Katedra aplikovanej matematiky a štatistiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Pravdepodobnosť a štatistika pri DNA-dôkazoch v kriminalistike

(Diplomová práca)

Bc. Katarína Prusáková

Bratislava, 2012

Katedra aplikovanej matematiky a štatistiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Pravdepodobnosť a štatistika pri DNA-dôkazoch v kriminalistike

(Diplomová práca)

Bc. Katarína Prusáková

Vedúci: Mgr. Ján Somorčík, PhD.
Študijný program: Ekonomická a finančná matematika
Študijný odbor: 1114 Aplikovaná matematika

Bratislava, 2012

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Katarína Prusáková
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Pravdepodobnosť a štatistika pri DNA-dôkazoch v kriminalistike

Cieľ: Kriminalisti na základe DNA-analýz následne uskutočnujú pravdepodobnostné výpočty, ktoré priradia výsledku DNA-analýzy takzvanú 'value of evidence', ktorá má následne pôsobiť na súd pri rozhodovaní o vine či nevine podozrivého. Cieľom práce by malo byť podrobné matematické popísanie kriminalistických scenárov, ktoré sú v súčasnej literatúre analyzované len okrajovo.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky

Dátum zadania: 13.01.2011

Dátum schválenia: 14.01.2011

prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestne prehlasujem, že som túto diplomovú prácu vypracovala samostatne s použitím citovaných zdrojov.

Bratislava 15. 4. 2012

. . . . . .

Poďakovanie

V prvom rade by som chcela poďakovať vedúcemu svojej diplomovej (aj bakalárskej) práce doktorovi Jánovi Somorčíkovi za to, že mi ušil skvelú tému na mieru, za úlohy, ktoré pre mňa vymýšľal, za jeho trpezlivosť a všetok čas, ktorý mi venoval napriek tomu, že sám ho má málo. Pedagógov ako on už dnes veľa nie je a ja som mala úžasné šťastie, že som na neho narazila.

Za dáta veľmi pekne ďakujeme zamestnancom Oddelenia biológie a genetickej analýzy Kriminalistického a expertízneho ústavu Policajného zboru SR. Bez ich príspevku by celá tretia kapitola bola len suchá teória.

Za neoceniteľnú radu pri grafickom znázorňovaní výsledkov ďakujeme profesorovi Davidovi J. Baldingovi z University College London, autorovi článkov [5] a [4], ktorého model sme v práci použili a bol ochotný sa s nami podeliť o svoje know-how.

Z celého srdca ďakujem môjmu manželovi Miškovi, ktorý dokázal, že stojí pri mne v zdraví aj v chorobe, bez ktorého pomoci by som bola nútená prerušiť štúdium a k písaniu diplomovej práce by som sa ani nedostala.

Veľká vďaka patrí aj ockovi, mamke, starkým a Zuzke, ktorí ma podporovali počas celého môjho štúdia a dali mi silu vždy, keď som to potrebovala.

A ešte môjmu psíkovi Atosovi.

Abstrakt

Autor:	Bc. Katarína Prusáková
Názov práce:	Pravdepodobnosť a štatistika pri DNA-dôkazoch v kriminalistike
Škola:	Univerzita Komenského v Bratislave
Fakulta:	Fakulta matematiky, fyziky a informatiky
Katedra:	Katedra aplikovanej matematiky a štatistiky
Vedúci:	Mgr. Ján Somorčík, PhD.
Miesto a rok:	Bratislava, 2012
Rozsah práce:	71 strán
Kľúčové slová:	value of evidence, náhodná zhoda, koeficient príbuznosti

Určovať value of evidence DNA-dôkazového materiálu znamená určovať pravdepodobnosť náhodnej zhody genotypu podozrivého a páchateľa trestného činu. Situácia sa môže skomplikovať, ak existuje alternatívny podozrivý, ktorý je pokrvným príbuzným pôvodného podozrivého.

Ďalšia komplikácia nastáva, ak páchateľ aj podozrivý pochádzajú zo špecifickej subpopulácie v danom regióne. Vtedy treba brať do úvahy koeficient príbuznosti odlišujúci navzájom jednotlivé subpopulácie, ktorých členovia majú na základe prítomnosti spoločných predkov podobné genotypy.

Práca sa zaoberá matematickým pozadím analýzy DNA a výpočtami podmienenej pravdepodobnosti. Druhá časť práce sa venuje odhadu koeficientu príbuznosti, ktorý je na Slovensku zatiaľ takmer neznámym pojmom.

Abstract

Author:	Bc. Katarína Prusáková
Title:	Probability and statistics in DNA-proofs in criminology
School:	Comenius university, Bratislava
Faculty:	Faculty of mathemathics, physics and computer science
Department:	Department of applied mathematics and statistics
Supervisor:	Mgr. Ján Somorčík, PhD.
Place and year:	Bratislava, 2012
Length of thesis:	71 pages
Keywords:	value of evidence, random match, co-ancestry coefficient

Determining value of evidence of a DNA sample in crime investigation means computing the match probability of suspects genotype with the one of the offender. The situation can become complicated when there is an alternative suspect who is blood-related to the original one.

Another complication occurs when the suspect and the offender originate from a specific subpopulation in the given region. Then co-ancestry coefficient must be regarded which enables us to distinguish between populations with common ancestors and therefore similar genetic profiles.

This thesis focuses on mathematical background of DNA analysis and calculations of conditional probability. Its second part is devoted to estimation of co-ancestry coefficient which is an almost unknown concept in Slovakia.

Predhovor

Vzorka DNA ako dôkazový materiál je často kľúčovým krokom k vyriešeniu policajného prípadu. Okrem genetiky a biológie zohráva v procese jej použitia dôležitú úlohu aj matematika, presnejšie poznatky z pravdepodobnosti a štatistiky. Dôležité je napríklad vedieť vypočítať pravdepodobnosť, že genotypy dvoch osôb v nejakom pokrvnom vzťahu sa zhodujú, alebo sú v istom zmysle podobné. Zaujímavé je aj rozlišovanie subpopulácií žijúcich na určitom území pomocou koeficientu príbuznosti, pretože je to možnosť lepšie špecifikovať páchateľa a znížiť pravdepodobnosť obvinenia nesprávnej osoby. Navyše, vďaka koeficientu príbuznosti môžeme na výpočet value of evidence používať stále celoslovenskú databázu, pretože on ju adaptuje na podmienky daného regiónu.

Cieľom tejto diplomovej práce je priniesť čitateľovi pohľad do zákulisia vyhodnocovania sily a relevancie dôkazového materiálu a poukázať na zaujímavé prípady, s ktorými sa v praxi možno stretnúť. Ďalej prináša odhady koeficientu príbuznosti pre populácie žijúce v jednotlivých krajoch. Na záver práca poukazuje aj na chybu, ktorej sa možno dopustiť, ak odborník nemá na zreteli aj matematickú stránku problému a praktický dôsledok, ak sa nesprávne vyhodnotený materiál použije ako dôkaz v súdnom procese.

Práca okrajovo načrtáva aj mnohé iné zaujímavé miesta, kde sa v DNA-analýze matematika využíva, no pre krátkosť času a nedostatok priestoru sa im nevenuje podrobnejšie.

Obsah

Kapitola 1

Úvod

V Spojených štátoch amerických sa analýza DNA začala v kriminalistike využívať v polovici osemdesiatych rokov minulého storočia. Odvtedy sa vďaka rozvoju vied (najmä genetiky) a výpočtovej techniky jej používanie zdokonalilo a rozšírilo do všetkých krajín sveta vrátane Slovenskej republiky. Pre vyšetrovateľov je DNA mnohokrát kľúčovým dôkazom pri usvedčení páchateľa trestného činu. V úvode práce zhrnieme poznatky, ktoré budeme počas práce potrebovať a ilustrujeme, ako sa táto analýza v praxi realizuje.

1.1 Analýza DNA

Na začiatok si spomeňme, čo vieme o DNA zo strednej školy. Pamätáme si, že je pre každého jedinca (DNA majú aj rastliny, hoci v trochu inej podobe ako ľudia) rôzna, čo z nej robí jednoznačný identifikačný znak, akým je u ľudí aj odtlačok prsta, alebo dúhovka v oku. Nachádza sa v každej bunke a je v nej uložená ako dvojzávitnicová špirála.

Popisu štruktúry DNA sa nebudeme podrobne venovať, lebo to nie je predmetom tejto práce a zároveň je to skôr genetická problematika. Bude nám stačiť predstava, že reťazec DNA sa skladá z častí, ktoré sa nazývajú *locusy*. Vo vnútri každého locusu sú dve *alely*, ktoré sú priamo nositeľom genetickej informácie. Tieto pochádzajú od rodičov, jedna od otca a druhá od matky. Pre ľahšiu predstavu majme locus, kde je uložená informácia o farbe očí. Nech sú v ňom dve alely, z ktorých jedna je pre hnedú (od otca) a druhá pre modrú farbu (od matky). Poznatky z genetiky hovoria, že gén pre hnedé oči je dominantný voči génu pre modré oči, preto dieťa bude mať oči „po otcovi“.

V rámci analýzy DNA sa neskúma celý reťazec, ale len vybrané locusy, resp. alely v nich. Výber toho, ktoré locusy sa skúmajú, nie je náhodný. Je to však komplikovaná téma,

ktorej by sa mohla venovať ďalšia diplomová práca, preto sa ňou podrobne zaoberať nebudeme. Stačí nám vedieť, že sa vyberajú tak, aby sa minimalizovalo riziko náhodnej zhody, o ktorej sa zmienime neskôr. Predpokladáme teda, že máme pevne dané, ktoré locusy do analýzy zahrnúť. Výstupné dáta takejto analýzy (t.j. rozbor vzorky získanej od podozrivých osôb) môžu vyzerať napríklad takto:

názov locusu	$TH01$		VWA		$D21S11$	
	1. alela	2. alela	1. alela	2. alela	1. alela	2. alela
osoba 1	9	9.3	16	18	29	32.2
osoba 2	9	9.3	18	18	29	32.2
osoba 3	9	9	14	18	29	29
osoba 3	7	9	16	16	28	29

Vidíme, že každý locus má svoj názov a obsahuje dve alely, ktoré v praxi reprezentujeme číslami. Sú to počty opakovaní rovnakých základných stavebných jednotiek DNA – *nukleotidov* na skúmanom locuse[1]. Postupnosť locusov v tvare, v akom ho vidíme v tabuľke, sa nazýva *genotyp* osoby.

1.2 Použitie analýzy DNA v kriminalistike

V tejto podkapitole sa budú vyskytovať tri slová, ktorých význam si čitateľ môže nechtiac zameniť a preto môže dôjsť k mylnému vysvetleniu popísanej problematiky. Preto pokladáme za potrebné najskôr objasniť presné významy týchto pojmov.

- *páchateľ* – osoba, ktorá spáchala trestný čin
- *podozrivý* – osoba, ktorej na základe dôkazov vyšetrovatelia pripisujú daný trestný čin
- *obvinený / obžalovaný* – osoba, ktorá je postavená pred súd za vykonanie daného trestného činu (t.j. podozrivý, proti ktorému má vyšetrovateľ najjasnejšie dôkazy)

Teraz si vysvetlíme, ako sa dá DNA využiť ako dôkazový materiál pri usvedčení podozrivého. V celej práci budeme predpokladať situáciu, že na mieste činu sa našla vzorka (kúsok krvi alebo vlas), ktorá nepatrí obeti, a teda zrejme patrí páchateľovi. Predpokladajme ďalej, že polícia na základe výpovedí svedkov alebo iných dôkazov nájde človeka, ktorého prehlási za podozrivého. Jeho DNA sa následne porovná so vzorkou nájdenou na mieste činu.

[1]Podrobnejšie vysvetlenie môže zvedavý čitateľ nájsť napríklad na stránke *en.wikipedia.org/wiki/Short_tandem_repeat*.

Nakoľko analýza nepreskúma celú DNA ale len niektoré jej časti, vzniká tu riziko *náhodnej zhody*, a síce toho, že podozrivý v skutočnosti nie je páchateľom, no jeho DNA sa v skúmaných locusoch zhoduje so vzorkou z miesta činu. Tento problém by sa dal vyriešiť analýzou celého DNA reťazca, no to je prakticky (aj z finančného hľadiska) v súčasnosti nemožné. V praxi sa preto aplikuje prístup, ktorý je založený na znalosti podmienenej pravdepodobnosti a matematickej štatistiky.

Úlohou tohto prístupu je nejakým spôsobom kvantifikovať silu (relevanciu) dôkazového DNA materiálu v prebiehajúcom vyšetrovaní a následnom súdnom procese. Táto sila sa nazýva *value of evidence* a dá sa presne definovať nasledovným vzorcom.

Definícia: *Value of evidence* je pomer vierohodnosti (likelihood ratio) v tvare

$$LR = \frac{P(E|H_p)}{P(E|H_d)},$$

kde E (evidence) je genotyp vzorky z miesta činu (predpokladaného páchateľa) a dvojica H_p a H_d sú hypotézy obžaloby (prosecution) a obhajoby (defence), s ktorými sa pracuje na súde. V najzákladnejšom prípade je to dvojica tvrdení *vzorka pochádza od obvineného* vs. *vzorka nepochádza od obvineného, jej zdrojom je iný človek.*

Aby sme to lepšie pochopili, prestavme si konkrétnu situáciu. Napríklad nech sa v trezore prepadnutej banky našiel vlas s genotypom X, ktorý[2] nepatrí žiadnemu zo zamestnancov, je teda predpoklad, že ho tam nechal zlodej. Polícia na základe opisu a výpovedí svedkov nájde muža, ktorého obviní a je vykonaná analýza DNA, ktorá v skúmaných locusoch prinesie pozitívny výsledok. Tento výsledok môžeme slovne interpretovať takto:

$$\frac{P(\text{genotyp vlasu je } X|\text{ zanechal ho tam obvinený})}{P(\text{genotyp vlasu je } X|\text{ nezanechal ho tam obvinený, ale iný človek})} \tag{1.1}$$

Čitateľ zlomku je v tomto prípade rovný jednej, lebo ak obvinenému bude odobratá DNA, bude sa zhodovať s vlasom s trezora (analýza to potvrdila). Na druhej strane výraz v menovateli zodpovedá pravdepodobnosti náhodnej zhody, t.j. pravdepodobnosti, že genotyp, ktorý pochádza od iného, náhodne vybraného človeka, sa zhoduje s genotypom páchateľa (vlasom). Ak hodnota tohto pomeru je k, znamená to, že je k-krát pravdepodobnejšie, že genotyp zhodný s vlasom má obvinený, ako to, že ho má náhodne vybraný človek. Ak je číslo k veľké, value of evidence tohto dôkazu je veľká (t.j. pravdepodobnosť náhodnej zhody je malá). Všimnime si, že k bude stále väčšie ako 1, nakoľko v čitateli je jednotka a v menovateli hodnota pravdepodobnosti, ktorá je určite menšia ako 1.

[2]pod X si, samozrejme, predstavíme dlhú postupnosť jednotlivých locusov a alel v nich

1.3 Predpoklady pre výpočet

Ako sme v predošlej kapitole naznačili, pri vyčísľovaní value of evidence budeme poväčšine počítať podmienené pravdepodobnosti. Tomuto bude venovaná prvá polovica práce, pričom si ukážeme možné komplikácie, ktoré môžu nastať (chýbajúci podozrivý, iná H_d ako tá štandardná atď.) a odvodíme vzorce, ktoré možno použiť v týchto prípadoch. Model ľudskej populácie, s ktorým budeme pracovať, a následný výpočet má však tri dôležité predpoklady, ktoré musíme na úvod spomenúť.

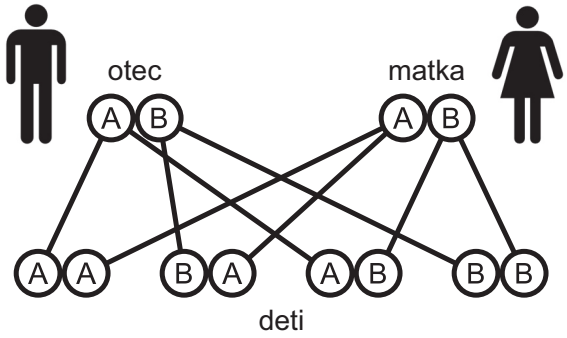
1. *Poznáme frekvencie výskytu jednotlivých alel.* V praxi tento predpoklad môžeme považovať za platný, znamená, že na základe dát, ktoré máme k dispozícii napríklad pre Slovenskú republiku, vieme odhadnúť frekvencie výskytu alel na locusoch u Slovákov. Napríklad odhadneme, že v locuse $TH01$ má 10% populácie alelu 9 a 8% populácie alelu 9.3.
2. *Linkage ekvilibrium* je ďalší platný predpoklad, znamená nezávislosť medzi jednotlivými locusmi. V praxi, ak napríklad locus $TH01$ obsahuje u nejakého jedinca alely 9 a 10, tak locus VWA u toho istého jedinca môže mať akékoľvek alely, ich rozdelenie je dané len ich frekvenciami a nie je nijako podmienené alelami na locuse $TH01$. Musíme ešte poznamenať, že druhov alel, ktoré sa môžu vyskytnúť na locuse, je len nejaký konkrétny počet (nie je to tak, že opakovaní môže byť ľubovoľne veľa – napríklad pri locuse $TH01$ sme v slovenskej vzorke našli len 6 možných alel – $6, 7, 8, 9, 9.3$ a 10).
3. *Hardyho – Weinbergovo ekvilibrium* hovorí, že populácia je nekonečná a výber partnera funguje náhodne. Inými slovami, výskyt alel (ktoré sú teda zdedené od náhodne vybraných rodičov) na locuse jedinca je nezávislý[3], t.j. ak jedna z alel na locuse $TH01$ je 9, tá druhá môže byť ktorákoľvek. Tento predpoklad v praxi úplne neplatí, bude mu preto venovaná druhá časť tejto práce. V nej „upravíme“ vzorce z prvej časti tak, aby platili aj pri porušení Hardyho – Weinbergovho ekvilibria.

Aby sme mohli pristúpiť k počítaniu, musíme si ešte uvedomiť dôležitú vec spomínanú v úvode práce a síce, že jednu z alel zdedí jedinec od matky a druhú od otca. To, ktorú alelu od rodiča zdedí, je náhodné, preto môže dostať ktorúkoľvek z jeho dvojice s pravdepodobnosťou 1/2. Tento fakt je veľmi dôležitý a budeme ho často využívať.

Keďže pracovať s locusmi v podobe názvov a alel v podobe čísel by bolo pre čitateľa veľmi neprehľadné, rozhodli sme sa to trochu zjednodušiť. V prvom rade budeme pracovať s genotypom, ktorý sa skladá z jedného locusu. Value of evidence by sme tak

[3]Všimnime si rozdiel oproti Linkage ekvilibriu – to hovorí o nezávislosti medzi locusmi, kým Hardyho – Weinbergovo ekvilibrium znamená nezávislosť dvojice alel v rámci jedného locusu.

vypočítali pre každý locus zvlášť a konečný výsledok dostaneme vynásobením výsledkov pre jednotlivé alely (môžeme tak urobiť vďaka linkage ekvilibriu). Ďalej, alely budeme označovať písmenami namiesto čísel. Budeme teda hovoriť, že osoba má genotyp napríklad (A, B) alebo (A, A). V prvom prípade, keď sú alely v genotype rôzne, budeme osobe hovoriť *heterozygot*, v druhom, keď sú alely rovnaké, ju nazveme *homozygot*. Na nasledujúcom obrázku vidíme, ako funguje spomínaná „dedičnosť“ alel, ktorej výsledkom je, že heterozygotný pár s genotypom (A, B) bude mať dieťa s genotypom (A, B) s pravdepodobnosťou $1/2$.



Obrázok 1.1. Schéma vzniku možných potomkov z páru heterozygotných rodičov

Kapitola 2

Výpočet value of evidence

V tejto kapitole si ukážeme, ako počítať value of evidence v rôznych zaujímavých situáciách a odvodíme vzorce, ktoré sa dajú všeobecne použiť. Podotýkame, že výsledné vzorce sme mali k dispozícii v [1], avšak boli tam uvedené bez akéhokoľvek náznaku odvodenia.

2.1 Základný prípad

Budeme postupovať ďalej pre príklad, ktorým sme začali v prvej kapitole. Nech teda genotyp vlasu nájdeného v trezore banky je (A, B). Analýza DNA podozrivého ukázala, že aj jeho genotyp je (A, B) a iného podozrivého nemáme, preto je dotyčný postavený pred súd. Poďme teraz vyčísliť hodnotu výrazu (1.1), ak na súde pracujeme s dvojicou klasických hypotéz *vlas pochádza od obvineného* (H_p) a *vlas nepochádza od obvineného, jeho zdrojom je iný človek* (H_d).

Výsledný pomer vierohodnosti bude

$$LR = \frac{P(\text{genotyp vlasu je } (A,B)|H_p)}{P(\text{genotyp vlasu je } (A,B)|H_d)} = \frac{1}{2f_A f_B}, \tag{2.1}$$

kde prítomnosť jednotky v čitateli sme si vysvetlili v minulej kapitole a menovateľ získame ako pravdepodobnosť, že genotyp náhodného človeka je (A, B), čiže pravdepodobnosť, že jedna jeho alela je A a druhá B, pričom na ich poradí nezáleží. Výraz v menovateli sme vyjadrili pomocou známych frekvencií výskytu daných alel, ktoré sme označili f_A a f_B. Všimnime si, že ak sú tieto frekvencie vysoké (veľa ľudí má genotyp obsahujúci alely A alebo B), výsledná value of evidence vyjde relatívne malá, čo zodpovedá vysokej pravdepodobnosti náhodnej zhody. Naopak, ak sú frekvencie malé, a teda ide o zriedkavý genotyp (t.j. pravdepodobnosť náhodnej zhody je maličká), výsledná value of evidence je veľká.

2.2 Prípad bez podozrivého

Po spáchaní trestného činu nie je neobvyklé, že podozrivý unikne polícii a teda nie je možné od neho odobrať vzorku DNA. Aj v tomto prípade však možno počítať value of evidence a to vtedy, ak vzorku DNA poskytnú pokrvní príbuzní podozrivého. Majme napríklad vzorku od rodičov podozrivého, ktorí majú obaja genotyp (A, B) a nech má takýto genotyp (naďalej) aj vlas z miesta činu. V takomto prípade sa nám zmení pôvodná H_p z *vlas pochádza od podozrivého* na *vlas pochádza od syna týchto rodičov*. Vo vzťahu (2.1) sa to prejaví v čitateli (v menovateli ostáva vždy pravdepodobnosť náhodnej zhody, čiže vôbec nezáleží na tom, aké vzorky DNA rodičov máme k dispozícii). V čitateli teda budeme hľadať pravdepodobnosť, že *vlas má genotyp* (A, B) *ak patrí synovi daných rodičov*, inými slovami pravdepodobnosť, že syn týchto rodičov bude mať genotyp (A, B). Situáciu sme si prehľadne znázornili na obrázku v závere časti 1.3, takže môžeme rovno dosadiť $1/2$ a dostaneme

$$LR = \frac{1/2}{2f_A f_B}.$$

Len pre doplnenie dodáme, že keby vlas z trezora mal genotyp (A, A), bol by čitateľ rovný $1/4$ (tiež z obrázka z konca prvej kapitoly) a menovateľ $(f_A)^2$, čo je pravdepodobnosť náhodnej zhody pre homozygota.

Ešte si všimneme zaujímavú vec. V praxi sa dá hneď pri analýze vzorky z miesta činu rozlíšiť pohlavie páchateľa, takže keby mali daní rodičia len jedného syna a jednu dcéru, na základe vzorky z trezora by sme vedeli jedného z nich vylúčiť. Keby však mali dvoch synov, obaja by boli podozriví v rovnakej miere (pre každého z nich platí, že je synom daného páru rodičov) a nie je možnosť, ako ich rozlíšiť (kým by sami neodovzdali vzorku svojej DNA).

Rodičia však nie sú jediná možnosť na odber DNA. Pomer vierohodnosti sa dá vyčísliť aj vtedy, ak máme napríklad vzorku od súrodenca podozrivého, dokonca aj vtedy, ak je ten súrodenec nevlastný (s podozrivým majú len jedného spoločného rodiča). Všeobecne môžeme povedať, že sa dá počítať so vzorkou DNA od akéhokoľvek pokrvného príbuzného. To, ako vypočítať podmienené pravdepodobnosti takéhoto typu a ako ďalej postupovať v zložitejšej situácii, si ukážeme v nasledujúcich častiach. V tejto chvíli si zapamätáme hlavne to, že pri chýbajúcom podozrivom sa nám oproti klasickému vzorcu (2.1) mení čitateľ.

