Norwegian University
of Life Sciences

Master’s Thesis 2016   60 ECTS
Department of Chemistry, Biotechnology and Food Science

Estimation of small probabilities with applications to forensic genetics

Naomi Azulay
Bioinformatics and Applied Statistics

Abstract

When determining kinship relations based on genetic material, there is a need to quantify the probability for doing errors. These probabilities are typically very small, but nonetheless important, and needs to be controlled. It is no trivial task to estimate such small probabilities, and so the main purpose of this thesis has been to explore how they can best and most accurately be estimated. The applications in this thesis has been kinship cases in forensic genetics, but the methods are relevant also for other areas where calculations rely on the values of small probabilities.

To estimate these probabilities, the simulation method of importance sampling was used. This method works by sampling from a more beneficiary distribution than the one of original interest, and then correcting for it. The precision of the estimates has also been of interest in this thesis, and is quantified in terms of the variance and MSE. These measures were used to evaluate different possible sampling distributions. The importance sampling method worked better for smaller probabilities than the straightforward simulation, but there are some trade-offs; there seems to be a difference between which distribution gives the best estimates, and which ones manage to make non-zero estimates for the smallest probabilities.

As is intuitively understood, using more markers when genotyping makes it easier to correctly determine relationships, but it does however also make it harder to estimate the probabilities involved. In conclusion, we have found that relevant estimates can be found, and that these estimates can be used to add understanding to conventionally used verbal statements when evidence is presented in court. It is possible to accompany such verbal statements, like “extremely strong evidence”, to tangible probabilities.

Sammendrag

Når slektskapsforhold skal bestemmes på grunnlag av genetisk materiale er det et behov for å tallfeste sannsynligheten for å gjøre feil. Disse sannsynlighetene er typisk veldig små, men likevel viktige, og må kontrolleres. Det er ingen triviell oppgave å estimere slike små sannsynligheter, og hensikten i denne avhandlingen har hovedsakelig vært å undersøke hvordan disse kan estimeres på beste og mest presise måte.

For å estimere disse sannsynlighetene har simuleringsmetoden importance sampling blitt brukt. Denne metoden virker ved å trekke utvaget fra en mer gunstig fordeling enn den som opprinnelig er av interesse, for så å korrigere for dette. Presisjonen til estimatene har også vært av interesse i denne oppgaven, og har blitt tallfestet gjennom varianse og MSE. Disse målene ble brukt til å evaluere forskjellige alternative fordelinger. Metoden importance sampling virket bedre for mindre sannsynligheter enn direkte simulering, men noen avveininger må gjøress; det ser ut til å være en forskjell mellom hvilken fordeling som gir det beste estimatet, og hvilken som får til å produsere et resultat som ikke er null for de aller minste sannsynlighetene.

Som man intuitivt kan forstå så blir det enklere å bestemme slektskap på en korrekt måte når flere markører tas i bruk, men samtidig blir det vanskeligere å estimere de involverte sannsynlighetene. For å konkludere, så har vi funnet at de relevante estimatene kan bestemmes, og at disse estimatene videre kan bli brukt til å øke forståelsen for konvensjonelt brukte verbale uttalelser ved presentasjon av bevis i retten. Det gir en mulighet for at slike verbale uttalelser som “ekstremt sterkt bevis” kan ledsages av håndfaste sannsynligheter.

Acknowledgements

This thesis was written at the Department of Chemistry, Biotechnology and Food Science at the Norwegian University of Life Sciences (NMBU) for the Biostatistics group. I would like to thank my supervisor Thore Egeland for all his incredible help and support throughout the work with this thesis. I would also like to thank Trygve Almøy, both for teaching my first course in statistics in such an enjoyable manner, and for having the door open for me at any time ever since.

In addition I would like to thank Marguerethe Stenersen for lending me an office at the paternity laboratory in Oslo. I am also very grateful to Daniel Kling and the others at the laboratory for their hospitality, and for being the best of lunch partners.

To my friends and family who have encouraged me through the hard times, and especially Youssef, thank you for always being there for me.

Oslo, May 2016

Naomi Azulay

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General background

First we wish to explain the thesis' title to a general audience. An important key word is 'forensics', by which we mean the science that is relevant for the law and court. It is not surprising that many areas of science can be relevant for court cases. A much used example is forensic pathology, that is typically used to estimate time and find the cause of death. If we use the term "digital forensics" to refer to new technology in the service of the court, that is another example. It can include many methods from statistics and computer science, like pattern recognition and image analysis, that are used to analyze pictures or the like, which may be of relevance for the court. One relevant example is that this evidence may come from surveillance cameras from terrorist attacks.

In this thesis the broad area of application is biology, more specifically we will deal with genetical (DNA) applications in forensics. Forensic genetics is a large field, and normally we distinguish between crime cases and kinship cases, see figure 1.1. This thesis is relevant for both applications, but our examples will be drawn from kinship cases. Kinship cases are largely the determination of familial relationships based on DNA profiles. By the term 'kinship' we indicate that we may be looking into more distant relationships than a parent-offspring relationship, but we will mainly focus on paternity cases. The objective of this is to determine whether a specific man is the biological father of a child or not. This enterprise is important both for the children involved, since they may have an obvious interest in their biological

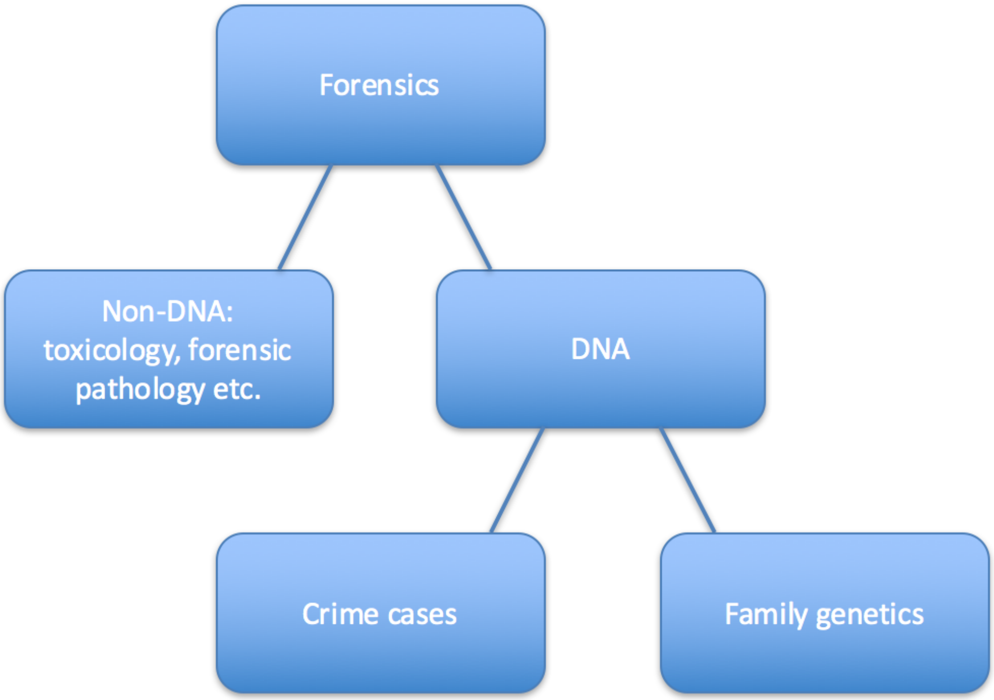


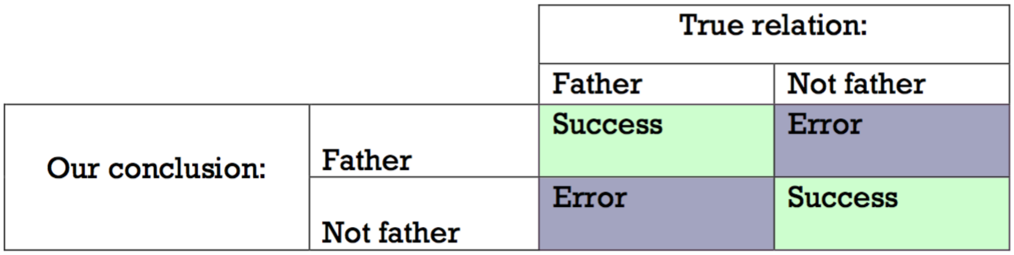
Figure 1.1: Overview of the field of forensics

origin, but also for the general society, as these things are often linked to social benefits etc. This may of course differ between countries.

There are also other types of applications for kinship testing, e.g. inheritance claims, where establishing paternity status may have considerable economic importance for the family involved. A classical example is that after the death of a rich man it is claimed that he has unacknowledged children, and a dispute around the inheritance arises. Immigration is another area where it is important to determine familial relationships precisely, as this can have consequences for their possibility of family reunion. In many countries a person that can prove a close relationship, like being the child of somebody that has been granted immigration, has the right to reunite. Again the law and practice differs between countries. A third and important application is identification of victims after mass disasters (DVI). An early example of identification of victims based on their living relatives is presented in [12]. This paper discusses the identification of 141 persons after a plain crash at Svalbard in 1996. 139 out of the 141 victims were identified based on the comparison of DNA recovered from the disaster area and DNA from living relatives; the two last were identified by other means as no relatives were found.

To cover the genetical parts of this thesis there will inevitably be presented some biological theory. For this purpose the 'Glossary for non-biologist' from [3] can be helpful. It can be found through searching up the book title at `https://books.google.com`, page 229.

Turning to the first part of the title, we note that statistical methods are relevant for the mentioned applications in forensic genetics. When conducting these kinds of tests there is always a risk of doing one of two types of errors. Formulated in terms of paternity testing, we can either wrongly declare that the true biological father is not the father, or we can declare that a man who is not the father to the child is the biological father (see figure 1.2). The probability of committing any of these errors is small, but certainly not 0, and so we have to accept a non-zero chance of committing an error. In this thesis we would like to calculate the probabilities of these errors as precisely as possible. This explains the first part of the title, particularly the word 'small'. Estimating small probabilities is typically a non-trivial task. In this thesis our main method will be based on stochastic simulation. In particular we will be using a method called importance sampling, which is especially well suited for applications of this kind.



		True relation:	
		Father	**Not father**
Our conclusion:	**Father**	**Success**	**Error**
	Not father	**Error**	**Success**

Figure 1.2: The two types of errors we can make in a paternity case.

1.2 Literature

There have for a long time been methods available to *exclude* paternities based on the ABO blood system. Alleged fathers could be excluded from paternity if their blood type were incompatible with the child's, but could not be confirmed fathers. The works of Erik Essen-Möller [4] are considered fundamental since they were the first to present the statistical methods to *confirm* paternity. The paper [4] gives a historical account of the work of Essen-Möller and also explains the relevance of forensic genetics. There is a large literature on forensic genetics and applications to kinship testing. We will mention some of the more recent text books, like [1, 3, 5], that are relevant for forensic genetics. Next we will include some references for more statistical literature. For this thesis the most relevant references are those that deal with stochastic simulation. A classical book on this topic is [13]. This book describes variance reduction and parts of the book on importance sampling is of particular relevance for this thesis. Regarding scientific literature this thesis builds directly on [9], which describes how importance sampling can be used in forensics.

1.2.1 Software

The author of [9] also made available an R-package, 'DNAprofiles', which provides functions for importance sampling and exact methods for the type of applications we have in mind. We have used this R-package in the thesis, but we have also implemented some new functions in the R-package 'naomi'.

1.3 Purpose

In Norway alone, there are about 2000 paternity cases that come up a year (personal communication with Thore Egeland. More information can be found at `http://www.farskap.no`). In these cases important decisions are made based on the methods presented, and our main aim is thus to avoid errors. In this thesis we will focus on the probability for the second type of error mentioned above, namely concluding that a man is the biological father of a child when he is not. The statistical evidence in these cases are based on the DNA profiles of the persons involved, and is summarized with what is called the 'Likelihood Ratio' (LR). The interpretation of the LR is that it states how much more probable the collected data is if the man is the father compared to if he is unrelated to the child. The larger the LR gets, the more compatible the DNA profiles are. A large enough LR, say $LR > 10^6$, would lead to a conclusion of paternity. We would like to know the probability of getting such a large LR, or larger, if the man is in fact unrelated to the child, i.e. $\Pr(LR > 10^6 \mid \text{not father})$. This probability is called an exceedance probability. It is usually very small, but important to quantify because it says something about how often we make that type of mistake. When probabilities get this small, normal simulation methods often fail. For this reason we will investigate if the simulation method importance sampling is a good alternative.

Other possible methods include an exact approach, asymptotic approximation to the normal distribution and direct, Monte Carlo, simulation, which will all be discussed. In [9] Kruijver presents both a method for exact calculation of these probabilities and for using importance sampling. We have chosen to follow up on the importance sampling method, since this in principal could work generally also for a high number of markers, while the exact approach has some upper boundary. We wanted to expand upon the method in [9] so that it would also contain methods for evaluating the estimates.

Another objective in this thesis is to find out, in the case of importance sampling, what distribution is best to sample from. The method of importance sampling revolves around finding a good alternative distribution, a *biasing distribution*, to sample from. The added methods for evaluation of the estimates was used for this purpose.

1.4 Organization of the thesis

The thesis is organized as follows: In chapter 2 on methods and data we present the theoretical basis for the thesis. We start by explaining the standard way of testing hypotheses that are typically taught in beginner courses in statistics. We then introduce some of the biological concepts that are needed to understand the forensic aspect of the thesis, and contrast the traditions of statistics in forensics and in general. In section 2.3 we take an in depth look at the likelihood ratio and its distribution, and explore different approaches for finding exceedance probabilities.

At the end of the chapter we look at how to evaluate estimates and a section on how the thesis was implemented in R. In chapter 3 we lay out the results in the form of four examples, that are discussed in chapter 4. We also join in on the discussion around how the statistical evidence is best presented in court.

Chapter 2

Methods and data

2.1 Classical hypothesis testing

We start by reviewing the classical approach for hypothesis testing. In classical statistics, the method of hypothesis testing is widely used to choose between two competing claims, or hypotheses, on the basis of collected data. If, for example, someone wants to introduce a new medical drug to the market, we need to decide how much better the new drug has to be compared to the old one, before it is safe to switch the general practice to this new drug. This decision is based on the value of a test statistic.

One of the hypothesis is called the null hypothesis, H_0, and represents the established knowledge, or the most conservative position to hold. It typically states that there is nothing new, there is no effect of what we are testing. We assume that this is true, and so the burden of proof lies with anyone claiming the opposite.

The distinguishing feature of the classical approach is the model. In most cases a precise model is defined including probability distributions for the data involved. The model typically contains some parameters: The data from the experiment is assumed to follow a known distribution, in the above example it could make sense to assume a normal distribution. If so, the parameters are the expectation μ and the standard deviation σ. For convenience we now assume σ to be known. If the drug is supposed to lower a person's cholesterol level, we define:

$$x_i = \text{measured cholesterol level for person } i \text{ with the new drug}$$
$$x_i \sim NID(\mu, \sigma)$$

In many cases the old mean is known, because the old way have been in place for such a long time. Let us assume that the mean with the old drug is known, we call it μ_0. Now we can formulate the hypotheses in terms of parameters from the model:

$$H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0$$

In forensics, as we will discuss in more detail in section 2.2.2, hypotheses are not formulated in terms of parameters but rather statements like "AF is the biological father".

We want to reject the null hypothesis if the sample mean $\bar{X}$ is sufficiently larger than μ_0, but how much is 'sufficiently'? Instead of looking at the distribution of X, we can use what we know from the standard normal distribution. We know for example that there is a 5 % chance of observing a z greater than 1.64. In our example, however, the data is not standard normally distributed, and so we need to find a way to compare $\bar{x}$ to 1.64:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

We use this to make a rule for when to choose which hypothesis. First we have to decide how much of a risk we are willing to take, i.e. specify an α. We can then summarize the test by the decision rule: Reject if $Z > Z_\alpha$, where Z_α is is the number that delineate the upper (or lower) $\alpha \cdot 100$ percentile in the standard normal distribution.

Above we chose $\alpha = 0.05$, which gives us $Z_\alpha = 1.64$. This means that we can reject H_0 if $z > 1.64$, or equivalently if $\bar{x} > \mu_0 + 1.64 \cdot \frac{\sigma}{\sqrt{n}}$.

When we have found our z-value, we can calculate a p-value. The p-value is $\Pr(Z > z|H_0)$, the probability of getting at least this z, given that the null hypothesis is true. We compare the p-value with the α we chose, and reject H_0 if the p-value is smaller. This means that we are taking a smaller risk than our upper limit.

There are two different types of mistakes we can make now, called type I and type II errors. A type I error is to reject the null hypothesis when it is in fact true, and the type II error is to keep the null hypothesis when the alternative is true. The type I error is usually considered the most severe mistake, and we often compare this to the legal system. A person is always considered innocent until proven guilty, because we are afraid of miscarriages of justice. To judge an innocent person guilty is in the jurisprudence considered worse than letting a guilty person go. In that case the null hypothesis is that the person is not guilty, and it has to be proven “beyond reasonable doubt” that the person is guilty. Because of this the hypothesis are said to be asymmetrical.

