an argument in favor of the common source problem. Likewise, the first unknown source evidence in the common source problem could be seen as fixed, which would transform the setup to a specific source problem.

For the rare type match problem, the common source model leads to a more conservative value of evidence than the specific source problem. However, for the court most interest seems to lie in answering a specific source question, which would help provide a decision between guilt and innocence of a specific suspect (Ommen & Saunders, 2018). Since the choice between the common and specific source problem has a major impact on the value of evidence for the rare type match problem, the forensic expert should be aware of the consequences of this choice and carefully state the assumptions before evaluating the evidence.

# 6 Conclusion

For the beta-binomial model, which is frequently used in the rare type match problem, two values of evidence from the literature are discussed. The main question of interest here is whether or not one of the traces needs to be added to the background material to determine the plug-in estimate for the likelihood ratio. In this paper the sampling models for both the common source and specific source problem from (Ommen, 2017) are presented for the evaluation of discrete evidence. The underlying statistical models are made precise and the value of evidence is considered, using either the likelihood ratio or the Bayes factor. Using a fully Bayesian approach it is shown that one of the values from the literature corresponds to the identification of common source problem and

the other to the identification of specific source problem. This means that the question of interest reduces to the question whether a common source or specific source problem is under consideration. The value of evidence from the common source problem is found to be more conservative than the specific source Bayes factor. The difference between both values is especially noticeable in the rare type match problem.

# Appendix

## 6.1 Likelihood ratio and Bayes factor of the common source problem

Let $f$ denote the likelihood structure of the evidence. For the common source problem, the likelihood ratio can be found from

$$\begin{aligned} LR_{CS}(\theta_a; e) = \frac{f(e|\theta_a, H_p)}{f(e|\theta_a, H_d)} &= \frac{f(e_{u_1}, e_{u_2}|\theta_a, H_p) f(e_a|\theta_a, H_p)}{f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d) f(e_a|\theta_a, H_d)} \\ &= \frac{f(e_{u_2}|e_{u_1}, \theta_a, H_p) f(e_{u_1}|\theta_a, H_p)}{f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)} \end{aligned}$$

where we use the assumption that $f(e_a|\theta_a, H_p) = f(e_a|\theta_a, H_d)$ and the rule of conditional probability $f(x, y) = f(x|y)f(y)$. Note that this expression of the likelihood ratio is slightly different from Equation (3.6) in (Ommen, 2017), since we want to condition $e_{u_2}$ on $e_{u_1}$ so that we can use that according to the prosecution both sets of common source evidence are equal with probability 1.

Using the stochastic model corresponding to the common source rare type match problem, as introduced in Section 2, we translate this in terms of statements in $Y$:

$$\begin{aligned} LR_{CS}(\theta_a; e) &= \frac{P(Y_{u_2} = y_{u_2}|Y_{u_1} = y_{u_1}, \theta_a, H_p) P(Y_{u_1} = y_{u_1}|\theta_a, H_p)}{P(Y_{u_1} = y_{u_1}|\theta_a, H_d) P(Y_{u_2} = y_{u_2}|\theta_a, H_d)} \\ &= \frac{P(Y_{u_1} = y_{u_1}|\theta_a, H_p)}{P(Y_{u_1} = y_{u_1}|\theta_a, H_d) P(Y_{u_2} = y_{u_2}|\theta_a, H_d)} \\ &= \frac{g(y_{u_1}|\theta_a)}{g(y_{u_1}|\theta_a) g(y_{u_2}|\theta_a)} = \frac{1}{g(y_{u_2}|\theta_a)}, \end{aligned}$$

where we use that under the prosecution model $Y_{u_2}$ is equal to $Y_{u_1}$ with probability 1, so that

$$P(Y_{u_2} = y_{u_2}|Y_{u_1} = y_{u_1}, \theta_a, H_p) = 1.$$

In (Ommen, 2017), the following alternative expression of the Bayes factor is derived:

$$BF_{CS}(e) = \frac{\int f(e|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} = \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p)\ d\Pi(\theta_a|e_a, H_p)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)}.$$

