Sandsynlighedsregning i biologi

Om begrebet sandsynlighed

Hvis vi kaster en almindelig, symmetrisk terning, er det klart for de fleste af os, hvad vi mener, når vi siger, at sandsynligheden for at få en femmer er 1/6.
Hvis vi får at vide, at sandsynligheden for, at en mand er rød-grøn farveblind, er 5%, er vi også umiddelbart klar over, hvad det betyder.
Alligevel er de to sandsynligheder meget forskellig af natur. I tilfældet med terningen behøver vi ikke at kaste den for at kunne udtale os om sandsynligheden for at få en femmer; den er nærmest givet på forhånd. En sådan sandsynlighed kaldes en *a priori sandsynlighed*.
I tilfældet med farveblindheden forholder det sig anderledes. For at vi skal have en chance for at kunne udtale os om sandsynligheden her, må vi have nogle statistiske oplysninger om rød-grøn farveblindhed for mænd. En sandsynlighed, der bestemmes på denne måde, kaldes en *frekventiel sandsynlighed*.

> Øvelse 1
> Hvad er sandsynligheden for at få plat ved kast med en mønt?
>
> Øvelse 2
> Vi kaster en tændstikæske på gulvet. Hvad er sandsynligheden for, at den lander på svovlet?
>
> Øvelse 3
> 25 ud af 10.000 kvinder er rød-grøn farveblinde.
> a) Hvad er sandsynligheden for, at en kvinde er rød-grøn farveblind.
> b) Hvad kan årsagen være til forskellen på farveblindhed hos mænd og kvinder?
>
> Øvelse 4
> a) Overvej i hvert tilfælde i øvelse 1 til 3, om der er tale om en a priori sandsynlighed eller en frekventiel sandsynlighed.
> b) Find selv nogle eksempler på henholdsvis a priori sandsynligheder og frekventielle sandsynligheder.

Betinget sandsynlighed

> Øvelse 5
> Anders og Bent er kokke på restaurant Bayes i henholdsvis 6 dage og 1 dag om ugen.
> Anders mad er god i 90% af tilfældene, mens Bents mad er god i 50% af tilfældene.
> En aften serverer Bayes et elendigt måltid.
> a) Synes du umiddelbart, det er rimeligt at konkludere, at det er Bent, der har lavet maden den aften?
> b) I den uge er der blevet serveret 700 måltider på Bayes. Fordel de 700 måltider i følgende skema:
>
> | | Godt | Dårligt |
> |---|---|---|
> | Anders | | |
> | Bent | | |
>
> c) Hvor mange dårlige måltider er der blevet serveret i den uge?
> d) Hvor mange af dem har Bent lavet?
> e) Hvad er sandsynligheden for, at et dårligt måltid er lavet af Bent?

Den sandsynlighed, vi finder i sidste spørgsmål i øvelse 5 kaldes en betinget sandsynlighed. Vi finder nemlig sandsynligheden for, at det er Bent, der har lavet maden, når vi ved, at måltidet er dårligt. Denne sandsynlighed vil vi betegne P(Bent│Dårligt) eller kort P(B│D).
I skemaet i øvelse 5 bør I have fundet frem til, at Anders har lavet 60 og Bent 50 dårlige måltider. Der er altså i alt lavet 110 dårlige måltider den uge, og sandsynligheden for, at et tilfældigt måltid er dårligt, er

således 110/700. Denne sandsynlighed kalder vi P(D). Sandsynligheden for at få et måltid, der både er lavet af Bent og er dårligt, er 50/700, og denne sandsynlighed kalder vi P(B∩D).
Der gælder nu ifølge sidste spørgsmål i øvelse 5, at

$$P(B \mid D) = \frac{50}{110} = \frac{\frac{50}{700}}{\frac{110}{700}} = \frac{P(B \cap D)}{P(D)}.$$

Vi har altså indført en betinget sandsynlighed P(B | D) og har fundet en formel, som kan bruges til beregning af den.

