

# Forensic identification: the Island Problem and its generalisations

Klaas Slooten\* and Ronald Meester†

April 26, 2010

## Abstract

In forensics it is a classical problem to determine, when a suspect  $S$  shares a property  $\Gamma$  with a criminal  $C$ , the probability that  $S = C$ . In this paper we give a detailed account of this problem in various degrees of generality. We start with the classical case where the probability of having  $\Gamma$ , as well as the a priori probability of being the criminal, is the same for all individuals. We then generalize the solution to deal with heterogeneous populations, biased search procedures for the suspect,  $\Gamma$ -correlations, uncertainty about the subpopulation of the criminal and the suspect, and uncertainty about the  $\Gamma$ -frequencies. We also consider the effect of the way the search for  $S$  is conducted, in particular when this is done by a database search. We concentrate on the so-called posterior odds, but also show how conditioning is of importance when one wants to quantify the “weight” of the evidence by a likelihood ratio.

KEYWORDS: Island problem, Forensic identification, Weight of evidence, Posterior odds, Bayes’ rule.

## 1 Introduction

In 1968, a couple stood to trial in a notorious case, known as “People of the State of California vs. Collins”. The pair had been arrested since it matched

---

\*Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands; k.slooten@nfi.minjus.nl

†VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands; rmeester@few.vu.nl

eye-witness descriptions. It was estimated by the prosecution that only one in twelve million couples would match this description. The jury were invited to consider the probability that the accused pair were innocent and returned a verdict of guilty.

Later, the verdict was overturned, essentially because of the flaws in the statistical reasoning. The case sparked interest in the abstraction of this problem, which became known as the island problem, following terminology introduced by Eggleston [4]. Its formulation is the following. A crime has been committed by an unknown member of a population of  $N+1$  individuals. It is known that the criminal has a certain property  $\Gamma$ . Each individual has  $\Gamma$  (independently) with probability  $p$ . A random member of the population is tested and observed to have  $\Gamma$ . What is the probability that it is the criminal?

This problem has been quite extensively studied in the literature. For example, Balding and Donnelly [1] give a detailed account of the island problem as well as of its generalization to inhomogeneous populations or (alternatively) uncertainty about  $p$ . They also discuss the effects of a database search or a sequential search (i.e., a search which stops when the first  $\Gamma$ -bearer is found). Dawid and Mortera have studied the generalization of the island problem to the case where the evidence may be unreliable [2, 3].

The current paper has a twofold nature: it is expository in the sense that some of the above mentioned results are reproduced - albeit presented in a somewhat different way, see the next two paragraphs - and a research article in the sense that we consider generalizations which to our knowledge have not appeared elsewhere.

Most texts typically focus on the “likelihood ratio”, the quantity that transforms “prior” odds on guilt, that is, before seeing the evidence, into “posterior” odds after seeing the evidence. The likelihood ratio is often viewed as the weight of the evidence - it is the quantity of interest for a forensic lab, which is unable or not allowed to compute prior or posterior odds, this being the domain of the court.

However, we mostly focus on so called “posterior probabilities”, that is, the probability that a suspect is guilty, given background information and/or evidence. There is a good reason to do so. Indeed, the weight of the evidence depends on which of the available information is regarded as background information or as evidence (and of course also on the propositions that one is interested in proving). For the island problem, the background information is relevant as soon as the population has subpopulations, each with their own  $\Gamma$ -frequency. In that case, considering the information that the criminal has  $\Gamma$  as information on the one hand or as evidence on the other, leads to different likelihood ratios, but the posterior odds are (of course) the same.

We will go into this phenomenon in some detail, considering subpopulations simultaneously with uncertainty about to which subpopulation the criminal and the suspect belong, together with uncertainty about the  $\Gamma$ -frequencies in each of the subpopulations. Another possibility which we will consider is that of  $\Gamma$ -correlation or a biased search (i.e., the choice of suspect depends on the true identity of the criminal).

The outline of this paper is as follows. In Section 2, we review the classical island problem. We then consider in Section 3 the effect of having a biased search protocol, and of having  $\Gamma$ -correlations; we show that these two different types of having dependencies are strongly related to each other. In Section 4, we treat the case where the population is a disjoint union of subpopulations, each with their own  $\Gamma$ -frequency and prior probability of having issued the criminal. In Section 5, we consider the effect of uncertainty of the  $\Gamma$ -frequencies, both in a homogeneous and heterogeneous population. In addition, we investigate the effect on the likelihood ratio of uncertainty about the criminal's and the suspect's subpopulations. Section 6 deals with the case in which a suspect is found through a match in a database. Finally, in Section 7 we present a significant number of numerical examples.

We have tried to include all details of the computations, but at the same time to state our conclusions in a non-technical and accessible way. Our main conclusions can be recognised in the text as bulleted ( $\bullet$ ) lists. As such, we hope that our contribution is interesting and useful both for mathematicians and forensic scientists.

## 2 The classical case

Our starting point is a collection  $X$  of  $N + 1$  individuals. All forthcoming random variables are defined on a (non-specified) probability space with probability measure  $P$ . The random variables  $C$  and  $S$  take values in  $X$  and represent the criminal and the suspect respectively. Furthermore, we have a characteristic  $\Gamma$ , for which we introduce indicator random variables  $\Gamma_x$ , taking value 1 if  $x \in X$  has the characteristic  $\Gamma$  and 0 otherwise. The  $\Gamma_x$  are independent of  $(S, C)$  in the sense that  $P(\Gamma_x = 1 \mid C = y, S = z) = P(\Gamma_x = 1)$  for all  $x, y, z$ . The number of  $\Gamma$ -bearers is written as  $U = \sum_{x \in X} \Gamma_x$ .

We are primarily interested in the conditional probability

$$P(C = s \mid S = s, \Gamma_C = \Gamma_s = 1);$$

often we follow the habit of stating the so-called *posterior odds* in favour of guilt, that is,

$$\frac{P(C = s \mid S = s, \Gamma_C = \Gamma_s = 1)}{P(C \neq s \mid S = s, \Gamma_C = \Gamma_s = 1)}. \quad (2.1)$$

Since we will often be working conditional on  $\{S = s\}$  we introduce the notation

$$P_s(\cdot) = P(\cdot \mid S = s).$$

We define the events  $I := \{\Gamma_C = 1\}$ ,  $G := \{S = C\}$ ,  $E := \{\Gamma_S = 1\}$ ,  $E_x := \{\Gamma_x = 1\}$  and  $G_x = \{x = C\}$ . We will sometimes refer to the event  $I$  (or similar events) as “information”, and to  $E$  (or similar events) as “evidence”; this is just colloquial use of language, and sometimes we will view  $I$  as part of the evidence.

With this notation, (2.1) reads

$$\frac{P_s(G \mid I, E)}{P_s(G^c \mid I, E)},$$

which can be rewritten as

$$\frac{P_s(G \mid I, E)}{P_s(G^c \mid I, E)} = \frac{P_s(E \mid G, I)}{P_s(E \mid G^c, I)} \cdot \frac{P_s(G \mid I)}{P_s(G^c \mid I)} \quad (2.2)$$

$$= \frac{P_s(I, E \mid G)}{P_s(I, E \mid G^c)} \cdot \frac{P_s(G)}{P_s(G^c)}. \quad (2.3)$$

The left hand side of this equation is called the *posterior odds*. In (2.2), we arrive at the posterior odds in two different ways. In the first way, we “started” with background information  $I$ , and the quotient  $P_s(G \mid I)/P_s(G^c \mid I)$ , called the *prior odds*, is transformed into the posterior odds by multiplication with  $P_s(E \mid G, I)/P_s(E \mid G^c, I)$ . This latter quotient is called the *likelihood ratio*, and transforms prior odds into posterior ones. In the second line, we “started” from prior odds  $P_s(G)/P_s(G^c)$ , that is, we interpreted both  $I$  and  $E$  as evidence. The likelihood ratio in that case is in general different as well. From this it already follows that depending on which part of the available material is considered evidence or background information, the likelihood ratio will be different. The posterior odds do not change, and this is the main reason to focus on these.

In this section treating the classical case, we assume that  $C$  and  $S$  are independent and that  $C$  is uniformly distributed on  $X$ . Furthermore, the  $\Gamma_x$  are independent and identically Bernoulli distributed with success probability  $p$ . The outcomes in this section do not depend on the particular  $s$  we condition on, but for the sake of consistency, we do write  $P_s$  instead of  $P$ . We abbreviate  $E := E_s$ .

The independence between  $S$  and  $C$  now implies that  $P_s(G) = 1/(N+1)$  and it easily follows that

$$\frac{P_s(G \mid I, E)}{P_s(G^c \mid I, E)} = \frac{1}{p} \cdot \frac{P_s(G \mid I)}{P_s(G^c \mid I)} = \frac{1}{p} \cdot \frac{P_s(G)}{P_s(G^c)} = \frac{1}{Np}. \quad (2.4)$$

In this case, the likelihood ratio (the quantity that transfers prior odds into posterior odds) is the same and equal to  $1/p$  regardless of whether we view  $I$  as evidence or background information upon which we condition.

In the next two subsections we will examine (for the classical case) how  $P_s(G | I, E)$  is related to the random variable  $U$ . It turns out that we may express  $P_s(G | I, E)$  both as the inverse of the expectation of  $U$  and as the expectation of the inverse of  $U$ , as long as we condition correctly.

## 2.1 Expected number of $\Gamma$ -bearers.

Before anyone is tested for  $\Gamma$ ,  $U$  has a  $\text{Bin}(N + 1, p)$ -distribution. When the crime is committed and it is observed that the criminal has  $\Gamma$ , we condition on  $\Gamma_C = 1$  and obtain

$$\begin{aligned} P_s(U = k + 1 | I) &= \frac{P_s(I | U = k + 1)P_s(U = k + 1)}{P_s(I)} \\ &= \frac{\frac{k+1}{N+1} \binom{N+1}{k+1} p^{k+1} (1-p)^{N-k}}{p} \\ &= \binom{N}{k} p^k (1-p)^{N-k}. \end{aligned}$$

It follows that the probability that  $U = k + 1$ , given  $I$ , is equal to the probability that a random variable with a  $\text{Bin}(N, p)$ -distribution takes value  $k$ , i.e.,  $U | I$  is distributed as  $1 + \text{Bin}(N, p)$ . Hence, writing  $E_s$  for expectation with respect to  $P_s$ , we have

$$E_s(U | I) = 1 + Np.$$

Thus, the posterior probability of guilt is given by the inverse of the expected number of  $\Gamma$ -bearers, where this expectation takes into account that there is a specific individual - the criminal - who has  $\Gamma$ :

$$P_s(G | E, I) = \frac{1}{E_s(U | I)}. \quad (2.5)$$

Intuitively this makes sense: the criminal is a  $\Gamma$ -bearer, any one of the  $\Gamma$ -bearers is equally likely to be the criminal, and we have found one of them. So we have to compute the expected number of  $\Gamma$ -bearers, given the knowledge that  $C$  is one of them.