This expression will be used to derive the Bayes factor for the common source problem:

$$\begin{aligned} BF_{CS}(e) &= \frac{\int f(e_{u_1}, e_{u_2}|\theta_a, H_p)\ d\Pi(\theta_a|e_a, H_p)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)} \\ &= \frac{\int f(e_{u_2}|e_{u_1}, \theta_a, H_p) f(e_{u_1}|\theta_a, H_p)\ d\Pi(\theta_a|e_a, H_p)}{\int f(e_{u_1}|\theta_a, H_d) f(e_{u_2}|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)}. \end{aligned}$$

Again, we use the rule of conditional probability $f(x, y) = f(x|y)f(y)$ to condition $e_{u_2}$ on $e_{u_1}$ given $H_p$. For the probability model corresponding to the rare type match problem, this results in

$$\begin{aligned} BF_{CS}(e) &= \frac{\int P(Y_{u_2} = y_{u_2}|Y_{u_1} = y_{u_1}, \theta_a, H_p) P(Y_{u_1} = y_{u_1}|\theta_a, H_p)\ d\Pi(\theta_a|e_a, H_p)}{\int P(Y_{u_1} = y_{u_1}|\theta_a, H_d) P(Y_{u_2} = y_{u_2}|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)} \\ &= \frac{\int g(y_{u_1}|\theta_a)\ d\Pi(\theta_a|e_a, H_p)}{\int g(y_{u_1}|\theta_a) g(y_{u_2}|\theta_a)\ d\Pi(\theta_a|e_a, H_d)}, \end{aligned}$$

where we use that under the prosecution model $Y_{u_2}$ is equal to $Y_{u_1}$ with probability 1 and hence

$$P(Y_{u_2} = y_{u_2}|Y_{u_1} = y_{u_1}, \theta_a, H_p) = 1.$$

Assuming that the prior distribution of $\theta_a$ given the background material is the same for both hypotheses, the common source Bayes factor is given by

$$BF_{CS}(e) = \frac{\int g(y_{u_1}|\theta_a)\ d\Pi(\theta_a|e_a)}{\int g(y_{u_1}|\theta_a) g(y_{u_2}|\theta_a)\ d\Pi(\theta_a|e_a)}.$$

Note that this derivation is analogous to the development used in (Taroni et al., 2010).

## 6.2 Likelihood ratio and Bayes factor of the specific source problem

Let $f$ denote the likelihood structure of the evidence. For the specific source problem, the likelihood ratio can be found from

$$\begin{aligned} LR_{SS}(\theta_a; e) = \frac{f(e|\theta_a, H_p)}{f(e|\theta_a, H_d)} &= \frac{f(e_u, e_s|H_p) f(e_a|\theta_a, H_p)}{f(e_u|\theta_a, H_d) f(e_s|H_d) f(e_a|\theta_a, H_d)} \\ &= \frac{f(e_u|e_s, H_p) f(e_s|H_p)}{f(e_u|\theta_a, H_d) f(e_s|H_d)} = \frac{f(e_u|e_s, H_p)}{f(e_u|\theta_a, H_d)}, \end{aligned}$$

where we use the assumption that $f(e_a|\theta_a, H_p) = f(e_a|\theta_a, H_d)$ and $f(e_s|H_p) = f(e_s|H_d)$. Moreover, we use the rule of conditional probability $f(x, y) = f(x|y)f(y)$ to condition $e_u$ on $e_s$ so that we can use that according to the prosecution the unknown source evidence is equal to the specific source evidence with probability 1. Note that this expression of the likelihood ratio is slightly different from Equation (3.9) in (Ommen, 2017) since we omit the parameter $\theta_s$, which is used to model the 'within-specific source' variation that is not present in the discrete setup.

