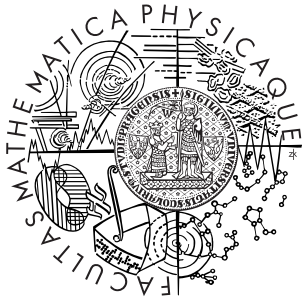
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Helena Kubátová

Statistické metody pro interpretaci forenzních DNA směsí

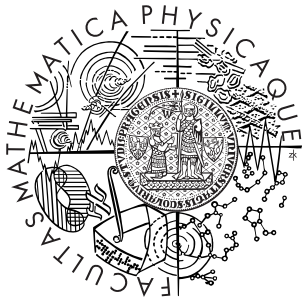
Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	Prof. RNDr. Jana Zvárová, DrSc.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán:	Teorie pravděpodobnosti a náhodné procesy

2010

Charles University in Prague
Faculty of Mathematics and Physics

DIPLOMA THESIS



Helena Kubátová

Statistical methods for interpreting forensic DNA mixtures

Department of Probability and Mathematical Statistics

Supervisor:	Prof. RNDr. Jana Zvárová, DrSc.
Study programme:	Mathematics
Study branch:	Probability, Mathematical Statistics and Econometrics
Study plan:	Probability Theory and Random Processes

2010

I would like to thank my supervisor, Prof. RNDr. Jana Zvárová, DrSc., for the patient guidance, support and advice she has provided me during this work.

I declare that I have written all of the thesis on my own and that I have cited all used sources of information. I agree with public availability and lending of the thesis.

Prague, 6 August 2010 Helena Kubátová

Contents

Název práce: Statistické metody pro interpretaci forenzních DNA směsí
Autor: Helena Kubátová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Jana Zvárová, DrSc.
e-mail vedoucího: zvarova@euromise.cz

Abstrakt: V předložené práci studujeme interpretaci forenzních DNA směsí z pohledu teorie pravděpodobnosti a matematické statistiky. Provádíme detailní matematickou formalizaci dané problematiky, což nám umožňuje formulovat matematicky přesné a konzistentní výsledky. Zabýváme se vyhodnocováním dané evidence prostřednictvím věrohodnostního poměru, který porovnává dvě hypotézy dané genotypy známých přispěvatelů a počtem neznámých přispěvatelů do směsi DNA. Dále rozebíráme případ strukturované populace, přičemž provádíme důkladnou revizi vzorců pro homozygotní a heterozygotní genotypy, jejich zobecněné varianty a dále věty pro výpočet pravděpodobnosti, že předpokládaný počet náhodně vybraných jedinců vysvětluje danou genetickou evidenci.

Klíčová slova: DNA (genetická) evidence, věrohodnostní poměr, struktura populace, coancestry koeficient

Title: Statistical methods for interpreting forensic DNA mixtures
Author: Helena Kubátová
Department: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Jana Zvárová, DrSc.
Supervisor's e-mail address: zvarova@euromise.cz

Abstract: In the present work we study interpretation of forensic DNA evidence from the point of view of probability theory and mathematical statistics. We provide a detailed mathematical formalization of the problem, which enables us to formulate mathematically accurate and consistent results. We deal with evaluation of a given evidence in terms of a likelihood ratio, which compares two hypotheses specified by genotypes of known contributors and the number of unknown contributors to the DNA mixture. We further analyze the case of a subdivided population, performing a thorough revision of the formulas for homozygous and heterozygous genotypes, their general version and a theorem for calculating the probability that an assumed number of random individuals explain a given genetic evidence.

Keywords: DNA (genetic) evidence, likelihood ratio, population substructure, coancestry coefficient

Chapter 1

Introduction

Various mathematical and especially statistical models are used in almost all spheres of human activity. They become applied in numerous branches, which originally had nothing in common with mathematics - for example, biology, medicine or forensic science.

Evaluation of forensic DNA evidence is one of the cases where probabilistic and statistical models come into play, together with genetics, forensic science and also the law and justice.

As in many other interdisciplinary branches, we soon come across the issue that one person is usually a great professional in one of the disciplines, while having only an overview of the other ones. Particularly, mathematicians typically do not have sufficient genetic education, and vice versa. As a result, many sources written by geneticists or also statistical geneticists are mathematically inaccurate, incomplete and often even incorrect.

This thesis is definitely more mathematical than most of the material published on this topic. It deals with a common issue of DNA mixtures, both under the assumption of Hardy-Weinberg equilibrium and in subdivided populations, performing a thorough mathematical revision of the published results and giving suggestions to several adjustments.

After providing a brief introduction to genetics, which explains the necessary terms, and a general description of the situation we deal with in Chapter 2, we start in Chapter 3 by a proper mathematical formalization of the independence case, followed by a theorem for the probability that x random individuals together explain the unexplained part of a given genetic evidence. In Chapter 4 we extend the considerations to the case of a subdivided population, where exact allele proportions are unknown and

we therefore need to represent the population substructure more generally. Finally, we provide an extension of the theorem from Chapter 3 to the subdivided population case, including its proof, computer implementation and examples of use. A small English – Czech dictionary of key genetic and forensic terms is available in Appendix 1.

Chapter 2

Basic terms and facts

2.1 Genetic basics

Genetics is a scientific discipline dealing with heredity and variation in living organisms. By **heredity** we mean the process of passing information from ancestors to offspring during reproduction.

Genetic information is stored in the nucleus of each cell on **chromosomes**. These are actually thin threads of **DNA (deoxyribonucleic acid)**, surrounded by protein and some other material. Every human cell contains 23 chromosome pairs, except for sex cells (gametes), which only contain one set of 23 chromosomes.

A **gene** is the basic unit of the inherited genetic information. It is a segment of the DNA thread of various length. The position of a particular gene on the chromosome is called a **locus**. Every gene has several alternative forms called **alleles**.

As we said, genes are stored on chromosome pairs, thus every individual carries two alleles of each gene. A complete set of an individual's genes is referred to as his or her **genotype**, but very often the meaning of this term is shifted to denoting only one particular allele pair. If both alleles in a pair are of the same type, we call such individual a **homozygote** (or we say he or she is carrying a homozygous genotype), otherwise we call him or her a **heterozygote** (heterozygous genotype).

During sexual reproduction, each parent passes one randomly selected allele from each allele pair to the offspring, so that every individual receives a complete genotype, which is however different from both parents' genotypes. This way, genetic variation is ensured in the population. Since both alleles

from a parent's genotype have an equal chance to be passed to the genotype of the offspring, simple combinatorial principles apply to the passing of alleles to following generations.

A **population** can be defined as a group of individuals living in the same geographical area, so that sexual reproduction is possible between any pair of individuals within this area and is more probable than sexual reproduction between a pair of an individual from this area and an individual from another area. A population may be divided into **subpopulations**, determined usually geographically or racially. A population which does not contain any subpopulations is usually referred to as **homogeneous**.

If a homogeneous population fulfills several criteria like being enough large, with no migration, no mutations and no selection, we say that **Hardy-Weinberg equilibrium** has been established. It is a simplifying assumption, allowing us to consider the two alleles at a particular locus mutually independent. Knowing the allele proportions in such population, we can then calculate probabilities of genotypes very easily.

2.2 DNA forensics

DNA profiling, which was introduced by Jeffreys *et al.* in 1985 [5], is a very powerful method for human identification, since there are no two people in the world with exactly the same genotypes (with the exception of identical twins).

However, practically a unique identification cannot be assured, since the DNA profiling techniques use only some genetic markers and thus do not provide a complete description of an individual's genotype. Furthermore, sampling errors may occur, which must be taken into account. Therefore DNA fingerprinting can never provide a 100 % certainty about the identity of the individual from whom the sample comes. Nevertheless, it can still give us a lot of information: if a given genetic sample and a particular individual's genotype share very rare alleles or share a large amount of alleles, then it is definitely highly likely that the sample comes from that particular individual.

Suppose that a crime has been committed and a DNA evidence has been collected from the crime scene. It can be sometimes very easy to obtain DNA samples in such case, since DNA can be found in every cell of the human body, including blood, hair, skin, bones, semen, sweat, saliva etc. The problem, however, is, that this way we often collect mixed material from the victim and the perpetrator(s) and possible other contributors, having then

no way to determine the particular genetic profiles of the contributors, or even their number. By sampling we only obtain the genetic profile of the evidence, i.e. a list of distinct alleles in it, but neither we can assess the frequencies with which each of them occurred there, nor their original configurations in the genotypes. Assessing the weight of such evidence against an identified suspect (or suspects) can become a complex problem.

Moreover, to make our models reflect reality properly, we often need to take into account the population substructure, which means that we cannot consider alleles in genotypes and genetic profiles independent. The calculations then become more complicated, as they try to incorporate the uncertainty we have about the allele proportions in a subdivided population. Ignoring this uncertainty would sometimes lead to overstating of the evidence weight, which is naturally unfavourable to innocent suspect. It is therefore of crucial importance to establish a consistent and conservative model for dealing with such situations.

Chapter 3

Independence case

In the whole work we study the situation at a single locus, supposing that genetic profiles at different loci are mutually independent. We only consider alleles of a discrete type, denoting them usually with capital letters from the beginning of the alphabet.

3.1 Formal description

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a finite probability space with $\Omega = \{A_1, A_2, ..., A_n\}$ representing all alleles that might occur at the observed locus; $\mathcal{F}$ being the sigma algebra of all subsets of Ω and P being a probability measure on Ω given by allele frequencies estimated from a relevant database, i.e. $\mathsf{P}\{A_i\} = p_{A_i}$ for $i = 1, 2, \ldots, n$, where p_{A_i} is the estimated frequency of allele A_i in the population.[1]

Note 3.1. To be able to define precisely the probability of genotypes - i.e. pairs of alleles (but neither ordered pairs nor two-element sets) - we would need to introduce another probability space $(\Omega_g, \mathcal{F}_g, \mathsf{P}_g)$ with $\Omega_g = \{(A_i, A_j) : A_i, A_j \in \Omega; i \leq j\}$, $\mathcal{F}_g$ being again the sigma algebra of all its subsets and P_g being a probability measure on $(\Omega_g, \mathcal{F}_g)$. However, for the needs of this paper we will identify the notation of these two probability spaces as $(\Omega, \mathcal{F}, \mathsf{P})$. (The situation is similar to throwing two dices and examining the probability of possible results - usually the same notation is used as for one dice.)