## 2.2 Expected inverse number of $\Gamma$ -bearers

As we have seen,  $U | I$  is distributed as  $1 + \text{Bin}(N, p)$ . Therefore, one expects  $\mathbb{E}_s(U | I) = 1 + Np$  bearers of  $\Gamma$ . If we in addition also condition on  $\Gamma_S = 1$ , we compute

$$\begin{aligned} P_s(U = k | E, I) &= \frac{P_s(E | U = k, I)P_s(U = k | I)}{P_s(E | I)} \\ &= \frac{\frac{k}{N+1}P_s(U = k | I)}{\frac{1+Np}{N+1}} \\ &= \frac{k}{1 + Np}P_s(U = k | I). \end{aligned}$$

We use this calculation to obtain:

$$\mathbb{E}_s(U^{-1} | E, I) = \sum_{k=1}^{N+1} \frac{1}{k} P_s(U = k | E, I) \quad (2.6)$$

$$= \sum_{k=1}^{N+1} \frac{1}{k} \frac{k}{1 + Np} P_s(U = k | I) \quad (2.7)$$

$$= \frac{1}{1 + Np}. \quad (2.8)$$

Summarizing,

$$P_s(G | E, I) = (\mathbb{E}_s(U | I))^{-1} = \mathbb{E}_s(U^{-1} | I, E). \quad (2.9)$$

So  $P_s(G | I, E)$  is in fact also equal to the expectation of  $U^{-1}$ , however, not of  $U | I$  but of  $U | I, E$ . This can be understood in an intuitive way: both  $S$  and  $C$  have  $\Gamma$ , they have been sampled with replacement, so the probability that they are equal is the inverse of the number of  $\Gamma$ -bearers. This number is unknown, so we have to take expectations, given knowledge of  $S$  and  $C$ .

When we compare this explanation with the one of (2.5), we see the importance of careful conditioning.

## 2.3 Effect of a search, Yellin's formula

So far,  $S$  and  $\Gamma_S$  were supposed to be independent of each other. In this subsection, we consider a different situation. The random variable  $C$  representing the criminal is still supposed to be uniformly distributed, but the definition of  $S$  is different: we repeatedly select from  $X$  - with or without replacement - until a  $\Gamma$ -bearer is found, without keeping any records on the

search itself, such as its duration. The  $\Gamma$ -bearer found this way is denoted by  $S$ ; if there is no  $\Gamma$ -bearer in the population, we set  $S = *$ , and define  $\Gamma_* = 0$ . As before we write  $E = \{\Gamma_S = 1\}$  and note that in this situation  $I \subseteq E$ .

As above, we are interested in  $P_s(G | E, I)$  which, since  $I \subseteq E$ , reduces to  $P_s(G | I)$ , and this conditional probability is easy to compute:

$$\begin{aligned} P_s(G | E, I) &= P_s(G | I) = \sum_{k=0}^N k^{-1} P_s(U = k | I) \\ &= E_s(U^{-1} | I). \end{aligned} \tag{2.10}$$

This formula was published by Yellin in [10] as the solution to this version of the island problem with a search. Sometimes, however, it is incorrectly quoted in the literature (e.g. in [1]) as an incorrect solution to the island problem without search as we have discussed it.

## 2.4 Conclusions

- The classical version of the island problem is not difficult to solve, but the relation between the probability of guilt and the expected number of  $\Gamma$ -bearers is rather subtle. The basic formula is

$$P_s(G | I, E) = (E_s(U | I))^{-1} = E_s(U^{-1} | I, E) = \frac{1}{1 + Np}.$$

- In the case of a search we have  $I \subseteq E$  and this leads to

$$P_s(G | E, I) = E_s(U^{-1} | I).$$

These outcomes are independent of  $s$ .

- For the value of the likelihood ratio, it does not matter whether or not one interprets  $I$  as background information or as evidence - in both cases the value is  $1/p$ .
- The prior odds, the likelihood ratio and (hence) the posterior odds are all independent of  $s$ .

## 3 Dependencies

In this section we relax the condition that the  $\Gamma_x$  are independent random variables or that  $S$  and  $C$  are independent. To this end, we define

$$c_{x,y} = P(\Gamma_x = 1 | \Gamma_y = 1), \tag{3.1}$$

$$\sigma_{x,y} = P(S = x | C = y, I). \tag{3.2}$$

### 3.1 Independent $\Gamma_x$

First we assume that the  $\Gamma_x$  are independent (not necessarily identically distributed) random variables. We write  $p_x$  for  $P(\Gamma_x = 1)$ . Now (2.1) becomes

$$\begin{aligned}
 \frac{P_s(E | G, I)P_s(G | I)}{P_s(E | G^c, I)P_s(G^c | I)} &= \frac{P(E | G, S = s, I) P(G | S = s, I)}{P(E | G^c, S = s, I) P(G^c | S = s, I)} \\
 &= \frac{1 P(G, S = s | I)}{p_s P(G^c, S = s | I)} \\
 &= \frac{1 P(S = s | C = s, I) P(C = s | I)}{p_s P(S = s | C \neq s, I) P(C \neq s | I)} \\
 &= \frac{1}{p_s} \frac{\sigma_{s,s} P(C \neq s | I) P(C = s | I)}{\sum_{y \neq s} \sigma_{s,y} P(C = y | I) P(C \neq s | I)} \quad (3.3)
 \end{aligned}$$

In this last expression (3.3), the first term  $1/p_s$  is the likelihood ratio in case of a search such that  $\sigma_{s,s} = \sigma_{s,y}$  for all  $y \neq s$ , i.e., such that the probability of selecting  $s$  is independent of  $C$ . In particular, this holds for a search where  $S$  is uniformly random but other distributions of  $(S, C)$  may also satisfy this criterion.

The middle term in (3.3) is the term that accounts for the bias of the search, i.e., it expresses the effect of the dependence between  $S$  and  $C$  in the case  $S = s$ .

The last term of (3.3) is the “prior odds”, the odds in favour of  $C = s$ , when  $I$  is taken into account. It is of course also possible to start from “prior odds”  $P(C = s)/P(C \neq s)$ ; this will yield the same posterior odds, but a different expression for the likelihood ratio. We will make this explicit for some special cases later on.

### 3.2 Arbitrary $\Gamma_x$

We now assume again that  $S$  and  $C$  are independent, but we drop the assumption that the  $\Gamma_x$  are independent. In that case, we can write

$$\begin{aligned}
\frac{P_s(G | I, E)}{P_s(G^c | I, E)} &= \frac{P_s(E | G, I) P_s(G | I)}{P_s(E | G^c, I) P_s(G^c | I)} \\
&= \frac{P_s(G | I)}{P_s(E, G^c | I)} \\
&= \frac{P_s(G | I)}{\sum_{y \neq s} P_s(E, C = y | I)} \\
&= \frac{P_s(G | I)}{\sum_{y \neq s} P_s(E | C = y, I) P_s(C = y | I)} \\
&= \frac{P_s(G | I)}{\sum_{y \neq s} c_{s,y} P_s(C = y | I)} \tag{3.4}
\end{aligned}$$

$$= \frac{1}{p_s} \frac{P_s(G^c | I)}{\sum_{y \neq s} \frac{c_{s,y}}{p_s} P_s(C = y | I)} \frac{P_s(G | I)}{P_s(G^c | I)}, \tag{3.5}$$

where we have used that

$$P_s(E | I, C = y) = P(\Gamma_s = 1 | \Gamma_y = 1) = c_{s,y},$$

since we have assumed that the  $\Gamma_i$  are independent of  $S$  and  $C$ .

As for the case of a biased search, the term  $1/p_s$  is the likelihood ratio that we obtain in the case where the  $\Gamma$ -correlations do not play a role, i.e., when  $c_{s,y} = p_s$  for all  $y \neq s$ . The middle term, analogously to (3.3),

$$\frac{P_s(C \neq s | I)}{\sum_{y \neq s} \frac{c_{s,y}}{p_s} P_s(C = y | I)} = \frac{P_s(C \neq s | I)}{\sum_{y \neq s} \frac{c_{y,s}}{p_y} P_s(C = y | I)} \tag{3.6}$$

accounts for the  $\Gamma$ -correlations, and the last term

$$\frac{P_s(G | I)}{P_s(G^c | I)} = \frac{P(C = s | I)}{P(C \neq s | I)}$$

describes the prior odds, conditional on  $I = \{\Gamma_C = 1\}$ . If we remove this conditioning, we get

$$\frac{P_s(G | I, E)}{P_s(G^c | I, E)} = \frac{P_s(G)}{\sum_{y \neq s} c_{y,s} P_s(C = y)} \tag{3.7}$$

$$= \frac{1}{p_s} \frac{P_s(C \neq s)}{\sum_{y \neq s} \frac{c_{y,s}}{p_s} P_s(C = y)} \frac{P_s(G)}{P_s(G^c)}. \tag{3.8}$$

As for (3.5), the last line contains three terms: the likelihood ratio  $1/p_s$  in the uncorrelated case, the term due to the correlation and the prior odds.

Finally, note that (3.4) and (3.7) imply

$$\frac{P_s(G | I, E)}{P_s(G^c | I, E)} = \frac{P_s(G | I)}{\sum_{y \neq s} c_{s,y} P_s(C = y | I)} = \frac{P_s(G)}{\sum_{y \neq s} c_{y,s} P_s(C = y)} \quad (3.9)$$

(or equivalently, the symmetry between the middle terms in (3.5) and (3.8)): the way the correlation between the  $\Gamma_i$  appear in the posterior odds depends on whether or not one considers  $I = \{\Gamma_C = 1\}$  to be evidence, or an event upon which everything is conditional.

### 3.3 Comparison of biased search and $\Gamma$ -correlations

When we compare the posterior odds (3.3) and (3.5) of the two situations, we see that the expressions are very similar. Both have a correction factor in the denominator. In fact, when  $S$  and  $C$  are independent, then in (3.5)  $P_s$  can be replaced with  $P$ , and the two cases reduce to each other if  $\sigma_{x,y}/\sigma_{x,x} = c_{x,y}$  for all  $x \neq y$ . A trivial example of this is obtained when  $C$  is uniform on  $X$  and the  $\Gamma_x$  are independent Bernoulli random variables. More generally, every case of a biased search without  $\Gamma$ -correlations where the correlation coefficients between criminal and suspect are such that  $0 \leq p_y \frac{\sigma_{x,x}}{\sigma_{x,y}} \leq 1$  is equivalent (as far as the probability of guilt is considered) to a case where the search is unbiased but the  $\Gamma_x$  are correlated with coefficients  $c_{x,y} = p_y \frac{\sigma_{x,x}}{\sigma_{x,y}}$ .

