MVE235 Matematisk Orientering: Matematisk statistik med tillämpningar från AI till forensik

Petter Mostad

Chalmers

December 8, 2019

Innehåll

- Matematisk statistik.
- Bayesiansk statistik.
- Tillämpningsområde: AI och maskinlärning.
- Tillämpningsområde: Forensisk statistik.
  - Medicinsk åldersbedömning
  - DNA-tester för släktsskap
- Vetenskapsteori.

Matematisk statistik

- Statistik är sammanställning och uppsummering av data.
- *Matematisk* statistik är att använda probabilistiska modeller *och* data för att göra probabilistiska *prediktioner*.
- Det finns olika *paradigmer* för hur man gör detta bäst, e.g., frekventistisk (klassisk) eller Bayesiansk statistik.
- Hur man gör prediktioner baserad på data är nåt många i dag förbinder med *maskininlärning* (ML) och *Artificiell Intelligens* (AI). Grunderna i dessa teknologier kan sägas vara probabilistiska modeller, ofta Bayesiansk statistik.

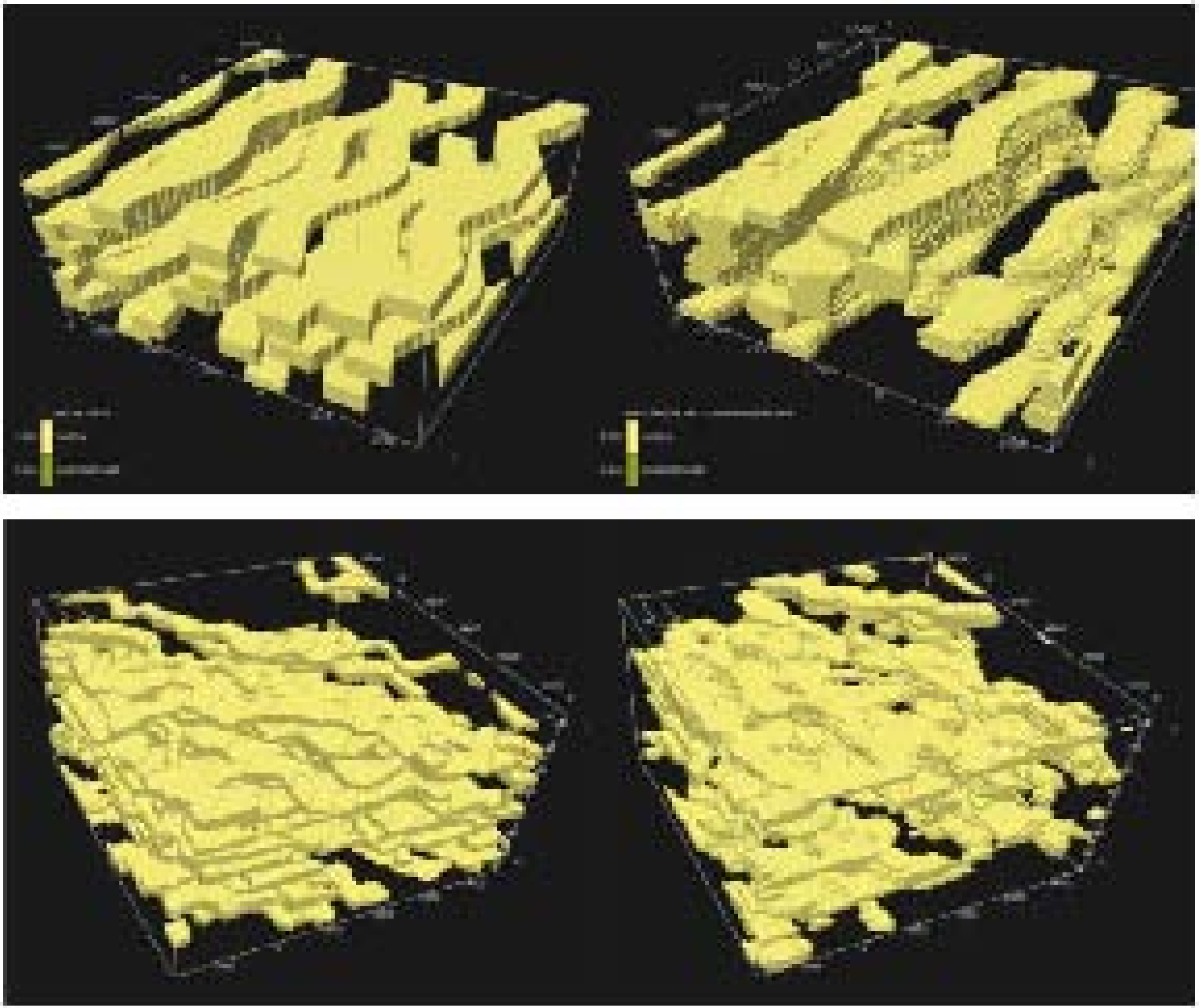
Stokastiska modeller och Bayesiansk staistik

- En *stokastisk* (eller probabilistisk) modell är en samling slumpvariabler som representerar observerbara delar av någon begränsad del av verkligheten.
- De variablerna som representerar nåt som inte är observerad representerar då en *probabilistisk prediktion*.
- Hur tar man fram en stokastisk modell från data? De vanligaste sätten är *frekventistisk* och *Bayesiansk*.
- Det frekventistiska sättet är att ta fram en modell med en *parametervektor* θ, och sen *estimera* denna från data.
- Det Bayesianska sättet är att presentera en modell där både θ, data, och det man vill predikera ingår som slumpvariabler. Man tar sen fram den *betingade modellen* där data-variablerna har fixerats till de observerade värdena.

