Niels Richard Hansen

Matematisk Statistik

supplerende noter

INSTITUT FOR MATEMATISKE FAG
KØBENHAVNS UNIVERSITET

INSTITUT FOR MATEMATISKE FAG
KØBENHAVNS UNIVERSITET
UNIVERSITETSPARKEN 5
2100 KØBENHAVN Ø

Forord

I en årrække har der været undervist i matematisk statistik på Københavns Universitet efter 3. udgave af bogen "Introduktion til Matematisk Statistik" (IMS) af Ernst Hansen (EH). Pensum i faget har imidlertid taget en drejning væk fra indholdet af IMS, og det har derfor været nødvendigt at udvikle nyt materiale til en del af kurset. Bind 1 af IMS bruges således ikke mere, mens kapitlerne 9–14 i IMS, bind 2, fortsat benyttes.

Materialet i disse noter er et supplement, og kapitel 1 fungerer som en introduktion til IMS, bind 2, der så kan læses uafhængigt af IMS, bind 1. Men disse supplerende noter og IMS, bind 2, bør stadig læses sammen med en generel behandling af matematisk statistik. I kurset vil bogen "Basic Mathematical Statistics" (BMS) af Steffen Lauritzen blive benyttet. Kapitel 1 indeholder desuden resultater om transformationer af normalfordelte variable, som tidligere kunne findes i bogen "Measure Theory", ligeledes af EH.

Kapitel 2 i disse noter er skrevet med en vis inspiration fra kapitel 15 og 16 i IMS, bind 2, såvel som materiale om eksponentielle familier af Søren Tolver Jensen. Denne tilføjelse afspejler på samme måde som BMS en ændring af pensum i kurset, og er desuden baseret på resultater om eksponentielle familier, som kan findes i BMS.

Niels Richard Hansen
København, januar, 2021

Indhold

Kapitel 1

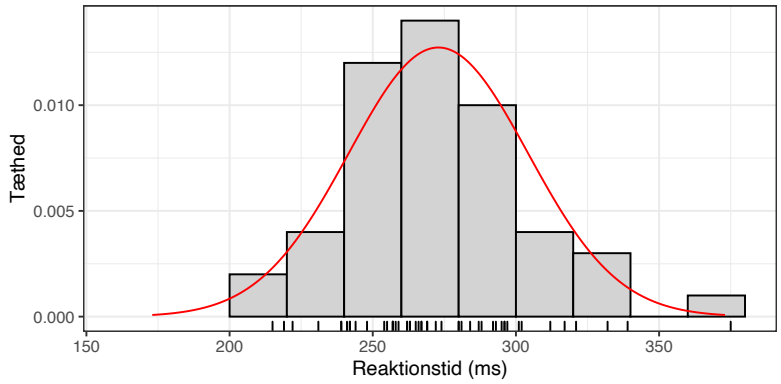
Statistik, likelihood og normalfordelingen

Dette kapitel giver en kortfattet introduktion til matematisk statistik med normalfordelingen. Kapitlet skal tjene som motivation og grundlag for de efterfølgende kapitlers udførlige analyser af statistiske modeller baseret på normalfordelingen.

Kapitlet introducerer nogle helt grundlæggende begreber såsom en parametriseret statistisk model, en log-likelihoodfunktion og en maksimaliseringsestimator ved hjælp af et normalfordelingseksempel, men gør intet forsøg på abstrakt og generelt at definere alle relevante begreber. Til gengæld diskuteres det i nogen detaljegrad, hvad vi skal med en model, hvordan vi fortolker sandsynligheder og modeller, og hvad det er statistik handler om. Alt sammen illustreret gennem normalfordelingseksempler.

1.1 En statistisk model

Normalfordelingen er kendt fra indledende kurser i statistik og sandsynlighedsregning, og den dukker ofte op som en model for data. Figur 1.1 viser et histogram af $N = 50$ reaktionstider i millisekunder, $x_1, \ldots, x_N$, målt for en forsøgsperson i et eksperiment udført i forbindelse med kurset Statistik 1TS i 2002. Målingerne er foretaget ved at forsøgspersonen ser et signal og derefter skal trykke på en knap. Reaktionstiden er tiden fra signalet vises til der trykkes på knappen. Figuren viser også tætheden



Figur 1.1: Histogram og rug plot af 50 reaktionstider. Den røde kurve er tætheden for normalfordelingen med samme middelværdi og varians som data.

for en normalfordeling, som har samme middelværdi og varians som data. Middelværdi og varians for data er her beregnet som gennemsnittene

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} x_i = 273 \tag{1.1}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\alpha})^2 = 982, \tag{1.2}$$

som også er middelværdi og varians for den empiriske fordeling

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}.$$

Umiddelbart er der en rimelig overensstemmelse mellem data og normalfordelingsmodellen. Men hvad skal vi med modellen?

Før vi forsøger at besvare, hvad modellen skal bruges til, bør vi også overveje flere andre spørgsmål.

- Hvorfor har vi indsamlet data? Hvilke spørgsmål ønsker vi, at data skal hjælpe os med at svare på?

- Hvad er en model overhovedet?
- Hvad er det for en mekanisme, der har frembragt data? Hvordan kan vi beskrive og formalisere mekanismen?
- Hvad mener vi egentlig med en "rimelig overensstemmelse" mellem data og model, og hvordan formaliserer vi det?

I faget *matematisk statistik* bruger vi matematik til at formalisere og besvare ovenstående spørgsmål og andre tilsvarende spørgsmål, der handler om, hvordan vi bruger og analyserer data.

I eksperimentet med reaktionstider kunne formålet simpelt hen være at måle en persons middelreaktionstid. Gennemsnittet af de målte reaktioner for forsøgspersonen er 273 ms. Til sammenligning har dupreeh fra det profesionelle danske CS-team Astralis en middelreaktionstid på 162 ms.[1] For professionelle CS-spillere er middelreaktionstid en oplagt interessant størrelse at måle, men i andre sammenhænge kan den også være interessant. En langsom reaktionstid kan være udtryk for at personen er påvirket af alkohol eller andre stoffer. Eller at personen er syg, og målingerne kunne således indgå som en del af en diagnostisk test. Vi kunne også foretage målinger på flere forskellige personer eller på den samme person under forskellige forhold, og så kunne vi være interesserede i at lave sammenligninger. Er dupreeh f.eks. hurtigere end dev1ce?

Middelreaktionstid afspejler imidlertid ikke hele sandheden om en persons reaktionstid. Det fremgår af histogrammet i figur 1.1 at der er en betydelig variation fra måling til måling i personens reaktionstid. Reaktionstid er derfor en *fordeling* og ikke et tal! De 50 specifikke målinger er eksperimentets empiriske fordeling, men vi kan næppe tænke på disse 50 tal som en fuldstændig karakterisering af personens reaktionstid. De er måske nok repræsentative, men andre reaktionstider er mulige, så hvordan opsummerer vi målingerne på en måde, der tillader generalisering fra de 50 målinger? Histogrammet er i sig selv en opsummering af de rå datapunkter, og det er tættere på en generaliserbar karakterisering, end den empiriske fordeling er. Normalfordelingen er en grovere opsummering, som ligeledes generaliserer fra de observerede datapunkter, og det er en *model*, som karateriserer personens reaktionstidsfordeling på en simpel måde.