## 4 Heterogeneous populations

In this section we consider the situation where the population consists of several subpopulations, each with their own  $\Gamma$ -frequency and each with their own probability of containing the criminal. To model this, we write  $X$  as a disjoint union of subpopulations  $X_i$ :

$$X = X_1 \cup \dots \cup X_m, \quad (4.1)$$

with  $X_i \cap X_j = \emptyset$  whenever  $i \neq j$ . If  $x \in X_i$ , we say that  $x$  is in subpopulation  $i$  and write  $i = X(x)$ . Let  $N_i = |X_i|$  be the size of subpopulation  $X_i$ . We write  $N_x = N_i$  if  $i = X(x)$ . Let

$$P(C \in X_i) = \beta_i, \quad (4.2)$$

where the  $\beta_i$ 's are positive and satisfy  $\sum_{i=1}^m \beta_i = 1$ . We assume that the random variables  $\Gamma_x$  are independent Bernoulli variables with probability of success  $p_{X(x)}$ ; hence they are not identically distributed as their distribution varies for different subpopulations.

## 4.1 Posterior probability of guilt

It follows from the above that we have  $c_{x,y} = p_x$  for all  $x, y \in X$ . Therefore, it follows from (3.5) and (3.7) that

$$\frac{P_s(G | I, E)}{P_s(G^c | I, E)} = \frac{1}{p_s} \frac{P_s(G | I)}{P_s(G^c | I)} = \frac{P_s(G^c)}{\sum_{i=1}^m p_i \beta_i - p_s P_s(G)} \frac{P_s(G)}{P_s(G^c)} \quad (4.3)$$

We can work this out in more detail in the case where  $P(C = x | C \in X(x)) = 1/N_x$ , and  $S$  and  $C$  are independent. First, we define  $\alpha_i$  to be the probability that  $C \in X_i$ , given that  $C$  has  $\Gamma$ :

$$\alpha_i = P(C \in X_i | I) = \frac{P(I | C \in X_i)P(C \in X_i)}{P(I)} = \frac{p_i \beta_i}{\sum_{j=1}^m p_j \beta_j}. \quad (4.4)$$

Now,  $P(C = x) = \alpha_x/N_x$  and  $P(C = x | I) = \beta_x/N_x$ , and (4.3) can be rewritten as

$$\frac{P_s(G | I, E)}{P_s(G^c | I, E)} = \frac{1}{p_s} \frac{\alpha_s}{N_s - \alpha_s} = \frac{1}{N_s \sum_{i=1}^m p_i \frac{\beta_i}{\beta_s} - p_s}. \quad (4.5)$$

## 4.2 Likelihood ratios

It follows from (4.5) that, whether  $S$  and  $C$  are independent or not, the likelihood ratio conditioned on  $I$  is always given by

$$\frac{P_s(E | G, I)}{P_s(E | G^c, I)} = \frac{1}{p_s}. \quad (4.6)$$

If we assume independence of  $S$  and  $C$  and that  $C$  restricted to each sub-population is uniform, then

$$\frac{P_s(I, E | G)}{P_s(I, E | G^c)} = \frac{N_s - \beta_s}{N_s \sum_{j=1}^m p_j \beta_j - p_s \beta_s}. \quad (4.7)$$

To get some feeling for this latter quantity, note that when  $N_s$  is large, (4.7) is approximately equal to

$$\frac{1}{\sum_{j=1}^m p_j \beta_j}; \quad (4.8)$$

this is independent of  $s$ . Hence, contrary to the classical case described in Section 2, the weight of the evidence depends on whether or not we consider the fact that the criminal has  $\Gamma$  to be evidence or background information. If we take into account that  $C$  has  $\Gamma$  when we formulate odds for  $G$ , then

only the improbability of finding  $\Gamma$  in  $S$  determines the posterior odds, and the weight of the evidence is  $1/p_s$ . If we do not take this into account and consider both  $I$  and  $E$  as evidence, then the weight of this evidence is given by (4.8) for large populations. In that case, it does not depend on the suspect's subpopulation but only on the weighted average  $\Gamma$ -frequency, where the weights are the a priori probabilities that the criminal belongs to the various subpopulations.

So, two likelihood ratios can be used to answer the informal question "What is the weight of the evidence that the suspect has the same characteristic as the criminal?". This articulates the fact that one should be very careful with the use of such likelihood ratios, and that one should primarily be interested in posterior odds rather than in likelihood ratios. A similar warning in a different situation can be found in [6] and [7].

### 4.3 Expected number of $\Gamma$ -bearers

If we choose  $\beta_i = N_i/N$  as we did in the classical case, then we can again express the posterior probability of guilt as the inverse of the expected number of  $\Gamma$ -bearers. We compute  $E_s(U | I, C \in X_s) = \sum_i N_i p_i + 1 - p_s = N_s \sum_i p_i \frac{N_i}{N_s} - p_s + 1$ , and from (4.5) it follows that

$$P_s(G | I, E) = \frac{1}{E_s(U | I, C \in X_s)}, \quad (4.9)$$

which is the analogue of (2.5). The reader may check that similarly,

$$P_s(G | I, E) = E_s(U^{-1} | I, E, C \in X_s).$$

This is the analogue of (2.9).

### 4.4 Without conditioning on $S = s$

Assume that  $S$  is uniformly distributed on  $X$ , and suppose we do not condition on  $\{S = s\}$ . Concentrating on the conditional probability of  $G$  we obtain

$$P(G | I, E) = \sum_{s \in X} P(G | I, E, S = s) P(S = s | I, E). \quad (4.10)$$

The first term in the summation is computed above already, so we need only to compute  $P(S = s | I, E)$ . Since information about  $S$  and its  $\Gamma$ -status does

not say anything about  $\Gamma_C$ , we have that

$$\begin{aligned} P(S = s \mid I, E) &= P(S = s \mid \Gamma_S = 1) \\ &= \frac{P(\Gamma_S = 1 \mid S = s)P(S = s)}{\sum_{s \in X} P(\Gamma_s = 1 \mid S = s)P(S = s)} \\ &= \frac{p_s}{\sum_{s \in X} p_s}. \end{aligned}$$

Hence it follows that

$$P(G \mid E, I) = \frac{\sum_{s \in X} p_s Z_s}{\sum_{s \in X} p_s},$$

where

$$Z_s = P_s(G \mid I, E) = \frac{\alpha_s}{p_s(N_s - \alpha_s) + \alpha_s}.$$

Hence the posterior probability of guilt is a weighted average of the conditioned ones, with weights  $p_s$ .

## 4.5 Conclusions

- The probability of guilt in this situation is equal to

$$P_s(G \mid I, E) = \frac{\alpha_s}{p_s(N_s - \alpha_s) + \alpha_s},$$

and this answer depends on  $s$  via the frequency of  $\Gamma$  in the subpopulation of  $s$ , the distribution of  $C$  and the size of the subpopulation of  $s$ . The sizes of the other subpopulations do not play a role other than in the assessment of the  $\beta_i$  and thereby of the  $\alpha_i$ , i.e., in the distribution of  $C$ .

- For the value of the likelihood ratio, it does matter whether or not  $I$  is interpreted as background information or evidence. For the probability of guilt this distinction is - of course - irrelevant.
- The probability of guilt, conditioning only on the fact that the suspect has  $\Gamma$  but not on the identity (subpopulation) of the suspect, is the weighted average of the individual conditional probabilities, with weight factors  $p_s$ . The sizes of the subpopulations and the distribution of  $C$  do not play a role in the weights.

## 5 Uncertainty about the frequency of $\Gamma$

In this section we assume that the  $\Gamma$ -frequency  $P(\Gamma_x = 1)$  is not known with certainty. Instead, we describe the frequency with a probability distribution.

### 5.1 Classical case

We assume that there are no subpopulations. The random variable  $C$  is uniform on  $X$ , and  $S$  and  $C$  are independent. To model the uncertainty of the  $\Gamma$ -frequency, we assume that there is a random variable  $W$ , taking values in  $[0, 1]$  and with density  $\chi$ , such that conditional on  $W = r$ , the  $\Gamma_x$  are independent Bernoulli variables with  $P(\Gamma_x = 1) = r$ . We let  $p$  denote the expectation of  $W$  and  $\sigma^2$  its variance. We again condition on  $S = s$  whenever we compute odds, but all results in this section are independent of  $s$ .

**Definition 5.1.** *The distribution of  $W$  is called the prior-to-crime distribution and the distribution of  $W$  conditioned on  $I$  is called the prior-to-suspect distribution. Finally, the distribution of  $W$  conditioned on both  $I$  and  $E$  is called the post-match distribution. The densities of these three random variables are denoted by  $\chi$ ,  $\chi_I$  and  $\chi_{I,E}$  respectively.*

Since

$$P(I) = \int_0^1 P(I | W = t)\chi(t)dt = \int_0^1 t\chi(t)dt = p,$$

the continuous version of Bayes' theorem implies that

$$\chi_I(t) = \frac{P(I | W = t)\chi(t)}{P(I)} = \frac{t}{p}\chi(t). \quad (5.1)$$

Furthermore, we have

$$\chi_{I,E}(t) = \frac{1 + Nt}{1 + N(p + \sigma^2/p)}\chi_I(t). \quad (5.2)$$

To see this, note that

$$\chi_{I,E}(t) = \frac{P(E | W = t, I)\chi_I(t)}{P(E | I)} \quad (5.3)$$

and compute the denominator:

$$P(E | I) = \int_0^1 P(E | W = t, I) \chi_I(t) dt \quad (5.4)$$

$$= \int_0^1 \frac{1 + Nt}{1 + N} \frac{t}{p} \chi(t) dt \quad (5.5)$$

$$= \frac{1}{(1 + N)p} (p + N(p^2 + \sigma^2)) \quad (5.6)$$

$$= \frac{1 + N(p + \frac{\sigma^2}{p})}{1 + N}. \quad (5.7)$$

From this, the claim readily follows.

The expectation of  $W$  given  $I$  is expressed in terms of  $\chi$  by

$$p' := E(W | I) = \int_0^1 t \chi_I(t) dt = \int_0^1 \frac{1}{p} t^2 \chi(t) dt = \frac{1}{p} (p^2 + \sigma^2). \quad (5.8)$$

The expected number of  $\Gamma$ -bearers, given  $I$  is now given by

$$E(U | I) = \int_0^1 E(U | I, W = t) \chi_I(t) dt = \int_0^1 (1 + Nt) \chi_I(t) dt = 1 + Np'. \quad (5.9)$$

As in the classical case where  $\sigma^2 = 0$  (cf. (2.5)), the inverse of this expression is equal to the posterior probability of guilt, since

$$P_s(G | I, E) = \int_0^1 P_s(G | I, E, W = t) \chi_{I,E}(t) dt \quad (5.10)$$

$$= \int_0^1 \frac{1}{1 + Nt} \frac{1 + Nt}{1 + N(p + \sigma^2/p)} \chi_I dt \quad (5.11)$$

$$= \frac{1}{1 + N(p + \sigma^2/p)}. \quad (5.12)$$

As in the classical case (cf. (2.8)), we also have  $P_s(G | I, E) = E_s(U^{-1} | I, E)$ . Indeed,

$$E_s(U^{-1} | I, E) = \int_0^1 E_s(U^{-1} | I, E, W = t) \chi_{I,E}(t) dt \quad (5.13)$$

$$= \int_0^1 \frac{1}{1 + Nt} \frac{1 + Nt}{1 + N(p + \sigma^2)} \chi_I(t) dt \quad (5.14)$$

$$= \frac{1}{1 + N(p + \sigma^2/p)}. \quad (5.15)$$

The expectation  $p'$  only depends on  $\chi$  and not on the population size. This is to be expected, since learning that a (randomly chosen) population member has  $\Gamma$  is not informative about the population size. This changes when we learn  $E$ , the fact that a randomly selected islander has  $\Gamma$  as well. Indeed, in a small population this is more likely to happen since we are more likely to accidentally select the criminal. In the extreme case where  $N = 0$ ,  $E$  can not offer any new information, but for other  $N$ , it does. It follows from (5.2) that

$$\begin{aligned} p'' &:= \mathbb{E}_s(W | I, E) = \int_0^1 t \chi_{I,E}(t) dt \\ &= \frac{1}{1 + N(p + \sigma^2/p)} \int_0^1 t(1 + Nt) \frac{t}{p} \chi(t) dt \\ &= \frac{1}{1 + N(p + \sigma^2/p)} \left( p + \frac{\sigma^2}{p} + \frac{N}{p} \int_0^1 t^3 \chi(t) dt \right). \end{aligned}$$

We can also write

$$p'' = \frac{1}{1 + Np'} (p' + N(\sigma_{W|I}^2 + p'^2)),$$

if we want to express  $p''$  in terms of  $\chi_I$ , where  $\sigma_{W|I}^2$  denotes the variance of  $\chi_I$ . The above formula can be rewritten as

$$p'' = p' \frac{1 + N(\sigma_{W|I}^2/p' + p')}{1 + Np'} \geq p',$$

with equality only if  $\sigma_{W|I}^2 = 0$  or  $N = 0$  (as expected, cf. the remark above).