[1] Since the frequency f_A of the allele A in a sample of N alleles from the population has a binomial distribution $\mathsf{Bi}(N, p_A)$, the relative frequency $\frac{f_A}{N}$ is a maximum likelihood estimation of p_A.

Under the assumption of Hardy-Weinberg equilibrium, i.e. when all alleles are considered to be statistically independent, the probability of a homozygous genotype occurring at the observed locus can be evaluated as

$$\mathsf{P}(A_iA_i) = p_{A_i}^2 \tag{3.1}$$

and the probability of a heterozygous genotype is

$$\mathsf{P}(A_iA_j) = 2p_{A_i}p_{A_j}, i \neq j. \tag{3.2}$$

The factor 2 can be explained by two possible ways of inheritance of such genotype - A_i from the mother and A_j from the father, or vice versa.

Note 3.2. Another approach to evaluating probabilities of genotypes could be considering them (theoretically) as ordered pairs of alleles, e.g. always *('allele from mother','allele from father')* (so that Ω_g from Note 3.1 would be $\{(A_i, A_j) : A_i, A_j \in \Omega\}$), and then identifying genotypes containing the same alleles, i.e. $(A_i, A_j) = (A_j, A_i)$. This would also satisfactorily explain the factor 2 in the probability of a heterozygous genotype.

From this point of view, genetic *evidence* is a finite (usually small) set of alleles; we will usually denote it as E. Supposing we know the number of contributors to the evidence, $|E| \leq 2k$, where k is the number of contributors ($k = 1, 2, \dots$). An issue we often need to deal with is the fact that there is no way to determine the number of contributors and their exact genotypes, knowing only the evidential set of alleles. Therefore, when determining the probability of a given evidence having arisen (under certain circumstances), we have to take into account all possible ways of its occurrence. More precisely, we need to evaluate the probability of the union of all random events from Ω, which would under these circumstances lead to finding the given evidence.

For the purpose of this work, we will suppose that we know the number of contributors to the given evidence.

Example 3.3. Let $E = \{A, B, C\}$ (we will usually denote this as $E = \{ABC\}$ or even $E = ABC$ for short) and suppose we know that it comes from two contributors. It means that the two contributors must carry alleles A, B and C and no other alleles in their genotypes. The probability of this evidence

having arisen can then be evaluated as follows:[2]

$$\begin{aligned}\mathsf{P}(E) &= 2\mathsf{P}(AA,BC) + 2\mathsf{P}(AB,AC) + 2\mathsf{P}(AB,BC) + 2\mathsf{P}(AB,CC) \\ &\quad + 2\mathsf{P}(AC,BB) + 2\mathsf{P}(AC,BC) \\ &= 2 \cdot (p_A^2 \cdot 2p_Bp_C + 2p_Ap_B \cdot 2p_Ap_C + 2p_Ap_B \cdot 2p_Bp_C \\ &\quad + 2p_Ap_B \cdot p_C^2 + 2p_Ap_C \cdot p_B^2 + 2p_Ap_C \cdot 2p_Bp_C) \\ &= 2 \cdot 2p_Ap_Bp_C \cdot (3p_A + 3p_B + 3p_C) \\ &= 12p_Ap_Bp_C(p_A + p_B + p_C)\end{aligned}$$

△

3.2 Likelihood ratio

To be able to evaluate the chance that a certain suspect is the perpetrator of a given crime (in other words, it is the chance that a certain person contributes to a given DNA evidence), we usually need to determine the probability of that evidence having arisen under different hypotheses - typically "the suspect is the perpetrator" or "the suspect is innocent". Formally, it means that we need to evaluate the conditional probability of E under a hypothesis H. To enumerate the chance of the suspect's guilt we then use so called *likelihood ratio*, defined as

$$LR = \frac{\mathsf{P}(E|H_1)}{\mathsf{P}(E|H_2)}. \tag{3.3}$$

It tells us, that the evidence E is LR-times more likely to have arisen under the hypothesis H_1 (often the "guilt" hypothesis) then under H_2 (which is often the hypothesis of the suspect's innocence).[3]

According to [4], "probabilities in the magnitude of one in millions or one in billions are commonly heard in court cases." However, we have to realize that these values are obtained from the investigation of large numbers of loci, whereas this work only deals with calculations on one locus.

[2]To be precisely formal, we would again need to introduce a new probability space - a space of pairs of genotypes. To stay simple, we will keep the previously introduced notation, with $\mathsf{P}(AB,CD)$ denoting the probability that two randomly selected individuals will carry the genotypes AB and CD, regardless of order.

[3]We adopted the usual terminology of "hypotheses" as the conditions, which must not be confused with statistical hypothesis testing, as it is a completely different branch.

Example 3.4. Let us illustrate the usage of likelihood ratio on a trivial example. Let $E = AB$, coming from one perpetrator only (such evidence might be obtained for example from a blood stain found on a broken window, used by the perpetrator to access or leave the crime scene). That is, the perpetrator's profile $W = E = AB$. Suppose we have one suspect, whose profile matches the evidence, i.e. $S = AB$. This match does not necessarily have to mean that the suspect is the perpetrator of the crime, but it definitely makes this alternative more likely.

Precisely: If $H_1 =$ *the suspect is the perpetrator*, then

$$\mathsf{P}(E|H_1) = \mathsf{P}(W = E|W = S) = \mathsf{P}(E = S) = 1;$$

if $H_2 =$ *the perpetrator is an unknown person with a (random) genetic profile* R, then

$$\mathsf{P}(E|H_2) = \mathsf{P}(W = E|W = R) = \mathsf{P}(E = R) = \mathsf{P}(R = AB).$$

The likelihood ratio is then

$$LR = \frac{1}{\mathsf{P}(AB)} = \frac{1}{2p_A p_B},$$

which in case of allele frequencies $p_A = 0.1$ and $p_B = 0.2$ would mean that the evidence is 25 times more likely to have arisen under the hypothesis H_1 (= *the suspect is the perpetrator*) than under H_2. △

Very often we deal with genetic evidence in the form of a mixed DNA sample, knowing that it comes from the perpetrator(s) and the victim. (Of course, other situations may also occur - for example, in rape cases the evidential sample may also contain alleles of the victim's current partner.) E is then the set of all distinct alleles from the victim's profile and from the perpetrators' profiles. Let us take the usual situation when $E = V \cup W$, where V denotes the victim's genetic profile and W the perpetrator's profile (supposing that the crime was committed by a single perpetrator).[4] We have one suspect, whose genetic profile will be denoted by S, and want to determine the chance that this suspect is the perpetrator of the crime. Suppose that $\{S\} \subset E$ and $\{V\} \cup \{S\} = E$, so that the suspect cannot be excluded from the range of possible perpetrators directly by his/her genotype.

[4]We need to be careful about the union operation here, as V and W are not sets in general - it should therefore be taken as $E = \{V\} \cup \{W\}$, where $\{V\}$ and $\{W\}$ are the sets of distinct alleles from V and W, respectively.

Under the hypothesis of guilt, i.e. H_1*: the contributors to the evidence are the victim and the suspect*, we have

$$\mathsf{P}(V \cup W = E | W = S) = \mathsf{P}(V \cup S = E) = 1,$$

which corresponds to the intuitive result that the given evidence would surely have arisen if it was contributed by the victim and the suspect.

Under the hypothesis of innocence, i.e. H_2*: the contributors to the evidence are the victim and an unknown person with a genetic profile* R, we have

$$\mathsf{P}(V \cup W = E | W = R) = \mathsf{P}(V \cup R = E) = \mathsf{P}(E \setminus V \subset R \subset E),$$

which can take any value from the interval $(0, 1]$, depending on the relation between $\{V\}$ and $\{E\}$ and on the particular allele frequencies.

Evaluation of expressions of the type $\mathsf{P}(U \subset R \subset E)$ for given allele sets $E, U \subset E$ and a random allele set R will be discussed further on.

3.3 Evaluating $\mathsf{P}_x(U, E, K)$

When calculating likelihood ratios in cases with a larger number of contributors to the given evidence, we often need to evaluate the probability that x random individuals will carry all alleles from the evidence E which are not contained in the genotypes of known contributors and at the same time they will not carry any other alleles except for those contained in E.
First, let us briefly summarize the used notation:

- E the evidential set of alleles,
- K the set of all distinct alleles contained in the genotypes of known contributors,
- x the number of unknown contributors to the evidence (suppose it is known from other circumstances of the crime),
- R the set of all distinct alleles contained in the genotypes of the unknown contributors[5],

[5]Which actually means the set of distinct alleles obtained by $2x$ random allele draws from the population.

U the unexplained part of the evidence E, i.e. $U = E \setminus K$ in this case.

For the given evidence E to arise it is necessary that $E \setminus K \subset R \subset E$, i.e. $U \subset R \subset E$.

Supposing that there are exactly x unknown contributors, let $\mathsf{P}_x(U, E, K)$ denote the probability that their profiles will together "fit" the evidence (in the sense explained above).

Zoubková in [9] gives a formula (with proof; both based on [7]) for the evaluation of this probability for a fixed x:

Theorem 3.5. *For $x \in \mathbb{N}$ and $E = \{1, \ldots, e\}$ [6], $K \subset E$ and $U \subset E$; $U = E \setminus K$; being allele sets as declared above, it is*

$$\begin{aligned}\mathsf{P}_x(U, E, K) = (T_0)^{2x} - \sum_{i \in U} (T_{1i})^{2x} + \sum_{i,j \in U; i<j} (T_{2ij})^{2x} \\ - \sum_{i,j,k \in U; i<j<k} (T_{3ijk})^{2x} + \cdots + (-1)^u (T_{uU})^{2x},\end{aligned} \tag{3.4}$$

where $u = |U|$ and

$$T_0 = \sum_{l \in E} p_l,$$

$$T_{1i} = \sum_{l \in E \setminus \{i\}} p_l, \quad i \in U,$$

$$T_{2ij} = \sum_{l \in E \setminus \{i,j\}} p_l, \quad i, j \in U,$$

$$T_{3ijk} = \sum_{l \in E \setminus \{i,j,k\}} p_l, \quad i, j, k \in U,$$

$$\vdots$$

$$T_{uU} = \sum_{l \in K} p_l.$$

[6] The notation of allele types is shifted here from $A_1, A_2, \ldots, A_e$ to $1, 2, \ldots, e$ for brevity (e is the number of distinct alleles in E, i.e. $|E| = e$).