2.1.1 Power function

With the classical hypothesis testing approach we are in control of the chance of doing a type I error, but the α does not control the probability of making a type II error – failing to reject H_0 when H_1 is true. To deal with this problem we can calculate the probability of getting a sample mean in the acceptance region despite the true mean being different from μ_0. This is called β and is calculated as $\beta = \Pr(\bar{X} < k \mid \mu)$. However, it is more common to compute the *power* of a test, $1 - \beta$. The power of a test is the probability of rejecting the null hypothesis when the true mean is μ (see figure 2.1 on the following page). The power function $\gamma(\mu) = \Pr(\bar{X} \geq k \mid \mu)$ depends both on the true (unknown) value of the parameter μ and the number of observations n:

$$\begin{aligned}\gamma(\mu) &= \Pr(\bar{X} \geq k) = \Pr(Z \geq \frac{k-\mu}{\sigma/\sqrt{n}}) \\ &= \Pr(Z \geq z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}) = 1 - \phi(z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}})\end{aligned}$$

This is based on $k = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$. An example of a power function is given in figure 2.2 on page 11. The power function is a useful and intuitive way of presenting the properties of statistical tests. We can easily study the effect of increasing sample size, for example. Unfortunately, there is no simple analogue to power functions in the forensic application since we do not have a parametric model. However, we will later study how $\Pr(LR > t|H)$ depends on the number of markers.

2.2 Review of forensic genetics

In this section we provide the necessary genetical background. The first thing that is important to know is the way DNA is inherited from parents to their children. A child inherits roughly half of its DNA from the mother and the other half from the father. This means of course that each parent only passes on half of their DNA to their child. What parts that gets passed on is chosen in a random way. We will get more into this in section 2.2.1,

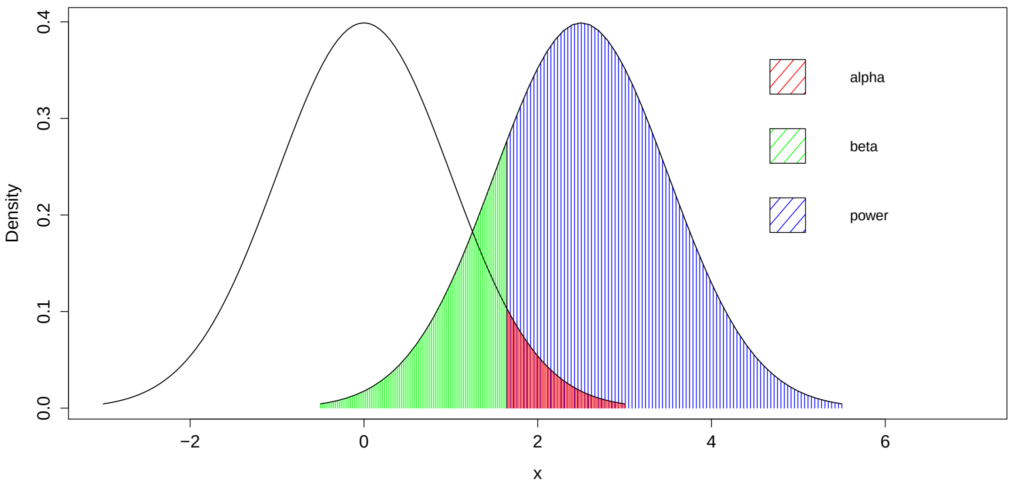


Figure 2.1: The curve on the left is the distribution under H_0 with $\mu_0 = 0$, while the curve on the right is the true distribution with $\mu = 2.5$. Both with $\sigma = 1$. The power is the probability of rejecting the null hypothesis when the true mean is μ.

but the process of meiosis will not be explained in any more detail. A more thorough background can be found in [15]

It is normal to differentiate between coding and non-coding parts of the genome. A (maybe overly) simplistic explanation is that the coding parts, the genes, encode protein sequences and so gives rise to different traits called *phenotypes*. However, the vast majority of the genome consists of non-coding parts, at least for more complex organisms, including humans. For some of these parts we know the biological function, one example being regulation of genes, but for much of it we do not even know if it has any function.

In the applications of genetics in forensics that we have in mind, only a small fraction of these non-functional, non-coding parts are utilized. For this reason, we use the term ‘marker’ (or locus, plural loci) rather than ‘gene’ for the genetic information obtained, as the latter may imply that some phenotypic information is involved. This way of choosing markers is the traditional one in forensics. There has, however, been an increase in the use of phenotypic markers, the objective being for example to determine hair color. Such problems, i.e. the prediction of phenotypes, will not be considered in this thesis.

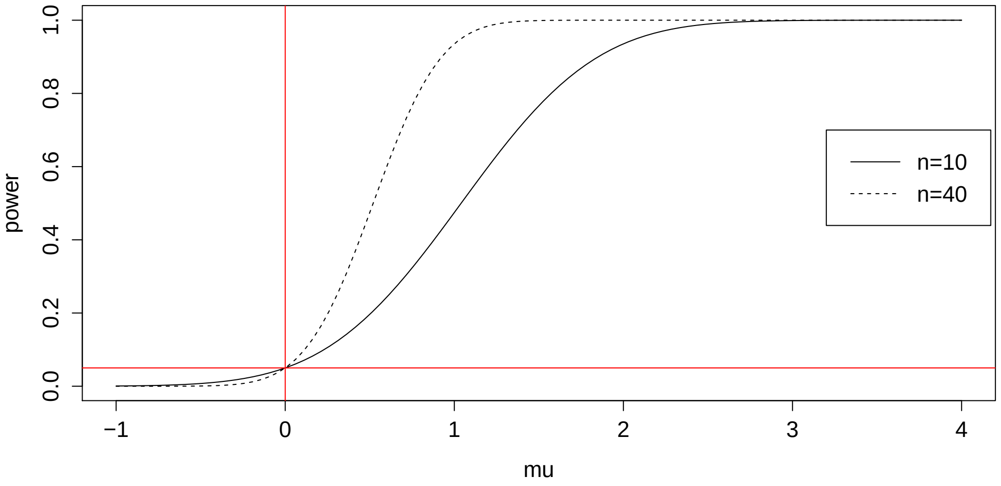


Figure 2.2: Power function for $\mu_0 = 0$, $\alpha = 0.05$ and $\sigma = 2$. When $\mu = \mu_0$, the power is the same as α (see figure 2.1 on the preceding page); we see that $\gamma(0) = 0.05$ where the red lines meet. With 10 observations we have an 80% change of detecting a difference if the real $\mu = 1.6$. With 40 observations we reach 80% certainty already at $\mu = 0.8$.

There are mainly two reasons for the traditional choice of markers; first, we do not want to collect phenotypic information, as it could be of a sensitive nature or in other ways be difficult to process. For example, if it is found that a person is a carrier of a harming mutation, it could be a moral obligation to inform the individual, which could be difficult.

Second, coding parts of the genome are more susceptible to variations between populations, and in forensics we prefer to deal with markers that are homogeneous, meaning that the *genetical variation* is reasonably constant between populations. Specifically, it is convenient if we can use the same databases for many or all populations, as will be explained in more detail later on.

In addition to being homogeneous, we also prefer the genetical markers to be *polymorphic*, i.e. variable. The meaning being that there exist many variants of it in the population, which is something that makes a marker good at distinguishing between people. The different variants of a marker are called alleles, and if there exists a lot of alleles for a marker, it has greater identifying power.

There are many different types of genetic markers, but *STRs*, short tandem repeats, are the most commonly used in forensic genetics [3]. These are places in the genome where patterns in the genetical code occurs, and these patterns can be repeated a different number of times. How many times a specific pattern is repeated determines which allele it is, see figure 2.3. Another much used type of marker is *SNPs*, single nucleotide polymorphisms. These are variations in a single nucleotide in the genome, and they usually consist of only two alleles. Because of this one needs to use a greater number of markers compared to STRs to get the same confidence. The advantage with SNPs is that they are generally less prone to mutations.

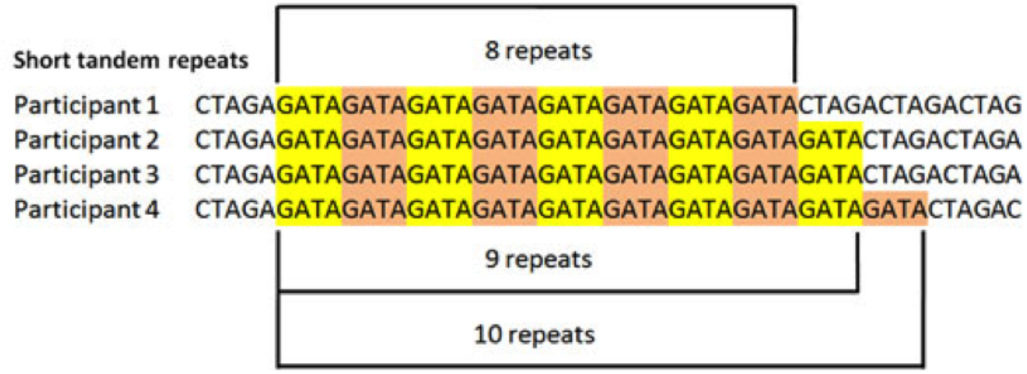

Short tandem repeats
8 repeats
Participant 1 CTAGAGATAGATAGATAGATAGATAGATAGATAGATACTAGACTAGACTAG
Participant 2 CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATACTAGACTAGA
Participant 3 CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATACTAGACTAGA
Participant 4 CTAGAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTAGAC
9 repeats
10 repeats

Figure 2.3: Each nucleotide in the genome is denoted with one of the letters 'A', 'T', 'C' or 'G'. The way these are put together is what makes our genomes different from one another. At some loci the nucleotides form repeating patterns. Here is a representation of an STR marker made up of the pattern 'GATA'. The number of times the pattern is repeated determines which allele the participant has. Here we see the alleles 8, 9 and 10 for this specific marker. Source: `http://www.stewartsociety.org/bannockburn-genetic-genealogy-project.cfm`.

Different countries use different markers, but the overlap is considerable in the sense that quite a few markers are in common. An effort has been made to standardize and use the same markers in the databases in a large number of countries, as described in [6]. This makes it possible to perform searches across different countries. For instance, if a stain recovered from a crime sample in Norway, does not give a match against the Norwegian database, searches can be made against neighboring countries according to the Prüm agreement (`https://en.wikipedia.org/wiki/Pr%C3%BCm_Convention`).

2.2.1 Standard paternity case

In this thesis we will concern ourselves with the theory surrounding relationship inference, and the simplest form of relationship inference we find in paternity cases. Here the goal is to say on a satisfyingly solid basis whether some alleged father is in fact the father of a child or not. A standard paternity case is illustrated in figure 2.4 (this figure and figure 2.5 are made using the R package 'paramlink' developed by M.D. Vigeland). We get DNA-data from the parties involved, in this simple example only from one marker. If we only have data from the alleged father and the child, it is called a *duo case*. If we also have data from the mother it is a *trio case*.

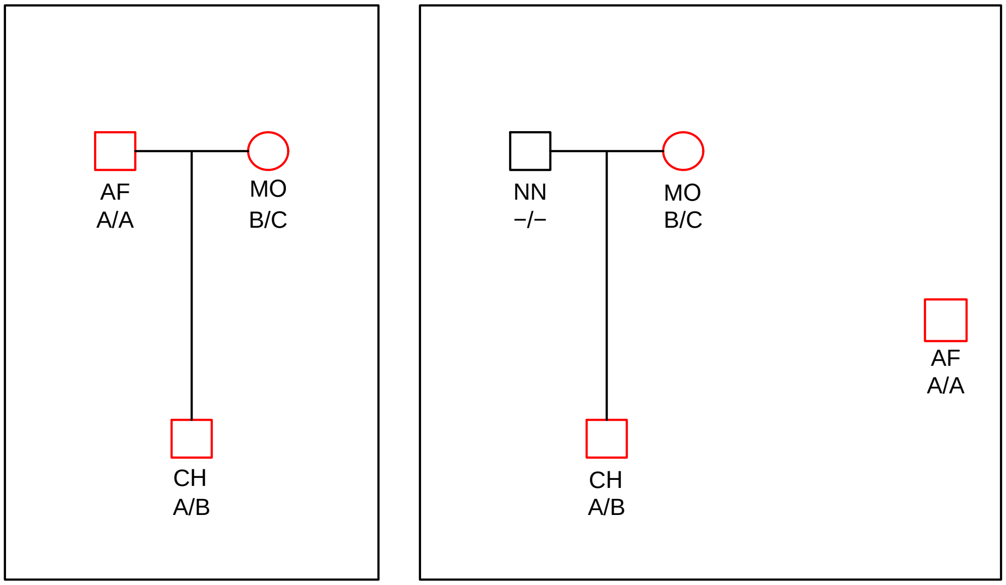


Figure 2.4: A standard paternity case: On the left hand side, AF and mother are parents of the child, whereas in the right hand side AF is an unrelated man. AF is the alleged father, MO is the (undisputed) mother, CH is the child.

The figure shows three individuals: AF (alleged father), MO (mother) and CH (a child). Females are conventionally shown using circles, males squares. In the left hand part, AF and mother are parents of the child, whereas AF is an unrelated individual in the right hand side. One marker is shown in the figure. For each (autosomal) marker, a person has two alleles, one inherited from the mother and one from the father. (For X and Y chromosomes and the mitochondrial DNA, inheritance patterns differ, however, these markers are not discussed in this thesis.) For instance, we see that the alleged father's genotype (G_{AF}) is A/A and the child's genotype (G_{CH}) is A/B.

In accordance with Mendelian genetics, each parent passes on one of their two alleles randomly to the child. In the case shown in figure 2.4 AF will have to pass on A since he is *homozygous*, whereas the mother would pass on B or C, each with probability 0.5. The figure indicates data consistent with paternity, but it could also be due to chance. To find out which of the two scenarios are most likely to have generated the data, we calculate a likelihood ratio (LR), as will be explained in section 2.2.3. But first we need to take a look at how the hypothesis are sat up in the forensic setting.

2.2.2 Hypotheses in forensics

Previously, in section 2.1, we discussed hypothesis testing as it is normally done in the classical way. Next, the approach in the forensic genetic setting is introduced and we will compare and contrast these traditions.

As mentioned previously, the formulation of hypotheses differentiates the forensics field from classical statistics. Instead of writing them in the form of parameters, they are formulated verbally. For this reason it is crucial that the hypotheses are precisely formulated. A typical example in a paternity case will be (also see [3]):

$$\begin{aligned} &H_1 : \text{AF is the biological father of the child.} \\ &H_2 : \text{Someone unrelated to AF is the biological father of the child.} \end{aligned} \tag{2.1}$$

We have to bear in mind that we always test a hypothesis against some alternative, and our conclusions will vary with different hypothesis. If we include relatives of AF in H_2, say the brother of AF, our evaluation of the evidence will not be unaffected.

Another difference is that the hypotheses are said to be *symmetrical*. This is because we think of it as equally bad to commit either of the two possible errors (see figure 1.2 on page 3): to wrongly claim that someone is the father of a child or to claim that the persons are unrelated when they are not. This is in contrast to the classical testing, were we are most concerned about falsely rejecting H_0, the type I error. Since we are equally worried about falsely rejecting either hypothesis, the concepts of type I and type II errors fade. Furthermore, we do not denote either of the hypotheses with the name H_0, as this implies that a position is assumed without evidence.

As a consequence, p-values are no longer used to draw conclusions. The p-value only control the chance of a type I error. Regardless, because of the lack of a parametric formulation we do not have a good analogue to these values, and so we need a different criterion to evaluate the claims. In the next section we go into the Likelihood Ratio and how it is used to evaluate evidence in forensic genetics.

In summary, if we compare classical hypothesis testing to the tradition in the forensics field, three differences stand out: first is the framework of parameterized models that we do not have in forensics, second the symmetry of the hypotheses, and third is the lack of the critical value and p-value to conclude the testing. By "critical value" we understand a value anchored in calculations; such calculations are not presented to justify what *threshold*, t, is used to conclude in forensics.

2.2.3 Likelihood Ratio

As described in the previous section, we do not use p-values to conclude the testing in forensics. Instead we summarize the evidence with the likelihood ratio. It is defined as:

$$LR = \frac{\Pr(\text{Data}|H_1)}{\Pr(\text{Data}|H_2)}$$

As seen from the formula, one has to state the two competing hypothesis to be able to calculate the LR. In a paternity case they will usually be as shown in equation 2.1 on the preceding page.

The LR is a summary of the statistical evidence, and is interpreted as how much more likely the data is if H_1 is true compared to if H_2 is true. The 'data' is the information from the genetic markers.