Enkelt exempel

- Anta du gentar liknande oberoende försök 8 gångar. Antag sannolikheten för success är θ i varje försök. Antag 3 av 8 försök gav success. Vad är sannolikheten för success i nionde försöket?
- Frekventistisk lösning: $3/8$.
- Bayesiansk lösning:
  - En *apriori* sannolikhetsfördelnign för θ etableras, baserat på *kontexten* försöken görs i.
  - En *posteriori* sannolikhetsfördeling för θ tas fram, betingat på observerade data.
  - En prediktion för nionde försöket görs baserad på posteriorifördelningen för θ.

Svårare exempel: Modellering av oljereservoar



Bayesianer vs. frekventistister

- Filosofi: Vad är *sannolikhet*? (Existerar den objektivt eller bara subjektivt?)
- Frekventister: "Den Bayesianska lösningen är inte vetenskaplig, då den baserar sig på annat än data. Speciellt: Man kan faktiskt få *villket resultat som hellst* baserat på hur man väljer prior."
- Bayesianer: "Man *vill* faktiskt anpassa prediktionen till kontexten. Till exempel i det enkla exemplet över: Prediktionen borde bero på annat än 8 observationer (speciellt om tex. alla observationerna är "success").
- Många är "agnostiker", och använder metoder som de tycker passar till uppgiften.

Grundläggande verktyg i Bayesiansk statistik

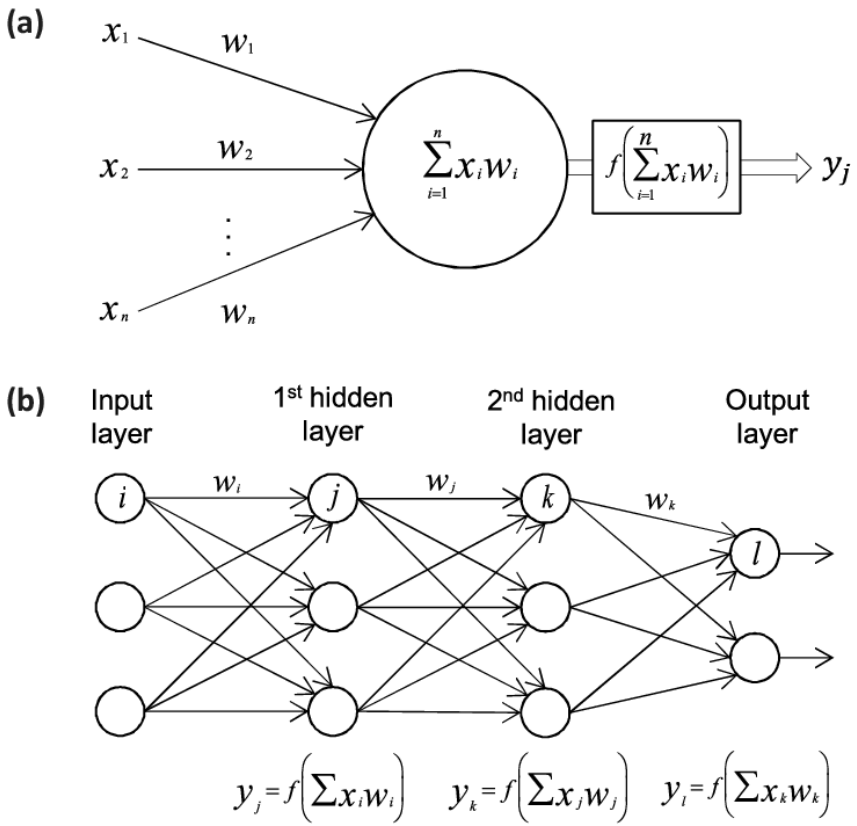
- Analytiska beräkningar (bara för enklaste modeller).
- Numeriska beräkningar (e.g., numerisk integration...)
- Simuleringar: Markov Chain Monte Carlo (MCMC), Sequential Monte Carlo, etc.
- Några nyare approximativa metoder, oftast för mera speciella modeller.
- Några kurser (där jag är lärare):
  - MVE550 Stokastiska processer och Bayesiansk inferens
  - MVE187 Beräkningsmetoder för Bayesiansk statistik
- Exempel (från MVE550): Kryptografiproblem löst med MCMC.

Tillämpning: Artificiell Intelligens och maskinlärning

- Detta är en grupp teknologier som är i framgång.
- Exempel: AlphaZero.
- Ingredienser som gör AI-framgångarna möjliga:
  - Datorkapasitet.
  - Sensorer / kommunikation.
  - Metoder för processering av stora datamängder.
  - Teoretiska / matematiska modeller för hur lärning kan göras.
- Några matematiska byggstenar:
  - Neurala nät.
  - Reinforcement learning.
  - Optimering.

Neurala nät

- Olika typer nätverk, som convolutional neural networks (CNN) etc.
- Genom att derivera hela den sammansatta funktionen kan man optimera vikterna (approximera Maximum Likelihood vikter).
- Kan sägas vara en generalisering av logistisk regression.
- När man har mycket data kan ofta "minibatching" användas.
- Stochastic Gradient Descent.
- Metoderna över verkar generellt undvika "overfitting".



Neurala nät, några matematiska problemställningar

- Algoritmerna för att "träna" neurala nät är främst utvecklat genom trial-and-error. *Varför* fungerar dessa algoritmer?
- Modellval: Hur skall man välja typ och storlek av nätverk, och hur man tränar nätverket?
- Hur kan man effektivt hitta nätverk som har färre variabler men fungerar lika bra?