It is perhaps worth mentioning that one can reconstruct  $\chi_I$  from  $\chi_{I,E}$  and  $\chi$  from  $\chi_I$ . Indeed we have

$$\chi_I(t) = \frac{\chi_{I,E}(t)}{1 + Nt} \left( \int_0^1 \frac{\chi_{I,E}(s) ds}{1 + Ns} \right)^{-1} \quad (5.16)$$

and

$$\chi(t) = \left( t \int_0^1 \frac{\chi_I(x)}{x} dx \right)^{-1} \chi_I(t). \quad (5.17)$$

To see this, note that from (5.2) we have

$$\chi_I(t) = \frac{1 + Np'}{1 + Nt} \chi_{I,E}(t). \quad (5.18)$$

On the other hand, it follows from (5.15) that

$$E_s(U^{-1} | I, E) = \int_0^1 \frac{1}{1 + Ns} \chi_{I,E}(s) ds = \frac{1}{1 + Np'},$$

and the first claim (5.16) follows.

For (5.17) we simply note from (5.1) that

$$\chi(t) = \frac{p}{t} \chi_I(t), \quad (5.19)$$

where  $p = \int_0^1 t \chi(t) dt$ . Integrating this equation gives

$$1 = p \int_0^1 \frac{\chi_I(t)}{t} dt = 1$$

and this expresses  $p$  in terms of  $\chi_I$ . Substituting this into (5.19) gives (5.17).

As a conclusion, we have seen that

$$p'' \geq p' \geq p,$$

so one has

$$\frac{1}{1 + Np''} \leq \frac{1}{1 + Np'} \leq \frac{1}{1 + Np}.$$

### 5.1.1 Conclusions

- The basic formula of Conclusions 2.4 still holds: using (5.9), (5.12) and (5.15), we see that the probability of guilt is given by

$$P_s(G | I, E) = E_s(U^{-1} | I, E) = E_s(U | I)^{-1} = \frac{1}{1 + Np'}.$$

- The conditional probability of guilt expressed in terms of  $\chi$  is

$$P_s(G | I, E) = \frac{1}{1 + N(p + \sigma^2/p)}. \quad (5.20)$$

Therefore, ignoring the uncertainty (i.e., using  $p$  instead of  $p'$ ), is unfavourable to the suspect. If, on the other hand, one incorrectly assumes that there is uncertainty, then this is favourable to the suspect.

- The conditional probability of guilt expressed in terms of  $\chi_I$  is

$$P_s(G | I, E) = \frac{1}{1 + Np'}. \quad (5.21)$$

In this case, the uncertainty in  $\chi_I$  is irrelevant in the sense that its variance plays no role.

- The conditional probability of guilt expressed in terms of  $\chi_{I,E}$  is

$$P_s(G | I, E) = \int_0^1 \frac{1}{1 + Nt} \chi_{I,E}(t) dt. \quad (5.22)$$

Ignoring the uncertainty in  $\chi_{I,E}$  (obtaining  $P_s(G | I, E) = 1/(1 + Np'')$ ) would be favourable to the suspect.

## 5.2 Uncertainty about the criminal's subpopulation

Suppose that, as in Section 4, the population is divided into subpopulations  $X = X_1 \cup \dots \cup X_m$ , and that  $C$  has characteristic  $\Gamma$ . We let  $W_i$  be the random variable modelling the frequency of  $\Gamma$  in  $X_i$ . The expectation resp. variance of  $W_i$  are denoted by  $p_i$  resp.  $\sigma_i^2$ . So, if  $X(x) \neq X(y)$  then  $\Gamma_x$  and  $\Gamma_y$  are independent, and furthermore conditional on  $W_i = p_i$  the  $\Gamma_x$  for  $x \in X_i$  are independent Bernoulli variables with probability of success  $p_i$ . We write  $E = E_s$  and  $G = G_s$  as before.

### 5.2.1 Unconditioned on $I$

We first interpret  $I$  as evidence, not as background information. The posterior probability of guilt, given that  $S = s \in X_i$  and  $C$  has  $\Gamma$ , is

$$P_s(G | I, E) = P_s(G | C \in X_i, I, E) P_s(C \in X_i | I, E).$$

The first term in the RHS equals (see (5.12))

$$P_s(G | C \in X_i, I, E) = \frac{1}{1 + (N_i - 1)(p_i + \sigma_i^2/p)},$$

since we are now back in the setting of a homogeneous population. The second term equals (cf. (5.7),(4.4))

$$\begin{aligned} P_s(C \in X_i | I, E) &= \frac{P_s(E | C \in X_i, I) P_s(C \in X_i | I)}{P_s(E | I)}, \\ &= \frac{\frac{1 + (N_i - 1)(p_i + \sigma_i^2/p_i)}{N_i} \frac{p_i \beta_i}{\sum_{j=1}^m p_j \beta_j}}{P_s(E | I)}. \end{aligned}$$

It remains to compute  $P_s(E | I)$ :

$$\begin{aligned}
P_s(E | I) &= \sum_{j=1}^m P_s(E | C \in X_j) P_s(C \in X_j | I) \\
&= \sum_{j=1, j \neq i}^m p_i \frac{p_j \beta_j}{\sum_{k=1}^m p_k \beta_k} + \frac{1 + (N_i - 1)(p_i + \sigma_i^2/p_i)}{N_i} \frac{p_i \beta_i}{\sum_{k=1}^m p_k \beta_k} \\
&= \frac{N_i \sum_{j=1, j \neq i}^m p_i \beta_j p_j + (1 + (N_i - 1)(p_i + \sigma_i^2/p_i)) \beta_i p_i}{N_i \sum_{k=1}^m \beta_k p_k}.
\end{aligned}$$

Putting the parts together yields

$$P_s(G | I, E) = \frac{1}{1 + N_i \sum_{j=1}^m \frac{\beta_j}{\beta_i} p_j + (N_i - 1) \frac{\sigma_i^2}{p_i} - p_i}. \quad (5.23)$$

This is the analogue of (4.5). For large  $N_i$ , the probability of guilt is roughly equal to

$$P_s(G | I, E) \approx \frac{1}{N_i (\sum_{j=1}^m \frac{\beta_j}{\beta_i} p_j + \sigma_i^2/p_i)}. \quad (5.24)$$

The odds on guilt are then roughly equal to

$$\frac{P_s(G | I, E)}{P_s(G^c | I, E)} = \frac{1}{N_i (\sum_{j=1}^m \frac{\beta_j}{\beta_i} p_j + \frac{\sigma_i^2}{p_i})} = \frac{\beta_i}{N_i} \frac{1}{\sum_{j=1}^m \beta_j p_j + \beta_i \frac{\sigma_i^2}{p_i}}. \quad (5.25)$$

The weight of the evidence, the likelihood ratio, is given by

$$\frac{P_s(E, I | G)}{P_s(E, I | G^c)} = \frac{N_i - \beta_i}{N_i \sum_{j=1}^m p_j \beta_j + \frac{\sigma_i^2}{p_i} \beta_i (N_i - 1) - p_i \beta_i},$$

which for large populations is roughly equal to

$$\frac{1}{\sum_{j=1}^m \beta_j p_j + \beta_i \sigma_i^2/p_i}, \quad (5.26)$$

as is also clear from (5.25). This should be compared to (4.8).

### 5.2.2 Conditional on $I$

Now,  $I$  is interpreted as background information. Let, as in (4.4),

$$\alpha_i = P_s(C \in X_i | I) = \frac{p_i \beta_i}{\sum_{k=1}^m p_k \beta_k}.$$

Then (5.23) can be rewritten as

$$P_s(G | I, E) = \frac{1}{\sum_{j \neq i} N_j p_j \frac{\alpha_j}{\alpha_i} + 1 + (N_i - 1)p'_i},$$

where

$$p'_i = p_i + \frac{\sigma_i^2}{p_i}$$

is the expectation of  $W_i$  given  $C \in X_i$ .

Since the prior odds, conditional on  $I$ , in favour of guilt of  $x \in X_i$  are

$$\frac{P(G | I)}{P_s(G^c | I)} = \frac{\alpha_i}{N_i - \alpha_i},$$

the corresponding likelihood ratio is equal to

$$\begin{aligned} \frac{P_s(E | G, I)}{P_s(E | G^c, I)} &= \frac{N_i - \alpha_i}{N_i p_i (1 - \alpha_i) + \alpha_i (1 + (N_i - 1)p'_i)} \\ &= \frac{N_i - \alpha_i}{N_i \alpha_i (p'_i - p_i) + N_i p'_i \alpha_i - p'_i \alpha_i}. \end{aligned}$$

Of course, this leads to the same posterior probability of guilt as in (5.24). Notice that when  $p'_i = p_i$ , then this reduces to  $1/p_i$ , i.e., we retrieve (4.6).

For large  $N_i$ , the likelihood ratio is roughly equal to

$$\frac{1}{p_i + \alpha_i (p'_i - p_i)} = \frac{p_i}{p_i^2 + \alpha_i \sigma_i^2}. \quad (5.27)$$

### 5.2.3 Conclusions

- As in the case without uncertainty about the  $\Gamma$ -frequencies, we obtain two likelihood ratios that quantify the weight of the evidence: for large populations these are (5.26) if the evidence is taken to be  $(I, E)$  and (5.27) if the evidence is taken to be only  $E$ .
- Only the uncertainty about the frequency of  $\Gamma$  in the suspect's subpopulation plays a role in the likelihood ratio and the posterior probability of guilt, the uncertainty in the other subpopulations does not. The effect of this uncertainty is weighted by the probability that the true culprit belongs to this subpopulation.
- As in the classical case, if one conditions on  $I$  then the likelihood ratio given by (5.27) for large populations, only contains quantities associated to the suspect's subpopulation.