Although the formula seems rather complicated, it can be implemented easily using any common programming language. Zoubková in Appendix 2 of [9] gives a complete implementation in R.

Let us now illustrate the usage of 3.4 on a few examples.

Example 3.6. Suppose that the evidence profile is $E = ABC$, the victim's profile has been identified as $V = AB$ and we have one suspect with a genetic profile $S = CC$.

Assuming that there was only one perpetrator of the crime, we may set the usual hypotheses, H_1*: the contributors to the evidence are the victim and the suspect* and H_2*: the contributors to the evidence are the victim and an unknown person*. We immediately see that $\mathsf{P}(E|H_1) = 1$.

For the evaluation of $\mathsf{P}(E|H_2)$ we could either sum up all possibilities how the evidence could have arisen under H_2, or - as a shorter way - we can use 3.4. Here we have $E = \{A, B, C\}$, $K = V = \{A, B\} \Rightarrow\ U = E \setminus K = \{C\}$; $x = 1$. Thus

$$\mathsf{P}(E|H_2) = \mathsf{P}_1(C, ABC, AB) = \left(\sum_{l\in\{A,B,C\}} p_l\right)^2 - \sum_{i\in\{C\}} \left(\sum_{\substack{l\in\{A,B,C\}\setminus\{i\}\\ i\in\{C\}}} p_l\right)^2$$

$$= (p_A + p_B + p_C)^2 - (p_A + p_B)^2,$$

which directly describes the fact that for the evidence $E = ABC$ to arise, the unknown individual could have carried alleles A, B or C in his or her genotype, but we must exclude the possibility that he or she would not have carried any C alleles (as that case would not explain the evidence).

The likelihood ratio is hence

$$LR = \frac{\mathsf{P}(E|H_1)}{\mathsf{P}(E|H_2)} = \frac{1}{(p_A + p_B + p_C)^2 - (p_A + p_B)^2},$$

which for example in case that $p_A = 0.1$ and $p_B = p_C = 0.2$ means, that the evidence is 6.25 times more likely to have arisen under H_1 than under H_2. △

Example 3.7. Let us take the situation from the previous example, but assume now that there were two perpetrators of the crime.

The hypotheses will therefore be as follows: H_1*: the contributors to the evidence are the victim, the suspect and one unknown person* and H_2*: the contributors to the evidence are the victim and two unknown people.*

Under H_2, the only change from the previous example is that $x = 2$ instead of $x = 1$, so that

$$\mathsf{P}(E|H_2) = \mathsf{P}_2(C, ABC, AB) = (p_A + p_B + p_C)^4 - (p_A + p_B)^4.$$

Under H_1, the victim and the suspect together explain the evidence, but we still need the one unknown person to "fit" it. More precisely: $E = \{A, B, C\}$, $K = V \cup S = \{A, B, C\} \Rightarrow\ U = E \setminus K = \emptyset$ and $x = 1$. That is

$$\mathsf{P}(E|H_1) = \mathsf{P}_1(\emptyset, ABC, ABC) = \left(\sum_{l \in \{A,B,C\}} p_l \right)^2 = (p_A + p_B + p_C)^2,$$

which actually says nothing more than that the unknown person must only carry alleles of types A, B or C in his or her genotype, regardless of whether or not some particular allele types occur in there.

The likelihood ratio in this case is

$$LR = \frac{\mathsf{P}(E|H_1)}{\mathsf{P}(E|H_2)} = \frac{(p_A + p_B + p_C)^2}{(p_A + p_B + p_C)^4 - (p_A + p_B)^4},$$

which for the values from the previous example means that E is 4.6 times more likely to have arisen under H_1 than under H_2. $\triangle$

Example 3.8. To provide an example with a larger unknown part of the evidence, let us suppose that $E = ABCD$, the victim's profile is homozygous $V = AA$ and we have one suspect typed $S = AB$. Let us (already briefly) analyze two sets of hypotheses:

- *H_1: the contributors to the evidence are the victim, the suspect and one unknown person* and *H_2: the contributors to the evidence are the victim and two unknown people*

$$\begin{aligned}\mathsf{P}(E|H_1) &= \mathsf{P}_1(CD, ABCD, AB) \\ &= (p_A + p_B + p_C + p_D)^2 - (p_A + p_B + p_C)^2 - (p_A + p_B + p_D)^2 \\ &\quad + (p_A + p_B)^2 \\ &= 2p_C p_D\end{aligned}$$

$$
\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_2(BCD, ABCD, A) \\
&= (p_A + p_B + p_C + p_D)^4 - (p_A + p_B + p_C)^4 - (p_A + p_B + p_D)^4 \\
&\quad - (p_A + p_C + p_D)^4 + (p_A + p_B)^4 + (p_A + p_C)^4 \\
&\quad + (p_A + p_D)^4 - p_A^4 \\
&= 12 \cdot (2p_A + p_B + p_C + p_D) p_B p_C p_D
\end{aligned}
$$

- *H_1: the contributors to the evidence are the victim, the suspect and two unknown people* and *H_2: the contributors to the evidence are the victim and three unknown people*

$$
\begin{aligned}
\mathsf{P}(E|H_1) &= \mathsf{P}_2(CD, ABCD, AB) \\
&= (p_A + p_B + p_C + p_D)^4 - (p_A + p_B + p_C)^4 - (p_A + p_B + p_D)^4 \\
&\quad + (p_A + p_B)^4 \\
&= 2p_C p_D (6p_A^2 + 6p_B^2 + 2p_C^2 + 2p_D^2 + 12p_A p_B + 6p_A p_C + 6p_A p_D \\
&\quad + 6p_B p_C + 6p_B p_D + 3p_C p_D)
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_3(BCD, ABCD, A) \\
&= (p_A + p_B + p_C + p_D)^6 - (p_A + p_B + p_C)^6 - (p_A + p_B + p_D)^6 \\
&\quad - (p_A + p_C + p_D)^6 + (p_A + p_B)^6 + (p_A + p_C)^6 \\
&\quad + (p_A + p_D)^6 - p_A^6 \\
&= 30 p_B p_C p_D (2p_A + p_B + p_C + p_D)(2p_A^2 + p_B^2 + p_C^2 + p_D^2 \\
&\quad + 2p_A p_B + 2p_A p_C + 2p_A p_D + p_B p_C + p_B p_D + p_C p_D)
\end{aligned}
$$

$\triangle$

Chapter 4

Substructured population

4.1 Situation description

The previous chapter was dealing with the interpretation of DNA mixtures in the case that the observed population meets Hardy-Weinberg equilibrium. However, this assumption is seldom correct - deviations arise mainly due to population substructure as noted i.a. by Balding and Nichols [1].

In every population there exist subpopulations which have a specific genetic structure different from the structure of the general population. For example, in the U.S. population we can find the Black subpopulation, the Hispanic/Latino subpopulation, the subpopulations of Asian Americans, American Indians, Alaska Natives etc. Allele frequencies in such subpopulations are obviously different from the ones of the whole population.

This is caused primarily by the fact that the members of these subpopulations share a recent ancestry. Therefore there are generally fewer distinct types of alleles in a small subpopulation and their particular frequencies are higher than in the general population, which also means that there are (relatively) more homozygotes in a subpopulation than in the general population.

Mainly it is necessary to realize that random draws from a subpopulation cannot be considered independent from the point of view of the general population, and so we cannot apply directly the results from the previous chapter. We could only do that if we knew the allele frequencies of the observed subpopulation and supposed that it meets Hardy-Weinberg equilibrium - then we could simply move all the previously derived calculations to the "new" population and continue with the "new" allele frequencies.

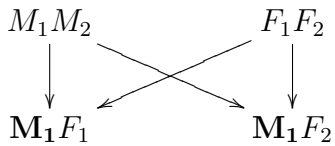
The problem is that the subpopulation frequencies are usually unknown. Thus, we need to manage with the general frequencies and take the substructure into account another way. Ignoring the uncertainty caused by population substructure may lead to significant overstating of the weight of evidence against the defendant and is therefore unfavourable to innocent suspects ([1], [3]).

4.2 Coancestry coefficient

As indicated above, the effect of recent common ancestry on a subpopulation lies in a higher probability that two individuals selected randomly from this subpopulation share the same progenitor. Alternatively, it means a higher probability that two randomly selected alleles can be traced back to one common original allele, i.e. they are both copies of the same allele from the past. Such alleles are referred to as *IBD alleles*, where IBD stands for *identical by descent*.

Example 4.1. We can illustrate the meaning of IBD on a very simple example of two full siblings - each of them inherits one randomly selected allele from their mother's genotype and one randomly selected allele from their father's genotype. With a probability of 50 % they both receive the same allele from the mother - then these two alleles are declared to be IBD. Of course, the situation is absolutely the same for alleles inherited from the father.

For even better illustration we attach a diagram of a concrete situation, where alleles received from the mother are IBD, but alleles from the father are not:



In subpopulations it is typical that the common original allele is found in an earlier generation than at the parents. As a consequence, it is possible that two IBD alleles appear in the genotype of one (homozygous) individual. △

To quantify the effect of the presence of IBD alleles in the subpopulation on its genetic structure, we use so called *coancestry coefficient* θ, defined as the probability that two randomly selected alleles from the subpopulation

are IBD. In some sources, coancestry coefficient is also denoted by F and sometimes it is also called *the degree of population subdivision*. According to [4], θ can be regarded as a measure of the variation in subpopulation allele proportions.

Common values of θ are usually between 0 and 0.1, with $\theta = 0.01$ considered as a conservative value for most populations and $\theta = 0.03$ being often considered appropriate for smaller isolated populations.

Very often, θ is defined as the probability that an allele randomly selected from the genotype of one random person is IBD with an allele randomly selected from the genotype of *another* random person (at the same locus). However, it is unnecessarily complicated to define it this way: the probability that two alleles in the genotype of one person are IBD is actually the same as the probability that an allele randomly selected from the genotype of their mother is IBD with an allele randomly selected from the genotype of their father, while the parents can certainly be considered as two random persons (the influence of sex is insignificant for large enough populations). Therefore our simple definition is equivalent to the more common, but more complicated one. Zoubková [9] on p. 31 provides a more detailed numerical derivation of the equivalence of the definitions.