If we use the data from figure 2.4 on page 13 and start with assuming AF to be the father, he has to pass on the allele A with probability 1 (disregarding mutations and other artifacts like genotyping error). If a random man is the father the allele frequencies in the general population have to be used to estimate how probable the child's profile is. We denote the frequency of allele A in the population as p_A. The mother is assumed to be the mother in both cases, and passes down a B with probability 0.5.

$$LR = \frac{\Pr(\text{Data}|H_1)}{\Pr(\text{Data}|H_2)} = \frac{\Pr(G_{AF}, G_{MO}, G_{CH}|H_1)}{\Pr(G_{AF}, G_{MO}, G_{CH}|H_2)}$$

$$= \frac{\Pr(G_{CH}|G_{AF}, G_{MO}, H_1)}{\Pr(G_{CH}|G_{AF}, G_{MO}, H_2)} \frac{\Pr(G_{AF}, G_{MO}|H_1)}{\Pr(G_{AF}, G_{MO}|H_2)}$$

$$= \frac{\Pr(G_{CH}|G_{AF}, G_{MO}, H_1)}{\Pr(G_{CH}|G_{MO}, H_2)} = \frac{1 \cdot 0.5}{p_A \cdot 0.5} = \frac{1}{p_A}$$

Here we see that the strength of the evidence depends on how common the allele p_A is. This is quite intuitive, as it is more probable to see a common feature in a random person than a rare one. This is also the reason why we prefer polymorphic markers. For example, if p_A is close to zero, the LR goes to infinity. If AF is the only, or one of a very few people, with this feature, it is much more likely that this allele comes from him compared to a random person. If, on the other hand, $p_A = 0.5$, which means that half the population have this allele, then the LR = 2. The interpretation for the LR is then that it is only twice as likely that the data is a result of AF being the father compared to some random man unrelated to him. This means that the LR is heavily dependent on our estimates of allele frequencies being accurate.

So far only one marker has been considered, but more markers are needed to obtain reliable conclusions. The problem then arises: how should information from different markers be combined? Importantly, this is achieved by selecting markers to be independent. This will be the case if markers are on different chromosomes or sufficiently far apart on the same chromosome. We know from basic probability theory, that independence is a fundamental property that allows multiplication. From every marker we calculate an LR, and because we use independent markers we can combine these by multiplication to one overall LR, i.e.,

$$LR = LR_1 \cdot LR_2 \cdot \ldots \cdot LR_n$$

Example 2.1. A duo case with three markers is shown in figure 2.5. In this example we go through how the partial LR_is from each markers are combined to one overall LR in practice, based on the profiles in the figure. The standard assumptions are in force (i.e. no mutations, Hardy Weinberg Equilibrium, no genotyping errors).

Marker D3S1358:

$$LR_1 = \frac{0.5p_{17}}{p_{17}^2} = \frac{0.5}{0.2040} = 2.45$$

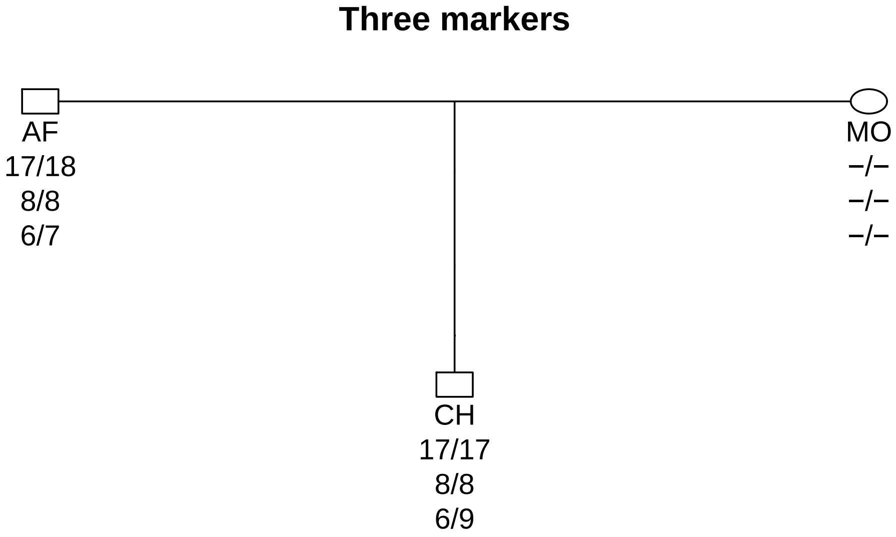


Figure 2.5: Pedigree for a duo case with three markers. The markers used are the three first in the table A.1 in appendix A.2.

Marker TPOX:

$$LR_2 = \frac{1p_8}{p_8^2} = \frac{1}{0.5539} = 1.81$$

Marker TH01:

$$LR_3 = \frac{0.5p_9}{2p_6p_9} = \frac{0.5}{2 \cdot 0.2093} = 1.19$$

Total:

$$LR = 2.450 \cdot 1.805 \cdot 1.195 = 5.285$$

The markers are the three first from the table A.1 in appendix A.2, which is similar to table 2.5 in [3]. The rest of the allele frequencies can be found at `http://familias.name/Table2.5.fam`.

⋆

2.2.4 Match probabilities

Sometimes, for example in crime cases, we want to compare two profiles to find out if they come from the same person. If we have a stain from a crime scene and a reference sample from a suspect, we can calculate a match probability, i.e. the probability that the profiles come from two randomly selected individuals that happens to have identical genotypes by chance. This is called RMP, random match probability, and can be used as an alternative to the LR. If there is a match between the two samples, the RMP for a given marker is calculated as the observed genotype's frequency. For example, if both stain and suspect has a genotype A/B, we calculate the RMP as $\Pr(G = A/B)$. If there is no match, the suspect is considered excluded, and no number is reported.

2.2.5 Hardy-Weinberg equilibrium

Motivated by the two previous sections, we understand that it is necessary for the computations that we know the genotype frequencies in the general population. As these are generally unknown, we use databases to estimate them. Because there are so many different combinations of the alleles, especially with highly polymorphic markers, and the databases are not large enough to encounter them all, we have to settle for estimates of the allele frequencies. An example of why it is often impossible to estimate genotype frequencies can be seen from the marker SE33 in the NorwegianFrequencies database in the R package 'Familias'. This particular marker has 55 different alleles, which means there are $\binom{55}{2} + 55 = 1540$ different genotypes.

This means that we need to estimate allele frequencies. Once we have estimates of the allele frequencies we can assume *Hardy-Weinberg equilibrium* (HWE). This makes it possible to estimate the genotype frequencies from these allele frequencies. If we have a diallelic marker with alleles 1 and 2 that have frequencies p and q respectively, the genotypes would occur with these probabilities under HWE:

$$\Pr(G = 1/1) = p^2$$

$$\Pr(G = 1/2) = 2pq$$

$$\Pr(G = 2/2) = q^2$$

We also observe that $p^2 + 2pq + q^2 = 1$, as a consequence of there being only two alleles, so that $p + q = 1$. In general, if we have $p_1, ..., p_n$ alleles, the genotype frequencies are found from the expansion of $(p_1 + ... + p_n)^2$. For instance with three alleles:

$$(p_1 + p_2 + p_3)^2 = 1 \quad \Rightarrow \quad p_1^2 + p_2^2 + p_3^2 + 2p_1p_2 + 2p_1p_3 + 2p_2p_3 = 1$$

2.2.6 Testing of markers

A lot of tests are done on the markers before they are used, including tests for HW equilibrium. There are several methods of doing this, here we will focus on the chi-square test. This test examines the difference between observed genotype counts and the counts expected under HWE.

Example 2.2. An example of a dataset for one marker is given in table 2.1 on the next page. The test's null hypothesis is H_0: HWE applies, against the alternative H_1: HWE does not apply. We assume the null hypothesis to be true, and we use the test to see if there are significant deviations from what we would then expect to see. The conventional chi-square test-statistic for the data in table 2.1 on the following page is as follows:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(30-25)^2}{25} + \frac{(43-50)^2}{50} + \frac{(27-25)^2}{25}$$
$$= 2.14$$

Since we have $k = 3$ categories (genotypes), the degrees of freedom are $3 - 1 = 2$. This assumes that we disregard uncertainty in the estimate of p. If p had to be estimated from table 2.1, the degrees of freedom would have been 3-1-1=1. With $\alpha = 0.05$ we have to compare our calculated χ^2 with 5.99. Since 2.14 is less than 5.99 we can not reject the null hypothesis of HWE. This result is often used to argue that one can use the HW method for estimating genotype frequencies from allele frequencies.

⋆

The p-value for the test in example 2.2 is 0.343, which is the probability of observing a χ^2 of 2.14 or higher, if it is true that HW equilibrium applies.

Table 2.1: One SNP marker simulated for 100 people, and the corresponding expected counts under HWE. The allele frequencies are $p = q = 0.5$.

Genotype	Observed	Expected
1/1	30	25
1/2	43	50
2/2	27	25

When we test in this way, the consequence is that the burden of proof lies with anyone claiming the system to be out of HWE. It is not obvious that it is correct to assume HWE in the testing. This is especially tricky without an accompanying power calculation to tell us if we would actually discover a deviation from HWE had there been one.

Another problem we quickly encounter when performing this test, is a problem with sparse tables. The chi-square test presented here uses an asymptotic approach that relies on there being enough observations in each cell. A common rule of thumb is that it should be at least 5 observations in each cell, but this will often not be the case with real data. There are however other methods of testing for HWE that do not have this problem, but we will not go into them in any detail in this thesis.

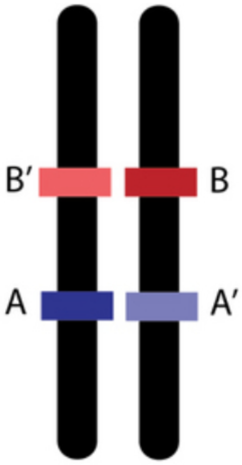


Figure 2.6: Two different haplotypes for the same chromosome. Source: `http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/HAPSEG/1?print=yes`

Another type of test of the markers is a test for linkage disequilibrium (LD). Unlike the HWE test that is concerned with independence *within* markers, the test for LD examines independence *between* different markers. Nevertheless, the test for LD is analoguous to the HWE test, as we also here compare observed and expected counts. The null hypothesis here is H_0: there is linkage equilibrium, against the alternative H_1: there is linkage disequilibrium.

To carry out the test we need to observe *haplotypes*. A haplotype is a specific combination of alleles from different markers located on the same chromosome [3], see figure 2.6 on the preceding page for an illustration. For simplicity we use male X-chromosomes in this example, as it is easy to know the haplotypes due to the fact that males only have the one X-chromosome.

Table 2.2: A haplotype is the combination of alleles from different markers located on the same chromosome. For any given chromosome we denote the allele frequencies as p_{ij} for allele j from marker number i. The table shows an example with a chromosome with 2 markers, each with 2 alleles. These can combine to give 4 different haplotypes, as seen in table 2.3.

	Marker 1	Marker 2
Allele 1	p_{11}	p_{21}
Allele 2	p_{12}	p_{22}

Example 2.3. Consider two SNP markers on the X-chromosome observed in 100 males. The alleles are denoted 1 and 2 for both markers, with allele frequencies of $p_{11} = 0.4$ and $p_{21} = 0.3$ (see table 2.2 for notation). The individuals each have one of the four possible haplotypes stated in table 2.3. If there is linkage equilibrium (LE), the frequencies of these haplotypes can be found from the marginal, for example $\Pr(1-1) = p_{11}p_{21}$. The expected haplotype count is found by multiplying the expected haplotype frequency with the total number of observations. Once you have the expected counts it is the same procedure as with the test for HWE:

$$\begin{aligned}\chi^2 &= \frac{(10-12)^2}{12} + \frac{(25-28)^2}{28} + \frac{(30-18)^2}{18} + \frac{(35-42)^2}{42} \\ &= 9.82\end{aligned}$$

Here we have 4 - 1 = 3 degrees of freedom (again assuming that the parameters are known from other data), and with an α of 0.05 the critical value is 7.81. Since the calculated $\chi^2 = 9.82$ is greater than the critical value 7.81, we can reject the null hypothesis of linkage equilibrium. This means that we can not use the marginal distribution to estimate the haplotype frequencies. More severe is the fact that this will affect all our calculations, including the LR [8]. This is because the combination of markers with multiplication depends on the markers being independent, as discussed in section 2.2.3. Section 7.4 of [3] presents a model, the so called lambda model, that can be used to estimate haplotype frequencies in the presence of potential LD.

Table 2.3: Example haplotype counts for 100 males for two SNP markers on the X-chromosome (made up data). The allele frequencies used to calculate expected counts are p_{11} = 0.4 and p_{21} = 0.3. For example, if the system is in linkage equilibrium we expect $100 \cdot 0.4 \cdot 0.7 = 28$ *of the haplotype "1 - 2".*

Haplotype	Observed	Expected
1 - 1	10	12
1 - 2	25	28
2 - 1	30	18
2 - 2	35	42

⋆

Unfortunately there is a great deal of confusion related to independence in genetics (see [16]). On one side, independence is related to markers being *unlinked*, which means that they are passed on independently in a pedigree, as opposed to as a unit (linkage). Linkage typically comes into play when meioses in one individual affect several individuals, e.g. two brothers. On the other side, there can also be dependence on the population level, in which case it is called linkage disequilibrium.

It is not straightforward to formulate simple rules for when we must assume no linkage to be allowed to multiply the LRs. In general, we require both that markers are not linked and in linkage equilibrium for independence to hold, but for paternity cases it is often sufficient with LE. In there is LD, however, we would need the requirement of no linkage to be met.

2.3 The distribution of the LR

Now we turn to the more mathematical part of the thesis, focusing on the LR. After we have calculated an LR, what interests us is the exceedance probability $\Pr(LR \geq t|H_i)$. This is the probability of getting an LR of t or higher given the hypothesis H_i. In this thesis we will write these probabilities as $\Pr(\mathbb{X}_i \geq t)$.

As mentioned in the introduction, there are two types of mistakes we can make in a paternity test. One type is to conclude that a person is not the biological father of a child when in fact he is. The probability of making this error, $\Pr(\mathbb{X}_1 < t)$, is considered easy to compute and we will therefore not go into any detail about this. On the other hand, the probability $\Pr(\mathbb{X}_2 \geq t)$, of committing the other error of claiming that an unrelated man is the father, is not trivial to find. This is the general problem of estimating small probabilities that thesis focuses on.

There are different approaches to finding these probabilities, and in this section we will go more into detail on four such methods. Here we think of the LR as a random variable and approximate or find its distribution with the various methods.

2.3.1 Exact distribution

We start with the exact approach. In other words, this is not an approximation, but a way of unambiguously finding the true probabilities for all possible values the LR can take on. This is a manageable task when the number of markers is low, but as we will see, rapidly becomes a complex problem whit increasing numbers of markers.

Example 2.4. Consider a paternity case with only one marker, and with the hypotheses still as stated in equation 2.1 on page 14. The marker used is diallelic with allele frequencies p and q. All possible combinations of AF-CH genotypes and the resulting LRs are stated in table 2.4 on the following page.

In the column H_1 we find the probabilities $\Pr(G_{CH}|G_{AF}, H_1) \cdot \Pr(G_{AF}|H_1)$, whereas $\Pr(G_{CH}|G_{AF}, H_2) \cdot \Pr(G_{AF}|H_2)$ is stated in the H_2 column. Since the hypotheses does not affect the probability of AF's genotype, $\Pr(G_{AF}|H_1) = \Pr(G_{AF}|H_2) = \Pr(G_{AF})$. There are some symmetric cases in the table that gives the same LR, e.g. line 2 and 4, but we have chosen to list both cases.

Table 2.4: Duo case with one diallelic marker with allele frequencies $p_A = p$ and $p_B = q$. Column three and four shows $\Pr(G_{CH}|G_{AF}, H_i) \cdot \Pr(G_{AF}|H_i)$ for the specified hypothesis.

G_{AF}	G_{CH}	H_1	H_2	LR
(A/A)	(A/A)	$p \cdot p^2$	$p^2 \cdot p^2$	1/p
(A/A)	(A/B)	$q \cdot p^2$	$2pq \cdot p^2$	1/(2p)
(A/A)	(B/B)	0	$q^2 \cdot p^2$	0
(A/B)	(A/A)	$0.5p \cdot 2pq$	$p^2 \cdot 2pq$	1/(2p)
(A/B)	(A/B)	$0.5(p+q) \cdot 2pq$	$2pq \cdot 2pq$	1/(4pq)
(A/B)	(B/B)	$0.5q \cdot 2pq$	$q^2 \cdot 2pq$	1/(2q)
(B/B)	(A/A)	0	$p^2 \cdot q^2$	0
(B/B)	(A/B)	$p \cdot q^2$	$2pq \cdot q^2$	1/2q
(B/B)	(B/B)	$q \cdot q^2$	$q^2 \cdot q^2$	1/q

For example, consider the first line. Under H_1, when AF is the father, the child inherits an A from him with probability 1. The other A allele the child gets from the mother with probability p from the general population, as she is not genotyped. The probability for AF's genotype A/A is p^2 if we assume HWE. That gives us $1p \cdot p^2$. Under H_2, $\Pr(G_{CH}|G_{AF}, H_2) = \Pr(G_{CH}|H_2)$. We use the HW probabilities for the genotypes, i.e. p^2 for both the father's and child's genotype.