- Contrary to the classical case, if one considers the evidence to be  $(I, E)$  then the likelihood ratio (given by (5.26)) depends on the suspect's subpopulation, through the uncertainty about the  $\Gamma$ -frequency in this population.
- Regardless of whether one lets the evidence be  $I, E$  or only  $E$ , the greater the uncertainty, the lower the weight of the evidence.

### 5.3 Uncertainty about the suspect's and the criminal's subpopulation

Suppose now that it is also unknown to which subpopulation  $s$  belongs. In that case we can no longer condition on  $S = s$ , but we can use the results of the previous section by writing

$$P(G | I, E) = \sum_{i=1}^m P(G | S \in X_i, I, E)P(S \in X_i | I, E). \quad (5.28)$$

We have determined the  $P(G | S = s, I, E)$  in (5.23), and it is not difficult to see that this is equal to  $P(G | S \in X_i, I, E)$  whenever  $s \in X_i$ . Hence, we only need to compute  $P(s \in X_i | I, E)$ .

The distribution of  $S$  plays a role now, and we define

$$\epsilon_i = P(s \in X_i)$$

to be the probability that  $S$  belongs to  $X_i$ . Then the a priori probability of guilt is

$$P(G) = P(C = s) = \sum_{i=1}^m P(s \in X_i)P(C = s | s \in X_i) = \sum_{i=1}^m \epsilon_i \frac{\beta_i}{N_i}.$$

Recall that  $\beta_i$  is the probability that  $C \in X_i$  and that we assume a uniform distribution over each subpopulation.

We now compute  $P(S \in X_i | I, E)$ :

$$\begin{aligned} P(S \in X_i | I, E) &= \frac{P(E | S \in X_i, I)P(S \in X_i | I)}{P(E | I)} \\ &= \frac{P(E | S \in X_i, I)P(S \in X_i | I)}{\sum_{j=1}^m P(E | S \in X_j, I)P(S \in X_j | I)}. \end{aligned} \quad (5.29)$$

It remains to compute  $P(E | S \in X_j, I)$  and  $P(S \in X_i | I)$ . The latter is easy: since  $I$  is information about  $C$  and not about  $S$ , we have

$$P(S \in X_i | I) = \epsilon_i.$$

The former can be computed as follows:

$$P(E | S \in X_i, I) = \sum_{j=1}^m P(E | C \in X_j, S \in X_i, I) P(C \in X_j | S \in X_i, I).$$

Now  $P(C \in X_j | S \in X_i, I)$  is the probability that  $C$  belongs to  $X_j$ , given that  $S$  has been selected from  $X_i$  and that  $C$  has  $\Gamma$ . However, nothing is given about  $S$ 's  $\Gamma$ -status and therefore  $S \in X_i$  can not be informative about  $C$  at all, hence

$$P(C \in X_j | S \in X_i, I) = P(C \in X_j | I) = \frac{p_j \beta_j}{\sum_{k=1}^m p_k \beta_k}.$$

It remains to evaluate the terms  $P(E | C \in X_j, S \in X_i, I)$ . If  $i \neq j$  then  $S$  and  $C$  belong to different populations. If  $i = j$  then (5.7) applies, so

$$P(E | S \in X_i, C \in X_j, I) = \begin{cases} p_i & i \neq j, \\ \frac{1 + (N_i - 1)(p_i + \sigma_i^2/p_i)}{N_i} & i = j. \end{cases}$$

If we put these ingredients together, we obtain after some computations:

$$P(E | S \in X_i, I) = \frac{p_i}{\sum_{k=1}^m p_k \beta_k} \left( \sum_{j=1}^m p_j \beta_j + \frac{\beta_i}{N_i} (1 - p_i + (N_i - 1) \sigma_i^2 / p_i) \right).$$

Plugging this into (5.29), we obtain

$$\begin{aligned} P(S \in X_i | I, E) &= \frac{p_i (\sum_{k=1}^m p_k \beta_k + \frac{\beta_i}{N_i} (1 - p_i + (N_i - 1) \sigma_i^2 / p_i)) \epsilon_i}{\sum_{j=1}^m p_j (\sum_{k=1}^m p_k \beta_k + \frac{\beta_j}{N_j} (1 - p_j + (N_j - 1) \sigma_j^2 / p_j)) \epsilon_j} \\ &= \frac{p_i \epsilon_i \beta_i}{N_i P_s(G | I, E, S \in X_i)} \frac{1}{\sum_{j=1}^m \frac{p_j \epsilon_j \beta_j}{N_j P_s(G | I, E, S \in X_j)}} \quad (5.30) \end{aligned}$$

Substituting this expression into (5.28), we arrive at the posterior probability of guilt:

$$P(G | I, E) = \frac{\sum_{i=1}^m \frac{p_i \epsilon_i \beta_i}{N_i}}{\sum_{i=1}^m \frac{p_i \epsilon_i \beta_i}{N_i P_s(G | I, E, S \in X_i)}}. \quad (5.31)$$

Although this is not immediately obvious in the above presentation, the expression (5.31) is symmetric in  $\epsilon$  and  $\beta$ . To show this, notice that we only

have to prove it for the denominator. Denoting

$$\begin{aligned} f(\epsilon, \beta) &= \sum_{i=1}^m \frac{p_i \epsilon_i \beta_i}{N_i P_s(G | I, E, S \in X_i)} \\ &= \sum_{i=1}^m \frac{p_i \beta_i \epsilon_i}{N_i} \left( 1 + N_i \sum_{j=1}^m p_j \beta_j / \beta_i + (N_i - 1) \sigma_i^2 / p_i - p_i \right), \end{aligned}$$

we compute

$$\begin{aligned} f(\epsilon, \beta) - f(\beta, \epsilon) &= \sum_{i=1}^m p_i \beta_i \left( \sum_{j=1}^m p_j \beta_j / \beta_i - \sum_{j=1}^m p_j \epsilon_j / \epsilon_i \right) \\ &= \sum_{i=1}^m p_i \left( \epsilon_i \sum_{j=1}^m p_j \beta_j - \beta_i \sum_{j=1}^m p_j \epsilon_j \right) \\ &= \sum_{i,j=1}^m (p_i \epsilon_i p_j \beta_j - p_i \beta_i p_j \epsilon_j) \\ &= 0. \end{aligned}$$

Intuitively, it is clear that (5.31) must possess this symmetry. Indeed, we have an unknown criminal  $C$  and a suspect  $S$ , both with  $\Gamma$ . The probability that  $S = C$  depends, as far as  $\epsilon$  and  $\beta$  are concerned, on how they allow for  $S$  and  $C$  to be issued from the same subpopulation. Exchanging the distributions  $\epsilon$  and  $\beta$  should not make a difference.

To conclude this section we sketch the behaviour of (5.30) in extreme situations.

### 5.3.1 Probability that $S \in X_i$ for extreme situations

- If all  $\sigma_j^2 = 0$  and the  $N_j$  are very large (compared to the  $p_j^{-1}$ ), then (5.30) is approximately equal to

$$P(S \in X_i | I, E) \approx \frac{p_i \epsilon_i}{\sum_{j=1}^m p_j \epsilon_j} = P(S \in X_i | E).$$

This is reasonable, since if  $p_i$  is big compared to  $1/N_i$ , then it is very unlikely that  $C = S$  even when  $\Gamma$  is taken into account. In this case, knowing that  $C$  has  $\Gamma$  does not really alter our belief about  $S$ 's subpopulation which we have based on  $E$ .

- If all  $\sigma_j^2 = 0$  and the  $p_j$  are small compared to the  $1/N_j$ , then

$$P(S \in X_i | I, E) \approx \frac{p_i \frac{\beta_i}{N_i} \epsilon_i}{\sum_{j=1}^m p_j \frac{\beta_j}{N_j} \epsilon_j} = \frac{1}{\sum_{j=1}^m \frac{p_j \beta_j N_i \epsilon_j}{p_i \beta_i N_j \epsilon_i}}. \quad (5.32)$$

If  $\epsilon_i = N_i/N$ , then (5.32) reduces to

$$P(S \in X_i | I, E) \approx \frac{p_i \beta_i}{\sum_{i=1}^m p_j \beta_j} = P(C \in X_i | I), \quad (5.33)$$

which is also reasonable, since for very small  $\Gamma$ -frequencies it is quite likely that  $C = S$ .

- If also  $\beta_i = N_i/N$ , then (5.32) reduces to

$$P(S \in X_i | I, E) \approx \frac{N_i p_i}{\sum_{i=1}^m N_j p_j}, \quad (5.34)$$

and this is also understandable: if there is no information about the identity of  $C$  or  $S$ , then the probability that  $S \in X_i$  is proportional to the expected number of  $\Gamma$ -bearers in that subpopulation.

## 6 Database search

In this section we suppose that there is a database  $\mathcal{D} \subset X$  containing the  $\Gamma$ -status of individuals  $x_1, \dots, x_n$ . After possibly renumbering, we write  $X = \{x_1, \dots, x_{N+1}\}$  and let  $\mathcal{D} = \{x_1, \dots, x_n\}$ . Suppose that  $\sum_{d \in \mathcal{D}} \Gamma_d = k$ , that is, there are  $k$  matches in the database. Let the evidence  $E_{\mathcal{D}}$  be given by

$$E_{\mathcal{D}} = \{\Gamma_{x_1} = \dots = \Gamma_{x_k} = 1, \Gamma_{x_{k+1}} = \dots = \Gamma_{x_n} = 0\}.$$

We also assume that  $P(C = x_i | I) = \alpha_i$  and that each individual has  $\Gamma$  with probability  $p$ .

There are several pairs of propositions whose support by the data can be considered. These propositions all give rise to their own likelihood ratios or posterior probabilities, which has caused considerable confusion in the literature; see [6] for an account on this. Some of the forthcoming discussion also appears in [6] but we recall it here for completeness.

We will discuss three ways of looking at database matches. The most interesting case is where the database search produces a single match. Indeed, if there are no matches then the inquiry comes to an end as far as the database is concerned and if there are several matches, then it is clear that chance matches have occurred:

1. Database-focused: in this case, the quantity of interest is  $P(C \in \mathcal{D} | E_{\mathcal{D}}, I)$ , the probability that the criminal is in the database;

2. Individual-focused: in this case, the quantity of interest is  $P(C = x_1 | E_{\mathcal{D}}, I)$ , the conditional probability that  $C = x_1$  supposing that  $x_1$  has  $\Gamma$ ;
3. Database effectiveness: in this case, the quantity of interest is  $P(S = C | E_1, I)$ , the probability that  $S = C$  where  $E_1$  denotes the event that  $k = 1$  (a unique match, but not specified with whom), and where  $S$  is the label of the matching individual.