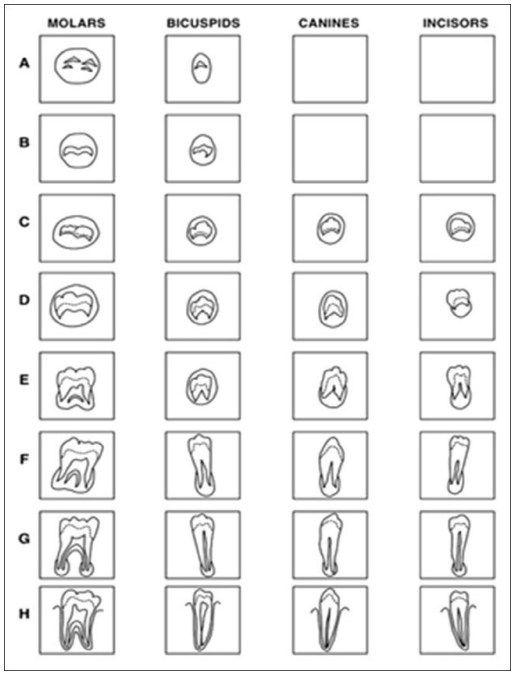
Tillämpning: Forensisk statistik

- Definition: Användning av statistik som verktyg inom *forensiska vetenskaper*, alltså vetenskap använt på *juridiska* frågeställningar.
- Några exempel:
  - DNA-spår i kriminalfall: Hur kan man hitta rätt match? Hur vad är *beviskraften* i en match?
  - Andra spår i kriminalfall.
  - DNA testing av släktskapsförhållanden.
  - Bestämning av dödstidspunkt vid dödsfall.
  - Åldersbedömning av asylsökande.
- En generell fråga är om Bayesiansk eller frekventistisk paradigm används. Inom Bayesiansk paradigm kan man använda teori för beslut under osäkerhet.
- På Nationalt Forensisk Center (NFC) i Linköping används i ökande grad Bayesiansk tänk. Men det varierar mycket mellan olika tillämpningsområden.

Medicinsk åldersbedömning

- Observation av medicinska karakteristika (”indikatorer”) som ändras vid hyfsat fasta åldrar.
- Exempel:
  - Tänder
  - Olika delar av skelettet
  - Pubertetsindikatorer, vikt, längd, ...
  - Psyko-social mognad
  - DNA-data, e.g., telomerlängd.
- Många olika syften
- Välj indikatorer som ändras mycket runt åldern relevant för syftet.
- Syftet här: Bedöma över/under 18 år. Ofta använda indikatorer: Tänder, handledsmognad, nyckelbensmognad.

Exempel: Visdomständer



- Man tittar på rötternas utveckling, och använder röntgenbilder.
- Klassificeringsschema: Demirjian (finns även andra)
- Speciellt visdomständer ändrar sig till sin ”mogna” form (H) i slutet av tonåren.
- Åldern då en person får ”mogna visdomständer” varierar med ett par-tre år.
- Åldern beror även på kön. Beroenden på genetisk bakgrund och uppväxtvillkor diskuteras.

Exempel: Knän

- Man använder NMR undersökning, inte röntgenbilder.
- Undersökning av distal femur, och dens tillväxtzon. Metoder varierar.
- Åldern för mognad är oftast i sena tonåren.
- Variation, och beroenden på kovariater som kön, genetisk bakgrund, och uppväxtvillkor är inte mycket undersökt.

Åldersbedömning av asylsökare i Sverige

- Under 2014-15 ansökte 244.178 personer om asyl i Sverige. Bland dessa: 42.418 "ensamkommande".
- För 2016-17 sjönk siffrorna till 52.667 och 3435.
- Behandling av en asylansökan är starkt beroende av om personen är över eller under 18 år.
- Juridiskt åligger det asylsökaren att "styrka" sin "identitet", inkluderat ålder.
- Personer från Afghanistan/Somalia/... saknar ofta dokumentation som intyger ålder. Om dokumentation finns så anses den inte trovärdig av Migrationsverket.
- Några asylsökare tog tidigare själva initiativ till medicinsk åldersbedömning.

Standardiserad åldersbedömning via Rättsmedicinalverket

- Sedan 2017 erbjuds asylsökare standardiserad åldersbedömning via Rättsmedicinalverket (RMV) som alternativ till att Migrationsverket fastställer åldern. Andra åldersbedömningar accepteras inte.
- RMV "outsourcer" insamling av data till olika laboratorier: Röntgenbilder av visdomständer och NMR av knän.
- Experter, två för varje datatyp, bestämmer om åldersindikatorn är **mogen**, **inte mogen**, eller **inte bedömbar**.
- Båda experter behöver bedöma indikatorn som mogen för att den skall anses vara mogen.
- Åt andra hållet så bedöms personen vara över 18 år om minst en av indikatorna är mogen (gäller killar).
- MÄRK: RMV producerar olika textliga konklutioner i några olika fall. Migrationsverkets beslut görs dock i regel bara på grundlag denna konklutionen, och på ett sätt som motsvarar beskrivningen över.

Problemer med denna beslutsprocedur

- Oftast ser man bort från all osäkerhet i metoden. Beslut om ålder baseras bara på RMVs konklution. Ingen annan information i fallet tas hänsyn till.
- Metodens egenskaper som beslutsregel är högst oklara: **Ingen valideringsstudie, där metoden har använts på personer med känd ålder, har publicerats.**
- Ett antal gångar har det framkommit information som gör det naturligt att ifrågasätta RMVs egen beskrivning av metodens egenskaper. T.ex.:
  - Bedömning av tjejer.
  - Second-opinion värdering av knä-data gav nytt resultat i 55% av 137 fall.
  - Antalet killar med moget knä och omogen tand är 4-5 gångar så många som antalet med omoget knä och mogen tand. Svårt att förklara om, som RMV har angett, tänder mognar tidigare än knän.