Although coancestry coefficient gives us some information about the genetic structure of a subpopulation, it still gives us no information about the particular allele frequencies. It is only a "measure of similarity" of the subpopulation members' genotypes. Hence, two subpopulations may have the same θ-value if their evolutionary history was similar, but their allele frequencies can be entirely different. Therefore we need to use some additional information when trying to calculate probabilities of genotypes.

4.3 Homozygous genotypes in substructured population

Balding and Nichols in [1] proposed a method for taking into account the population substructure using the coancestry coefficient. This method has then been taken over by many other authors and according to Fung, Hu [3], it has been used for example in UK courts. Unfortunately, the method was published without proper mathematical derivation and thus we have several concerns about its correctness.

First, Balding and Nichols specify a formula for the probability of a ho-

mozygous genotype AA occurring in a subpopulation with a given θ. They literally state: "Since initially nothing is known about the sub-population frequencies, the probability that an allele drawn in the sub-population is of type A is $P_r[A|p_A] = p_A \ldots$" This is certainly not absolutely correct, because we do not know the probability of drawing an A allele - it would be more precise to say that $\mathsf{P}(A)$ can be *estimated* by the general population frequency p_A. Although this estimate will not always be very good (for example, some alleles may not appear in the subpopulation at all, while their general frequency can be quite large), we may accept it as the only information we have about the particular allele frequency. Possible issues evoked by the estimation error will be discussed later in this section.

Balding and Nichols further state that the probability of a homozygous genotype is (rewritten using our notation)

$$\mathsf{P}(AA) = p_A(\theta + (1-\theta)p_A), \tag{4.1}$$

adding an explanation that "The observation of one A allele in the subpopulation makes it likely that A is more common in the sub-population than in the general population ..." and then trying to justify the equation based on a few statistical and genetic arguments. As we did not find these justifications very clear, we will precise the mathematical consideration leading to the formula.

We are drawing randomly one allele from the subpopulation, knowing a priori nothing about its allele frequencies, and therefore we estimate the probability that this allele is an A by p_A. When drawing the second random allele, knowing that the first one was an A, we cannot consider this draw independent from the first one. The other allele can either be IBD with the first one, which happens with a probability of θ, and it is then clear that they are both of the type A, or it is not IBD and then we again need to estimate the probability that it will be an A based on the general population frequencies. We can summarize this consideration into a simple schematic formula (in which we are actually using the complete probability theorem)[1]:

$$\mathsf{P}(AA) = \mathsf{P}(A)\big(\mathsf{P}(\text{IBD}) \cdot \mathsf{P}(A|\text{IBD}) + \mathsf{P}(\text{nonIBD}) \cdot \mathsf{P}(A|\text{nonIBD})\big). \tag{4.2}$$

As already stated, we agree with Balding and Nichols in the estimation of $\mathsf{P}(A)$ by p_A; values of $\mathsf{P}(\text{IBD})$, $\mathsf{P}(A|\text{IBD})$ and $\mathsf{P}(\text{nonIBD})$ are clear, but

[1]Since this formula is really just schematic, we are using a very simplified notation, with "IBD" and "nonIBD" denoting the random events that the allele selected in the second draw **is IBD** and **is not IBD** with the allele selected in the first draw, respectively.

we disagree with the estimation of $\mathsf{P}(A|\text{nonIBD})$ by p_A again: it would mean allowing for any A allele from the subpopulation to be selected in the second draw, including those IBD with the one selected in the first draw - but we wanted to select only **nonIBD** ones. This way, alleles IBD with the first one are counted twice in the equation, and so for formal correctness they should be omitted in the second term. Therefore we propose the following adjustment:

Denote $s \in (0, 1)$ the proportion of the subpopulation in the general population - we suppose that usually an accurate enough estimate is available. We have already drawn one A allele from the subpopulation. Then the proportion of all alleles IBD with this one in the subpopulation is θ and their proportion in the whole population is hence $s\theta$. The probability of the next allele being an A, given that it is not IBD with the first one, can then be estimated by[2]

$$\frac{p_A - s\theta}{1 - s\theta} \tag{4.3}$$

(i.e. "A alleles from the general population with the IBD alleles excluded / all alleles from the general population with the IBD alleles excluded"). Finally, the formula for the probability of a homozygous genotype will be

$$\mathsf{P}(AA) = p_A\big(\theta \cdot 1 + (1 - \theta) \cdot \frac{p_A - s\theta}{1 - s\theta}\big). \tag{4.4}$$

Although we find it important to perform precise mathematical considerations to obtain formally correct formulas, in this case numerical results show that neglecting the double counting of IBD alleles has quite little practical effect for common values of s, θ and p_A. The obtained probability values usually differ in the third or fourth decimal place, with the results of the Balding-Nichols equation being usually just several percent larger than the results of the adjusted equation. The concrete numbers for some particular common values of s, θ and p_A are given in Table 4.1.

One more issue we have identified with the Balding-Nichols method of calculation (both the original and the adjusted one) is the fact that it returns an increased probability of homozygotes for all types of alleles which appear in the general population, although some of them can be extremely rare in the subpopulation. It might therefore not reflect reality correctly -

[2]Having an allele of type A in the subpopulation, there are θ ($\times$ the subpopulation size) alleles IBD with it - thus also of type A. Therefore we can suppose that $p_A \geq \theta \geq s\theta$.

p_A	s	θ	p_A^2	P_1	P_2	P_1/P_2
0.01	0.01	0.01	0.0001	0.0002	0.0002	1.0050
0.01	0.01	0.03	0.0001	0.0004	0.0004	1.0073
0.01	0.01	0.05	0.0001	0.0006	0.0006	1.0080
0.01	0.01	0.10	0.0001	0.0011	0.0011	1.0083
0.01	0.05	0.01	0.0001	0.0002	0.0002	1.0253
0.01	0.05	0.03	0.0001	0.0004	0.0004	1.0377
0.01	0.05	0.05	0.0001	0.0006	0.0006	1.0413
0.01	0.05	0.10	0.0001	0.0011	0.0010	1.0428
0.01	0.10	0.01	0.0001	0.0002	0.0002	1.0519
0.01	0.10	0.03	0.0001	0.0004	0.0004	1.0785
0.01	0.10	0.05	0.0001	0.0006	0.0005	1.0863
0.01	0.10	0.10	0.0001	0.0011	0.0010	1.0900
0.01	0.20	0.01	0.0001	0.0002	0.0002	1.1095
0.01	0.20	0.03	0.0001	0.0004	0.0003	1.1710
0.01	0.20	0.05	0.0001	0.0006	0.0005	1.1900
0.01	0.20	0.10	0.0001	0.0011	0.0009	1.2002
0.05	0.01	0.01	0.0025	0.0030	0.0030	1.0016
0.05	0.01	0.03	0.0025	0.0039	0.0039	1.0035
0.05	0.01	0.05	0.0025	0.0049	0.0049	1.0047
0.05	0.01	0.10	0.0025	0.0073	0.0072	1.0059
0.05	0.05	0.01	0.0025	0.0030	0.0030	1.0080
0.05	0.05	0.03	0.0025	0.0039	0.0039	1.0180
0.05	0.05	0.05	0.0025	0.0049	0.0048	1.0238
0.05	0.05	0.10	0.0025	0.0073	0.0070	1.0305
0.05	0.10	0.01	0.0025	0.0030	0.0029	1.0161
0.05	0.10	0.03	0.0025	0.0039	0.0038	1.0366
0.05	0.10	0.05	0.0025	0.0049	0.0046	1.0488
0.05	0.10	0.10	0.0025	0.0073	0.0068	1.0633
0.05	0.20	0.01	0.0025	0.0030	0.0029	1.0327
0.05	0.20	0.03	0.0025	0.0039	0.0036	1.0763
0.05	0.20	0.05	0.0025	0.0049	0.0044	1.1031
0.05	0.20	0.10	0.0025	0.0073	0.0064	1.1368
0.10	0.01	0.01	0.0100	0.0109	0.0109	1.0008
0.10	0.01	0.03	0.0100	0.0127	0.0127	1.0021
0.10	0.01	0.05	0.0100	0.0145	0.0145	1.0030
0.10	0.01	0.10	0.0100	0.0190	0.0189	1.0043
0.10	0.05	0.01	0.0100	0.0109	0.0109	1.0041
0.10	0.05	0.03	0.0100	0.0127	0.0126	1.0104
0.10	0.05	0.05	0.0100	0.0145	0.0143	1.0150
0.10	0.05	0.10	0.0100	0.0190	0.0186	1.0219
0.10	0.10	0.01	0.0100	0.0109	0.0108	1.0083
0.10	0.10	0.03	0.0100	0.0127	0.0124	1.0211
0.10	0.10	0.05	0.0100	0.0145	0.0141	1.0305
0.10	0.10	0.10	0.0100	0.0190	0.0182	1.0450
0.10	0.20	0.01	0.0100	0.0109	0.0107	1.0167
0.10	0.20	0.03	0.0100	0.0127	0.0122	1.0433
0.10	0.20	0.05	0.0100	0.0145	0.0136	1.0633
0.10	0.20	0.10	0.0100	0.0190	0.0173	1.0953
0.25	0.01	0.01	0.0625	0.0644	0.0644	1.0003
0.25	0.01	0.03	0.0625	0.0681	0.0681	1.0008
0.25	0.01	0.05	0.0625	0.0719	0.0718	1.0012
0.25	0.01	0.10	0.0625	0.0813	0.0811	1.0021
0.25	0.05	0.01	0.0625	0.0644	0.0643	1.0014
0.25	0.05	0.03	0.0625	0.0681	0.0679	1.0040
0.25	0.05	0.05	0.0625	0.0719	0.0714	1.0063
0.25	0.05	0.10	0.0625	0.0813	0.0804	1.0105
0.25	0.10	0.01	0.0625	0.0644	0.0642	1.0029
0.25	0.10	0.03	0.0625	0.0681	0.0676	1.0081
0.25	0.10	0.05	0.0625	0.0719	0.0710	1.0126
0.25	0.10	0.10	0.0625	0.0813	0.0795	1.0214
0.25	0.20	0.01	0.0625	0.0644	0.0640	1.0058
0.25	0.20	0.03	0.0625	0.0681	0.0670	1.0164
0.25	0.20	0.05	0.0625	0.0719	0.0701	1.0257
0.25	0.20	0.10	0.0625	0.0813	0.0778	1.0443

$$P_1 = p_A(\theta + (1-\theta)\cdot p_A)$$
$$P_2 = p_A(\theta + (1-\theta)\cdot \tfrac{p_A - s\theta}{1-s\theta})$$

Table 4.1: Numerical comparation of the results of the original and adjusted version of Balding-Nichols formula for the probability of homozygous genotypes

although there are relatively more homozygotes in the subpopulation than in the general population, it is actually caused by the fact that fewer distinct types of alleles appear in the subpopulation, which also means fewer distinct types of homozygotes. Increasing the probability of any homozygote is therefore obviously improper and the question is whether Balding-Nichols model really gives more accurate results then the simple calculation based on the assumption of independence.