[1] `https://www.dr.dk/ligetil/sport/er-du-lige-saa-hurtig-som-astralis`

Normalfordelingen med middelværdi α og varians σ^2 er en sandsynlighedsfordeling på $\mathbb{R}$ med tæthed

$$f_{\alpha,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\alpha)^2}.$$

Som *sandsynlighedsmodel* kan den for givne værdier af parametrene α og σ^2 bruges til at kvantificere udsagn om reaktionstiden, og vi kan f.eks. udtrykke sandsynligheden for at få en reaktionstid, X, mindre end 200 ms som

$$P(X \leq 200) = \int_{-\infty}^{200} f_{\alpha,\sigma^2}(x)\mathrm{d}x.$$

Den stokastiske variabel X siger vi er normalfordelt med middelværdi α og varians σ^2, og vi benytter ofte betegnelsen $X \sim \mathcal{N}(\alpha, \sigma^2)$.

Forskellige parametre vil resultere i forskellige sandsynligheder. Hvis $(\alpha, \sigma^2) = (273, 982)$ som på figur 1.1 kan vi beregne ovenstående sandsynlighed til $P(X \leq 200) \simeq 0,010$, som altså er ret lille, mens $(\alpha, \sigma^2) = (220, 950)$ giver en noget større værdi på $P(X \leq 200) \simeq 0,26$. Man kan indvende, at en reaktionstid selvfølgelig ikke kan være normalfordelt, fordi den ikke kan være negativ! Men for relevante valg af parametre vil normalfordelingen tillægge negative værdier en helt ubetydelig sandsynlighed – for $(\alpha, \sigma^2) = (273, 982)$ er den 1.5×10^{-18}. I praksis er det derfor intet problem at bruge normalfordelingen selvom den af principielle grunde må være lidt forkert.

Selve udregningen af en sandsynlighed for en given model er ukontroversiel, men fortolkningen af en sandsynlighed er et stridspunkt indenfor statistik. Alle fortolkninger er nogenlunde enige om, at sandsynligheder kvantificerer en *forventning* eller en *grad af tiltro* til udfald eller udsagn, der er usikkerhed omkring. Hvis $(\alpha, \sigma^2) = (273, 982)$ forventer vi ikke en reaktionstid mindre end 200 ms da sandsynligheden for det er ret lille, men hvis $(\alpha, \sigma^2) = (220, 950)$ har vi en vis forventning til at observere en reaktionstid mindre end 200 ms. Men her stopper enigheden. For hvad betyder forventning og usikkerhed helt præcist? Er forventning en objektiv størrelse, karakteristisk for forsøget, eller er det en iagtagers subjektive opfattelse? Og er usikkerhed iagtagerens mangel på viden eller en irreducibel egenskab ved forsøget. Vi diskuterer hovedlinjerne i den bayesianske og den frekventistiske fortolkning nedenfor, og hvilke konsekvenser disse fortolkninger har for, hvordan vi laver statistik, men i første omgang stiller vi os tilfredse med en lidt løs *forventningsfortolkning*.

Normalfordelingen ovenfor er en model for en enkelt observation i $\mathbb{R}$. Data består derimod af N observationer:

$$x_1, \ldots, x_N \in \mathbb{R}.$$

En enkelt reaktionstid kan modelleres med en normalfordeling på $\mathbb{R}$, men hvad med de $N = 50$ reaktionstider, som eksperimentet resulterede i? Hvordan skal vi koble sandsynlighedsmodeller for de enkelte observationer sammen til en model for alle N observationerne? En simpel matematisk kobling er produktmålet, f.eks. fordelingen på $\mathbb{R}^N$ med tæthed

$$f_{(\alpha,\sigma^2)}^{\otimes N}(x) = \prod_{i=1}^{N} f_{\alpha,\sigma^2}(x_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\alpha)^2}, \quad x = (x_1, \ldots, x_N)^T \in \mathbb{R}^N$$

mht. lebesguemålet, hvor alle marginalfordelingerne er $\mathcal{N}(\alpha, \sigma^2)$. Den sandsynlighedsteoretiske fortolkning af produktmålet er, at observationerne er *uafhængige*, dvs. at vores forventning til X_i ikke ændres af viden om at $X_j = x_j$ for $j \neq i$. Antagelsen om uafhængighed tillader os altså at løfte modeller for de enkelte observationer til en model for hele datasættet, og dermed kan vi udtrykke forventninger til udsagn vedrørende alle datapunkter. Vi kan f.eks. beregne

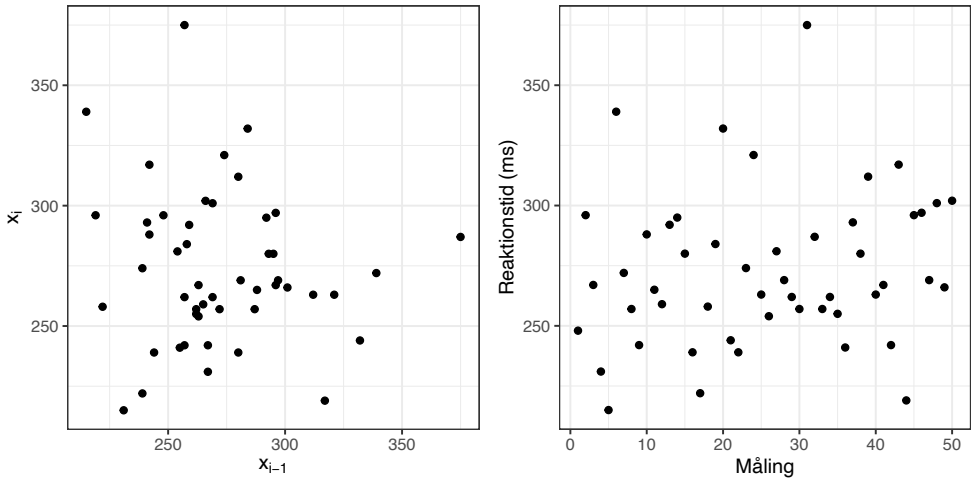
$$P\left(\frac{1}{N}(X_1 + \ldots + X_N) \in A\right)$$

for $A \subseteq \mathbb{R}$, hvis vi ønsker det, og dermed finde fordelingen af gennemsnittet, jf. opgave 1.1.

For de fleste er det lidt af en udfordring at få en intuition omkring uafhængighed. Produktmålet er måske nok en rimelig formalisering af uafhængighed, men hvornår kan vi i praksis antage uafhængighed? Hvilke argumenter kan vi bruge for at retfærdiggøre sådan en antagelse?

Lad os først formulere en teori omkring mekanismerne bag reaktionstiderne. Af neurofysiologiske grunde må der nødvendigvis gå lidt tid fra signal til reaktion – signalet skal jo have tid til at nå fra øje til hjerne og blive omsat til reaktionen, hvor forsøgspersonen trykker på knappen. Det er fint i overensstemmelse med at reaktionstiderne ligger i intervallet 215 til 375, og i hvert fald holder sig pænt væk fra 0. Der er dog tydeligvis en del variation fra måling til måling. Ydre omstændigheder ved forsøget, som ændrer sig fra måling til måling, kunne principielt forklare noget af variationen. En telefon kunne forstyrre under måling 11 til 17. Eller måske skulle forsøgspersonen bruge højre hånd i de første 25 test og venstre i de sidste 25. Det er også muligt, at noget af variationen kunne forklares med, at forsøgspersonen blev udmattet og langsommere til sidst, eller med at han havde lært noget og blev hurtigere.

Ændringer i ydre omstændigheder kan skabe afhængighed. Rimeligheden af at antage uafhængighed er i høj grad baseret på, hvorvidt vi kender nogle ydre omstændigheder,



Figur 1.2: Lag plot af reaktionstiderne (venstre) og reaktionstidene plottet mod målingsnummer (højre).

som kan forklare en eventuel afhængighed. En praktisk tilgang til antagelsen om uafhængighed er at spørge "hvorfor ikke?". Hvis vi ikke kan finde en mekanisme, der kan skabe afhængighed, hvorfor så ikke antage uafhængighed? Og hvis vi kan finde en potentiel mekanisme, så kan vi også undersøge, om den skaber afhængighed.