$$P(B \mid D) = \frac{P(B \cap D)}{P(D)}.$$

Skrivemåden $B \cap D$ svarer til, at vi ser på de måltider, der lavet af Bent og er dårlige. På samme måde svarer $A \cap G$ til de måltider, der er lavet af Anders og er gode. Tallene i skemaet fra øvelse 5 kan oversættes til sandsynlighederne

	Godt	Dårligt
Anders	$P(A \cap G)$	$P(A \cap D)$
Bent	$P(B \cap G)$	$P(B \cap D)$

eller

	Godt	Dårligt
Anders	$\frac{540}{700}$	$\frac{60}{700}$
Bent	$\frac{50}{700}$	$\frac{50}{700}$

Øvelse 6
Ligesom P(D) i ovenstående er sandsynligheden for at få et dårligt måltid, er P(A) sandsynligheden for, at et tilfældigt måltid er lavet af Anders.
a) Find P(A).
b) Find også P(B) og P(G) og forklar med ord, hvad disse sandsynligheder betyder.
c) Find P(A | D), P(B | G).og P(A | G) og forklar med ord, hvad disse sandsynligheder betyder.

Definition Ved *den betingede sandsynlighed* P(A | B) forstår vi sandsynligheden for, at A sker under forudsætning af, at B er sket, og P(A | B) er givet ved formlen

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ hvor det er forudsat, at } P(B) > 0.$$

Eksempel 1
På Minikøbing Gymnasium og HF fordelte antallet af elever sig et år således

	Gymnasieelever	Hf-elever	I alt
Drenge	64	49	113
Piger	71	66	137
I alt	135	115	250

Med G forstår vi, at en elev er gymnasieelev, og med D forstår vi, at en elev er en dreng. Vi ser af skemaet, at

P(G) = 135/250, P(D) = 113/250 og P(G∩D) = 64/250.

Den betingede sandsynlighed P(G | D) er så:

$$P(G \mid D) = \frac{P(G \cap D)}{P(D)} = \frac{\frac{64}{250}}{\frac{113}{250}} = \frac{64}{113} \approx 56{,}6\%.$$

Tager vi en tilfældigt valgt dreng på Minikøbing Gymnasium og HF, er der altså knap 57% chance for, at han går i gymnasiet.

> Øvelse 7
> Udregn P(D | G) og beskriv med ord, hvad denne sandsynlighed betyder.
>
> Øvelse 8
> Vis ved hjælp af definitionen ovenfor, at
> $$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

Fælles for de indledende eksempler og øvelser er, at der eksisterer en række alternative muligheder, hvis sandsynlighed man udtaler sig om. Disse alternativer kalder man også for *udfald*, og alle udfaldene udgør tilsammen *udfaldsrummet*. Delmængder af udfaldsrummet kaldes *hændelser*. To hændelser kaldes *disjunkte*, hvis de ikke har nogen udfald til fælles.
I eksempel 1 udgøres udfaldsrummet af de 250 elever, der går på Minikøbing Gymnasium og HF, og G er et eksempel på hændelsen, at en tilfældigt valgt elev er en gymnasieelev. Hvis H betegner hændelsen, at en tilfældigt valgt elev er en Hf-elev, er G og H disjunkte. Man siger også, at G og H udelukker hinanden. Man kan desuden se, at G og H tilsammen udgør hele udfaldsrummet.

Sætning *(Bayes formel)*
Hvis et udfaldsrum er delt op i n hændelser H_1, H_2, , H_n, som er parvis disjunkte, og der er givet en hændelse A med P(A) > 0, så vil

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \ldots + P(A|H_n)P(H_n)},$$

for i = 1, 2, … , n.

Bevis:
Vi kan dele udfaldene i A op i de udfald, der ligger i H_1, de udfald, der ligger i H_2, de udfald, der ligger i H_3 osv. Derfor bliver

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + \ldots + P(A \cap H_n).$$

Af definitionen på betinget sandsynlighed kan vi omskrive dette til

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \ldots + P(A|H_n)P(H_n)$$

Fra øvelse 8 ved vi, at

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A)},$$

og indsætter man nu udtrykket for P(A) i nævneren, fås sætningen.