4.4 Heterozygous genotypes in substructured population

Analogous considerations as those on calculating probabilities of homozygotes can be of course performed to obtain formulas for heterozygous genotypes. Since there are still some differences, we will briefly summarize the results.

Balding and Nichols state that the probability of a heterozygous genotype AB occurring in a subpopulation with coancestry coefficient θ is

$$\mathsf{P}(AB) = p_A p_B (1 - \theta). \tag{4.5}$$

The precise mathematical consideration which leads to the heterozygote-formula should be

$$\begin{aligned}\mathsf{P}(AB) = \mathsf{P}(A)\big(&\mathsf{P}(\text{IBD}) \cdot \mathsf{P}(B|\text{IBD}) + \mathsf{P}(\text{nonIBD}) \cdot \mathsf{P}(B|\text{nonIBD})\big) \\ + \mathsf{P}(B)\big(&\mathsf{P}(\text{IBD}) \cdot \mathsf{P}(A|\text{IBD}) + \mathsf{P}(\text{nonIBD}) \cdot \mathsf{P}(A|\text{nonIBD})\big).\end{aligned} \tag{4.6}$$

The values (or their estimates) of $\mathsf{P}(A)$, $\mathsf{P}(B)$, $\mathsf{P}(\text{IBD})$ and $\mathsf{P}(\text{nonIBD})$ are clear; $\mathsf{P}(B|\text{IBD})$ (i.e. the probability that the second allele is a B given that it is IBD with the A allele already drawn) is obviously 0, as well as $\mathsf{P}(A|\text{IBD})$ in the second part of the equation. But again we propose an adjustment to the estimation of $\mathsf{P}(B|\text{nonIBD})$ (the probability of drawing a B allele, knowing that it is not IBD with the A allele already drawn) by omitting the IBD alleles from the calculation:

$$\mathsf{P}(B|\text{nonIBD}) = \frac{p_B}{1 - s\theta}. \tag{4.7}$$

This time we only need to exclude IBD alleles from the whole population, as there are obviously no alleles IBD with an A among the B alleles in the

subpopulation. The situation for $\mathsf{P}(A|\text{nonIBD})$ in the second part of the equation is naturally the same.

As a result, we obtain the adjusted formula for the probability of a heterozygous genotype in the subpopulation:

$$\mathsf{P}(AB) = 2p_A p_B \frac{1-\theta}{1-s\theta}. \tag{4.8}$$

In this case, our formula gives larger results than its original Balding-Nichols version, mainly because Balding and Nichols omitted the factor 2 for some reason - but then $\mathsf{P}(AA) + \mathsf{P}(AB) + \mathsf{P}(BB)$ do not sum up to 1 in the case that $p_A + p_B = 1$, which is obviously incorrect, but the error can be fixed just by adding the missing factor. Besides that, our results are also larger because, informally said, selecting a B allele from alleles not IBD with an A allele is "easier" than selecting a B from all alleles in the subpopulation.

4.5 Extended formula

The above considerations can be extended to the case of randomly drawing more than two alleles from a subpopulation, which becomes useful in the evaluation of DNA mixtures. Balding and Nichols in [1] give a formula for the conditional probability of an A allele being selected from a subpopulation, given that r alleles of type A and s alleles of type B have already been selected:

$$\mathsf{P}(A^{r+1}, B^s | A^r, B^s) = \frac{r\theta + p_A(1-\theta)}{1 + (r+s-1)\theta}. \tag{4.9}$$

Since no p_B appears in the r. h. s. of the equation, it seems clear that the probability of an A allele occurring in the $(r+s+1)^{\text{th}}$ draw from the subpopulation only depends on the fact that there were exactly r alleles of type A selected in the previous draws, regardless of the types of the other s alleles. Therefore a more appropriate form of the formula can be given as follows:

Theorem 4.2. *Let $\theta \in [0,1]$, $r, s \in \mathbb{N}$ with (r,s) denoting the event that exactly r A-alleles have been drawn out of $r+s$ random draws from a subpopulation. Then it holds*

$$\mathsf{P}(r+1, s | r, s) = \frac{r\theta + p_A(1-\theta)}{1 + (r+s-1)\theta}. \tag{4.10}$$

Some justifications of this formula have been provided by Balding and Nichols in [2]; a complete proof however was not shown. Although we have some mathematical concerns regarding those partial justifications, a complete revision would require a more genetic approach than is the one of this work, and so we are leaving it as a suggestion for further research.

Accepting the extended formula in the above form is further supported by Slovák [6], who obtains it as a result of a mathematical derivation based on [8].

Moreover, we can see that $\mathsf{P}(r+1, s|r, s)$ is a probability with an expected behaviour in special cases (which is of course not a proof of the formula correctness, but it allows us to treat it as a probability, assuming that the value is correct):

- Clearly $\mathsf{P}(\mathbf{r}+\mathbf{1}, \mathbf{s}|\mathbf{r}, \mathbf{s}) \geq \mathbf{0}$, since it is a ratio of two positive numbers for any $r, s \in \mathbb{N}$, $\theta \in [0, 1]$ and $p_A \in [0, 1]$.

- $\mathsf{P}(\mathbf{r}+\mathbf{1}, \mathbf{s}|\mathbf{r}, \mathbf{s}) \leq \mathbf{1}$, since for any $r, s \in \mathbb{N}$, $\theta \in [0, 1]$ and $p_A \in [0, 1]$ it holds:

$$\begin{aligned} p_A(1-\theta) &\leq 1-\theta \\ \Rightarrow p_A(1-\theta) &\leq 1-\theta+s\theta \\ \Rightarrow r\theta + p_A(1-\theta) &\leq 1+(r+s-1)\theta \end{aligned}$$

- **σ-additivity**: the probability, that an allele of type A will be selected in the next draw + the probability that an allele of type B will be selected must be equal to the probability that A OR B will be selected, since "A selected" and "B selected" are disjoint random events. That is, having already drawn r A-alleles and s B-alleles out of $r+s+t$ random draws, it must be

  $$\mathsf{P}(r+1, s+t|r, s+t) + \mathsf{P}(s+1, r+t|s, r+t) = \mathsf{P}(r+s+1, t|r+s, t),$$

  which is fulfilled, as we can see from the expansion of all three terms:

  $$\frac{r\theta + p_A(1-\theta)}{1+(r+s+t-1)\theta} + \frac{s\theta + p_B(1-\theta)}{1+(r+s+t-1)\theta} = \frac{(r+s)\theta + (p_A+p_B)(1-\theta)}{1+(r+s+t-1)\theta}$$

- The probability of all possible random events sums up to 1:

  $$\mathsf{P}(A \text{ selected}) + \mathsf{P}(A \text{ not selected}) = \mathsf{P}(r+1, s|r, s) + \mathsf{P}(r, s+1|r, s) = 1,$$

because

$$\frac{r\theta + p_A(1-\theta)}{1+(r+s-1)\theta} + \frac{s\theta + (1-p_A)(1-\theta)}{1+(r+s-1)\theta} = \frac{r\theta + s\theta + 1 - \theta}{1+(r+s-1)\theta} = 1.$$

- $\theta = 0$ is the independence case:

$$\frac{r \cdot 0 + p_A(1-0)}{1-(r+s-1)\cdot 0} = p_A.$$

4.6 $\mathsf{P}_x(U, E, K)$ in substructured population

The methods described in Sections 3.2 and 3.3 may naturally apply also to the substructured population case. However, the enumeration of $\mathsf{P}_x(U, E, K)$ becomes more complicated without the assumption of independence, as Theorem 4.2 (i.e. the Balding-Nichols formula) needs to be applied. Fung and Hu [3] derived an extension of the formula 3.4, including its proof. Since not all of their statements are absolutely mathematically correct or complete, we provide a detailed revision.

First, for the purpose of incorporating θ in the calculation, we will define a special type of "power function":

Definition 4.3. For a fixed $\theta \in [0,1]$, for $p \in \mathbb{R}$ and $m, k \in \mathbb{N}$ define

$$\begin{aligned} p^{(m)}(k) &= [k\theta + (1-\theta)p] \cdot [(k+1)\theta + (1-\theta)p] \cdot \ \ldots \ \cdot [(k+m-1)\theta + \\ &\quad + (1-\theta)p] \\ &= \prod_{i=0}^{m-1} [(k+i)\theta + (1-\theta)p] \end{aligned} \tag{4.11}$$

for $m > 0$, and

$$p^{(0)}(k) = 1 \tag{4.12}$$

for $m = 0$ and for any $p \in \mathbb{R}$, $k \in \mathbb{N}$.

Specially,

$$\begin{aligned} 1^{(m)}(k) &= [1 + (k-1)\theta] \cdot [1 + k\theta] \cdot \ \ldots \ \cdot [1 + (k+m-2)\theta] \\ &= \prod_{i=0}^{m-1} [1 + (k+i-1)\theta], \ m > 0. \end{aligned} \tag{4.13}$$

Note 4.4. It can be easily seen that for $\theta = 0$ we obtain $p^{(m)}(k) = p^m$ for any k, which can actually represent reduction to the independence case with no θ incorporated.

We will now derive an extension of the binomial and multinomial expansion theorems, applying to the function from Definition 4.3.