Hvis vi f.eks. tror at en hurtig reaktion tildels forhindrer en efterfølgende hurtig reaktion, kan vi undersøge det ved at plotte x_i mod x_{i-1}. Figur 1.2 viser sådan et lag plot, og der er ikke noget der tyder på en afhængighed mellem X_i og X_{i-1}. Små værdier af x_{i-1} giver ikke systematisk anledning til større eller mindre værdier af x_i end store værdier af x_{i-1}.

Figur 1.2 viser også reaktionstiderne plottet i den rækkefølge de er målt. Vi kan heller ikke her se nogen systematik. Reaktionstiderne ser ikke ud til at blive hverken hurtigere eller langsommere som forsøget skrider frem, og der ser heller ikke ud til at være en "klump" af målinger, der er systematisk mindre eller større end resten. Alt i alt understøtter figur 1.2 at målingerne er såvel uafhængige som identisk fordelte – men den er på ingen måde et bevis for, at det er tilfældet.

Forsøget er gennemført således at de ydre omstændigheder burde være konstante gennem hele forsøget. Vi kender altså ikke umiddelbart nogen mekanisme, der kan skabe

afhængighed mellem målingerne eller ændre fordelingen af målingerne over tid. Men på trods af de konstante omstændigheder er der en betydelig uforklaret variation fra måling til måling. Vi kan henvise til opmærksomhedsteorier fra den kognitive psykologi for modeller og beskrivelser af, hvordan vores opmærksomhed ikke kan holde 100% fokus på forsøget hele tiden. Det kan nok forklare variationen som fænomen, men vi kan ikke forklare hvorfor den enkelte måling er enten stor eller lille. Den tilbageværende variation er uforklarlig og usystematisk, men den kan beskrives statistisk ved en fordeling.

Om variationen er *tilfældig* er meget svært at afgøre uden en formel definition af tilfældighed, og mindre kan gøre det. Statistiske anvendelser benytter i udpræget grad fordelinger på en pragmatisk måde til at beskrive uforklarlig og usystematisk variation uden at tage filosofisk stilling til, om variationen er tilfældig. Antagelsen om uafhængighed er dermed i sidste ende en antagelse om, at de uforklarlige og usystematiske variationer fra måling til måling ikke har konspireret om, i hvilken retning de påvirker udfaldene. Antagelsen om at de er identisk fordelte er en antagelse om, at det uforklarlige og usystematiske opfører sig på samme måde for alle målingerne.

Vi skal ikke tage let på antagelsen om uafhængighed. Men den må heller ikke paralysere os. I al anvendt statistik vil der et eller andet sted gemme sig en antagelse om, at nogle størrelser er uafhængige. De modeller, vi bygger, vil forsøge at fange alle de former for systematisk variation og afhængighed, som vi kender eller kan tænke os til, mens den resterende usystematiske variation er modelleret som uafhængige stokastiske størrelser. Og vi vil næppe være i stand til at argumentere overbevisende for, hvorfor disse størrelser vitterligt er uafhængige. Vi må stille os tilfredse med, at vi ikke kan finde grunde til, at de skulle være afhængige. Der bør i praksis altid være en nagende tvivl, men med erfaring følger også en bedre forståelse af hvad uafhængighed på forskellige niveauer betyder. Så når vi spørger "hvorfor ikke antage uafhængighed?", skal vi selvfølgelig tænke os om og analysere situationen, men derefter må vi acceptere nogle nødvendige uafhængighedsantagelser og komme videre med vores analyser.

Modellen for data givet ved tætheden $f_{(\alpha,\sigma^2)}^{\otimes N}$ beskriver målingerne som uafhængige og marginalt identisk normalfordelte. Det er en *parametrisk model*, hvor fordelingerne i modellen er parametriserede af middelværdiparameteren α og variansparameteren σ^2. Hvis data er indsamlet med det formål at estimere middelreaktionstid, vil vi sige at α er interesseparameteren. Tilstedeværelsen af σ^2 er nødvendig for en fuld modelspecifikation men er teknisk set en plage, og den kaldes derfor en nuisanceparameter.

I et andet forsøg laver vi måske målingerne N gange før og N gange efter at forsøgspersonen har indtaget alkohol. Modellen med tæthed

$$f_{(\alpha_1,\sigma^2)}^{\otimes N} \otimes f_{(\alpha_2,\sigma^2)}^{\otimes N}$$

beskriver de $2N$ målinger som uafhængige og marginalt normalfordelte, alle med samme varians men med forskellige middelværdier før og efter indtagelse af alkohol. Interesseparameteren er *forskellen*

$$\alpha_2 - \alpha_1,$$

som angiver hvor meget langsommere personen bliver efter indtagelse af alkohol. Modellen er naturligt parametriseret i termer af $(\alpha_1, \alpha_2, \sigma^2)$, men den kan reparametriseres i termer af

$$(\alpha_2 - \alpha_1, \alpha_1, \sigma^2),$$

hvor (α_1, σ^2) kan opfattes som en to-dimensional nuisanceparameter. Generelt vil interesseparameteren være en funktion af den fulde parameter, og "resten" kan opfattes som nuisanceparametre. Hvad interesseparameteren er afhænger ikke af modellen men af hvad vi vil med data.

Den parametriserede model giver os en fuldstændig sandsynlighedsteoretisk karakterisering af mulige datagenererende mekanismer. For givne parametre kan vi simulere nye data lige så tosset vi vil. Den parametriserede model gør det samtidigt muligt at udtrykke et svar på vores grundlæggende spørgsmål i termer af parametrene gennem en parameterfunktion, som vi kalder vores interesseparameter. Parametrene er godt nok ukendte, men den parametriserede model danner en ramme for at konvertere data til udsagn om parametrene (estimater) og dermed ultimativt at svare på vores spørgsmål. Modellen er et værktøj; en løftestang som i kraft af sine beskrivende egenskaber giver os en mulighed for at analysere, fortolke og generalisere fra data.

1.2 Statistik

Faget statistik opstod på basis af et praktisk behov for indsamling og opsummering af befolkningsdata såvel som et teoretisk behov for forståelsen af tilfældige fænomener, som f.eks. målefejl eller udfaldet af et terningkast.

Gennem 17- og 1800-tallet udviklede staten en behov for et mere indgående kendskab til beskrivende data om befolkningen, og betegnelsen *stat-istik* har sin oprindelse herfra. Indsamling af data ved folketællinger og den efterfølgende behandling

og tabulering var oprindeligt en manuel og arbejdskrævende opgave, og det katalyserede i slutningen af 1800-tallet udviklingen af mekaniske procedurer. Dermed var statistik helt fra starten med til at drive computerteknologien fremad. Herman Hollerith introducerede f.eks. i 1889 en revolutionerende maskine baseret på hulkort for US Census Bureau, og han lagde hermed grunden til virksomheden IBM. I Danmark varetages opgaven med indsamling og håndtering af befolkningsdata primært af Danmarks Statistik, og mange vil nok forbinde ordet statistik med denne klassiske form for dataindsamling og opgørelse.

Udviklingen af den teoretiske statistik var knyttet til sandsynlighedsteorien, som bl.a. tog udgangspunkt i hasardspil, altså spil hvor udfaldet primært er styret af tilfældigheder. Gennem 1800-tallet udvikledes en matematisk forståelse af målefejl, og omkring starten af 1900-tallet blev grundlaget for den moderne matematiske statistik lagt.