Kan statistiska metoder öka kunskapen om egenskaperna till RMVs metod?

- Jag önskade ta reda på hur mycket det går att säga om metoden, och om undersökta asylsökares ålder, med den information som finns.
- Tillgängliga data: Klassificeringsdata för killar, 2017:

	Moget knä	Omoget knä	Inga data	SUMMA
Mogen tand	4176	348	187	4711
Omogen tand	1735	1087	83	2905
Inga data	1364	237	63	1664
SUMMA	7275	1672	333	9280

I tillägg all information som finns i literaturen om åldersindikatorerna.

- Jag ansökte Juni 2017 om mera specifika data från RMV. Jag har mottagit vissa data november 2018, och mera kompletta data augusti 2019.
- **Mostad, Tamsen: Error Rates for Unvalidated Medical Age Assessment Procedures** publicerat i **International Journal of Legal Medicine**. ▸ Link

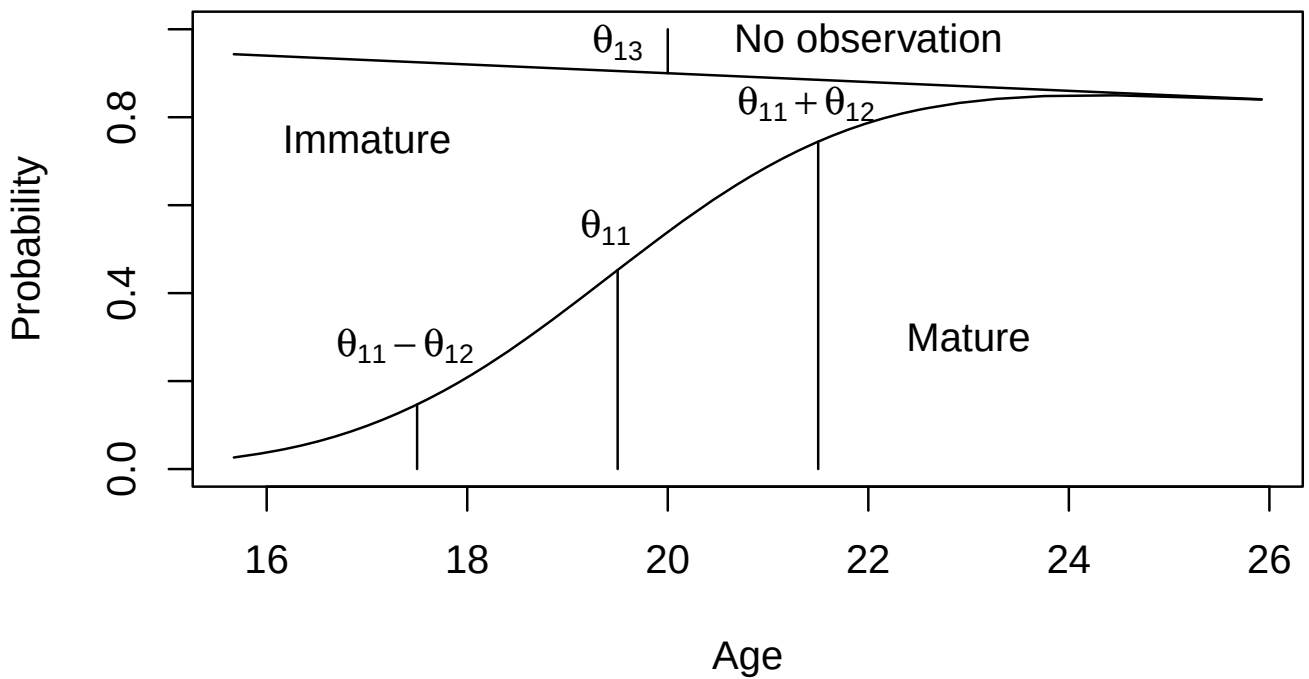
Mognad av en åldersindikator som funktion av ålder

Parametrar $\theta_k = (\theta_{k1}, \theta_{k2}, \theta_{k3}, \theta_{k4})$ beskriver relationen mellan kronologisk ålder x och åldersindikator k ($k = 1$: tand, $k = 2$: knä).

$$
\begin{aligned}
p_{k1}(x) &= (1 - p_{k3}(x))\, \Phi\left(\tfrac{x-\theta_{k1}}{\theta_{k2}}\right) && \text{P(mogen)} \\
p_{k2}(x) &= (1 - p_{k3}(x)) \left(1 - \Phi\left(\tfrac{x-\theta_{k1}}{\theta_{k2}}\right)\right) && \text{P(omogen)} \\
p_{k3}(x) &= \theta_{k3} + \theta_{k4}(x - 20) && \text{P(inga data)}
\end{aligned}
$$

Below: $\theta_{11} = 19.5$, $\theta_{12} = 2$, $\theta_{13} = 0.1$, and $\theta_{14} = 0.01$.

An age indicator model:



Modellvariabler

- $\theta = (\theta_1, \theta_2) = ((\theta_{11}, \ldots, \theta_{14}), (\theta_{21}, \ldots, \theta_{24}))$: Parametrar för modeller för åldersindikatorer.
- $\psi = (\psi_1, \ldots, \psi_{100})$: Sannolikhetsvektor med sannolikheter att testade personer har specifika åldrar $x_1, \ldots, x_{100}$. ($x_i \in [15, 30]$).
- $\tau = \{\tau_{ij}\}$, $i = 1, \ldots, 100$; $j = 1, \ldots, 9$: Antal personer med ålder x_i klassificerat av RMV till kategori j:
(mogen/mogen, mogen/omogen, ..., inga data / inga data)
- $y = (y_1, \ldots, y_9)$: Observerade data, alltså det totala antalet personer klassificerat av RMV till vaje kategori 1,...,9.