In the cases where $p = q = 0.5$, the LR sample space is $\{0, 1, 2\}$. Table 2.5 shows the probability of getting each of these LRs given the two hypotheses.

Table 2.5: Table for the basic introductory example

t	$\Pr(\mathbb{X}_1 = t)$	$\Pr(\mathbb{X}_2 = t)$
0	0.00	0.125
1	0.75	0.75
2	0.25	0.125

The numbers in table 2.5 are computed like this:

$$\Pr(\mathbb{X}_1 = 0) = 0 + 0 = 0$$

$$\Pr(\mathbb{X}_1 = 1) = qp^2 + qp^2 + pq + pq^2 + pq^2 = 0.75$$

$$\Pr(\mathbb{X}_1 = 2) = p^3 + q^3 = 0.25$$

$$\Pr(\mathbb{X}_2 = 0) = p^2q^2 + p^2q^2 = 2 \cdot 0.0625 = 0.125$$

$$\Pr(\mathbb{X}_2 = 1) = 2qp^3 + 2qp^3 + 4p^2q^2 + 2pq^3 + 2pq^3 = 0.75$$

$$\Pr(\mathbb{X}_2 = 2) = p^4 + q^4 = 0.125$$

In other words, each probability in table 2.5 is the sum of the chances of getting an LR of t given the hypothesis in question.

⋆

The problems for the exact approach arises when we combine several markers, each with its own partial LR, by multiplication. When we want to calculate the exact probability of getting a specific LR, we have to consider every combination of the partial LR_is from each marker that can result in the LR. For the highest and lowest values, we can easily calculate the exact probability by hand. Let:

$$X = \text{number of markers where } \mathrm{LR}_i = 0$$

$$p = \Pr(LR_i = 0)$$

Then X is binomially distributed, i.e.:

$$X \sim Bin(p, n)$$

$$\Pr(\mathbb{X}_2 = 0) = \Pr(X > 0) = 1 - \Pr(X = 0) = 1 - (1 - p)^n$$

Similarly

$$\Pr(\mathbb{X}_2 = LR_{max}) = \Pr(LR_i = LR_{i_{max}})^n$$

The second lowest and second highest values can also be calculated by hand. The second lowest value occurs if all markers have their second lowest LR (assuming the lowest value is 0):

$$LR_{sl} = \text{The second lowest LR}$$

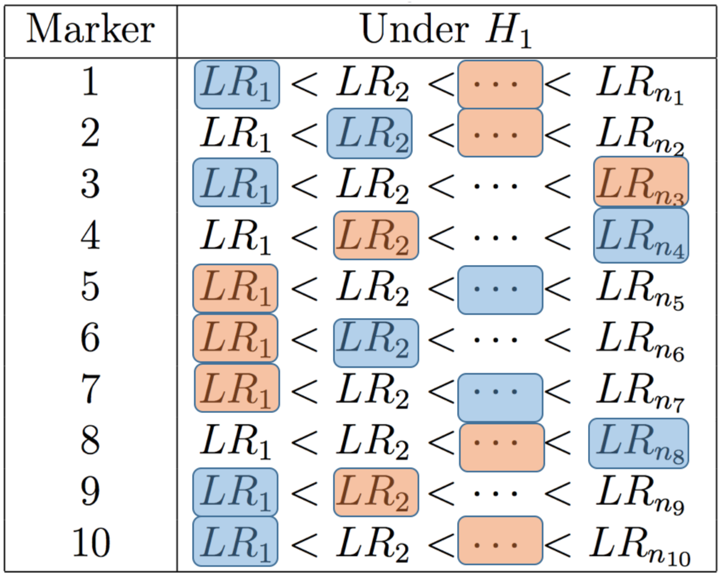
$$\Pr(\mathbb{X}_2 = LR_{sl}) = \Pr(LR_i = LR_{i_{sl}})^n$$

To get the second highest LR, all markers will have to have the highest value, except one that has the second highest value:

$$LR_{sh} = \text{The second highest LR}$$

$$\Pr(\mathbb{X}_2 = LR_{sh}) = n \Pr(LR_i = LR_{i_{max}})^{n-1} \Pr(LR_i = LR_{i_{sh}})$$

The probability of getting the maximum LR, $\Pr(\mathbb{X}_2 = LR_{max})$, is easy to calculate even when the markers have different distributions: as every marker will have to result in its maximum LR, $\Pr(\mathbb{X}_2 = LR_{max}) = \Pi_{i=1}^{n} \Pr(\mathbb{X}_2 = LR_{i_{max}})$



Marker	Under H_1
1	$LR_1 < LR_2 < \cdots < LR_{n_1}$
2	$LR_1 < LR_2 < \cdots < LR_{n_2}$
3	$LR_1 < LR_2 < \cdots < LR_{n_3}$
4	$LR_1 < LR_2 < \cdots < LR_{n_4}$
5	$LR_1 < LR_2 < \cdots < LR_{n_5}$
6	$LR_1 < LR_2 < \cdots < LR_{n_6}$
7	$LR_1 < LR_2 < \cdots < LR_{n_7}$
8	$LR_1 < LR_2 < \cdots < LR_{n_8}$
9	$LR_1 < LR_2 < \cdots < LR_{n_9}$
10	$LR_1 < LR_2 < \cdots < LR_{n_{10}}$

Figure 2.7: $\Pr(LR = t|H_i)$ *is found as the sum of the probabilities for all paths whose product is* t*. The number of terms in the sum is very big when we use many markers, because there are so many possible paths to the same LR. Shown in the figure are two different paths (under* H_1*), marked in red and blue, that nonetheless can result in the same total LR.*

For the rest of the possible LRs, however, it can be much more complicated. In figure 2.7 we have illustrated this problem for 10 markers. This can be 10 markers following the same distribution, or 10 different distributions with n_i possible values each. If we call specific combinations of LR_is from different markers for paths, then each path will occur with the probability $\prod_{i=1}^{n} \Pr(LR_i)$, assuming all markers to be independent. $\Pr(LR = t|H)$ is found as the sum of the probabilities for all paths whose product is t. When many markers are used, the number of possible paths can get very high. To find all possible paths with a brute force method is very computer intensive, which is illustrated in [2]. In [9] Kruijver presents a more efficient method for exact computations, but this also has an upper limit with respect to number of markers. The result is that this approach only works in simple cases with a small number of markers, but will generally not work. Hence, alternative methods are needed.

2.3.2 Asymptotic

Another possible approach to the problem of small probabilities is to use an approximation to the normal distribution. The central limit theorem states that the mean of independent random variables from the same distribution is approximately normally distributed if the sample size is big enough. As the samples size increases, the distribution of the mean approach the normal distribution asymptotically. This relation holds regardless of the distribution of the random variables themselves. That means that if:

$$E(X) = \mu \qquad \wedge \qquad SD(X) = \sigma \qquad \text{then}$$

$$E(\bar{X}) = \mu \qquad \wedge \qquad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The theorem is also valid for sums of independent random variables. Then the outcome is as follows:

$$E(\sum_{i=1}^{n} X_i) = n\mu \qquad SD(\sum_{i=1}^{n} X_i) = \sqrt{n}\sigma$$

Applied to the likelihood ratio, it means that the LR is not covered by the central limit theorem, but U = log(LR) is approximately normally distributed if enough markers are used, with $E(U) = n\mu$ and $SD(U) = \sqrt{n}\sigma$ in accordance with the theorem.

$$LR = \prod_{i=1}^{n}(LR_i)$$

$$U = \log(LR) = \log(\prod_{i=1}^{n}(LR_i)) = \sum_{i=1}^{n}(\log(LR_i))$$

Now $\mu = E(\log(LR_i))$ and $\sigma^2 = Var(\log(LR_i))$, and can be found through the formulas for discrete distributions.

Example 2.5. Let LR_i be the marker from table 2.5 on page 24 under H_1 and lr_i be a specific value that LR_i can take. Then

$$\begin{aligned}
\mu = E(\log(LR_i)) &= \sum_{i=1}^{n} \log(lr_i) \Pr(LR_i = lr_i) \\
&= \log(1) \cdot 0.75 + \log(2) \cdot 0.25 \\
&= 0.1732868 \\
\sigma^2 = Var(\log(LR_i)) &= \sum_{i=1}^{n} \log(lr_i)^2 \Pr(LR_i = lr_i) - \mu^2 \\
&= \log(1)^2 \cdot 0.75 + \log(2)^2 \cdot 0.25 - 0.1732868^2 \\
&= 0.09008494
\end{aligned}$$

If we use 5 such markers, then

$$\begin{aligned}
E(U) &= n\mu = 5 \cdot 0.1732868 = 0.866434 \\
SD(U) &= \sqrt{n\sigma^2} = \sqrt{5 \cdot 0.09008494} = 0.6711369
\end{aligned}$$

⋆

We find approximations to the exceedance probabilities through the normal distribution:

$$\begin{aligned}
\Pr(LR > t|H) &= \Pr(\log(LR) > \log(t)|H) = 1 - \Pr(\log(LR) \leq \log(t)|H) \\
&\approx 1 - \phi(\frac{\log(t)\text{-}n\mu}{\sqrt{n}\sigma})
\end{aligned}$$

In figure 2.8 we used this method for 10 iid markers under H_1 following the distribution in table 2.5 on page 24, and compared it to the exact approach from [9], and we see that the approximation is not too good even for the higher probabilities.

There are however some problems with this asymptotic approach. One of the problems is that we are primarily interested in the distribution of the LR under H_2. The problem with this lies in that 0 is often a possible value for the LR under H_2, but we can not take the logarithm of 0. We would

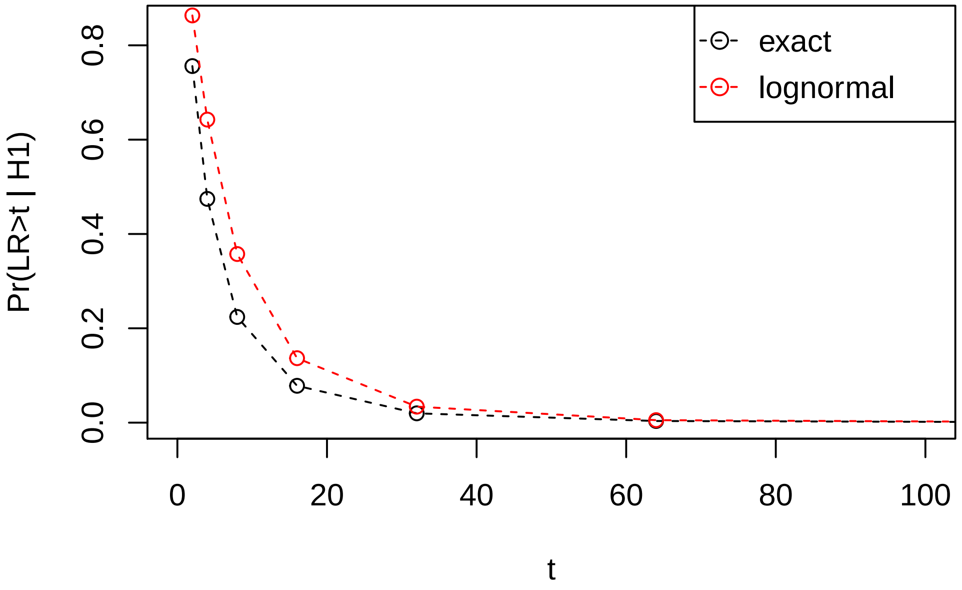


Figure 2.8: $\Pr(LR > t|H_1)$*, found by the normal approximation and the exact method. Both methods have been used on the same ts, so the red and black points go together pairwise vertically. The data is based on 10 markers, all following the distribution in table 2.5 on page 24.*

then be forced to make an asymptotic mixture distribution of the normal distribution and one where 0 is a possibility.

A second, maybe bigger, problem is that the central limit theorem, at least in its simplest form, requires the random variables to follow the same distribution. This means that all markers we use have to have the same allele frequencies for the theorem to apply, which is not very realistic. It could be possible to use more advanced versions of the theorem, but we have chosen to not follow this lead. It would anyway be dubious to rely on estimates from these types of methods when the markers are not iid.

Another reason to give up this approach is that log-normal distributions are known to be heavy tailed, and it is exactly the tails of the distribution that we take interest in. This is especially a problem when the sample size is not big enough. It is highly uncertain that n = 16 markers, which is a normal number of markers, is enough for the distribution to converge to the normal distribution. We therefore move on to look at other possible approaches.

2.3.3 Monte Carlo simulation

By Monte Carlo simulation we mean straightforward simulation as is explained below. Simulation can be a good tool for solving problems that are too complex to solve analytically. As we saw in section 2.3.1 on the exact approach, our problem should then be a good candidate for a simulation study. It can be used to approximate probabilities by running multiple mock trials on a computer. Monte Carlo simulation (often used synonymously with Stochastic simulation [13]) usually involves dividing the problem into smaller subproblems where we know the distribution of the stochastic variables involved. With the distributions and a random number generator, the computer can mimic the stochastic variables so that we get a view of the different possible outcomes of the system as a whole.

We have a somewhat different application in mind, as we will use simulation to find probabilities. These probabilities can be written as sums (for the discrete case) or as integrals (continuous case). However, as we have seen in the previous section, these sums and integrals are too hard to calculate analytically and therefore we resort to simulation.

Example 2.6. We can for example use simulation to estimate the probability of getting a sum smaller than seven when rolling two dice. Let S be the sum of both dice and assume H = the throws are independent. We want to find $\Pr(S < 7|H)$ by simulation. In R it could look like this:

```
set.seed(5)
N=10
s1 <- sample(1:6, N, replace=TRUE)
s2 <- sample(1:6, N, replace=TRUE)
s <- s1 + s2
p <- mean(as.integer(s<7))

>p
[1] 0.3
```

Here we 'throw' two dice 10 times each and find the frequency of hits. We get an estimate of 0.3, but it is not a very good one. The answer can in this case easily be computed exactly by hand, it is approximately 0.4167. As is intuitively understood (and as a result of the law of large numbers), the number of trial runs is crucial to the accuracy of the estimate, and 10 runs is far from enough. Another try:

```
dice <- function(nsim, N, seed = FALSE) {
  if (seed) {
    set.seed(seed)
  }
  p.vec <- NULL
  for (i in 1:nsim){
    s1 <- sample(1:6, N, replace=TRUE)
    s2 <- sample(1:6, N, replace=TRUE)
    s <- s1 + s2
    p <- mean(as.integer(s<7))
    p.vec <- c(p.vec, p)
  }
  return(p.vec)
}

nsim <- 100
N <- 10000
res <- dice(nsim, N, 8)
```

This code makes 100 estimates, each based on 10 000 throws. A higher number of simulations not only increases the accuracy of the estimate, but it also lets us study its variability with greater precision. With several estimates it is possible to look at the distribution of the estimates, and that gives us more information of the quality of the estimate. Figure 2.9 on the following page shows a histogram of the 100 estimates made in the code above to show how they are distributed.

When evaluating simulation methods it is always a good idea to start with small problems where it is possible to find the exact value. That way we have a way to know how much the estimate misses. We will come back to methods for evaluating the estimates in section 2.4.

⋆

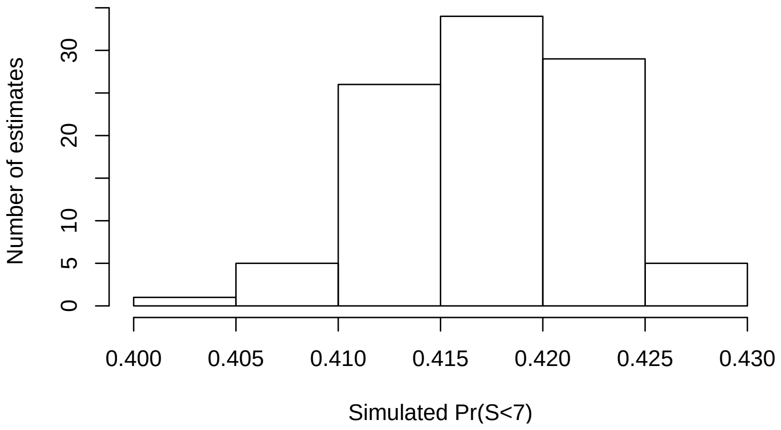


Figure 2.9: The distribution of the 100 estimates made in the code above.

However, not all problems are suitable for this kind of simulation. This issue is something that will be explored closely in this thesis, more specifically when the reason is that the probabilities are too small to be detected.