Sandsynlighedsteori danner i dag i udpræget grad den matematiske ramme for faget statistik og derigennem for den praktiske forståelse og fortolkning af data. Anvendelserne er mange og statistik er det metodiske grundlag for behandling af en række videnskabelige problemstillinger såsom: målefejl og variation i designede eksperimenter; usikkerhed forbundet med stikprøveudtagelse; forventninger og fremskrivninger baseret på historiske data; og risici forbundet med handlinger. Men anvendelserne strækker sig også langt ud over det akademiske, og statistik indgår centralt i beslutningsprocesserne indenfor f.eks. sundhedsektoren og de offentlige som de private dele af den finansielle sektor. Statistik og machine learning er kommet til at spille en afgørende rolle – på godt og ondt – i den moderne medie- og reklameindustri, og spiller en fortsat større og større rolle i det meste af produktionsindustrien. Mantraet er, at beslutninger skal være drevet af data. Spørgsmålet er, hvordan data skal drive beslutningerne?

1.2.1 Induktion: Den bayesianske og frekventistiske tilgang

I statistik beskæftiger vi os med den grundlæggende videnskabelige problemstilling *induktion*. Dvs. muligheden for at uddrage generel viden fra specifikke eksempler. Kvantitative modeller og fortolkningsrammer, der tillader induktion, er centrale i statistik, og faget er mere optaget af praktiske metoder og en forståelse af deres egenskaber end af den filosofiske mulighed for (eller umulighed af) induktion. Men der er alligevel nogle grundlæggende uenigheder indenfor faget om hvordan induktion kan og bør foretages i praksis. Den største uenighed berører selve fortolkningen af sandsynligheder, og de implikationer forskellige fortolkninger har for, hvordan vi udvikler

statistisk metode til analyse af data. De to hovedfortolkninger er den subjektivistisk bayesianske fortolkning og den frekventistiske fortolkning.

I den subjektivistisk bayesianske fortolkning udtrykker sandsynligheder en intern og personlig forventning. Givet den information man som individ har til rådighed, hvad er så ens subjektive forventning til udfaldet? Usikkerhed er i den sammenhæng et udtryk for en individuel usikkerhed begrundet med subjektets begrænsninger med den givne information. I den frekventistiske fortolkning udtrykker sandsynligheder en forventning til udfaldets værdier ved (hypotetiske) gentagelser. Usikkerhed er så en iboende egenskab ved udfaldet og ikke udtryk for en eventuel iagtagers begrænsninger.

Der findes en række variationer over den bayesianske såvel som den frekventistiske fortolkning af sandsynligheder, og de forskellige fortolkninger har været debateret intensivt i flere hundrede år. En pragmatisk holdning er i dag udbredt blandt mange statistikere, som accepterer at flere fortolkninger kan sameksistere. Hvis vi udtaler os om en helt specifik begivenhed, som f.eks. sandsynligheden for at Danmark kvalificerer sig til det næste VM i fodbold, er den bayesianske fortolkning næsten uundgåelig. Man kan selvfølgelig kvalificere et bud på sådan en sandsynlighed med historisk og faktuel viden, men det er ekstremt hypotetisk at tænke sig til et stort antal gentagelser af kvalifikationskampene og en frekventistisk fortolkning af sandsynligheden for dansk kvalifikation. Men for simple mekanismer, som mønt- og terningekast og deraf afledte hændelser såsom data indsamlet ved en tilfældig stikprøve eller fra et randomiseret eksperiment, er det nemt at forestille sig gentagelser og en frekventistisk fortolkning af sandsynlighederne. Trods udbredt pragmatisme støder de to hovedskoler for fortolkningen af sandsynligheder dog stadig sammen i formaliseringen af, hvordan vi laver statistik. Altså, hvordan vi analyserer data.

Såvel en bayesiansk som en frekventistisk model beskriver mulige sandsynlighedsfordelinger for et udfald X, og det er faktisk ikke så meget fortolkningen af disse sandsynligheder, som adskiller bayesiansk og frekventistisk statistik. Forskellen består i, hvordan vi beskriver og udtrykker vores manglende viden om parametrene i modellen. For normalfordelingsmodellen kender vi ikke α og σ^2, men vi ønsker at sige noget om α og σ^2 baseret på data. I den bayesianske tilgang vil man udtrykke den manglende viden gennem sandsynligheder på præcis den samme måde, som man udtrykker manglende viden om udfaldet X. Derved kan man i princippet udregne en såkaldt *a posteriori* fordeling af parametrene givet observationen $X = x$ og på den måde udtrykke en forventning til parametrene baseret på data. I den frekventistiske tilgang er der derimod en asymmetri i den måde som udfaldet, X, og parametrene,

(α, σ^2), optræder på, og der findes flere metoder til at konvertere en observation af $X = x$ til et udsagn om de ukendte parametre. Forskellige frekventistiske argumenter kan så benyttes for at understøtte de forskellige metoder.

Den frekventistiske tilgang tilbyder ingen universel procedure til at konvertere observationen $X = x$ til et udsagn om parametrene. Det er veldokumenteret at denne manglende principfasthed for frekventistiske procedurer kan lede til inkonsistente konklusioner. Hvor den bayesianske statistik har et klart *induktionsprincip*, så mangler den frekventistiske tilgang sådan et princip. Den rent bayesianske tilgang mangler på den anden side de frekventistiske værktøjer til at analysere den bayesianske procedure. Den pragmatiske statistiker vil gerne se den analyse sammen med en tilsvarende analyse af alternative frekventistiske metoder.

Et ofte fremført argument mod bayesiansk statistik i en videnskabelig sammenhæng er, at det er en subjektiv metode. Heroverfor skulle den frekventistiske tilgang være objektiv. Det argument tillægger ordene "objektiv" og "subjektiv" en noget vildledende og værdiladet betydning. Tillægsordet "subjektiv" forklarer, fra hvilket perspektiv sandsynlighedsudsagn skal fortolkes. Det betyder ikke automatisk, at den statistiske analyse er forudindtaget eller skævvredet grundet personlige synspunkter eller interesser. Begge tilgange til statistik indeholder en række metodiske valg, som er delvist subjektive, og hvad der er langt vigtigere for objektiviteten og redeligheden er gennemskuelighed i metodevalg. Meget ofte viser det sig, at en metode har en dualistisk fortolkning som både en bayesiansk og en frekventistisk procedure, så *hvad* vi har gjort er betydeligt vigtigere end vores filosofiske udgangspunkt. Men det er selvfølgelig stadig vigtigt, at vi har en korrekt forståelse af fortolkningerne, og at vi udtrykker os sprogligt korrekt og præcist givet den fortolkning, som vi nu engang benytter.

En grundlæggende – og måske den væsentligste – forskel på en frekventistisk og en bayesiansk dataanalyse er, at den frekventistiske analyse tillader det hypotetiske spørgsmål:

Hvilke observationer kunne vi have fået i stedet for dem vi fik?

Et kompromisløst bayesiansk standpunkt er, at svaret på dette frekventistiske spørgmål ikke bidrager med relevant indsigt. Derfor er der ikke behov for at svare på spørgsmålet i en rent bayesiansk analyse. Udgangspunktet i denne bog er, at spørgsmålet er relevant, og at svaret både giver en nyttig kvantificering af usikkerhed og en mulighed for at undersøge om modellen er i overensstemmelse med data.

