Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Dalibor Slovák

Statistické metody stanovení váhy evidence v procesu identifikace jedince

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

2009

Děkuji své rodině a přátelům za podporu ve studiu a své vedoucí Prof. RNDr. Janě Zvárové, DrSc. za zajímavé téma i odborný dohled.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 4.8.2009 Dalibor Slovák

Obsah

Název práce: Statistické metody stanovení váhy evidence v procesu identifikace jedince
Autor: Dalibor Slovák
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Jana Zvárová, DrSc.
e-mail vedoucího: zvarova@euromise.cz

Abstrakt: V předložené práci se věnujeme identifikaci pachatele a stanovení váhy evidence proti němu. Nejprve zformulujeme jednoduchý model zvaný ostrovní problém a odvodíme vzorec pro stanovení váhy evidence v jeho základním tvaru. V dalších kapitolách rozebíráme jednotlivé modifikace ostrovního problému a problémy s tím spojené. Hledáme, jak se vypořádat s neznalostí základních parametrů modelu jako například velikost populace. Zjišťujeme, jak zahrnout do modelu možnost různých chyb nebo vliv příbuznosti a subpopulační struktury. V závěru se podrobněji věnujeme také směsím DNA, včetně odvození a naprogramování příslušných vzorců.

Klíčová slova: identifikace, ostrovní problém, stanovení váhy evidence, beta-binomická formule

Title: Statistical methods for determination of the weight of evidence in the identification process
Author: Dalibor Slovák
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 an identification of culprit and assesment of evidence against him. At first we define a simple model called the island problem and we derive the weight-of-evidence formula in its basic form. In the next chapters we analyse several modifications of island problem and related issues. We find how we can deal with uncertainty about basic parametres of model, like size of population. We investigate possibility of inclusion of different errors or influence of relatedness and subpopulation structure into model. At the close we enlarge mixtures of DNA, including deriving and programming of appropriate formulas.

Keywords: identification, island problem, weight-of-evidence, beta-binomial sampling formula

Kapitola 1

1.1 Úvod

Matematika dnes promlouvá do mnoha oborů lidské činnosti, o kterých by si to před několika desítkami let stěží někdo pomyslel. To na jedné straně klade vyšší nároky na zájemce o příslušný obor, na straně druhé to dává matematikům větší možnost uplatnit své schopnosti v oblasti svých dalších zájmů. V neposlední řadě takováto interdisciplinární spolupráce vyžaduje také dobrou komunikaci mezi pracovníky z jednotlivých vědních disciplín.

Jedním z takovýchto oborů je i kriminalistika a forenzní (soudní) medicína. Klasických i nových výsledků z oblasti pravděpodobnosti a statistiky se používá jednak k samotné identifikaci pachatele zločinu (případně k ospravedlnění neprávem obviněné osoby), jednak k lepšímu chápání genetických dat a jejich zpracování.

V této práci podrobně rozebereme jednu z možností, jak stanovit pravděpodobnost viny podezřelého. Ukážeme, které okolnosti ovlivňují výslednou pravděpodobnost a zda jejich zanedbání svědčí ve prospěch či neprospěch podezřelého. Některé potřebné informace mohou být neznámé, příslušné statistické metody a postupy by však vydaly na samostatnou práci, a proto u většiny parametrů předpokládáme, že jsou alespoň přibližně známy.

Všechny odvozené vzorce jsou takového tvaru, že je lze snadno naprogramovat. U některých z nich jsme k tomu skutečně přistoupili a příslušné příkazy v programu R jsou obsaženy v příloze (kapitola 7).

Tato práce je spíše matematického zaměření, a předpokládá proto u čtenáře alespoň základní znalosti z teorie pravděpodobnosti. Na druhou stranu používá z pochopitelných důvodů i některé pojmy z genetiky, jež budou vysvětleny v následující části.

1.2 Pojmy z genetiky

Vzhledem k zaměření práce jsou pojmy vysvětlovány konkrétně u člověka a nemusí platit obecně u všech organismů. Pokud se čtenáři některé

pojmy budou zdáti vysvětleny nedostatečně, mohu doporučit k prostudování například [5].

Za zakladatele genetiky je považován (jistě k nemalé pýše českých vlastenců) opat brněnského kláštera Johann Gregor Mendel (1822 - 1884) díky své studii dědičnosti u hrachu. *Dědičností* rozumíme předávání genetických informací o základních znacích a vlastnostech druhu z rodičů na potomky. U člověka při pohlavním rozmnožování vzniká nový jedinec, který dědí polovinu genetické informace od otce a polovinu od matky.

Genetická informace je uložena v *chromozomech* jádra každé lidské buňky. Na jejich stavbě se význačně podílí *kyselina deoxyribonukleová (DNA)*, která je tvořena chemickými sloučeninami zvanými nukleotidy; právě pořadí nukleotidů má zásadní význam pro přenos genetické informace. V každé buňce je 23 párů chromozomů. Dva chromozomy, které spolu vytvářejí pár, se nazývají *homologické*.

Základní jednotkou genetické informace je *gen*. Místo, kde je na chromozomu gen uložen, se nazývá *lokus*. Z hlediska molekulární biologie je gen souvislý úsek makromolekuly DNA, obsahující informaci pro vznik určitého znaku, např. barvy očí nebo struktury nějakého enzymu. Jiná definice říká, že gen je vloha zodpovědná za vznik jednoho konkrétního nejmenšího samostatně vymezitelného rozdílu mezi dvěma jedinci v populaci. Z této formulace je zřejmé, že u každého genu musí existovat alespoň dvě varianty; často jich však existuje mnohem více. Tyto varianty se nazývají *alely*. Jestliže na lokusu vymizí všechny alely kromě jedné, říkáme, že došlo k *fixaci* příslušné alely.

V buňce je každý gen zastoupen dvěma alelami na homologických chromozomech. Jednu z alel zdědí potomek po otci, druhou po matce. Obě alely nesené rodičem mají stejnou pravděpodobnost, že budou předány potomku, proto se předávání alel rodičů do další generace řídí jednoduchými kombinatorickými pravidly.

Soubor všech genů jedince se nazývá *genotyp*, někdy je však tento pojem používán v užším smyslu jako označení jednoho alelového páru. Jelikož v této práci budeme obvykle pracovat na jednom lokusu, genotypem budeme rozumět povětšinou právě příslušnou dvojici alel. Jsou-li alely na homologických chromozomech totožné, jedinec se nazývá *homozygot* (pro daný gen), jedinec s rozdílnými alelami se nazývá *heterozygot* (pro daný gen). Jestliže jsou dvě alely kopií jedné konkrétní alely od společného předka, nazýváme je z anglického termínu *identical by descent* zkratkou *ibd*.

Soubor jedinců, kteří jsou spolu spjati potenciální možností pohlavního rozmnožování, nazýváme *populace*. Populace může být rozdělena do *subpopulací*, obvykle na základě rasy či geografického umístění. Pokud v populaci žádné subpopulace nejsou, nazýváme ji *homogenní*. Pro zkoumání vztahů uvnitř populace se zavádí mnohé zjednodušující podmínky. Jednou z nich je *Hardyova-Weinbergova rovnováha (Hardy-Weinberg*

equilibrium, HWE), která označuje stav, při kterém genotypy vznikají nezávislým kombinováním alel. To znamená, že pravděpodobnost výskytu homozygotního genotypu AA je p_A^2 a pravděpodobnost výskytu heterozygotního genotypu AB je $2p_Ap_B$, kde p_A a p_B značí pravděpodobnost, že alela náhodně vybraná z populace je typu A, resp. B.

Ke značení alel budeme obvykle používat - tak jako v předchozím odstavci - velká písmena z počátku abecedy, ale v některých případech, obzvláště pokud není předem jasný počet alel, využijeme značení $A_1, A_2, \ldots$ Co se týče matematického značení, ve většině případů budeme průnik jevů značit pouhou čárkou; výjimečně však pro větší přehlednost použijeme i standardní matematické značení $\cap$.

Kapitola 2

Ostrovní problém

2.1 Úvod do ostrovního problému

Uvažujme existenci jistého vzácného rysu - označme jej Υ. To, zda někdo má či nemá Υ, není vidět na první pohled. Přesvědčit se o tom lze specifickým testem, o němž předpokládáme, že je bezchybný.

Na nepřístupném ostrově se 101 obyvateli byl spáchán zločin. Na počátku nemáme žádné informace o pachateli, a tak každému z ostrovanů přidělíme stejnou (apriorní) pravděpodobnost spáchání zločinu. Je zjištěno, že pachatel je nositelem znaku Υ, a u podezřelého byl tento znak rovněž nalezen. Jak moc si můžeme být jisti, že námi nalezený podezřelý je skutečně pachatel?

Odpověď na tuto otázku závisí na pravděpodobnosti výskytu Υ. Předpokládejme, že podezřelý a pachatel (kteří mohou, ale nemusí být tatáž osoba) jsou jedinými osobami na ostrově, u nichž je znám stav Υ (tj. zda Υ mají či nemají). Na blízké pevnině byl v populaci velkého rozsahu proveden výzkum, jenž ukázal, že Υ má přibližně 1 % populace. Na základě tohoto výzkumu položíme pravděpodobnost výskytu znaku Υ u libovolného jedince v populaci rovnu $0,01$. Předpokládáme, že stejnou pravděpodobnost má Υ i na okolních ostrovech, tedy i na tom našem. Posledním předpokladem je, že pravděpodobnost, že kterýkoli ostrovan má znak Υ, není ovlivněna znalostí toho, že má tento znak podezřelý.

Zapišme nyní přehledně všechny předpoklady:

- Všech 101 ostrovanů je navzájem nepříbuzných.
- Stav Υ u libovolného ostrovana není nijak ovlivněn stavem Υ u ostatních ostrovanů (nezávislost výskytu Υ).
- Pro libovolného obyvatele ostrova je apriorní pravděpodobnost, že právě on je pachatelem, rovna $\frac{1}{101}$.
- Pachatel má Υ.

- Podezřelý má Υ.
- Stav Υ u ostatních ostrovanů je neznámý.
- Pravděpodobnost výskytu Υ na ostrově je $0,01$.

Řešení je následující:
Informaci o stavu Υ máme pouze u jedné osoby, a tou je podezřelý (pachatel je neznámou osobou). U zbylých 100 ostrovanů můžeme počet nositelů Υ pouze odhadnout. Jelikož počet osob se znakem Υ má binomické rozdělení s parametry $p = 0,01$ a $n = 100$, odhadneme jej střední hodnotou. Ta se vypočítá jako $n \cdot p = 100 \cdot 0,01 = 1$. Kromě podezřelého tedy očekáváme na ostrově ještě jednoho nositele Υ. S překvapením tak zjišťujeme, že ačkoli znak Υ je poměrně vzácný, pravděpodobnost, že jsme identifikovali správnou osobu a že náš podezřelý je skutečně pachatelem, je pouhých 50 %.

2.2 Potenciální selhání modelu

V předchozím povídání jsme ostrovní problém zatížili velkým množstvím předpokladů. Podívejme se, kde všude náš model může selhat:

- *Bezchybnost testu na znak* Υ
  Kromě toho, že test může v malém procentu dávat chybné výsledky, je možné uvažovat i chyby způsobené takzvaně "lidským faktorem": kontaminace či záměna vzorku, z nějž je stav Υ zjišťován, chybné vyhodnocení výsledku či dokonce záměrná dezinterpretace.

- *Nepřístupnost ostrova*
  Tímto požadavkem je myšlena uzavřenost vyšetřované populace vůči migraci. Selhat může například tehdy, pokud se pachateli podaří uniknout nepozorovaně z místa činu a odcestovat do takové vzdálenosti, aby nebyl zařazen do populace podezřelých osob. V jistých případech je však uzavřenost populace přímo dána situací - například při identifikaci osob po leteckém neštěstí je obvykle pevně dán seznam cestujících, a náš ostrov (tj. populace možných osob) je tedy skutečně "nepřístupný".

- *Počet obyvatel* N
  Velikost populace N je často pouze odhadnuta. Tento problém souvisí i s výše zmiňovanou nepřístupností ostrova. Pokud dochází u vyšetřované populace k migraci, je třeba při stanovení počtu obyvatel počítat s o to větší nejistotou.