## 6.1 Database-focused

First, we consider the proposition, found e.g. in [8],

$$C \in \mathcal{D},$$

and its negation  $C \notin \mathcal{D}$ . The prior odds in favour of  $C \in \mathcal{D}$  are

$$\frac{P(C \in \mathcal{D} | I)}{P(C \notin \mathcal{D} | I)} = \frac{\alpha_1 + \cdots + \alpha_n}{\alpha_{n+1} + \cdots + \alpha_{N+1}},$$

where  $\alpha_i = P(C = x_i | I)$  is the probability of guilt of  $x_i$ , given that  $C$  has  $\Gamma$ . Clearly,

$$P(E_{\mathcal{D}} | C \notin \mathcal{D}, I) = p^k(1 - p)^{n-k}.$$

Similarly, it is easy to see that

$$P(E_{\mathcal{D}} | C \in \mathcal{D}, I) = \frac{p^{k-1}(1 - p)^{n-k}(\alpha_1 + \cdots + \alpha_k)}{\alpha_1 + \cdots + \alpha_n},$$

and therefore the likelihood ratio of evidence  $E_{\mathcal{D}}$  in favour of  $C \in \mathcal{D}$  is equal to

$$\frac{P(E_{\mathcal{D}} | C \in \mathcal{D}, I)}{P(E_{\mathcal{D}} | C \notin \mathcal{D}, I)} = \frac{\alpha_1 + \cdots + \alpha_k}{p(\alpha_1 + \cdots + \alpha_n)}. \quad (6.1)$$

The posterior odds in favour of  $C \in \mathcal{D}$  are

$$\frac{P(C \in \mathcal{D} | E_{\mathcal{D}}, I)}{P(C \notin \mathcal{D} | E_{\mathcal{D}}, I)} = \frac{\alpha_1 + \cdots + \alpha_k}{p(\alpha_{n+1} + \cdots + \alpha_{N+1})}. \quad (6.2)$$

If  $k = 1$ ,  $C \in \mathcal{D}$  becomes logically equivalent to  $C = x_1$ , and we have

$$\frac{P(C = x_1 | E_{\mathcal{D}}, I)}{P(C \neq x_1 | E_{\mathcal{D}}, I)} = \frac{\alpha_1}{p(\alpha_{n+1} + \cdots + \alpha_{N+1})} = \frac{1}{p} \frac{P(C = x_1 | I)}{P(C \notin \mathcal{D} | I)}. \quad (6.3)$$

In fact, it is not difficult to show that (6.3) also holds when the probability of having  $\Gamma$  differs among the individuals in the database. In that case,  $p$  in (6.3) should be replaced with  $p_1 = P(\Gamma_{x_1} = 1 | I)$ . Therefore, the weight of the evidence is not influenced by the presence in the database of people of different ethnic origin other than by the determination of the  $\alpha_i$ .

## 6.2 Individual-focused

Of course, the proposition  $C \in \mathcal{D}$  is not really of interest to a court. Rather, presented with an individual  $x$  such that  $\Gamma_x = 1$ , a court is interested in  $P(C = x \mid E_{\mathcal{D}}, I)$ . Therefore, suppose as above that there are  $k$  hits in the database, namely  $x_1, \dots, x_k$ . A computation analogous to the above one shows that the posterior odds in favour of  $C = x_1$  are

$$\frac{P(C = x_1 \mid E_{\mathcal{D}}, I)}{P(C \neq x_1 \mid E_{\mathcal{D}}, I)} = \frac{\alpha_1}{\alpha_2 + \dots + \alpha_k + p(\alpha_{n+1} + \dots + \alpha_{N+1})}. \quad (6.4)$$

Notice that, if  $k = 1$ , we retrieve (6.3), as we should.

## 6.3 Database effectiveness

The most interesting case is when the database produces a unique hit. In that case, as we have seen, the posterior odds in favour of  $S = C$  are given by (6.3). In this section we investigate a related, but different probability, namely the probability that if we have a unique database hit, that it is with the true culprit. This probability represents the long term effectiveness of the database in selecting the correct individual in the cases where it produces a unique match. We let  $E_1$  denote the event that there is exactly one  $\Gamma$ -bearer in the database, and we will calculate

$$P(S = C \mid E_1, I),$$

where  $S$  is the unique individual in the database with  $\Gamma$ . To do so, we write

$$P(S = C \mid E_1, I) = \sum_{i=1}^n P(S = C \mid \Gamma_{x_i} = 1, E_1, I)P(\Gamma_{x_i} = 1 \mid E_1, I).$$

First notice that (6.3) gives

$$P(S = C \mid E_1, \Gamma_{x_i} = 1, I) = \frac{\alpha_i}{\alpha_i + pP(C \notin \mathcal{D} \mid I)},$$

and it remains to compute  $P(\Gamma_{x_i} = 1 \mid E_1, I)$ :

$$\begin{aligned} P(\Gamma_{x_i} = 1 \mid E_1, I) &= P(C = x_i \mid E_1, I) + \frac{1}{n}P(C \notin \mathcal{D} \mid E_1, I) \\ &= \frac{P(E_1 \mid C = x_i, I)P(C = x_i \mid I) + \frac{1}{n}P(E_1 \mid C \notin \mathcal{D}, I)P(C \notin \mathcal{D} \mid I)}{P(E_1 \mid I)} \\ &= \frac{P(E_1 \mid C = x_i, I)P(C = x_i \mid I) + \frac{1}{n}P(E_1 \mid C \notin \mathcal{D}, I)P(C \notin \mathcal{D} \mid I)}{P(E_1 \mid C \in \mathcal{D}, I)P(C \in \mathcal{D} \mid I) + P(E_1 \mid C \notin \mathcal{D}, I)P(C \notin \mathcal{D} \mid I)} \\ &= \frac{\alpha_i + pP(C \notin \mathcal{D} \mid I)}{P(C \in \mathcal{D} \mid I) + npP(C \notin \mathcal{D} \mid I)}, \end{aligned}$$

where in the last step we used that  $P(E_1 | C \in \mathcal{D}, I) = (1 - p)^{n-1}$  and  $P(E_1 | C \notin \mathcal{D}, I) = p(1 - p)^{n-1}$ . It follows that

$$\begin{aligned} P(S = C | E_1, I) &= \sum_{i=1}^n \frac{\alpha_i}{\alpha_i + pP(C \notin \mathcal{D} | I)} \frac{\alpha_i + pP(C \notin \mathcal{D} | I)}{P(C \in \mathcal{D} | I) + npP(C \notin \mathcal{D} | I)} \\ &= \frac{P(C \in \mathcal{D} | I)}{P(C \in \mathcal{D} | I) + npP(C \notin \mathcal{D} | I)}. \end{aligned}$$

which can also be written in odds form:

$$\frac{P(S = C | E_1, I)}{P(S \neq C | E_1, I)} = \frac{1}{np} \frac{P(C \in \mathcal{D} | I)}{P(C \notin \mathcal{D} | I)}. \quad (6.5)$$

If the database is comprised of individuals coming from different subpopulations, then (6.5) does not hold. However, in that case one may view the database as a disjoint union  $\mathcal{D} = \mathcal{D}_1 \cup \dots \cup \mathcal{D}_m$ , where  $\mathcal{D}_m$  is the subset of  $\mathcal{D}$  containing individuals from subpopulation  $i$ . For each of these separately, (6.5) holds.

It is rather interesting to see what happens with the odds on  $S = C$  (given  $E_1$  and  $I$ ) when the size of the database grows. It may seem from (6.5) that as  $n$  grows, the odds on  $S = C$  decrease. However, this is not true in general, since  $P(C \in \mathcal{D} | I)$  may also depend on  $n$ . It does, however, mean that enlarging a database does not necessarily improve its effectiveness, in the sense of increasing the odds (6.5) on a unique match being with the true offender. For example, suppose that a database  $\mathcal{D}_n$  of size  $n$  yields  $P(C \in \mathcal{D}_n | I) = q_n$ , and that a larger database  $\mathcal{D}_{2n}$  of size  $2n$  yields  $P(C \in \mathcal{D}_{2n} | I) = q_{2n}$ . If  $\mathcal{D}_n \subset \mathcal{D}_{2n}$  then naturally  $q_{2n} \geq q_n$ , but the probability that  $S = C$  given a unique match in  $\mathcal{D}_{2n}$  is greater than the probability that  $S = C$  given a unique match in  $\mathcal{D}_n$  only when

$$\frac{q_{2n}}{1 - q_{2n}} > 2 \frac{q_n}{1 - q_n}.$$

This can be explained intuitively: if one adds many people who are unlikely to be  $C$  to the database, then the probability of a chance match with one of these new individuals outweighs the fact that the probability that  $C$  has been added to the database has increased in the sense that it becomes less likely that a unique match actually is a match with the criminal.

Hence the value of a unique match may increase or decrease with the size of the database, and it is not hard to see that the probability of a unique match itself may (independently) decrease or increase.

## 6.4 Conclusions

- If it is known with whom the match is, say with  $x_i$ , then (cf. (6.3)) the posterior probability of guilt is given by

$$\frac{P(C = x_i | E_{\mathcal{D}}, I)}{P(C \neq x_i | E_{\mathcal{D}}, I)} = \frac{\alpha_i}{pP(C \notin \mathcal{D} | I)}.$$

Notice that this quantity only depends on  $\alpha_i = P(C = x_i | I)$ , on the likelihood  $p$  of a chance match with  $x_i$  and on the a priori probability that the database contains the criminal. As the database increases,  $P(C \notin \mathcal{D} | I)$  decreases but depending on  $\alpha_i/p$  the posterior probability  $P(C = x_i | E_{\mathcal{D}}, I)$  may be greater or smaller than for a smaller database.

- If it is not specified with which individual the match is, and the probability of having  $\Gamma$  is  $p$  for everyone in the database, then the posterior probability that the match is with the criminal is given by, cf. (6.5),

$$\frac{P(S = C | E_1, I)}{P(S \neq C | E_1, I)} = \frac{1}{np} \frac{P(C \in \mathcal{D} | I)}{P(C \notin \mathcal{D} | I)}.$$

These odds describe the long-term behaviour of the database, i.e., the proportion in the long run of unique matches that are matches with the true criminal. Naturally, enlarging the database always increases the probability that the criminal is contained in it. But the probability of a unique match may increase or decrease, and (independently) the value of a unique match may increase or decrease. In many cases, in an enlarged database the probability of a unique match increases, but the probability of a unique match being with the true offender decreases.

## 7 Examples

In this section we illustrate the obtained results by considering some examples. We have chosen to cast most of these examples in a DNA-setting, as this provides one of the few types of forensic evidence that are so well understood that more or less exact computations can be performed.

The uncertainty surrounding DNA-profile frequency estimates depends on the size of the database from which allele frequencies are estimated. A possible model is to define a prior distribution of allele frequencies, and to update this distribution with the database to obtain a posterior distribution. An often used approach is to use Dirichlet distributions (see [9] for an account

of the method and a discussion on the sensitivity for the choice of prior). Doing this for a database containing alleles of 230 persons (for many forensic labs the actual size of their database is a few hundred individuals), it seems (based on simulations for DNA-profiles with six or seven loci and frequencies between  $10^{-10}$  and  $10^{-7}$ ) reasonable to use a standard deviation  $p/3 \leq \sigma \leq 2p/3$  in the below examples.