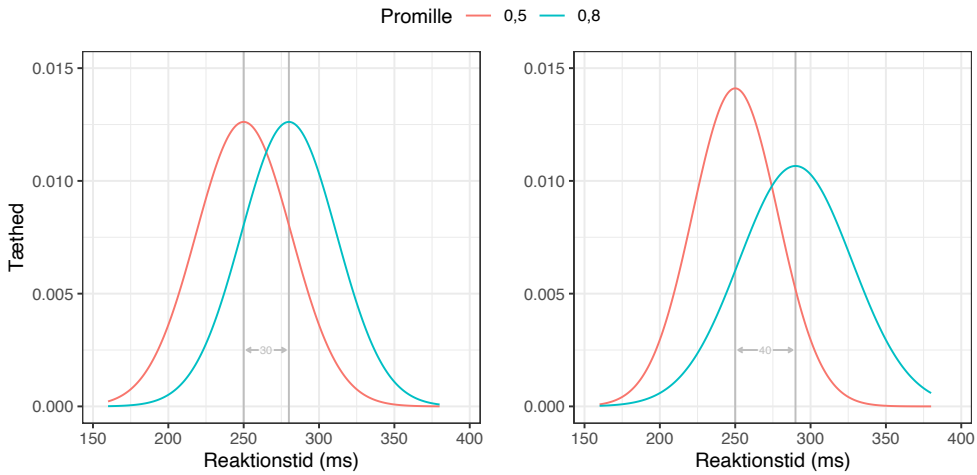
Som nævnt ovenfor findes der ikke et universelt, frekventistisk induktionsprincip, men den likelihoodbaserede metode, specielt via maksimaliseringsestimation, er et udbredt anvendt induktionsprincip for parametriserede modeller. Og denne bogs hovedfokus er da også på en frekventistisk behandling af likelihoodbaserede statistiske metoder for normalfordelingsmodeller. Behandlingen leder til en række procedurer til analyse af data. Disse procedurer er velkendte og udbredte analyseværktøjer i anvendt statistik, og bør derfor også være velkendte for enhver statistiker.

Hensigten med bogen er at skabe en solid forståelse af normalfordelingsmodellerne. Dels af de forudsætninger, som modellerne gør sig om data og de mekanismer, som har frembragt data, og dels af de værktøjer, som vi benytter i analysen af modellerne, og som i sidste ende leder til praktiske procedurer til dataanalyse. Vi vil fokusere på det matematiske grundlag for at forstå disse procedurer, mens en filosofisk eller matematisk retfærdiggørelse spiller en mindre rolle. Bogen bør derfor læses sammen med anden litteratur, der sætter metoderne i bogen i perspektiv, og som belyser alternative metoder, se f.eks. Lauritzen (2021), Hansen (2012), MacKay (2003), Gelman et al. (2014), Efron and Hastie (2016).

1.2.2 Inferens

Hvad mener vi så helt konkret med "dataanalyse", "induktion" og "generalisering"? Overordnet kan vi mene to lidt forskellige ting, som vi kalder henholdsvis *prædiktion* og *inferens*. Med prædiktion mener vi en forudsigelse af X, før udfaldet x observeres, f.eks. forudsigelse af en reaktionstid før målingen foretages. Med inferens mener vi en fuldstændig eller delvis karakterisering af fordelingen af X. Der er ingen modstrid mellem prædiktion og inferens; en fuldstændig karakterisering kan f.eks. bruges til prædiktion. Man kan imidlertid lave prædiktion uden at have en fuldstændig karakterisering, og man kan have en delvis karakterisering, som ikke kan bruges til at prædiktere med. Forskellen mellem prædiktion og inferens skal findes i de konkrete anvendelser, og de spørgsmål som vi der ønsker svar på.

Det virker ikke specielt relevant at kunne prædiktere reaktionstid som et mål i sig selv, men vi kunne spørge, hvordan reaktionstid afhænger af alkoholindtagelse. Promillegrænsen i Danmark blev i 1998 sænket fra 0,8 til 0,5 – begrundet med at personer med en promille mellem 0,5 og 0,8 var overrepræsenterede i uheldsstatistikkerne sammenlignet med personer, som ikke havde indtaget alkohol. Bliver din reaktionstid så langsommere, hvis du har en promille på 0,8 i forhold til en promille på 0,5?



Figur 1.3: Eksempler på modeller for en persons reaktionstid ved en promille på 0,5 såvel som 0,8. I den ene model (venstre) er forskellen i middelreaktionstid 30 ms mens spredningen er den samme (31,6 ms) uanset promille. I den anden model (højre) er forskellen i middelreaktionstid 40 ms og spredningen er 28,3 ms og 37,4 ms ved en promille på hhv. 0,5 og 0,8.

Og hvis det er tilfældet, hvor stor er forskellen? Hvis vi kender din reaktionstidsfordeling, når du har en promille på 0,8 hhv. 0,5, kan vi svare på spørgsmålene ved at sammenligne de to fordelinger. I praksis må vi lave et eksperiment og på basis af data *drage inferens* om forskellen, altså udlede en generel konklusion på basis af, hvad data fra eksperimentet viser.

Figur 1.3 viser forskellige normalfordelingsmodeller, som kan bruges til at sammenligne reaktionstider ved to forskellige alkoholpromiller. I den ene model er der tre parametre,

$$(\alpha_1, \alpha_2, \sigma^2) = (250, 280, 1000).$$

Middelværdiparameteren afhænger af alkoholpromille men variansen er uanset promillen $\sigma^2 = 1000$ (spredningen er altså $\sigma = \sqrt{1000} \simeq 31,6$). Interesseparameteren er forskellen

$$\alpha_2 - \alpha_1 = 280 - 250 = 30.$$

Da variansen er konstant er fordelingen ved en promille på 0,8 altså en translation på 30 ms af fordelingen ved en promille på 0,5.

Den anden model har fire parametre,

$$(\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2) = (250, 290, 800, 1400),$$

hvor variansen nu også afhænger af promillen. Vi kunne stadig fokusere på forskellen i middelværdi som interesseparameter, og den er 40 ms, men fortolkningen er blevet mere kompliceret. Variansen går også op ved den højere promille.

Uanset om variansen er konstant eller afhænger er promillen er det muligt at drage inferens om forskellen i middelværdi, $\alpha_2 - \alpha_1$, på basis af modellen. Det viser sig at være teknisk lidt vanskeligere at forstå præcist, hvilke observationer vi ellers kunne have fået, hvis variansen ikke er konstant. Og dermed bliver det lidt vanskeligere at kvantificere usikkerheden på et estimat af forskellen. Men hvad der er langt vigtigere er, at fortolkningen og kommunikationen bliver mere mudret, hvis variansen ikke er konstant. I eksemplet ovenfor med konstant varians kan vi med rimelighed sige at reaktionstiden er 12% større ved en promille på 0,8 end ved en promille på 0,5. I eksemplet med forskellig varians er middelreaktionstiden 16% større ved en promille på 0,8, men en fuld beskrivelse af forskellen kræver, at vi også fortæller hvorledes variationen i reaktionstid er større ved den høje promille.

Når variansen er konstant kan vi udtrykke forskelle i normalfordelingsmodeller alene i termer af middelværdier. Generaliserbar inferens på basis af data kan dermed udtrykkes på en let fortolkelig måde. Af den grund er fokus i bogen på modeller med konstant varians. Vi kan ikke forvente at konstant varians er en naturlov, men hvis modeller med konstant varians ikke fitter data, kan vi ofte finde en variansstabiliserende transformation af data. Og så vil vi foretrække at modellere og analysere data på den transformerede skala.

Bemærk at det er den fuldstændige karakterisering af fordelingen af data, som tillader os at opdage, at forskellen i middelværdi ikke giver et komplet billede. Vi kunne vælge at fokusere snævert på forskellen i middelværdi som vores interesseparameter, og vi kunne så drage inferens fra data om den parameter. Men uden en modelramme vil vi ikke opdage, at en 16% forøgelsen i reaktionstiden kun beskriver forskellen delvist. Normalfordelingsmodeller med konstant varians er en ideel modelramme. *Hvis* sådan en model fitter data kan vi på en meget klar og præcis måde studere forskelle via interesseparametre udtrykt i termer af middelværdier.