Stokastisk modell

$$\pi(y, \tau, \psi, \theta) = \pi(y \mid \tau)\pi(\tau \mid \psi, \theta)\pi(\psi)\pi(\theta)$$

- $\pi(y \mid \tau)$ är deterministisk: Summerar över åldrarna.
- $\pi(\tau \mid \psi, \theta)$ är Multinomialfördelad, eftersom ψ och θ tillsammans specificerar sannolikheten för varje kategori.
- $\pi(\theta)$ är trunkert multivariat normalfördelad, anpassad med data från ett antal publikationer.
- $\pi(\psi)$ är Dirichlet-fördelad. Vi sprider åldrarna $x_1, \ldots, x_{100}$ *ojämnt* över intervallet $[15, 30]$ så att den mest sannolika åldersfördelningen är en Gamma$(4, 1)$ fördelning försjutet så den startar vid 15 och är trunkerad vid 30. Stor möjlig variation runt denna åldersprofil används.

Parameterestimater från literaturen

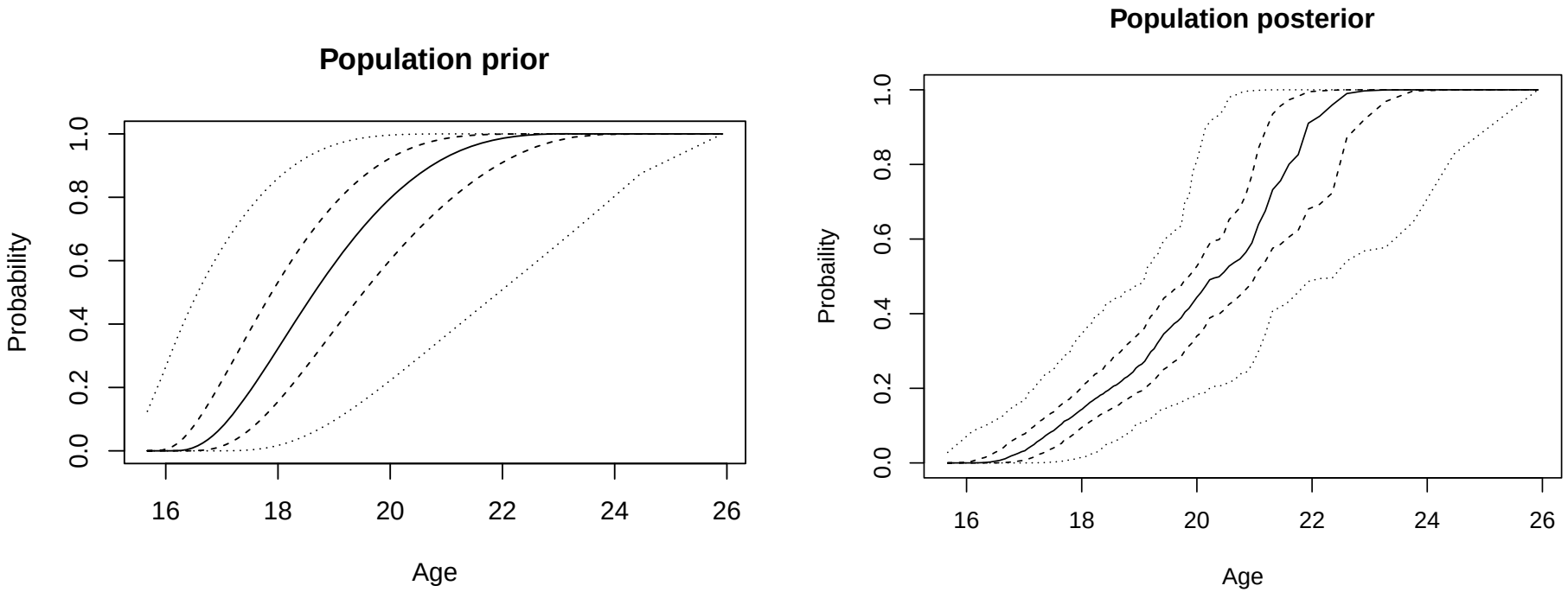
- Tand-parametrar estimeras från följande publikationer / databaser:
  - DARL: https://www.dentalage.co.uk/rds-uk-caucasian
  - Lucas et al (2016) "Dental age estimation: ..."
  - Mincer et al (1993) "The ABFO study..."
  - Haglund et al (2018) "A systematic review and meta-analysis..."

	DARL	Lucas	Mincer	Haglund	Prior
θ_{11}	19.5	18.6	19.9	20.9	19.5
θ_{12}	1.6	0.8	2.2	2.5	1.6

- Knä-parametrar estimeras från följande publikationer:
  - Soc.s.:Socialstyrelsen (2018) Om magnetkamera vid bedömning av ålder.
  - Ottow et al (2017) "Forensic age estimation by magnetic resonance imaging of the knee..."
  - Adj. Ott.: Using adjusted data from Ottow et al.