We will in each example freely use the notation introduced in the section that it illustrates.

## 7.1 Classical island problem with uncertain $\Gamma$ -frequency

We start with the simple version of a homogeneous population  $X$  of size  $N + 1$  and profile frequency  $p$ . As we have seen (cf. (5.20) and (5.21)), the posterior probability of guilt is equal to  $P_s(G | I, E) = 1/(1 + N(p + \sigma^2/p)) = 1/(1 + Np')$ . With  $p/3 \leq \sigma \leq p$ , we get  $p' \in [9p/8, 13p/9]$ . Thus, the effect of the uncertainty about  $p$  is to effectively increase  $p$ , or equivalently, to decrease the likelihood ratio associated to  $I, E$  or to  $E$ . It may be prudent to use  $\sigma = p$ . For example, with  $N = 10^7, p = 10^{-8}, \sigma = p$ , we have  $P_s(G | I, E) = 83\%$  instead of 91%.

## 7.2 Subpopulations and Likelihood Ratios

We now illustrate the results of Section 4. Suppose that a crime has been committed in a heterogeneous population  $X = X_1 \cup X_2$ , with  $N_1 = 10^7$  and  $N_2 = 10^5$ . Prior to DNA-analysis it is estimated that the crime could equally probably have been committed by a member of  $X_1$  as by a member of  $X_2$ , i.e.,  $\beta_1 = \beta_2 = 0.5$ . Now a DNA-trace of the criminal is found, giving rise to a profile  $\Gamma$ . The forensic lab calculates  $p_1 = 10^{-9}$  and  $p_2 = 10^{-8}$ .

### 7.2.1 Unconditioned on the profile

The likelihood ratio (4.8) (taking both the fact that the criminal and the suspect have  $\Gamma$  as evidence) equals  $1/(p_1\beta_1 + p_2\beta_2) = 1.8 \times 10^8$ . This likelihood ratio holds for any suspect  $s$ , as long as  $S$  is independent of  $C$ .

With this likelihood ratio we obtain, for  $s \in X_1$ , posterior odds in favour of guilt equal to

$$(p_1\beta_1 + p_2\beta_2)^{-1}\beta_1/N_1 \approx 9,$$

corresponding to (cf. (4.5))  $P_s(G | I, E) = 90\%$ . For  $s \in X_2$  the posterior odds are

$$(p_1\beta_1 + p_2\beta_2)^{-1}\beta_2/N_2 \approx 910,$$

such that  $P_s(G | I, E) = 99,9\%$ .

### 7.2.2 Conditional on the profile

Given the fact that  $C$  has  $\Gamma$  and the frequencies  $p_1, p_2$ , we can also first calculate  $P(C \in X_i | I) = \alpha_i$ . This gives  $\alpha_1 = 9\%$  and  $\alpha_2 = 91\%$ : since the profile  $\Gamma$  is rarer in  $X_1$ , it is much more likely that the criminal is from  $X_2$ . The odds on  $C$  belonging to  $X_1$  are  $\alpha_1/\alpha_2 = 10$ . If this is taken as information relative to which everything else is conditioned, then the likelihood ratio associated to having  $\Gamma$ , is the inverse random match probability for the suspect:  $1/p_1$  or  $1/p_2$ . This gives rise to the same  $P_s(G | I, E)$ : if  $s \in X_1$  then the posterior odds are

$$p_1^{-1}\alpha_1/(N_1 - \alpha_1) \approx \alpha_1/(N_1 p_1) = 0.09/(10^{-9}10^7) = 9,$$

as above. Similarly, for  $s \in X_2$ , we get posterior odds

$$p_2^{-1}\alpha_2/(N_2 - \alpha_2) \approx \alpha_2/(N_2 p_2) = 0.91/(10^{-8}10^5) = 910,$$

as above.

### 7.2.3 Consequences of errors

When statements are made regarding the subpopulation to which  $C$  belongs, one has to be careful to note whether or not  $I$  has been taken into account. Indeed taking  $\alpha_i$  equal to  $\beta_i$ , that is,  $\alpha_1 = \alpha_2 = 0.5$ , we overestimate posterior odds in favour of guilt with a factor 10 for suspects from  $X_1$  and underestimate them with the same factor for suspects from  $X_2$ . This is a serious overestimate of the actual odds for suspects from  $X_1$ . In this example, it leads to a posterior probability of guilt of 98% (instead of 90%).

Finally, we note that if the forensic lab assumes  $p_2 = p_1 = 10^{-9}$  for both populations, e.g. because it always uses the population frequencies of the dominant population  $X_1$ , then we arrive at  $\alpha_i = \beta_i$ . The posterior odds in favour of guilt will in that case be calculated to be  $p_1^{-1}\alpha_i/(N_i - \alpha_i) \approx \alpha_i/(p_1 N_i)$  for  $s \in X_i$ . In this example, these odds are 50 for  $s \in X_1$  and 5000 for  $s \in X_2$  which is an overestimate in both cases.

## 7.3 Subpopulations: general case

We next illustrate the results that we have obtained for the case where the populations is heterogeneous w.r.t.  $\Gamma$ -probability, and there is uncertainty about the profile frequency in each population, as well as uncertainty about the subpopulation to which an individual belongs. This is described in section 5.3. Since there are many parameters that can be varied, we will keep some of them fixed throughout. We assume that the population consists of three

disjoint subpopulations  $X_1, X_2, X_3$ , where  $X_1$  is the dominant one, and the others are much smaller. We set  $N_1 = 20 \cdot 10^6, N_2 = 10^6, N_3 = 10^5$  and  $\sigma = p/2$ . We will compare the true posterior probability of guilt  $P(G | I, E)$  with the probability obtained assuming that for  $X_2, X_3$  the same  $\Gamma$ -frequency  $p_1$  is used as for  $X_1$ . This allows one to judge what the consequences are of having a subpopulation without knowing so. For example, there may be a region of the country with a relatively high  $\Gamma$ -frequency due to its relative isolation in the past. In practice it can be difficult to say with certainty if a given individual belongs to that subpopulation.

We compute for several choices of  $p_i, \epsilon_i$  and  $\beta_i$  the true probability of guilt and compare it to what one would obtain if  $p_2, p_3$  would be ignored, namely (5.20) with  $N + 1 = N_1 + N_2 + N_3$  and  $p = p_1$ . We denote this result with  $P^{\text{hom}}(G | I, E)$  and call it the *naive* probability of guilt.

**Example 7.1.** Let  $p_1 = 10^{-8}, p_2 = 10^{-7}, p_3 = 10^{-6}$ . We keep the  $\epsilon_i$  fixed to a choice where it is 90% certain that  $S \in X_1$ , not knowing  $I$  or  $E$ . The results are summarized in Table 1. Notice that the true probability of guilt may be

Table 1: Guilt probabilities for  $p_1 = 10^{-8}, p_2 = 10^{-7}, p_3 = 10^{-6}$

$(\epsilon_1, \epsilon_2, \epsilon_3)$	$(\beta_1, \beta_2, \beta_3)$	$P(G I, E)$	$P^{\text{hom}}(G I, E)$
(0.9,0.05,0.05)	(0.999,0.0005,0.0005)	0.50	0.79
(0.9,0.05,0.05)	uniform	0.70	0.79
(0.9,0.05,0.05)	(0.99,0.005,0.005)	0.74	0.79
(0.9,0.05,0.05)	(0.9,0.05,0.05)	0.84	0.79

smaller or greater than the naive probability. In the first line with  $\beta_1 = 0.999$ , there is considerable uncertainty as to the subpopulation to which  $S$  belongs given  $I, E$ ; in fact  $P_s(S \in X_1 | I, E) = 0.40, P_s(S \in X_3 | I, E) = 0.56$ . Since for this choice of parameters  $P_s(G | I, E, S \in X_3)$  (given by (5.23)) is only 0.32, we get a probability of guilt equal to 0.50, much smaller than the naive probability. However, as  $\beta_1$  decreases, so does  $P_s(S \in X_1 | I, E)$ , and  $P_s(S \in X_3 | I, E)$  grows. In the last line of Table 1,  $P_s(S \in X_3 | I, E)$  is large (equal to 0.95), so the posterior probability of guilt is predominantly given by (5.23) applied to  $s \in X_3$ , which is 0.89 for these parameters.

**Example 7.2.** As observed above, we obtain the same probabilities  $P(G | I, E)$  (and of course, the same naive probability of guilt), when in the above example  $\epsilon$  and  $\beta$  are exchanged. The explanation for these probabilities is somewhat different. In the first line of Table 1 (now with  $\epsilon_1 = 0.999$ ), it is quite likely that  $S$  belongs to  $X_1$  given  $I, E$ ; in fact  $P_s(S \in X_1 | I, E) = 0.79, P_s(S \in X_3 | I, E) = 0.21$ . Since for this choice of parameters  $P_s(G |$

$I, E, S \in X_1$ ) (given by (5.23)) is only 0.40, we get a probability of guilt equal to 0.50, much smaller than the naive probability. However, exactly as for Example 7.1, as  $\epsilon_1$  decreases, so does  $P_s(S \in X_1 | I, E)$ , and  $P_s(S \in X_3 | I, E)$  grows.

These examples show that the effect of having subpopulations can be considerable when the profile is more common among the smaller subpopulations, even when both  $S$  and  $C$  are likely issued from the largest subpopulation. The magnitude and the direction of the subpopulation effect depend strongly on the a priori probabilities for  $S$  and  $C$  to belong to each of the subpopulations.

**Example 7.3.** Letting  $S$  and  $C$  be likely issued from  $X_2$  or  $X_3$ , we get a posterior probability of guilt between 0.80 and 0.85 which does not depend strongly on the precise choice of  $\epsilon_i$  and  $\beta_i$ . This is understandable since these choices all make  $P(S \in X_1 | I, E)$  small, and (5.23) applied to  $X_2$  and  $X_3$  yields 0.89 for both populations (note that they have the same expected number of  $\Gamma$ -bearers).

**Example 7.4.** Consider the case where  $p_2$  and  $p_3$  are smaller than  $p_1$ , for example  $p_1 = 10^{-8}, p_2 = 10^{-9}, p_3 = 10^{-10}$ . The population as a whole then has a smaller number of expected  $\Gamma$ -bearers compared to when  $p_2 = p_3 = p_1$ . The true probability of guilt exceeds the naive probability unless one is almost sure that  $S$  and  $C$  are from different subpopulations, as illustrated in Table 2.