- *Pravděpodobnost p výskytu znaku Υ v populaci*
  Rovněž hodnota p je obvykle neznámá, a proto se odhaduje na základě relativní četnosti výskytu Υ v podobné populaci, o níž máme více informací. Ovšem tato pomocná data mohou být již zastaralá nebo vystihují naši populaci jen zčásti.

- *Výběr podezřelého*
  Podezřelý obvykle není vybírán z populace náhodně, ale na základě dalších indicií, které zvyšují pravděpodobnost viny. Jinou možností je vybírání podezřelého na základě testování osob z populace na přítomnost znaku Υ. Tímto způsobem může dojít k vyloučení osob, u nichž znak Υ nebyl nalezen, a tím ke zmenšení velikosti populace podezřelých osob.

- *Příbuznost a příslušnost ke stejné subpopulaci*
  Pokud je podezřelý (nebo jiná testovaná osoba) nositelem Υ a zároveň jsou v populaci zahrnuti nějací jeho příbuzní, v případě profilu DNA se díky dědičnosti zvyšuje pravděpodobnost výskytu Υ. Nezvykle vysoká relativní četnost obvykle vzácného znaku se často vyskytuje i v rámci stejné subpopulace, ačkoli u jeho nositelů nemusí být identifikován žádný společný předek.

- *Stejná apriorní pravděpodobnost spáchání zločinu*
  Ačkoli tento požadavek intuitivně odpovídá všeobecné presumpci neviny, můžeme různým osobám přiřadit rozdílnou apriorní pravděpodobnost, kupříkladu na základě vzdálenosti od místa činu, časové dostupnosti nebo možnému alibi.

Později se na každý z těchto problémů zaměříme podrobněji.

2.3 Posuzování viny na základě evidence

Předpokládejme, že na místě činu byl nalezen materiál (většinou biologické povahy), u něhož vyšetřování naznačuje, že jej tam zanechal pachatel. Z tohoto materiálu můžeme získat profil pachatele. Profilem míníme kupříkladu barvu vlasů, velikost nohy (tento profil hrál podstatnou roli v pohádce o Popelce), otisky prstů nebo genetický vzorek - právě genetický profil budeme uvažovat nejčastěji. Profil získaný na místě činu a profil podezřelého spolu s dalšími okolnostmi zločinu (například místo a způsob provedení zločinu nebo výpovědi případných svědků) nazýváme souhrnně *evidence*. Při pevně dané evidenci E sestavme dvě komplementární hypotézy:

$$\begin{aligned} G &: \text{podezřelý je vinen (guilty)} \\ I &: \text{podezřelý je nevinen (innocent).} \end{aligned}$$

Nyní tyto pojmy převedeme do matematické symboliky; např. $\mathsf{P}(G|E)$ bude značit pravděpodobnost, že platí hypotéza G při dané evidenci E. Pak podle Bayesovy věty platí

$$\mathsf{P}(G|E) = \frac{\mathsf{P}(E|G)\mathsf{P}(G)}{\mathsf{P}(E|G)\mathsf{P}(G) + \mathsf{P}(E|I)\mathsf{P}(I)}. \tag{2.1}$$

Avšak výraz $\mathsf{P}(E|I)$ nelze spočítat přímo. Označme $\mathcal{I}$ populaci alternativních podezřelých, tj. všechny obyvatele ostrova vyjma podezřelého, a C_i jev, že pachatelem je osoba $i \in \mathcal{I}$. Podezřelý je nevinen právě tehdy, když existuje index $i \in \mathcal{I}$, že nastává jev C_i. Jev I je tedy ekvivalentní s jevem $\cup_{i\in\mathcal{I}}\, C_i$ a díky disjunktnosti jevů C_i platí

$$\mathsf{P}(I) = \mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \sum_{i\in\mathcal{I}} \mathsf{P}(C_i).$$

Odtud

$$\begin{aligned}
\mathsf{P}(I)\mathsf{P}(E|I) &= \mathsf{P}\left(\cup_{i\in\mathcal{I}}\, C_i\right) \mathsf{P}\left(E|\, \cup_{i\in\mathcal{I}}\, C_i\right) = \\
&= \mathsf{P}\left(\cup_{i\in\mathcal{I}}\, C_i\right) \frac{\mathsf{P}\left(E \cap \left(\cup_{i\in\mathcal{I}}\, C_i\right)\right)}{\mathsf{P}\left(\cup_{i\in\mathcal{I}}\, C_i\right)} = \\
&= \mathsf{P}\left(\cup_{i\in\mathcal{I}}\, (E \cap C_i)\right) = \sum_{i\in\mathcal{I}} \mathsf{P}\left(E \cap C_i\right) = \\
&= \sum_{i\in\mathcal{I}} \mathsf{P}(C_i)\mathsf{P}(E|C_i).
\end{aligned}$$

Předpokládejme, že máme dánu počáteční evidenci E_0, kterou tvoří například základní informace o místě a způsobu zločinu, a chceme zahrnout nově získanou evidenci E. Definujme **věrohodnostní poměr**

$$R_i\left(E|E_0\right) = \frac{\mathsf{P}(E|C_i, E_0)}{\mathsf{P}(E|G, E_0)}, \tag{2.2}$$

jenž vyjadřuje, kolikrát je při znalosti E_0 pravděpodobnost vzniku evidence E větší za podmínky, že pachatelem je osoba i, než za podmínky, že pachatelem je podezřelý.
Dále definujme **věrohodnostní váhy**

$$w_i\left(E_0\right) = \frac{\mathsf{P}(C_i|E_0)}{\mathsf{P}(G|E_0)};$$

celý vzorec (2.1) se potom dá přepsat ve tvaru

$$\mathsf{P}(G|E, E_0) = \frac{1}{1 + \sum_{i\in\mathcal{I}} w_i\left(E_0\right) R_i\left(E|E_0\right)}. \tag{2.3}$$

Vzorec (2.3) se obvykle nazývá **vzorec pro stanovení váhy evidence (the weight-of-evidence formula)**. Pokud budou jasně určeny evidence E a E_0, budeme místo $R_i(E|E_0)$ a $w_i(E_0)$ psát pouze R_i, w_i a vzorec (2.3) používat ve tvaru

$$\mathsf{P}(G|E) = \frac{1}{1 + \sum_{i \in \mathcal{I}} w_i R_i}. \tag{2.4}$$

V následujícím odstavci si ukážeme, jak vypadá vzorec (2.4) v případě ostrovního problému z počátku této kapitoly.

2.4 Stanovení váhy evidence v ostrovním problému

Nyní budeme aplikovat vzorec (2.4) na ostrovní problém ze sekce 2.1, ale nejprve jej zformulujeme o něco obecněji:

- Na ostrově žije $N + 1$ nepříbuzných osob.
- Všechny osoby mají stejnou apriorní pravděpodobnost viny.
- Pachatel má znak Υ.
- Podezřelý má znak Υ.
- Stav Υ u ostatních ostrovanů je neznámý.
- Pravděpodobnost, že osoba na ostrově má znak Υ, je rovna p.

Jde o speciální případ použití (2.4). Evidencí E je v tomto případě myšlena přítomnost znaku Υ u pachatele a podezřelého. Počáteční evidence E_0 je pro všechny stejná, a proto položíme $w_i = 1 \ \forall i$. Nyní se podívejme na hodnotu R_i. Jestliže je pachatelem podezřelý, je pravděpodobnost vzniku evidence E rovna 1. Pokud je pachatelem osoba $i \in \mathcal{I}$, pravděpodobnost vzniku evidence E je rovna p. Podle vzorce (2.2)

$$R_i\,(E|E_0) = \frac{\mathsf{P}(E|C_i, E_0)}{\mathsf{P}(E|G, E_0)} = \frac{p}{1} = p.$$

Velikost populace alternativních podezřelých $\mathcal{I}$ je rovna N, proto ze vzorce (2.4) plyne

$$\mathsf{P}(G|E) = \frac{1}{1 + N \cdot p}\ . \tag{2.5}$$

Dosadíme-li $N = 100$ a $p = 0,01$, pak

$$\mathsf{P}(G|E) = \frac{1}{1 + 100 \cdot 0,01} = \frac{1}{2}, \text{ tedy } 50\ \%,$$

což odpovídá původnímu výsledku.

Kapitola 3

Obecnější modely ostrovního problému

Jak se změní vzorec (2.4), potažmo (2.5), porušíme-li některý z předpokladů, jak to bylo naznačeno v sekci 2.2? Pro větší názornost porušíme vždy jen jeden předpoklad.

3.1 Další evidence

U ostrovního problému, jak jsme jej popsali v kapitole 2.4, jsme předpokládali, že není k dispozici žádná jiná evidence než informace o stavu Υ. Tato situace v praxi takřka nikdy nenastává. I kdyby neexistovaly kromě Υ-evidence žádné další stopy vedoucí k pachateli, vždy máme k dispozici základní informace o povaze zločinu - místo, čas apod., díky nimž můžeme některé osoby považovat za pravděpodobnější pachatele než jiné. Ukážeme si teď, jak postupovat v případě, že evidence se skládá z více komponent (podobnou situaci zmiňujeme i v kapitole 6). Věrohodnostní poměry a váhy budeme opět chvíli psát v plném tvaru.

Nezávislost jednotlivých složek evidence

Nechť se evidence E skládá ze dvou položek, E_1 a E_2. Celkový věrohodnostní poměr získáme dvěma ekvivalentními způsoby, které odpovídají dvěma různým pořadím jednotlivých složek evidence:

$$\begin{aligned} R_i(E|E_0) &= R_i(E_1,E_2|E_0) = \frac{\mathsf{P}\left(E_1,E_2|C_i,E_0\right)}{\mathsf{P}\left(E_1,E_2|G,E_0\right)} = \\ &= \frac{\mathsf{P}\left(E_2|C_i,E_1,E_0\right)}{\mathsf{P}\left(E_2|G,E_1,E_0\right)}\frac{\mathsf{P}\left(E_1|C_i,E_0\right)}{\mathsf{P}\left(E_1|G,E_0\right)} = \\ &= R_i(E_2|E_1,E_0)R_i(E_1|E_0) \end{aligned}$$

a stejně lze získat

$$R_i(E|E_0) = R_i(E_1, E_2|E_0) = R_i(E_1|E_2, E_0)R_i(E_2|E_0). \tag{3.1}$$

Odtud vidíme, že nezáleží na pořadí, v jakém jsou jednotlivé položky zařazovány do evidence. Kromě toho, pokud jsou jednotlivé složky evidence navzájem nezávislé, můžeme celkový věrohodnostní poměr vyjádřit jako součin marginálních věrohodnostních poměrů:

$$R_i(E_1, E_2|E_0) = R_i(E_1|E_0)R_i(E_2|E_0). \tag{3.2}$$

Příklad

Předpokládejme, že jsou před soudem prezentovány dvě svědecké výpovědi, které označíme E_1 a E_2. Je určeno, že na základě každé z nich je pravděpodobnost spáchání zločinu podezřelým desetkrát větší než pravděpodobnost spáchání zločinu osobou i, tj. $R_i(E_1|E_0) = R_i(E_2|E_0) = 0,1$. Jestliže soudce předpokládá, že tyto výpovědi jsou na sobě nezávislé, podle vzorce (3.2) je

$$R_i(E_1, E_2|E_0) = R_i(E_1|E_0)R_i(E_2|E_0) = 0,1 \cdot 0,1 = 0,01.$$

Jindy naopak soudce může předpokládat vysokou korelovanost obou výpovědí, například pokud se jedná o dva kamarády, kteří zločin viděli společně, a je možné, že při následné diskuzi sladili všechny odlišnosti. V tom případě druhou výpovědí nezískáváme žádnou novou informaci, což vyjádříme jako $R_i(E_2|E_1, E_0) = 1$, a podle vzorce (3.1) je

$$R_i(E_1, E_2|E_0) = R_i(E_2|E_1, E_0)R_i(E_1|E_0) = 1 \cdot 0,1 = 0,1.$$

To potvrzuje naši intuici, že dvě nezávislé položky evidence mají společně větší výpovědní sílu než dvě položky, které jsou závislé, v krajním případě až do té míry, že pouze opakují stejnou informaci.