Example 2.7. We can look at an example of this: if it is stated that $U \sim N(0, 1)$ and we want to estimate the probability that U exceeds 1, $\Pr(U > 1)$, Monte Carlo simulation could be used. Random numbers from U's distribution are drawn and we count up the number of draws that are greater than 1. This would work pretty well because the probability of interest is of a 'reasonable' size. If instead we wanted to estimate the probability that this U would exceed 5, $\Pr(U > 5)$, we would with Monte Carlo simulation most likely not encounter a U bigger than 5 at all, and then the estimate would be 0. This is not because the event is impossible, it is just extremely rare.

⋆

In situations like the one in example 2.7, when the probability that is being simulated is very small, the Monte Carlo method often falls short.

2.3.4 Importance sampling

With this in mind, we have to look to other methods to more precisely estimate these kinds of small probabilities. One such method is importance sampling, and the basic method is described in [9] and [13]. The basic idea is to sample from another probability distribution than the one you originally are interested in, and then adjust to the distribution of interest, see figure 2.10 on the next page.

Put in general terms, we always want to estimate some expectation of the type $Ef(x)$ where $x \sim p$ [13]. The exact expression for this expectation is

$$Ef(x) = \int_{-\infty}^{\infty} f(x)p(x)dx$$

With Monte Carlo simulation we would estimate this as the mean of a sample:

$$Ef(x) \approx \frac{1}{N}\sum_{i=n}^{n} f(x_i) \qquad \text{where } x_i \sim p$$

To find the importance sampling estimator, we first rewrite the exact expression:

$$\begin{aligned} Ef(x) &= \int_{-\infty}^{\infty} f(x)p(x)dx \\ &= \int_{-\infty}^{\infty} f(x)\ \frac{p(x)}{q(x)}q(x)dx \\ &= \int_{-\infty}^{\infty} f_*(x)\ q(x)dx \qquad \text{where } f_*(x) = f(x)\frac{p(x)}{q(x)} \end{aligned}$$

Mathematically we do nothing illegal here, it is allowed for any probability distribution $q(x)$ such that $q(x) = 0 \Rightarrow p(x) = 0$. We can now estimate this expression in a Monte Carlo fashion, but it would now be called the importance sampling estimate. The index 'b' indicates that the biasing distribution is used.

$$\begin{aligned} E_b f(x) &\approx \frac{1}{N}\sum_{i=N}^{n} f_*(x_i) \\ &\approx \frac{1}{N}\sum_{i=N}^{n} f(x_i)\frac{p(x_i)}{q(x_i)} \end{aligned}$$

In other words, this is still an estimate of $Ef(x)$ with $x \sim p$, the difference is that we now draw the x_is from q's distribution, but importantly we correct

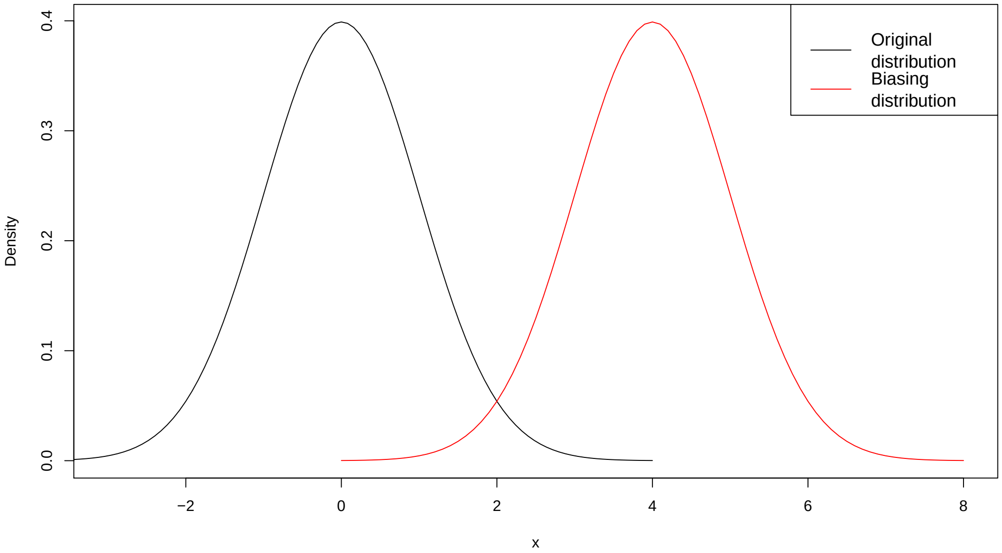


Figure 2.10: Say we want to know $\Pr(X > 5)$*, then the number of simulations would have to be very high for us to expect to get even one hit. If instead we sample from the distribution with for example* $\mu = 4$*, it is much more likely to see* $X > 5$*.*

for doing so. q is called the biasing distribution, and the term $\frac{p(x)}{q(x)}$ is called the importance weights.

Let $\theta = Ef(x)$. As with the Monte Carlo method, this will give an unbiased estimate, but they will not have the same variance:

$$Var(\hat{\theta}_{MC}) = \frac{1}{N} Var(f(x))$$

$$Var_b(\hat{\theta}_I) = \frac{1}{N} Var_b(f_*(x)) = \frac{1}{N} Var_b(f(x)\frac{p(x)}{q(x)})$$

Here $\hat{\theta}_{MC}$ and $\hat{\theta}_I$ are the estimators for the Monte Carlo and importance sampling methods, respectively.

Importance sampling is a method for *variance reduction*. This is achieved by choosing the biasing distribution q wisely, so that it reduces the variance of $f_*(x)$ such that $Var_b(f_*(x)) < Var(f(x))$. By reducing the estimator's variance, we improve its effectiveness. That means that less simulations are needed, which reduces the running time.

Example 2.8. Let us revisit example 2.7 from the previous section and try to find $\Pr(U > 5|\mu = 0) = E(I(U > 5))$, but now with importance sampling drawing from the distribution with $\mu = 5$. Since the problem is that the probability is small, we want to choose a biasing distribution where the desired outcome occurs more frequently. The Monte Carlo estimator would be:

$$E[I(U > 5)] \approx \frac{1}{N}\sum_{i=N}^{n} I(U_i > 5) \qquad\qquad \text{where } U_i \sim N(0, \sigma)$$

We rewrite the exact expression, using $\phi_\mu(U)$ for the normal density function with expectation μ:

$$\begin{aligned} E[I(U > 5)] &= \int_{-\infty}^{\infty} I(U > 5)\phi_0(U)dU \\ &= \int_{-\infty}^{\infty} I(U > 5)\frac{\phi_0(U)}{\phi_5(U)}\phi_5(U)dU \\ &= \int_{-\infty}^{\infty} f_*(U)\phi_5(U)dU \end{aligned}$$

Which means that the importance sampling estimate is

$$\begin{aligned} E_b[I(U > 5)] &\approx \frac{1}{N}\sum_{i=N}^{N} f_*(U_i) \\ &\approx \frac{1}{N}\sum_{i=N}^{N} I(U_i > 5)\frac{\phi_0(U_i)}{\phi_5(U_i)} \qquad\qquad \text{where } U_i \sim N(5, \sigma) \end{aligned}$$

This process, from sampling to estimate, is shown step-by-step in table 2.6 on the following page. Instead of drawing U_is from the distribution with $\mu = 0$ and hoping to encounter enough $U_i > 5$, we draw from the distribution with $\mu = 5$, where it of course is much more likely to see $U_i > 5$. This will be a more accurate estimate than the one we get with Monte Carlo simulation, because this U is more frequently occurring in the distribution it is drawn from.

Table 2.6: The process of making an estimate of $\Pr(U > 5|\mu = 0)$ *with importance sampling. The final estimate is the mean of the last column.*

	U	$I(U > 5)$	$\frac{\phi_0(U)}{\phi_5(U)}$	$I(U > 5)\frac{\phi_0(U)}{\phi_5(U)}$
1	6.22	1	8.38e-09	8.38e-09
2	4.88	0	6.94e-06	0
3	4.04	0	4.47e-04	0
4	4.92	0	5.54e-06	0
5	3.86	0	1.13e-03	0
6	3.94	0	7.36e-04	0
7	4.62	0	2.55e-05	0
8	5.61	1	1.78e-07	1.78e-07
9	6.67	1	8.70e-10	8.70e-10
10	3.46	0	8.12e-03	0

⋆

Next we explain how importance sampling is used in our context. Our hypotheses are as stated in section 2.2.2, namely H_1: AF is the father, versus H_2: a man unrelated to AF is the father. We would like to calculate $\Pr(LR > t|H_2)$, and this probability, the probability of declaring an unrelated man to be the father, is small.

Example 2.9. Say we set the threshold t to 10^6. If $LR > 10^6$ then we conclude that AF is the father, if it is less we conclude that he is not. We want to know the probability of wrongly concluding that he is the father, i.e. $\Pr(LR > t|H_2) = E[I(LR > t)|H_2]$ (the conditioning is dropped in the notation below). The Monte Carlo estimator and its variance is:

$$E[I(LR > t)] \approx \frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)$$

$$\begin{aligned} Var(\hat{\theta}_{MC}) &= \frac{1}{N}\left[E[\hat{\theta}^2_{MC}] - (E[\hat{\theta}_{MC}])^2\right] \\ &= \frac{1}{N}\left[E[I(LR_i > t]^2) - (E[I(LR_i > t)])^2\right] \\ &= \frac{1}{N}\left[E[I(LR_i > t)] - (E[I(LR_i > t)])^2\right] \end{aligned}$$

$$\begin{aligned}
\widehat{Var(\hat{\theta}_{MC})} &= \frac{1}{N}\left[\frac{1}{N}\sum_{i=N}^{N} I(LR_i > t) - \left(\frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)\right)^2\right] \\
&= \frac{1}{N}\left[\hat{\theta}_{MC} - \hat{\theta}^2_{MC}\right]
\end{aligned}$$

where LR_i now denotes the ith randomly drawn LR from H_2's distribution. For the importance sampling estimator we use the LR's distribution under H_1, where the bigger LRs have a higher probability. The estimator and its variance is:

$$\begin{aligned}
E_b[I(LR > t)] &\approx \frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)\frac{\Pr(\text{Data} > t|H_2)}{\Pr(\text{Data} > t|H_1)} \\
&= \frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)\frac{1}{LR_i} \\
Var_b(\hat{\theta}_I) &= \frac{1}{N}\left[E(\hat{\theta}_I^2) - E(\hat{\theta}_I)^2\right] \\
&= \frac{1}{N}\left[E[I(LR_i > t)^2\left(\frac{1}{LR_i}\right)^2] - E[I(LR_i > t)\frac{1}{LR_i}]^2\right] \\
&= \frac{1}{N}\left[E[I(LR_i > t)\frac{1}{LR_i^2}] - E[I(LR_i > t)\frac{1}{LR_i}]^2\right] \\
\widehat{Var_b(\hat{\theta}_I)} &= \frac{1}{N}\left[\frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)\frac{1}{LR_i^2} - \left(\frac{1}{N}\sum_{i=N}^{N} I(LR_i > t)\frac{1}{LR}\right)^2\right]
\end{aligned}$$

where the LRs are drawn according to H_1.

⋆

On the optimal choice of biasing distribution

How to choose the best biasing distribution q depends a great deal on why importance sampling is used. One reason to use the method is when the distribution p that is being studied is impossible or very difficult to sample from. In that case you often want to choose a q that resembles p as much as possible, with the key difference that it is actually possible to sample from. If, on the other hand, the reason to use importance sampling is to improve your estimate and make it more stable, you want to choose a biasing distribution that minimizes the estimates variance.

There is a theoretical optimal biasing distribution, but this can often be impossible or impractical to sample from. This has been much discussed in the literature, and the wikipedia page about importance sampling (`https://en.wikipedia.org/wiki/Importance_sampling`) has some suggestions, e.g. is 'scaling' and 'translation' mentioned. However, this is not necessarily directly applicable to our application in the forensic setting.

In the forensics field there has not been done much research into the selection of a biasing distribution. The default mode is to use the probability distribution of LR under the numerator (H_1) hypothesis, but there is no theoretical reason why this should be optimal. This is something we will look further into in the results section.

2.4 Evaluation

When we have the option to choose amongst several estimators, we need to have some criterion by which we choose the best or more precise one. First, it is important to know if your estimate is biased or unbiased, i.e. if it makes any systematic errors. An analogy is often drawn to a dartboard, where the bullseye is the true value of the parameter in question, and if your estimator is unbiased, your aim is on the bullseye [14].

As we see in figure 2.11, the variance also has a lot to say for the quality of the estimator. If it has a high variance it can be unreliable even though it is unbiased. In example 2.9 on page 36 we go through how to calculate the variance. The general formula is

$$Var(\hat{\theta}) = E[(X-\theta)^2] = E(X^2) - (E(X))^2$$

The problem with this approach is that if the estimate is biased, the variance does not necessarily tell us very much. Sometimes it can even be absurd if we only compare by the variance. An example of this is when we use Monte Carlo simulation for small probabilities and the estimate is 0, then the variance is also 0. This does not necessarily mean that it is a better estimate than an estimate that lies closer to the true value but has a higher

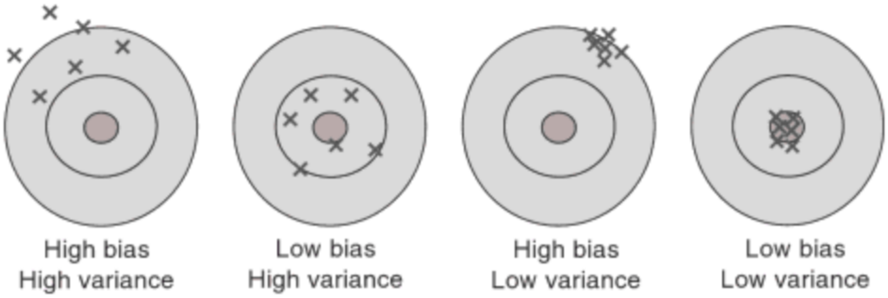


Figure 2.11: The bias of an estimator illustrated with a dartboard: an unbiased estimator aims on the 'bullseye', the parameter, whereas a biased estimator is skewed. The ultimate goal is to have an unbiased estimator with low variance. Source: [14], page 101.

variance. What is the best estimate depends on what the further use of it will be, as an estimate of 0 can clutter later calculations. This is discussed more extensively in chapter 4.

In many cases it can be useful with the mean squared error, MSE, instead of (or in addition to) the variance. This is particularly useful for biased estimators. The MSE, as the name suggests, calculates the squared distances between the estimate and the parameter value, i.e., by how much our estimate on average misses on the target.

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\ &= Var(\hat{\theta}) + [\text{bias}(\hat{\theta}, \theta)]^2 \end{aligned}$$

From the formula we see that for unbiased estimators, the MSE is always the variance of the estimate.

In table 2.7 on the following page the MSE is shown when estimating $\Pr(U > u|\mu = 0)$ using importance sampling with different biasing distributions. Two patterns emerges when we look at this table – first, the main diagonal is standing out as the smallest MSE for each row. Second, the MSE is decreasing as u is increasing. The reason for the former could in this case be that importance sampling works best when sampling from the distribution where u has the highest probability of occurring. The latter is connected

Table 2.7: MSE when estimating $\Pr(U > u|\mu = 0)$ *with importance sampling, sampling from the distribution with* $\mu = \mu_I$*. For the last line, the exact value gets so small that R rounds it of to 0. For the two first entries of that line, the estimate is also 0, which leads to the MSE also being is 0.*

u	$\mu_I = 0$	$\mu_I = 3$	$\mu_I = 5$	$\mu_I = 7$	$\mu_I = 9$
0	4.54e-06	4.23e-04	4.68e-02	2.45e-01	2.50e-01
3	1.41e-08	1.68e-10	4.62e-10	1.49e-07	1.74e-06
5	8.22e-14	8.28e-17	5.00e-18	1.22e-16	1.35e-14
7	1.64e-24	5.67e-25	1.63e-27	1.34e-28	1.09e-27
9	0.00	0.00	2.11e-38	1.31e-38	1.26e-38

to the terms absolute and relative error. As u increases further and further away from μ, the probability for U to be larger than u decreases. A small probability is easier to estimate just as an artifact of it being small. As a consequence, the *absolute* error gets smaller. The relative error on the other hand, takes into account the size of the estimate: $e_R = \frac{(\hat{\theta}-\theta)^2}{\theta}$

In table 2.8 we made the equivalent of table 2.7, but with the relative error instead of the MSE. It still has the same tendency of decreasing with increasing u, but to a lesser degree. As is seen in the last line of the table, the disadvantage of dividing by the true parameter value is that it crashes when the true value gets very small. Sometimes the sum of the absolute differences is used to avoid this.