2.3 Prípad inej alternatívnej hypotézy

Obhajoba sa nezakladá vždy na tvrdení, že namiesto obžalovaného spáchal trestný čin nejaký náhodný iný človek. Stáva sa, že obhajoba má k dispozícii „alternatívneho podozrivého“, ale nemá jeho vzorku DNA. Ak je ním osoba pokrvne príbuzná s pôvodným podozrivým, dá sa value of evidence prispôsobiť špeciálne pre danú situáciu. Postupne si ukážeme, ako vypočítať value of evidence pre príbuznosť na rôznych úrovniach.

2.3.1 Otec a syn

Pokračujme ďalej s prípadom lúpeže, kde máme k dispozícii vlas páchateľa s genotypom (A, B). Na prípravu na súdne pojednávanie máme k dispozícii dve hypotézy:

$$H_p : \text{vlas zanechal obvinený}$$

$$H_d : \text{vlas zanechal otec obvineného}$$

Genotyp podozrivého je tiež (A, B), no genotyp jeho otca nepoznáme. Vidíme, že v takomto prípade nebude v menovateli vzorca (2.1) vystupovať pravdepodobnosť náhodnej zhody, ale podmienená pravdepodobnosť v tvare

$$P(\text{genotyp stopy je } (A, B) \,|\, \text{vlas zanechal otec obvineného}),$$

inými slovami, pravdepodobnosť, že otec obvineného má genotyp (A, B), ak jeho syn (obvinený) má genotyp (A, B). Označíme si udalosť *genotyp otca je* (A, B) ako O a *genotyp syna je* (A, B) ako S. Všimnime si, že nevieme nič o matke, preto ju budeme pokladať za „náhodnú premennú“ a jej neznámy genotyp označíme μ. Pri odvodení vzorca pre value of evidence budeme vychádzať z klasickej Bayesovej vety o podmienenej pravdepodobnosti

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)} = \sum_i P(X|Y \wedge Z_i)P(Z_i),$$

kde $\sum_i P(Z_i) = 1$ a všetky Z_i sú navzájom disjunktné. Teraz a aj v ďalších prípadoch nám pomôže nasledujúca tabuľka. V riadku a stĺpci sa nachádzajú genotypy rodičov a hodnoty v nej sú pravdepodobnosti, že daný pár bude mať potomka s genotypom (A, B). Okrem alel A a B v nej vystupuje aj alela X zahŕňajúca všetky ostatné alely, ktoré sa okrem A a B môžu na skúmanom locuse vyskytnúť. Len pre úplnosť, frekvencia alely X je $1 - f_A - f_B$.

Otec/Matka	(A,A)	(A,B)	(B,B)	(A,X)	(B,X)	(X,X)
(A,A)	0	1/2	1	0	1/2	0
(A,B)	1/2	1/2	1/2	1/4	1/4	0
(B,B)	1	1/2	0	1/2	0	0
(A,X)	0	1/4	1/2	0	1/4	0
(B,X)	1/2	1/4	0	1/4	0	0
(X,X)	0	0	0	0	0	0

Tabuľka 2.1. Pravdepodobnosť vzniku potomka s genotypom (A,B)

Najskôr si prepíšeme Bayesovu vetu do podoby, v akej ju potrebujeme. Výrazy $P(S)$ a $P(O)$ majú rovnakú hodnotu $(2f_Af_B)$, preto ich môžeme vykrátiť a dostaneme

$$P(O|S) = \frac{P(S|O)P(O)}{P(S)} = P(S|O).$$

Toto je zaujímavý poznatok, ktorý ešte viackrát využijeme. Je to spôsob, ako sa dá $P(O|S)$ (ktorú nevieme priamo vypočítať) previesť na $P(S|O)$, ktorá sa počíta ľahko. Uvedomme si, že ak poznáme genotyp otca aj matky, vieme priamo vypočítať pravdepodobnosť akéhokoľvek genotypu dieťaťa. Keďže v tomto prípade genotyp matky nepoznáme, vypočítame pravdepodobnosť, že ich syn bude mať genotyp (A,B) ako sumu pravdepodobností cez možné genotypy matky (vychádzame z vety o úplnej pravdepodobnosti). Dostaneme

$$P(O|S) = P(S|O) = \sum_i P(S|O \wedge \mu_i)P(\mu_i), \tag{2.2}$$

kde μ_i je genotyp matky a $P(\mu_i)$ pravdepodobnosť jeho výskytu. Teraz už stačí len hodnoty z tabuľky dosadiť do vzorca (2.2), pričom nás bude zaujímať len riadok, kde genotyp otca je (A,B). Pravdepodobnosti výskytu jednotlivých genotypov matky dosádzame klasicky pre homozygota ako druhú mocninu frekvencie výskytu alely a pre heterozygota, napríklad (A,X), ako $2f_A(1-f_A-f_B)$. Výsledná pravdepodobnosť $P(O|S)$, resp. $P(O|S)$ bude

$$\frac{1}{2}f_A^2 + \frac{1}{2}\cdot 2f_Af_B + \frac{1}{2}f_B^2 + \frac{1}{4}\cdot 2f_A(1-f_A-f_B) + \frac{1}{4}\cdot 2f_B(1-f_A-f_B) =$$

$$= \frac{1}{2}f_A + \frac{1}{2}f_B.$$

Teraz si všimnime, že sme vlastne nevyriešili jednu úlohu, ale dve, pretože poznáme aj vzorec pre situáciu, keď hypotéza obhajoby H_d bude, že *vlas zanechal syn obvineného.*

V tomto prípade vieme totiž priamo dosadiť

$$P(S|O) = \sum_i P(S|O \wedge \mu_i)P(\mu_i).$$

Ešte sa pozrime na to, ako by vyzerala situácia, keby dôkazový materiál (vlas) a teda aj podozrivý mal homozygotný genotyp (A, A). Vtedy nemusíme brať vôbec do úvahy alelu B a môžeme ju zahrnúť do X, ktorej frekvencia preto bude $(1 - f_A)$. Tabuľka, s ktorou budeme pracovať, vyzerá takto (existujú len dva genotypy matky, ktoré nás reálne zaujímajú):

Otec/Matka	(A, A)	(A, X)	(X, X)
(A, A)	1	1/2	0
(A, X)	1/2	1/4	0
(X, X)	0	0	0

Tabuľka 2.2. Pravdepodobnosť vzniku potomka s genotypom (A, A)

Po dosadení do vzorca (2.2) dostaneme

$$P(O|S) = P(S|O) = 1f_A^2 + \frac{1}{2} \cdot 2f_A(1 - f_A) = f_A.$$

Pre opakovanie ešte dodáme, že výsledná value of evidence má v tvar $1/P(O|S)$ ak H_d je *vlas zanechal otec obvineného* resp. $1/P(S|O)$ ak H_d je *vlas zanechal syn obvineného*. Oba výsledky $P(O|S)$ pre homozygota aj pre heterozygota súhlasia s výsledkami nájdenými v [1], kde však (ako sme na začiatku kapitoly spomenuli) boli uvedené len výsledné hodnoty.

2.3.2 Bratia

Teraz sa pozrieme na prípad, keď alternatívnym podozrivým je brat. Postupovať budeme rovnako ako pri otcovi a synovi, teda potrebujeme odvodiť pravdepodobnosť, že ak podozrivý má nejaký daný genotyp, tak aj jeho brat bude mať rovnaký. Nech hľadaný genotyp je heterozygot (A, B). Pre homozygota bude platiť ten istý vzorec, len dosadenie pomocou tabuľky bude vyzerať inak. Znova budeme vychádzať zo základnej podmienenej pravdepodobnosti danej vzťahom

$$P(B|S) = \frac{P(B \cap S)}{P(S)}, \tag{2.3}$$

pričom len dopočítame jednotlivé časti. Označenie je analogické ako v minulom prípade, B znamená *genotyp brata je* (A, B) a S bude označovať udalosť *genotyp podozrivého*

(suspect) je (A, B). V menovateli sa nachádza $P(S)$, čo je v tomto prípade $2f_A f_B$. V čitateli máme pravdepodobnosť, že genotyp podozrivého a jeho brata zároveň je (A, B).

Kľúčovou myšlienkou, z ktorej budeme vychádzať, je, že bratia majú spoločných rodičov, teda dostali každý jednu alelu od rovnakej matky aj otca. Ďalej to, ktorú z dvoch alel od rodiča dostali, je náhodné. Inými slovami, ak napríklad matka bola (A, B) a jeden z bratov od nej dostal A, nevieme povedať, čo od nej dostal druhý brat (ani jedna z možností $\{A, B\}$ nie je viac pravdepodobná ako tá druhá). Rovnako je to aj pri otcovi.

Ak máme genotypy rodičov, vieme vyrátať, s akou pravdepodobnosťou z nich vznikne syn s nejakým genotypom. Zoberme si teda jednu dvojicu matka – otec a považujme ich genotypy za pevné. Ak túto dvojicu označíme ϕ_i, pravdepodobnosť, že z nich vznikne syn s genotypom (A, B) bude $P(S|\phi_i)$, resp. $P(B|\phi_i)$. Navyše z predošlého odstavca vieme, že vznik genotypov dvoch detí z konkrétneho páru rodičov je nezávislý, takže vieme povedať, že $P(S \cap B|\phi_i) = P(S|\phi_i)P(B|\phi_i)$.

Určili sme pravdepodobnosť, že z nejakého konkrétneho páru rodičov vzniknú dve deti s rovnakým genotypom. Ak chceme spočítať pravdepodobnosť, že vzniknú dve deti s rovnakým genotypom, musíme (podľa vety o úplnej pravdepodobnosti) sčítať pravdepodobnosti vzniku cez všetky možné genotypy rodičov. Dostaneme

$$P(B \cap S) = \sum_i P(B|\phi_i)P(S|\phi_i)P(\phi_i).$$

Toto už len dosadíme do vzorca (2.3) a dostaneme výsledný tvar

$$P(B|S) = \frac{P(B \cap S)}{P(S)} = \frac{\sum_i P(B|\phi_i)P(S|\phi_i)P(\phi_i)}{P(S)} = \\ = \frac{\sum_i P(B|\phi_i)^2 P(\phi_i)}{P(S)}. \tag{2.4}$$

Aby sme výpočet dokončili, musíme ešte dosadiť príslušné pravdepodobnosti. Možné genotypy rodičov a pravdepodobnosti toho, že kombináciou týchto genotypov vznikne dieťa s genotypom (A, B) sú uvedené v tabuľke 2.1. Za pravdepodobnosť $P(\phi_i)$ výskytu jednotlivých genotypov matky a otca budeme dosádzať klasicky druhú mocninu frekvencie alely, ak je rodič homozygot, a $2f_A f_B$ ak je rodič heterozygot s genotypom (A, B). Výslednú pravdepodobnosť páru rodičov dostaneme vynásobením pravdepodobností výskytu otca a matky, napr. pravdepodobnosť, že otec bol (A, A) a matka (A, B) je $4f_A^3 f_B$. Po dosadení dostaneme

$$P(B|S) \cdot (2f_A f_B) = \frac{1}{4} \cdot 2f_A f_B \cdot f_A^2 + 1 \cdot f_B^2 f_A^2 + \frac{1}{4} \cdot 2f_B(1 - f_A - f_B) \cdot f_A^2 +$$

$$+\frac{1}{4}f_A^2 \cdot 2f_Af_B + \frac{1}{4} \cdot 2f_Af_B \cdot 2f_Af_B + \frac{1}{4}f_B^2 \cdot 2f_Af_B + \frac{1}{16} \cdot 2f_A(1-f_A-f_B) \cdot 2f_Af_B +$$
$$+\frac{1}{16} \cdot 2f_B(1-f_A-f_B) \cdot 2f_Af_B + 1 \cdot f_A^2f_B^2 + \frac{1}{4} \cdot 2f_Af_B \cdot f_B^2 + \frac{1}{4} \cdot 2f_A(1-f_A-f_B) \cdot f_B^2 +$$
$$+\frac{1}{16} \cdot 2f_Af_B \cdot 2f_A(1-f_A-f_B) + \frac{1}{4}f_B^2 \cdot 2f_A(1-f_A-f_B) +$$
$$+\frac{1}{16} \cdot 2f_B(1-f_A-f_B) \cdot 2f_A(1-f_A-f_B) + \frac{1}{4} \cdot f_A^2 \cdot 2f_B(1-f_A-f_B) +$$
$$+\frac{1}{16} \cdot 2f_Af_B \cdot 2f_B(1-f_A-f_B) + \frac{1}{16} \cdot 2f_A(1-f_A-f_B) \cdot 2f_B(1-f_A-f_B)$$

a po následnej úprave získame výslednú pravdepodobnosť v tvare

$$P(B|S) = \frac{1}{4}(1 + f_A + f_B + 2f_Af_B).$$

Keď hľadaný genotyp je homozygot (A, A), dosádzame tiež do vzorca (2.4), no použijeme tabuľku (2.2). Dostaneme

$$P(B|S) \cdot f_A^2 = 1 \cdot f_A^2f_A^2 + \frac{1}{4} \cdot 2f_A(1-f_A) \cdot f_A^2 + \frac{1}{4}f_A^2 \cdot 2f_A(1-f_A) +$$
$$+\frac{1}{16} \cdot 2f_A(1-f_A) \cdot f_A(1-f_A)$$

a teda

$$P(B|S) = \frac{1}{4}(1 + 2f_A + f_A^2).$$

2.3.3 Nevlastný brat

Skúmajme ďalej situáciu s genotypom (A, A), ak obvinený a jeho brat majú spoločného len jedného rodiča (nech je to otec). Označíme si udalosť *genotyp nevlastného brata je* (A, A) ako N a *genotyp podozrivého je* (A, A) ako S. Možné genotypy ich spoločného otca označíme ν_i. O matkách nevieme nič, len to, že obe dali svojmu dieťaťu alelu A (túto udalosť označíme M_1, resp. M_2 pre jednotlivé matky, kde prvá je matka podozrivého). Rovnako vieme, že aj otec dal obom synom alelu A. Pravdepodobnosť, že obvinený (alebo jeho brat) má genotyp (A, A) je teda súčin pravdepodobností, že mu otec a matka dali alelu A. Pravdepodobnosť, že otec dal synovi alelu A, závisí od jeho genotypu. Ak je genotyp pevný, tak pravdepodobnosť, že otec dal podozrivému alelu A je $P(S|\nu_i)$. Keďže o ani jednej z matiek nevieme nič, pravdepodobnosť, že dali synovi alelu A sa rovná pravdepodobnosti, že sa alela A vyskytne v náhodnom genotype, a teda je to f_A. Potom pravdepodobnosť, že otec aj matka dali podozrivému každý alelu A je $P(S|\nu_i \wedge M_1)f_A$. Naviac, medzi nevlastnými bratmi platí rovnaká nezávislosť ako v predošlej časti. Ak pokladáme genotyp otca za pevný, dostaneme

$$P(N \cap S|\nu_i) = P(N|\nu_i \wedge M_2)f_AP(S|\nu_i \wedge M_1)f_A,$$

čo po zosumovaní cez všetky možné genotypy otca dáva výraz

$$P(N \cap S) = f_A^2 \sum_i P(N|\nu_i \wedge M_2)P(S|\nu_i \wedge M_1)P(\nu_i),$$

ktorý po dosadení do (2.3) dá výsledný vzorec

$$P(N|S) = \frac{f_A^2 \sum_i P(N|\nu_i \wedge M_1)P(S|\nu_i \wedge M_1)P(\nu_i)}{P(S)}. \tag{2.5}$$

Aby sme dostali vzťah, kde budú vystupovať len frekvencie alel, dosadíme ešte hodnoty z tabuľky 2.2. Hodnota podmienenej pravdepodobnosti pre nevlastného brata v prípade homozygota je teda

$$P(N|S) = \frac{1}{f_A^2} f_A^2 \left[1 \cdot f_A^2 + \frac{1}{4} \cdot 2f_A(1 - f_A)\right] = \frac{1}{2}(f_A + f_A^2).$$

V prípade heterozygota s genotypom (A, B) bude dosádzanie trochu komplikovanejšie, pretože budeme musieť zvážiť prípady, keď matka (aj otec) dali synom alelu A alebo B. Prehľadný rozbor prípadov, ktoré môžu nastať, je v nasledujúcej tabuľke.

Otec	P(jeho výskytu)	alely, ktoré dá synom	P(že dá tie alely)	alely, ktoré dajú matky	súčin frekvencií
(A, A)	f_A^2	A,A	1	B,B	f_B^2
(B, B)	f_B^2	B,B	1	A,A	f_A^2
(A, X)	$2f_A(1 - f_A - f_B)$	A,A	1/4	B,B	f_B^2
(B, X)	$2f_B(1 - f_A - f_B)$	B,B	1/4	A,A	f_A^2
(A, B)	$2f_A f_B$	A,A	1/4	B,B	f_B^2
(A, B)	$2f_A f_B$	B,B	1/4	A,A	f_A^2
(A, B)	$2f_A f_B$	A,B	1/4	B,A	$f_B f_A$
(A, B)	$2f_A f_B$	B,A	1/4	A,B	$f_A f_B$

Aby sme dostali výsledný tvar vzorca na podmienenú pravdepodobnosť pre nevlastného brata s genotypom (A, B), stačí sčítať členy, ktoré vzniknú ako súčin stĺpcov tejto tabuľky, začínajúc prvým riadkom $1f_A^2 f_B^2$ a končiac posledným v tvare $2f_A^2 f_B^2/4$. Po sčítaní a úprave dostaneme, že

$$P(B|S) = f_A f_B + \frac{1}{4}(f_A + f_B).$$

Všimnime si, že symetria, ktorá platila v podmienených pravdepodobnostiach pri synovi a otcovi tu platí tiež, čiže

$$P(N|S) = \frac{P(S|N)P(N)}{P(S)} = P(S|N),$$

nakoľko hodnoty $P(S)$ a $P(N)$ sa rovnajú a teda sa vykrátia. Logický dôvod, prečo táto symetria platí je aj ten, že nevlastní bratia sú v rodostrome na rovnakej úrovni.

2.3.4 Starý otec a vnuk

Rozhodujúce v tejto situácii bude to, akým spôsobom dedí vnuk alely od starých rodičov. Uvedomme si, že dedí jednu zo štyroch alel, ktoré majú dokopy jeho starí rodičia (nech sú to starí rodičia z otcovej strany) a druhú alelu získa od matky. Tá je v tomto prípade neznáma, teda bude rovnako ako v predošlom prípade reprezentovaná len frekvenciou alely, ktorú dala svojmu potomkovi. Majme znova najskôr ľahší prípad homozygota s genotypom (A, A) a pozrime sa bližšie na spomínaných starých rodičov. Odvodíme si tabuľku pravdepodobností, z ktorej teraz využijeme len časť, no neskôr sa nám ešte zíde. Výsledná tabuľka vyzerá takto:

počet alel A spolu v genotype starých rodičov	P(vnuk od nich dostane A)
1	1/4
2	1/2
3	3/4
4	4/4

Tabuľka 2.3. Pravdepodobnosť zdedenia alely A od starých rodičov

Uvedené pravdepodobnosti dostaneme nasledovne: ak majú starí rodičia spolu

- 1 alelu A, pravdepodobnosť, že ju dostane ich syn, je $1/2$ a že ju dostane ich vnuk je potom $(1/2)^2 = 1/4$.
- 2 alely A, môžu mať genotypy (A, X) a (A, X) alebo (A, A) a (X, X) s rovnakou pravdepodobnosťou. Ich syn môže byť v prvom prípade (A, A), (A, X), (X, A) alebo (X, X) (každé s pravdepodobnosťou $1/4$) a v druhom prípade len (A, X). Potom pravdepodobnosť, že vnuk dostane alelu A je

$$\frac{1}{2}\left(\frac{1}{4}(1+\frac{1}{2}+\frac{1}{2}+0)+\frac{1}{2}\right)=\frac{1}{2}.$$

- 3 alely A, majú genotypy (A, A) a (A, X). Ich syn môže mať s rovnakou pravdepodobnosťou genotyp (A, A) alebo (A, A). Teda pravdepodobnosť, že ich vnuk dostane od ich syna alelu A je

$$\frac{1}{2}\left(1+\frac{1}{2}\right)=\frac{3}{4}.$$

- 4 alely A (obaja sú homozygoti), ich syn musí byť homozygot a ich vnuk od neho určite dostane A, lebo iné mu ani dať nemôže.

Teraz môžeme pristúpiť k odvodeniu vzorca. Označíme si klasicky udalosť *genotyp vnuka je* (A, A) ako V a *genotyp starého otca je* (A, A) ako D (dedko). Potom platí rovnaká symetria ako v prípade s otcom a synom, a síce

$$P(D|V) = \frac{P(V|D)P(D)}{P(V)} = P(V|D),$$

takže zase riešime dve úlohy naraz. Prvou úlohou je výpočet podmienenej pravdepodobnosti prislúchajúci H_d v tvare *škvrnu zanechal starý otec obvineného* a druhou výpočet pre prípad, keď H_d je *škvrnu zanechal vnuk obvineného*. Symetria opäť platí vďaka rovnosti výrazov $P(D)$ a $P(V)$.

Keďže genotyp starého otca máme pevne daný ako (A, A) a genotyp matky reprezentujeme frekvenciou, jedinou premenlivou časťou bude genotyp starej mamy, ktorý označíme μ_i. Ďalej si označme udalosť *vnuk zdedil od starých rodičov alelu* A ako $\hat{V}$. Potom $P(\hat{V}|D \wedge \mu_i)$ je pravdepodobnosť, že vnuk zdedí od starého otca a konkrétnej starej mamy s genotypom μ_i alelu A. S výpočtom pravdepodobnosti, že vnuk zdedí alelu A od konkrétneho páru starých rodičov nám pomôže tabuľka 2.3.

Táto situácia je veľmi podobná príkladu so synom a otcom, lebo platí

$$P(D|V) = P(V|D) = f_A \sum_i P(\hat{V}|D \wedge \mu_i)P(\mu_i), \tag{2.6}$$

kde f_A reprezentuje pravdepodobnosť, že vnuk dostal alelu A od svojej matky a suma reprezentuje úplnú pravdepodobnosť toho, že dostal alelu A aj od svojich starých rodičov. Konkrétne v prípade homozygota pri dosádzaní do tohto vzorca použijeme nasledujúce prípady (pamätajme na to, že genotyp starého otca poznáme):

genotyp starej mamy μ_i	(A, A)	(A, X)	(X, X)
P(vnuk dostane od starých rodičov alelu A)	1	3/4	1/2

Výsledný vzorec bude mať tvar

$$P(D|V) = f_A \left[1 \cdot f_A^2 + \frac{3}{4} \cdot 2f_A(1 - f_A) + \frac{1}{2} \cdot (1 - f_A)^2\right] =$$

$$= \frac{1}{2}(f_A + f_A^2).$$

V prípade heterozygota s genotypom (A, B) je situácia trochu komplikovanejšia. Vzorec (2.6) bude totiž obsahovať dve sumy, pretože starí rodičia môžu dať vnukovi alelu A alebo B (matka potom „doplní“ tú druhú). Máme teda pevný genotyp starého otca (A, B) a nasledovné pravdepodobnosti:

genotyp starej mamy	(A, A)	(A, B)	(B, B)	(A, X)	(B, X)	(X, X)
P(vnuk zdedí A)	3/4	1/2	1/4	1/2	1/4	1/4
P(vnuk zdedí B)	1/4	1/2	3/4	1/4	1/2	1/4

Potom po dosadení do vzorca[1] v tvare

$$P(D|V) = f_B \sum_i P(\hat{V}|D \wedge \mu_i)P(\mu_i) + f_A \sum_j P(\hat{V}|D \wedge \mu_j)P(\mu_j) \tag{2.7}$$

a následnej úprave dostaneme výslednú hodnotu podmienenej pravdepodobnosti

$$P(D|V) = P(V|D) = f_A f_B + \frac{1}{4}(f_A + f_B)\cdot \tag{2.8}$$

Táto sa zhoduje s výsledkom v [1], no celé jej odvodenie je originálne (ako aj všetky ostatné odvodenia v tejto časti práce).

2.3.5 Bratranci

Túto úlohu už nebudeme riešiť tak podrobne ako predošlé, lebo v nej len postupne použijeme „triky“, ktoré sme používali doteraz a trochu tým precvičíme čitateľovu pozornosť. V situácii, kde alternatívnym podozrivým je bratranec, budeme teda najskôr hľadať paralely s už vyriešenými príkladmi.

- bratranci sú v rodostrome navzájom na „rovnakej úrovni“, takže vzorec sa bude nápadne podobať na prípad s bratmi (platí aj rovnaká nezávislosť pri dedení alel)
- spoločný majú zdroj jednej alely – starých rodičov, pričom pravdepodobnosti, s akými od nich dedia alely už poznáme z prípadu vnuk – dedko
- zdroje druhých alel sú rovnaké ako pri nevlastnom bratovi – sú nimi náhodné matky, ktorých prítomnosť vo vzorci nahradíme frekvenciou alely, ktorú potomkovi dali (f_A, resp. f_B)

Budeme postupovať rovnako ako doteraz, čiže vychádzame zo základného vzorca (2.3). Označíme *podozrivý zdedil od starých rodičov alelu* (A) ako $\hat{S}$ a *bratranec podozrivého zdedil od starých rodičov alelu* A ako $\hat{B}$. Možné genotypy starých rodičov označíme ϕ_i a predpokladáme, že bratranci sú príbuzní cez otcov (ako sme už podotkli, matky sú neznáme).

Poďme teda vypočítať $P(B|S)$, pričom sa pozrieme na vznik genotypov bratrancov ako na navzájom nezávislé udalosti (pri pevnom páre spoločných starých rodičov) ako to

[1] Indexy i prislúchajú genotypom v situácii, keď vnuk dedí od starých rodičov alelu A a indexy j keď dedí alelu B.

bolo pri bratoch. Predpokladajme, že sú homozygoti, bude nás teda zaujímať len alela A a pravdepodobnosť, že ju dostanú od svojho otca (ktorý ju dostal od svojich rodičov). Toto máme prehľadne zhrnuté v tabuľke 2.3 pre všetky možné genotypy starých rodičov. Pravdepodobnosť získania alely A od matiek nahradíme frekvenciou výskytu alely A. Potom ak uvažujeme pevný pár starých rodičov s genotypom ϕ_i, pravdepodobnosť vzniku podozrivého s genotypom (A, A) ak je to ich vnuk je $P(\hat{S}|\phi_i)f_A$, kde f_A je pravdepodobnosť, že jeho matka mu dala alelu A. Potom pravdepodobnosť vzniku dvoch vnukov s rovnakým genotypom je

$$P(B \cap S|\phi_i) = P(\hat{B}|\phi_i)f_A P(\hat{S}|\phi_i)f_A,$$

čo po zosumovaní podľa vety o úplnej pravdepodobnosti (ako v prípade bratov) dáva

$$P(B \cap S) = f_A^2 \sum_i P(\hat{B}|\phi_i)P(\hat{S}|\phi_i)P(\phi_i).$$

Po doplnení do vzorca (2.3) dostaneme podmienenú pravdepodobnosť v tvare

$$P(B|S) = \frac{f_A^2 \sum_i P(\hat{B}|\phi_i)P(\hat{S}|\phi_i)P(\phi_i)}{P(S)} \tag{2.9}$$

a následným dosadením frekvencií a pravdepodobností získame finálny tvar

$$P(B|S) = f_A^2 + f_A(1 - f_A)/4,$$

pričom sme sa riadili nasledovnou tabuľkou:

Dedko/Babka	(A, A)	(A, X)	(X, X)
(A, A)	1	3/4	1/2
(A, X)	3/4	1/2	1/4
(X, X)	1/2	1/4	0

Tabuľka 2.4. Pravdepodobnosť zdedenia alely A od starých rodičov

V prípade bratrancov s heterozygotným genotypom (A, B) budeme postupovať ako v predošlom prípade, t.j. rozdelíme si situáciu na viaceré sumy podľa toho, ktoré alely dávajú bratrancom matky a ktoré spoloční starí rodičia. Po dosadení príslušných pravdepodobností a nasledovnej úprave dostaneme výsledok

$$P(B|S) = 2f_A f_B + \frac{1}{8}(f_A + f_B - 4f_A f_B)\cdot$$

2.3.6 Synovec a strýko

Podobne ako pri bratrancovi aj tu budeme vychádzať z poznatkov, ktoré sme už použili. Najdôležitejšie je nájsť spojenie medzi strýkom a synovcom čo sa týka zdrojov alel. Predstavme si znova situáciu, že strýko a synovec sú príbuzní cez chlapcovho otca. To znamená, že chlapcova matka je neznáma a jeho otec a strýko majú spoločných rodičov. Znova budeme teda pracovať so starými rodičmi, lebo strýko má v genotype dve z ich štyroch alel a synovec má jednu (druhú nahradíme znova len jej frekvenciou). Začnime znova ľahším, homozygotným prípadom a označme si *synovec má genotyp* (A, A) *ako* S a *strýko má genotyp* (A, A) *ako* U *(ujo)*. Ak považujeme pár starých rodičov (označme jeho genotyp znova ϕ_i) za pevný, pravdepodobnosť, že vznikne syn a zároveň vnuk (ktorý nie je jeho syn) s genotypom (A, A), je

$$P(S \cap U|\phi_i) = P(\hat{S}|\phi_i) f_A P(U|\phi_i),$$

pričom $P(\hat{S}|\phi_i)$ je pravdepodobnosť, že vnuk zdedil od starých rodičov alelu A, f_A je pravdepodobnosť, že matka (ktorá je neznáma) mu dala tiež alelu A a $P(U|\phi_i)$ je pravdepodobosť, že syn daných starých rodičov (*strýko*) má genotyp (A, A). Po zosumovaní cez možné genotypy starých rodičov podľa vety o úplnej pravdepodobnosti (ako v prípade bratrancov a bratov) dostaneme

$$P(S \cap U) = f_A \sum_i P(\hat{S}|\phi_i) P(U|\phi_i) P(\phi_i).$$

Po doplnení do vzorca (2.3) získame výslednú podmienenú pravdepodobnosť v tvare

$$P(S|U) = \frac{f_A \sum_i P(\hat{S}|\phi_i) P(U|\phi_i) P(\phi_i)}{P(U)}, \tag{2.10}$$

do ktorej už stačí len dosadiť kombináciu pravdepodobností z tabuliek 2.2 a 2.4. Po úprave dostaneme

$$P(S|U) = \frac{f_A}{f_A^2} \left[1 \cdot f_A^4 + \frac{1}{2} \cdot \frac{1}{4} \cdot 4 f_A^2 (1 - f_A)^2 + 2 \cdot (\frac{3}{4} \cdot \frac{1}{2} \cdot 2 f_A^3 (1 - f_A)) \right] =$$

$$= \frac{1}{2}(f_A + f_A^2).$$

V prípade heterozygotných genotypov synovca a strýka budeme postupovať klasicky – rozdelíme vzorec (2.10) na viac súm a dostaneme

$$P(S|U) = f_A f_B + \frac{1}{4}(f_A + f_B).$$

Záver časti 2.3

Na záver sa vráťme k prípadu bez podozrivého, ktorým sme sa zaoberali v časti 2.2. Podmienené pravdepodobnosti, ktoré sme tu vypočítali, možno použiť aj v spomínanej 2.2, konkrétne pri zmene v čitateli value of evidence. Modifikáciu týchto vzorcov použijeme aj v nasledujúcej časti, ktorá sa bude zaoberať prípadmi, keď nenastane úplná, ale len čiastočná zhoda vzorky podozrivého a páchateľa.

2.4 Prípad čiastočnej zhody

Ako sme už naznačili, táto časť sa zaoberá prípadom, keď sa vzorka z miesta činu nezhoduje úplne s genotypom podozrivého. On je teda ako páchateľ vylúčený, no je pravdepodobné (ak odchýlka je malá[2]), že sa so vzorkou z miesta činu bude zhodovať niektorý jeho pokrvný príbuzný. Ak získame vzorku DNA od daného príbuzného, postupujeme ako v predošlých príkladoch, pričom príbuzný sa v tomto prípade stane podozrivým. Ak sa vzorku získať nepodarí, vieme na základe vzorky od pôvodného podozrivého vypočítať pravdepodobnosť, že niektorý konkrétny jeho pokrvný príbuzný sa bude úplne zhodovať so vzorkou z miesta činu.

Znova si to ukážeme na príklade. Predpokladajme, že sa stala vražda, kde sa za nechtami obete (ženy) našli kúsky kože, ktoré zrejme pochádzajú od páchateľa. Nech genotyp získaný z tejto kože je (A, A). Na základe výpovedí svedkov podozrieva polícia jej manžela a je mu odobratá DNA. Jeho genotyp je však (A, B), preto je ako páchateľ vylúčený. Rozdiel genotypov je však „malý“, takže je možné, že páchateľom je osoba pokrvne príbuzná s manželom, pričom podľa svedkov ako alternatívny podozrivý prichádza do úvahy jeho brat (švagor obete). Ten ale náhle odišiel z mesta, takže odber vzorky DNA je nemožný.

Mohli by sme postupovať ako na začiatku kapitoly v prípade s chýbajúcim podozrivým, ak by sme mali k dispozícii napríklad rodičov. Predpokladajme ale, že manžel obete a jeho brat sú jediní, od koho je možné odobrať vzorku DNA (iní pokrvní príbuzní neexistujú). Vieme síce vypočítať pravdepodobnosť, že švagor obete má rovnaký genotyp ako manžel, ale to nám tu nepomôže. Na základe rovnakej idey (ako v spomínanom prípade rovnakých genotypov) však vieme odvodiť vzorec, ktorý rieši našu situáciu. Odvodenie, ktoré bude nasledovať, je možné nájsť aj v knihe [2].

Majme teda k dispozícii stopu s genotypom (A, A) a podozrivého s genotypom (A, B).

[2]Nezabúdajme na to, že v praxi pracujeme s viacerými locusmi, nie len s jedným.

V takomto prípade máme na súde klasickú dvojicu hypotéz:

$$H_p : \text{škvrnu zanechal príbuzný podozrivého}$$

$$H_d : \text{škvrnu zanechal náhodný muž}$$

Odvodenie tvaru pomeru vierohodnosti (value of evidence) urobíme všeobecne pre príbuzného. V úlohách, ktoré budú nasledovať, naznačíme výpočet pre konkrétnych členov rodiny. Označíme *genotyp podozrivého je* (A, B) ako S (suspect), *genotyp dôkazového materiálu je* (A, A) ako E (evidence) a *genotyp príbuzného je* (A, A) ako R (relative). Pomer vierohodnosti v tomto prípade napíšeme ako

$$LR = \frac{P(S, E|H_p)}{P(S, E|H_d)}.$$

Môžeme ho ďalej upraviť na súčin dvoch zlomkov a prepísať do tvaru

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \frac{P(S|E, H_p)}{P(S|E, H_d)} = \frac{P(S|R)}{P(S)}.$$

Úpravy, ktoré sme urobili, musíme samozrejme odôvodniť:

- prvý zlomok súčinu môžeme položiť rovný 1, lebo bez poznania genotypu obvineného a toho, či vôbec má brata, je pravdepodobnosť genotypu dôkazu rovnaká, nech ju podmienime čímkoľvek
- čitateľ druhého zlomku môžeme napísať skrátene v tvare $P(S|R)$, lebo vieme (z H_p), že zdrojom dôkazového materiálu je príbuzný obvineného a dôkaz má zároveň genotyp (A, A) – preto do podmienky stačí dať, že „genotyp príbuzného je (A, A)“
- v menovateli druhého zlomku máme H_d a teda neexistuje spojitosť medzi obvineným a tým, kto zanechal dôkazový materiál – preto môžeme podmienku celkom vypustiť

Vráťme sa teraz k vražde, ktorú sme uviedli ako príklad. Budeme počítať pravdepodobnosť, že s genotypom stopy sa zhoduje genotyp brata podozrivého. Pomôže nám vzorec (2.4), ktorý sme odvodili pre alternatívnu hypotézu, že páchateľom je brat, ktorý má tvar

$$P(B|S) = \frac{P(B \cap S)}{P(S)} = \frac{\sum_i P(B|\phi_i)P(S|\phi_i)P(\phi_i)}{P(S)}.$$

Do tohto vzorca budeme dosádzať presne tak, ako v spomínanom prípade, až na to, že tu bratia majú rôzny genotyp. Možné genotypy rodičov, pravdepodobnosť ich výskytu a pravdepodobnosť vzniku genotypu bratov je v nasledujúcej tabuľke. (Môžeme vynechať genotypy, kde z žiadny z rodičov neobsahuje A, lebo vtedy nemôže vzniknúť syn $((A, A))$.

Otec	Matka	P(ich výskytu)	P(syn je (A, A))	P(syn je (A, B))
(A, A)	(A, A)	f_A^4	1	0
(A, A)	(A, B)	$2f_A^3 f_B$	1/2	1/2
(A, A)	(A, X)	$2(1 - f_A - f_B)f_A^3$	1/2	0
(A, B)	(A, A)	$2f_A^3 f_B$	1/2	1/2
(A, B)	(A, B)	$4f_A^2 f_B^2$	1/4	1/2
(A, B)	(A, X)	$4(1 - f_A - f_B)f_A^2 f_B$	1/4	1/4
(A, X)	(A, A)	$2(1 - f_A - f_B)f_A^3$	1/2	0
(A, X)	(A, B)	$4(1 - f_A - f_B)f_A^2 f_B$	1/4	1/4
(A, X)	(A, X)	$4(1 - f_A - f_B)^2 f_A^2$	1/4	0

Tento výsledok už nebudeme podrobne upravovať ako tie predošlé. Nepokladáme to za potrebné, lebo spôsob, akým sa to robí, poznáme. Pozrime sa radšej v skratke na iné prípady príbuzenstva. Čitateľ už asi tuší, že použijeme vzorce, ktoré sme odvodili v časti 2.3 o iných alternatívnych hypotézach. Musíme si však uvedomiť jeden dôležitý fakt, a síce, že v tomto prípade nefunguje symetria, pomocou ktorej sme v spomínanej kapitole zjednodušovali vzorce. Napríklad v prípade syna a otca bude výsledný vzorec pre prípad čiastočnej zhody mať tvar

$$P(O|S) = \frac{P(S|O)P(O)}{P(S)} = \frac{\sum_i P(S|O \wedge \mu_i)P(\mu_i)P(O)}{P(S)}, \tag{2.11}$$

lebo $P(S)$ a $P(O)$ sa vo vzorci nevykrátia, nakoľko sa vo všeobecnosti nerovnajú (napríklad to môžu byť pravdepodobnosti, že genotyp otca je (A, A) a genotyp syna je (A, B)). Podobne to bude aj v ostatných prípadoch (starý otec – vnuk, strýko – synovec atď.). Samotné sumy a spôsob odvodenia však ostávajú rovnaké.

Kapitola 3

Koeficient príbuznosti

3.1 Čo robiť ak predpoklady neplatia

V úvode práce sme zhrnuli a vysvetlili predpoklady o stave a vývine populácie, ktoré sme následne používali pri odvodení vzorcov. Vyzdvihli sme jeden, ktorý v praxi takmer nikdy neplatí, čo potom samozrejme vplýva na správnosť vzorcov odvodených v predošlej kapitole. Na nasledujúcich stranách sa k tejto problematike vrátime, ukážeme si, ako opraviť prípadné chyby a aká je situácia na Slovensku.

Hardyho – Weinbergovo ekvilibrium hovorí, že výber partnera v populácii funguje náhodne, t.j. pre každého muža a ženu (zrejme sa bavíme o osobách v približne rovnakom veku) je rovnako pravdepodobné, že budú mať spoločných potomkov. Ak sa na tento fakt pozrieme na dlhom časovom úseku, znamená to vlastne, že výskyt dvoch alel, ktoré má jedinec na niektorom konkrétnom locuse, je navzájom nezávislý (jeho matka a otec boli náhodne vybraný pár, ich rodičia pred tým tiež atď.).

Vieme, že toto v praxi nie je pravda. Ľudia si vyberajú partnera podľa mnohých kritérií ako napríklad etnická či náboženská príslušnosť, či úroveň vzdelania. Vo väčšine prípadoch však má na výber partnera rozhodujúci vplyv zemepisná poloha a vzdialenosti. Ako príklad si predstavme typickú slovenskú obec s povedzme 1000 obyvateľmi, kde sa ľudia roky sobášili (vo veľkej miere) medzi sebou. V tejto obci sú genotypy jej obyvateľov navzájom „dosť podobné“, nakoľko pri pohľade do minulosti často nachádzame spoločných predkov aj u dnes naoko vôbec nie príbuzných ľudí. Treba však povedať, že v dnešnej dobe sa z dôvodu častej migrácie mladých ľudí situácia trochu upravuje v prospech Hardyho – Weinbergovho ekvilibria.

Ukážme si ale aj pozitívny príklad. Ten dostaneme, ak sme budeme skúmať napríklad populáciu študentov na internátoch v Mlynskej doline. Nakoľko študenti sú z celého

Slovenska (zemepisná poloha ich bydliska nie je rozhodujúce kritérium) a nemajú (až na malé výnimky) spoločných predkov, mohli by sme považovať Hardyho – Weinbergovo ekvilibrium za platné, a teda vzorce z predošlej kapitoly sú použiteľné. Podobná situácia je napr. aj v krajinách, kde človek cestuje po štáte za prácou a často mení bydlisko (Slovensko takouto krajinou nie je).

V tejto kapitole sa teda budeme snažiť odhadnúť mieru toho, ako veľmi majú ľudia žijúci v určitom regióne spoločné genetické vlastnosti vyplývajúce z existencie spoločných predkov. Matematicky ju budeme interpretovať ako akúsi koreláciu medzi výskytom dvoch alel na locuse. Formálne sa definuje nasledovne:

Definícia: *Koeficient príbuznosti* je pravdepodobnosť, že dve alely náhodne vybrané z danej subpopulácie sú rovnaké z dôvodu spoločného pôvodu.

V knihe [3] a článkoch [4] a [5] sa toto číslo nazýva *co-ancestry coefficient* a označuje sa ako F. Na nasledujúcich stranách čitateľovi naznačíme spôsob korekcie vzorcov z predošlej kapitoly a následne sa pokúsime odhadnúť hodnoty F pre subpopulácie slovenských krajov.

Pre úplnosť by sa ešte patrilo spomenúť, aké môže mať spomínaná nepresnosť vzorcov praktické následky. Predstavme si, že máme prípad vraždy, ktorým sme sa zaoberali v minulej kapitole a určujeme value of evidence pri dvojici klasických hypotéz. Nech genotyp obvineného aj páchateľa je (A, A). Pravdepodobnosť náhodnej zhody (ktorá je v menovateli vzorca (1.1)) by sme chceli počítať ako f_A^2. Môžeme to urobiť?

Vysvetlili sme si, že ak páchateľ pochádza z nejakej konkrétnej obce, ľudia, s ktorými má spoločných predkov, majú podobné genotypy ako on. V danej obci je teda pravdepodobnosť náhodnej zhody väčšia ako na zvyšku územia Slovenska. Určite teda neplatí, že je rovná f_A^2 (vypočítame ju v ďalších odsekoch).

Ak vo vzorci (1.1) zvýšime pravdepodobnosť náhodnej zhody, zmenší sa value of evidence (delíme jednotku väčším číslom). Ak obvinený pochádza z danej obce, takto „prilepšíme“ jeho obhajobe (obžaloba má dôkaz s nižšou value of evidence). Naproti tomu, ak by sme používali klasický spôsob a považovali Hardyho – Weinbergovo ekvilibrium za platné, poškodili by sme obvinenému, nakoľko vypočítaná value of evidence by bola určite vyššia ako je v skutočnosti. A tu prichádzame k druhej otázke, ktorú budeme v tejto kapitole riešiť. Koľko je správna hodnota F?

3.2 Pravdepodobnosť náhodnej zhody

Ako sme už v minulom odstavci naznačili, koeficient príbuznosti bude vo vzorcoch z druhej kapitoly vplývať na pravdepodobnosť náhodnej zhody. Logicky na nič iné vplývať nemôže, nakoľko ostatné vlastnosti populácie vyplývali z iných predpokladov. Ešte konkrétnejšie vieme vplyv F definovať ako vplyv na očakávanie, aká bude alela, ktorú z populácie „uvidíme“ ako ďalšiu. Ak predpokladáme, že platí Hardyho – Weinbergovo ekvilibrium a prvá pozorovaná alela bola A, tak druhá alela bude A, B, C atď. s takou pravdepodobnosťou, aká je frekvencia jej výskytu. Ak má ale populácia, s ktorou pracujeme, nejaký kladný koeficient príbuznosti, očakávania o ďalšej pozorovanej alele sa zmenia. Ukážme si to najskôr pre dve alely na jednom locuse.

Odvoďme tvar pravdepodobnosti náhodnej zhody, ak predpokladáme kladný koeficient F (pre úplnosť treba poznamenať, že jeho záporné hodnoty uvažovať nebudeme, lebo nemajú praktickú interpretáciu).

Na začiatok majme jedinca, ktorý má v pozorovanom locuse alelu A, a poďme vypočítať pravdepodobnosť, že jeho druhá alela na tomto locuse je tiež A. Takáto situácia môže nastať v dvoch prípadoch – buď je druhá alela A, pretože je to náhoda (jeho rodičov a ich rodičov môžeme považovať za náhodne vybrané páry), alebo to tak nie je a v histórii rodiny jedinca sú prítomní spoloční predkovia pre niektoré páry. Druhú situáciu nazveme *špeciálne okolnosti* a pravdepodobnosť, s ktorou sa vyskytnú, označíme F. Potom pravdepodobnosť, že špeciálne okolnosti sa v blízkom rodokmeni jedinca nevyskytli, je $(1 - F)$.

Pravdepodobnosť, že druhá alela na locuse jedinca je A, je teda súčet pravdepodobností v prípadoch bez a s výskytom špeciálnych okolností. Ak je to naozaj náhoda, hľadaná pravdepodobnosť bude $(1-F)f_A$ (nie sú prítomné špeciálne okolnosti, no druhá alela je aj tak A). Oproti tomu, alela je A v dôsledku špeciálnych okolností s pravdepodobnosťou F (presnejšie $1F$, teda pravdepodobnosť výskytu špeciálnych okolností krát istota, že alela je v takomto prípade A).

Celková pravdepodobnosť, že jedinec v takejto populácii bude mať genotyp (A, A) je potom $f_A((1 - F)f_A + F)$, alebo inak $f_A^2(1 - F) + f_A F$.

V prípade heterozygotného genotypu (A, B) je to jednoduchšie. Jedna z alel je A a druhá nie je A (je to B), lebo špeciálne okolnosti sa nevyskytli. Potom pravdepodobnosť, že jedinec má genotyp (A, B) je $2f_A f_B(1 - F)$.

Keby sme počítali pravdepodobnosť, že ďalšia pozorovaná alela z populácie (na locuse ďalšieho jedinca) bude A, ak sme videli doteraz dve alely A (jedného homozygotného

jedinca), museli by sme brať do úvahy to, že sme daného jedinca videli. Pri počítaní pravdepodobnosti výskytu ďalšej alely musíme brať do úvahy všetky predchádzajúce[1]. Problémom sa budeme detailnejšie zaoberať v časti 3.3.2 a ukážeme si, ako sa takéto pravdepodobnosti dajú rekurentne vyjadriť pre postupnosť alel.

Odvodzovaním vzorcov z kapitoly 2 pomocou tejto korekcie sa v práci zaoberať nebudeme. Zvedavý čitateľ môže jeho náznak nájsť v článku [4] alebo v knihe [2].

Keďže už máme presnú predstavu, na čo bude koeficient príbuznosti slúžiť (na korekciu Hardyho – Weinbergovského výpočtu pravdepodobnosti náhodnej zhody), treba ešte odpovedať na otázku, či je naozaj potrebný. Pri výpočte pravdepodobnosti náhodnej zhody štandardne používame frekvencie alel, ktoré máme odhadnuté z celoslovenskej vzorky. Ak navyše vieme, že páchateľ aj obvinený pochádzajú napríklad z Košického kraja, vezmeme koeficient príbuznosti pre Košický kraj a pomocou neho upravíme vzorce. Tu ale vyvstáva otázka: nebolo by jednoduchšie a lepšie odhadnúť rovno frekvencie alel obyvateľov Košického kraja a použiť ich vo vzorcoch namiesto upravovania vzorcov s celoslovenskými frekvenciami? Odpoveď je nie. Po prvé, bolo by zdĺhavé odhadovať frekvencie pre každý kraj a locus a potom ich vkladať do výpočtov value of evidence. Navyše, nemáme dostatok dát v krajských vzorkách, aby sme tak urobili (toto je hlavný dôvod). Naším cieľom je používať odhady frekvencií alel, ktoré už máme (z celoslovenskej vzorky) a to bez ohľadu na to, v akom regióne bol trestný čin spáchaný. Koeficient príbuznosti nám slúži na splnenie tohto cieľa – aby sme vzorec len „prispôsobili“ danému kraju.

3.3 Správna hodnota F

Pri odhadovaní koeficientu príbuznosti sa sá postupovať mnohými spôsobmi a každý z nich je v istom zmysle dobrý. Napríklad sporné by mohlo byť už aj to, aké subpopulácie skúmať, pretože neexistuje spôsob ako nejakú subpopuláciu presne vymedziť. Na Slovensku by sme mohli za ukážkový príklad subpopulácie považovať napríklad rómsku menšinu, lebo sa aj fyzickými znakmi líši od zvyšku obyvateľov Slovenska. Iné subpopulácie sa však tak ľahko vymedziť nedajú. Napríklad pri geografickom členení sa nedá stanoviť presná hranica tak, aby obyvatelia obcí na jej dvoch stranách neboli v nejakom zmysle geneticky príbuzní.

Druhým problémom pri odhadovaní je často absencia použiteľných dát, prípadne ich zlá

[1] Všimnime si, že závislosť v tomto prípade je veľmi silná. Napríklad pri Markovovskej postupnosti závisí ďalší člen len od jedného predošlého, kým tu závisí od všetkých predošlých.

štruktúra. Týmto by sme chceli ešte raz vyjadriť veľkú vďaku zamestnancom Oddelenia biológie a genetickej analýzy Kriminalistického a expertízneho ústavu Policajného zboru SR za to, že nám sprístupnili slovenskú DNA-databázu a dokonca nám poskytli dáta roztriedené po krajoch.

Ďalším problémom je, že aj keď máme k dispozícii dáta poskytujúce dostatočnú informáciu, nemáme záruku, že dobre popisujú daný región. O osobách v databáze totiž nevieme, či ich predkovia žili v danom kraji už dlhší čas, alebo sa napríklad ich rodičia do daného kraja prisťahovali z nejakej vzdialenej oblasti. Toto sú ale nedostatky, ktoré nie je v našich silách odstrániť, preto sa nimi zaoberať nebudeme.

V neposlednom rade je sporný aj prístup ku koeficientu ako takému. Je jasné, že ide o to prostriedok na rozlíšenie subpopulácií medzi sebou. Môžeme však povedať, že nejaká subpopulácia „má svoje F“ ako jednu charakteristickú konštantu? Alebo sa hodnota F vzťahuje aj na konkrétny locus a teda subpopulácia má svoj vektor F, kde jeho zložky sú koeficienty príbuznosti pre konkrétne locusy? Toto sú otázky, ktoré sa v tejto časti pokúsime zodpovedať pre Slovensko. Spomenieme ešte, že pre iné štáty ako napríklad Veľkú Britániu alebo Taliansko už boli urobené analýzy tohto typu, no pre Slovensko ich podľa našich informácií ešte nikto nerobil.

Budeme používať kombináciu metodiky z článkov [5] a [7], kde sú popísané viaceré metódy na odhad koeficientu príbuznosti. Začneme tým, že odhadneme frekvencie výskytu alel na locusoch, ktoré máme k dispozícii. Použijeme na to celoslovenskú vzorku v počte 247 jedincov s genotypom pozostávajúcim z ôsmich alel ($D3$, $D8$, $D16$, $D18$, $D21$, VWA, FGA a $TH01$). Vzorka, ktorú sme dostali, obsahovala viac locusov, no ostatné boli disjunktné s krajskou databázou, ktorú sme dostali neskôr. Zmysel malo brať do úvahy len tie locusy, ktoré sa nachádzali v oboch databázach – kvôli tomu, že výpočty používajú obe databázy naraz. Frekvencie sme podľa [7] pre každý locus odhadovali ako

$$\hat{f}_i = \frac{x_i + 1}{2n + m}, \tag{3.1}$$

kde i je alela, x_i je počet výskytov alely i na tomto locuse, n je veľkosť vzorky (u nás 247) a m je počet rôznych alel, ktoré sa vyskytli na danom locuse. Výraz $2n$ sa vo vzorci nachádza kvôli tomu, že každý z n jedincov má 2 alely ($2n$ je teda celkový počet pozorovaných alel). Výstupom pre jednotlivé locusy sú tabuľky v nasledovnom tvare:

alela	16	15	17	18	14	19
frekvencia	0.272	0.265	0.194	0.165	0.090	0.014

Tabuľka 3.1. Locus $D3$

alela	12	11	12.5	10	14	13	9
frekvencia	0.143	0.099	0.004	0.046	0.250	0.336	0.012
alela	15	8	16	7			
frekvencia	0.074	0.012	0.018	0.006			

Tabuľka 3.2. Locus $D8$

alela	11	12	13	10	9	14	8
frekvencia	0.285	0.287	0.178	0.072	0.138	0.030	0.010

Tabuľka 3.3. Locus $D16$

alela	13	15	14	16	18	17	12	19
frekvencia	0.117	0.131	0.159	0.187	0.069	0.116	0.099	0.045
alela	21	20	10	16.2	24	11	22	
frekvencia	0.011	0.029	0.009	0.004	0.003	0.013	0.008	

Tabuľka 3.4. Locus $D18$

alela	29	32.2	28	31	31.2	30.2	34.2	30
frekvencia	0.249	0.119	0.136	0.062	0.079	0.072	0.019	0.190
alela	33.2	32	27	29.2	24.2			
frekvencia	0.034	0.018	0.016	0.003	0.003			

Tabuľka 3.5. Locus $D21$

alela	16	18	14	17	15	19	13
frekvencia	0.191	0.201	0.119	0.270	0.119	0.078	0.004
alela	20	11	12	21			
frekvencia	0.012	0.002	0.002	0.002			

Tabuľka 3.6. Locus VWA

alela	20	23	25	21	22	19	24
frekvencia	0.178	0.114	0.055	0.158	0.162	0.090	0.154
alela	26	22.2	18	16	21.2	27	23.2
frekvencia	0.017	0.019	0.023	0.004	0.004	0.004	0.006
alela	17	32	24.1	24.2	26.2		
frekvencia	0.004	0.002	0.002	0.002	0.002		

Tabuľka 3.7. Locus FGA

alela	9	9.3	7	6	8	10	8.3	5
frekvencia	0.200	0.294	0.140	0.230	0.124	0.006	0.004	0.002

Tabuľka 3.8. Locus $TH01$

Samotné hodnoty F nebudeme z dát odhadovať priamo, ale simulovať pomocou Metropolisovho – Hastingsovho algoritmu. Pri jeho programovaní sme postupovali podľa knihy [6]. Našou pôvodnou ideou však nebolo použiť simulačný algoritmus. To, prečo sme ho nakonieco použili, sa čitateľ dozvie v časti 3.3.3. Najskôr však podrobne vysvetlíme, ako spomínaný algoritmus funguje, aby dôvody uvedené v spomínanej časti boli lepšie pochopiteľné.

3.3.1 Metropolisov – Hastingsov algoritmus

Tento algoritmus vo všeobecnosti slúži na generovanie postupnosti hodnôt z požadovaného rozdelenia $D(x)$. Použijeme ho, pretože naše $D(x)$ nepatrí medzi štandardné rozdelenia, z ktorých program R generuje priamo. Ideou algoritmu je, že ak dosť dlho generujeme Markovovskú postupnosť hodnôt, po nejakom čase (u nás je to po prvých 1000 iteráciách) dostávame už hodnoty, ktoré pochádzajú z rozdelenia $D(x)$. Vstupom do algoritmu je teda požadovaná hustota $D(x)$ parametra, ktorý chceme generovať (u nás F), hustota rozdelenia $Q(x)$, z ktorej generujeme novú hodnotu parametra, a prvá hodnota parametra (prvý člen postupnosti). U nás bude $Q(x)$ normálne rozdelenie so strednou hodnotou rovnou predošlému F a štandardnou odchýlkou 0.002 (neskôr vysvetlíme, prečo sme sa rozhodli pre takúto hodnotu). Podľa literatúry je takáto voľba $Q(x)$ najpoužívanejšia. Algoritmus pracuje v piatich krokoch:

1. Vygeneruje novú hodnotu parametra F^{new} z rozdelenia $Q(x)$ – normálneho rozdelenie so strednou hodnotou rovnou predošlému F a štandardnou odchýlkou 0.002.
2. Vygeneruje hodnotu α z rovnomerného rozdelenia na intervale $[0, 1]$.

3. Porovná hodnotu α a výrazu

$$\frac{D(F^{new})Q(F^{new}, F)}{D(F)Q(F, F^{new})},$$

kde $Q(F, F^{new})$ je hodnota hustoty normálneho rozdelenia so strednou hodnotou F v bode F^{new}. Tu si všimnime, že vďaka symetrickosti normálneho rozdelenia sa $Q(F^{new}, F)$ a $Q(F, F^{new})$ vykrátia (to vysvetľuje častú voľbu normálneho rozdelenia za $Q(x)$).

4. Ak je hodnota výrazu väčšia ako hodnota α, algoritmus považuje F^{new} za vyhovujúce a pridá ho do postupnosti hodnôt za posledné F.
5. Ak je hodnota výrazu menšia ako α, F^{new} je nevyhovujúce a ako ďalší člen pridá do postupnosti znova hodnotu predošlého F (táto hodnota tam teda bude dvakrát za sebou).

Výstupom z tohto algoritmu je postupnosť hodnôt F, ktorá má 10000 členov (toľko nagenerovali aj autori článku [5]), z ktorých prvých 1000 odstránime. Podľa článku a aj podľa nášho pozorovania približne toľko trvá, kým program začne generovať postupnosť, ktorá spĺňa naše podmienky.

Ešte sa vráťme k štandardnej odchýlke 0.002. Aby bol algoritmus použiteľný, musí generovať rôzne hodnoty, t.j. musí prijať F^{new} „dosť často“. V literatúre sa uvádza, že algoritmus by mal prijať F^{new} v aspoň polovici krokov. Táto jeho vlastnosť sa dá regulovať tým, že meníme disperziu rozdelenia $Q(x)$, z ktorého F^{new} generuje. Čím je táto menšia, tým bližšie je F^{new} k strednej hodnote rozdelenia $Q(x)$, ktorá je rovná predošlej prijatej hodnote F.

3.3.2 Použitie Metropolisovho – Hastingsovho algoritmu

Aby sme mohli použiť Metropolisov – Hastingsov algoritmus, musíme mu zadať všetky potrebné vstupy, z ktorých najdôležitejšia je informácia, z akého rozdelenia má generovať. Toto rozdelenie skonštruujeme Bayesovským prístupom, čiže budeme predpokladať nejaké apriórne rozdelenie parametra F, ktoré následne upravíme pomocou krajskej vzorky dát (lebo porovnávame F v rámci krajov). Výsledné aposteriórne rozdelenie potom do algoritmu vstúpi ako $D(x)$. Keď budeme mať po spustení algoritmu výslednú postupnosť 9000 hodnôt F (po odobratí prvej tisícky), nakreslíme jej histogram a graficky (v softvéri R [9]) odhadneme hustotu rozdelenia parametra F. Ako reprezentatívnu hodnotu F potom zoberieme podľa vzoru článku [5] modus tejto hustoty (t.j. hodnotu, ktorej výskyt je najpravdepodobnejší, nakoľko v ňom nadobúda funkcia hustoty najväčšiu hodnotu).

Podľa vzoru spomínaných článkov budeme predpokladať, že F má apriórne rozdelenie β s parametrami 1.5 a 50. Zvolili sme ho, lebo:

- Očakávame hodnotu F určite menšiu ako 0.1 (takýto vysoký koeficient príbuznosti by už zodpovedal pokrvnej príbuznosti, napríklad F pre synovca a strýka je[2] 1/8).
- Doteraz pozorované hodnoty boli maximálne 0.03. Rozdelenie $\beta(1.5, 50)$ má medián 0.023 a strednú hodnotu 0.029, čo je väčšie ako predpokladaná hodnota F. Navyše predpokladáme, že korekcia apriórnej hustoty pomocou dát zníži hodnotu modusu pod 0.03 (z výslednej hustoty bude zjavné, ako veľmi dáta vplývali na jej tvar).

Váhu dát budeme do výpočtu aposteriórnej hustoty zahŕňať podľa Bayesovej vety v tvare

$$h(F|X) \sim h(F)h(X|F), \tag{3.2}$$

kde $h(F|X)$ je hustota parametra F podmienená vzorkou dát X, $h(F)$ je už spomínaná apriórna hustota a $h(X|F)$ je člen, ktorý zodpovedá váhe (likelihood) dát. Všimnime si, že vzorec nehovorí o presnom tvare aposteriórnej hustoty, ale len o úmere, t.j. nepoznáme konštantu, ktorou treba pravú stranu vynásobiť, aby sme dostali ľavú. Platí teda

$$h(F|X) = K \cdot h(F)h(X|F). \tag{3.3}$$

Toto však vôbec nevadí Metropolisovmu – Hastingsovmu algoritmu. Stačí si uvedomiť, že pri vyhodnocovamí výrazu v tvare

$$\frac{D(F^{new})Q(F^{new}, F)}{D(F)Q(F, F^{new})}, \tag{3.4}$$

kde aposteriórna hustota vystupuje, je $D(x)$ činiteľom v čitateli a aj v menovateli zlomku (len pre úplnosť $D(F) = h(F|X) = K \cdot h(F)h(X|F)$). To znamená, že konštanta K, ktorú nepoznáme, sa vykráti, a teda nepotrebujeme vedieť jej presnú hodnotu (pričom logicky vieme, že to nie je nula kvôli vlastnostiam F, takže jej prítomosť v menovateli je tiež v poriadku). Táto skutočnosť sa považuje za jednu z kľúčových predností Metropolisovho – Hastingsovho algoritmu.

Aby sme teda dostali výslednú hustotu, musíme ešte zarátať váhu dát, resp. vynásobiť apriórnu hustotu pravdepodobnosťou v tvare $h(\mathrm{X}|F)$, kde X sú naše dáta. To je vlastne „pravdepodobnosť, že sme dostali presne také dáta, ako sme dostali“, pričom dátami

[2]Túto hodnotu dostaneme porovnaním tvaru vzorca odvodeného pre synovca a strýka v predošlej kapitole a tvaru tohto vzorca z manuálu [1] , ktorý je $f_A^2 + 4f_A(1 - f_A)F$ pre homozygota a $2f_Af_B + 2(f_A + f_B - 4f_Af_B)F$ pre heterozygota.

myslíme postupnosť alel na locusoch ľudí z krajskej vzorky. Už sme si odvodili pravdepodobnosť, že ak prvá alela na locuse A, tak aj druhá bude A (v sekcii 3.2). Rovnako sa to dá vypočítať aj pre ďalšie, resp. vieme vyjadriť, aká je pravdepodobnosť, že ak z n doteraz vybraných alel bolo n_A alel A, tak aj ďalšia vybraná alela bude A. Výsledný tvar vzorca je

$$P_n(A) = \frac{n_A F + f_A(1-F)}{1+(n-1)F}, \tag{3.5}$$

pričom tento vzorec nebudeme odvádzať, nakoľko je to náročné a nie je to predmetom tejto práce. Čitateľ môže nájsť jeho odvodenie napríklad v článku [4] alebo naznačené v knihe [2].

Pre potreby nášho algoritmu musíme tento vzorec ešte zovšeobecniť, a to tak, aby nerátal len s jednou alelou, ale so všetkými, teda s kompletnou vzorkou. Hľadáme teda vzorec na výpočet pravdepodobnosti, že ak z N doteraz pozorovaných alel bolo N_i alel i, tak aj ďalšia vybraná alela bude i, platiaci zároveň pre každé i. Túto pravdepodobnosť vyjadríme rekurentne. Výsledný vzorec má tvar

$$P_{N+1}(N_1, N_2, \ldots, N_{i+1}, \ldots, N_m) =$$

$$= P_N(N_1, N_2, \ldots, N_i, \ldots, N_m)\frac{N_i F + \hat{f}_i(1-F)}{1+(N-1)F}, \tag{3.6}$$

kde N_i sú počty alely i pozorované do aktuálneho kroku, n je veľkosť vzorky a F je koeficient príbuznosti (v algoritme je to posledná hodnota F, ktorá bola prijatá).

Hodnota prvého člena je $P_1(0,0,\ldots,0) = 1$. Pri výpočte tejto rekurencie algoritmus počíta alely, keď prechádza vzorkou, ktorú sme mu dodali, a pre každú z nich aplikuje vzorec (3.5). S každou novou pozorovanou alelou sa niektoré N_i zväčší o jedna. Napríklad pre locus $TH01$ bude počítadlo na začiatku nulový vektor, kde alely idú v poradí ako v tabuľke 3.8. Keď algoritmus začne prechádzať dáta a prvá alela, ktorú nájde, je 10, počítadlo sa zmení na $(0,0,0,0,0,1,0)$. Takýmto spôsobom prejde celou vzorkou dát a výraz P_N bude súčin zlomkov v tvare, aký sme uviedli. Vzorec, podobne ako predošlý, nebudeme odvádzať, jeho podobu sme prebrali z článku [7].

K použitiu Metropolisovho – Hastingsovho algoritmu už dodáme len to, že prvý člen postupnosti (prvú hodnotu F) sme vygenerovali z už spomínaného apriórneho rozdelenia $\beta(1.5, 50)$ (aj keď na tomto veľmi nezáleží, keďže prvých 1000 hodnôt sme pri výslednom odhade nepoužili).

3.3.3 Pôvodná idea

Naším pôvodným nápadom bolo určiť aposteriórne rozdelenie (teda hustotu) parameftra F priamym výpočtom podľa (3.2). Mali sme v pláne zobrať apriórne rozdelenie $\beta(1.5, 50)$ a vynásobiť ho postupnosťou zlomkov podľa (3.6), čím by sme dostali pravú stranu vzorca (3.2). Následne by sme numerickým integrovaním spočítali plochu S pod získanou krivkou. Keďže výsledkom má byť hustota (čiže plocha pod krivkou má byť rovná jednej), vypočítali by sme potrebnú konštantu K ako $1/S$. Narazili sme však na obrovský problém.

Pri výpočte člena $h(X|F)$ podľa (3.6) sme ako výsledok dostali numericky nulu. Len pre zopakovanie, prvý člen rekurencie bol $P_1(0, 0, \ldots, 0) = 1$ a nasledujúce boli zlomky vždy menšie ako jednotka, pričom tieto zlomky sa postupne násobili medzi sebou. Pozrime sa na výstup z programu, ktorý počítal výslednú hodnotu P_N (súčin všetkých zlomkov):

```
0.1577909
0.01788051
0.001591931
          ...
2.672804e-141
5.937063e-142
```

V každom kroku sa priebežná hodnota súčinu zmenšila cca desaťkrát. Vidíme, že na konci výpočtu bola už hodnota $h(X|F)$ a teda aj celá hustota parametra F podľa programu prakticky rovná nule, čo znemožňuje určiť konštantu zo vzorca (3.2).

Potom sme ale objavili článok [5] a v ňom Metropolisov – Hastingsov algoritmus. Ak si podrobnejšie rozpíšeme podiel, ktorý vyčísľujeme v (3.4), dostaneme

$$\frac{D(F^{new})Q(F^{new}, F)}{D(F)Q(F, F^{new})} = \frac{K \cdot h(F^{new}) \cdot h(X|F^{new}) \cdot Q(F^{new}, F)}{K \cdot h(F) \cdot h(X|F) \cdot Q(F, F^{new})}. \tag{3.7}$$

Ak použijeme náš program, v čitateli aj menovateli (počítame ich zvlášť) vyjde nula. Keďže sa ale v čitateli aj v menovateli vzorca (3.7) jedná o súčin, môžeme ho počítať postupne, ak si uvedomíme, že $h(X|F^{new})$ aj $h(X|F)$ je súčinom rovnakého počtu zlomkov. Postupovať budeme nasledovne.

- Najskôr vykrátime konštantu K a dvojicu $Q(F, F^{new})$ a $Q(F^{new}, F)$ (o tom, prečo sa to dá, sme sa už zmienili).
- Potom vypočítame podiel $h(F^{new})/h(F)$, čo je podiel hodnôt hustoty rozdelenia $\beta(1.5, 50)$ v dvoch rôznych bodoch.

- Na záver vypočítame podiel $h(X|F^{new})/h(X|F)$ postupne po jednotlivých zlomkoch, teda vydelíme prvý člen rekurencie pre F^{new} prvým členom rekurencie pre staré F, potom druhý vydelíme druhým atď. Vzťah (3.7) potom nadobudne tvar

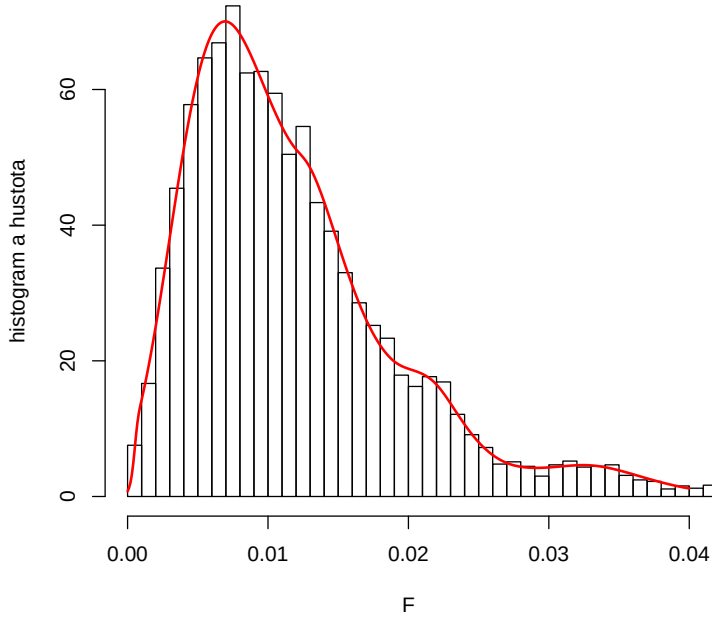
$$\frac{D(F^{new})Q(F^{new},F)}{D(F)Q(F,F^{new})} = \frac{\beta(F^{new},1.5,50)\cdot P_1^{F^{new}}\cdot P_2^{F^{new}}\cdot P_3^{F^{new}}\ldots P_N^{F^{new}}}{\beta(F,1.5,50)\cdot P_1^{F}\cdot P_2^{F}\cdot P_3^{F}\ldots P_N^{F}} =$$
$$= \frac{\beta(F^{new},1.5,50)}{\beta(F,1.5,50)}\cdot\frac{P_1^{F^{new}}}{P_1^F}\cdot\frac{P_2^{F^{new}}}{P_2^F}\cdot\frac{P_3^{F^{new}}}{P_3^F}\cdots\frac{P_N^{F^{new}}}{P_N^F}.$$

Keďže budeme zakaždým najskôr deliť malé číslo malým a len potom tieto podiely násobiť, zaistíme, aby nám výsledná hustota neimplodovala do nuly. Takto sme vyriešili naoko neriešiteľný problém použitím Metropolisovho – Hastingsovho algoritmu: zachránilo nás, že sme nemuseli vyčísliť $D(F^{new})$ a $D(F)$, ale len ich podiel.

3.3.4 Výsledky

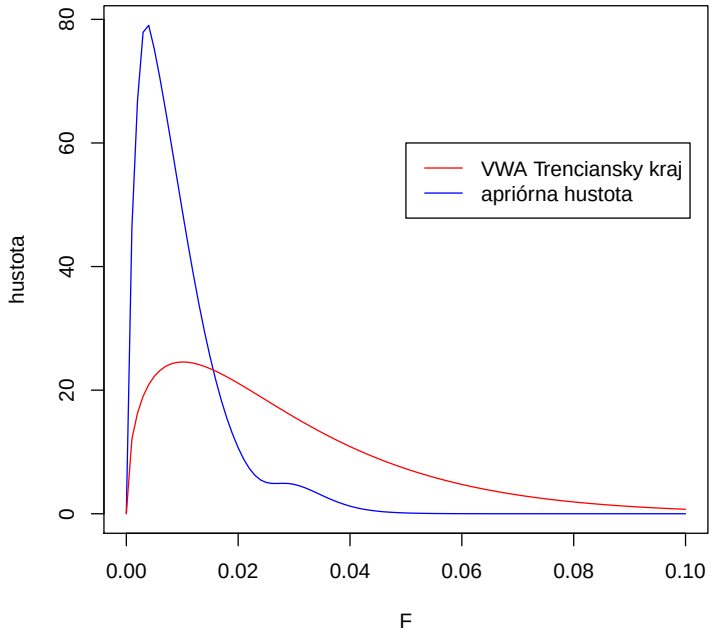
Vráťme sa však k Metropolisovmu – Hastingsovmu algoritmu a tomu, čo sme pomocou neho dostali. Program sme použili na výpočet F pre konkrétne dvojice kraj – locus (teda napríklad FGA v Banskobystrickom kraji). Vstupnými parametrami sú frekvencie alel na locuse, ktoré máme odhadnuté z celoslovenskej vzorky, kde vystupuje daný locus a alely v ňom. Ako výstup pre locus FGA v Banskobystrickom kraji sme dostali nasledovný histogram a hustotu.[3]

[3]Hustotu sme aproximovali pomocou balíka *polspline* v softvéri R [9].



Obrázok 3.1. Histogram a hustota F na locuse FGA v Banskobystrickom kraji

Zaujímavé je aj porovnanie s apriórnou hustotou, kde vidíme, aký veľký vplyv má na výslednú hustotu korekcia dátami. Zdá sa, že platí, že čím viac máme dát (teda dodatočnej informácie, ktorou upravíme apriórne rozdelenie), tým má výsledné odhadnuté rozdelenie parametra F užší kopec. V jazyku odhadov to znamená, že máme lepšiu informáciu o tom, kde sa nachádza presná hodnota F – čím špicatejšie je totiž výsledné rozdelenie, tým presnejšie sme odhadli interval, v ktorom sa asi skutočné F nachádza (interval je šírka kopca). Inými slovami, špicatejší, užší vrchol znamená, že nasimulované hodnoty F sú viac koncentrované okolo skutočnej hodnoty parametra, pre ktorú máme úzky intervalový odhad.



Obrázok 3.2. Porovnanie hustoty F na locuse VWA a apriórnej hustoty $\beta(1.5, 50)$ v Trenčianskom kraji

Pre jednotlivé kraje sme následne samostatne vypočítali odhady F pre každý z locusov VWA, FGA a $TH01$, zhrnutie je v tabuľke 3.9. „Celkové“ F sme počítali zo všetkých 8 locusov (do člena, ktorým korigujeme apriórne rozdelenie nevstúpili dáta o jednom, ale o ôsmich locusoch v danom kraji), ktoré sa zhodovali v celoslovenskej a krajských vzorkách. Hodnoty „celkového“ F vyšli vždy niekde medzi hodnotami na jednotlivých locusoch, čo je logicky správny výsledok.

Musíme však dodať, že nie vždy je správne počítať hodnotu „celkového“ F. Nemôžeme totiž predpokladať, že na všetkých locusoch je hodnota koeficientu príbuznosti rovnaká. V polovici krajov (Banskobystrickom, Trnavskom, Košickom a Trenčianskom) sú podľa nášho výpočtu totiž hodnoty F (ich modusy, niekedy aj tvary hustôt) na jednotlivých locusoch príliš rozdielne nato, aby to malo zmysel.

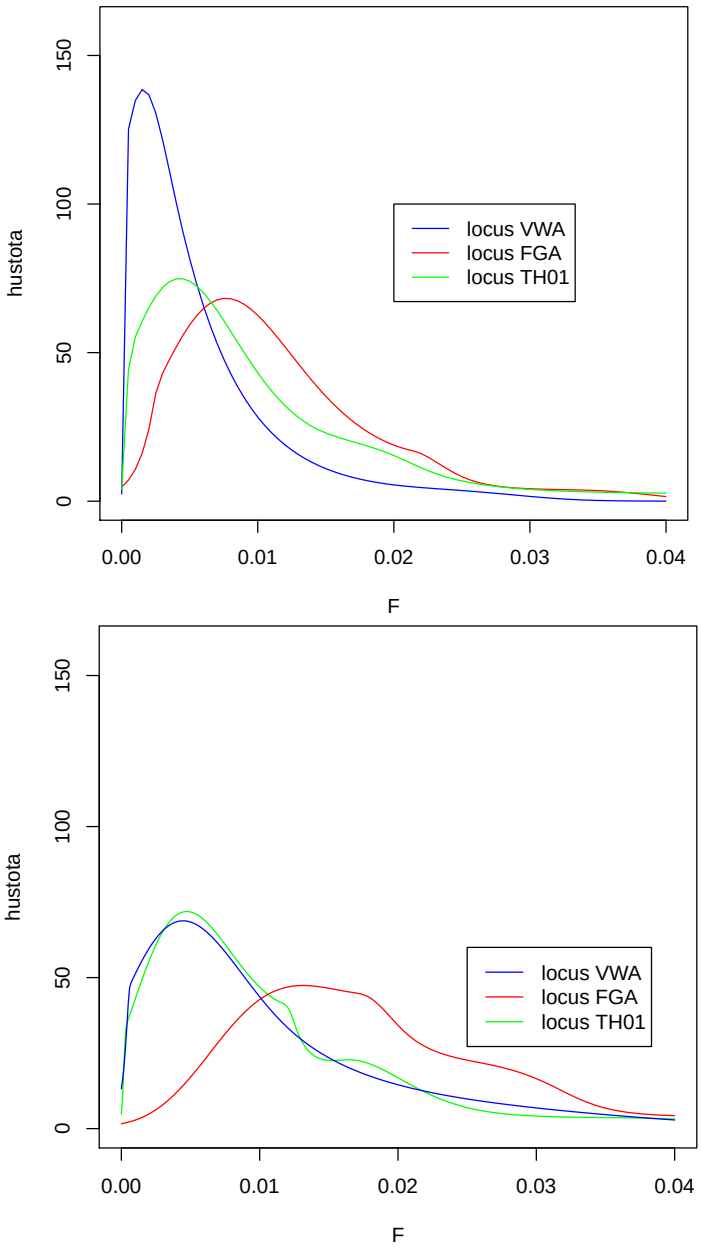
Kraj / Locus	veľkosť vzorky	VWA	FGA	$TH01$	„celkové F“
Bratislavský	65	0.00424	0.00318	0.00397	0.00404
Banskobystrický	79	0.00182	0.00709	0.00413	0.00235
Košický	82	0.00981	0.00375	0.00948	0.01883
Nitriansky	76	0.00263	0.00277	0.00358	0.00275
Prešovský	93	0.00418	0.00597	0.00410	0.00445
Trenčiansky	73	0.00392	0.00308	0.00965	0.00389
Trnavský	62	0.00419	0.01376	0.00464	0.00316
Žilinský	82	0.00170	0.00447	0.00215	0.00431

Tabuľka 3.9. Modus hodnoty F pre niektoré locusy podľa krajov

V ostatných štyroch krajoch (Prešovskom, Bratislavskom, Nitrianskom a Žilinskom) sme na pozorovaných locusoch zistili navzájom veľmi podobné hodnoty modusu koeficientu príbuznosti a teda úvaha o „celkovom“ F by mala zmysel.

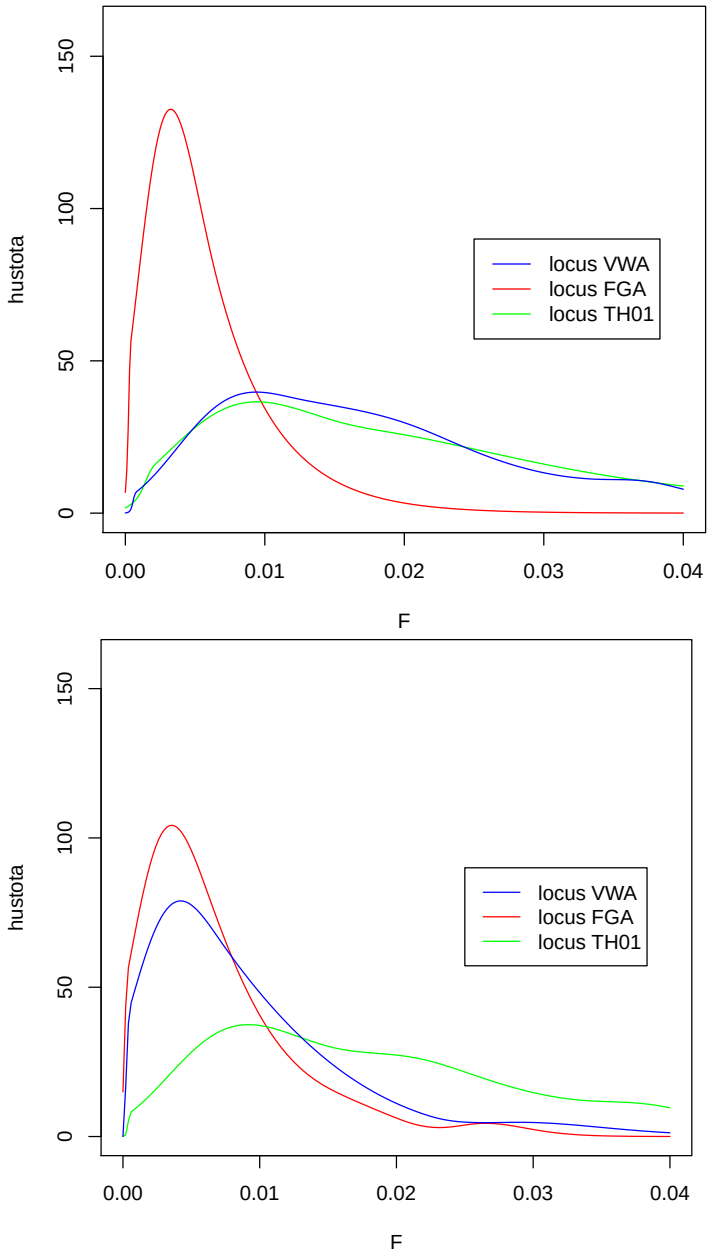
Na nasledujúcich stranách uvidíme výsledné hustoty v krajoch (pre tri locusy VWA, FGA a $TH01$)[4]. Platí, že čím má hustota tenší (špicatejší) kopec, tým presnejší je odhad modusu pre dané F, ktorý je v tabuľke 3.9. Podľa tohto kritéria sme najpresnejšie odhadli koeficientu príbuznosti pre locus VWA v Nitrianskom kraji. Vidíme, že miera presnosti odhadu, ak by sme ju definovali len šírkou kopca, je v jednotlivých krajoch veľmi rozdielna.

[4]Vybrali sme preto, lebo ich často ako reprezentatívne používali aj autori článkov, z ktorých sme vychádzali.

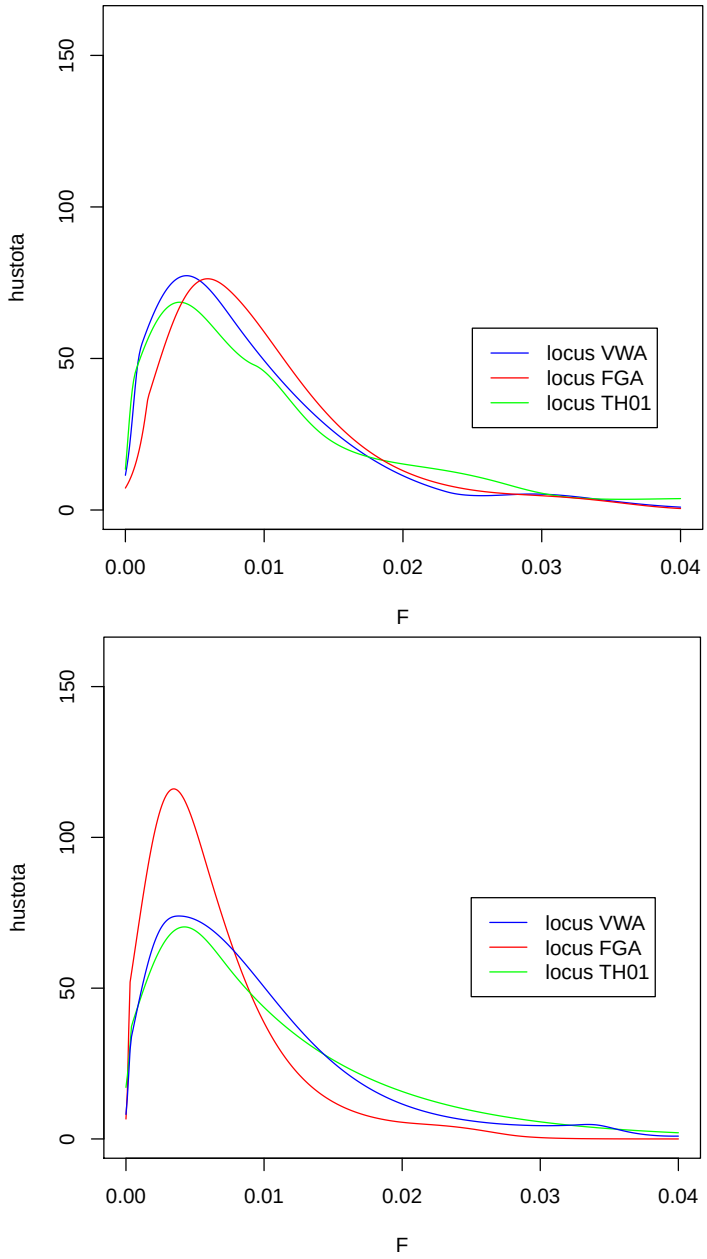


Obrázok 3.3. Banskobystrický (hore) a Trnavský (dole) kraj

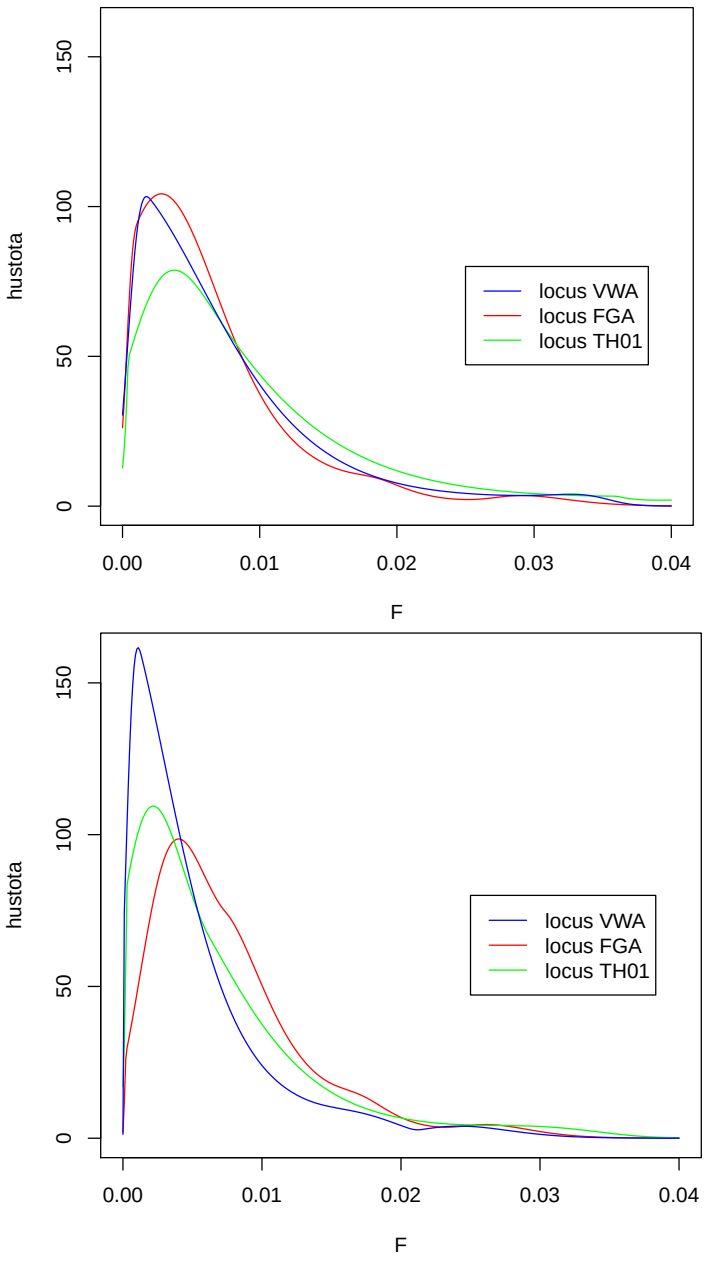
V Košickom, Trenčianskom a Bratislavskom kraji vidíme výrazne tenší kopec (hustotu koncentrovanú blízko nuly) na locuse FGA.



Obrázok 3.4. Košický (hore) a Trenčiansky (dole) kraj

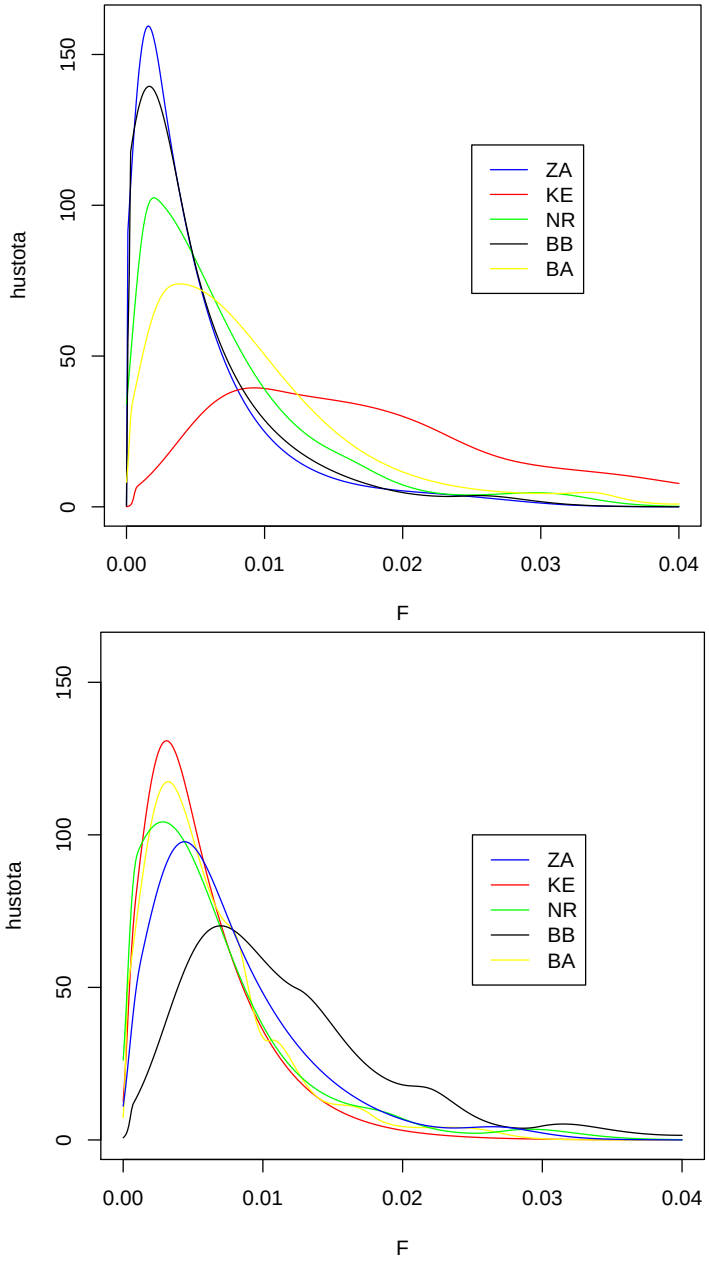


Obrázok 3.5. Prešovský (hore) a Bratislavský (dole) kraj



Obrázok 3.6. Nitriansky (hore) a Žilinský (dole) kraj

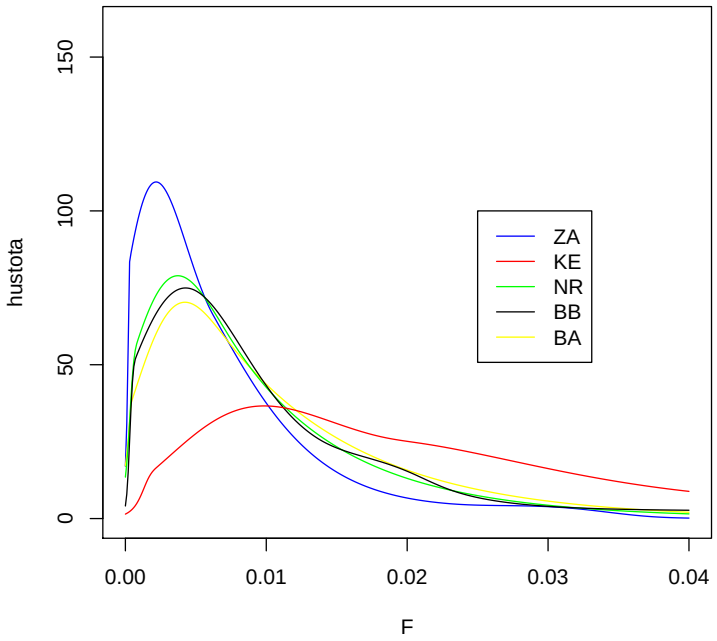
Okrem otázky, či sa dá určiť pre kraj jedna hodnota F spoločná pre všetky locusy (pre niektoré áno, pre iné nie) sme skúmali aj to, či ten ktorý locus má podobný modus a hustotu F v rôznych krajoch, teda, či vieme povedať, že napríklad locus FGA má na celom území Slovenska hodnoty, ktoré môžeme považovať za rovnaké. Na nasledujúcich obrázkoch je situácia pre tri skúmané locusy.



Obrázok 3.7. Locusy VWA (hore) a FGA (dole) v piatich krajoch

Vidíme, že na locuse VWA nie sú medzi štyrmi z načrtnutých krajov[5] výrazné rozdiely v polohe modusu a teda v koncentrácii hustoty. Košický kraj je ale viditeľne iný, tam má odhadnutá aposteriórna hustota príliš široký kopec a preto je odhad koeficientu príbuznosti menej presný ako pre zvyšné štyri kraje. Rovnaká situácia sa v tomto kraji

[5]Pôvodne sme chceli znázorniť všetkých osem krajov, no obrázok by bol veľmi neprehľadný. Preto sme zvolili vzorovú päticu tak, aby reprezentovala celé Slovensko.



Obrázok 3.8. Locus $TH01$ v piatich krajoch

opakuje aj pri locuse $TH01$. Tieto dva odhady, ako neskôr uvidíme v tabuľke 3.10, majú väčšiu výberovú disperziu ako ostatné, čo je len logický dôsledok toho, čo vidíme na obrázkoch.

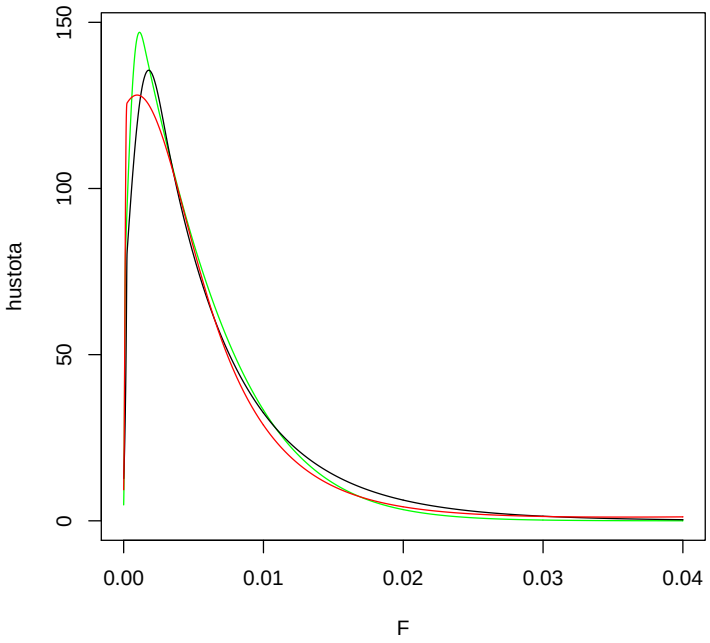
3.4 Nepresnosti odhadu

Keďže sme použili simulačný algoritmus, výsledky pri jeho opakovanom spustení sa navzájom líšia. V tejto časti sa pokúsime vyhodnotiť mieru chybovosti zapríčinenej dvoma základnými faktormi a odstrániť chyby, ktoré mohli vzniknúť pri výpočte.

3.4.1 Počet iterácií

Nakoľko pri odhade parametra F odhadujeme jeho hustotu a následne jej modus pomocou simulačného algoritmu, je jasné, že hodnota modusu nebude pri opakovanom použití algoritmu rovnaká. Navyše počas iterácií ide, hlavne pri vyčísľovaní hodnoty P_N zo vzorca (3.6), o malé čísla a preto je možná aj určitá nepresnosť spôsobená výpočtovou technikou. V neposlednom rade sa odlišnosť výstupu pri opakovanom spustení predpokladá aj v samotnom algoritme, nakoľko ide o generovanie náhodnej postupnosti z nejakého rozdelenia.

Po niekoľkých spusteniach algoritmu sme pozorovali, že hodnota modusu je rôzna a niekedy sa líši už aj na treťom desatinnom mieste. Túto vlastnosť sme sa snažili odstrániť zvyšovaním počtu iterácií[6].Tento pokus však nepriniesol žiadny výsledok, lebo hodnota modusu sa nezačala „ustáľovať“. Preto sme skúsili graficky porovnať, ako vyzerá hustota parametra F pri viacerých spusteniach algoritmu. Dostali sme (pri troch spusteniach) nasledujúci obrázok.



Obrázok 3.9. Výsledky troch spustení simulácie

Usúdili sme, že rozdiel v hustotách (hlavne ich tvaroch) nie je závažný a keďže so zvyšujúcim sa počtom iterácií sme nedosiahli zlepšenie, ďalej sme sa problému nevenovali.

V rámci hodnotenia presnosti odhadu sme ešte zistili aposteriórne výberové disperzie koeficientu príbuznosti pre každú kombináciu kraj – locus, s ktorou sme pracovali. Hodnoty disperzií sú v nasledujúcej tabuľke, rádovo sa pohybujú v 10^{-5}, čiže sú to relatívne malé čísla. Uvádzame ich preto, lebo v ďalšej časti budeme tento výsledok porovnávať s druhým modelom a disperziu potrebujeme ako indikátor presnosti odhadu.

[6]V článkoch [5] a [7] robili autori 10000 iterácií a analýze nepresnosti odhadu sa vôbec nevenovali.

Kraj / Locus	VWA	FGA	$TH01$
Bratislavský	$2.610 \cdot 10^{-5}$	$6.522 \cdot 10^{-5}$	$9.697 \cdot 10^{-5}$
Banskobystrický	$5.785 \cdot 10^{-5}$	$3.295 \cdot 10^{-5}$	$8.928 \cdot 10^{-5}$
Košický	$1.924 \cdot 10^{-5}$	$15.301 \cdot 10^{-5}$	$18.763 \cdot 10^{-5}$
Nitriansky	$3.851 \cdot 10^{-5}$	$4.901 \cdot 10^{-5}$	$8.225 \cdot 10^{-5}$
Prešovský	$5.198 \cdot 10^{-5}$	$5.957 \cdot 10^{-5}$	$10.445 \cdot 10^{-5}$
Trenčiansky	$3.451 \cdot 10^{-5}$	$6.251 \cdot 10^{-5}$	$17.800 \cdot 10^{-5}$
Trnavský	$10.349 \cdot 10^{-5}$	$9.753 \cdot 10^{-5}$	$8.669 \cdot 10^{-5}$
Žilinský	$3.392 \cdot 10^{-5}$	$3.075 \cdot 10^{-5}$	$4.722 \cdot 10^{-5}$

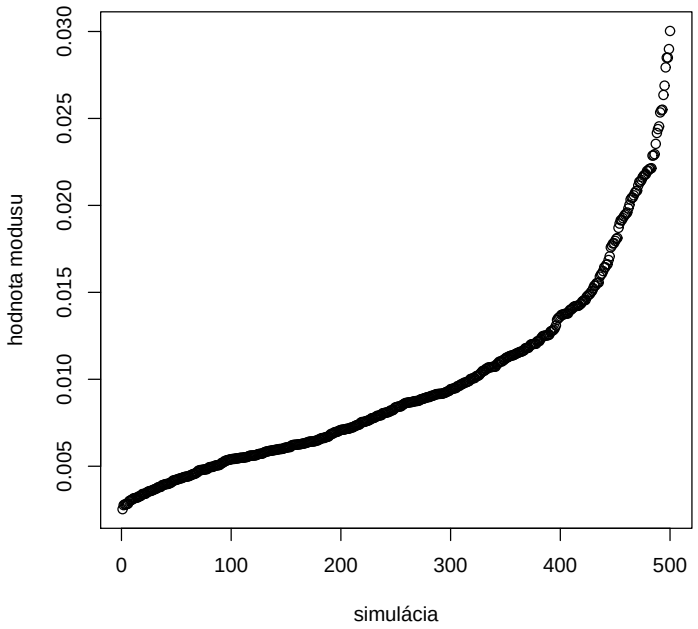
Tabuľka 3.10. Výberová disperzia odhadu F pomocou prvého modelu

3.4.2 Vplyv Bayesovskej korekcie

Druhým, závažnejším problémom, ktorý nastáva, je problém korekcie dátami. V aproximácii hustoty parametra F násobíme apriórne $\beta(1.5, 50)$ rozdelenie pravdepodobnosťou, že dostaneme danú vzorku krajských dát. Keďže dáta prechádzame postupne a po jednom, záleží na každom z nich. Ak by sme niektorý skúmaný locus s alelami $(9, 10)$ nahradili napríklad locusom s alelami $(9, 8)$, výsledná aposteriórna hustota $D(x)$, ktorá vstupuje do Metropolisovho – Hastingsovho algoritmu by sa zmenila (lebo sa zmení výsledok (3.6)). V dôsledku tohto sa iste zmení aj modus a možno aj tvar výslednej hustoty. Kvôli tomuto sme sa rozhodli urobiť akúsi analýzu citlivosti odhadu na vzorku podľa vzoru článku [8], kde autorka generuje vlastné vzorky dát z pôvodnej vzorky pomocou metódy *bootstrap* a následne určuje interval, v ktorom sa nachádza skutočná hodnota parametra.

Ako sme postupovali, ukážeme na príklade. Zobrali sme vzorku z Banskobystrického kraja o počte 79 jedincov, z ktorých každý má locus VWA (a v ňom dve alely). Z tejto vzorky sme náhodným výberom s opakovaním vybrali 79 jedincov a týchto sme prehlásili za novú vzorku. Z každej krajskej vzorky sme takto urobili 500 nových vzoriek, na ktorých sme spustili náš simulačný algoritmus. Musíme podotknúť, že výpočet bol časovo náročný, pre jednu päťstovku (jeden locus v jednom kraji) trval výpočet približne 9 hodín. Výsledkom výpočtu bolo 500 modusov pre každú kombináciu locus – kraj (dokopy sme mali 24 takýchto kombinácií).

Členy postupnosti 500 modusov sme si potom usporiadali podľa veľkosti a dostali sme (pre každú dvojicu kraj – locus) obrázky ako ten nasledujúci.



Obrázok 3.10. Usporiadaná postupnosť 500 modusov na locuse $TH01$ v Banskobystrickom kraji

Vidíme, že modus sa výrazne mení vplyvom zmeny dát vo vzorke. Pomocou metódy bootstrap však vieme pekne odhadnúť interval, v ktorom sa skutočný modus (teda skutočná „reprezentatívna“ hodnota F) nachádza. Pre každý kraj a locus sme si tento interval popísali kvantilmi získanými z postupnosti spomínaných 500 modusov usporiadaných podľa veľkosti.

Kraj / Locus	5%-ný kvantil	priemer	medián	95%-ný kvantil
Bratislavský				
VWA	0.0035	0.0095	0.0080	0.0196
FGA	0.0025	0.0098	0.0086	0.0219
$TH01$	0.0043	0.0114	0.0098	0.0241
Banskobystrický				
VWA	0.0028	0.0076	0.0063	0.0158
FGA	0.0046	0.0136	0.0126	0.0256
$TH01$	0.0036	0.0097	0.0084	0.0216
Košický				
VWA	0.0050	0.0144	0.0130	0.0275
FGA	0.0024	0.0088	0.0075	0.0195
$TH01$	0.0053	0.0137	0.0125	0.0251
Nitriansky				
VWA	0.0029	0.0084	0.0069	0.0184
FGA	0.0031	0.0103	0.0093	0.0202
$TH01$	0.0035	0.0086	0.0074	0.0175
Prešovský				
VWA	0.0033	0.0098	0.0082	0.0208
FGA	0.0031	0.0118	0.0106	0.0234
$TH01$	0.0041	0.0085	0.0069	0.0176
Trnavský				
VWA	0.0046	0.0101	0.0086	0.0201
FGA	0.0063	0.0193	0.0183	0.0354
$TH01$	0.0038	0.0099	0.0080	0.0216
Žilinský				
VWA	0.0024	0.0063	0.0053	0.0138
FGA	0.0030	0.0102	0.0089	0.0210
$TH01$	0.0031	0.0084	0.0073	0.0172

Tabuľka 3.11. Rozmedzie hodnôt F meniacich sa v závislosti od vzorky

Kraj / Locus	5%-ný kvantil	priemer	medián	95%-ný kvantil
Trenčiansky				
VWA	0.0032	0.0096	0.0084	0.0212
FGA	0.0033	0.0115	0.0098	0.0251
$TH01$	0.0053	0.0137	0.0123	0.0264

Tabuľka 3.12. Rozmedzie hodnôt F meniacich sa v závislosti od vzorky

Vďaka údajom v tabuľke vieme nadobudnúť trochu lepšiu predstavu o skutočnej hodnote F ako pomocou použitia hodnoty modusu z jedinej simulácie, nakoľko modus je len bodový odhad a údaje v tabuľke popisujú interval. Vidíme, že najväčší 95%-ný kvantil (horný odhad) zo všetkých locusov a krajov je v Trnavskom kraji na locuse FGA a dosahuje hodnotu vyše 0.035 (v článkoch sa často používa aj percentuálne vyjadrenie 3.5%).

V praxi (t.j. v západnej Európe a USA, lebo u nás sa koeficient príbuznosti zatiaľ zanedbáva) sa na výpočet value of evidence používa hodnota 3%. Je to akýsi kompromis medzi veľkým poškodením obžalovanému (ak by sme F úplne zanedbali) a prehnanou pomocou jeho obhajobe. Tento údaj však vychádza stále z lokálnych dát a preto nemôžeme s istotou povedať, že na Slovensku by to bola tiež správna hodnota. Pri pohľade na tabuľku by sme sa možno skôr priklonili k hodnotám okolo 2.5%.

Všeobecne však pri pohľade na tabuľku vidíme, že aj maximálne hodnoty modusu F sú relatívne malé čísla. Tu vyvstáva otázka, či takéto hodnoty koeficientu príbuznosti sú naozaj „malé“ a teda, či ich môžeme zanedbať. Alebo môžeme túto otázku formulovať aj praktickejšie: aká veľká chyba sa môže stať pri vyčísľovaní value of evidence, ak koeficient príbuznosti zanedbáme? Na túto otázku odpovieme o pár strán ďalej v časti 3.6.

3.5 Druhý model pre F

Na odhad koeficientu príbuznosti pomocou Metropolisovho – Hastingsovho algoritmu sme vyskúšali použiť aj trochu iný model uvedený v článku [5]. V tomto modeli nevieme presne pomenovať alebo pekne analyticky vyjadriť apriórnu hustotu parametra F ako takého[7], ale podľa článku [5] predpokladáme, že koeficient príbuznosti spĺňa pre daný kraj i a locus j vzťah

$$F_{ij} = \frac{1}{1 + \alpha_i + \beta_j}, \tag{3.8}$$

[7] V minulom modeli sme ju poznali, bolo to $\beta(1.5, 50)$.

kde α_i a β_j sú lognormálne rozdelené a ich logaritmus má strednú hodnotou 3.5 a disperziu 1.5. Tieto dva parametre sú navzájom nezávislé a prestavujú vplyv kraja a locusu. Niekomu by sa možno zdalo, že v tomto modeli niečo chýba. Podľa skúseností z iných modelov by sme čakali, že bude prítomný aj člen vyjadrujúci interakciu medzi prvými dvoma. V článku [5] autor hovorí, že jeho pôvodný model obsahoval aj interakčný člen γ_{ij}, no rozhodol sa ho z modelu vylúčiť, nakoľko jeho hodnoty boli zanedbateľne malé (z nasledujúceho popisu prvých dvoch členov to aj logicky vyplýva).

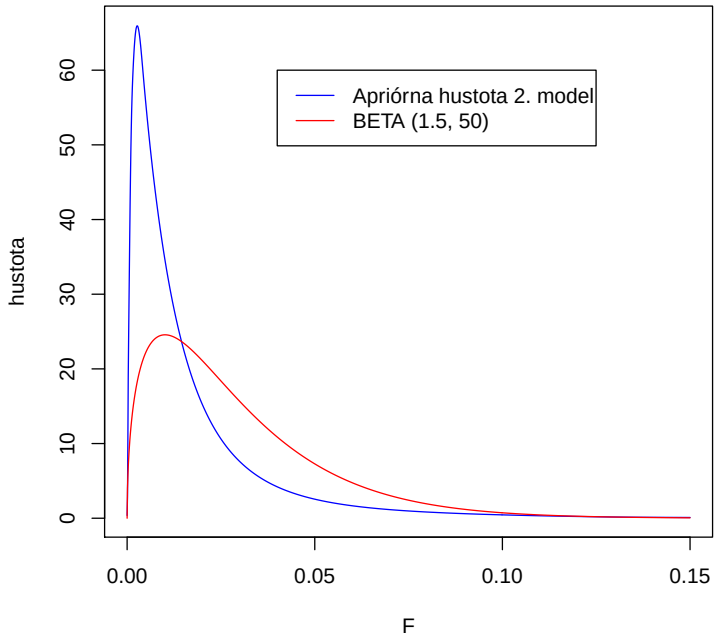
Vplyv kraja α_i definujeme v zmysle miery migrácie jeho obyvateľov a teda pribúdania a odbúdania genotypov jedincov. Všimnime si, že ak miera migrácie v kraji stúpne, prichádza veľa nových ľudí, a teda pribúdajú nové genotypy (alebo viac ľudí odchádza a tým odchádzajú aj „podobné“ genotypy). Týmto sa populácia kraja stáva rôznorodejšou – koeficient príbuznosti v danom kraji preto klesá. Ak sa pozrieme va vzťah (3.8), vidíme, že zväčšujúce sa číslo α_i v menovateli naozaj zníži hodnotu F.

Vplyv locusu β_j sa spája s mutáciu génov. Jednoduché vysvetlenie tohto vplyvu je, že počas vzniku nového jedinca spojením genotypu matky a otca môže nastať mutácia, napríklad alela 9 sa môže zmeniť na 9.5 alebo 10. Týmto môže vzniknúť nový druh alely, ktorý sa bude v ďalšej generácii dediť a tým spestrí množinu možných genotypov v kraji – koeficient príbuznosti teda klesne. Vo vzorci znova vidíme, že takéto vysvetlenie je korektné aj matematicky, pretože ak sa zväčší β_j, zmenší sa F.

Apriórnu hustotu parametra F v tomto modeli nevieme analyticky vyjadriť. Pre lepšiu predstavu si ju nasimulujeme a porovnáme s apriórnou hustotou z pôvodného modelu. Simulovať budeme nasledovne:

- Vygenerujeme α_i a β_j a následne z neho vypočítame F podľa (3.8).
- Toto zopakujeme 100000 krát a zostrojíme histogram a následne odhadneme hustotu.

Ak výslednú hustotu porovnáme s apriórnou hustotou z prvého modelu, dostaneme nasledovný obrázok. Vidíme, že oproti prvému modelu je táto apriórna hustota koncentrovaná bližšie k nule (teda má menší modus).



Obrázok 3.11. Porovnanie apriórnych hustôt dvoch modelov

Čo sa týka samotného simulovania pomocou tohto modelu, Metropolisov – Hastingsov algoritmus je rovnaký ako v minulom prípade, až na dva rozdiely:

1. Namiesto priameho generovania novej hodnoty parametra F^{new} z rozdelenia Q, generujeme dva parametre α_i^{new} a β_j^{new} z dvoch nezávislých normálnych rozdelení Q_1 a Q_2 so strednými hodnotami rovnými α_i a β_j z predošlého kroku a disperziou prispôsobenou s ohľadom na funkčnosť algoritmu (aby prijímal nové hodnoty dosť často). Následne z nich vyjadríme F^{new} podľa (3.8), pomocou ktorého vypočítame hodnotu výsledného člena rekurencie (3.6). Toto je najdôležitejší poznatok – musíme si uvedomiť, že F je v tomto prípade len akýsi „bočný produkt“, teda výsledok výpočtu, kým priamo generované hodnoty sú α_i a β_j.
2. V závere algoritmu, keď vyhodnocujeme podiel (3.7), tak za výrazy $h(F^{new})$ a $h(F)$ dosadíme namiesto hodnoty hustoty $\beta(1.5, 50)$ v bodoch F^{new} a F
   - súčin hustôt dvoch lognormálnych rozdelení so strednou hodnotou 3.5 a disperziou 1.5 v bodoch α_i^{new} a β_j^{new} ako $h(F^{new})$,
   - súčin hustôt dvoch lognormálnych rozdelení so strednou hodnotou 3.5 a disperziou 1.5 v bodoch α_i a β_j ako $h(F)$.

Obe tieto zmeny oproti prvému modelu vyplývajú priamo z definície algoritmu v časti 3.3.1, pričom sa jedná o rozdiely spôsobené zmenou v popise apriórnej hustoty rozdelenia $D(x)$. Pre úplnosť ešte dodáme presný popis parametrov:

Kraj / Locus	parameter
Bratislavský	α_1
Banskobystrický	α_2
Košický	α_3
Nitriansky	α_4
Prešovský	α_5
Trenčiansky	α_6
Trnavský	α_7
Žilinský	α_8
VWA	β_1
FGA	β_2
$TH01$	β_3

Tabuľka 3.13. Priradenie parametrov ku krajom a locusom v druhom modeli

3.5.1 Použitie modelu

Pri predošlom modeli sme spomínali, že sme ho použili na odhad koeficientu príbuznosti pre dvojice locus – kraj a ako jeho vstup sme použili vzorku daného locusu v danom kraji. Takto sme pomocou neho postupne samostatne odhadli 24 parametrov F pre tri locusy v ôsmich krajoch (model bol saturovaný). Všimnime si, že v druhom modeli, keďže neodhadujeme F priamo, ale pomocou α_i a β_j, je počet parametrov menší, presnejšie ich bude 11 (8 parametrov α_i pre kraj plus 3 parametre β_j pre locusy). Z tohto vyplýva, že odhady koeficientu príbuznosti pre dvojice locus – kraj pomocou druhého modelu budú závislé (kým tie z prvého boli navzájom nezávislé).

Ďalšia vec, ktorú treba spomenúť je, že už nemôžeme na odhad použiť len vybranú časť vzorky, ako tomu bolo v prvom modeli (napríklad pre odhad F v Košickom kraji na locuse FGA sme použili len vzorku dát o locuse FGA z Košického kraja). Vysvetlíme si to na príklade.

Nech α_1 prislúcha Bratislavskému kraju a β_1 locusu VWA. Povedzme, že na odhad použijeme len dáta o locuse VWA z Bratislavského kraja a odhadneme tak α_1 a β_1. Potom ale zoberieme ďalší locus FGA v tomto kraji a chceme odhadnúť koeficient príbuznosti. Ak odhadneme α_1 a β_2, zistíme, že pre α_1 už máme dva odhady a nevieme ich nijako porovnať, ani posúdiť, ktorý je lepší. Z tohto príkladu je jasné, že tento model nebudeme môcť použiť z hľadiska dát presne rovnako, ako ten prvý. Nemôžeme odhadovať parametre α_i a β_j postupne, nakoľko nevieme určiť správne poradie ich odhadovania (tento problém sme pri prvom modeli nemali, nakoľko tam boli odhadované parametre

navzájom nezávislé).

Riešením tohto problému je odhadovať všetky parametre naraz v jednej simulácii. Takto nebudeme musieť dávať pozor na poradie dát, keďže do modelu vstúpia všetky dáta naraz, a zároveň využijeme informáciu z nich v najväčšej možnej miere. V prvom modeli, keď sme na odhad napr. koeficientu príbuznosti na locuse FGA v Košickom kraji, sme použili *len* dáta o FGA z Košického kraja. Tu použijeme (lebo vstúpia do modelu naraz) *všetky dáta o locuse* FGA a *všetky dáta z Košického kraja.*

Algoritmus zostavený podľa kritérií z predošlých dvoch odstavcov bude pracovať nasledovne:

1. Vygeneruje počiatočné hodnoty $\alpha_1, \ldots, \alpha_8$ a $\beta_1, \ldots, \beta_3$ z lognormálneho rozdelenia so strednou hodnotou 3.5 a disperziou 1.5 a následne vypočíta 24 začiatočných hodnôt F_{ij}.
2. Vygeneruje nové hodnoty $\alpha_1^{new}, \ldots, \alpha_8^{new}$ a $\beta_1^{new}, \ldots, \beta_3^{new}$, každú z normálneho rozdelenia so strednou hodnotou rovnou predošlej hodnote daného parametra a disperziou[8] 15, a dopočíta F_{ij}^{new}.
3. Vygeneruje hodnotu α z rovnomerného rozdelenia na na intervale $[0, 1]$.
4. Porovná hodnotu α a výrazu

$$V = \frac{D(F_{ij}^{new}) \prod_{i=1}^{8} Q(\alpha_i^{new}, \alpha_i) \prod_{j=1}^{3} Q(\beta_j^{new}, \beta_j)}{D(F_{ij}) \prod_{i=1}^{8} Q(\alpha_i, \alpha_i^{new}) \prod_{j=1}^{8} Q(\beta_j, \beta_j^{new})},$$

kde výrazy Q nás nezaujímajú (vykrátia sa, pretože sú to znova symetrické normálne rozdelenia) a $D(x)$ je aposteriórne rozdelenie vektora $(\alpha_1, \ldots, \alpha_8, \beta_1, \beta_2, \beta_3)$ – súčin apriórneho rozdelenia a členov, ktoré zodpovedajú korekcii dátami. Rozpíšme si ho podrobnejšie.

$$D(F_{ij}) = \prod_{i=1}^{8} \text{lognorm}(\alpha_i, 3.5, 1.5) \cdot \prod_{j=1}^{3} \text{lognorm}(\beta_j, 3.5, 1.5) \cdot \prod_{i=1}^{8} \prod_{j=1}^{3} P_N^{ij},$$

kde P_N^{ij} je rekurentný súčin (3.6) pre dáta j-teho locusu v i-tom kraji a $\text{lognorm}(\alpha_i, 3.5, 1.5)$ je hustota lognormálneho rozdelenia s parametrami 3.5 a 1.5 v bode α_i.

Apriórne rozdelenie je teda súčin jedenástich hustôt lognormálnych rozdelení s parametrami 3.5 a 1.5 v bodoch $\alpha_1^{new}, \ldots, \alpha_8^{new}$ a $\beta_1^{new}, \ldots, \beta_3^{new}$ v prípade $D(F_{ij}^{new})$, a súčin jedenástich takýchto hustôt v bodoch $\alpha_1, \ldots, \alpha_8$ a $\beta_1, \ldots, \beta_3$ v prípade $D(F_{ij})$. Korekciu

[8] Ako v prípade prvého modelu, číslo zvolené kvôli funkčnosti algoritmu, aby prijímal nové hodnoty parametrov dosť často.

dátami urobíme podľa vzorca (3.6) pre každú kombináciu parametrov α_i a β_j, teda pre každé F_{ij} s časťou dát prislúchajúcou danej kombinácii kraj – locus (i, j). Bude to teda súčin 24 skupín zlomkov násobiacich sa navzájom ako v (3.6).

5. Ak hodnota výrazu V je väčšia ako hodnota α, algoritmus považuje všetky α_i^{new} a β_j^{new} (a teda aj F_{ij}^{new}) za vyhovujúce a pridá F_{ij}^{new} do postupností hodnôt za posledné F_{ij}.
6. Ak je hodnota výrazu V menšia ako α, nové hodnoty sú nevyhovujúce a ako ďalšie členy pridá do postupností znova hodnoty predošlých F_{ij}.

Vidíme, že idea a aj funkčnosť algoritmu ostala rovnaká ako v prvom modeli, zmenil sa len jeho vstup.

3.5.2 Výsledky

Na nasledujúcich stranách v krátkosti zhrnieme výsledky, ktoré sme dostali použitím tohto modelu, nakoľko spôsob, ako sa na ne pozerať a ich možnú interpretáciu sme kompletne ukázali pri tom predošlom. Začneme odhadnutými modusmi (tabuľka 3.14), ktoré aj napriek hojnému použitiu v článkoch podľa nás nemajú sami osebe veľkú výpovednú hodnotu.

Kraj / Locus	VWA	FGA	$TH01$
Bratislavský	0.00379	0.00168	0.00169
Banskobystrický	0.00243	0.00478	0.00174
Košický	0.00116	0.00178	0.00100
Nitriansky	0.00213	0.00344	0.00169
Prešovský	0.00254	0.00585	0.00169
Trenčiansky	0.00233	0.00392	0.00158
Trnavský	0.00245	0.00580	0.00169
Žilinský	0.00241	0.00373	0.00168

Tabuľka 3.14. Modusy F podľa druhého modelu

Vidíme, že nemožno urobiť jednoznačný záver čo sa týka porovnania číselného výsledku dvoch metód, pretože v niektorých prípadoch sú modusy odhadnuté druhou metódou väčšie ako pri tej prvej (tabuľka 3.9) a inokedy menšie. Viac je ale prípadov, keď sú menšie. Oproti výsledkom prvého modelu chýba v tabuľke stĺpec s „celkovým F“, pretože počítať ho v tomto prípade nemá zmysel. Definícia modelu (3.8) totiž predpokladá automaticky rôzne alfy a bety, a teda žiadne rovnaké koeficienty nepripúšťa.