Postupné přidávání položek evidence

V následujícím výpočtu si trochu procvičíme podmiňování:

$$\begin{aligned}
R_i\left(E_1, E_2|E_0\right) w_i\left(E_0\right) &= \frac{\mathsf{P}\left(E_1, E_2|C_i, E_0\right)}{\mathsf{P}\left(E_1, E_2|G, E_0\right)} \frac{\mathsf{P}(C_i|E_0)}{\mathsf{P}(G|E_0)} = \\
&= \frac{\mathsf{P}\left(E_1, E_2, C_i|E_0\right)}{\mathsf{P}\left(E_1, E_2, G|E_0\right)} = \\
&= \frac{\mathsf{P}\left(E_1, C_i|E_2, E_0\right)}{\mathsf{P}\left(E_1, G|E_2, E_0\right)} \frac{\mathsf{P}(E_2|E_0)}{\mathsf{P}(E_2|E_0)} = \\
&= \frac{\mathsf{P}\left(E_1|C_i, E_2, E_0\right)}{\mathsf{P}\left(E_1|G, E_2, E_0\right)} \frac{\mathsf{P}(C_i|E_2, E_0)}{\mathsf{P}(G|E_2, E_0)} = \\
&= R_i\left(E_1|E_2, E_0\right) w_i\left(E_2, E_0\right).
\end{aligned}$$

Celkově tedy platí

$$R_i\left(E_1, E_2|E_0\right) w_i\left(E_0\right) = R_i\left(E_1|E_2, E_0\right) w_i\left(E_2, E_0\right). \tag{3.3}$$

(3.3) nám říká, že vzorec pro stanovení váhy evidence dává stejný výsledek, ať už aplikujeme složky E_1, E_2 najednou nebo postupně. V celé této kapitole budeme předpokládat, že veškerá evidence je aplikována najednou, ačkoli v praxi tomu bývá často jinak, neboť jednotlivé položky se aplikují na vzorec (2.4) ve chvíli, kdy jsou získány.

Shrňme si výsledky sekce 3.1 na příkladě.

Příklad

Nechť máme pevně zvoleného podezřelého a populaci alternativních podezřelých tvoří 1000 osob. Protože apriorní evidence E_0 nám nedává žádné bližší informace, stanovíme $w_i = 1$ pro všechna i. Evidence E se skládá z DNA profilu a jednoho svědectví, označme tyto položky jako E_{DNA} a E_{sved}. DNA profily z místa činu a od podezřelého jsou shodné, a tak je stanoven věrohodnostní poměr odpovídající DNA-evidenci jako $R_i^{DNA} = 10^{-6}\ \forall\ i$. Předpokládáme, že svědectví na DNA-evidenci nijak nezávisí, tudíž můžeme podle vzorce (3.2) věrohodnostní poměr vyjádřit v součinovém tvaru. Uvažujme dvě možné situace:

- Svědecká výpověď hovoří v neprospěch podezřelého, proto je stanoven $R_i^{sved} = 1/100$ pro všechna i. Aplikací na vzorec (2.3) dostaneme

$$\begin{aligned} \mathsf{P}(G|E_{DNA}, E_{sved}, E_0) &= \frac{1}{1 + \sum_{i=1}^{1000} w_i R_i^{DNA} R_i^{sved}} = \\ &= \frac{1}{1 + 1000 \cdot 10^{-6} \cdot 10^{-2}} \approx 0,99999 \end{aligned}$$

- Svědecká výpověď hovoří ve prospěch podezřelého, $R_i^{sved} = 100$ pro všechna i a pro aposteriorní pravděpodobnost viny platí:

$$\mathsf{P}(G|E_{DNA}, E_{sved}, E_0) = \frac{1}{1 + 1000 \cdot 10^{-6} \cdot 100} \approx 0,91$$

Z toho vidíme, že ačkoli je svědecká evidence mnohem slabší než DNA-evidence, její efekt na celkový výsledek je poměrně velký.

3.2 Výběr podezřelého

Pojďme se nyní podívat na způsob, jakým jsme našli podezřelého. Velmi často používanou metodou je vyhledávání v databázích DNA. O zřízení rozsáhlých národních databází se začalo uvažovat v 90. letech dvacátého století. První národní databáze vznikla v Anglii a Walesu v roce 1995, v ČR došlo k vytvoření národní databáze roku 2002.

V současnosti nejrozšířenějším databázovým systémem je systém CODIS (Combined DNA Analysis System), který vyvinula americká FBI speciálně pro kriminalistické účely a policejním sborům cizích států jej na základě vzájemných dohod poskytuje bezplatně. CODIS rozděluje získané genetické profily do dvou dílčích databází. Ve *forenzní databázi* se ukládají biologické vzorky získané na místě činu, v *databázi odsouzených* potom figurují genetické profily osob v minulosti odsouzených (v poslední době ale dochází stále častěji k tomu, že jsou do této databáze ukládány všechny genetické vzorky získané v průběhu vyšetřování, tedy i od osob, jež byly později osvobozeny). Tyto dvě databáze jsou pak vzájemně porovnávány a případné shody profilů jsou prověřeny kvalifikovanými odborníky. Podle použité metody se liší počet použitých úseků DNA - u nás se využívá patnácti úseků, zatímco například systém CODIS pracuje pouze s třinácti. Podle údajů z USA ze srpna roku 2006 obsahovala v té době forenzní databáze přibližně 150 000 profilů a databáze odsouzených přes $3,5$ milionu profilů (viz [4]). Národní databázi Velké Británie v současnosti tvoří přes čtyři miliony profilů a měsíčně přibývá 40 až 50 tisíc nových. O úspěšnosti tohoto přístupu svědčí i to, že po vytvoření databáze DNA vzrostl počet objasněných trestných činů ve Velké Británii z 24 % na 43 %, díky čemuž se systém databází těší podpoře veřejnosti. Na druhou stranu se z DNA dají získat velice citlivé osobní údaje, a proto je potřeba zajistit důkladnou ochranu databází proti zneužití.

Dalším z možných přístupů je sekvenční vyhledávání. Jednotliví příslušníci populace možných podezřelých jsou postupně náhodně vybíráni a testováni na přítomnost znaku Υ. V momentě, kdy je u prvního z nich znak Υ nalezen, je tato osoba označena jako podezřelý a testování je přerušeno. Předpokládejme, že před podezřelým bylo testováno k osob. U nich znak Υ nalezen nebyl, a tedy mohou být vyřazeny z populace podezřelých osob. Velikost populace se tím ovšem zmenšila a (2.5) je tvaru

$$\mathsf{P}(G|E) = \frac{1}{1 + (N-k)p} \ .$$

Je snadno nahlédnutelné, že takováto modifikace hovoří v neprospěch podezřelého. V mezní situaci, kdy by byl pachatel jediným nositelem Υ na ostrově a my bychom jej testovali jako posledního v pořadí, vyřadili bychom předtím všech N osob testovaných negativně a pachatele bychom mohli označit s jistotou, neboť

$$\mathsf{P}(G|E) = \frac{1}{1 + (N-N)p} = 1.$$

Pro ilustraci, jestliže $N = 100$, $p = 0,01$ a podezřelý je nalezen jako jedenáctá osoba v pořadí (tedy $k = 10$), $\mathsf{P}(G|E) = 0,5263158$, tedy $52,63$ %. Pravděpodobnost viny tak vzrostla o necelá tři procenta.

Jindy může být podezřelý zajištěn kvůli podezřelému chování, chybějícímu alibi apod. Z okolností zločinu, například místa, času nebo způsobu provedení, jsou však některé osoby podezřelejší než jiné. Zatímco dosud jsme pokládali $w_i = 1$ pro každé i, nyní na základě indicií přiřadíme různým osobám různé váhy w_i. Jednou z možností je těm, kdož bydlí místu zločinu blíže než podezřelý, přiřadit hodnotu $w_i > 1$ a těm, kdo bydlí dál, přiřadit $w_i < 1$. Jiným důvodem, proč přidělovat osobám různé váhy, může být svědecká výpověď, hovořící například o přibližném vzhledu či stáří pachatele nebo o jeho dopravním prostředku. Ve všech těchto případech je

$$\mathsf{P}(G|E) = \frac{1}{1 + p\sum_{i=1}^{N} w_i}.$$

Připomeňme, že hodnoty w_i neudávají celkovou (absolutní) pravděpodobnost, že pachatelem je osoba i, ale pravděpodobnost relativně vzhledem k podezřelému.

3.3 Nejistota ohledně N

Nejistota ohledně velikosti populace možných alternativních podezřelých působí na apriorní pravděpodobnost $\mathsf{P}(G)$. Nechť velikost populace $\tilde{N}$ je náhodná veličina se střední hodnotou N. Apriorní pravděpodobnost viny podmíněně při hodnotě $\tilde{N}$ je

$$\mathsf{P}(G|\tilde{N}) = 1/(\tilde{N} + 1),$$

ale protože $\tilde{N}$ neznáme, použijeme střední hodnotu:

$$\mathsf{P}(G) = \mathbb{E}\left[G|\tilde{N}\right] = \mathbb{E}\left[\frac{1}{\tilde{N} + 1}\right]$$

Funkce $1/(\tilde{N} + 1)$ není symetrická, ale je alespoň na intervalu $(0, \infty)$ konvexní. Podle Jensenovy nerovnosti pro konvexní funkci f platí

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x]).$$

Z $\mathbb{E}[\tilde{N}] = N$ plyne

$$\mathsf{P}(G) = \mathbb{E}\left[\frac{1}{\tilde{N} + 1}\right] \geq \frac{1}{N + 1}\ .$$

Opomenutí nejistoty ohledně hodnoty N tedy působí ve prospěch obžalovaného. Navíc je tento efekt obvykle velice malý; pojďme si to ukázat na konkrétních případech.

Malá nejistota ve velikosti populace

Položme pro $\varepsilon \in (0; 0,5)$

$$\tilde{N} = \begin{cases} N-1 & \text{s pravděpodobností } \varepsilon \\ N & \text{s pravděpodobností } 1-2\varepsilon \\ N+1 & \text{s pravděpodobností } \varepsilon \end{cases}$$

Potom

$$\begin{aligned} \mathsf{P}(G) &= \mathsf{E}\left[\frac{1}{\tilde{N}+1}\right] = \frac{\varepsilon}{N} + \frac{1-2\varepsilon}{N+1} + \frac{\varepsilon}{N+2} = \\ &= \frac{1}{N+1} + \frac{2\varepsilon}{N(N+1)(N+2)} \geq \frac{1}{N+1} \end{aligned}$$

a položíme-li $\varepsilon = 0,25$ a $N = 100$, potom $\mathsf{P}(G)$ je větší než $1/(N+1)$ o pouhých $0,000000485$.

Podívejme se, co způsobí nejistota ve velikosti populace ve vzorci (2.5):

$$\begin{aligned} \mathsf{P}(G|E) &= \frac{1}{1+\sum_i R_i \frac{\mathsf{P}(C_i)}{\mathsf{P}(G)}} = \frac{1}{1+p\frac{1}{\mathsf{P}(G)}\underbrace{\sum_i \mathsf{P}(C_i)}_{=1-\mathsf{P}(G)}} = \\ &= \frac{1}{1+p\frac{N(N+1)(N+2)}{N^2+2N+2\varepsilon}\left(1-\frac{N^2+2N+2\varepsilon}{N(N+1)(N+2)}\right)} = \\ &= \frac{1}{1+Np\frac{N^3+2N^2-2\varepsilon}{N^3+2N^2+2N\varepsilon}} = \frac{1}{1+Np\left(1-2\varepsilon\frac{N+1}{N^3+2N^2+2N\varepsilon}\right)}. \end{aligned}$$

Dosadíme-li opět $\varepsilon = 0,25$ a $N = 100$, vychází $\mathsf{P}(G|E) = 0,5000124$, což se i přes vysokou hodnotu ε liší od původního výsledku 50 %, při jehož výpočtu jsme hodnotu N brali jako pevnou, v řádu pouhé jedné tisíciny procenta. Pokud budeme chtít přesto počítat s nejistotou ohledně N, lze jako velice dobrou aproximaci brát

$$\mathsf{P}(G|E) \approx \frac{1}{1+Np\left(1-2\varepsilon/N^2\right)};$$

v našem příkladě dává tato aproximace výsledek $\mathsf{P}(G|E) = 0,5000125$, tedy $50,00125$ %.

Balding v [1] používá řádově horší aproximaci

$$\mathsf{P}(G|E) \approx \frac{1}{1+Np\left(1-4\varepsilon/N^3\right)},$$

která dává v našem příkladě hodnotu $\mathsf{P}(G|E) = 0,5000003$, to znamená $50,00003$ %.