Table 2: Guilt probabilities for  $p_1 = 10^{-8}, p_2 = 10^{-9}, p_3 = 10^{-10}$

$(\epsilon_1, \epsilon_2, \epsilon_3)$	$(\beta_1, \beta_2, \beta_3)$	$P(G I, E)$	$P^{\text{hom}}(G I, E)$
uniform	uniform	0.80	0.79
(0.9,0.05,0.01)	(0.9,0.05,0.05)	0.80	0.79
(0.2,0.6,0.2)	(0.2,0.6,0.2)	0.98	0.79
(0.1,0.3,0.6)	(0.3,0.3,0.4)	0.97	0.79
(0.9,0.09,0.01)	(0.01,0.01,0.98)	0.88	0.79
(0.99,0.009,0.001)	(0.001,0.001,0.998)	0.57	0.79

## 7.4 $\Gamma$ -correlation: relatedness

Suppose that  $C$  has DNA-profile  $\Gamma$  with a population frequency of  $10^{-7}$ , i.e.,  $p_x = 10^{-7}$  for all  $x \in X$ . Now we select  $s$  from  $X$ , and  $\Gamma_s = 1$ . Suppose that  $X = \{s, y_1, y_2, y_3, z_1, \dots, z_N\}$  and  $N = 10^6$ , such that  $c_{y_i, s} = 10^{-3}$

and  $c_{z_i,s} = p_{z_i} = 10^{-7}$ . Here we model a situation in which the suspect  $s$  has three brothers, whose  $\Gamma$ -probability is  $10^{-3}$  given  $\Gamma_s = 1$ , and that the rest of the population is unrelated to  $s$ . If the a priori probabilities are  $P_s(C = s | I) = 0.4, P_s(C = y_i | I) = 0.1, P_s(C = z_j | I) = 0.3/10^6$  then the correction factor (3.6) for the  $\Gamma$ -correlation is equal to

$$\frac{0.6}{3 \cdot 10^4 \cdot 0.1 + 10^6 \cdot 1 \cdot 0.3 \cdot 10^{-6}} = \frac{0.6}{0.3 + 3000} \approx \frac{1}{5000},$$

meaning that the likelihood ratio associated to  $\Gamma_s = 1$  has been made 5000 times smaller, reducing it from  $1/p_s = 10^7$  to 2000.

For the posterior probability of guilt  $P_s(G | I, E)$ , this means that it is reduced from

$$\frac{\frac{2}{3}10^7}{1 + \frac{2}{3}10^7} \approx 1 - \frac{3}{2}10^{-7}$$

that we would obtain without  $\Gamma$ -correlation, to approximately  $1-3/4000$ .

## 7.5 Biased search

We recast the example given in Section 7.4 in the setting of a biased search, to demonstrate the equivalence noted in Section 3.3. As in 7.4,  $p_x = 10^{-7}$  for all  $x \in X$ , we suppose that  $S = s$  has been selected and that  $\Gamma_s = 1$ , and that there are  $y_1, y_2, y_3 \in X$  such that  $\sigma_{x,y_i} = 10^4 \sigma_{x,x}$  for  $i = 1, 2, 3$  and  $\sigma_{x,z_i} = \sigma_{x,x}$  for all  $i$ . The prior odds are as in Section 7.4. Then the likelihood ratio associated to the evidence  $\Gamma_s = 1$  is reduced by a factor of about 5000, as in Example 7.4. In that example, the value was decreased since finding  $\Gamma$  in  $s$  made it more probable that population members  $y_1, y_2, y_3$ , which have non-negligible prior probabilities of guilt, also have  $\Gamma$ . In this situation, it is due to the fact that the selection procedure is such that if  $s$  is selected, it becomes less likely that  $s$  is guilty:

$$P(C = s | S = s, I) = \frac{P(C = s | I)}{\sum_{y \in X} \frac{\sigma_{s,y}}{\sigma_{s,s}} P(C = y | I)} \approx \frac{1}{7500},$$

which is considerably less than 0.4. The fact that  $s$  has  $\Gamma$  then raises the probability of guilt to approximately  $1 - 3/4000$  as above.

## 7.6 Database effectiveness

In (6.5) we have computed the odds in favour of a unique database match being with the true criminal. If the database is a random sample of the

population in the sense that  $P(C \in \mathcal{D} \mid I) = |\mathcal{D}|/|X| = n/N$ , then this equation reads

$$\frac{P(S = C \mid E_1, I)}{P(S \neq C \mid E_1, I)} = \frac{1}{p(N - n)},$$

which is monotonically increasing in  $n$ , going from  $1/((N - 1)p)$  for  $n = 1$  to infinity for  $n = N$ . It is not hard to derive this directly: since  $n - 1$  persons have been shown not to possess  $\Gamma$ , the population that can not be excluded has size  $N - n + 1$ . In that population, only the  $\Gamma$ -status of one individual (the one that matched in  $\mathcal{D}$ ) is known. Since  $\mathcal{D}$  was a random sample as defined above, the classical solution (2.4) applies.

If the database is not a random sample from the population in the above sense, then the situation is more interesting and quite different.

**Example 7.5.** Let  $p = 10^{-7}$  and suppose that with  $n = 10^5$  one has  $P(C \in \mathcal{D} \mid I) = 0.2$ . For example, this may be because the database consists of previously convicted individuals and based on the probability of a rightful conviction and of recidivism one arrives at such an estimate. For database  $\mathcal{D}$ , the odds that a unique match is with  $C$  are 25 to one, or equivalently,  $P(S = C \mid I, E_1) = 96\%$ .

It may be possible to enlarge  $\mathcal{D}$  to  $\mathcal{D}'$  with  $|\mathcal{D}'| = n'$  such that  $P(S = C \mid I, E_1) = 0.5$ , but only at the cost of adding very many individuals into  $\mathcal{D}'$ , e.g. with  $n' = 2 \cdot 10^6$ . In that case, the odds (6.5) on a unique match being with the offender decrease to 5, i.e., one in six of such matches will be with an innocent person.

The probability of actually obtaining a unique match is given by

$$P(E_1) = P(C \in \mathcal{D} \mid I)(1 - p)^n + P(C \notin \mathcal{D} \mid I)np(1 - p)^{n-1}.$$

For database  $\mathcal{D}$ , this evaluates to 20,6% and for database  $\mathcal{D}'$  to 49,1%. Thus, in  $\mathcal{D}'$  a search with a DNA-profile with population frequency  $10^{-7}$  will yield a unique match about half of the time, but only 5 out of 6 of these will be with the true offender. About 10% of such searches will result in two or more hits, and about 40% will not result in any hit.

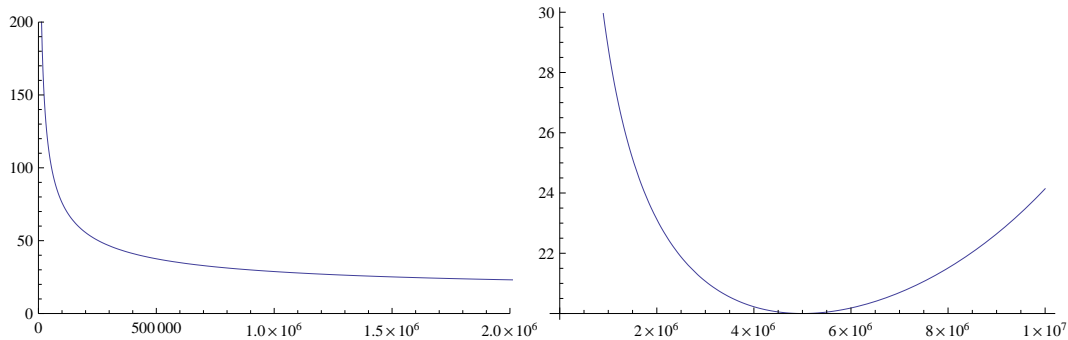
When multiple matches are found, it is more likely that one of them is with the true offender but not a near certainty: e.g., in case two matches are found (which happens with probability 9%), about one in ten of such double matches are both coincidental.

For the original database  $\mathcal{D}$ , about 20,6% of searches result in a unique match, almost all of which are with the offender; in the remaining cases one almost always has no hits: the probability of having more than one match being 0.2%.

**Example 7.6.** Suppose that the database is set up and expanded such that if it has size  $n$  then  $P(C \in \mathcal{D}) = \sqrt{n/N}$ . This is a model for a database in which individuals with higher prior probability of guilt are put in the database with higher probability. For example, if  $\mathcal{D}$  contains the DNA-profiles of 10% of the population, then it contains  $C$  with probability 31%. If  $\mathcal{D}$  is enlarged to contain 30% of the population, then it contains  $C$  with probability 55%.

In that case, the odds on a unique match being with the criminal in a database of size  $n$  are minimal for  $n = N/4$ . An example for  $N = 2 \cdot 10^7$ ,  $p = 10^{-8}$  is given in Figure 1. With  $n = N/4$  the odds in favour of a unique match being with the true offender are 20. As the plots show, when the database is relatively small the odds on a match being with the true offender decrease rapidly, e.g. from 105 if  $n = 50.000$  to 55 if  $n = 200.000$ . As  $n$  grows further, the odds decrease (slowly) to 20 for  $n = 5 \cdot 10^6 = N/4$ . When  $n$  grows further, the odds increase again. When 50% of the population is included ( $n = 10^7$ ), they are 24.

Figure 1: Database effectiveness with  $p = 10^{-8}$ ,  $N = 2 \cdot 10^7$ ,  $P(C \in \mathcal{D}) = \sqrt{n/N}$



Thus, enlarging a database may at the same time increase the chance of obtaining a unique match from it, and diminish the value of such a match in the sense that the probability of it being with the true offender decreases. These examples suggest that the idea that the larger the database, the better, needs relativation. It is of course true that enlarging a database increases the probability that the criminal is included. It is also obvious that given a unique match in the database, the probability that it is with the criminal increases when the database is expanded and does not yield additional matches. But as we have seen, it does not follow that hits in larger databases are stronger evidence for guilt than hits in smaller databases.

## References

- [1] D.J. Balding and P. Donnelly, Inference in Forensic Identification, *Journal of the Royal Statistical Society, Series A* **158** (1995), no. 1, 21–53.
- [2] A.P. Dawid and J. Mortera, Coherent Analysis of Forensic Identification Evidence, *Journal of the Royal Statistical Society. Series B (Methodological)* **58** (1996), no. 2, 425–443.
- [3] A.P. Dawid and J. Mortera, Forensic Identification with Imperfect Evidence, *Biometrika* **85** (1998), no. 4, 835–849.
- [4] R. Eggleston, Evidence, proof and probability, *Law in Context*, Weidenfeld and Nicolson, London, 1978.
- [5] S.L. Lauritzen and D.J. Spiegelhalter, Computations with Probabilities on Graphical Structures and Their Application to Expert Systems, *Journal of the Royal Statistical Society, Series B (Methodological)* **50** (1988), no. 2, 157–224.
- [6] R. Meester and M. Sjerps, The evidential value in the DNA database search controversy and the two-stain problem, *Biometrics* **59** (2003) 727–732.
- [7] R. Meester and M. Sjerps, Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence, *Law, Probability and Risk* **3** (2004) 51-62.
- [8] A. Stockmarr, Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search, *Biometrics* **55** (1999) 671-677.
- [9] C.M. Triggs and J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, *Science & Justice*, **46** (3), (2006) 169-178.
- [10] J. Yellin, Review of Evidence, Proof and Probability (by Richard Eggleston), *Journal of Economic Literature* **17** (1979), no. 2, 583–584.