Retsgenetik- case.
Case: Alle børn skal have en fader.

Biologisk baggrundsstof:
- DNA molekylet.
- Cellekernen med 46 kromosomer.
- Meiose (ikke overkrydsning)
- Et-gensnedarvning: cystisk fibrose, Huntingtons Chorea, blodtyper ABO, rhesus og blødersygdom/farveblindhed

Øvelse 9
- Hvor mange forskellige kønsceller kan dannes ved en meiotisk deling uden overkrydsning og uden mutationer?
- Hvad er frekvensen af cystisk fibrose i den danske befolkning? (find det evt. på nettet)

Et ægtepar har fået to normale børn og et barn med cystisk fibrose og venter et fjerde barn.
- Hvilken genotype har faderen og moderen?
- Opskriv et nedarvningsskema.
- Hvad er sandsynligheden for, at det ventede barn får cystisk fibrose?
- Hvordan kan man undersøge om fosteret har cystisk fibrose?

Øvelse 10
En kvinde på 25 år venter et barn. Kvindens far på 51 år har lige fået konstateret Huntingtons Chorea.
- Hvad er sandsynligheden for at kvinden har arvet genet for sygdommen?
- Vil du råde kvinden til en genetisk test?

Øvelse 11
- Analyse af stamtræ med blødersygdom f.eks. dronning Victorias.
- Udregning af sandsynligheden for udvalgte personers genotype.

Forsøg:
- DNA opsamling f.eks. fra løg + DNA elektroforese.
- Et-gennedarvning hos byg.
- Blodtypebestemmelse.
- Farveblindhedsundersøgelse.

Øvelse 12
Faderskabssag:
Moder har blodtype A, barnet har B. Følgende mænd gør krav på faderskabet: en med blodtype A, en med blodtype O, og en med AB.
Hvilken genotype har moderen og barnet?
Hvilke mænd kan straks udelukkes?
Hvor stor er sandsynligheden for barnets genotype, givet at faderen har genotypen $I^B I^B$, $I^B i$ og $I^A I^B$?
Hvor stor er sandsynligheden for, at de ovenstående mænd er far til barnet når nedenstående genotypefordeling er gældende?

Tabel over genotypefordelingen i den danske befolkning:

Genotype	$I^A I^A$	$I^A I^B$	$I^A i$	$I^B I^B$	$I^B i$	ii
Fænotype	A	AB	A	B	B	0
Frekvens %	7.84	4.48	35,84	0,64	10,24	40,96

Kontroller dine resultater på www.hugin.com.

Er det rimeligt at tilkende faderskabet på dette grundlag?

Anvendelse af Bayes formel: (f= farther, c=child, vi skriver A i stedet for I^A og B i stedet for I^B, og c i stedet for cBi på højresiderne af ligningen).

$$P(fAB|cBi) = P(fAB|c) = \frac{P(c|fAB)P(fAB)}{P(c|fBi)P(fBi) + P(c|fBB)P(fBB) + P(c|fAB)P(fAB) + P(c|fii)P(fii) + P(c|fAA)P(fAA) + P(c|fAi)P(fAi)}$$

Teori: PCR
Bruges til at opformere et lille stykke DNA i mange kopier. Stykket udvælges ved at finde to stykker DNA, primers, der ligger på hver side af det ønskede stykke DNA.
Det svarer til, at man i den danske sangskat leder efter stykket mellem "Mariehønen Evigglad gik tur" og "Evigglad til madam Snegl". Det vil entydigt give den sang, og ikke noget som helst andet. Hvis man leder mellem "vinter" og "vår", så får man sikkert mange forskellige stykker ud af det.
Der findes flere forskellige animationer af PCR på nettet, f.eks.
http://www.dnalc.org/shockwave/pcranwhole.html (hvis man har chock wave)
http://www.people.virginia.edu/~rjh9u/pcranim.html (hvor man dog ikke kan se forskel på korte og halvlange DNA stykker).

Øvelse 13

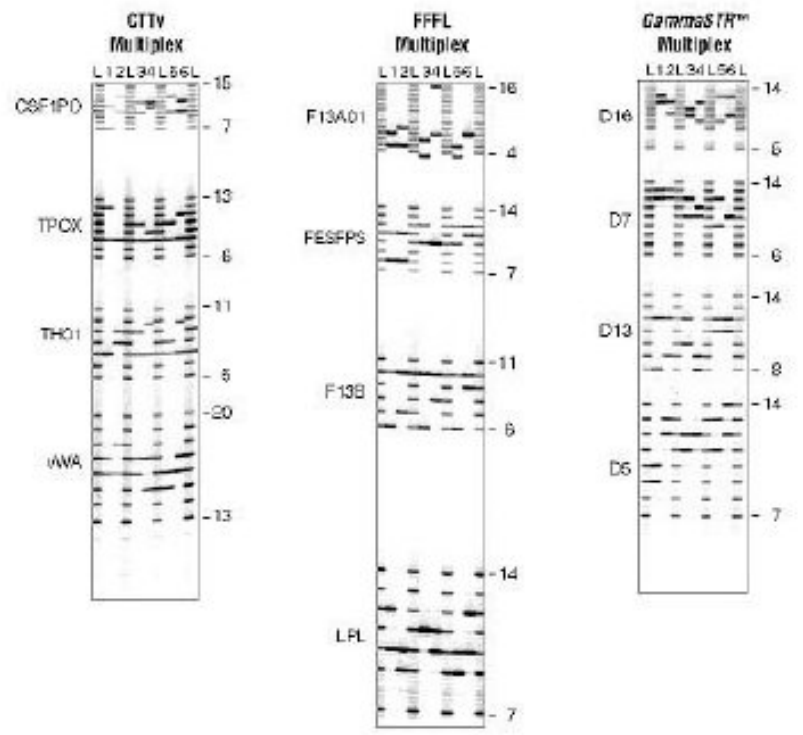
Gør rede for, at efter n kopieringer ud fra 1 langt, dobbeltstrengede DNA, er der $2^n - 2n$ stykker af det søgte (korte) dobbeltstrengede DNA. (Hint: Tæl enkeltstrengene i stedet for dobbeltstrengene, og hold styr på, hvor mange af strengene, der har den oprindelige længde, hvor mange der er halvlange, og hvor mange der er korte, men som sidder på de halvlange.)

Teori: STR
Man regner med, at vi bruger 0,5% af vores DNA. Ind imellem det livsvigtige DNA er der mange stykker DNA, hvori der kan ske ændringer uden at det får konsekvenser for individet. En af disse typer ændringer er, at et stykke DNA repeteres. Et kopieret stykke kunne f.eks. bestå af stykket GAGGCA, der optræder i et helt antal gentagelser på et bestemt sted på et bestemt kromosom. STR står for Short Tandem Repeats, og en STR på 8 betyder, at stykket optræder i 8 kopier. Man har fundet flere steder på det menneskelige kromosom, hvor sådanne gentagelser optræder. Heraf udvælger man de steder, hvor mennesker varierer meget, og hvor der sker meget få mutationer fra generation til generation, idet en høj mutationsrate vil ødelægge anvendelsen i faderskabssager. Der er typisk mellem 6 og 30 kopier af en bestemt sekvens. Hvis 2 mennesker er forskellige i sådan et STR sted, vil forskellen derfor være et multiplum af længden af stykket. Da de stykker DNA, som man udvinder vha. PCR, er 100-350 baser lange, er man faktisk i stand til at adskille 2 stykker DNA, der afviger med bare 4 baser.

Det DNA, der er dannet i PCR, kan eventuelt klippes med restriktionsenzymer. Så skal enzymerne klippe primer-enderne af, men de må ikke klippe i selve STR stedet. Til forskel for andre analyser, er STR nemlig ikke bestemt af, hvor restriktionsenzymerne klipper, men kun af hvor mange gentagelser der er. Man skal så være sikker på, at restriktionsenzymerne klipper ens hos alle mennesker.

En persons genotype for en bestemt STR region angives med et talsæt, f. eks. (8,12), hvor man har arvet et stykke med 8 sekvenser fra den ene forælder og et stykke med 12 sekvenser fra den anden forælder.

Teori: Elektroforese
Da DNA er negativt ladet, vil DNA stykker vandre mod den positive pol i et elektrisk spændingsfelt. I en agarose- eller polyacrylamid-gel, vil små stykker DNA vandre hurtigst, og derfor længst. Med farvning vil man derfor kunne se forskellige bånd af DNA stykker. Man sætter også en prøve, der indeholder mange forskellige DNA stykker med kendt længde i elektroforesen, så har man en lineal. På figur 1 er 12 STR sites valgt ud og inddelt i 3 grupper, således at man laver 4 tests sammen i hver elektroforese. På grund af stykkernes størrelser, vil de adskille sig pænt i 4 veldefinerede grupper i hver af de 3 elektroforeserne.



Figur 1
På hver af de tre elektroforeser ses 4 lodrette 'linealer'. Hver eneste vandret streg er en portion DNA med et bestemt antal gentagelser. Den øverste til venstre har således mellem 7 og 16 gentagelser, den næste er mellem 6 og 13.

I de tre mellemrum mellem linealerne er der i alt 6 personer L1 – L6. Bemærk, at hver person har 1 eller 2 streger i hvert STR site.

Kilde: New approaches to DNA Fingerprinting Analysis af J.W. Schumm, *Notes magazine* Number 58, 1996 p. 12.

Øvelse 14
Prøv at aflæse STR talsættene for personerne L1 – L6 i en af de 12 STR regioner på figuren. Klassen kan fordele regionerne eller personerne mellem sig, så alle personerne får aflæst deres 12 talsæt.

Øvelse 15
Mor er (10,12) i en bestemt STR region, barnet er (12,12), ægtemanden er (10,12). Kan ægtemanden være far til barnet? Kan man med sikkerhed sige, at han er far til barnet?

Øvelse 16
I den danske befolkning er P(10)=0,28425, P(12)=0,25942, P(X)=0,45634 (X er alt andet).
Hvad er sandsynligheden for, at en tilfældig person i befolkningen er (10,12)?

Definition:

$$\text{Faderskabskvotient} = \frac{P(\text{moderen og den givne mand får dette barn})}{P(\text{moderen og en tilfældig mand fra populationen får dette barn})}$$

Eksempel 2
I opgaven ovenfor har barnet fået nr 12 fra både mor og far. Sandsynligheden for at ægtemanden giver 12 til barnet er 0,5, mens sandsynligheden for, at en tilfældig mand giver 12 til barnet er 0,25942. Vi definere så faderskabskvotient som

$\frac{0,5}{0,25942} = 1,9274$. Det betyder, at ægtemanden er 1,92 gange så sandsynlig som far til barnet end en tilfældig mand af dansk afstamning.

Eksempel 3
Mor er (10,12), barnet er (10,12); ægtemanden er (12,12). Hvis ægtemanden er far giver han 12 og moderen giver 10, men hvis en tilfældig mand er far, kan moderen give enten 10 eller 12 og manden giver så 12 eller 10. Vi får altså at:

$$\text{Faderskabskvotienten} = \frac{0{,}5 \cdot 1}{0{,}5 \cdot 0{,}25942 + 0{,}5 \cdot 0{,}28425} = 1{,}9$$

Øvelse 17
Et andet sted på kromosomerne sidder en anden STR region, hvor P(14)=0,508, P(16)=0,373 og P(X)=0,119 =P(alt muligt andet end 14 og 16)
Hvis mor er (14,16), barnet er (x,16) og far er (x,16), Hvad er så faderskabskvotienten?