Table 2.8: Relative error when estimating $\Pr(U > u|\mu = 0)$ *with importance sampling that uses a biasing distribution with* $\mu = \mu_I$*. For the last line, the exact value gets so small that R rounds it of to 0. For the two first entries of that line, the estimate is also 0, and '0/0' gives NaN in R. For the three last entries of the same line we divide non-zero estimates by 0, which gives 'Inf' in R.*

u	$\mu_I = 0$	$\mu_I = 3$	$\mu_I = 5$	$\mu_I = 7$	$\mu_I = 9$
0	4.09e-06	6.12e-04	2.56e-03	4.90e-01	5.00e-01
3	6.62e-07	7.90e-08	2.12e-08	9.23e-06	1.29e-03
5	2.87e-07	1.12e-10	1.23e-12	3.07e-10	1.58e-08
7	1.28e-12	8.31e-14	2.97e-16	1.08e-18	1.25e-16
9	NaN	NaN	Inf	Inf	Inf

2.5 Implementation

Here we present the software that has been used throughout the work with the thesis. All the work has been implemented in R version 3.3.0 (2016-05-03). M. Kruijver's R-package ‘DNAprofiles’ has been used extensively, both for simulation, exact probabilities, datasets for allele frequencies and distributions of LR$_i$s. We have also used the package ‘paramlink’ for plotting of pedigrees, ‘HardyWeinberg’ (presented in [7]) to simulate genotype frequencies in connection with testing for Hardy Weinberg Equilibrium and the NorwegianFrequencies database from ‘Familias’ to look into the marker SE33.

In addition we have made an R package called ‘naomi’. This package is the main contribution in this thesis as far as implementation goes. The documentation of the functions made can be found in appendix A.1, and the code needed to make data for the tables and figures in this thesis is included as examples there.

Those who are interested in a closer look at the package can email the author at naomin.azulay@gmail.com

Chapter 3

Results

Primarily, we are interested in the probability $\Pr(\mathbb{X}_2 \geq t)$ for a big t, because this is the probability of saying that there is a familial relationship when there in fact is not. We will try to simulate this with importance sampling with a varying number of markers.

Example 3.1. Let us now revisit table 2.5 on page 24. When we have one marker with this distribution, the highest possible LR is 2, so we want to simulate $\Pr(\mathbb{X}_2 = 2)$. We simulate with importance sampling using the R-function showed in figure 3.1 on the next page. The code can also be found in appendix A.1.

With the seed 123 we get an estimate of 0.12462 which is close to the exact value of 0.125. We used option 3 in the function and got that the estimate's variance is 4.677986e-07 and the MSE is 6.121986e-07.

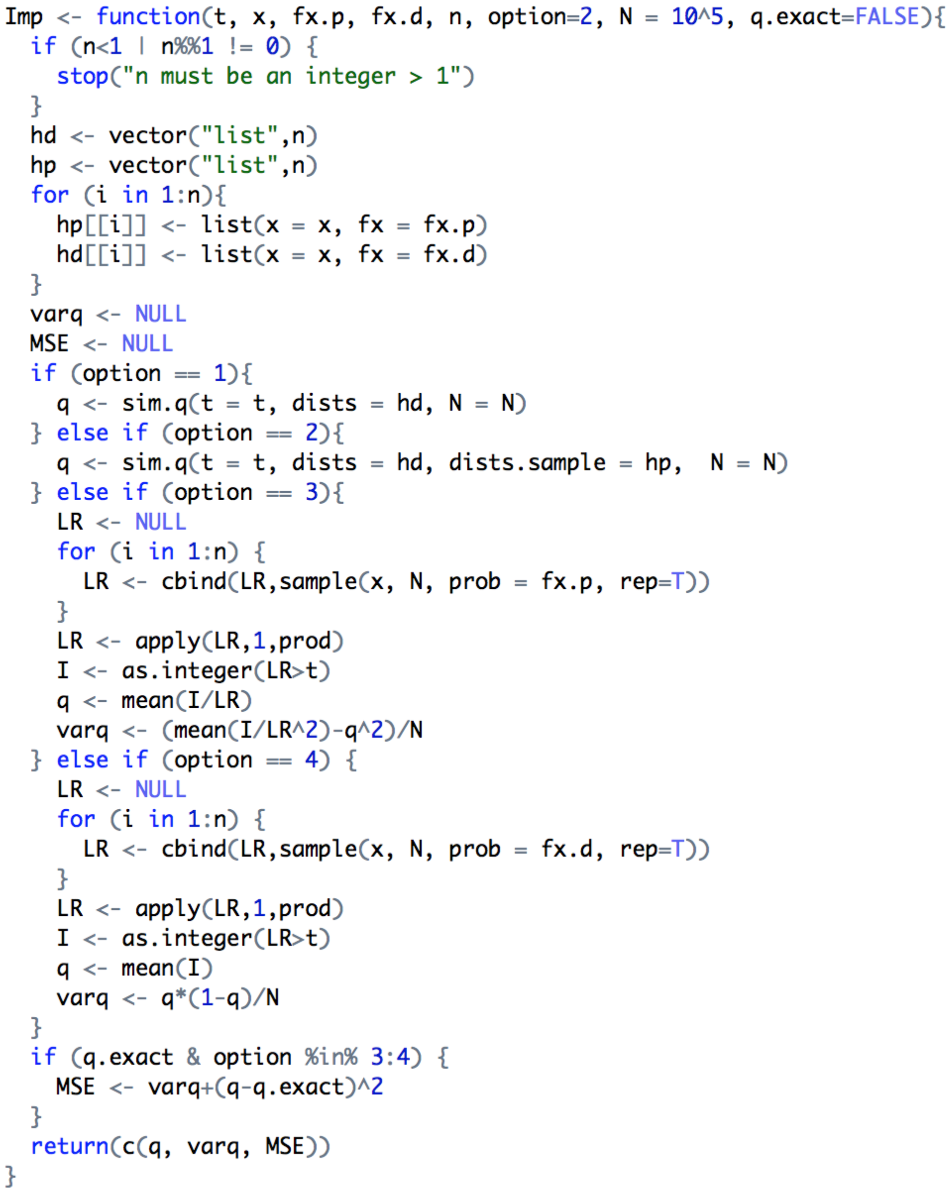
⋆

Example 3.2. Now we consider $\Pr(\mathbb{X}_2 > t)$ when the LR is based on two markers that follow the distribution in table 2.5 on page 24.

The simulation results in table 3.1 on page 45 are produced using the Imp-function from example 3.1 with option 3 and 4 respectively for importance sampling and Monte Carlo simulation. The function finds $\Pr(\mathbb{X}_i > t)$, and this is how the point probabilities was derived from the exceedance probabilities:

$$\Pr(\mathbb{X}_2 = 0) = 1 - \Pr(\mathbb{X}_2 > 0)$$

$$\Pr(\mathbb{X}_2 = 1) = \Pr(\mathbb{X}_2 > 0) - \Pr(\mathbb{X}_2 > 1)$$



```r
Imp <- function(t, x, fx.p, fx.d, n, option=2, N = 10^5, q.exact=FALSE){
  if (n<1 | n%%1 != 0) {
    stop("n must be an integer > 1")
  }
  hd <- vector("list",n)
  hp <- vector("list",n)
  for (i in 1:n){
    hp[[i]] <- list(x = x, fx = fx.p)
    hd[[i]] <- list(x = x, fx = fx.d)
  }
  varq <- NULL
  MSE <- NULL
  if (option == 1){
    q <- sim.q(t = t, dists = hd, N = N)
  } else if (option == 2){
    q <- sim.q(t = t, dists = hd, dists.sample = hp,  N = N)
  } else if (option == 3){
    LR <- NULL
    for (i in 1:n) {
      LR <- cbind(LR,sample(x, N, prob = fx.p, rep=T))
    }
    LR <- apply(LR,1,prod)
    I <- as.integer(LR>t)
    q <- mean(I/LR)
    varq <- (mean(I/LR^2)-q^2)/N
  } else if (option == 4) {
    LR <- NULL
    for (i in 1:n) {
      LR <- cbind(LR,sample(x, N, prob = fx.d, rep=T))
    }
    LR <- apply(LR,1,prod)
    I <- as.integer(LR>t)
    q <- mean(I)
    varq <- q*(1-q)/N
  }
  if (q.exact & option %in% 3:4) {
    MSE <- varq+(q-q.exact)^2
  }
  return(c(q, varq, MSE))
}
```

Figure 3.1: The function returns an estimate of $\Pr(\mathbb{X}_2 > t)$ *for n iid markers made with either importance sampling or Monte Carlo simulation. For option 3 and 4 it also outputs the estimate's variance, and if q.exact is specified, it return the MSE. The code is included because it is very general, and has been used extensively in the thesis. The restriction in this code is that the markers used have to be iid.*

$$\Pr(\mathbb{X}_2 = 2) = \Pr(\mathbb{X}_2 > 1) - \Pr(\mathbb{X}_2 > 2)$$

$$\Pr(\mathbb{X}_2 = 4) = \Pr(\mathbb{X}_2 > 3.99)$$

Since there is no likelihood between 2 and 4, any $2 \leq t < 4$ could have been used to find $\Pr(\mathbb{X}_2 = 4)$. For two markers the calculations are still possible to do by hand (see section 2.3.1), and so we have the benefit that we have the true values for comparison. From table 3.1 we see that when the number of markers is low, both Monte Carlo simulation and importance sampling produces good results. When summing up the MSEs from each row in the table, we get 3.434186e-06 for importance sampling and 1.043244e-05 for Monte Carlo simulation.

Table 3.1: $\Pr(\mathbb{X}_2 = t)$ *for two markers following the distribution in table 2.5. In the variance estimates in line 2 and 3 we ignored the covariance term. The effect is likely to be small, but the covariance term should be included in future versions.*

t	Exact	Importance	Variance	Monte Carlo	Variance
0	0.234375	0.2352000	7.410846e-07	0.2359600	1.802829e-06
1	0.562500	0.5624350	1.302311e-06	0.5632200	3.408596e-06
2	0.187500	0.1872625	6.005034e-07	0.1869900	1.770134e-06
4	0.015625	0.0155125	3.637487e-08	0.0157000	1.545351e-07

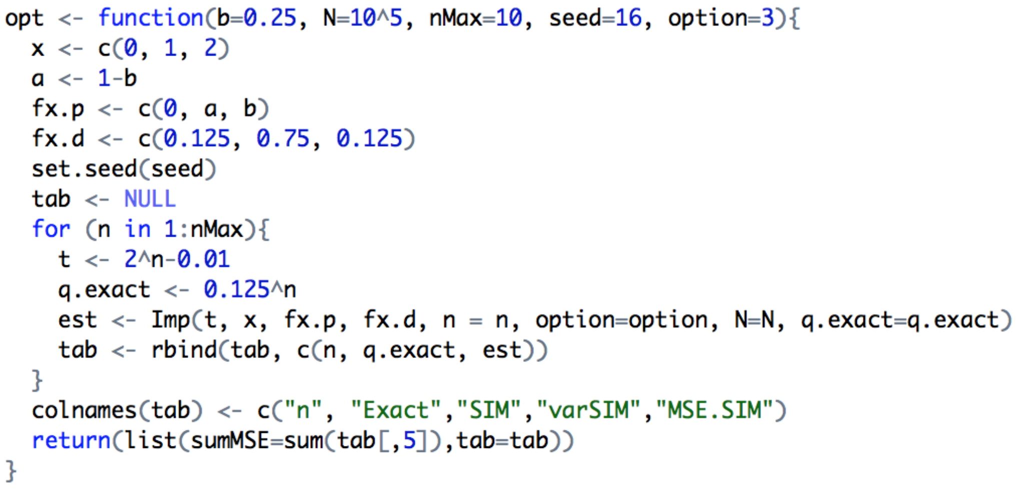
⋆

Example 3.3. In table 3.2 on the following page we estimated $\Pr(\mathbb{X}_2 = LR_{max})$ with Monte Carlo simulation and importance sampling for up to 10 markers, all following the distribution from table 2.5. We chose to look at the probability of getting LR_{max} partly because it is an easy probability to calculate by hand, making it possible to include the MSE. The table was made with the function 'opt' shown in figure 3.2, which uses the 'Imp' function in figure 3.1 to make the estimates. The sum of the MSEs was 3.24129e-06 and 9.335999e-07 for Monte Carlo simulation and importance sampling respectively.

The results show how the estimate made from Monte Carlo simulations quickly goes to zero for a small number of markers, while the estimate found with importance sampling lasts for a higher number of markers.

Table 3.2: $\Pr(\mathbb{X}_2 = 2^n)$ *for n iid markers following the distribution in table 2.5. The estimates are found with importance sampling and Monte Carlo simulation, the simulation was run 100 000 times.*

n	Exact	Importance	Variance	Monte Carlo	Variance
1	1.25e-01	1.26e-01	4.70e-07	1.26e-01	1.10e-06
2	1.56e-02	1.54e-02	3.62e-08	1.53e-02	1.51e-07
3	1.95e-03	1.93e-03	2.37e-09	2.00e-03	2.00e-08
4	2.44e-04	2.41e-04	1.50e-10	2.30e-04	2.30e-09
5	3.05e-05	3.16e-05	9.85e-12	3.00e-05	3.00e-10
6	3.81e-06	4.38e-06	6.83e-13	0.00	0.00
7	4.77e-07	7.81e-07	6.10e-14	0.00	0.00
8	5.96e-08	7.81e-08	3.05e-15	0.00	0.00
9	7.45e-09	1.95e-08	3.81e-16	0.00	0.00
10	9.31e-10	0.00	0.00	0.00	0.00



```
opt <- function(b=0.25, N=10^5, nMax=10, seed=16, option=3){
  x <- c(0, 1, 2)
  a <- 1-b
  fx.p <- c(0, a, b)
  fx.d <- c(0.125, 0.75, 0.125)
  set.seed(seed)
  tab <- NULL
  for (n in 1:nMax){
    t <- 2^n-0.01
    q.exact <- 0.125^n
    est <- Imp(t, x, fx.p, fx.d, n = n, option=option, N=N, q.exact=q.exact)
    tab <- rbind(tab, c(n, q.exact, est))
  }
  colnames(tab) <- c("n", "Exact","SIM","varSIM","MSE.SIM")
  return(list(sumMSE=sum(tab[,5]),tab=tab))
}
```

Figure 3.2: The R-function used to vary the biasing density when using importance sampling. The output is an nmax × *4 data.frame with the exact value, estimate, variance and MSE.*

We used the same function 'opt' to look at different biasing distributions when estimating $\Pr(LR > LR_{max}|H_2)$. This was done for 1 up to 20 markers of the type in table 2.5 on page 24. Earlier we used the distribution under H_1 as a biasing distribution, but now we have experimented with distributions on the form (0, 1-b, b). The b was varied over 100 different values from 0.01 to 1. The resulting MSEs are plotted in figure 3.3.

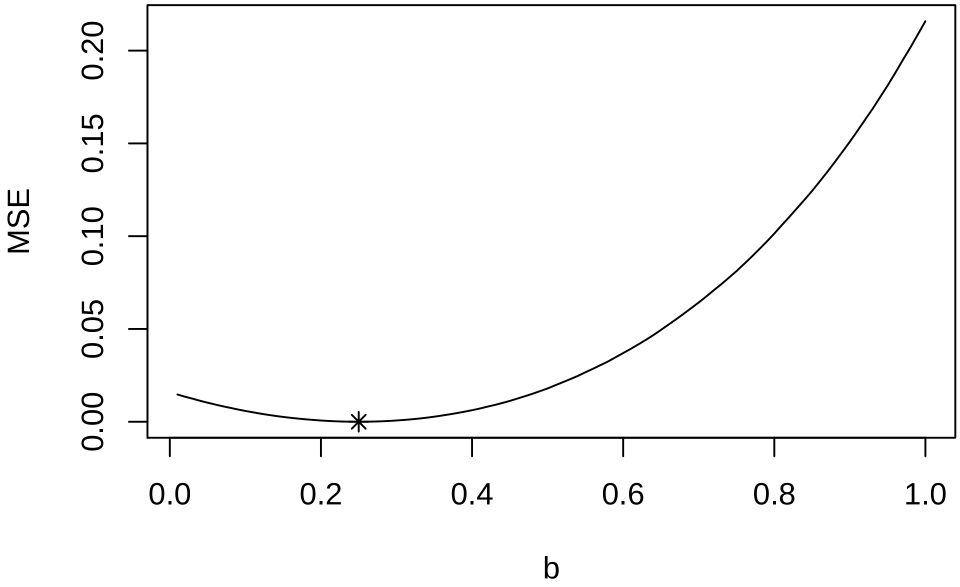


Figure 3.3: The effect on the MSE of varying the biasing distribution when estimating $\Pr(LR > LR_{max}|H_2)$ *with importance sampling. This was done for 1 up to 20 markers of the type in table 2.5 on page 24, but instead of using the distribution under* H_1 *as earlier, (0, 0.75, 0.25), we experimented with distributions on the form (0,* $1-b$*, b). We used 100 different bs that were evenly spaced in the interval [0.01, 1]. The minimal MSE is plotted with a star and occurred for* $b = 0.25$*, the distribution under* H_1*.*

In figure 3.4 we plotted the logarithm of the exact value and three different estimates of $\Pr(\mathbb{X}_2 = LR_{max})$ for up to 12 markers. One estimate was based on Monte Carlo simulation and two on importance sampling. The one named 'IMP' in the figure used H_1s distribution as biasing distribution while the other one, named 'IMP 0.5', used $b = 0.5$.