Using the stochastic model corresponding to the specific source rare type match problem, as introduced in Section 2, we translate this in terms of statements in $Y$:

$$LR_{SS}(\theta_a; e) = \frac{P(Y_u = y_u|y_s, H_p)}{P(Y_u = y_u|\theta_a, H_d)} = \frac{1}{g(y_u|\theta_a)},$$

where we use that under the prosecution model $Y_u$ is equal to $y_s$ with probability 1, so that

$$P(Y_u = y_u|y_s, H_p) = 1.$$

The Bayes factor for the specific source problem follows from:

$$\begin{aligned}
BF_{SS}(e) &= \frac{\int f(e|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{\int f(e_u, e_s|H_p) f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e_u|\theta_a, H_d) f(e_s|H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{f(e_u|e_s, H_p) f(e_s|H_p) \int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{f(e_s|H_d) \int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= \frac{f(e_u|e_s, H_p) \int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{\int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \cdot \frac{f(e_a|H_d)}{f(e_a|H_p)} \\
&= f(e_u|e_s, H_p) \cdot \frac{\int f(e_a|\theta_a, H_p)\ d\Pi(\theta_a|H_p)}{f(e_a|H_p)} \cdot \frac{f(e_a|H_d)}{\int f(e_u|\theta_a, H_d) f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)} \\
&= f(e_u|e_s, H_p) \cdot \frac{f(e_a|H_p)}{f(e_a|H_p)} \cdot \left[ \int f(e_u|\theta_a, H_d) \frac{f(e_a|\theta_a, H_d)\ d\Pi(\theta_a|H_d)}{f(e_a|H_d)} \right]^{-1} \\
&= \frac{f(e_u|e_s, H_p)}{\int f(e_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)},
\end{aligned}$$

where we use the assumptions that $f(e_s|H_p) = f(e_s|H_d)$ and $f(e_a|H_d) = f(e_a|H_p)$. This derivation is inspired by Derivation (3.11) from (Ommen, 2017), but again the parameter $\theta_s$ is omitted. Moreover, we choose to condition $e_u$ on $e_s$ given $H_p$ by applying the rule of conditional probability $f(x, y) = f(x|y)f(y)$. For the probability model corresponding to the rare type match problem,

this results in

$$BF_{SS}(e) = \frac{P(Y_u = y_u|y_s, H_p)}{\int P(Y_u = y_u|\theta_a, H_d)\ d\Pi(\theta_a|e_a, H_d)} = \frac{1}{\int g(y_u|\theta_a)\ d\Pi(\theta_a|e_a, H_d)},$$

where we use that under the prosecution model $Y_u$ is equal to $y_s$ with probability 1 so that

$$P(Y_u = y_u|y_s, H_p) = 1.$$

Assuming that the prior distribution of $\theta_a$ given the background material is the same for both hypotheses, the specific source Bayes factor is given by

$$BF_{SS}(e) = \frac{1}{\int g(y_u|\theta_a)\ d\Pi(\theta_a|e_a)}.$$

# References


Aitken, C., Zadora, G., & Lucy, D. (2007). A two-level model for evidence evaluation. *Journal of forensic sciences*, *52*(2), 412–419.

Brenner, C. (2010). Fundamental problem of forensic mathematics—the evidential value of a rare haplotype. *Forensic Science International: Genetics*, *4*(5), 281–291.

Cereda, G. (2017). Bayesian approach to LR assessment in case of rare type match. *Statistica Neerlandica*, *71*(2), 141–164.

Dawid, A. (2017). Forensic likelihood ratio: Statistical problems and pitfalls. *Science & Justice*, *57*(1), 73–75.

Dawid, A., & Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 425–443.

Ommen, D. (2017). *Approximate statistical solutions to the forensic identification of source problem* (PhD thesis). South Dakota State University.

Ommen, D., & Saunders, C. (2018). Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, *17*(2), 179–197.

Ommen, D., Saunders, C., & Neumann, C. (2017). The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *Journal of Statistical Computation and Simulation*, *87*(8), 1608–1643.

Robertson, B., & Vignaux, G. (1993). Probability—the logic of the law. *Oxford Journal of Legal*

*Studies*, *13*(4), 457–478.

Taroni, F., Bozza, S., Biedermann, A., & Aitken, C. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, *15*(1), 1–16.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., & Aitken, C. (2010). *Data analysis in forensic science: a Bayesian decision perspective* (Vol. 88). John Wiley & Sons.

van Dorp, I. (2018). *Statistical modelling of forensic evidence* (Master's thesis, Delft University of Technology, The Netherlands). Retrieved from `http://resolver.tudelft.nl/uuid:26b62fb7-97ed-438f-88e1-f7995ab4c73c`

Weir, B. (1996). *Genetic data analysis ii*. Sinauer Associates, Sunderland.