1.2.3 Prædiktion eller inferens

Selvom prædiktion af reaktionstid ikke er så interessant, er der mange andre eksempler, hvor prædiktion er det væsentlige. Dit email spamfilter prædikterer om en email er spam eller ej. Det er nyttigt, og det betyder ikke så meget, om du forstår og kan fortolke det, bare det virker! Prædiktion er særligt interessant, når målet er at automatisere processer og beslutninger sådan som spamfilteret automatiserer sorteringen af dine emails. Vi har ligeledes brug for prædiktion – og det ligger nærmest i ordet – når vi skal lave fremskrivninger/prognoser, f.eks. demografiske fremskrivninger af befolkningssammensætningen de næste 50 år, eller fremskrivninger af den globale temperatur og øvrige klimaudvikling. Hvis du bor tæt ved vandet og skal bygge et dige, er du interesseret i at prædiktere den fremtidige vandstand, og dit forsikringsselskab er interesseret i at prædiktere sandsynligheden for, at du alligevel får oversvømmelse, og hvad det så vil koste dem. Og hvis du af lægen får stillet en alvorlig kræftdiagnose, ønsker du nok også en prognose af den resterende levetid.

Når prædiktion bruges til automatisering bruger vi i dag ofte betegnelsen kunstig intelligens. Det er vigtigt at automatiserede processer er meget pålidelige – ellers vil vi i hvert fald ikke opfatte dem som særligt intelligente. Spamfilteret skal gerne ramme korrekt i langt de fleste tilfælde, og når SKAT får sine ejendomsvurderinger til at virke, skal de gerne ramme tæt på faktiske salgspriser. Når vi automatiserer har vi brug for *punktprædiktioner*, dvs. en forudsigelse af værdien af en bestemt stokastisk variabel givet den information, vi nu engang har til rådighed. Stor præcision betyder at vi har reduceret den uforklarlige variation til næsten ingenting og usikkerheden er stort set forsvundet. En sådan prædiktiv model generaliserer, hvis den kan opretholde sin høje præcision igen og igen og igen.

I andre sammenhænge kan vi glemme alt om høj præcision af punktprædiktioner. Vi kan ikke forudsige en bestemt reaktionstid særligt præcist, og viden om hvorvidt personen har en promille på 0,8 eller 0,5 hjælper kun lidt. Vores interesseparameter – forskellen i reaktionstid – betyder ikke så meget for den enkelte reaktionstid, men på basis af data kan vi drage inferens om, hvordan fordelingen af reaktionstider afhænger af alkoholpromille. Den form for induktion generaliserer, hvis vi vil genfinde de samme forskelle i *fordelingerne* af reaktionstider igen og igen og igen.

Vi kan heller ikke præcist prædiktere, hvad den maksimale vandstand vil være i 2030, men 99%-fraktilen for fordelingen af den årlige maksimale vandstand kunne være en relevant interesseparameter. Hvis vi vil bygge vores dige netop så højt, så har vi brug

for at drage inferens om den interesseparameter. Hvis den årlige maksimale vandstand havde været normalfordelt $\mathcal{N}(\alpha, \sigma^2)$, så kunne vi have udtrykt interesseparameteren som

$$\alpha + 2,33\sigma,$$

men normalfordelingen er sikkert en rædsom model for maksimal vandstand. En ekstremværdifordeling, som kan være skæv og have tunge haler, er formodentlig en meget bedre model.

Der er en flydende overgang fra situationer, hvor vi har brug for rå punktprædiktioner af høj præcision, til situationer, hvor vi har brug for at identificere en specifik interesseparameter, som vi ønsker at drage inferens om. Og da de samme modeller ofte bruges i begge tilfælde kan det være forvirrende, hvad formålet egentlig er. I de følgende kapitler vil fokus være på at drage inferens om en interesseparameter, eller i hvert fald at fitte statistiske modeller til data med det formål. At den resulterende statistiske model kan bruges til at lave punktprædiktioner er i den sammenhæng en sidegevinst.

1.3 Likelihood

Når vi i praksis skal drage inferens har vi brug for en metode til at konvertere en observation til et udsagn om parametrene. Og her er den likelihoodbaserede metode både udbredt i praksis og teoretisk velfunderet. I dette afsnit skal vi se, hvordan likelihoodfunktionen for en normalfordelingsmodel både kan bruges til at udlede estimatorer for parametrene og undersøge om en model fitter data.

I det følgende betegner X en reel stokastisk variabel med fordeling μ_0, som *ikke* nødvendigvis er en normalfordeling. For normalfordelingsmodellen $\mathcal{N}(\alpha, \sigma^2)$ defineres log-likelihoodfunktionen som logaritmen til tætheden evalueret i X, dvs.

$$\begin{aligned}\ell_X(\alpha, \sigma^2) &= \log\left(f_{\alpha,\sigma^2}(X)\right) \\ &= -\frac{1}{2\sigma^2}(X-\alpha)^2 - \frac{1}{2}\log(2\pi\sigma^2).\end{aligned}$$

Notationen indikerer, at log-likelihoodfunktionen skal opfattes som en funktion af parametrene

$$(\alpha, \sigma^2) \mapsto \ell_X(\alpha, \sigma^2).$$

Men den kan også opfattes som en stokastisk variabel (en transformation af X)

$$X \mapsto \ell_X(\alpha, \sigma^2)$$

for faste værdier af parametrene. Hvis vi vil understrege opfattelsen af log-likelihoodfunktionen som en funktion af parametrene for en given observation $X = x$, bruger vi notationen

$$\ell_x(\alpha, \sigma^2) = -\frac{1}{2\sigma^2}(x - \alpha)^2 - \frac{1}{2}\log(2\pi\sigma^2).$$

Vi kan tænke på log-likelihoodfunktionen som et mål for, hvor godt modellen $\mathcal{N}(\alpha, \sigma^2)$ stemmer overens med en observation af X. I det følgende afsnit om krydsentropi understøttes den fortolkning af Gibbs' ulighed.

1.3.1 Krydsentropi

Definition 1.3.1 (Krydsentropi)**.** Lad $(\mathcal{X}, \mathbb{F}, \nu)$ være et målrum, lad μ_0 være et sandsynlighedsmål på $(\mathcal{X}, \mathbb{F})$, og lad f være en sandsynlighedstæthed mht. ν. Krydsentropien af fordelingen $\mu = f \cdot \nu$ relativt til μ_0 er

$$H(\mu_0, f) = -\int \log f(x) \mathrm{d}\mu_0(x), \tag{1.3}$$

givet at $(\log f(x))^+ = \max\{\log f(x), 0\}$ er integrabel mht. μ_0.

Vi tillader at krydsentropien antager værdien $+\infty$, mens integrabilitetsantagelsen på positivdelen $(\log f(x))^+$ dels sikrer at krydsentropien er veldefineret, og dels sikrer at krydsentropien ikke kan antage værdien $-\infty$. Dvs.

$$H(\mu_0, f) \in (-\infty, \infty].$$

Hvis $\mu_0 = f_0 \cdot \nu$ vil vi også bruge betegnelsen $H(f_0, f) = H(\mu_0, f)$, og vi ser at

$$H(f_0, f) = -\int f_0(x) \log f(x) \mathrm{d}\nu(x).$$

Vi bemærker endvidere, at hvis X har fordeling μ_0 er

$$H(\mu_0, f) = -E(\log f(X)).$$

For μ_0 en fordeling på $\mathbb{R}$ og f tætheden for normalfordelingen, f_{α,σ^2}, mht. lebesguemålet har vi altså

$$H(\mu_0, f_{\alpha,\sigma^2}) = -E(\ell_X(\alpha, \sigma^2)),$$

eller i ord; krydsentropien af normalfordelingen med middelværdi α og varians σ^2 relativt til μ_0 er den forventede negative log-likelihood. Hvis vi vil præcisere at tætheden er mht. lebesguemålet kan vi bruge betegnelsen *differential(kryds)entropi*, jf. også opgave 1.5 for en behandling af, hvorledes krydsentropien afhænger af ν.

Vi viser nu Gibbs' ulighed for $\mu_0 = f_0 \cdot \nu$, som fortæller os at $H(f_0) = H(f_0, f_0)$ er en nedre grænse på krydsentropien $H(f_0, f)$. Størrelsen $H(f_0)$ kaldes også entropien af μ_0, og Gibbs' ulighed er en helt central ulighed i statistik og informationsteori. Vi vil nedenfor benytte uligheden som et springbræt til at introducere maksimaliseringsestimatoren.

Sætning 1.3.2. *Hvis $H(f_0) < \infty$ gælder Gibbs' ulighed*

$$H(f_0, f) \geq H(f_0)$$

med lighedstegn hvis og kun hvis $f = f_0$ μ_0-n.o.

Bevis. Sæt $A = \{x \in \mathcal{X} \mid f_0(x) > 0\}$ og observer at $\mu_0(A) = 1$. Bemærk endvidere, at funktionen $z \mapsto -\log(z)$ er konveks. Vi kan derfor bruge Jensens ulighed i fjerde skridt nedenfor, og vi finder deraf Gibbs' ulighed:

$$\begin{aligned}
H(f_0, f) &= -\int_A \log f(x) \mathrm{d}\mu_0(x) \\
&= -\int_A \log \frac{f(x)}{f_0(x)} \mathrm{d}\mu_0(x) - \int_A \log f_0(x) \mathrm{d}\mu_0(x) \\
&= -\int_A \log \frac{f(x)}{f_0(x)} \mathrm{d}\mu_0(x) + H(f_0) \\
&\geq -\log\Big(\int_A \frac{f(x)}{f_0(x)} f_0(x) \mathrm{d}\nu(x) \Big) + H(f_0) \\
&\geq -\log\Big(\underbrace{\int f(x) \mathrm{d}\nu(x)}_{=1} \Big) + H(f_0) \\
&= H(f_0).
\end{aligned}$$

Vi ved endvidere at der er lighedstegn i Jensens ulighed netop hvis $f(x)/f_0(x) = 1$ for μ_0-n.a. x, og deraf følger at der er lighedstegn i Gibbs' ulighed netop hvis $f = f_0$ μ_0-n.o. □

Hvis $X \sim \mathcal{N}(\alpha_0, \sigma_0^2)$ er krydsentropien

$$H(f_{\alpha_0,\sigma_0^2}, f_{\alpha,\sigma^2}) = \frac{1}{2\sigma^2} E\left((X-\alpha)^2\right) + \frac{1}{2}\log(2\pi\sigma^2),$$

som altså minimeres over (α, σ^2) i (α_0, σ_0^2). Dvs. den forventede log-likelihood maksimeres i (α_0, σ_0^2). Det kunne vi indse ved direkte udregning, f.eks. via lemma 1.3.4 nedenfor, men det følger altså også af den generelle sætning 1.3.2. Minimum for krydsentropien er (differential)entropien for normalfordelingen,

$$H(f_{\alpha_0,\sigma_0^2}) = \frac{1}{2\sigma_0^2} E\left((X-\alpha_0)^2\right) + \frac{1}{2}\log(2\pi\sigma_0^2) = \frac{1}{2}\log(2\pi e\sigma_0^2), \tag{1.4}$$

som kun afhænger af variansen σ_0^2.

Selv hvis $X \sim \mu_0$ ikke er normalfordelt, kan vi stadig fortolke krydsentropien som et mål for hvor langt $\mathcal{N}(\alpha, \sigma^2)$ er fra μ_0, og

$$(\alpha_0, \sigma_0^2) = \arg\max_{\alpha,\sigma^2} E(\ell_X(\alpha, \sigma^2)),$$

der maksimerer den forventede log-likelihood, kan derfor fortolkes som de parametre, der giver den bedste normalfordelingsapproksimation af μ_0. Ikke overraskende er $\alpha_0 = E(X)$ og $\sigma_0^2 = V(X)$, jf. opgave 1.3.

1.3.2 Maksimaliseringsestimatoren

Minimering af krydsentropien giver os en metode til at finde de parametre, der giver en fordeling, der bedst stemmer overens med fordelingen af X. Det springende punkt er selvfølgelig, at vi i praksis *ikke* kender fordelingen, μ_0, af X. Derimod har vi data, $x_1, \ldots, x_N$, fra denne fordeling, og vi kan opfatte den empiriske fordeling

$$\hat{\mu}_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$$

som en approksimation af μ_0. Ved at indsætte $\hat{\mu}_0$ i krydsentropien får vi

$$H(\hat{\mu}_0, f_{\alpha,\sigma^2}) = -\frac{1}{N}\sum_{i=1}^{N} \log f_{\alpha,\sigma^2}(x_i) = \frac{1}{2N\sigma^2}\sum_{i=1}^{N}(x_i-\alpha)^2 + \frac{1}{2}\log(2\pi\sigma^2). \tag{1.5}$$

Den empiriske krydsentropi givet ved (1.5) kan nu minimeres for at finde den normalfordeling, der bedst passer med den empiriske fordeling $\hat{\mu}_0$ og dermed med data. Men inden vi forfølger den ide, vil vi se en anden vej til højreside af (1.5).

Hvis $X = (X_1, \ldots, X_N)^T$ betegner sammenbundtningen af de N stokastiske variable, og $x = (x_1, \ldots, x_N)^T$ ligeledes betegner den N-dimensionale vektor af observationer, så er produktmålet med tæthed $f_{(\alpha,\sigma^2)}^{\otimes N}$ en model for X, hvor $X_1, \ldots, X_N$ altså antages uafhængige og identisk $\mathcal{N}(\alpha, \sigma^2)$-fordelte. Med den model er log-likelihoodfunktionen for observationen $X = x$

$$\begin{aligned}\ell_x(\alpha, \sigma^2) &= \log\left(\prod_{i=1}^{N} f_{\alpha,\sigma^2}(x_i)\right) \\ &= \sum_{i=1}^{N} \log f_{\alpha,\sigma^2}(x_i) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \alpha)^2 - \frac{N}{2}\log(2\pi\sigma^2).\end{aligned}$$

Vi ser altså, at

$$H(\hat{\mu}_0, f_{\alpha,\sigma^2}) = -\frac{1}{N}\ell_x(\alpha, \sigma^2), \qquad x = (x_1, \ldots, x_N)^T$$

dvs. den empiriske krydsentropi er op til konstanten $-1/N$ lig med log-likelihoodfunktionen for hele datasættet under en antagelse om uafhængighed. Minimering af den empiriske krydsentropi er derfor ækvivalent med maksimering af log-likelihoodfunktionen.

Definition 1.3.3. Maksimaliseringsestimatoren (MLE) er den værdi af parametrene, der maksimerer ℓ_x.

Maksimaliseringsestimatoren giver et princip for, hvordan vi konverterer en observation til et estimat for de ukendte parametre:

$$x \mapsto (\hat{\alpha}, \hat{\sigma}^2) = \arg\max_{\alpha,\sigma^2} \ell_x(\alpha, \sigma^2).$$

Estimatoren kan ligeledes findes ved at minimere $-\frac{1}{N}\ell_x(\alpha, \sigma^2)$, og teoretisk er der selvfølgelig ingen forskel på, hvorvidt vi minimerer den empiriske krydsentropi eller

maksimerer log-likelihoodfunktionen. Vi kunne for den sags skyld også maksimere likelihoodfunktionen $L_x = \exp(\ell_x)$. For normalfordelingsmodellen er det funktionen

$$L_x(\alpha, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^N (x_i-\alpha)^2},$$

som blot er tætheden opfattet som en funktion af parametrene for en given observation. I forbindelse med optimering af log-likelihoodfunktionen for normalfordelinger er følgende lemma nyttigt.