Ak by modusy vyšli pre všetky locusy a všetky kraje približne rovnako, malo by zmysel uvažovať o jedinej hodnote α a jedinej hodnote β a tým pádom o jednom spoločnom F. To by ale znamenalo, že koeficient príbuznosti je rovnaký pre všetky kraje a pre všetky locusy. Mohli by sme ho smelo zanedbať, nakoľko by nebola potreba korigovať vzorce z druhej kapitoly (frekvencie alel na locusoch by boli na celom území Slovenska rovnaké a ich odhad by bol teda platný pre všetky kraje a locusy[9]). Už výsledok prvého modelu však nasvedčoval tomu, že koeficient príbuznosti nie je jedno pevné číslo, takže sa touto alternatívou nebudeme ďalej zaoberať.

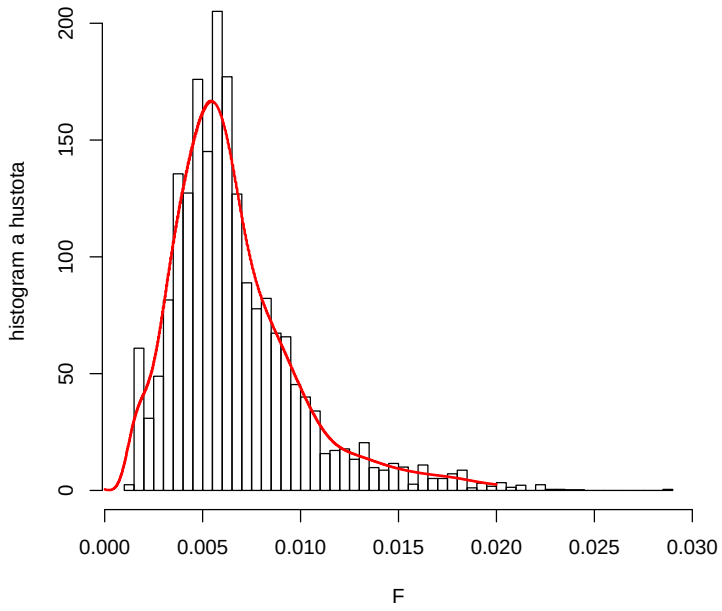
Poďme sa teraz pozrieť na odhadnuté hustoty. Na ich odhad sme nemohli použiť splajny ako v prvom modeli, pretože histogramy boli príliš členité a odhadnutá krivka vyzerala veľmi škaredo. Po vyskúšaní viacerých metód sme nakoniec napísali profesorovi Davidovi J. Baldingovi, autorovi článkov [5] a [4], aby nám prezradil, ako zobrazoval výsledky on. Podľa jeho rady sme napokon použili nasledovnú transformáciu:

- Umocnili sme odhadnuté hodnoty F na $2/3$ (označme $F^{2/3} = Z$), aby sme z histogramu F, ktorý je asymetrický (má pozitívny sklon) dostali symetrický histogram. Jadrové odhady, ktoré softvér R používa, totiž fungujú lepšie na symetrických rozdeleniach.
- Odhadli sme hustotu h_Z pomocou funkcie *density* v softvéri R [9].
- Spätnou transformáciou sme získali aposteriórnu hustotu h_F v tvare

$$h_F(F) = h_Z(F^{2/3}) \cdot \frac{2}{3} \cdot F^{-1/3}.$$

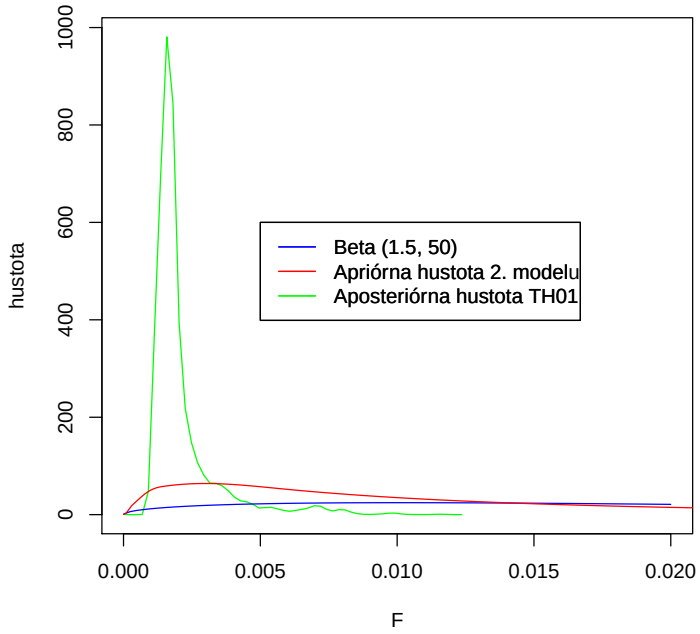
Takto sme postupne získali funkcie hustôt jednotlivých koeficientov príbuznosti. Výsledná krivka a histogram pre locus FGA v Banskobystrickom kraji je na nasledujúcom obrázku.

[9]Odhady frekvencií dosadené do vzorcov na výpočet value of evidence korigujeme koeficientom príbuznosti práve preto, že frekvencie nie sú v každom kraji a na každom locuse rovnaké a teda ich odhad z celoslovenských dát nie je ich skutočná hodnota. Ak by ale boli na celom území Slovenska frekvencie rovnaké, ich odhad by bol správna hodnota a nebola by potrebná korekcia.



Obrázok 3.12. Histogram a hustota F na locuse FGA v Banskobystrickom kraji

Na ďalšom obrázku vidíme porovnanie apriórnej a aposteriórnej odhadnutej hustoty. Pre lepšiu predstavu je tam znázornená aj apriórna hustota z prvého modelu. Vidíme, že využitím maximálnej informácie z dát sa nám pomocou druhého modelu podarilo veľmi zúžiť kopec aposteriórneho rozdelenia, čo znamená presnejší odhad koeficientu príbuznosti.

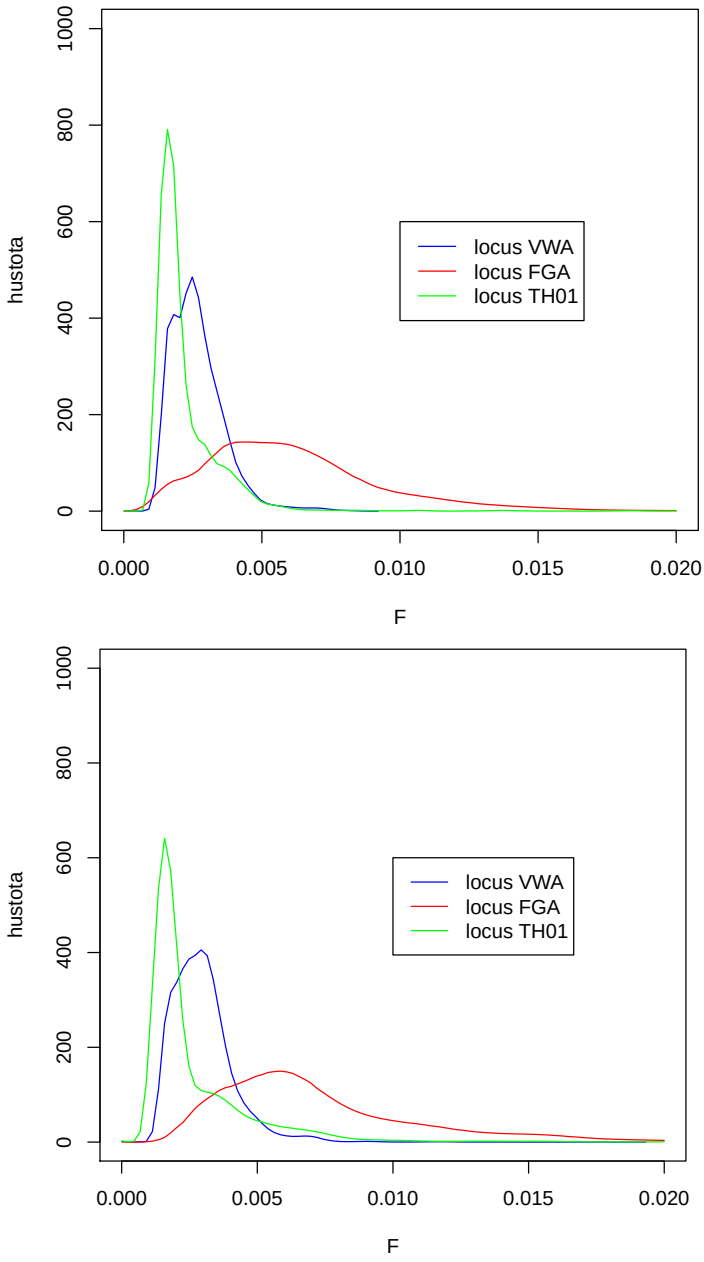


Obrázok 3.13. Porovnanie apriórnych hustôt dvoch modelov a aposteriórnej hustoty F na locuse $TH01$ v Bratislavskom kraji

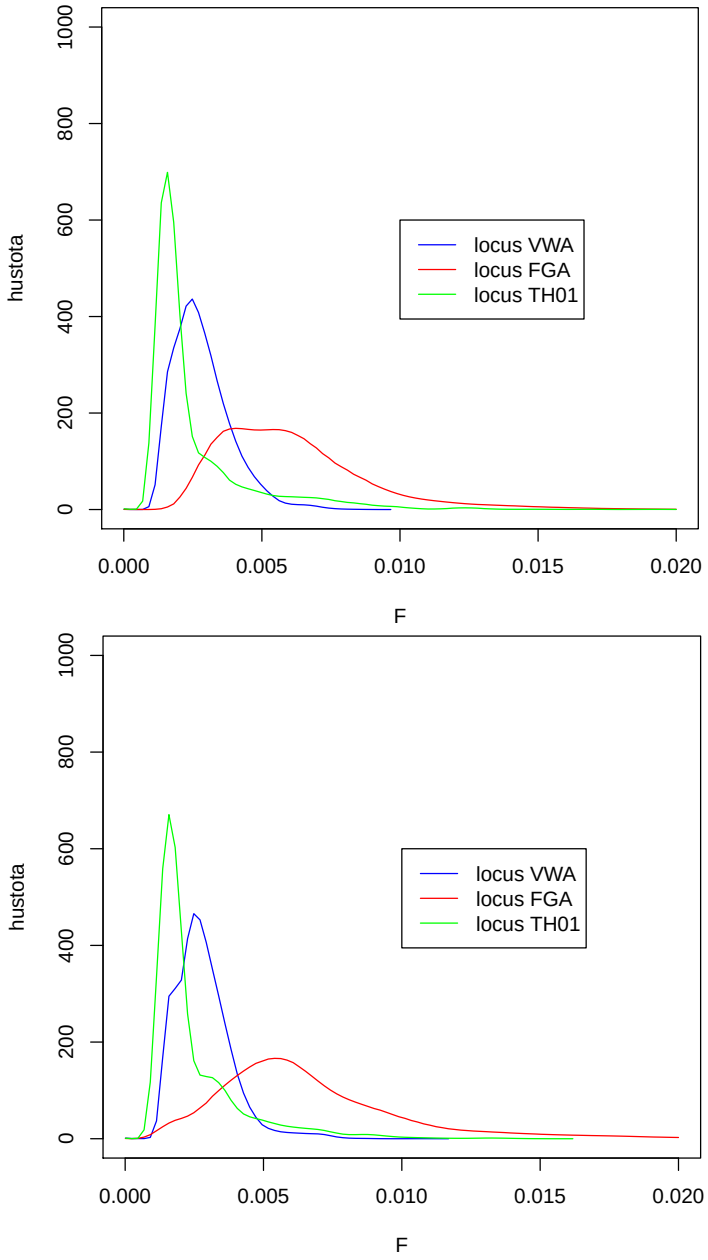
Rovnako ako pri prvom modeli sme sa pozerali najskôr na to, ako vyzerá situácia v rámci krajov – či daný kraj má navzájom podobné modusy a hustoty koeficientu príbuznosti na troch skúmaných locusoch. Zistili sme, že to tak nie je. Locus FGA je vo väčšine krajov výrazne iný ako zvyšné dva. Zaujímavé je aj to, že obrázky krajov sú si navzájom veľmi podobné (okrem Košického kraja), čo sa týka modusov ale aj tvarov hustôt pre jednotlivé locusy. Už aj v tabuľke 3.14 sme si mohli všimnúť, že modusy sú pre VWA a $TH01$ medzi krajmi navzájom veľmi podobné.

Tieto vlastnosti odhadnutých hustôt a modusov sa dajú vysvetliť vzájomným vzťahom parametrov α_i a β_j. Všimnime si, že na locusoch VWA a $TH01$ sú odhadnuté modusy takmer rovnaké. Dôvodom sú veľké β_1 (VWA) a β_3 ($TH01$), ktoré prevážia vplyv krajských α_i. Z tohto vidíme, že vplyv locusu na výsledný koeficient príbuznosti je veľmi podstatný, pri týchto dvoch locusoch oveľa podstatnejší ako vplyv kraja. Záver z tohto pozorovania je nasledujúci: podľa druhého modelu môžeme hodnoty koeficientu príbuznosti na locusoch VWA a $TH01$ považovať za rovnaké na celom území Slovenska.

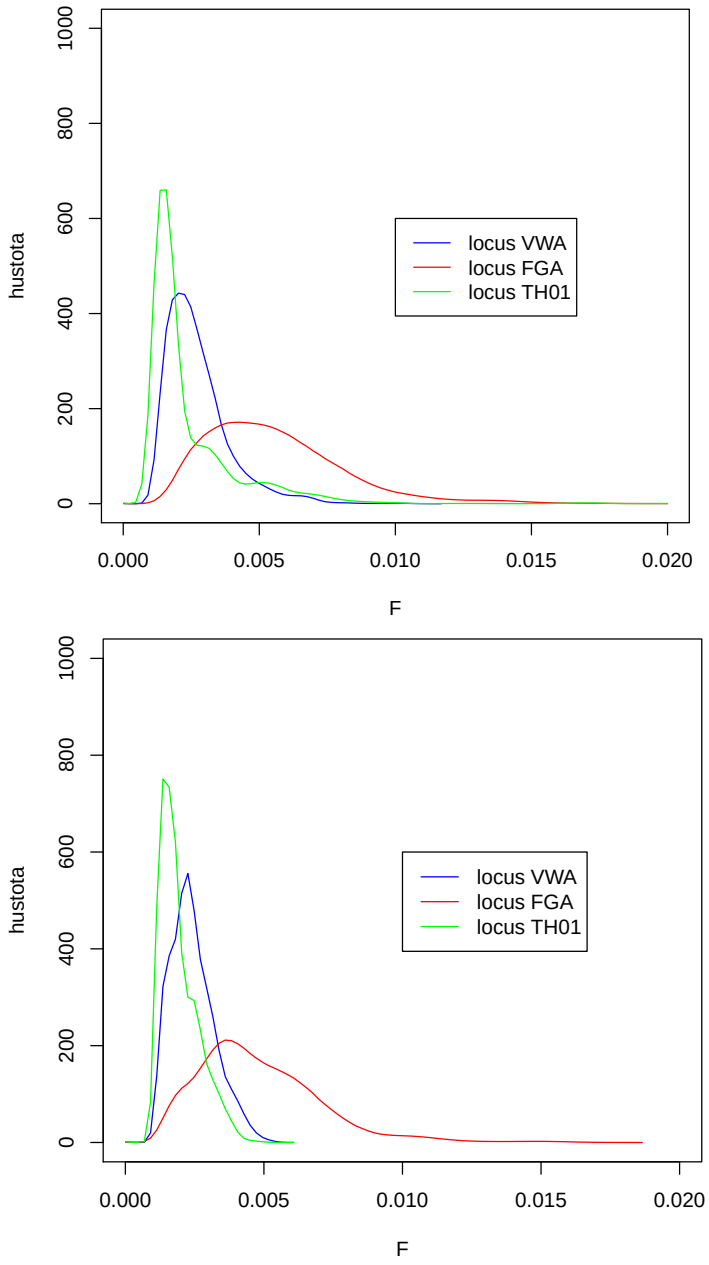
Oproti tomu na locuse FGA vidíme rôzne modusy. Toto vieme vysvetliť tak, že β_2 je malé číslo a krajské vplyvy α_i ho prevážia. To znamená, že na locuse FGA sú rôzne hodnoty koeficientu príbuznosti v rámci slovenských krajov. Na nasledujúcich obrázkoch vidíme odhadnuté aposteriórne hustoty F v ôsmich slovenských krajoch.



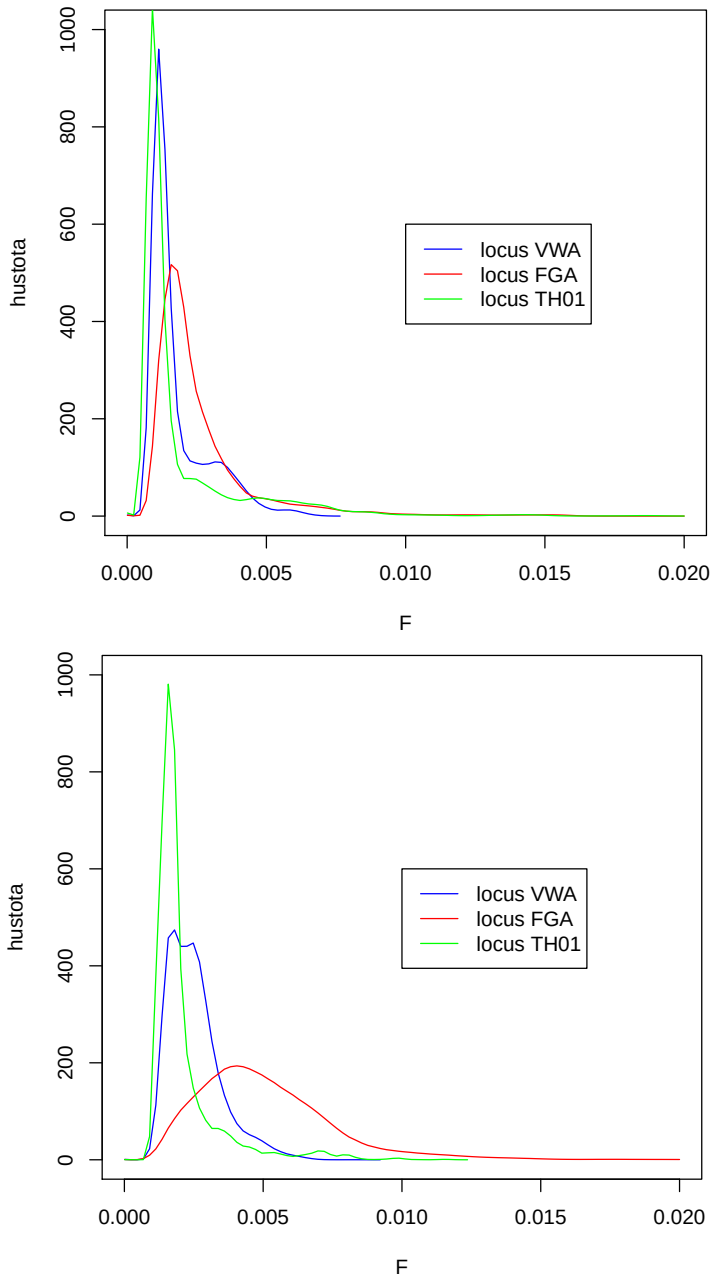
Obrázok 3.14. Žilinský (hore) a Trnavský (dole) kraj



Obrázok 3.15. Prešovský (hore) a Banskobystrický (dole) kraj



Obrázok 3.16. Trenčiansky (hore) a Nitriansky (dole) kraj



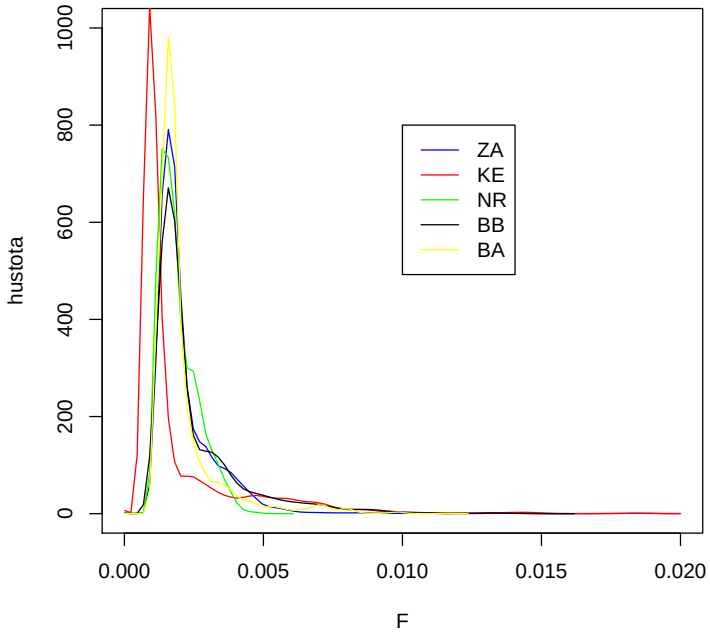
Obrázok 3.17. Košický (hore) a Bratislavský (dole) kraj

Všimnime si ešte aposteriórne hustoty pre Košický kraj. Vidíme, že na všetkých troch locusoch sú hustota a aj modus navzájom veľmi podobné. Toto znamená veľký vplyv kraja, čiže veľké α_3, ktoré preváži vplyv locusov β_j. Aj v tabuľke 3.14 vidíme, že modusy pre tri skúmané locusy sú v Košickom kraji veľmi podobné.

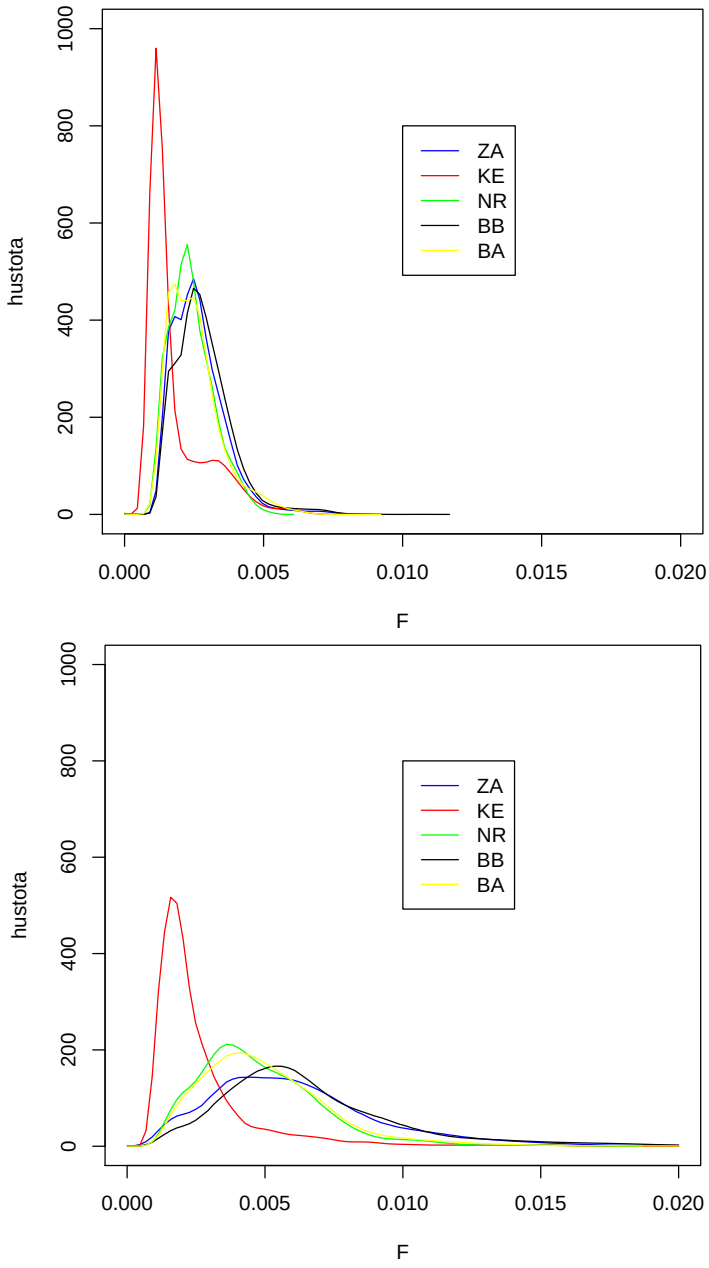
Rovnako ako pri prvom modeli sme sa pozerali nielen na porovnanie locusov v rámci kraja, ale aj na porovnanie krajov, ak ide o ten istý locus. To, čo sme zistili, znova

približne korešponduje s tabuľkou modusov – pre locusy VWA a $TH01$ je vplyv kraja (α_i) prevážený vplyvom locusu, kým pri locuse FGA sa kraje od seba dosť líšia. Je to spôsobené malou hodnotou β_2 pre locus FGA, ktorá je menšia ako jednotlivé krajské α_i (preto sa vplyv kraja výrazne prejavil).

Všimnime si tiež červené krivky (Košický kraj), ktoré sú výrazne naľavo od ostatných pre každý locus. Tu vidíme znova vplyv α_3, ktorý je väčší ako vplyvy ostatných krajov, čo spôsobuje koncentráciu apriórnej hustoty koeficientov príbuznosti pre tento kraj bližšie k nule.



Obrázok 3.18. Locus $TH01$



Obrázok 3.19. Locusy VWA (hore) a FGA (dolu)

Analýze nepresnosti simulačného odhadu pomocou metódy bootstrap sme sa nevenovali, nakoľko by to bolo výpočtovo a časovo náročné a vieme si predstaviť, že výsledky by boli podobné. Spočítali sme ale výberové disperzie nagenerovaných koeficientov príbuznosti vo dvojiciach locus – kraj. Výsledky sú v nasledujúcej tabuľke a sú rádu 10^{-6}. Z tohto vyplýva, že odhady pomocou druhej metódy by mali byť presnejšie (nakoľko majú menšiu disperziu).

Kraj / Locus	VWA	FGA	$TH01$
Bratislavský	$1.418 \cdot 10^{-6}$	$5.113 \cdot 10^{-6}$	$2.916 \cdot 10^{-6}$
Banskobystrický	$1.680 \cdot 10^{-6}$	$9.963 \cdot 10^{-6}$	$4.396 \cdot 10^{-6}$
Košický	$2.321 \cdot 10^{-6}$	$5.694 \cdot 10^{-6}$	$6.351 \cdot 10^{-6}$
Nitriansky	$96.061 \cdot 10^{-6}$	$4.475 \cdot 10^{-6}$	$79.793 \cdot 10^{-6}$
Prešovský	$1.677 \cdot 10^{-6}$	$7.456 \cdot 10^{-6}$	$0.107 \cdot 10^{-6}$
Trenčiansky	$1.951 \cdot 10^{-6}$	$6.029 \cdot 10^{-6}$	$9.408 \cdot 10^{-6}$
Trnavský	$1.642 \cdot 10^{-6}$	$0.138 \cdot 10^{-6}$	$7.388 \cdot 10^{-6}$
Žilinský	$1.208 \cdot 10^{-6}$	$8.536 \cdot 10^{-6}$	$2.547 \cdot 10^{-6}$

Tabuľka 3.15. Výberová disperzia odhadu F podľa druhého modelu

3.5.3 Porovnanie s prvým modelom

Pri celkovom pohľade na porovnanie druhej metódy s tou prvou vidíme dva podstatné rozdiely, ktoré sa v tejto časti pokúsime vysvetliť.

Menšie modusy. Mohli by sme ich vysvetliť jednoducho obrázkom 3.11, na ktorom vidíme, že už aj apriórne rozdelenie druhého modelu je koncentrované bližšie k nule ako $\beta(1.5, 50)$ z prvého modelu. Preto nie je prekvapivé, že ak sa vplyvom dát zúži kopec výslednej aposteriórnej hustoty a modus sa posunie ešte bližšie k nule, bude menší ako výsledný modus z prvého modelu. Ďalej, už sme si povedali, že dodatočná informácia zužuje interval, v ktorom predpokladáme prítomnosť skutočnej hondoty F. Nakoľko druhá simulácia odhaduje všetky parametre naraz, využíva maximálnu možnú informáciu, ktorú dáta vedia poskytnúť. Prvý model už zo spôsobu svojho použitia berie do úvahy menšie množstvo informácie. Ak teda platí „viac informácie znamená užší kopec“, je aj tento fakt dôvodom, prečo je výsledný modus v druhom modeli menší ako ten v prvom.

Menšia disperzia[10]. Disperzia sa spája s presnosťou odhadu, a teda by sme mali predpokladať, že odhad pomocou druhého modelu je presnejší. (Analýzy o kolísaní modusu pri opakovanom spustení simulácie sme nerobili.) Oproti tomu, druhý model odhaduje parametre, ktorých je menej ako pri prvom modeli a sú závislé – tento fakt hovorí proti

[10] Aby sme mohli porovnávať výsledky modelov z hľadiska disperzie, musíme skontrolovať, či apriórne hustoty pri oboch modeloch mali rovnakú disperziu. Pri prvom modeli (rozdelenie $\beta(1.5, 50)$) bola výberová disperzia 0.0005383886, kým pri druhom modeli to bolo 0.0006410189. Vidíme, že aj keď disperzia apriórneho rozdelenia bola trochu menšia pri druhom modeli, váha dát ju zmenšila tak veľmi, že výberová disperzia výsledného odhadu bola desaťkrát menšia ako v prvom modeli. Toto nás len utvrdzuje v tom, že miera korekcie dátami a teda využitá informácia je v druhom modeli väčšia.

názoru, že odhad pomocou druhého modelu by mal byť presnejší. Rozhodujúca je ale aj miera informácie, ktorá sa z dát využíva a jej rozsah vieme porovnať, nakoľko oba modely využívajú tie isté dáta, len rôznym spôsobom. Máme teda dva argumenty, prečo by druhý model mal mať presnejší odhad (disperzie a informáciu) a jeden argument, prečo by presnejší mal byť ten prvý (počet parametrov). Prikláňame sa k názoru, že práve miera využitej informácie robí z druhého modelu ten presnejší, samozrejme, ak druhý model platí pre populácie slovenských krajov.

Ohľadom platnosti druhého modelu vieme argumentovať nasledovne. O prvom modeli sme vedeli, že mal 24 nezávislých parametrov (bol saturovaný). O saturovanom modeli je známe, že platí vždy. Nakoľko výsledné odhady a hustoty z druhého modelu sa od prvého veľmi výrazne líšia, vedie to k podozreniu, že model z článku [5] v slovenských pomeroch pre krajské populácie asi neplatí (čo je škoda, lebo jeho výsledky boli veľmi pekne interpretovateľné).

3.6 Vplyv F na value of evidence

V tejto časti si ukážeme chybu, ktorá vznikne pri výpočte value of evidence, ak zanedbáme koeficient príbuznosti. Pre tento účel si vytvoríme vlastnú populáciu ľudí podľa návodu v článku [5], ktorá má koeficient príbuznosti rovný 0.01 (môžeme si ho zvoliť akýkoľvek). Vlastnú populáciu si tvoríme preto, aby sme hodnotu F v nej presne poznali a vedeli s ňou počítať.

Z takto vytvorenej populácie vyberieme sto jedincov s náhodnými genotypmi a porovnáme pre ne pravdepodobnosti náhodnej zhody (prevrátené hodnoty value of evidence v prípade klasickej nulovej hypotézy ako v (2.1)) vypočítané bez a s koeficientom príbuznosti. U jedinca z našej populácie budeme skúmať genotyp pozostávajúci zo štyroch locusov, z ktorých každý môže obsahovať dve z 15 možných alel. Pri generovaní jedincov budeme postupovať nasledovne (zatiaľ pre prvý locus):

- Vygenerujeme pravdepodobnosti p_i, s ktorými sa každá z 15 alel nachádza na danom locuse tak, aby mali rovnomerné rozdelenie na intervale $[0, 1]$ a súčet 1 (pretože sú to frekvencie).

Keďže nevieme priamo vygenerovať 15 čísel z rovnomerného rozdelenia tak, aby mali súčet 1, vygenerujeme na intervale $[0, 1]$ štrnásť bodov a ako frekvencie zoberieme dĺžky úsekov medzi nimi.

- Zahrnieme vplyv koeficientu príbuznosti. Podľa návodu v článku [5] použijeme Dirichletovo $D(\alpha_1, \ldots, \alpha_{15})$ rozdelenie, kde prvky vektora dostaneme ako $\alpha_i =$

$p_i \cdot \alpha_0$ a $\alpha_0 = 1/F - 1 = 99$. Potom zo spomínaného Dirichletovho rozdelenia vygenerujeme jeden vektor – frekvencie alel $\hat{p}_i$.

- Z rozdelenia, kde alela i má frekvenciu $\hat{p}_i$ teraz vygenerujeme dve alely pre každého zo sto jedincov tak, aby každý bol heterozygot (druhú vygenerovanú alelu porovnávame s tou prvou a ak sa rovnajú, generujeme ju znova – takto postupujeme vo všetkých 100 prípadoch). Prípad homozygota je príliš náročný (je tam možnosť tzv. nulových alel[11]), preto ho neskúmame.
- Rovnaký postup zopakujeme aj pre ďalšie tri locusy, čím dostaneme sto heterozygotných genotypov pozostávajúcich z ôsmich alel. Pre každý locus teda vygenerujeme najskôr frekvencie p_i (bez vplyvu koeficientu príbuznosti), potom frekvencie $\hat{p}_i$ (upravené o koeficient príbuznosti) a napokon s takto určenými frekvenciami vygenerujeme alely pre jedincov. Z teórie podľa [5] vyplýva, že v takto vygenerovanej populácii sa genotypové frekvencie neriadia Hardyho – Weinbergovým ekvilibriom, ale odrážajú nenulový koeficient príbuznosti (u nás rovný 0.01).

Pre každý z týchto genotypov budeme teraz skúmať pravdepodobnosť náhodnej zhody. Predpokladáme teda, že genotyp páchateľa máme pevne daný (ako jeden zo sto) a máme aj obvineného, ktorého genotyp sa s prvým zhoduje. Ak predpokladáme, že Hardyho – Weinbergovo ekvilibrium platí, bude pravdepodobnosť náhodnej zhody rovná $P((I,J)|(I,J)) = 2p_ip_j$, kde p_i a p_j sú frekvencie alel I a J na danom locuse. Ak vieme, že páchateľ aj podozrivý pochádzajú z tej istej subpopulácie, v ktorej dokonca poznáme hodnotu koeficientu príbuznosti, vieme sa na problém pozrieť aj inak a na výpočet použiť vzorec (3.5).

Aby sme správne dosadili do vzorca, stačí si uvedomiť, v akej sme situácii: máme k dispozícii jeden genotyp (páchateľa), v ktorom sa nachádza jedna alela I a jedna alela J, a ideme vypočítať pravdepodobnosť, že ďalšie dve vybrané alely (genotyp obvineného) budú I a J. Inými slovami: doteraz sme videli dve alely (vo vzorci (3.5) je to n), z ktorých jedna (vo vzorci n_i) bola I. Teda pravdepobnosť, že ďalšia alela bude tiež I je

$$\frac{1 \cdot F + (1-F)p_i}{1 + (2-1)F}.$$

Podobne pravdepodobnosť, že ďalšia (štvrtá) alela bude J, ak jedna z troch doterajších bola J je

$$\frac{1 \cdot F + (1-F)p_j}{1 + (3-1)F}.$$

[11]Nulová alela je miesto v DNA, kde chýba gén, alebo je na jeho mieste jeho mutácia.Viac sa môže čitateľ dozvedieť napríklad na stánke *en.wikipedia.org/wiki/Null_allele*.

Všimnime si, že na poradí I a J nezáleží a keby sme skôr počítali s alelou J a až potom s I, vymenili by sa v zlomkoch iba menovatele. Výsledná pravdepodobnosť náhodnej zhody, ak berieme do úvahy koeficient príbuznosti, je potom

$$P((I,J)|(I,J)) = \frac{2(F+(1-F)p_i)(F+(1-F)p_j)}{(1+F)(1+2F)}, \tag{3.9}$$

keďže výskyt alel na locusoch je nezávislý (linkage ekvilibrium z 1.3). Teraz už len zostáva porovnať, čo je lepšie na odhad frekvencie genotypu[12]: či databázové frekvencie použiť vo vzorci (3.9), alebo ich dosadiť do klasickej formuly danej Hardyho – Weinbergovým ekvilibriom ($2p_ip_j$). Skutočné frekvencie alel $\hat{p}_i$ a tým pádom aj genotypov ovplyvnené koeficientom príbuznosti totiž nepoznáme.

Pre účel našej úlohy (keďže predpokladáme, že vzorka úplne korešponduje s populáciou) odhadneme skutočné frekvencie genotypov na jednotlivých locusoch[13] ako pomer $k/100$, kde k je počet výskytov daného páru alel na danom locuse v stočlennej populácii. Napríklad ak na prvom locuse bude mať genotyp (I,J) frekvenciu 0.2, znamená to, že 20 ľudí z našich 100 malo na prvom locuse dvojicu alel (I,J). Takto postupne odhadneme frekvencie všetkých kombinácií alel na všetkých štyroch locusoch. Potom frekvencia celého genotypu (pozostávajúceho zo všetkých štyroch locusov) bude súčin frekvencií dvojíc alel na jednotlivých locusoch. Takto získané frekvencie nazveme *pozorovanie*, frekvencie vypočítané podľa (3.9) budú *korekcia* a tie vypočítané klasicky ako $2p_ip_j$ pomenujeme *Hardy – Weinberg*.

Porovnanie urobíme presne ako v článku [5]. Jednotlivé (štvorlocusové) genotypy znázorníme graficky ako body, ktorých x-ová súradnica bude

$$log_{10}\left(\frac{\text{pozorovanie}}{\text{Hardy – Weinberg}}\right)$$

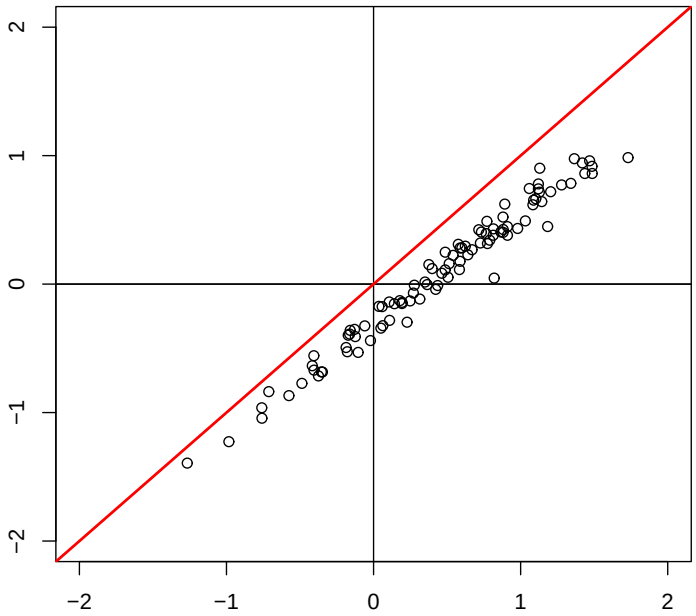
a y-ová bude

$$log_{10}\left(\frac{\text{pozorovanie}}{\text{korekcia}}\right).$$

Samotný obrázok vyzerá takto (čiernou sú dokreslené osi a červená priamka je $x = y$):

[12]Lebo počítať pravdepodobnosť náhodnej zhody vlastne znamená počítať frekvenciu genotypu v populácii.

[13]Genotypom na locuse myslíme dvojicu alel, ktorá sa na ňom nachádza.



Obrázok 3.20. Porovnanie dvoch spôsobov výpočtu value of evidence pre 100 vybraných genotypov

Interpretácia obrázka nie je veľmi zložitá. Začneme tým, že desiatkový logaritmus je tam len „pre krásu“, lebo jeho jediná funkcia je jasnejšie rozlišovať čísla menšie a väčšie ako jedna. Ak je číslo menšie ako 1, jeho logaritmus bude záporný, ak je väčšie, kladný. Zaujímavejšie sú ale samotné čísla, ktoré do logaritmu vstupujú. Tieto sú pomermi pravdepodobností náhodnej zhody. Údaj *pozorovanie* považujeme za reálny stav a teda pravdepodobnosť náhodnej zhody vypočítaná pomocou týchto údajov je tá správna. Zhrňme si teraz význam jednotlivých súradníc.

1. Ak x-ová súradnica bodu je záporná, pomer pozorovanie/Hardy – Weinberg je menší ako jedna, a teda pravdepodobnosť náhodnej zhody vypočítaná pomocou Hardyho – Weinbergovho ekvilibria je väčšia ako skutočná. Value of evidence je teda menšia ako v skutočnosti.
2. Ak x-ová súradnica bodu je kladná, pomer pozorovanie/Hardy – Weinberg je väčší ako jedna, a teda pravdepodobnosť náhodnej zhody vypočítaná pomocou Hardyho – Weinbergovho ekvilibria je menšia ako skutočná. Value of evidence je teda väčšia ako v skutočnosti.
3. Ak y-ová súradnica bodu je záporná, pomer pozorovanie/korekcia je menší ako jedna, a teda pravdepodobnosť náhodnej zhody vypočítaná pomocou (3.9) je väčšia ako skutočná. Value of evidence je teda menšia ako v skutočnosti.
4. Ak y-ová súradnica bodu je kladná, pomer pozorovanie/korekcia je väčší ako jedna,

a teda pravdepodobnosť náhodnej zhody vypočítaná pomocou (3.9) je menšia ako skutočná. Value of evidence je teda väčšia ako v skutočnosti.

Ideálne by sme chceli, aby náš výpočet value of evidence a teda aj pravdepodobnosti náhodnej zhody bol presný. Nie väčší ani menší ako skutočná hodnota. Chceli by sme teda mať pomer rovný jednej, čo znamená nulový logaritmus. Ak sa pozrieme na obrázok, vidíme, že body sa nenachádzajú na priamke $x = 0$ alebo $y = 0$ (až na pár náhodných genotypov). Takže ani jeden zo spôsobov výpočtu pravdepodobnosti náhodnej zhody, ktoré máme na výber, nedáva spoľahlivo tú správnu hodnotu.

Z obrázka ale vidno pár veľmi dôležitých vecí. Podľa polohy bodov vzhľadom na priamku vidíme, že pre každý vygenerovaný genotyp platí

$$log_{10}\left(\frac{\text{pozorovanie}}{\text{Hardy} - \text{Weinberg}}\right) > log_{10}\left(\frac{\text{pozorovanie}}{\text{korekcia}}\right),$$

čiže Hardy – Weinberg $<$ korekcia. Inými slovami, ak použijeme na výpočet pravdepodobnosti náhodnej zhody databázové frekvencie, dostaneme vždy menšiu pravdepodobnosť náhodnej zhody, ako keď uvažujeme koeficient príbuznosti. A menšia pravdepodobnosť náhodnej zhody znamená automaticky väčšiu value of evidence. Toto korešponduje s logickou úvahou. Ak uvažujeme koeficient príbuznosti, počítame so subpopuláciou, kde majú jedinci navzájom podobné genotypy. Pravdepodobnosť náhodnej zhody v subpopulácii je určite väčšia ako pravdepodobnosť náhodnej zhody vo väčšej skupine ľudí (kam patria aj jedinci mimo spomínanej subpopulácie).

Ďalej, v 68 prípadoch na obrázku je x-ová súradnica kladná, čo znamená, že value of evidence vypočítaná pomocou databázových frekvencií je väčšia ako v skutočnosti a že súd tak poškodzuje obvinenému. Naopak, pomáha mu v zvyšných 32 prípadoch, keď je x-ová súradnica záporná. Pre úplnosť, v 53 prípadoch je y-ová súradnica záporná a v 47 kladná. Z tohto ale vyplýva, že ak používame koeficient príbuznosti, prípadov, keď obžalovanému pomôžeme alebo poškodíme je odhadom približne rovnaký počet. Naproti tomu, ak F zanedbáme, vo väčšine prípadov to znamená zaujatosť proti obžalovanému.

Záver, ktorý môžeme z tejto časti vyvodiť je jasný: ak chceme mať spravodlivejšie súdne procesy, je lepšie pri vyhodnocovaní DNA-analýzy nezanedbávať koeficient príbuznosti.

Záver

V prvej a druhej kapitole práce sme zhrnuli základné poznatky o analýze DNA a určovaní value of evidence, resp. matematických postupoch, ktoré sa počas nej používajú. Odvodili sme vzorce, ktorými možno value of evidence vypočítať v špeciálnych prípadoch a naznačili sme mnohé zaujímavé triky a úvahy, ktoré možno vo všeobecnosti použiť pri matematickom popisovaní dedičnosti genetických znakov.

Následne sme zaviedli koeficient príbuznosti a odhadovali sme jeho hodnoty na území Slovenska pomocou simulačného Metropolisovho – Hastingsovho algoritmu, pričom sme porovnávali tieto hodnoty medzi slovenskými krajmi a jednotlivými locusmi. Tu sme použili slovenskú DNA-databázu, ktorú nám poskytlo Oddelenie biológie a genetickej analýzy Kriminalistického a expertízneho ústavu Policajného zboru SR.

Pomocou metódy bootstrap na generovanie vlastných vzoriek sme získali pomerne dobrú predstavu o hodnotách koeficientu príbuznosti v slovenských krajoch. Nakoľko sme boli prví, kto kedy na Slovensku takúto analýzu robil, nemáme naše výsledky s čím porovnať. Na rozdiel od v Západnej Európe a USA, kde sa bežne používa hodnota 3% sa na základe našej analýzy pre Slovensko prikláňame ku koeficientu príbuznosti 2.5% (hodnoty odhadu sú v tabuľkách 3.11 a 3.12).

Prvý použitý model, ktorý odhadoval každý z 24 koeficientov príbuznosti pre kombinácie locus – kraj každý samostatne z inej vzorky dát, sme ďalej porovnali s modelom z článku [5], kde sa zmenšil počet odhadovaných parametrov na 11, no zvýšila sa miera využitia informácie z dát. Tento model priniesol veľmi pekne interpretovateľné výsledky a odhady mali menšiu disperziu, čo nasvedčovalo, že sú presnejšie. Nakoľko ale prvý použitý model bol saturovaný (takže platí vždy) a výsledné odhady sa dosť líšili od toho druhého, viedlo to k pochybnostiam o platnosti druhého modelu v slovenských podmienkach.

V závere práce sme ukázali na praktickom príklade, k akým chybám môže dôjsť (a momentálne, žiaľ, aj dochádza) pri zanedbaní koeficientu príbuznosti.

Myslíme si, že oblasť DNA-analýzy a vecí s ňou spojených je veľmi vhodná a zaujímavá pre ďalší výskum, nakoľko z matematickej stránky sa jej na území Slovenska doteraz

takmer nikto nevenoval. Z otvorených otázok spomeňme napríklad problém výberu locusov, ktoré do DNA-analýzy zahrnúť, aby bola pravdepodobnosť náhodnej zhody čo najmenšia.

Zaujímavý je aj fakt, že počas spojenia genotypu otca a matky (vzniku genotypu dieťaťa) môže nastať mutácia alel, napríklad alela 9 sa môže zmeniť na 10 (vytvorí sa jedno opakovanie nukleotidov naviac) alebo naopak. V tejto práci sme s možnosťou mutácie vôbec nepočítali.

Literatúra

[1] Scientific Working Group on DNA Analysis Methods (SWGDAM). *SWGDAM Interpretation Guidelines for Autosomal STR Typing*. SWGDAM, USA, 2010.

[2] F. Taroni, C. Aitken, P. Garbolino, and A. Biedermann. *Bayesian Networks and Probabilistic Inference in Forensic Science*. John Wiley & Sons, Chichester, West Sussex, 2006.

[3] David Lucy. *Introduction to Statistics for Forensic Scientists*. John Wiley & Sons, Chichester, West Sussex, 2005.

[4] 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. *Genetica*, 96:3 –12, 1995.

[5] D. J. Balding and R. A. Nichols. Significant genetic correlations among caucasians at forensic dna loci. *Heredity*, 78:583 – 589, 1997.

[6] Sheldon M. Ross. *Simulation*. Elsevier Academic Press, Burlington, USA, 2006.

[7] L. A. Foreman, J. A. Lambert, and I. W. Evett. Regional genetic variation in caucasians. *Forensic Science International*, 95:27 – 37, 1998.

[8] L. A. Foreman, A. F. M. Smith, and I. W. Evett. A bayesian approach to validating str multiplex databases for use in forensic casework. *International Journal of Legal Medicine*, 110:244 – 250, 1997.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.