Větší nejistota ve velikosti populace

Zkusme nyní předpokládat o něco větší nejistotu. Pro $\varepsilon \in (0; 1/3)$ položme

$$\tilde{N} = \begin{cases} N-2 & \text{s pravděpodobností } \varepsilon/2 \\ N-1 & \text{s pravděpodobností } \varepsilon \\ N & \text{s pravděpodobností } 1-3\varepsilon \\ N+1 & \text{s pravděpodobností } \varepsilon \\ N+2 & \text{s pravděpodobností } \varepsilon/2 \end{cases}$$

Potom

$$\begin{aligned} \mathsf{P}(G) &= \mathsf{E}\left[\frac{1}{1+\tilde{N}}\right] = \frac{\varepsilon}{2(N-1)} + \frac{\varepsilon}{N} + \frac{1-3\varepsilon}{N+1} + \frac{\varepsilon}{N+2} + \frac{\varepsilon}{2(N+3)} \\ &= \frac{6\varepsilon(N^2+2N-1) + N(N-1)(N+2)(N+3)}{N(N-1)(N+1)(N+2)(N+3)} \end{aligned}$$

a následně

$$\mathsf{P}(G|E) = \frac{1}{1 + Np(1 - 6\varepsilon\frac{(N+1)(N^2+2N-1)}{N^2(N-1)(N+2)(N+3)+6N\varepsilon(N^2+2N-1)})}.$$

Vhodnou aproximací je výraz

$$\mathsf{P}(G|E) \approx \frac{1}{1 + Np\,(1 - 6\varepsilon/N^2)}.$$

Pokud dosadíme $\varepsilon = 0,2$ a $N = 100$, vychází přesný výsledek

$$\mathsf{P}(G|E) = 0,5000297, \text{ tj. } 50,00297\ \%$$

a aproximace

$$\mathsf{P}(G|E) = 0,50003, \text{ tj. } 50,003\ \%.$$

Obecný případ

Předpokládejme, že se nyní velikost populace může lišit od střední hodnoty N až o číslo k, přičemž pravděpodobnost takovéto konkrétní hodnoty velikosti populace klesá s každým dalším krokem o polovinu:

$$\tilde{N} = \begin{cases} N-k & \text{s pravděpodobností } \frac{\varepsilon}{2^{k-1}} \\ \vdots & \\ N-1 & \text{s pravděpodobností } \varepsilon \\ N & \text{s pravděpodobností } 1 - \frac{2^k-1}{2^{k-2}}\varepsilon \\ N+1 & \text{s pravděpodobností } \varepsilon \\ \vdots & \\ N+k & \text{s pravděpodobností } \frac{\varepsilon}{2^{k-1}} \end{cases}$$

pro $\varepsilon \in \left(0; \frac{2^{k-2}}{2^k-1}\right)$. Horní hranice tohoto intervalu klesá k 1/4. Vhodnou aproximací je výraz

$$\mathsf{P}(G|E) \approx \frac{1}{1 + Np\left(1 - \frac{\varepsilon}{N^2}\sum_{i=1}^{k} i^2 2^{2-i}\right)}.$$

Jak jsme viděli v předchozím, koeficient u $\frac{\varepsilon}{N^2}$ je pro $k = 1$ roven dvěma a pro $k = 2$ roven šesti. Pro rostoucí k se velice rychle blíží limitní hranici 24; už od $k = 10$ výše lze tento koeficient nahradit přímo touto limitou.

Jako optimální ε jsme volili takovou hodnotu, aby i hodnota pravděpodobnosti u N byla dvojnásobná oproti hodnotě pravděpodobnosti u $N - 1$ a $N + 1$. V tomto obecném případě získáme optimální hodnotu ze vzorce $\frac{2^{k-2}}{3*2^{k-1}-1}$. Snadno se lze přesvědčit, že pro $k = 1$ je tato hodnota 1/4 a pro $k = 2$ je rovna 1/5. S rostoucím k tato hodnota velice rychle klesá k 1/6.

Naprogramované vzorce z této sekce jsou v kapitole 7.

3.4 Příbuznost a příslušnost k subpopulaci

Jak si lze snadno uvědomit, všichni lidé jsou navzájem příbuzní. Počet našich předků totiž exponenciálně roste. Každý máme dva rodiče, čtyři prarodiče, osm praprarodičů... O dvacet generací zpátky je počet našich předků přibližně milion. S každým člověkem máme proto takřka jistě společného nějakého předka. V reálném uvažování ovšem takováto příbuznost nemá žádný význam. V genetice se obvykle jako nejvzdálenější příbuzný bere bratranec (resp. sestřenice). K vyjádření míry příbuznosti mezi dvěma osobami se používá tzv. *coancestry koeficient*, který udává pravděpodobnost, že pokud na pevně zvoleném lokusu vybereme jednu alelu od každé z osob, tyto alely jsou shodné a pochází od společného předka (tedy jedná se o ibd alely). Například u sourozenců je tento koeficient roven 1/4, pro bratrance 1/16; samozřejmě pouze tehdy, jestliže jejich rodiče (resp. u bratranců prarodiče) nejsou v příbuzenském vztahu. Pak by byl tento koeficient ještě vyšší.

Z toho plyne, že pokud v populaci alternativních podezřelých jsou i příbuzní podezřelého, nalezení znaku Υ zvyšuje pravděpodobnost dalšího výskytu tohoto znaku. Z důvodu společného vývoje platí totéž i pro příslušníky stejné subpopulace, ke které patří podezřelý. Populace podezřelých osob může být velmi rozsáhlá, a tak se jeví velmi nepraktické počítat věrohodnostní poměr pro každou osobu zvlášť. Obvykle se stanoví několik skupin přibližně stejné míry příbuznosti, např. následující:

1. jednovaječné dvojče (to je jediný případ, kdy je DNA dvou osob stejné[1])
2. sourozenci (včetně dvojvaječných dvojčat)
3. rodiče a děti podezřelého
4. příbuzní druhého stupně, např. strýcové, synovci nebo nevlastní sourozenci
5. příbuzní třetího stupně, např. bratranci a sestřenice
6. lidé, kteří nejsou příbuznými podezřelého, ale náleží ke stejné subpopulaci
7. lidé, kteří nejsou příbuznými podezřelého a náleží k jiné subpopulaci

Jednotlivé skupiny se od sebe liší hodnotou věrohodnostního poměru. Platí, že čím větší je číslo skupiny, do níž osoba i patří, tím je hodnota R_i menší. Pokud váhy w_i ponecháváme rovny jedné, lze (2.4) psát ve tvaru

$$\mathsf{P}(G|E) = \frac{1}{1 + \sum_{i=1}^{N} R_i}. \tag{3.4}$$

Pokud bychom nahradili všechna R_i hodnotou p, dostaneme vzorec (2.5). Je však třeba zdůraznit, že tato záměna poškozuje podezřelého.

Jak již bylo zmíněno v kapitole 3.2, někdy je potřeba stanovit nestejné váhy. Stačí-li nám znát dolní hranici (a to nám obvykle stačí, neboť nahrazení aposteriorní pravděpodobnosti viny její dolní hranicí mluví ve prospěch podezřelého), (2.4) můžeme modifikovat jako

$$\mathsf{P}(G|E) \geq \frac{1}{1 + R_A \sum_{i \in A} w_i + \ldots + R_G \sum_{i \in G} w_i}, \tag{3.5}$$

kde $A, \ldots, G$ je rozdělení populace do skupin (počet skupin samozřejmě může být libovolný jiný) a $R_A, \ldots, R_G$ jsou maximální hodnoty věrohodnostního poměru v jednotlivých skupinách. Pokud hodnoty w_i stanovíme uvnitř každé skupiny stejně, lze vzorec (3.5) dále zjednodušit:

$$\mathsf{P}(G|E) \geq \frac{1}{1 + N_A R_A w_A + \ldots + N_G R_G w_G},$$

kde $N_A, \ldots, N_G$ je počet osob v příslušné skupině. V případě, že neznáme přesný počet osob v jednotlivých skupinách, lze na každou z nich použít

[1] Jednovaječná dvojčata mají DNA shodnou při narození, s přibývajícím věkem a životem v rozdílných podmínkách dochází k drobným odchylkám. Při analýze DNA však lze odhalit, že k odchylce došlo, a proto pro naše účely můžeme předpokládat jejich shodnost.

přístup ze sekce 3.3. Často jsou věrohodnostní poměry v každé skupině takřka stejné, dolní hranice je pak velice blízko přesné hodnotě.

Ignorování příbuzných a příslušníků stejného etnika poškozuje podezřelého, často velice výrazně. Uvažujme znovu populaci $N = 100$ osob, položme $p = 0,01$ a ponechme váhy $w_i = 1$ pro všechna i. Nyní získáme informaci, že v populaci jsou rovněž dva bratři podezřelého. Jejich věrohodnostní poměr položíme roven 1/4. Potom z (3.4) plyne

$$\mathsf{P}(G|E) = \frac{1}{1 + 2 \cdot 1/4 + 98 \cdot 0,01} = 0,4032258, \text{ tj. } 40,3\ \%$$

oproti původním 50 %. Nechť je jeden z bratrů nalezen a je ochoten se podrobit testu, jenž je posléze shledán jako negativní, a můžeme ho tedy z populace podezřelých vyloučit. Pravděpodobnost viny podezřelého tím výrazně vzroste:

$$\mathsf{P}(G|E) = \frac{1}{1 + 1/4 + 98 \cdot 0,01} = 0,4484305, \text{ tj. } 44,8\ \%.$$

Blíže se budeme populační strukturou zabývat ještě v kapitole 4.

3.5 Chyby a selhání

Chyby při testování

Při testování osob na přítomnost znaku Υ může dojít ke dvěma chybám testu. Pravděpodobnost, že osoba není nositelem Υ, a přesto je chybně testována jako pozitivní, označme κ_1. Podobně pravděpodobnost, že osoba, jež je nositelem Υ, je chybně testována jako negativní, označme κ_2. Předpokládejme, že tyto pravděpodobnosti zůstávají stejné při testování profilu podezřelého i profilu z místa činu a že chyby se vyskytují zcela nezávisle.

Nejprve předpokládejme, že podezřelý a pachatel jsou dvě různé osoby. Potom ke shodě profilů může dojít třemi způsoby:

- Podezřelý i pachatel jsou nositeli Υ (každý s pravděpodobností p) a k testovací chybě ani u jednoho nedošlo (u každého s pravděpodobností $1 - \kappa_2$) - celková pravděpodobnost je tedy $p^2(1 - \kappa_2)^2$.

- První z osob je nositelem Υ a test dává správný výsledek (tato událost má pravděpodobnost $p(1-\kappa_2)$), druhá osoba znak Υ nemá (s pravděpodobností $1-p$), ale je chybně otestována jako pozitivní (s pravděpodobností κ_1). Obě osoby můžeme zaměnit, a tak celková pravděpodobnost je $2p(1-p)\kappa_1(1-\kappa_2)$.

- Ani jedna z osob znak Υ nemá[2], ale obě jsou chybně testovány jako pozitivní - příslušná pravděpodobnost je $(1-p)^2\kappa_1^2$.

Nyní předpokládejme, že podezřelý a pachatel jsou jedna a tatáž osoba. Ke shodě profilů může dojít dvojí cestou:

- Pachatel je nositelem Υ (s pravděpodobností p) a ani při jednom ze dvou testů nedošlo k chybě (to má pravděpodobnost $(1-\kappa_2)^2$) - celkově tedy $p(1-\kappa_2)^2$.

- Pachatel není nositelem Υ (s pravděpodobností $1-p$), ale v obou testech byl chybně označen jako pozitivní (tato pro pachatele smolná shoda má pravděpodobnost κ_1^2) - celkově tedy $(1-p)\kappa_1^2$.

První tři možnosti odpovídají čitateli věrohodnostního poměru R_i, druhé dvě jeho jmenovateli:

$$\begin{aligned} R_i &= \frac{p^2(1-\kappa_2)^2 + 2p(1-p)\kappa_1(1-\kappa_2) + (1-p)^2\kappa_1^2}{p(1-\kappa_2)^2 + (1-p)\kappa_1^2} = \\ &= \frac{(p+\kappa_1 - p(\kappa_1+\kappa_2))^2}{p(1-\kappa_2)^2 + (1-p)\kappa_1^2} \approx \frac{(p+\kappa_1)^2}{p} \end{aligned}$$

Aproximace platí, pokud je každá z hodnot p, κ_1, κ_2 malá. Všimněme si, že v tom případě výraz nezávisí na κ_2.