⋆

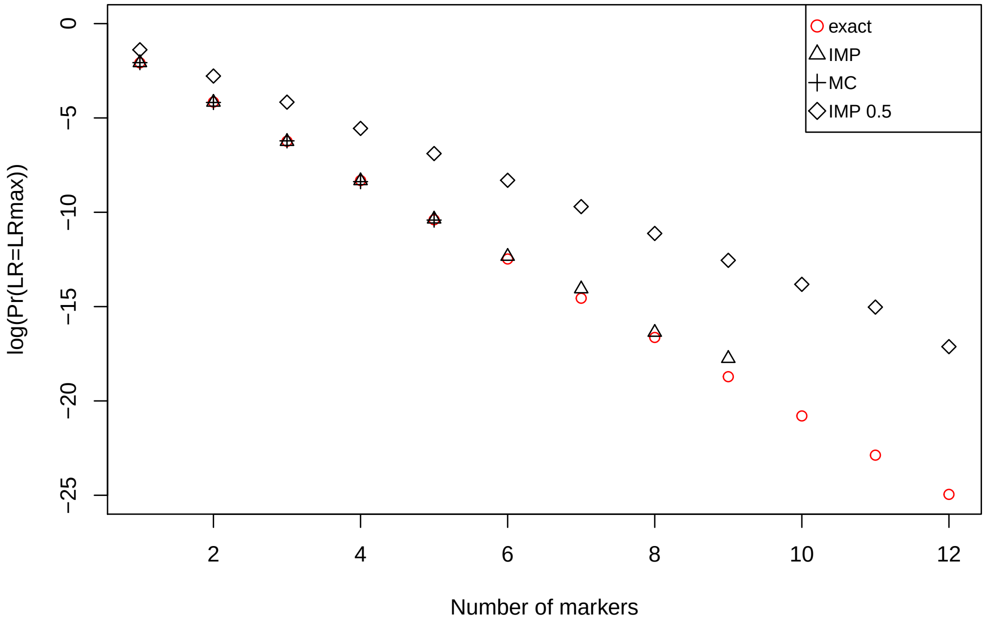


Figure 3.4: Plotted are the number of markers against the logarithm of the different estimates and the exact value. 'IMP' and 'MC' are the estimates from table 3.2 on page 46, while 'IMP 0.5' is the estimate resulting from using importance sampling with $b = 0.5$.

Example 3.4. Often times there are problems when forensic science meets the courtroom. For the layperson without a background in statistics or probability theory, the LR is easily misunderstood. It does not help that the LR, in contrast to probabilities, does not have a clear range, but depends upon the number of markers. To make the LR more interpretable, a standardized score as seen in table 3.3, was proposed in [11]. This was developed and implemented at NFC, Sweden (`http://nfc.polisen.se/en/English/`), and is now in use in there, meaning that the evidence is presented using the scale instead of the LR.

Table 3.3: The scale developed and implemented at NFC, Sweden. It is now in use in Sweden for reporting of evidence in court.

Score	LR	Explanation: The results are...
4	LR$> 10^6$	...at least a million times more expected ...
3	6000 <LR$\leq 10^6$	...at least 6000 times more expected ...
2	100 <LR$\leq$ 6000	...at least 100 times more expected ...
1	6 <LR$\leq$ 100	...at least 6 times more expected ...
0	1/6 <LR$\leq$ 6	...approximately as expected ...
		...if the main hypothesis is true compared to if the alternative hypothesis is true.
-1	1/100 <LR$\leq$ 1/6	...at least 6 times more expected ...
-2	1/6000 <LR$\leq$ 1/100	...at least 100 times more expected ...
-3	$1/10^6$ <LR$\leq$ 1/6000	...at least 6000 times more expected ...
-4	LR$\leq 1/10^6$	...at least a million times more expected ...
		...if the alternative hypothesis is true compared to if the main hypothesis is true.

We examined the probabilities of getting each score in the table when looking at different relationships and with different number of markers. Table 3.4 and 3.5 show the results for 10 and 15 markers for two pair of hypothesis; parent-offspring and full siblings, both against the persons being unrelated. We used the function sim.q from the package DNAprofiles to find the probabilities. Consider for instance the first line in table 3.4. This says that if AF is truly the biological father of the child, the chance of getting an LR bigger than 1 million is 0.021 when using 10 markers, and 0.34 when using 15 markers. A considerable difference, in other words. When AF is not the biological father, the probability of getting such a large LR is only

8.6e-09 and 9.9e-08 respectively for 10 and 15 markers. It is interesting to note how this is actually smaller for 10 markers than 15 markers for both pairs of relations. We have not studied this in any detail, or if the difference is significant for that matter. It could be an artifact of using more markers.

Table 3.4: The probability of getting each score on the Swedish scale when looking at a parent-offspring (PO) vs unrelated (UN) relationship with 10 and 15 markers. The allele frequencies used are Dutch and can be found through the R package 'DNAprofiles'.

Number of markers:	10		15	
True relationship:	PO	UN	PO	UN
Pr(LR> 10^6)	0.021	8.7e-09	0.34	9.9e-08
Pr(6000 <LR≤ 10^6)	0.41	2.3e-05	0.61	1.4e-05
Pr(100 <LR≤ 6000)	0.55	0.0008	0.049	2.8e-05
Pr(6 <LR≤ 100)	0.022	0.00047	9.9e-05	1.6e-06
Pr(1/6 <LR≤ 6)	4.4e-05	1.6e-06	1e-07	1.1e-06
Pr(1/100 <LR≤ 1/6)	0.00	0.00	0.00	0.00
Pr(1/6000 <LR≤ 1/100)	0.00	0.00	0.00	0.00
Pr(1/10^6 <LR≤ 1/6000)	0.00	0.00	0.00	0.00
Pr(0 ≤LR≤ 1/10^6)	0.00	0.99869	000	0.99

It can also be interesting to see the cumulative probabilities, so these are shown in tables 3.6 and 3.7. Here we see the probability of getting a certain score or higher. Consider the second line in table 3.6. Here we see that if the true relation is full siblings and 10 markers are used, the probability of getting an LR bigger than 6000 is 0.43. When using 15 markers the equivalent number is 0.95. If the persons are truly unrelated, the probability of getting an LR over 6000 is 2.3e-05 and 1.4e-05 respectively for 10 and 15 markers.

⋆

Table 3.5: The probability of getting each score on the Swedish scale when looking at a full sibling (FS) vs unrelated (UN) relationship with 10 and 15 markers. The allele frequencies used are Dutch and can be found through the R package 'DNAprofiles'.

Number of markers:	10		15	
True relationship:	FS	UN	FS	UN
Pr(LR> 10^6)	0.059	1.8e-08	0.27	5.3e-08
Pr(6000 <LR≤ 10^6)	0.28	1.2e-05	0.37	1.2e-05
Pr(100 <LR≤ 6000)	0.37	0.00088	0.24	0.00048
Pr(6 <LR≤ 100)	0.18	0.009	0.08	0.0037
Pr(1/6 <LR≤ 6)	0.089	0.094	0.033	0.035
Pr(1/100 <LR≤ 1/6)	0.012	0.26	0.0048	0.11
Pr(1/6000 <LR≤ 1/100)	0.0013	0.5	0.00075	0.38
Pr(1/10^6 <LR≤ 1/6000)	9.3e-06	0.14	1.6e-05	0.42
Pr(0 ≤LR≤ 1/10^6)	0.00	0.00055	1e-07	0.053

Table 3.6: The cumulative probabilities of getting a given score or higher when looking at a parent-offspring (PO) vs unrelated (UN) relationship using 10 and 15 markers.

Number of markers:	10		15	
True relationship:	PO	UN	PO	UN
Pr(LR> 10^6)	0.021	8.6e-09	0.34	9.9e-08
Pr(LR> 6000)	0.43	2.3e-05	0.95	1.4e-05
Pr(LR> 100)	0.98	0.00082	1.00	4.2e-05
Pr(LR> 6)	1.00	0.0013	1.00	4.3e-05
Pr(LR> 1/6)	1.00	0.0013	1.00	4.5e-05
Pr(LR> 1/100)	1.00	0.0013	1.00	4.5e-05
Pr(LR> 1/6000)	1.00	0.0013	1.00	4.5e-05
Pr(LR> 1/10^6)	1.00	0.0013	1.00	4.5e-05
Pr(LR≥ 0)	1.00	1.00	1.00	1.00

Table 3.7: The cumulative probabilities of getting a given score or higher when looking at a full siblings (FS) vs unrelated (UN) relationship using 10 and 15 markers.

Number of markers:	10		15	
True relationship:	FS	UN	FS	UN
Pr(LR$>10^6$)	0.059	1.8e-08	0.27	5.3e-08
Pr(LR$>$ 6000)	0.34	1.2e-05	0.64	1.3e-05
Pr(LR$>$ 100)	0.71	0.0009	0.88	0.0005
Pr(LR$>$ 6)	0.9	0.0099	0.96	0.0042
Pr(LR$>$ 1/6)	0.99	0.1	0.99	0.039
Pr(LR$>$ 1/100)	1.00	0.36	1.00	0.15
Pr(LR$>$ 1/6000)	1.00	0.86	1.00	0.53
Pr(LR$>1/10^6$)	1.00	1.00	1.00	0.95
Pr(LR$\geq$ 0)	1.00	1.00	1.00	1.00

Chapter 4

Discussion

In this thesis, our main application for the methods have been on the determination of kinship. We want to make as few errors as possible, and to know the probability of making these errors. This can be challenging because the probabilities can get so small that the normal Monte Carlo simulation method fails. For this reason we decided to explore if an alternative simulation method, importance sampling, could be a good option for this application.

There are, however, other applications where the methods discussed could be useful. An example can be taken from the area of construction, where a bridge or oil rig have a design load for wave heights of some size, say 25 meter. It is then important to quantify the probability of observing a wave higher than 25 m in the construction's future life time. When this is not done, the constructions tend to have a grossly exaggerated design load, with the economical implications for the builder that it has [10]. This is the case since a sequence of conservative, overly safe decisions tend to be made in a deterministic framework.

Yet another example is that of an insurance company facing the results of extreme weather or other disasters. In such circumstances they can get many claims in at the same time, and it is therefore important to know that their resources can cover such a scenario. More specific, if the amounts claimed are $S_1, ..., S_n$ and the resources that the company has available are r, they want to be sure that the combined claim $X = S_1 + ... + Sn$ can be met, given some conditions H. In other words, it is important to have an r such that $\Pr(X > r|H)$ is small.

4.1 Related research

Importance sampling is one of several methods for variance reduction in simulations, and is commonly used for problems that involves small probabilities. However, it had apparently not been used in the forensics field until [9], where the author was the first to use it in this setting. The paper can in a sense be viewed as a response to [2], where Dørum et al. presented an algorithm to find exceedance probabilities based on the exact distribution of the LR. The connection to p-values was also pointed out in the same paper. A search in `https://scholar.google.com` (May 12, 2016) shows that [9] has been cited 10 times, mostly from scientists associated with either the author or Dørum.

4.2 On the results and reporting of evidence

We found that the importance sampling method succeeds in finding smaller probabilities than the Monte Carlo method. This is shown for example in table 3.2 where the Monte Carlo method gives up after 5 markers with 10^5 simulations, while the importance sampling method lasts up to and including 9 markers. But we also see that there is not a big difference in the MSEs for the methods. When the probabilities get very small, 0 can be a very good approximation, close to the exact answer. An estimate of 0 is, however, difficult to use in further calculations, for example as the denominator of an LR. Sometimes, even an estimate which is farther from the true value is preferred to the 0 estimate.

We further made calculations based on the Swedish table, table 3.3. When using such tables it is important to know the probability of getting each score under the different hypotheses in question. We estimated these probabilities for a parent-offspring relation in table 3.4 and for full siblings in table 3.5, both for 10 and 15 markers. We see that since a parent-offspring relationship is closer than a sibling relationship, the probability is higher for getting the highest LRs.

For instance, we found that the probability that LR exceeds 100 in a duo case is 0.98 and 1.00 for 10 and 15 markers. This shows that 15 markers is preferable and indeed 15 markers is a lower bound on the number of markers typically used. The probability that LR exceeds 100 when the man is not the father should be close to 0. We found 0.00082 and 4.2e-05. While these numbers are small, it would obviously be an advantage to increase the number of markers to say, 23 markers, as is used in some labs. Obviously, including markers like SE33, that has 55 alleles, may help. The point is that we have to bear in mind that a large number of paternity cases are performed worldwide and therefore numbers close to 1 (when H_1 is true) and 0 (when H_2 is true) are called for. The number 100 above is rather arbitrary, but our calculations, and perhaps commons sense, indicate that it could be good to have a cutoff between 6000 and 10^6

It is worthwhile to note that there is an ongoing debate in the forensic science community on how the evidence is best presented in court. Some prefer a verbal scale like the Swedish one, reasoning that the LR is often difficult to explain in court. Others prefer to report the LR as it is, arguing that when a scale like the Swedish one is taken into use, judges and judiciary has to get accustomed to the new system, which can take even more time and effort. Furthermore, by introducing a verbal scale, the reporting officer may in a sense interfere with the job of the court and judges. One could argue that only numbers, based on well documented models and software, should be reported, and that statements like "very strong evidence" should be avoided. These and similar statements involve some amount of interpretation.

In some countries and labs it is more normal to report the *posterior probability* rather than the LR. The posterior probability has not been discussed in this thesis, but is explained for example in [3]. It has the advantage of being a probability with an easy interpretation and known range, but also the drawback of relying on some rather subjective input, namely the prior probabilities for the hypothesis. Normally a flat prior is used, i.e $\Pr(H_1) = \Pr(H_2) = {}^1/_2$, and then the posterior probability is $\Pr(H_1 \mid \text{data}) = {}^{LR}/_{(1+LR)}$. This approach is discussed further in [3].

Two of the R-functions made in connection with this thesis, 'Imp' and 'opt', warrant some extra attention. The former has an important extension compared to sim.q from 'DNAprofiles', namely the possibility to calculate the variance and MSE of the estimates. This makes it possible, for instance via the 'opt' function, to study the effects from choosing different biasing distributions.

In figure 3.3 we did this for 100 such distributions for an LR based on our standard marker from table 2.5 on page 24. From this it looks like using the probability distribution of LR under the numerator hypothesis, or a very similar one, is in fact the optimal choice. What is optimal depends on how important it is to not have a 0 estimate – a higher value of b makes the importance sampling method produce results for a higher number of markers, but then they tend to be overestimated. We see an example of this in figure 3.4; when we do importance sampling with $b = 0.25$, as is the probability under H_1, the estimate lies close to the exact value, while a $b = 0.5$ overestimates the probabilities. But then again, the former cannot produce results after 9 markers, while the latter gives estimates up to and including 18 markers (both numbers produced with the function's default seed of 16). The MSEs accompanying table 3.1 and 3.2 indicate that when using importance sampling with the optimal biasing distribution, this is preferred to the Monte Carlo simulation regardless of the above discussion. However, we have not checked if this difference in the MSEs is significant.

4.3 Future reaserch

It would be of interest to see how the results are for other, especially more distant, types of relationships that we have not investigated. An example is half siblings, either against a hypothesis of unrelatedness or full siblings. We would also like to know how transferable our results are to other related areas, like crime cases and mixtures. Specifically, tables like 3.6 could be useful for the common mixtures, i.e., those involving 2, or perhaps 3 persons. For this pupose, a carefully designed computer experiment, with several varying parameters, could be relevant. In addition, more simulations could be done for our applications, and parameters related to mutation, deviation from HWE, genotyping error and drop-out of alleles could be explored.

It would also be interesting to see an equivalent to table 3.4 where the mother has also been genotyped. We would then expect considerably improved results. The results of table 3.6 and similar tables could be reported as *ROC* curves. In this way the two types of errors can be studied in one figure.

There is also more to be done on the subject of the optimal biasing distribution when using importance sampling. It would be nice to figure out if there is some general rule to follow, or if we just have to find it trough trial and error in each situation. The figure 3.3 is made based on a parent-offspring relationship with iid markers, which could be expanded upon to apply to other relations and markers to see if the same relation holds true there. For this purpose it should not be too difficult to modify the function 'Imp' to accept markers with different distributions.

4.4 Final remarks

The main purpose of this thesis has been to explore how small probabilities can be estimated. The applications has been for kinship cases in forensic genetics, where there is a need to control the typically small probabilities for errors. The thesis builds directly on [9], and we have further extended on the results from it. Compared to the paper [9] and associated implementations in the R-package 'DNAprofiles', we have focused on the precision of the estimates, quantified as the variance and MSE, and how to use those measures of precision to choose a biasing distribution.