Eksempel 4
Hvis to forskellige STR regioner giver faderskabskvotienter på hhv. 1,355 og 1,281, er

$$\text{Faderskabsindekset} = 1{,}355 \cdot 1{,}281 = 1{,}735$$

Øvelse 18
Tag resultaterne fra øvelse 17 og eksempel 3. Hvad bliver det samlede faderskabsindex for de 3 sites?

Øvelse 19
1. Er der noget, der taler imod, at Peter Jensen (skemaet næste side) er far til barnet?
2. Beregn faderskabsindekset og overstreg *for* eller *imod* og *større* eller *mindre* i teksten.
3. Kommer manden til at betale for barnet ifølge retsmedicinernes regler?
4. Er 10.000 et rimeligt tal for faderskabsindekset sammenlignet med f.eks. 1000? Find ud af, hvor mange mennesker, der er i din by. Regn med at halvdelen er mænd og halvdelen af disse er i en "attraktiv" alder. Giv ud fra faderskabsindekset et skøn over, hvor mange af disse mænd, der kunne være fader til barnet.
5. Hvad er grunden til, at beregningerne ikke gælder ved sammenligning med en nær slægtning?

Retsgenetisk Afdeling ∴ Retsmedicinsk Institut

Rekvirent: Dommeren i Minikøbing

Sagstype: Faderskabssag *Rekvirents J.nr.*

Kvinde: Birte Jensen
Barn: Line Jensen
Mand: Peter Jensen

Blodprøver fra manden:

Resultater:

Genetisk system	Kvinde	Barn	Mand	
D3s1358	14 18	11 14	11 16	P(11)=0,201 P(14)=0,313 P(18)=0,345 P(X)=0,141
HumVWA31	16 17	15 16	15 17	P(15)=0,391 P(16)=0,041 P(17)=0,243 P(X)=0,325
FGA	21 23	21 22	20 22	P(20)=0,301 P(21)=0,121 P(22)=0,046 P(23)=0,386 P(X)=0,146
D8s1179	12 12	8 12	8 15	P(8)=0,156 P(12)=0,388 P(15)=0,256 P(X)=0,2
D21s11	31 32	30 31	30 30	P(30)=0,219 P(31)=0,332 P(32)=0,289 P(X)=0,16
D18s51	14 14	12 14	12 15	P(12)=0,385 P(14)=0,228 P(15)=0,209 P(X)=0,178
D5s818	8 12	11 12	10 11	P(8)=0,276 P(10)=0,391 P(11)=0,2 P(X)=0,133
D13s317	9 10	9 12	11 12	P(9)=0,433 P(10)=0,012 P(11)=0,17 P(12)=0,185 P(X)=0,2
D7s820	9 10	9 10	9 10	P(9)=0,129 P(10)=0,292 P(X)=0,579

Faderskabsindeks:

Resultaterne taler *for/imod* mandens faderskab til barnet med en vægt *større/mindre* end 10.000 til 1.

Resultaterne taler således *for/imod* at den undersøgte mand er *mere/mindre* end 10.000 gange mere sandsynlig som far til barnet end en tilfældig mand af dansk afstamning.

Beregningerne gælder ikke ved sammenligning med en nær slægtning til manden.

Odds

Odds for en hændelse defineres som sandsynligheden for hændelsen divideret med sandsynligheden for den komplementære hændelse. Faderskabskoefficienten er sandsynligheden for den givne far divideret med sandsynligheden for en hvilken som helst far, altså stort set faderens odds for faderskabet. Selv om faderen ikke specifikt er trukket ud af puljen til beregningen af nævneren, går vi ud fra, at puljen af mulige fædre er så stor, at det er underordnet om han er talt med eller ej.

Redigeret af Helle Aagaard-Hansen, oktober 2004