Vzorec (2.5) má při zahrnutí testovacích chyb tvar

$$\mathsf{P}(G|E) = \frac{1}{1 + N\frac{(p+\kappa_1-p(\kappa_1+\kappa_2))^2}{p(1-\kappa_2)^2+(1-p)\kappa_1^2}} \approx \frac{1}{1+N\frac{(p+\kappa_1)^2}{p}}.$$

Položme $p = 0,01$, $N = 100$ a $\kappa_1 = \kappa_2 = 0,005$. Potom pro pravděpodobnost viny platí

$$\begin{aligned} \mathsf{P}(G|E) &= \frac{1}{1 + 100\frac{(0,01+0,005-0,01(0,005+0,005))^2}{0,01(1-0,005)^2+(1-0,01)0,005^2}} = \\ &= 0,3089398, \text{ tedy } 30,89\ \% \end{aligned}$$

a aproximace dává

$$\mathsf{P}(G|E) \approx \frac{1}{1+100\frac{(0,01+0,005)^2}{0,01}} = 0,3076923, \text{ tedy } 30,77\ \%.$$

Zatímco aproximace se od výsledku příliš neliší, ukazuje se, že zanedbání testovacích chyb významně poškozuje podezřelého: v našem příkladě klesla aposteriorní pravděpodobnost viny po započtení možnosti testovací chyby z 50 % přibližně na 31 %.

[2]V celé práci sice předpokládáme, že pachatel je nositelem znaku Υ, ale tuto informaci jsme získali pouze z provedeného testu. Proto musíme uvažovat i situaci, že pachatel nositelem tohoto znaku není a byl pouze chybně testován.

Další možné chyby

Špatný výsledek nemusí vzniknout pouze chybou testu. Jak známo z každodenní zkušenosti, ani člověk není neomylný, a možností, jak manipulovat s výsledky analýzy DNA, je hned několik. Již při odběru biologických stop a zajištění vzorku DNA je nutné s určitostí potvrdit, že s místem činu nebylo pachatelem ani nikým jiným manipulováno, a tím vyloučit, že došlo k umístění falešné stopy (nedopalek cigarety, vlas, krev, sperma), která má odvést pozornost vyšetřovatelů od skutečného pachatele.

Ani po příjezdu policie na místo činu však nelze s jistotou tvrdit, že jsou biologické stopy v bezpečí a dojde k jejich korektnímu zajištění pro potřeby následné analýzy. Opět hrozí nebezpečí poničení stop, případně jejich přehlédnutí, objevuje se i riziko druhotného přenosu biologického materiálu. Představme si, že policie zatkne podezřelého s odůvodněním, že se jeho biologické stopy našly na místě činu. Přestože podezřelý tvrdí, že na místě činu nikdy nebyl, analýza DNA hovoří proti němu. Existují v zásadě čtyři možná vysvětlení:

- podezřelý lže, neboť chce uniknout trestu
- jeho DNA se na místo dostala až druhotně (např. přenosem vlasů na oblečení)
- došlo k záměrné manipulaci stop pachatelem
- stopa, dosvědčující vinu podezřelého, byla na místo činu podstrčena osobou, pohybující se na místě činu v souvislosti s vyšetřováním

Poslední varianta je sice obtížně představitelná, ale přesto ji nelze neuvažovat. Vyloučit poslední tři možnosti je úkolem policejních složek.

Pokud zajistíme odpovídajícím způsobem dopravu biologického materiálu až do laboratoře, vynořují se další rizika. Jedním z nejvážnějších je nevyhovující označení vzorků nebo nedostatečná dokumentace ohledně manipulace se vzorkem. Jinou hrozbou při zpracování vzorků DNA v laboratoři je jejich kontaminace. Toto riziko je větší, než se může zdát - údaje z USA hovoří o tom, že chybovost vlivem kontaminace tvoří až 2 %. Chybné zpracování vzorku jsme si již rozebrali, kromě toho může dojít i ke stanovení chybných závěrů znaleckého posudku či k chybné interpretaci soudem či porotou.

Nelze se tedy divit, že při zpracování biologického materiálu s cílem získat profil DNA je třeba dodržet řadu přísných kontrolních mechanismů, které zahrnují i vícenásobnou kontrolu prováděných kroků. Pokud soudce dojde k závěru, že některý z těchto mechanismů nebyl dodržen, může DNA-evidenci zcela vyřadit ze soudního projednávání, což je vzhledem k průkazní síle, kterou se analýza DNA pyšní, velice nepříjemné pro obžalobu.

Na závěr je třeba zmínit, že i když k některým zmíněným chybám často dochází, nemá to až tak zdrcující důsledky, jak by se mohlo zdát. To, co nás zajímá, totiž není pravděpodobnost jakékoli chyby, která nastane, ale pravděpodobnost chyby, která povede k chybné shodě obou profilů; a tato pravděpodobnost je mnohem menší. Dobře to ilustruje příklad pocházející od anglického filozofa Davida Humea: Tisková chyba v novinách má mnohem větší pravděpodobnost než výhra v loterii. Pokud však noviny otisknou, že vyhrálo moje číslo, je velmi pravděpodobné, že jsem skutečně vyhrál. Podstatná je totiž pravděpodobnost tiskové chyby, která povede k vytištění mého čísla, a tato konkrétní chyba má pravděpodobnost mnohem menší než výhra v loterii.

Kapitola 4

Zahrnutí vlivu subpopulace

V této kapitole si odvodíme výsledek, který nám umožní při výpočtu aposteriorní pravděpodobnosti viny vzít do úvahy také populační strukturu. Podobně jako u osob, mezi nimiž je blízký příbuzenský vztah, totiž platí, že genetické profily osob ze stejné subpopulace nemusí být nezávislé, ale díky společnému evolučnímu vývoji mohou mít společné alely ve větší míře, než by odpovídalo jejich výskytu v populaci. Jak jsme již uvedli na straně 20, k vyjádření míry příbuznosti se využívá *coancestry koeficientu*, který se obvykle značí θ. Ten udává pro pevně zvolený lokus pravděpodobnost, že alely x a y jsou ibd za podmínky, že x a y pochází od dvou osob náhodně vybraných z populace. $\theta = 0$ značí, že podíl alel je stejný ve všech subpopulacích, tedy že populace je homogenní. Naproti tomu z $\theta = 1$ vyplývá, že v rámci každé subpopulace bylo dosaženo fixace na daném lokusu, v různých subpopulacích však mohou být fixovány různé alely. Obecně θ mluví o vztahu dvou alel uvnitř subpopulace vzhledem ke vztahu dvou alel z různých subpopulací. Slouží proto také jako míra rozdílnosti mezi subpopulacemi.

4.1 Beta-binomická formule

Mějme na pevně zvoleném lokusu dány alely A_i, které mají v celé populaci pravděpodobnost p_i. Z populace nyní vyberme subpopulaci o n alelách. Předpokládejme, že θ je známé a nenulové. Wright v [7] dokázal, že relativní četnost jednotlivých alel v této subpopulaci má Dirichletovo rozdělení s parametry λp_i, kde $\lambda = (1-\theta)/\theta$. To znamená, že pravděpodobnost, že se v subpopulaci vyskytne m_i alel A_i ($\sum_i m_i = n$), je dána vzorcem

$$\mathsf{P}(\bigcap_i A_i^{m_i}) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+n)} \prod_i \frac{\Gamma(\lambda p_i + m_i)}{\Gamma(\lambda p_i)}. \tag{4.1}$$

Dialelický lokus

Pokud uvažujeme pouze dvě možné alely A_1, A_2, vzorec (4.1) můžeme psát ve tvaru

$$\mathsf{P}(m, n-m) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+n)} \frac{\Gamma(\lambda p+m)}{\Gamma(\lambda p)} \frac{\Gamma(\lambda(1-p)+n-m)}{\Gamma(\lambda(1-p))}, \qquad (4.2)$$

kde p je pravděpodobnost alely A_1 a $1-p$ pravděpodobnost alely A_2 v populaci.

Pro gama funkci platí známý vzorec

$$\Gamma(n) = (n-1)\Gamma(n-1)$$

a jeho opakovaným použitím získáme

$$\Gamma(\lambda+n) = \Gamma(\lambda) \prod_{i=0}^{n-1} (\lambda+i).$$

Potom (4.2) můžeme upravit:

$$\mathsf{P}(m, n-m) = \frac{\prod_{i=0}^{m-1}(\lambda p+i) \prod_{i=0}^{n-m-1}(\lambda(1-p)+i)}{\prod_{i=0}^{n-1}(\lambda+i)}.$$

Po dosazení $\lambda = (1-\theta)/\theta$ a úpravě dostáváme

$$\mathsf{P}(m, n-m) = \frac{\prod\limits_{i=0}^{m-1}((1-\theta)p+\theta i) \prod\limits_{i=0}^{n-m-1}((1-\theta)(1-p)+\theta i)}{\prod\limits_{i=0}^{n-1}(1-\theta+\theta i)}. \qquad (4.3)$$

Jmenovatel lze ještě upravit:

$$\begin{aligned} \prod_{i=0}^{n-1}(1-\theta+\theta i) &= (1-\theta)\cdot 1\cdot(1+\theta)\cdots(1+(n-2)\theta) = \\ &= (1-\theta)\prod_{i=1}^{n-2}(1+\theta i) \end{aligned}$$

Rovnost (4.3) je potom tvaru

$$\mathsf{P}(m, n-m) = \frac{\prod\limits_{i=0}^{m-1}((1-\theta)p+\theta i) \prod\limits_{i=0}^{n-m-1}((1-\theta)(1-p)+\theta i)}{(1-\theta)\prod\limits_{i=1}^{n-2}(1+\theta i)}. \qquad (4.4)$$

Vzorec (4.4) se nazývá **beta-binomická (výběrová) formule (beta-binomial sampling formula)**. Platí pro uspořádané výběry; pokud

bychom ji chtěli použít i pro výběry neuspořádané, stačí ji vynásobit binomickým číslem $\binom{n}{m}$.

V dalším však budeme využívat spíše důsledek beta-binomické formule, který si nyní odvodíme. Předpokládejme, že jsme dosud ze subpopulace vybrali n alel; m z nich typu A_1, $n-m$ typu A_2. Chtěli bychom zjistit podmíněnou pravděpodobnost $\mathsf{P}(A_1|A_1^m, A_2^{n-m})$, že příští alela bude typu A_1:

$$\begin{aligned}\mathsf{P}(A_1|A_1^m, A_2^{n-m}) &= \frac{\mathsf{P}\left(A_1^{m+1}, A_2^{n-m}\right)}{\mathsf{P}\left(A_1^m, A_2^{n-m}\right)} = \frac{\mathsf{P}\left(m+1, n-m\right)}{\mathsf{P}\left(m, n-m\right)} = \\ &= \frac{\frac{\prod_{i=0}^{m}((1-\theta)p+\theta i)}{\prod_{i=1}^{n-1}(1+\theta i)}}{\frac{\prod_{i=0}^{m-1}((1-\theta)p+\theta i)}{\prod_{i=1}^{n-2}(1+\theta i)}} = \frac{(1-\theta)\,p + m\theta}{1+(n-1)\,\theta}.\end{aligned}$$

Pokud bude ve vzorci

$$\mathsf{P}(A_1|A_1^m, A_2^{n-m}) = \frac{(1-\theta)\,p + m\theta}{1+(n-1)\,\theta} \tag{4.5}$$

m nebo $n-m$ rovno nule, budeme na levé straně příslušný člen pro jednoduchost vynechávat.

Všimněme si, že pokud dosadíme do (4.5) $\theta = 0$, pravděpodobnost vytažení alely A_1 se rovná p. Nebude už tedy záviset na dříve vytažených alelách a podmíněná pravděpodobnost se změní na nepodmíněnou.

Vzorec (4.5) jsme odvodili pro dvě alely, ale použít ho lze i v případě, že se na daném lokusu rozlišuje více alel. Jestliže počítáme pravděpodobnost, že příští alela bude A_i, víme-li, že z n předchozích jich bylo právě m typu A_i, můžeme všechny ostatní alely, jichž bylo $n-m$, zahrnout do souhrnného označení A_i^C.

Multialelický lokus

Pokud na lokusu existuje K alel $A_1, \ldots, A_K$ s pravděpodobnostmi $p_1, \ldots, p_K$, $\sum_{i=1}^{K} p_i = 1$, vzorec (4.1) lze stejným postupem jako u dialelického lokusu upravit na

$$\mathsf{P}(m_1, \ldots, m_K) = \frac{\prod_{k=1}^{K}\prod_{i=0}^{m_k-1}((1-\theta)\,p_k + \theta i)}{(1-\theta)\prod_{i=1}^{n-2}(1+\theta i)}, \tag{4.6}$$

kde $n = m_1 + \ldots + m_K$. Stejně jako (4.4), i vzorec (4.6) platí pro uspořádané výběry. Budeme-li chtít pracovat s neuspořádaným výběrem, je třeba výsledek vynásobit $\frac{n!}{m_1!\cdots m_K!}$.

Analogicky vzorci (4.5) lze rovněž získat podmíněnou pravděpodobnost

$$\mathsf{P}(A_j|A_1^{m_1},\ldots,A_j^{m_j},\ldots,A_K^{m_K}) = \frac{(1-\theta)\,p_j + m_j\theta}{1+(n-1)\,\theta}. \tag{4.7}$$

4.2 Aplikace beta-binomické formule

V sekci 2.3 jsme odvodili vzorec pro výpočet aposteriorní pravděpodobnosti viny podezřelého, v němž jsme použili věrohodnostní poměr definovaný výrazem

$$R_i\,(E|E_0) = \frac{\mathsf{P}(E|C_i,E_0)}{\mathsf{P}(E|G,E_0)}. \tag{4.8}$$

Předpokládejme nyní, že evidence E, kterou chceme zahrnout, je tvořena pouze informací o tom, že pachatel a podezřelý mají stejný DNA profil (označme jej písmenem D). Toho lze dosáhnout tím, že DNA evidence bude vyhodnocována jako poslední a veškerou dříve získanou evidenci i informace o okolnostech zločinu zahrneme do počáteční evidence E_0. Označme pachatele písmenem C a podezřelého písmenem S (z anglických termínů *culprit* a *suspect*); genotyp osoby budeme značit písmenem G s vyznačením příslušné osoby v dolním indexu, tedy např. u pachatele a podezřelého G_C a G_S. Písmenem G budeme značit i nadále také hypotézu, že obžalovaný je vinen, ale v tomto označení nefiguruje žádný dolní index, a proto by nemělo dojít k záměně. Evidenci E můžeme nyní psát ve tvaru $G_C = G_S = D$ a vzorec (4.8) jako

$$R_i = \frac{\mathsf{P}\,(G_C = G_S = D|C_i,E_0)}{\mathsf{P}\,(G_C = G_S = D|G,E_0)} \tag{4.9}$$

(pro větší přehlednost budeme psát označení věrohodnostního poměru bez zdůraznění počáteční a zahrnované evidence).

V čitateli předpokládáme, že pachatelem je osoba i, můžeme proto psát $G_i = G_S = D$ místo $G_C = G_S = D$. Podobně ve jmenovateli: je-li pachatelem podezřelý, jev $G_C = G_S = D$ je ekvivalentní s jevem $G_S = D$. (4.9) lze tedy zjednodušit na

$$\begin{aligned} R_i &= \frac{\mathsf{P}\,(G_i = G_S = D|E_0)}{\mathsf{P}\,(G_S = D|E_0)} = \\ &= \mathsf{P}(G_i = D|G_S = D,E_0). \end{aligned} \tag{4.10}$$

Pravděpodobnost na pravé straně (4.10) se nazývá *pravděpodobnost shody genotypu (match probability)*. Je to podmíněná pravděpodobnost, že osoba náhodně vybraná z populace ponese daný genotyp, víme-li, že tento genotyp byl již u nějaké osoby z populace pozorován. Vztahuje se tedy ke dvěma osobám a je vždy větší nebo rovna nepodmíněné

pravděpodobnosti nalezení daného genotypu. Rovnost nastává pouze a právě tehdy, když jsou dané genotypy nezávislé. Jak jsme však již zmínili na začátku této kapitoly, pokud uvažované osoby pocházejí ze stejné subpopulace, není nezávislost genotypů (a tudíž ani genetických profilů) splněna. K vyčíslení pravděpodobností shody genotypu proto využijeme výše odvozenou beta-binomickou formuli a její důsledek v podobě vzorce (4.5), resp. (4.7).

Předpokládejme nejprve, že pachatel má homozygotní profil A_jA_j. V souladu se vzorcem (4.10) bychom chtěli vypočítat, jaká je na základě znalosti této informace pravděpodobnost, že podezřelý má stejný homozygotní profil:

$$\begin{aligned} R_i &= \mathsf{P}(G_i = A_jA_j | G_S = A_jA_j) \equiv \mathsf{P}(A_j^2 | A_j^2) = \\ &= \mathsf{P}(A_j | A_j^3) \cdot \mathsf{P}(A_j | A_j^2) \end{aligned}$$

Tyto podmíněné pravděpodobnosti dokážeme vypočítat pomocí vzorce (4.5); nejprve do něj dosadíme $m = n = 2$, poté $m = n = 3$. Celkem tedy platí

$$R_i = \frac{[(1-\theta)\,p_j + 2\theta]\,[(1-\theta)\,p_j + 3\theta]}{(1+\theta)\,(1+2\theta)}. \tag{4.11}$$

Podobně postupujeme pro heterozygotní profil A_jA_k:

$$\begin{aligned} R_i &= \mathsf{P}(G_i = A_jA_k | G_S = A_jA_k) \equiv \mathsf{P}(A_jA_k | A_jA_k) = \\ &= \mathsf{P}(A_k | A_j^2A_k^1)\mathsf{P}(A_j | A_j^1A_k^1) = \mathsf{P}(A_j | A_j^1A_k^2)\mathsf{P}(A_k | A_j^1A_k^1). \end{aligned} \tag{4.12}$$

Pro vyčíslení v obou vzorcích na spodním řádku (4.12) dosadíme $m = 1, n = 2$ a $m = 1, n = 3$, pouze dojde k záměně hodnot p_j a p_k. Celkem vychází

$$R_i = 2\frac{[(1-\theta)\,p_j + \theta]\,[(1-\theta)\,p_k + \theta]}{(1+\theta)\,(1+2\theta)}. \tag{4.13}$$

Všimněme si, že oba postupy ze (4.12) daly stejný výsledek. Tato shoda vyplývá ze vzorce (4.5). Jmenovatel v něm totiž závisí pouze na n, a tudíž je v obou případech stejný. Podobně m leží v (4.5) pouze v čitateli, a ten tak závisí u každé alely pouze na tom, kolikrát byla daná alela v dosavadním průběhu vybrána z populace, a nikoli na tom, kolikrát se vyskytly alely jiné. Výše řečené platí i pro vzorec (4.7), takže můžeme tento odstavec shrnout následující rovností:

$$\begin{aligned} \mathsf{P}\left(A_1^{l_1}, \ldots, A_K^{l_K} | A_1^{m_1}, \ldots, A_K^{m_K}\right) &= \frac{(l_1 + \ldots + l_K)!}{l_1! \cdots l_K!} \cdot \\ \cdot \frac{[(1-\theta)p_{A_1} + m_1\theta] \cdots [(1-\theta)p_{A_1} + (m_1 + l_1 - 1)\,\theta]}{[1 + (m_1 + \cdots + m_K - 1)\,\theta]} &\cdots \\ \cdots \frac{[(1-\theta)p_{A_K} + m_K\theta] \cdots [(1-\theta)p_{A_K} + (m_K + l_K - 1)\,\theta]}{[1 + (m_1 + \cdots + m_K + l_1 + \cdots + l_K - 1)\,\theta]} & \end{aligned} \tag{4.14}$$

Ačkoli tento vzorec vypadá dosti složitě, jeho použití je snadné a pro rychlý výpočet vhodnější než jeho jednodušší formy (4.5) a (4.7). Ukážeme si použití (4.14) na konkrétním příkladě.

Příklad
Chceme spočítat pravděpodobnost $\mathsf{P}(A_1, A_2^2, A_3|A_1^6, A_2^2)$. Do (4.14) tedy dosadíme $K = 3$, $m_1 = 6$, $m_2 = 2$ a $m_3 = 0$ (celkem $m_1 + m_2 + m_3 = 8$); dále $l_1 = l_3 = 1$ a $l_2 = 2$:

$$\mathsf{P}\left(A_1, A_2^2, A_3|A_1^6, A_2^2\right) = \\ \frac{12\left[(1-\theta)\,p_{A_1} + 6\theta\right]\left[(1-\theta)\,p_{A_2} + 2\theta\right]\left[(1-\theta)\,p_{A_2} + 3\theta\right]\left[(1-\theta)\,p_{A_3}\right]}{\left[1+7\theta\right]\left[1+8\theta\right]\left[1+9\theta\right]\left[1+10\theta\right]}.$$

Máme 8 alel dříve vybraných a 4 alely, jejichž pravděpodobnost vybrání chceme spočítat. Ve jmenovateli i čitateli proto budou čtyři závorky. Ve jmenovateli se od sebe liší koeficientem u θ, který roste od $8 - 1 = 7$ do $7 + 4 - 1 = 10$. V čitateli odpovídají dvě závorky alele A_2 - ty se od sebe liší opět pouze koeficientem u θ - a po jedné závorce alelám A_1 a A_3. Koeficient u θ je roven počtu stejných alel již vybraných z populace. Zlomek závěrem násobíme 4! (odpovídající 4 alelám) a dělíme 1!2!1! (odpovídající počtům jednotlivých alel).

Kapitola 5

Směsi DNA

V některých případech se stává, že genetický materiál, jenž byl nalezen na místě činu, je směsí DNA několika osob; tuto směs označíme E_C. Počet přispěvatelů do této směsi může být znám či odhadován z okolností zločinu, ale obvykle je počet přispěvatelů stanoven jako polovina počtu pozorovaných alel. To znamená, že pozorujeme-li například ve směsi tři nebo čtyři alely, soudíme, že se jedná pravděpodobně o směs DNA dvou osob. V této kapitole si ukážeme nejčastější případy. Evidencí budeme nyní vždy rozumět směs z místa činu a genetické profily všech známých osob.

5.1 Oběť a podezřelý

Předpokládejme, že do směsi nalezené na místě činu přispěly oběť (z anglického termínu *victim* ji budeme značit V) a jedna neznámá osoba. Věrohodnostní poměr R_i definovaný vzorcem (2.2) můžeme nyní napsat jako

$$R_i = \frac{\mathsf{P}\left(E_C, G_S, G_V | C_i\right)}{\mathsf{P}\left(E_C, G_S, G_V | G\right)}. \tag{5.1}$$

(Pro přehlednost vynecháváme počáteční evidenci E_0.)

Jelikož profily G_V, G_S nezávisí na tom, zda podezřelý a oběť přispívají do směsi, lze (5.1) dále zjednodušit:

$$\begin{aligned} R_i &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} \frac{\mathsf{P}\left(G_S, G_V | C_i\right)}{\mathsf{P}\left(G_S, G_V | G\right)} = \\ &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} = \frac{\mathsf{P}\left(E_C | G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)}. \end{aligned} \tag{5.2}$$

Poslední úpravu jsme mohli provést díky tomu, že za předpokladu, že pachatelem je osoba i, podezřelý nepřispívá do směsi.

Směs čtyř alel

Nejprve se podíváme na případ, že je směs tvořena čtyřmi alelami. Předpokládejme, že platí následující podmínky:

1. Žádné dvě uvažované osoby nejsou v příbuzenském vztahu.
2. Populace je homogenní (tj. $\theta = 0$).
3. V populaci platí Hardyova-Weinbergova rovnováha.