Examples as those summarized in table 3.6 are also important, and we are not aware of similar calculations. Our methods have been implemented in the R-package 'naomi'; see appendix A.1 for documentation.

Appendix A

Appendix

A.1 R-code

The essential code prepared for this thesis is included in the R package ‘naomi’, available from the author (naomin.azulay@gmail.com). Below we include the automatically generated pdf documentation.

Some changes has been done in the example belonging to the function ‘Imp’ concerning the making of table 3.1:

```
## The importance sampling part of table 3.1:
n=2
N=10^5
exact <- c(1-0.875^2, 0.75^2, 2*0.75*0.125, 0.125^2)
set.seed(89)

#LR=0
p0 <-Imp(t = 0, x, fx.p, fx.d, n, option=3, N, q.exact=1-exact[1])
p00 <- c(1-p0[1], p0[2:3])

#LR=1
p1 <- c(Imp(t = 0, x, fx.p, fx.d, n, option=3, N,
                            q.exact=sum(exact[2:4])),
        Imp(t = 1, x, fx.p, fx.d, n, option=3, N,
                            q.exact=sum(exact[3:4])))
p11 <- c(p1[1] - p1[4], p1[2] + p1[5])
p11 <- c(p11, p11[2] + (p11[1] - exact[2])^2)
```

```
#LR=2
p2 <- c(Imp(t = 1, x, fx.p, fx.d, n, option=3, N,
                            q.exact=sum(exact[3:4])),
        Imp(t = 2, x, fx.p, fx.d, n, option=3, N,
                            q.exact=exact[4]))
p22 <- c(p2[1] - p2[4], p2[2] + p2[5])
p22 <- c(p22, p22[2] + (p22[1] - exact[3])^2)

#LR=4
p4 <-Imp(t = 3.99, x, fx.p, fx.d, n, option=3, N,
                            q.exact=exact[4])

importance <- matrix(c(p00, p11, p22, p4), ncol=3, byrow=T)
mse.imp <- sum(importance[,3])
##The mc part is made by changing the argument ‘option’ from 3 to 4
```

Package 'naomi'

May 17, 2016

Type Package

Title Naomis master package

Version 1.0

Date 2016-04-29

Author Naomi Azulay

Maintainer Naomi Azulay <naomi.azulay@nmbu.no>

Description The code used in Naomi Azulays masters thesis: Estimation of small pribabilties with applications to forensic genetics.

License GPL(>=2)

Imports DNAprofiles

R topics documented:

---

naomi-package *Naomis master package*

Description

The R-code developed during the work with Naomi's master's thesis. The most important, data-producing code is presented. The code used in Naomi Azulays masters thesis: Estimation of small pribabilties with applications to forensic genetics.

Details

The DESCRIPTION file: This package was not yet installed at build time.

Index: This package was not yet installed at build time.

Author(s)

Naomi Azulay Maintainer: Naomi Azulay <naomi.azulay@nmbu.no>

Imp *MC or imp simulation of Pr(LR>t|H)*

Description

Estimates the exceedance probability Pr(LR>t|H) with either Monte Carlo simulation or importance sampling for n iid markers. It is also possible to get the estimate's variance and the MSE.

Usage

```
Imp(t, x, fx.p, fx.d, n, option = 3, N = 10^5, q.exact=FALSE)
```

Arguments

`t`	Double. Threshold.
`x`	Vector of double. Possible values for the LR for one marker.
`fx.p`	Vector of double. Probabilities for the possible LR values under Hp (H1).
`fx.d`	Vector of double. Probabilities for the possible LR values under Hd (H2).
`n`	Integer. Number of iid markers.
`option`	Integer. 1 for MC with sim.q from DNAprofiles, 2 for imp with sim.q, 3 and 4 is a 'manual' simulation with imp and MC respectively, the variance of the estimate is included. For option 3 and 4 MSE is outputted if q.exact is specified.
`N`	Integer. Number of simulations. Default is 10^5.
`q.exact`	Double. The exact probability. Used for calculation of MSE.

Value

The estimate of Pr(LR >t |H), and potentially the variance and MSE. For option 1 and 2 it is a double (only the estimate), for option 3 and 4 this is a vector of doubles of length 2 or 3, depending on whether q.exact is specified. The order is estimate, variance, MSE.

Author(s)

Naomi Azulay

Examples

```
##Example of estimating with MC and getting variance and MSE:
t <- 1.99
x <-  c(0, 1, 2)
fx.p <- c(0, 0.75, 0.25)
fx.d <- c(0.125, 0.75, 0.125)
res <- Imp(t, x, fx.p, fx.d, n = 1, option = 4, q.exact=0.125)

## The code used in example 3.1:
set.seed(123)
res <- Imp(t, x, fx.p, fx.d, n = 1, option = 3, q.exact=0.125)
```

```
## The importance sampling part of table 3.1:
n=2
N=10^5
exact <- c(1-0.875^2, 0.75^2, 2*0.75*0.125, 0.125^2)
set.seed(89)

#LR=0
p0 <-Imp(t = 0, x, fx.p, fx.d, n, option=3, N, q.exact=exact[1])
p00 <- c(1-p0[1], p0[2:3])

#LR=1
p1 <- c(Imp(t = 0, x, fx.p, fx.d, n, option=3, N, q.exact=exact[1]),
        Imp(t = 1, x, fx.p, fx.d, n, option=3, N, q.exact=exact[2]))
p11 <- c(p1[1] - p1[4], p1[2] + p1[5], p1[6])

#LR=2
p2 <- c(Imp(t = 1, x, fx.p, fx.d, n, option=3, N, q.exact=exact[2]),
        Imp(t = 2, x, fx.p, fx.d, n, option=3, N, q.exact=exact[3]))
p22 <- c(p2[1] - p2[4], p2[2] + p2[5], p2[6])

#LR=4
p4 <-Imp(t = 3.99, x, fx.p, fx.d, n, option=3, N, q.exact=exact[4])
```

moments *E(LR) and Var(LR) for one marker.*

Description

Calculates the expectation and variance of LR, 1/LR and log(LR) for one marker with the formulas for discrete distributions.

Usage

```
moments(h)
```

Arguments

h	List of two vectors. x is the possible LRs, fx is the probability of getting each LR.

Value

Returns a data.frame with the expectation and variance of LR, 1/LR and log(LR).

Author(s)

Thore Egeland and Naomi Azulay

Examples

```
## data for figure 2.8:
library(DNAprofiles)
n <- 10

#asymptotic:
hp <- list(x = c(1,2), fx = c(0.75, 0.25))
m <- moments(hp)
mu <- m[1,3]
s<- sqrt(m[2,3])
t <- c(2^(1:n))
log.p <- 1-pnorm((log(t)-n*mu)/(sqrt(n)*s))

#exact values:
x.p <- c(1, 2)
fx.p <- c(0.75, 0.25)

dist <- vector("list", n)
for( i in 1:n)
  dist[[i]] <- list(x = x.p, fx = fx.p)
pair <-  dists.product.pair(dist)
cdfD <-  dist.pair.cdf(pair)

exact = 1-cdfD(t)


## The function is currently defined as:
function (x)
moments=function(h){
  LR=h$x
  p=h$fx
  E=sum(LR*p)
  E2=sum(LR^2*p)
  Var=E2-E^2
  EInverse=sum(LR^(-1)*p)
  EInverse2=sum(LR^(-2)*p)
  VarInverse=EInverse2-EInverse^2
  Elog=sum(log(LR)*p)
  varlog=sum((log(LR))^2*p)-Elog^2
  res=data.frame(LR=c(E,Var),LRInverse=c(EInverse,VarInverse),
                 LRlog=c(Elog,varlog))
  rownames(res)=c("E","Var")
  return(res)
}
```

`opt` *Vary the biasing distribution*

Description

Returns a list with two elemts - a table of estimated Pr(LR = LRmax) for 1 up to nmax iid markers, and the MSE summed over all rows. It is made specifically for the marker from table 2.5 in the thesis, with the possibility for varying the sampling distribution with the argument b.

Usage

```
opt(b=0.25, N=10^5, nMax=10, seed=16, option=3)
```

Arguments

`b` Double. Argument to vary the biasing distribution, on the form (0, 1-b, b). The default is Pr(LR = x | H1) for x = c(0, 1, 2), namely (0, 0.75, 0.25).

`N` Integer. Number of simulations, passed on to Imp. Default is 10^5.

`nMax` Integer. The table is made for 1 up to nmax number of markers.

`seed` Integer. The seed for the simulation. Default is 16.

`option` Integer. The option passed to Imp, the default is 3. Possible values is in the range from 1 to 4, but only 3 or 4 should be used here.

Value

Returns a list with two elements. tab is the table with the estimates, variance and MSEs. sumMSE is the sum of the column with MSEs.

Author(s)

Thore Egeland and Naomi Azulay

References

Naomi's master thesis. Used to make table 3.2, figure 3.3 and 3.4.

Examples

```
## Table 3.2 is based on:
imp.tab <- opt(nMax=20)
mc.tab <- opt(nMax=20, option=4)

## For figure 3.3 we added:
mod <- opt(b=0.5, nMax=12)$tab

## Figure 3.4:
y <- NULL
# we used this bb:
#bb <- seq(0.01, 1, 0.01)
#but for convenience we use this in the example:
bb <- seq(0.1, 1, 0.2)
res <- vector("list", length(bb))

for (i in 1:length(bb)){
  res[[i]] <- opt(bb[i], nMax=20)
  y <- c(y, res[[i]]$sumMSE)
}


## The function is currently defined as
opt <- function(b=0.25, N=10^5, nMax=10, seed=16, option=3){
  x <- c(0, 1, 2)
  a <- 1-b
  fx.p <- c(0, a, b)
  fx.d <- c(0.125, 0.75, 0.125)
  set.seed(seed)
  tab <- NULL
  for (n in 1:nMax){
    t <- 2^n-0.01
    q.exact <- 0.125^n
    est <- Imp(t, x, fx.p, fx.d, n = n, option=option, N=N, q.exact=q.exact)
    tab <- rbind(tab, c(n, q.exact, est))
  }
  colnames(tab) <- c("n", "Exact","SIM","varSIM","MSE.SIM")
  return(list(sumMSE=sum(tab[,5]),tab=tab))
}
```

swedishTable *Calculations on the Swedish table*

Description

Calculates the probability for each score in the Swedish table, or the cumulative probability of getting a score or higher, for hypothesis about relation r1 and r2.

Usage

```
swedishTable(r1, r2, n=10, Nsim=10^5,cumul=FALSE,  seed=17)
```

Arguments

r1	The relation under H1, e.g "FS" - full sibling or "UN" - unrelated. See ibdprobs(x) for full list of possible relations.
r2	The relation under H2, e.g "FS" - full sibling or "UN" - unrelated. See ibdprobs(x) for full list of possible relations.
n	Integer. Number of estimates to make. Default is 10.
Nsim	Integer. Number of simulations. Default is 10^5.
cumul	Logic
seed	Integer. The seed for the simulation. Default is 17.

Value

Returns a matrix with four columns - for both 10 and 15 markers we estimate the probability of getting the scores given each of the hypothesis.

Author(s)

Naomi Azulay

Examples

```
## Tables 3.4-3.7:
#We used n=100, but for convenience we use 10 here
library(DNAprofiles)
tab <- swedishTable(r1="FS", r2="UN")
tab <- swedishTable(r1="FS", r2="UN", cumul=TRUE)
tab <- swedishTable("PO", "UN", seed=78)
tab <- swedishTable("PO", "UN", seed=78, cumul=TRUE)
```

tableU *Estimates Pr(U>b | mu = 0) for U ~ N with imp.*

Description

Estimates Pr(U>b | mu = 0) for U ~ N(0, sd) with imp using a biasing distribution with mu = a. Returns list of 5 elements - a table with each step of making the imp estimate, the estimate, the exact value, the MSE and the relative error.

Usage

```
tableU(N=10, a=5, sd=1, b=5)
```

Arguments

N	Integer. Number of simulations. Default is 10^5.
a	Double. The mean for the biasing distribution.
sd	Double. The standard deviation for both distributions.
b	Double. Treshold for Pr(U>b).

Value

List of 5: tab - a data.frame with the imp method step-by-step, theta.hat - the estimate, theta.exact - the exact value, mse - the MSE, rel.err - the relative error.

Author(s)

Naomi Azulay

Examples

```
## Table 2.6:
set.seed(67)
imp.table <- tableU()$tab


## Tables 2.7 and 2.8:
N=10^5
set.seed(67)

MSE <- NULL
relERR <- NULL
for (u in c(0,3,5,7,9)) {
  for (mu in c(0,3,5,7,9)) {
    res <- tableU(N, a=mu, sd=1, b=u)
    MSE <- c(MSE, res$mse)
    relERR <- c(relERR, res$rel.err)
  }
}


## The function is currently defined as:
```

```
tableU <- function(N=10^5, a=5, sd=1, b=5) {
  simU <- rnorm(N, mean=a, sd=sd)
  Icol <- as.integer(simU>b)
  w <- dnorm(simU, 0, sd)/dnorm(simU, a, sd)
  z <- Icol*w
  tab <- data.frame(simU, Icol,  w, z)
  theta.hat <- mean(z)
  theta <- 1-pnorm(b, sd = sd)
  var.theta <- (mean(Icol*w^2)-theta.hat^2)/N
  mse <- var.theta + mean((theta.hat-theta)^2)
  rel.err <- (theta.hat-theta)^2/theta
  list(tab = tab, theta.hat = theta.hat, theta.exact = theta, mse=mse, rel.err=rel.err)
}
```

```
{ ~kwd1 }
{ ~kwd2 }
```

Index

A.2 Table for duo case

Table A.1: Genotype data for a child and alleged father (AF) along with LRs. The allele frequencies used can be found at `http://familias.name/Table2.5.fam`.

System	Child	AF	LR
D3S1358	17/17	17/18	2.450
TPOX	8/8	8/8	1.805
TH01	6/9	6/7	1.195
D21S11	29/30	28/29	1.096
D18S51	14/16	16/17	2.153
PENTA_E	7/11	11/16	2.408
D5S818	12/12	12/13	1.406
D13S317	8/8	8/11	4.042
D7S820	9/10	9/13	1.434
D16S539	13/14	11/14	8.312
CSF1PO	10/10	10/11	2.025
PENTA_D	8/11	8/13	11.989
VWA	19/19	17/19	5.565
D8S1179	13/16	11/16	9.651
FGA	21/22	21/21	2.956
D12S391	19/22	19/23	2.184
D1S1656	14/16	14/15	3.333
D2S1338	18/20	18/23	3.147
D22S1045	12/12	12/15	26.748
D2S441	10/13	10/15	1.446
D19S433	12/15	12/14	3.344
Total			0

Bibliography

[1] J. Buckleton, C. Triggs, and S. Walsh, editors. *Forensic DNA Evidence Interpretation.* CRC Press, Florida, USA, 2005.

[2] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, and T. Egeland. Exact computation of the distribution of likelihood ratios with forensic applications. *Forensic Science International: Genetics*, 9:93–101, 2014.

[3] T. Egeland, D. Kling, and P. Mostad. *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics.* Elsevier, 2016.

[4] T. Egeland, B. Kulle, and R. Andreassen. Essen-möller and identification based on dna. *Chance*, 19(2):27–31, 2006.

[5] I. Evett and B. Weir. *Interpreting DNA Evidence.* Sinauer, Sunderland MA, 1998.

[6] P. Gill, L. Fereday, N. Morling, and P. M. Schneider. The evolution of DNA databases—recommendations for new european STR loci. *Forensic Science International*, 156(2):242–244, 2006.

[7] J. Graffelman. Exploring diallelic genetic markers: The hardy weinberg package. *Journal of statistical software*, 64(3):1–23, 2015.

[8] D. Kling. *Computational challenges in family genetics.* PhD thesis, Norwegian University of Life Sciences, 2015.

[9] M. Kruijver. Efficient computations with the likelihood ratio distribution. *Forensic Science International: Genetics*, 14:116–124, 2015.

[10] H. O. Madsen and T. Egeland. Structural reliability: Models and applications. *International Statistical Review/Revue Internationale de Statistique*, pages 185–203, 1989.

[11] A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger. Scale of conclusions for the value of evidence. *Law, probability and risk*, 11:1–24, 2012.

[12] B. Olaisen, M. Stenersen, and B. Mevåg. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics*, 15:402–405, 1997.

[13] B. D. Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

[14] C. Sammut and G. I. Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[15] T. Strachan and A. Read. *Human molecular genetics*. Garland Science, 2010.

[16] A. O. Tillmar and D. Kling. Comments on "kinship analysis: assessment of related vs unrelated based on defined pedigrees" by s. turrina et al. *International Journal of Legal Medicine*, pages 1–3, 2016.

Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway