**УЧЕНИЧЕСКИ ИНСТИТУТ ПО МАТЕМАТИКА И ИНФОРМАТИКА**

**ЧЕТИРИНАДСЕТА УЧЕНИЧЕСКА КОНФЕРЕНЦИЯ УК’14**

**ТЕМА НА ПРОЕКТА:**

# Криминология и вероятности


**Автор:**

Гергана Марин Георгиева,
ПМГ "Акад. Никола Обрешков", гр. Бургас, 10 клас

**Научен ръководител (консултант):**

Владимир Георгиев, University of Pisa
Алесандра Ла Спина, University of Pisa

## Резюме

В настоящия проект показвам приложението на математиката в криминалистиката. Разработката постепенно ни въвежда в по-специфични области на теорията на вероятностите. Обясняват се някои термини от генетиката, които ще са ни нужни в по-нататъчните изследвания описани в проекта. Въвеждат се понятия като "ДНК профил" и "Cold Hit" търсене. Акцентира се на избрани криминални случаи, които са пряко свързани с целта на разработката.

## Abstract

In this project I present the application of mathematics in criminology. The study gradually introduce us in more specific parts of the Probability theory. There are some terms from genetics that are being explained, which are needed in further researches. Terms like "DNA profile" and "Cold hit" searching, are being introduced. It is emphasized on chosen criminal cases which are closely connected with the designation of this project.

## Съдържание

Разработката е продължение на проект представян на Пролетната Ученическа Секция на СМБ, като е провокирана от една от задачите, които бяха разгледани в него, която ще бъде припомнена малко по-късно в настоящия проект (1.5).

## Вероятност

В ежедневието си ние многократно и несъзнателно базираме различни свои решения на интуитивното си чувство, кое е вероятно и кое не е вероятно да се случи. Още от началните класове на различни математически състезания ни предизвикват да изчисляваме вероятности от типа на "Каква е вероятността от 10 разноцветни, еднакви по размер топчета, да изтеглим синьо?". Не веднъж сме отговаряли на такива въпроси и в ежедневието си, като обикновено предположението ни е провокирано от личната ни преценка, независимо дали тя е вярна или погрешна. Поради тази причина в математиката се е оформил отделен дял, наречен "Теория на вероятностите", изследващ и обосноваващ появата на всяко събитие.

Дефиниция 1.1.

Вероятност наричаме зададена функция P, която определя за всяко събитие A, принадлежащо на множеството S, реална стойност P(A), наречена вероятност на събитието A.

Дефиниция 1.2.

Елементарно събитие A наричаме подмножество на множеството S, състоящо се от точно един елемент.

Вероятност на дадено събитие изчисляваме като разделим броя на събитията които искаме да получим на общия брой елементарни събития които могат да се случат.

P (събития X) = X ÷ общия брой

Тъй като и събитието X и общия брой имат неотрицателни стойности, следва че и P(X) дава неотрицателна стойност.

По аналогия можем да кажем, че щом събитята X са част от общия брой изходи на които делим, то следва че са по-малко или равно на този брой. Тоест $P(X) \in [0;1]$

За да пресметнем сумата на събитие A и събитие Б, събираме вероятностите на двете събития и изваждаме общите им решения. Тоест P(A + Б)= P(A) + P(Б) – P(A.Б), като A.Б е неотрицателна величина. От което естествено следва, че $P(A + Б) \leq P(A) + P(Б)$. Ако тези две събития са несъвместими следва че те нямат общи решения и следователно формулата има вида P(A + Б)= P(A) + P(Б)

Съществува и условна вероятност. Тя се изразява в зависимостта между едно и друго събитие. Ако А зависи от Б, то значи че А се случва само когато Б се или не се изпълнява. Вероятността за настъпване на А ако Б се изпълнява се изобразява като $P(А_Б)$. За условната вероятност е вярна теоремата:

$P(АБ)= P(А).P(Б)_Б = P(А).P(Б)_А$

Събитията А и Б не принадлежат към условната вероятност, когато е вярно, че $P(А)_Б= P(А)$ и $P(Б)_А= P(Б)$