Lemma 4.5. *For any $p_1, p_2 \in \mathbb{R}$ and $m, k_1, k_2 \in \mathbb{N} \setminus \{0\}$, it holds*

$$\begin{aligned}(p_1 + p_2)^{(m)}(k_1 + k_2) &= p_1^{(m)}(k_1) + p_2^{(m)}(k_2) + \sum_{i=1}^{m-1} \binom{m}{i} p_1^{(i)}(k_1) p_2^{(m-i)}(k_2) \\ &= \sum_{\substack{i_1+i_2=m \\ i_1,i_2\in\mathbb{N}}} \binom{m}{i_1} p_1^{(i_1)}(k_1) p_2^{(i_2)}(k_2). \end{aligned} \tag{4.14}$$

Proof. We will use mathematical induction. For $m = 1$ we have:

$$\begin{aligned}(p_1 + p_2)^{(1)}(k_1 + k_2) &= (k_1 + k_2)\theta + (1 - \theta)(p_1 + p_2) \\ &= (k_1\theta + (1 - \theta)p_1) + (k_2\theta + (1 - \theta)p_2) \\ &= p_1^{(1)}(k_1) + p_2^{(1)}(k_2) \\ &= \sum_{\substack{i_1+i_2=1 \\ i_1,i_2\in\mathbb{N}}} \binom{1}{i_1} p_1^{(i_1)}(k_1) p_2^{(i_2)}(k_2).\end{aligned}$$

Assuming that the theorem holds for some $m \in \mathbb{N}, m \geq 1$ we get:

$$\begin{aligned}(p_1 + p_2)^{(m+1)}(k_1 + k_2) &= (p_1 + p_2)^{(m)}(k_1 + k_2) \cdot [(k_1 + k_2 + m)\theta \\ &\quad + (1 - \theta)(p_1 + p_2)] \\ &= \sum_{i=0}^{m} \binom{m}{i} p_1^{(i)}(k_1) p_2^{(m-i)}(k_2) \cdot [(k_1 + i)\theta + (1 - \theta)p_1 \\ &\quad + (k_2 + m - i)\theta + (1 - \theta)p_2] =: \star\end{aligned}$$

Here we just wrote the last factor separately and then used the induction hypothesis. By expanding the square brackets and realizing that $p^{(m+1)}(k) = p^{(m)}(k) \cdot [(k + m)\theta + (1 - \theta)p]$ we further obtain:

$$\star = \sum_{i=0}^{m} \binom{m}{i} p_1^{(i+1)}(k_1) p_2^{(m-i)}(k_2) + \sum_{i=0}^{m} \binom{m}{i} p_1^{(i)}(k_1) p_2^{(m-i+1)}(k_2)$$

Writing separately the m-th and the 0-th term from the first and the second sum, respectively, gives:

$$\begin{aligned}\star &= p_1^{(m+1)}(k_1) + p_2^{(m+1)}(k_2) + \\ &\quad + \sum_{i=0}^{m-1} \binom{m}{i} p_1^{(i+1)}(k_1) p_2^{(m-i)}(k_2) + \sum_{i=1}^{m} \binom{m}{i} p_1^{(i)}(k_1) p_2^{(m-i+1)}(k_2) \\ &= p_1^{(m+1)}(k_1) + p_2^{(m+1)}(k_2) + \\ &\quad + \sum_{i=1}^{m} \left(\binom{m}{i-1} + \binom{m}{i} \right) p_1^{(i)}(k_1) p_2^{(m-i+1)}(k_2)\end{aligned}$$

Now by just summing up the binomial coefficients we receive the theorem for $m+1$:

$$\star = p_1^{(m+1)}(k_1) + p_2^{(m+1)}(k_2) + \sum_{i=1}^{m} \binom{m+1}{i} p_1^{(i)}(k_1) p_2^{(m+1-i)}(k_2)$$

□

Again we can see that $\theta = 0$ results in the usual binomial expansion theorem.

Applying induction once more, this time on the number of terms in the brackets, we immediately obtain

Lemma 4.6. *For $p_1, p_2, \ldots, p_l \in \mathbb{R}$, $l \in \mathbb{N}^+$, $m, k_1, k_2, \ldots, k_l \in \mathbb{N}^+$*

$$\begin{aligned}&(p_1 + p_2 + \cdots + p_l)^{(m)}(k_1 + k_2 + \cdots + k_l) = \\ &\quad = \sum_{\substack{i_1+i_2+\cdots+i_l=m \\ i_1,i_2,\ldots,i_l \in \mathbb{N}}} \binom{m}{i_1, i_2, \ldots, i_l} \prod_{j=1}^{l} p_j^{(i_j)}(k_j) \qquad (4.15)\end{aligned}$$

where

$$\binom{m}{i_1, i_2, \ldots, i_l} = \frac{m!}{i_1! i_2! \ldots i_l!}$$

are multinomial coefficients.

Before we formulate the theorem about evaluation of $\mathsf{P}_x(U, E, K)$ in a substructured population, we have to realize one important difference from the independence case: while in the independence case we used K to denote

the set of all distinct alleles from known contributors and U was directly determined by the relation $U = E \setminus K$, now - due to dependence of genotypes - we have to take into account **all** known genotypes, regardless of whether or not their carrier contributed to the evidence (under the specified hypothesis). Furthermore, we need to consider how many alleles of each type occur in the known genotypes, since each observed allele gives us some information about the genetic structure of the subpopulation. Therefore we have to change our notation in the following way:

- E the evidential set of alleles,
- x the supposed number of unknown contributors to the evidence,
- y the number of typed persons (regardless of whether they contribute to the evidence),
- $\mathcal{K} = (\kappa_1, \kappa_2, \ldots, \kappa_{2y})$ the $2y$-tuple of all alleles from the typed persons,
- K the set of all distinct alleles from $\mathcal{K}$,
- k_{A_i} the number of alleles of the type A_i contained in $\mathcal{K}$, including $k_{A_i} = 0$ when $A_i \notin K$ (then clearly $\sum_{A_i \in E} k_{A_i} = 2y$),
- R the set of all distinct alleles from the unknown contributors (in other words, the set of all distinct alleles from $2x$ random draws from the subpopulation),
- U the unexplained part of the evidence.

U is not directly determined by E and K any more, since it depends on who of the typed persons were contributors to the evidence.

Because of the dependence on k_{A_i}, we will now denote the probability that $U \subset R \subset E$ by $\mathsf{P}_x(U, E, \mathcal{K})$. Alleles from E will be again denoted briefly by $1, 2, \ldots, e$ instead of $A_1, A_2, \ldots, A_e$.

Theorem 4.7. *For $x, y, e \in \mathbb{N}$, $E = \{1, 2, \ldots, e\}$, $\mathcal{K} = (\kappa_1, \kappa_2, \ldots, \kappa_{2y})$, K being the set of all distinct values from $\mathcal{K}$, $U \subset E$ being the unexplained part of the evidence and θ being the coancestry coefficient of the subpopulation,*

it holds

$$\begin{aligned}\mathsf{P}_x(U,E,\mathcal{K}) = \frac{1}{1^{(2x)}(2y)} \cdot \big((T_0)^{(2x)}(t_0) - \sum_{i\in U}(T_{1i})^{(2x)}(t_{1i}) \\ + \sum_{i,j\in U; i<j}(T_{2ij})^{(2x)}(t_{2ij}) - \cdots + (-1)^u (T_{uU})^{(2x)}(t_{uU})\big)\end{aligned} \tag{4.16}$$

where $u = |U|$ *and*

$$t_0 = \sum_{l\in E} k_l, \quad T_0 = \sum_{l\in E} p_l,$$

$$t_{1i} = \sum_{l\in E\setminus\{i\}} k_l, \quad T_{1i} = \sum_{l\in E\setminus\{i\}} p_l, \quad i \in E,$$

$$t_{2ij} = \sum_{l\in E\setminus\{i,j\}} k_l, \quad T_{2ij} = \sum_{l\in E\setminus\{i,j\}} p_l, \quad i,j \in E; i<j,$$

$$\vdots$$

$$t_{uU} = \sum_{l\in E\setminus U} k_l, \quad T_{uU} = \sum_{l\in E\setminus U} p_l,$$

k_l being the number of l-alleles in $\mathcal{K}$ and p_l being the frequency of l in the population.

Proof. For the purpose of proving the theorem we establish a probability model which considers random allele draws from a subpopulation as consecutive random draws from boxes, each of which contains all possible alleles for the observed locus.

More specifically, let us have $2y+2x$ boxes with n balls labelled $1, 2, \ldots, n$ in each. Let $p_1, p_2, \ldots, p_n \in \mathbb{R}^+$, satisfying $\sum_{i=1}^n p_i = 1$. p_i represents the probability of drawing the ball labelled i from one box when performing an independent draw.

We now draw one ball from each box in sequence. We will denote by $\mathcal{K} = (\kappa_1, \kappa_2, \ldots, \kappa_{2y})$ the labels of the first $2y$ balls, i.e. κ_i is the label of the ball drawn from the i-th box, $i = 1, 2, \ldots, 2y$. Similarly, denote $\mathcal{R} = (r_1, r_2, \ldots, r_{2x})$ the labels of the balls drawn from the last $2x$ boxes, i.e. r_i is the label of the ball drawn from the $2y + i$-th box, $i = 1, 2, \ldots, 2x$. Further, denote i_j the number of balls labelled j in $\mathcal{R}$, $j = 1, 2, \ldots, n$ (so

that $\sum_{j=1}^{n} i_j = 2x$) and k_j will be the number of balls labelled j in $\mathcal{K}$, $j = 1, 2, \ldots, n$ (so that $\sum_{j=1}^{n} k_j = 2y$).