Nechť je směs tvořena alelami A, B, C, D se známými celkovými pravděpodobnostmi výskytu v populaci p_A, p_B, p_C, p_D; nechť podezřelý má alely A, B a oběť C, D. Jmenovatel ve vzorci (5.2) je roven jedné, čitatel je roven pravděpodobnosti pozorování osoby s alelami A, B, což za výše uvedených předpokladů je $2p_Ap_B$. Věrohodnostní poměr je tedy roven

$$R_i = 2p_Ap_B.$$

Předpokládejme nyní, že všechny tři uvažované osoby mají navzájem stejný stupeň příbuznosti vyjádřený coancestry koeficientem θ. Potom podle (4.14)

$$R_i = \mathsf{P}\left(AB|ABCD\right) = \frac{2\left[\left(1-\theta\right)p_A + \theta\right]\left[\left(1-\theta\right)p_B + \theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)}. \tag{5.3}$$

Směs tří alel

V případě výskytu tří alel ve vzorku je rovněž potřeba předpokládat minimálně dva přispěvatele do směsi. Uvažujme tedy alely A, B, C s pravděpodobnostmi výskytu v populaci p_A, p_B, p_C. Je-li oběť homozygot pro alelu C, pak dostaneme stejné výsledky jako v případě směsi čtyř alel.

Předpokládejme tedy, že oběť je heterozygot s alelami A, B. Nechť podezřelý je homozygot pro alelu C a jsou splněny podmínky 1 až 3. Jmenovatel vzorce (5.2) je opět roven jedné, čitatel je tentokrát roven pravděpodobnosti pozorování osoby, která má alelu C a zároveň nemá jinou alelu než A, B nebo C. Proto

$$R_i = \mathsf{P}(AC) + \mathsf{P}(BC) + \mathsf{P}(CC) = 2p_Ap_C + 2p_Bp_C + p_C^2\ . \tag{5.4}$$

K zahrnutí populační struktury opět využijeme vzorce (4.14):

$$\begin{aligned}
R_i &= \mathsf{P}\left(AC|ABCC\right) + \mathsf{P}\left(BC|ABCC\right) + \mathsf{P}\left(CC|ABCC\right) = \\
&= \frac{2\left[\left(1-\theta\right)p_A + \theta\right]\left[\left(1-\theta\right)p_C + 2\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)} + \\
&+ \frac{2\left[\left(1-\theta\right)p_B + \theta\right]\left[\left(1-\theta\right)p_C + 2\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)} + \\
&+ \frac{\left[\left(1-\theta\right)p_C + 3\theta\right]\left[\left(1-\theta\right)p_C + 2\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)} = \\
&= \frac{\left[\left(1-\theta\right)p_C + 2\theta\right]\left[\left(1-\theta\right)\left(2p_A + 2p_B + p_C\right) + 7\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)}.
\end{aligned} \tag{5.5}$$

V předcházejícím výpočtu jsme předpokládali, že podezřelý je homozygot pro alelu C. Je-li heterozygotem s alelami A a C, respektive B a C, za platnosti podmínek 1 až 3 vzorec (5.4) zůstává nezměněn; v případě zahrnutí populační struktury dostaneme stejným postupem v obou případech věrohodnostní poměr

$$R_i = \frac{[(1-\theta)\,p_C + \theta]\,[(1-\theta)\,(2p_A + 2p_B + p_C) + 8\theta]}{(1+3\theta)\,(1+4\theta)}. \tag{5.6}$$

5.2 Podezřelý a neznámá osoba

Některé vzorky z místa činu obsahují DNA od více než jedné osoby, ale pouze jedna známá osoba je podezřelá být přispěvatelem do směsi. Je třeba definovat, jak vypadají hypotézy G a C_i ze vzorce (2.2):

G : pachateli jsou podezřelý a neznámá osoba N
C_i : pachateli jsou osoba i a neznámá osoba N.

Podobně jako v sekci 5.1 dostaneme

$$R_i = \frac{\mathsf{P}\,(E_C|C_i)}{\mathsf{P}\,(E_C|G_S, G)}. \tag{5.7}$$

Směs čtyř alel

Uvažujme alely A, B, C, D s příslušnými pravděpodobnostmi p_A, p_B, p_C, p_D a nechť $G_S = AB$. Při splnění podmínek 1 až 3 je jmenovatel vzorce (5.7) roven pravděpodobnosti pozorování osoby N s alelami C, D, kterážto pravděpodobnost je $2p_Cp_D$.

Čitatel je za stejných podmínek roven pravděpodobnosti, že osoby i a N budou mít dohromady všechny čtyři alely $A, \ldots, D$. Takovýchto možností je šest a jsou uvedeny v následující tabulce:

G_i	G_N	$\mathsf{P}\,(G_i, G_N)$
AB	CD	$2p_Ap_B \cdot 2p_Cp_D$
AC	BD	$2p_Ap_C \cdot 2p_Bp_D$
AD	BC	$2p_Ap_D \cdot 2p_Bp_C$
BC	AD	$2p_Bp_C \cdot 2p_Ap_D$
BD	AC	$2p_Bp_D \cdot 2p_Ap_C$
CD	AB	$2p_Cp_D \cdot 2p_Ap_B$

Celkem tedy

$$R_i = \frac{24p_Ap_Bp_Cp_D}{2p_Cp_D} = 12p_Ap_B.$$

Pro zahrnutí populační struktury uvažujme opět stejnou míru příbuznosti mezi osobami i, N a S vyjádřenou coancestry koeficientem θ. Věrohodnostní poměr vypočteme podle (4.14):

$$\begin{aligned} R_i &= \frac{\mathsf{P}\left(ABCD|AB\right)}{\mathsf{P}\left(CD|AB\right)} = \\ &= \frac{\frac{24[(1-\theta)p_A+\theta][(1-\theta)p_B+\theta](1-\theta)p_C(1-\theta)p_D}{(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}}{\frac{2(1-\theta)p_C(1-\theta)p_D}{(1+\theta)(1+2\theta)}} = \\ &= \frac{12\left[\left(1-\theta\right)p_A+\theta\right]\left[\left(1-\theta\right)p_B+\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)} \end{aligned} \tag{5.8}$$

Směs tří alel

Mějme směs E_C tvořenou alelami A, B, C s příslušnými pravděpodobnostmi p_A, p_B, p_C a nechť jsou splněny podmínky 1 až 3. Nejprve předpokládejme, že podezřelý je homozygot pro alelu C. Jmenovatel vzorce (5.7) je potom roven pravděpodobnosti pozorování osoby N s alelami B, C, tj. $2p_Bp_C$.

Čitatel vypočítáme jako pravděpodobnost získání všech tří alel A, B, C od osob i a N. Všechny možnosti včetně příslušných pravděpodobností jsou uvedeny v následující tabulce:

G_i	G_N	$\mathsf{P}\left(G_i, G_N\right)$
AA	BC	$p_A^2 \cdot 2p_Bp_C$
BB	AC	$p_B^2 \cdot 2p_Ap_C$
CC	AB	$p_C^2 \cdot 2p_Ap_B$
AB	AC	$2p_Ap_B \cdot 2p_Ap_C$
AB	BC	$2p_Ap_B \cdot 2p_Bp_C$
AB	CC	$2p_Ap_B \cdot p_C^2$
AC	AB	$2p_Ap_C \cdot 2p_Ap_B$
AC	BB	$2p_Ap_C \cdot p_B^2$
AC	BC	$2p_Ap_C \cdot 2p_Bp_C$
BC	AA	$2p_Bp_C \cdot p_A^2$
BC	AB	$2p_Bp_C \cdot 2p_Ap_B$
BC	AC	$2p_Bp_C \cdot 2p_Ap_C$

Sečtením pravděpodobností ve třetím sloupci dostaneme

$$12p_Ap_Bp_C\left(p_A+p_B+p_C\right),$$

věrohodnostní poměr potom vychází

$$R_i = 6p_A\left(p_A+p_B+p_C\right).$$

Nyní prostřednictvím θ zahrňme do výpočtu také populační strukturu. Jmenovatel vzorce (5.7) vypočteme jako $P(BC|AA)$. Ze vzorce (4.14) dostaneme

$$\frac{2\left(1-\theta\right)^2 p_B p_C}{\left(1+\theta\right)\left(1+2\theta\right)}.$$

Nyní se zaměříme na čitatel. Od osob i a N chceme získat čtyři alely, mezi nimiž musí být každá z alel A, B, C a zároveň žádná jiná. Je zřejmé, že existují jen tři takovéto kombinace: $AABC$, $ABBC$ a $ABCC$. Opět známe pouze genotyp podezřelého AA, potřebujeme tedy spočítat podmíněné pravděpodobnosti $\mathsf{P}(AABC|AA)$, $\mathsf{P}(ABBC|AA)$ a $\mathsf{P}(ABCC|AA)$ a poté je sečíst. Dosazením do vzorce (4.14) dostaneme

$$\mathsf{P}(AABC|AA) = \frac{12\left[\left(1-\theta\right)p_A + 2\theta\right]\left[\left(1-\theta\right)p_A + 3\theta\right]\left(1-\theta\right)^2 p_B p_C}{\left(1+\theta\right)\left(1+2\theta\right)\left(1+3\theta\right)\left(1+4\theta\right)},$$

$$\mathsf{P}(ABBC|AA) = \frac{12\left[\left(1-\theta\right)p_A + 2\theta\right]\left[\left(1-\theta\right)p_B + \theta\right]\left(1-\theta\right)^2 p_B p_C}{\left(1+\theta\right)\left(1+2\theta\right)\left(1+3\theta\right)\left(1+4\theta\right)}$$

a

$$\mathsf{P}(ABCC|AA) = \frac{12\left[\left(1-\theta\right)p_A + 2\theta\right]\left[\left(1-\theta\right)p_C + \theta\right]\left(1-\theta\right)^2 p_B p_C}{\left(1+\theta\right)\left(1+2\theta\right)\left(1+3\theta\right)\left(1+4\theta\right)}.$$

Celý věrohodnostní poměr je potom tvaru

$$R_i = \frac{6\left[\left(1-\theta\right)p_A + 2\theta\right]\left[\left(1-\theta\right)\left(p_A + p_B + p_C\right) + 5\theta\right]}{\left(1+3\theta\right)\left(1+4\theta\right)}. \tag{5.9}$$

Je-li podezřelý heterozygotem s alelami A, B, zůstává čitatel vzorce (5.7) stejný a ve jmenovateli je pravděpodobnost, že neznámá osoba bude mít alelu C a zároveň nebude mít jinou alelu než A, B, C. Jmenovatel je proto roven součtu pravděpodobností vyskytu genotypů AC, BC a CC a věrohodnostní poměr vychází za podmínek 1 až 3

$$R_i = \frac{12 p_A p_B \left(p_A + p_B + p_C\right)}{2p_A + 2p_B + p_C}. \tag{5.10}$$

Pro zahrnutí populační struktury opět počítáme čitatel a jmenovatel zvlášť. Ve jmenovateli spočítáme pomocí vzorce (4.14) pravděpodobnosti $\mathsf{P}(AC|AB)$, $\mathsf{P}(BC|AB)$ a $\mathsf{P}(CC|AB)$; jejich sečtením dostaneme

$$\frac{2\left(1-\theta\right)p_C\left[\left(1-\theta\right)\left(2p_A + 2p_B + p_C\right) + 5\theta\right]}{\left(1+\theta\right)\left(1+2\theta\right)}.$$

V čitateli spočítáme a poté sečteme pravděpodobnosti $\mathsf{P}(AABC|AB)$, $\mathsf{P}(ABBC|AB)$ a $\mathsf{P}(ABCC|AB)$. Dostaneme

$$\frac{12\left[\left(1-\theta\right)p_A + \theta\right]\left[\left(1-\theta\right)p_B + \theta\right]\left[\left(1-\theta\right)\left(p_A + p_B + p_C\right) + 5\theta\right]\left(1-\theta\right)p_C}{\left(1+\theta\right)\left(1+2\theta\right)\left(1+3\theta\right)\left(1+4\theta\right)}$$

a věrohodnostní poměr poté vychází

$$R_i = \frac{12[(1-\theta)p_A + \theta][(1-\theta)p_B + \theta][(1-\theta)(p_A + p_B + p_C) + 5\theta]}{(1+3\theta)(1+4\theta)[(1-\theta)(2p_A + 2p_B + p_C) + 5\theta]}. \quad (5.11)$$

5.3 Dva podezřelí

V případě dvou podezřelých je několik možností, jak definovat hypotézy G a C_i. Ukážeme si tedy několik věrohodnostních poměrů a jejich interpretaci.