В средата на 18 век английският математик Томас Бейс съставя формула за пресмятане на вероятност за настъпването на дадено събитие, ако вече имаме част от информацията за него. Формулира я по следния начин:

Нека вероятността за настъпване на събитието А бележим с $P_А$, следователно вероятността за настъпването на събитието Б ще бележим с $P_Б$. Условната вероятност за настъпването на събитието Б, когато А е настъпило, ще бъде $P(Б)_А$, а условната вероятност за настъпването на събитието А, когато Б е настъпило, ще бъде $P(А)_Б$

Теорема 1.3. (Теорема на Бейс)

$$P(A)_Б = \frac{P(Б)_A .P(A)}{P(Б)}$$

Пример 1.4. (Парадокс на Бертрандските кутии) Има три кутии. В първата се съдържат две златни монети, във втората- една златна и една сребърна монета, а в третата- две сребърни монети. На случаен принцип избираме една от кутиите и теглим от нея монета. Ако се случи тя да е златна, то каква е вероятността монетата, останала в кутията също да е златна?

Решение: Търсим вероятността да изтеглим златна монета, след като вече сме изтеглили една такава. Това ще бъде условната вероятност $P(А)_Б$, където събитието А е вече изтеглената златна монета, а събитието Б- монетата оставаща в кутията. $P(Б)_А$ в нашата задача е равна на 1, защото вече сме изтеглили тази златна монета. Имаме общо три златни монети, като вероятностите да сме изтеглили всяка от тях са равни, следователно $P_А = 1/3$. Оставащата монета е или сребърна или златна, следователно $P_Б = 1/2$.

По този начин по теорема на Бейс изчисляваме, че:

$$P(A)_Б = \frac{P(Б)_A . P_A}{P_Б} = \frac{1\ .\ 1/3}{1/2} = \frac{2}{3}$$

Задача 1.5.

Тест за алкохол е 99% точен. Ако шофьорът има алкохол в системата си, в 99% от случаите тестът ще покаже положителен резултат. Ако шофьорът не е употребявал алкохол, в 99% от случаите тестът ще покаже отрицателен резултат. Нека 0,5% от шофьорите на които е направен този тест са пили. Ако тестът покаже положителен резултат, каква е вероятността човекът да има алкохол в системата си?

Решение:

Р(А) е вероятността шофьорът да е пил, която в случая е равна на 0,005

Р(Б) е вероятността шофьорът да е чист, тоест 1 - 0,005 = 0,995

Р(П) е вероятността тестът да даде положителен резултат, независимо дали човекът е пил или не е.

$Р(П)_А$ е вероятността тестът да даде положителен резултат, ако човекът е пил, т.е. 0,99

$Р(П)_Б$ е вероятността тестът да даде положителен резултат, а човекът да не е пил, т.е. 0,01

Търсената вероятност е $Р(А)_П$ , тоест вероятността шофьорът да е пил, ако тестът е положителен

$$P(A)_П = \frac{P(П)_A \,.\, P(A)}{P(П)_Б \,.\, P(Б) + \; P(П)_A \,.\, P(A)} = \frac{0{,}99\,.0{,}005}{0{,}01\,.0{,}995 + 0{,}99\,.\;0{,}005} = 0{,}3322147651$$

Тоест тестът дава едва 33% достоверност. Това се дължи на дръстичната разлика в броя на шофьорите, които не са пили и тези, които са. За да са достоверни резултатите от един такъв тест, той трябва да дава много повече от 99% сигурност. Това е само един от примерите, за това, че числата лесно могат да ни подведат, ако не сме убедени как да боравим правилно с тях.

Точно това Кийт Девлин- британец, по-настоящем доктор на математическите науки в университета Станфорд, неколкократно споменава в една от разработките си ("Scientific Heat about Cold Hits")- че поради невъзможността на човек да си представи по-големи числа, често подценява важни резултати, които се получават чрез тях.

За защитаване на тезата си в "Scientific Heat about Cold Hits" Девлин ни въвежда в криминалния случай на Денис Долингър, но преди да се запознаем с него ще разгледаме някои понятия от криминологията и генетиката.

## ДНК профилиране

ДНК молекулата се състои от две дълги нишки, усукани една около друга в познатата двуспирална структура, съединени във въжестълбовидна форма от химически изграждащи блокове, наречени бази (Двете нишки представляват "въжета" на "стълба", а връзките между тях "стълбите"). Има четири различни бази- аденин (А), тимин (Т), гуанин (G) и цитозин (С).

Човешкият геном е направен от поредица от около три милиарда от тези базови двойки. Изхождайки от ДНК молекулата, последователността от букви обозначаващи реда на подреждане на базите (част може да бъде...AATGGGCATTTTGAC...) осигурява "разчитане" на генетичния код на човек (или друг жив организъм). Точно това "разчитане" осигурява профилирането на ДНК.

ДНК-то е организирано в големи структурни тела, наречени хромозоми. Хората имат 23 двойки от хромозоми, които заедно формират човешкия геном. Едната хромозома във всяка двойка е наследена от майката, а другата- от бащата. Това значи, че индивидът ще има два пълни комплекта от генетичен материал.

Дефиниция 2.1.

Определено местоположение в една хромозома, например на даден ген или на даден биомаркер, наричаме "локус".

Гените могат да имат различни състояния.

Дефиниция 2.2.

Формите на състояния на гена се наричат "алели".

Двойка хромозоми имат един и същ локус по цялата си дължина, но могат да имат различни алели в някои от локусите. Алелите се характеризират със своите малко по-различни базови последователности. Някои от гените изследвани в тестове на смесено ДНК имат повече от 35 различни алели.

Повечето хора имат много сходни генни последователности, но някои части на ДНК последователност могат да се разбичават дръстично от човек до човек. Сравнявайки промяната в тези области, учените си позволяват да отговорят на въпроса дали две различни ДНК проби произлизат от един и същ човек.

Техниката за профилиране, използвана от ФБР и други правоприлагащи органи, зависи от факта, дали разликата се появява от различието в дължината, измерена чрез номера на базите, или от това колко пъти дадена последователност се е повтарила, между предварително определени позиции. Тази процедура дава две стойности за всяка проба, за всеки локус- една за страната на бащата и една за страната на майката. Дължината на ДНК фрагменти може да се определи с точност. При сравняване на две проби на даден локус, ако двойката стойности от една проба е същата като двойката стойности от друга, то се предполага, че профилите съвпадат в този локус. В противен случай, се смята, че не съвпадат в дадения локус. Ако двата профила съвпадат във всеки от изследваните локуси, се смята че

профилите съвпадат. Ако профилите не съвпадат в един или повече локуси, то профилите не съвпадат и е почти сигурно, че пробите не са взети от един и същ човек.

Съвпадение не означава със сигурност, че двете проби трябва със сигурност да са от един и същ източник. Всичко, което може да се заключи, е че доколкото теста може да определи, двата профила са идентични, но е възможно повече от един човек да има един и същ профил сред няколко локуса. За всеки даден локус, процентът на хората имащи ДНК фрагменти от дадена дължина, по отношение на базови двойки, е малък, но не нула. ДНК тестовете черпят сила от връзката между съвпаденията във всеки от няколкото локуса. Много рядко срещано е две проби, взети от индивиди без каквато и да е връзка, да покаже такова съответствие в много локуси.

## Системата CODIS

През 1994 г. , признавайки нарастващото значение на съдебномедицинския анализ на ДНК, Конгресът приема Закона за ДНК идентификацията, с която разрешава създаването на ДНК база данни на национално осъдените престъпници и създаването на ДНК Консултативен съвет (DAB- DNA Advisory Board) да съветва ФБР по различни въпроси. Членовете на DAB са назначени от директора на ФБР от списък с експерти, определени от Националната академия на науките и професионални криминалистични общества.

CODIS (COmbined DNA Index System), система на ФБР за ДНК профилиране е била започната като пилотна програма през 1990 година. Системата съчетава компютърни и ДНК-технологии, за да предостави мощен инструмент за борба с престъпността. Базата данни CODIS се състои от четири категории на ДНК записи :

- Осъдени престъпници - Записи на ДНК идентификация на лица, осъдени за престъпления;
- Криминалистика - Анализи на ДНК проби, събрани от местопрестъпления ;
- Неидентифициран човешки останки - Анализи на ДНК проби , събрани от неидентифицирани човешки останки;
- Роднини на изчезнали лица - Анализи на ДНК проби доброволно предоставени от роднини на безследно изчезнали лица.

Базата данни CODIS на осъдените извършители на престъпления през 2005 година съдържа над 2,7 милиона записа.

ДНК профилите, съхранявани в CODIS се основават на тринадесет специфични локуси, избрани, защото те показват значителни различия сред населението.

CODIS използва компютърен софтуер за автоматично търсене сред тези бази данни за съвпадение на ДНК профили.

CODIS също поддържа Население- база данни с анонимни ДНК профили, използвани за определяне на статистическата значимост на дадено съвпадение.

CODIS не е изчерпателна база данни за престъпленията, а по-скоро система от показатели; базата данни съдържа само информация, необходима за определянето на съвпадения. Профилите, съхранявани в CODIS, съдържат идентификатор образец, подпомагащ лабораторния индикатор, инициалите (или името) на ДНК притежателя, свързани с анализа и действителните характеристики на ДНК. CODIS не съхранява информация за криминалните истории, информация по различните случаи, номера на социални осигуровки или дати на раждане.

Когато две случайно избрани ДНК проби съвпадат напълно в голям брой области, както в тринайсетте използвани в системата на ФБР, вероятността те да са взети от двама души без каквато и да е връзка помежду си практически е нула. Този факт прави идентифицирането на ДНК изключително надеждно (когато е изпълнено правилно). Степента на надеждност обикновено е изчислена чрез теория на вероятностите, за да се провери вероятността за намирането на определен профил сред случайна селекция на населението.

Например, да разгледаме профил базиран на само три изследвани области. Вероятността нечие ДНК да съвпадне с някоя случайна ДНК проба в която и да е област е около 1/10. Следователно вероятността нечие ДНК да съвпадне в три области със случайна проба бе била около 1/1 000:

$$1/10 \times 1/10 \times 1/10 = 1/1\ 000$$

Прилагайки същото вероятностно изчисление за всички 13 области изследвани в системата на ФБР- CODIS би означавало, че вероятността за съвпадение с дадена случайна ДНК проба е около едно към десет трилиона:

$$(1/10)^{13} = 1/10\ 000\ 000\ 000\ 000$$

Това число е известно като вероятност за случайно съвпадение (RMP- Random Match Probability). Тъй като се изчислява като се използва принципа за умножаване на вероятностите, това предполага, че моделите са намерени на две различни и независими местоположения. В ранните години на използване на метода за ДНК профилиране, това е било причина за значителен брой дебати, но както изглежда, тези тревоги са изчезнали. На практика действителните вероятности са различни, в зависимост от няколко фактора, но цифрите, изчислени по-горе са взети като цяло и са доста надежден индикатор за вероятността от случайно съвпадение. Това означава, че RMP обикновено се приема като добър показател при определен ДНК профил от населението, въпреки че това трябва да се проверява внимателно. (Например , едноядрени близнаци споделят почти идентични ДНК профили.)

Все още има някои съмнения по отношение на използването на RMP, като надежден индикатор на случайно съвпадение, тъй като е базиран единствено на досегашните ни разбирания за генетиката. Изчисляването на RMP, в края на краищата, изисква математическа независимост на локусите- изключително взискателно условие, с цел да бъде в състояние да прилага правилото на продукта. Следва да се отбележи, че анализ на базата данни през 2005 година на Аризона на осъдените за виновни престъпници (база данни, която използва 13-локусната система CODIS) разкрива, че сред около 65 000 в списъка има 144 лица, чиито ДНК профили съвпадат в 9 локуси (включително едно съвпадение между хора от различни раси- единия от бялата раса, а другия афро-американец), още няколко , които съвпадат в 10 локуса, двама, които съвпадат в 11, и двама, които си съответстват в 12. Съвпаденията при 11 и 12 локуси са на братя и сестри, следователно не са случайни. Но съвпадения на 9 или 10 локуси сред толкова малка база данни от 65 000 в списъка, предизвиква значителни съмнения върху изчисления като "1 на 10 трилиона" за съвпадение, което проверява само с 3 или 4 допълнителни локуса.

## Случая "Дженкинс"

През юни 1999г., полицията на Вашингтон намира трупа на Долингър в дома му в Капитъл Хил. Той е бил промушкан неколкократно (поне 25 пъти) с отверка. Някои ценни вещи, като портфейл, пръстен и златна верижка, са му били откраднати. Няколко часа след убийството, в Александрия мъж на име Стивън Уотсън използва кредитната карта на жертвата. Той става основен заподозрян в случая, като всички доказателства сочат към него, и е осъден. През това време ФБР анализират пробата от местопрестъплението и кръвта на Уотсън, като те не дават съвпадение и той бива освободен.

Дефиниция 2.3.

"Cold hit"("студено попадение")- намиране на съвпадение между ДНК доказателството и човек, регистриран в базата данни (CODIS).

Тогава ФБР провежда "Cold hit" търсене сред досегашни престъпници, чрез системата CODIS. В системата Не се открива съвпадение и пробата е пратена в щата Вирджиния. Там компютър сравнява ДНК на 101, 905 престъпници по критерий 8 локуса. Този път е намерено съвпадение- на мъж на име Робърт Гарет, който по-късно става ясно е псевдоним на Реймънд Антъни Дженкинс. Взима му се проба, която вече се сравнява по 13 локуса с тази от местопрестъплението. Кръвта му дава съвпадение във всичките 13 локуса. На базата на това в началото на 2000 година ФБР арестуват Дженкинс. Всички страни по случая се съгласяват, че лабораторните изследвания са надеждни, и че ДНК-то на Дженкинс съвпада в 13 локуса с това, изследвано след престъплението. Това, върху което е спора, е какво показва съвпадението. И това се оказва, че е въпрос с математически характер.

През 1992г. Националния съвет за научни изследвания в САЩ съобщава, че чрез "Cold hit" търсене има много по-голяма вероятност да се открие съвпадение, отколкото когато се направи ДНК профил на вече заподозрян. Поради тази причина те изискват в съда да се представи вероятност друг човек да има същия ДНК профил като този открит от системата, базирана само на допълнителните съвпадения в локуси, намерени при повторното профилиране. Това в случая на Дженкинс са 5 локуса, т.е. вероятност:

$$1/\ 10^5 = 1/\ 100\ 000$$

Поради възражения от страна както на съда, така и на ФБР, се налага поправка в правилото. През 1993г. излиза второто съобщение на съвета, гласящо: "Когато е открит заподозрян чрез "Cold hit" търсене, вероятността RMP трябва да се умножи по N, където N е броят на хората в базата данни." Като често това се нарича DMP (Database Match Probability), което е необичайно, след като това не е вероятност. В случая на Дженкинс вероятността би изглеждала, след като в базата данни на Вирджиния е имало приблизително 100 000 души:

$$100\ 000 \,.\, \frac{1}{100\ 000\ 000} = \frac{1}{1\ 000}$$

По този начин на пресмятане лесно разбираме защо при проучването в Аризона имаме такъв голям брой души с идентични локуси.

$$65\ 000 \,.\, \frac{1}{100\ 000\ 000} = \frac{65}{100\ 000}$$

Не всички били съгласни с новия начин на представяне на ДНК идентифицирането в съда. Доктор Питър Донали, професор по статически науки в университета в Оксфорд, е на противоположно мнение със съвета. През 2004 година в писмена декларация до съвета, озаглавена "ДНК доказателство след попадение на базата данни". Донъли се аргументира като използва вероятност на Беязин, за да изчисли вероятността за случайно съвпадение, като описва в съвместен проект с Дейвид Балдинг методите си, под името "Изчисляване на доказателство ДНК профил, когато заподозрян е идентифициран чрез търсене в базата данни":

$P(S)_E$: условната вероятност обвиняемия да е виновен от дадените доказателства

$P(E)_S$: условната вероятност да е намерено доказателство приемайки, че обвиняемия е виновен

P(E) : вероятността да се намери доказателство сред населението

P(S) : вероятността обвиняемия да е виновен

$$P(S)_E = \frac{P(E)_S . P(S)}{P(E)}$$

*"Класическите статистици се опитват да избягват лична преценка, като вместо това се стремят да определят какво заключение може да бъде направено на базата на повторяемост на наблюденията. Беязинов подход- актуализираме вероятностите зададени за определено твърдение като имаме впредвид получаването на доказателство впоследствие- това е неприемливо за класическите статистици, защото зависи от субективна задача за вероятности при липса на обективно измерима информация..."*

Те разчитат на изчисление на вероятностното съотношение. Тяхната форма на статистически анализ, позната като "Беязин", изисква правенето на статистически допускания като по-важните разлики между индивидите. Те представят силен статистически аргумент, който все още има широко разпространение в съдилищата в Съединените щати.

Методът на Балдинг и Донъли се различава по три начина от това на Националния съвет за научни изследвания:

1. не се провеждат никакви тестове на допълнителните локуси
2. генетическите маркери използвани в търсенето на оригиналната база данни са включени в статистическите изчисления
3. размерът на базата данни (N), в която се търси, се взема впредвид.

Техния метод още се различава по един крайно основен начин: за тези групи от експерти, ефектът на големината на базата данни за значението на съвпадението е напълно противоположно- голямата база данни генерира най-изобличаващите статистики за обвиняемия, докато според Националния съвет- колкото по-голяма е базата данни, толкова по-незначителни стават статистиките за обвиняемия.

Една от причините Националния съвет да не е съгласен с изчисленията на Донъли и Балдинг, е че не намират за редно в началото на предположенията им всеки човек по света, да бъде считан за потенциален извършител на престъплението, което прави вероятността за всеки от нас малка (около 1 към 7 милиарда), но не 0.

Този спор още не е решен, като във Великобритани и някои страни от Европа е признат метода на Балдинг и Донъли, докато САЩ продължава да отрича постигнатия от тях резултат, като считат, че техниката Беязин е твърде специфична, за да бъде разбрана от нестатистици който е малко по-малък от вероятността за случайно съвпадение (RMP).

## Случая "Дженкинс"

През 2000 година, случаят отново стига до съда, като се допуска споменаването на RMP само ако бъде спомената и отношението изразено от DMP. Двете страни по случая водят специалисти, които да представят информацията от двете гледни точки, но съдът смята за неправилно да претегля доказателства, за които е нужно да си експерт, за да оцениш правилно, след като дори световни лидери в тези научни области не могат да достигнат до правилното решение. Съдът се опитва да разгледа и двете доказателства, но

когато се смесят различни начини за пресмятане на вероятност, нещата само се объркват още повече.

## Бъдещо развитие на проекта и Благодарности

За вбъдеще възнамерявам да се запозная по-обстойно със смесени кръвни проби и доказателствата, които те биха могли да дадат при изследването им. Искам да намеря отворени проблеми свързани с темата, които да се опитам да разработя и да се запозная с българската система за идентифициране, тъй като досегашните ми познания са ограничени от извличането на информация от източници на хиляди километри от мен.

Изказвам специални благодарности към научния си ръководител Владимир Георгиев, че ми даде възможността да се докосна до такава тема, като ме свърза с колежката си Алесандра Ла Спина и беше готов да ми помогне при всякакви затруднения (особено с италиянския). Благодаря на УчИМИ, че ни давате възможността да се самоусъвършенстваме и направлявате нашето развитие.

## Библиография

I. DNA in crime cases: some mathematical models for DNA mixture analysis and the "cold hit" problem, Alessandra La Spina, University of Pisa

II. Scientific Heat about Cold Hits (Draft), Keith Devlin, Stanford University, 2007

III. Evaluating DNA Profile Evidence When the Suspect Is Identified Through a Database Search, Peter Donnelly and David Balding, University of Oxford, 2004

IV. ДНК и криминалистика, Пламен Топалски, студент в Юридическия факултет на Софийски университет

V. https://onlinecourses.science.psu.edu/stat414/node/27

VI. http://pocketexpert.net/files/Joseph_20Berger_20Cold_20Hit_20Frye_20Motion_1_.pdf