Assuming that Theorem 4.2 applies when drawing balls from the boxes consequently, we obtain that

$$\mathsf{P}(\mathcal{R}|\mathcal{K}) = \frac{\prod_{j\in\{1,\ldots,n\}} \prod_{i=0}^{i_j-1}[(k_j+i)\theta + p_j(1-\theta)]}{\prod_{i=0}^{2x-1}[1+(2y+i-1)\theta]} = \frac{\prod_{j=1}^{n} p_j^{(i_j)}(k_j)}{1^{(2x)}(2y)}. \quad (4.17)$$

As $\mathsf{P}(\mathcal{R}|\mathcal{K})$ is obviously independent from the order of the balls in $\mathcal{R}$ (and also in $\mathcal{K}$), we may write:

$$\mathsf{P}(i_1, i_2, \ldots, i_n | k_1, k_2, \ldots, k_n) = \binom{2x}{i_1, i_2, \ldots, i_n} \frac{\prod_{j=1}^{n} p_j^{(i_j)}(k_j)}{1^{(2x)}(2y)}. \quad (4.18)$$

Using Lemma 4.6 we obtain that

$$\begin{aligned} \sum_{\mathcal{R}} \mathsf{P}(\mathcal{R}|\mathcal{K}) &= \sum_{i_1+i_2+\cdots+i_n=2x} \binom{2x}{i_1, i_2, \ldots, i_n} \frac{\prod_{j=1}^{n} p_j^{(i_j)}(k_j)}{1^{(2x)}(2y)} \\ &= \frac{(p_1+p_2+\cdots+p_n)^{(2x)}(k_1+k_2+\cdots+k_n)}{1^{(2x)}(2y)} \\ &= \frac{1^{(2x)}(2y)}{1^{(2x)}(2y)} = 1. \end{aligned} \quad (4.19)$$

Analogically, for any $S \subset \{1, 2, \ldots, n\}$ it holds

$$\sum_{\mathcal{R}: u_1\in S, u_2\in S, \ldots, u_{2x}\in S} \mathsf{P}(\mathcal{R}|\mathcal{K}) = \frac{\left(\sum_{l\in S} p_l\right)^{(2x)} \left(\sum_{l\in S} k_l\right)}{1^{(2x)}(2y)}. \quad (4.20)$$

Now let $S_E \subset E \subset \{1, 2, \ldots, n\}$ and we want to calculate the probability that $S_E \subset R \subset E$ given K, where R is the set of all distinct labels from $\mathcal{R}$.

For $M \subset E$ denote F_M the event that no ball with a label contained in M occurs in $\mathcal{R}$. Then, using 4.20, we have:

$$\mathsf{P}(R \subset E|\mathcal{K}) = \frac{\left(\sum_{l\in E} p_l\right)^{(2x)} \left(\sum_{l\in E} k_l\right)}{1^{(2x)}(2y)} = \frac{(T_0)^{(2x)}(t_0)}{1^{(2x)}(2y)},$$

$$\mathsf{P}((R \subset E) \cap F_{\{i\}}|\mathcal{K}) = \frac{\left(\sum_{l\in E\setminus\{i\}} p_l\right)^{(2x)} \left(\sum_{l\in E\setminus\{i\}} k_l\right)}{1^{(2x)}(2y)} = \frac{(T_{1i})^{(2x)}(t_{1i})}{1^{(2x)}(2y)}, \quad i \in E,$$

$$\mathsf{P}((R \subset E) \cap F_{\{i,j\}}|\mathcal{K}) = \frac{\left(\sum_{l\in E\setminus\{i,j\}} p_l\right)^{(2x)} \left(\sum_{l\in E\setminus\{i,j\}} k_l\right)}{1^{(2x)}(2y)} = \frac{(T_{2ij})^{(2x)}(t_{2ij})}{1^{(2x)}(2y)},$$
$$i,j \in E, i \neq j,$$

$$\vdots$$

Generally,

$$\mathsf{P}((R \subset E) \cap F_M|\mathcal{K}) = \frac{\left(\sum_{l\in E\setminus M} p_l\right)^{(2x)} \left(\sum_{l\in E\setminus M} k_l\right)}{1^{(2x)}(2y)} =: \frac{(T_{|M|M})^{(2x)}(t_{|M|M})}{1^{(2x)}(2y)}. \tag{4.21}$$

For the probability that $S_E \subset R \subset E$ given $\mathcal{K}$ we now have:

$$\mathsf{P}(S_E \subset R \subset E|\mathcal{K}) = \mathsf{P}\big((R \subset E) \bigcap_{i\in S_E} F^{\mathcal{C}}_{\{i\}}|\mathcal{K}\big) =: \diamond,$$

since all labels from S_E must have been drawn at least once from the last $2x$ boxes for S_E to be subset of R. ($F^{\mathcal{C}}$ denotes the complement of the event F.)

Using the principle of inclusion and exclusion we obtain:

$$\begin{aligned}\diamond &= \mathsf{P}(R \subset E|\mathcal{K}) - \mathsf{P}\big(\bigcup_{i\in S_E} [(R \subset E) \cap F_{\{i\}}]|\mathcal{K}\big) \\ &= \Big((T_0)^{(2x)}(t_0) - \sum_{i\in S_E} (T_{1i})^{(2x)}(t_{1i}) + \sum_{i,j\in S_E; i<j} (T_{2ij})^{(2x)}(t_{2ij}) - \dots \\ &\quad + (-1)^{|S_E|}(T_{|S_E|S_E})^{(2x)}(t_{|S_E|S_E})\Big) \cdot \frac{1}{1^{(2x)}(2y)}\end{aligned}$$

Putting now $S_E = U$ we get the theorem:

$$\mathsf{P}(U \subset R \subset E|\mathcal{K}) = \frac{1}{1^{(2x)}(2y)} \cdot \Big((T_0)^{(2x)}(t_0) - \sum_{i \in U} (T_{1i})^{(2x)}(t_{1i}) + \sum_{i,j \in U; i<j} (T_{2ij})^{(2x)}(t_{2ij}) - \ldots + (-1)^{|U|} (T_{|U|(U)})^{(2x)}(t_{|U|(U)}) \Big)$$

□

Note 4.8. Using the notation from 4.21, we could also write the theorem in the following form:

$$\mathsf{P}_x(U, E, \mathcal{K}) = \frac{1}{1^{(2x)}(2y)} \cdot \sum_{z=0}^{|U|} \left((-1)^z \cdot \sum_{\substack{M \subset U \\ |M|=z}} (T_{|M|M})^{(2x)}(t_{|M|M}) \right) \tag{4.22}$$

Note 4.9. From Note 4.4 we can see that $\theta = 0$ results in

$$\mathsf{P}_x(U, E, \mathcal{K}) = (T_0)^{2x} - \sum_{i \in U} (T_{1i})^{2x} + \sum_{i,j \in U; i<j} (T_{2ij})^{2x} - \cdots + (-1)^{|U|} (T_{|U|(U)})^{2x},$$

which is 3.4. Therefore Theorem 4.7 is a direct extension of Theorem 3.5 from the independence case to the case of a substructured population.

The computation of $\mathsf{P}_x(U, E, \mathcal{K})$ based on Theorem 4.7 can be easily implemented in any common programming software. In Figure 4.1 we give an example of a complete implementation in Mathematica 6 based on the alternative form of the theorem given in 4.22. The code particularly refers to Example 4.10 (with the alleles denoted $1, 2, 3$ instead of A, B, C in accordance with the notation from Theorem 4.7), but can be trivially changed to calculate $\mathsf{P}_x(U, E, \mathcal{K})$ in any other situation by just changing the values in the first two rows. The result of the computation can be obtained by calling

```
P[x_, U_, E_, K_]
```

with the desired values, that is e.g.

```
P[1, {3}, {1, 2, 3}, {1, 2, 3, 3}]
```

for the case of Example 4.10.

```
theta = 0.05;
db = SparseArray[{1 -> 0.1, 2 -> 0.2, 3 -> 0.2}];

ThetaPower[p_, m_, k_] :=
  If[m == 0,
   1,
   Product[
    (k + i)*theta + (1 - theta)*p,
    {i, 0, m - 1}]];

T[E_, M_] :=
  Sum[
   db[[l]],
   {l, Complement[E, M]}];

t[E_, M_, K_] :=
  Sum[
   Count[K, l],
   {l, Complement[E, M]}];

P[x_, U_, E_, K_] :=
  (
   y = Length[K]/2;
   1/(ThetaPower[1, 2*x, 2*y])*Sum[
     (-1)^z*Sum[
       ThetaPower[T[E, M], 2*x, t[E, M, K]],
       {M, Subsets[U, {z}]}],
     {z, 0, Length[U]}
     ]
   );
```

Figure 4.1: Implementation of Theorem 4.7 in Mathematica 6

We will now repeat calculations from Chapter 3 for the case of a substructured population to show the usage of Theorem 4.7.

Example 4.10 (Continuation of Example 3.6). We have $E = ABC$, $V = AB$ and $S = CC$, while the hypotheses are H_1*: the contributors to the evidence are the victim and the suspect* and H_2*: the contributors to the evidence are the victim and an unknown person.* Now $\mathcal{K} = (A, B, C, C) =: ABCC$, $K = A, B, C$.

Then again $\mathsf{P}(E|H_1) = 1$ and

$$\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_1(C, ABC, ABCC) \\
&= \frac{1}{1^{(2)}(4)} \cdot \big((T_0)^{(2)}(t_0) - \sum_{i\in\{C\}} (T_{1i})^{(2)}(t_{1i})\big) \\
&= \frac{1}{1^{(2)}(4)} \cdot [(p_A + p_B + p_C)^{(2)}(k_A + k_B + k_C) \\
&\quad - (p_A + p_B)^{(2)}(k_A + k_B)] \\
&= \frac{(4\theta + (1-\theta)(p_A + p_B + p_C))(5\theta + (1-\theta)(p_A + p_B + p_C))}{(1+3\theta)(1+4\theta)} \\
&\quad - \frac{(2\theta + (1-\theta)(p_A + p_B))(3\theta + (1-\theta)(p_A + p_B))}{(1+3\theta)(1+4\theta)} \\
&= \frac{(2\theta + (1-\theta)p_C)(7\theta + (1-\theta)p_C + 2(1-\theta)(p_A + p_B))}{(1+3\theta)(1+4\theta)} \,.
\end{aligned}$$

Supposing again that $p_A = 0.1$ and $p_B = p_C = 0.2$ and additionally $\theta = 0.05$, we find out that E is 4.29 times more likely to have arisen under H_1 than under H_2, compared to $LR = 6.25$ in the independence case. Here we can see that the uncertainty about the population genetic substructure weakens the strength of the evidence against the suspect. △

Example 4.11 (Continuation of Example 3.7). The situation is the same as in the previous example, but the hypotheses are now: H_1*: the contributors to the evidence are the victim, the suspect and an unknown person* and H_2*: the contributors to the evidence are the victim and two unknown persons.*