Označme S_1, S_2 podezřelé a i_1, i_2 alternativní podezřelé. Proti podezřelému S_1 uvažujme tři různé věrohodnostní poměry:

$$R_i^a = \frac{\mathsf{P}(E|i_1 \text{ a } S_2 \text{ jsou přispěvateli do směsi})}{\mathsf{P}(E|S_1 \text{ a } S_2 \text{ jsou přispěvateli do směsi})}$$

$$R_i^b = \frac{\mathsf{P}(E|i_1 \text{ a } i_2 \text{ jsou přispěvateli do směsi})}{\mathsf{P}(E|S_1 \text{ a } S_2 \text{ jsou přispěvateli do směsi})}$$

$$R_i^c = \frac{\mathsf{P}(E|i_1 \text{ a } i_2 \text{ jsou přispěvateli do směsi})}{\mathsf{P}(E|S_1 \text{ a } i_1 \text{ jsou přispěvateli do směsi})}.$$

R_i^a je použitelné pro stanovení váhy evidence proti podezřelému S_1 za předpokladu, že S_2 je přispěvatelem do směsi. Ovšem použití R_i^a u soudu by odpovídalo tomu, že S_2 je vinen, což není vhodné, pokud jsou považováni za spolupachatele a vyslýcháni souběžně. Tento věrohodnostní poměr je možné použít například tehdy, jestliže se S_2 přizná; výpočet je potom stejný jako v sekci 5.1.

R_i^b odpovídá stanovení váhy evidence proti S_1 a S_2 zároveň. Tento postup však neodpovídá potřebě rozhodnout zvlášť o vině S_1 a zvlášť o vině S_2, proto ani R_i^b není příliš vhodný.

R_i^c stanovuje váhu evidence proti podezřelému S_1 za předpokladu, že druhý přispěvatel je neznámý, aniž by bral v potaz, zda je S_2 pachatelem či nikoli. Proto je obvykle nejvhodnější a nadále budeme obvyklým značením R_i myslet právě jej.

Nebudeme nyní už zkoumat všechny možnosti jako v předchozích sekcích, neboť postup je velmi podobný. Ukážeme si pouze případ směsi tří alel a v jeho závěru rozebereme, jakou hodnotu má věrohodnostní poměr pro některé konkrétní hodnoty parametrů p a θ.

Nechť směs z místa činu obsahuje alely A, B, C, $G_{S_1} = AB$ a $G_{S_2} = CC$. Ekvivalentně vzorci (5.7) můžeme získat následující věrohodnostní poměr:

$$R_i = \frac{\mathsf{P}(E_C = ABC|G_{S_1} = AB, G_{S_2} = CC, i_1 \text{ a } i_2 \text{ jsou přispěvateli})}{\mathsf{P}(E_C = ABC|G_{S_1} = AB, G_{S_2} = CC, S_1 \text{ a } i_1 \text{ jsou přispěvateli})}.$$

Za předpokladu, že jsou splněny podmínky 1 až 3, věrohodnostní poměr vychází stejně jako ve vzorci (5.10), tedy

$$R_i = \frac{12p_A p_B\,(p_A + p_B + p_C)}{2p_A + 2p_B + p_C}.$$

Pokud položíme $p_A = p_B = p_C = p$, potom $R_i = 36p^2/5$. S rostoucím p roste i R_i, maximální je tedy pro $p = 1/3$; tehdy nabývá R_i hodnoty $0,8$. Jaké jsou hodnoty R_i pro některá p, ukazuje následující tabulka:

p	$1/3$	$0,2$	$0,1$	$0,05$	$0,01$
R_i	$0,8$	$0,288$	$0,072$	$0,018$	$0,00072$

S klesající hodnotou p tedy roste aposteriorní pravděpodobnost viny podezřelého S_1.

Předpokládejme tedy nyní, že všechny čtyři uvažované osoby jsou vybrány ze stejné subpopulace charakterizované coancestry koeficientem θ (ale mezi žádnou dvojicí není příbuzenský vztah). Čitatel je součtem podmíněných pravděpodobností $\mathsf{P}(AABC|ABCC)$, $\mathsf{P}(ABBC|ABCC)$ a $\mathsf{P}(ABCC|ABCC)$, jmenovatel potom získáme sečtením $\mathsf{P}(AC|ABCC)$, $\mathsf{P}(BC|ABCC)$ a $\mathsf{P}(CC|ABCC)$. Celkem vychází

$$R_i = \frac{12[(1-\theta)p_A + \theta][(1-\theta)p_B + \theta][(1-\theta)(p_A + p_B + p_C) + 7\theta]}{(1+5\theta)\,(1+6\theta)\,[(1-\theta)\,(2p_A + 2p_B + p_C) + 7\theta]}. \quad (5.12)$$

V následující tabulce jsou uvedeny hodnoty R_i pro některá p a θ:

$\theta \diagdown p$	$1/3$	$0,2$	$0,1$	$0,05$	$0,01$
0 %	$0,8$	$0,288$	$0,072$	$0,018$	$0,00072$
1 %	$0,768$	$0,292$	$0,083$	$0,0263$	$0,00356$
2 %	$0,739$	$0,295$	$0,093$	$0,0346$	$0,00775$
5 %	$0,668$	$0,301$	$0,120$	$0,0588$	$0,02489$
10 %	$0,582$	$0,304$	$0,152$	$0,0949$	$0,05797$

Vidíme, že pro některá pevná p věrohodnostní poměr s rostoucím θ roste, pro některá p naopak klesá; pro některé hodnoty p dokonce věrohodnostní poměr jakožto funkce parametru θ není monotónní (např. pro $p = 0,2$).

V [6] se uvádí, že konzervativní odhad hodnoty koeficientu θ pro velké subpopulace (např. celé USA) je $0,01$ a pro malé izolované subpopulace $0,03$. V tomto rozmezí takřka výhradně platí, že s rostoucím θ roste také hodnota věrohodnostního poměru. Zanedbání populační struktury tedy obvykle svědčí v neprospěch obžalovaného. Výjimkou je například případ, kdy jsou v nalezené směsi obsaženy všechny známé alely (jako tomu bylo v případě tří alel, z nichž každá měla pravděpodobnost $1/3$).

Kapitola 6

Rozšíření pro více lokusů

V příkladě na konci sekce 3.1 jsme měli dány dvě složky evidence - výpověď svědka a DNA-evidenci. Protože jsme předpokládali, že tyto dvě složky jsou vzájemně nezávislé, celkový věrohodnostní poměr jsme dostali vynásobením věrohodnostních poměrů příslušejících jednotlivým položkám. Podobné pravidlo (v anglicky psané literatuře nazývané *product rule*) bychom chtěli použít také v případě porovnání DNA profilů na více lokusech.

Postupem popsaným v předchozích kapitolách dostaneme věrohodnostní poměry pro jednotlivé zkoumané lokusy. Jejich zkombinování skrze vynásobení však předpokládá statistickou nezávislost a předpoklad statistické nezávislosti je zde ekvivalentní předpokladu nezávislosti genotypů na různých lokusech.

Genotypy dvou osob obvykle nezávislé nejsou, obzvláště z důvodu příbuznosti, ať už v rámci rodiny, či v rámci subpopulace (viz sekci 3.4). Nabízí se dvě možná řešení. Můžeme vybírat geny ležící na různých chromozomech; pak mezi nimi neexistuje žádná vazba a výslednou hodnotu dostaneme jako součin hodnot odpovídajících jednotlivým nezávislým lokusům. Druhou možností (obecnější, avšak složitější) je najít vnější faktor, pomocí kterého závislé lokusy vyrovnáme tak, abychom poté mohli použít výše popsané pravidlo o násobení. Tímto vnějším faktorem je míra příbuznosti, vyjádřená koeficientem θ. Obnovení alespoň přibližné nezávislosti pomocí vyrovnání vzhledem k míře příbuznosti je však tématem dosti obšírným a v současné odborné literatuře stále ještě hojně zkoumaným, proto jej zde nebudeme více rozebírat.

Kapitola 7

Příloha 1

V této příloze jsou naprogramovány některé vzorce z kapitol 4 a 5 ve statistickém programu **R**. Na jednotlivé vzorce je odkazováno pomocí jejich čísla v textu. Před jejich aplikací je potřeba zadat hodnoty parametru θ (zde označovaného jako t) a pravděpodobností p_j, p_k, resp. pA, pB, pC.

Vzorec (4.11):

```
Ri<-((1-t)*p_j+2*t)*((1-t)*p_j+3*t)/((1+t)*(1+2*t))
```

Vzorec (4.13):

```
Ri<-2*((1-t)*p_j+t)*((1-t)*p_k+t)/((1+t)*(1+2*t))
```

Vzorec (5.3):

```
Ri<-2*((1-t)*pA+t)*((1-t)*pB+t)/((1+3*t)*(1+4*t))
```

Vzorec (5.5):

```
Ri<-((1-t)*pC+2*t)*((1-t)*(2*pA+2*pB+pC)+7*t)/
  ((1+3*t)*(1+4*t))
```

Vzorec (5.6):

```
Ri<-((1-t)*pC+t)*((1-t)*(2*pA+2*pB+pC)+8*t)/
  ((1+3*t)*(1+4*t))
```

Vzorec (5.8):

```
Ri<-12*((1-t)*pA+t)*((1-t)*pB+t)/((1+3*t)*(1+4*t))
```

Vzorec (5.9):

```
Ri<-6*((1-t)*pA+2*t)*((1-t)*(pA+pB+pC)+5*t)/
  ((1+3*t)*(1+4*t))
```

Vzorec (5.11):

```
Ri<-12*((1-t)*pA+t)*((1-t)*pB+t)*((1-t)*(pA+pB+pC)+5*t)/
 ((1+3*t)*(1+4*t)*((1-t)*(2*pA+2*pB+pC)+5*t))
```

Vzorec (5.12):

```
Ri<-12*((1-t)*pA+t)*((1-t)*pB+t)*((1-t)*(pA+pB+pC)+7*t)/
 ((1+5*t)*(1+6*t)*((1-t)*(2*pA+2*pB+pC)+7*t))
```

Příloha 2

Připojíme zde rovněž vzorce ze sekce 3.3. Naprogramované byly v softwaru Mathematica. Nepotřebují žádné další zadání, jejich výsledkem je prvních deset požadovaných hodnot.

Tabulku koeficientů u členu ε/N^2 v doporučené aproximaci lze získat následujícím příkazem:

```
Table[Limit[n^2/e*Together[1 - Apart[Together[
1/(n*Together[Sum[e/(2^(Abs[k] - 1)
(n + k + 1)), {k, -g, g}] +
(1 - 4*e - e*Sum[1/2^k, {k, 0, g - 2}])/(n + 1)])*
(1 - Together[Sum[e/(2^(Abs[k] - 1)
(n + k + 1)), {k, -g, g}] +
(1 - 4*e - e*Sum[1/2^k, {k, 0, g - 2}])/(n + 1)])]]],
n -> Infinity], {g, 1, 10}]
```

Horní hranici intervalu, z nějž můžeme volit ε, získáme příkazem

```
Table[2^(g - 2)/(2^g - 1), {g, 1, 10}]
```

a tabulku optimálních ε příkazem

```
Table[2^(g - 2)/(3*2^(g - 1) - 1), {g, 1, 10}]
```

Literatura

[1] Balding D.J.: *Weight-of-evidence for forensic DNA profiles*, John Wiley & Sons, Ltd, 2005

[2] Ewett I.W., Weir B.S.: *Interpreting DNA evidence: Statistical genetics for forensic scientists*, Sinauer, Sunderland, 1998

[3] Flegr J.: *Zamrzlá evoluce aneb je to jinak, pane Darwin*, Academia, Praha, 2006

[4] Folda J.: články *Otevřete ústa, prosím...* a *Databáze DNA*, dostupné 31.7.2009 z internetové adresy

`http://www.uoou.cz/uoou.aspx?menu=287&submenu=288`

[5] Jelínek J.: *Biologie člověka a úvod do obecné genetiky*, FIN, Olomouc, 1994

[6] National Research Council: *The evaluation of forensic DNA evidence*, National Academy Press, Washington, DC., 1996

[7] Wright S.: *The genetical structure of populations*, Ann. Eugen. **15**, 1951, str. 323-354

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

[9] Zvárová J., Mazura I.: *Stochastická genetika*, Karolinum, Praha, 2001