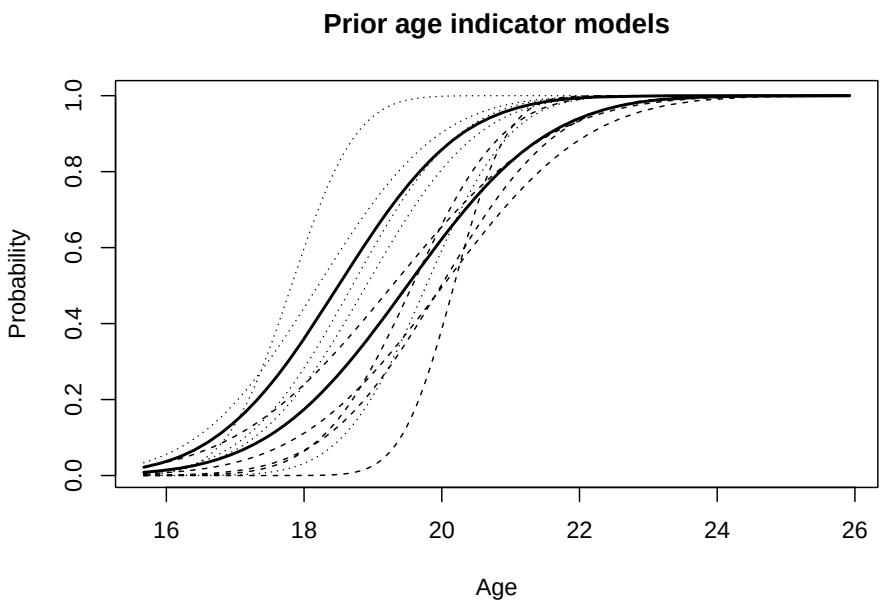
	Soc.s.	Ottow	Adj. Ott.	Prior
θ_{21}	18.5	18.5	17.7	18.5
θ_{22}	1.3	1.5	1.4	1.4

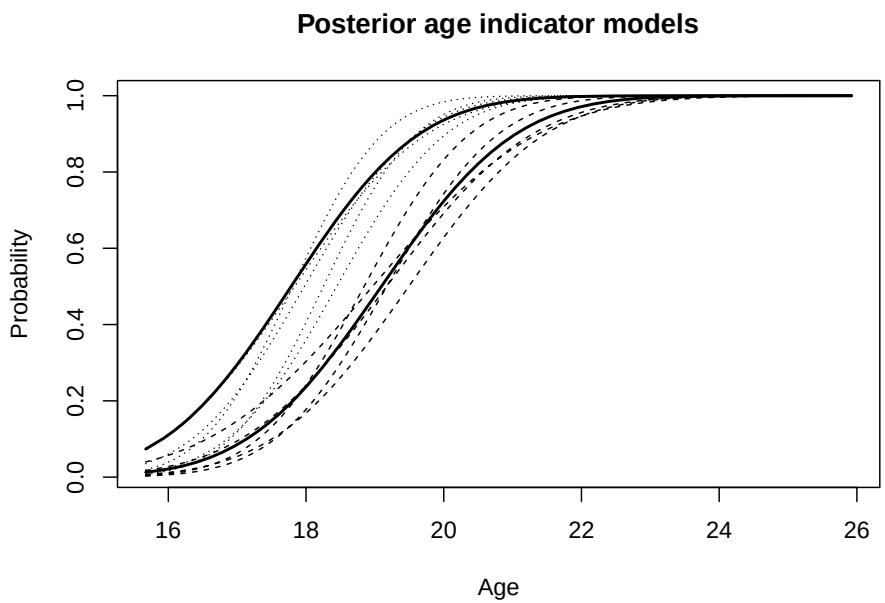
A priori och posteriori ålderfördelning: ψ



Figurerna visar kumulativ åldersfördelning: Apriori till vänster och posteriori till höger. De inre banden visar ett 50%-ig kredibilitetsintervall. De yttra banden visar ett 95%-ig kredibilitetsintervall.

A priori och posteriori parametrar θ för åldersindikatormodeller





Figurerna visar a priori (vänster) och posteriori (höger) åldersindikatormodeller. I varje plot reprecenterar höger heltrukna linje tänder medan vänster heltrukna linje representerar knän. De stiplade linjerna representerar möjliga modeller simulerade under varje fördelning.

Skattade resultat för killar och män testade under 2017

	Klass. som vuxna	Klass. som barn	Inte klass.	SUMMA
Vuxna	7260 (5908 – 7794)	581 (116 –1305)	59 (49 –63)	7900 (6102–8570)
Barn	550 (16 – 1902)	826 (102 –1291)	4 (0 – 14)	1380 (133 –3379)
SUMMA	7810	1407	63	9280

Tabellen visar den mest troliga siffran i varje grupp. Parenteserna visar 95%-iga kredibilitetsintervaller.

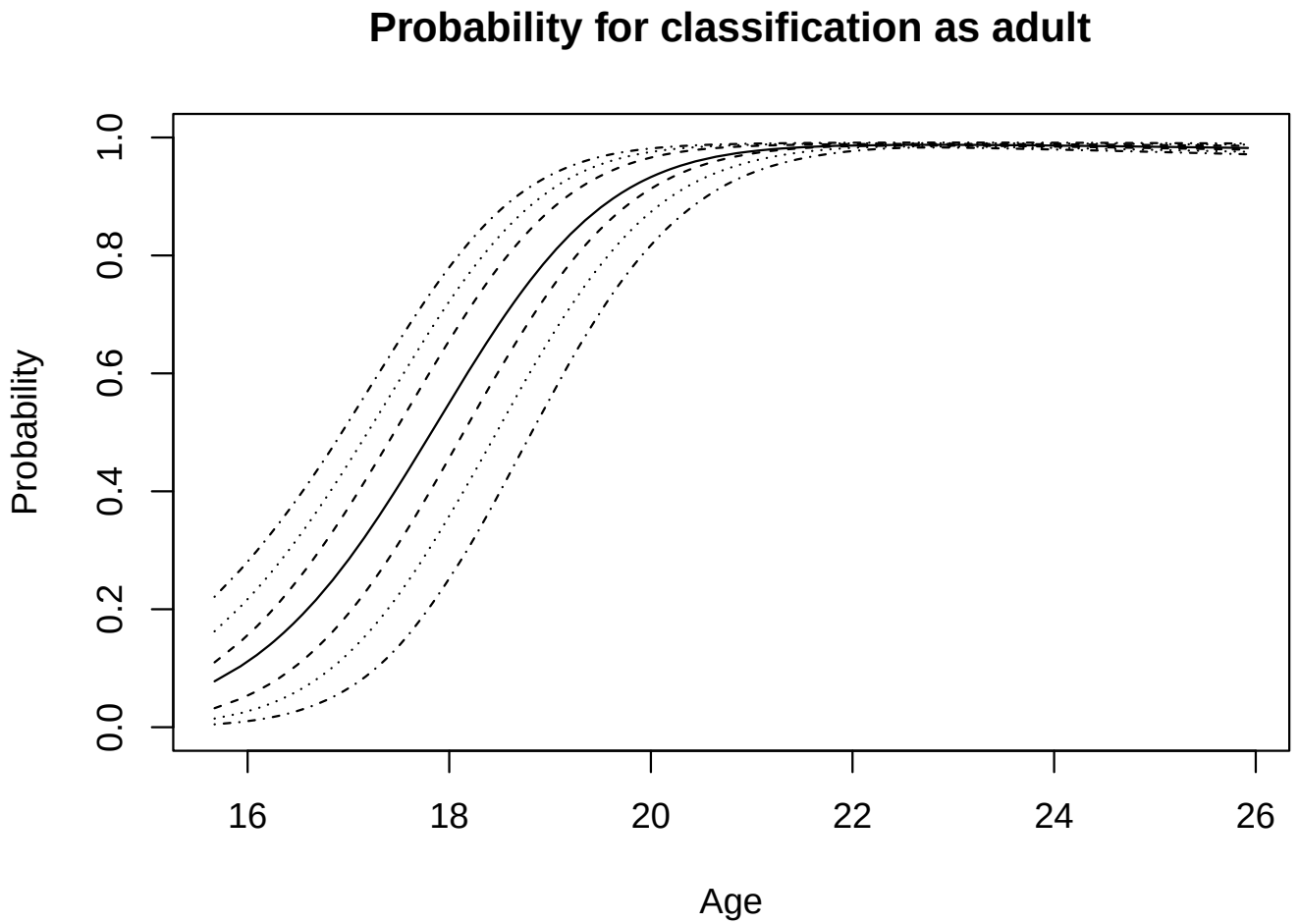
Sensitivitet 93% (CI: 86-98), specificitet 67% (CI: 39-94), Positivt prediktivt värde 93% (CI: 76-100), Negativt prediktivt värde 59% (CI: 7-92).

Andel barn i varje klassificeringsgrupp

	Moget knä	Omoget knä	Inga data knän	SUMMA
Mogna tänder	1 (0–8)	24 (8–78)	2 (0–9)	3 (0–12)
Omogna tänder	19 (1–64)	63 (8–95)	28 (2–70)	36 (4-74)
Inga data tänder	5 (0–17)	48 (4 –88)	7 (0–22)	11 (1–27)
SUMMA	6 (0–23)	53 (6–90)	9 (1–26)	15 (1–34)

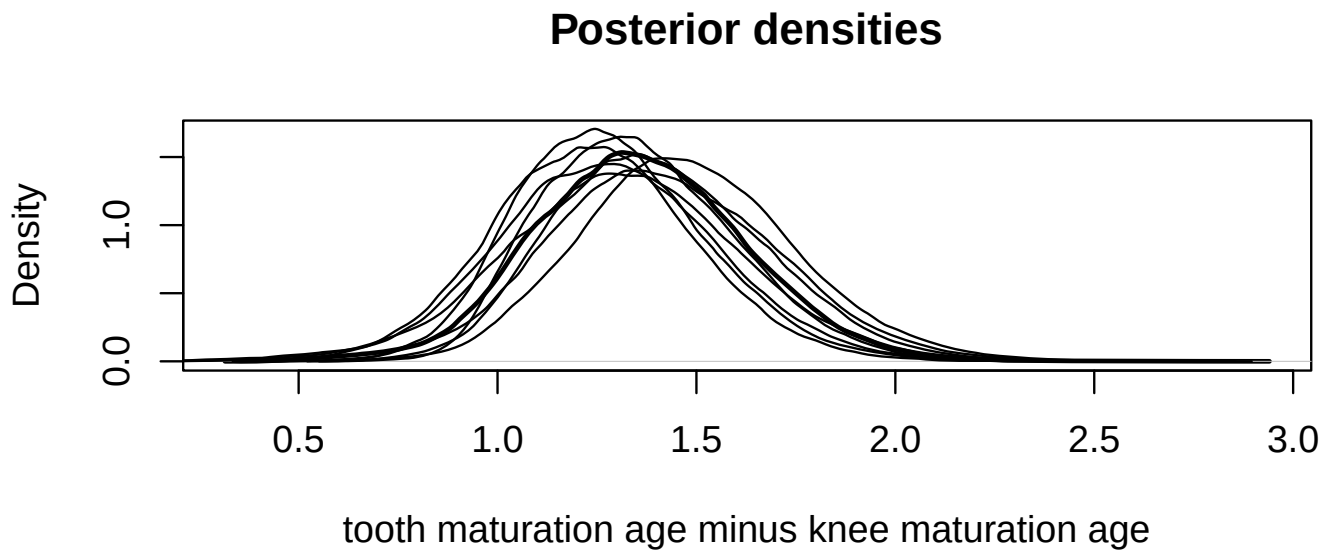
Procentandel barn i varje kategori (95%-iga kredibilitetsintervaller in parenteserna). Cellerna med grå bakgrund representerar de där RMVs procedur klassificerar killar/män som vuxna.

Konsekvenser av posteriorifördelningen för θ



De med ålder mellan 17 och 18 blir klassificerade som vuxna med sannolikhet 41%. Ett 95%-ig kredibilitetsintervall för denna siffran är 12%-70%.

Knän mognar före tänder



- RMV har gjort uttalanden om att knän generellt mognar efter tänder.
- Den kraftiga linjen över viser posteriori fördelning för differensen mellan åldern då 50% av pojkar har fått mogna tänder och åldern då 50% av pojkar har fått mogna knän.
- De andra linjerna representerar resultat om man använder alternativa aprioriförde lningar: Vi har testat ett antal alternativer för att utvärdera robustheten av våra konklutioner.

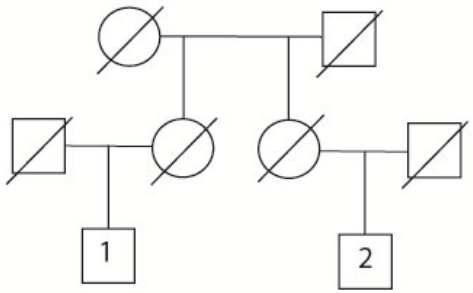
Kommentarer om beräkningarna

- Vi använder en MCMC (Markov chain Monte Carlo) algoritm för att simulera från posteriorifördelningen.
- Mera specifikt använder vi Gibbs sampling över de tre parametrarna θ, ψ och τ, med en random-walk förslagsfunktion för θ och direkt sampling från betingade fördelningar för ψ och τ.
- Konvergens var långsam, och därmed användes långa simuleringskädjor.
- En burn-in på 20.000 iterationer från totalt 1.000.000 iterationer blev använd för beräkningar av resultat.
- Ett antal kontroller blev gjort för att utforska robustheten i resultaten i relation till ändringar i aprioriförderlningen.

Viktigaste konklutioner om RMVs åldersbedömningar

- Stokastisk modellering gör det möjligt att få viss information både om hur åldersbedömningen fungerar, och åldern till de åldersbedömda.
- Några utvalda resultat:
  - 85% (66-92) av åldersbedömda killar under 2017 var över 18.
  - Bland de som bedömts som vuxna därför att de hade mogna tänder och omogna knän så var 24% barn (8-78).
  - Bland 17-åringar var sannolikheten för att bedömas som vuxen 41% (12-70).
  - Knän mognar generellt ungefär 1-1.5 år innan tänder.
- Tolknig och användning av RMVs bedömnignar har baserats på information från RMV om deras procedur. Delar av denna information har vi visat är felaktig. Detta har skapat en rättsosäker situation för asylsökare.

Tillämpning: Beviskraft vid DNA-tester för släktsskap



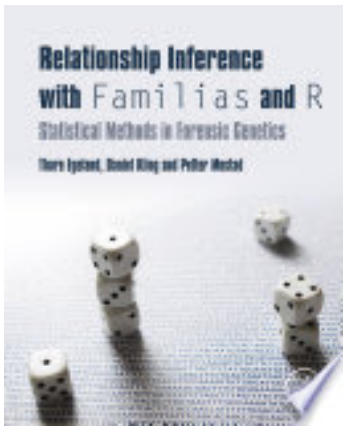
- Frågeställning: Givet DNA test data (för vissa "DNA-markörer"), vad är relativ beviskraft för olika släktsskapshypoteser, så som "kusiner", "orelaterade", "farbror", etc.
- Beräkningar involverar modellering av arv (enligt Mendel's lagar), populationseffekter, association och "linkage" mellan markörer, mutationer, och observationsfel.
- Den enklaste tillämpningen är faderskapssaker. Men mera avancerade frågor inkluderar t.ex. sök efter saknade personer.

Exempel på släktsskapsberäkning

- En rik man dör utan egna barn. En kvinna påstår sig vara hans brorsdotter, och vill dela på arvet i hop med andra släktningar. Det finns DNA tester av henne och dessa släktningar.
- För varje möjlig hypotes om familjförhållanden (oftast bara två) beräknas sannolikheten för observerade DNA test data. Kvoten av dessa (LR, likelihood ratio) är beviskraften för dessa DNA data i denna frågan.
- För vanliga faderskapssaker kan LR lätt bli över en miljon, och konklutionen räknas som "säker". För saker som den över, kan LR lätt bli mellan 0.01 och 100, och eventuella konklutioner är ganska osäkra.

Några resultat för släktskapstesting

- Programmet `Familias` för Windows (www.familias.no) och som R paket (www.familias.name).
- Boken "Relationship Inference with Familias and R".



- Exempel på användningsområde: Spårning av släkt till bortförda barn i Argentina.

Vetenskapsteori

- Hur tar man reda på och kommer överens om vad som är sanning?
- *Vetenskaplig metod* är helt centralt som grundlag för att ta fram sanningen. Är vi överens om vad vetenskaplig metod är?
- En formalisering av hur vetenskaplig metod fungerar kan använda matematisk statistik, och speciellt Bayesiansk statistik och beslutsteori, som ramvärk.
- Min åsikt: Vetenskapsteori, som matematiserad vetenskap, är underutvecklad, och ett viktigt framtida forskningsområde.