Under H_1 we have $x = 1$, $y = 2$, $\mathcal{K} = (A, B, C, C) =: ABCC$, thus

$$\begin{aligned}
\mathsf{P}(E|H_1) &= \mathsf{P}_1(\emptyset, ABC, ABCC) \\
&= \frac{1}{1^{(2)}(4)} \cdot \big((T_0)^{(2)}(t_0)\big) \\
&= \frac{1}{1^{(2)}(4)} \cdot [(p_A + p_B + p_C)^{(2)}(k_A + k_B + k_C)] \\
&= \frac{(4\theta + (1-\theta)(p_A + p_B + p_C))(5\theta + (1-\theta)(p_A + p_B + p_C))}{(1+3\theta)(1+4\theta)}
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_2(C, ABC, ABCC) \\
&= \frac{1}{1^{(4)}(4)} \cdot \big((T_0)^{(4)}(t_0) - \sum_{i \in \{C\}} (T_{1i})^{(4)}(t_{1i})\big) \\
&= \frac{1}{1^{(4)}(4)} \cdot [(p_A + p_B + p_C)^{(4)}(k_A + k_B + k_C) \\
&\quad - (p_A + p_B)^{(4)}(k_A + k_B)] \\
&= \frac{1}{1^{(4)}(4)} \cdot [(p_A + p_B + p_C)^{(4)}(4) - (p_A + p_B)^{(4)}(2)] \\
&= \frac{1}{(1+3\theta)(1+4\theta)(1+5\theta)(1+6\theta)} \cdot \\
&\quad \cdot \big((4\theta + (1-\theta)(p_A + p_B + p_C)) \cdot (5\theta + (1-\theta)(p_A + p_B + p_C)) \cdot \\
&\quad \cdot (6\theta + (1-\theta)(p_A + p_B + p_C)) \cdot (7\theta + (1-\theta)(p_A + p_B + p_C)) \\
&\quad - (2\theta + (1-\theta)(p_A + p_B)) \cdot (3\theta + (1-\theta)(p_A + p_B)) \\
&\quad \cdot (4\theta + (1-\theta)(p_A + p_B))(5\theta + (1-\theta)(p_A + p_B))\big)
\end{aligned}$$

The value of the likelihood ratio for $p_A = 0.1$, $p_B = p_C = 0.2$ and $\theta = 0.05$ would be 2.95, compared to 4.6 in the independence case, which again shows weakening of the evidence due to uncertainty. △

Example 4.12 (Continuation of Example 3.8). We have $E = ABCD$, $V = AA$, $S = AB$ and we are examining two sets of hypotheses:

- H_1*: the contributors to the evidence are the victim, the suspect and one unknown person* and H_2*: the contributors to the evidence are the victim and two unknown people.*

Then $\mathcal{K} = (A, A, A, B)$ and

$$\begin{aligned}
\mathsf{P}(E|H_1) &= \mathsf{P}_1(CD, ABCD, AAAB) \\
&= \frac{1}{1^{(2)}(4)} \cdot \big((T_0)^{(2)}(t_0) - \sum_{i\in\{C,D\}} (T_{1i})^{(2)}(t_{1i}) \\
&\quad + \sum_{i,j\in\{C,D\}} (T_{2ij})^{(2)}(t_{2ij})\big) \\
&= \frac{1}{1^{(2)}(4)} \cdot [(p_A + p_B + p_C + p_D)^{(2)}(k_A + k_B + k_C + k_D) \\
&\quad - (p_A + p_B + p_D)^{(2)}(k_A + k_B + k_D) \\
&\quad - (p_A + p_B + p_C)^{(2)}(k_A + k_B + k_C) \\
&\quad + (p_A + p_B)^{(2)}(k_A + k_B)] \\
&= \frac{1}{1^{(2)}(4)} \cdot [(p_A + p_B + p_C + p_D)^{(2)}(4) \\
&\quad - (p_A + p_B + p_D)^{(2)}(4) - (p_A + p_B + p_C)^{(2)}(4) \\
&\quad + (p_A + p_B)^{(2)}(4)],
\end{aligned}$$

$$\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_2(BCD, ABCD, AAAB) \\
&= \frac{1}{1^{(4)}(4)} \cdot \big((T_0)^{(4)}(t_0) - \sum_{i\in\{B,C,D\}} (T_{1i})^{(4)}(t_{1i}) \\
&\quad + \sum_{i,j\in\{B,C,D\}} (T_{2ij})^{(4)}(t_{2ij}) - \sum_{i,j,k\in\{B,C,D\}} (T_{3ijk})^{(4)}(t_{3ijk})\big) \\
&= \frac{1}{1^{(4)}(4)} \cdot [(p_A + p_B + p_C + p_D)^{(4)}(k_A + k_B + k_C + k_D) \\
&\quad - (p_A + p_C + p_D)^{(4)}(k_A + k_C + k_D) \\
&\quad - (p_A + p_B + p_D)^{(4)}(k_A + k_B + k_D) \\
&\quad - (p_A + p_B + p_C)^{(4)}(k_A + k_B + k_C) \\
&\quad + (p_A + p_D)^{(4)}(k_A + k_D) + (p_A + p_C)^{(4)}(k_A + k_C) \\
&\quad + (p_A + p_B)^{(4)}(k_A + k_B) - p_A^{(4)}(k_A)] \\
&= \frac{1}{1^{(4)}(4)} \cdot [(p_A + p_B + p_C + p_D)^{(4)}(4) \\
&\quad - (p_A + p_C + p_D)^{(4)}(3) - (p_A + p_B + p_D)^{(4)}(4)
\end{aligned}$$

$$
\begin{aligned}
&- (p_A + p_B + p_C)^{(4)}(4) + (p_A + p_D)^{(4)}(3) \\
&+ (p_A + p_C)^{(4)}(3) + (p_A + p_B)^{(4)}(4) - p_A^{(4)}(3)].
\end{aligned}
$$

- *H_1: the contributors to the evidence are the victim, the suspect and two unknown people* and *H_2: the contributors to the evidence are the victim and three unknown people.*

  Under both hypotheses, U, E, $\mathcal{K}$ and y remain the same as in the previous case, the only change occurs in the number of unknown contributors x. Thus

$$
\begin{aligned}
\mathsf{P}(E|H_1) &= \mathsf{P}_2(CD, ABCD, AAAB) \\
&= \frac{1}{1^{(4)}(4)} \cdot \big((T_0)^{(4)}(t_0) - \sum_{i \in \{C,D\}} (T_{1i})^{(4)}(t_{1i}) \\
&\quad + \sum_{i,j \in \{C,D\}} (T_{2ij})^{(4)}(t_{2ij})\big)
\end{aligned}
$$

  and

$$
\begin{aligned}
\mathsf{P}(E|H_2) &= \mathsf{P}_3(BCD, ABCD, AAAB) \\
&= \frac{1}{1^{(6)}(4)} \cdot \big((T_0)^{(6)}(t_0) - \sum_{i \in \{B,C,D\}} (T_{1i})^{(6)}(t_{1i}) \\
&\quad + \sum_{i,j \in \{B,C,D\}} (T_{2ij})^{(6)}(t_{2ij}) - \sum_{i,j,k \in \{B,C,D\}} (T_{3ijk})^{(6)}(t_{3ijk})\big).
\end{aligned}
$$

  We omit further details for brevity.

△

Bibliography

[1] D. J. Balding and R. A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64:125–140, 1994.

[2] D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Forensic Science International*, 96:3–12, 1995.

[3] W. K. Fung and Y. Q. Hu. Interpreting forensic DNA mixtures: allowing for uncertainty in population substructure and dependence. *Journal of the Royal Statistical Society: Series A*, 163:241–254, 2000.

[4] W. K. Fung and Y. Q. Hu. *Statistical DNA Forensics: Theory, Methods and Computation*. John Wiley & Sons, Ltd., 2008.

[5] A. J. Jeffreys, S. L. Thein, and V. Wilson. Individual-specific 'fingerprints' of human DNA. *Nature*, 316:76–79, 1985.

[6] D. Slovák. *Statistické metody stanovení váhy evidence v procesu identifikace jedince*. MFF UK, Praha, 2009.

[7] B. S. Weir, C. M. Triggs, L. Starling, L. I. Stowell, K. A. J. Walsh, and J. Buckleton. Interpreting DNA mixtures. *Journal of Forensic Sciences*, 42(2):213–222, 1997.

[8] S. Wright. The genetical structure of populations. *Ann. Eugen.*, 15:323–354, 1951.

[9] K. Zoubková. *Statistické metody ve forenzní genetice*. MFF UK, Praha, 2004.

Appendix

English - Czech dictionary of key terms

English	Czech	Czech explanation
allele	**alela**	konkrétní forma genu
ancestor	**předek**	
chromosome	**chromozom**	struktura buněčného jádra, nesoucí genetickou informaci
coancestry	**společný původ**	
contributor	**přispěvatel**	
crime	**zločin**	
crime scene	**místo činu**	
DNA (deoxyribonucleic acid)	**DNA; DNK**	deoxyribonukleová kyselina
equilibrium	**rovnováha**	
evidence	**důkaz**	
forensic	**forenzní, soudní**	
genetics	**genetika**	věda o dědičnosti a proměnlivosti
gene	**gen**	základní jednotka dědičné informace; úsek vlákna DNA
genotype	**genotyp**	soubor veškeré genetické informace organismu, resp. konkrétního znaku
heredity	**dědičnost**	

heterozygote	**heterozygot**	jedinec nesoucí na sledovaném lokusu dvě různé alely
homozygote	**homozygot**	jedinec nesoucí na sledovaném lokusu dvě stejné alely
IBD (identical by descent)	**společné původem**	alely pocházející ze společného předka
inherit	**dědit, zdědit**	
likelihood ratio	**věrohodnostní poměr**	
locus	**lokus**	umístění konkrétního genu na chromozomu
marker	**ukazatel, znak**	
mixture	**směs (DNA)**	
nucleus	**jádro (buněčné)**	
offspring	**potomek**	
pedigree	**rodokmen**	
perpetrator	**pachatel**	
sample	**vzorek (DNA)**	
suspect	**podezřelá osoba**	
variation	**proměnlivost**	
victim	**oběť**	
weight (of evidence)	**váha, síla (důkazu)**	