Lemma 1.3.4. *Lad $a, b > 0$. Funktionen $f : (0, \infty) \to \mathbb{R}$ givet ved*

$$f(y) = -\frac{a}{y} - b \log y$$

antager et entydigt globalt maksimum i $y = a/b$.

Bevis. Den afledte af f er

$$f'(y) = \frac{a}{y^2} - \frac{b}{y} = \frac{a - by}{y^2},$$

som er nul netop hvis $y = a/b$, positiv for $y \in (0, a/b)$ og negativ for $y \in (a/b, \infty)$. Dvs. f antager et globalt maksimum i a/b. □

Sætning 1.3.5. *Såfremt to af observationerne er forskellige er den entydige maksimaliseringsestimator (MLE) for (α, σ^2) i normalfordelingsmodellen givet ved*

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^N x_i$$
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \hat{\alpha})^2.$$

Værdien af log-likelihoodfunktionen i MLE er

$$\ell_x(\hat{\alpha}, \hat{\sigma}^2) = -\frac{N}{2} - \frac{N}{2}\log(2\pi\hat{\sigma}^2).$$

Bevis. Observer at

$$(x_i - \alpha)^2 = (x_i - \hat{\alpha})^2 + 2(x_i - \hat{\alpha})(\hat{\alpha} - \alpha) + (\hat{\alpha} - \alpha)^2,$$

så

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\alpha)^2 &= \sum_{i=1}^{N}(x_i-\hat{\alpha})^2 + \sum_{i=1}^{N}2(x_i-\hat{\alpha})(\hat{\alpha}-\alpha) + N(\hat{\alpha}-\alpha)^2 \\ &= \sum_{i=1}^{N}(x_i-\hat{\alpha})^2 + 2(\hat{\alpha}-\alpha)\underbrace{\left(\sum_{i=1}^{N}x_i - N\hat{\alpha}\right)}_{=0} + N(\hat{\alpha}-\alpha)^2 \\ &= \sum_{i=1}^{N}(x_i-\hat{\alpha})^2 + N(\hat{\alpha}-\alpha)^2.\end{aligned}$$

Vi omskriver log-likelihoodfunktionen

$$\begin{aligned}\ell_x(\alpha,\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\alpha)^2 - \frac{N}{2}\log(2\pi\sigma^2) \\ &= \underbrace{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\hat{\alpha})^2 - \frac{N}{2}\log(2\pi\sigma^2)\right)}_{=\ell_x(\hat{\alpha},\sigma^2)} - \frac{1}{2\sigma^2}N(\hat{\alpha}-\alpha)^2 \\ &\leq \ell_x(\hat{\alpha},\sigma^2).\end{aligned}$$

Uligheden er skarp medmindre $\alpha = \hat{\alpha}$. Såfremt to observationer er forskellige, er

$$a = \frac{1}{2}\sum_{i=1}^{N}(x_i-\hat{\alpha})^2 > 0,$$

og med $b = N/2 > 0$ giver lemma 1.3.4 at $\ell_x(\hat{\alpha},\sigma^2)$ antager sit entydige maksimum i

$$\hat{\sigma}^2 = \frac{a}{b} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\alpha})^2.$$

Dvs.

$$\ell_x(\alpha,\sigma^2) \leq \ell_x(\hat{\alpha},\hat{\sigma}^2) = -\frac{N}{2} - \frac{N}{2}\log(2\pi\hat{\sigma}^2)$$

med lighedstegn hvis og kun hvis $(\alpha,\sigma^2) = (\hat{\alpha},\hat{\sigma}^2)$. □

Sætning 1.3.5 er på sin vis skuffende banal. I normalfordelingsmodellen er MLE for α og σ^2 simpelt hen empirisk middelværdi og varians. For reaktionstiderne er MLE

altså $(\hat{\alpha}, \hat{\sigma}^2) = (273, 982)$, jf. (1.1) og (1.2). Når skuffelsen har lagt sig, kan vi forholde os til, hvad Sætning 1.3.5 faktisk fortæller os; at maksimaliseringsestimatoren giver rigtig god mening for en normalfordelingsmodel parametriseret ved netop middelværdi og varians!

I beviset stødte vi på den partielt optimerede funktion

$$\ell_x(\hat{\alpha}, \sigma^2) = \max_{\alpha} \ell_x(\alpha, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \hat{\alpha})^2 - \frac{N}{2} \log(2\pi\sigma^2),$$

som kaldes *profil log-likelihoodfunktionen* for σ^2. Vi fandt den ved, for fastholdt σ^2, at maksimere over α. Vi kunne også, ved brug af lemma 1.3.4 for fast α, have profileret σ^2 ud og fundet profil log-likelihoodfunktionen

$$\ell_x(\alpha, \hat{\sigma}^2(\alpha)) = \max_{\sigma^2} \ell_x(\alpha, \sigma^2) = -\frac{N}{2} - \frac{N}{2} \log\left(\frac{2\pi}{N} \sum_{i=1}^{N} (x_i - \alpha)^2 \right).$$

Da logaritmen er en monotont voksende funktion, maksimeres denne profil log-likelihoodfunktion over α på samme måde som i beviset ovenfor. Bemærk at når vi profilerer σ^2 ud først, kommer

$$\hat{\sigma}^2(\alpha) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \alpha)^2 \tag{1.6}$$

til at afhænge af α.

Det er lidt et temperamentsspørgsmål, hvorvidt teoretiske udregninger lettest foretages med likelihoodfunktionen eller med log-likelihoodfunktionen, men likelihoodfunktionen er som regel uegnet til numeriske udregninger. I praksis benytter vi derfor altid ℓ_x eller $-\frac{1}{N}\ell_x$ til numeriske udregninger, og det kan endvidere være bekvemt at minimere $-\frac{1}{N}\ell_x$ fremfor at maksimere ℓ_x. Faktoren $-1/N$ stabiliserer f.eks. de numeriske værdier af funktionen, som bliver mindre afhængige af størrelsen af datasættet, og for $N \to \infty$ sikrer den også, at funktionen konvergerer mod $H(\mu_0, f_{\alpha,\sigma^2})$. Derudover slipper vi for nogle minusser for normalfordelingsmodellerne, og optimeringsproblemet følger standarden i optimeringslitteraturen, hvor alle problemer er formuleret som minimeringsproblemer.

På samme måde som vi kan studere profil log-likelihoodfunktionerne, kan vi også

studere de to profilerede krydsentropi-funktioner,

$$H(\hat{\mu}_0, f_{\hat{\alpha},\sigma^2}) = \frac{1}{2N\sigma^2}\sum_{i=1}^{N}(x_i - \hat{\alpha})^2 + \frac{1}{2}\log(2\pi\sigma^2) = \frac{\hat{\sigma}^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

$$H(\hat{\mu}_0, f_{\alpha,\hat{\sigma}^2(\alpha)}) = \frac{1}{2} + \frac{1}{2}\log\left(\frac{2\pi}{N}\sum_{i=1}^{N}(x_i - \alpha)^2\right),$$

som begge er nedad begrænsede af den minimale krydsentropi for normalfordelingsmodellen

$$H(\hat{\mu}_0, f_{\hat{\alpha},\hat{\sigma}^2}) = \frac{1}{2}\log(2\pi e\hat{\sigma}^2).$$

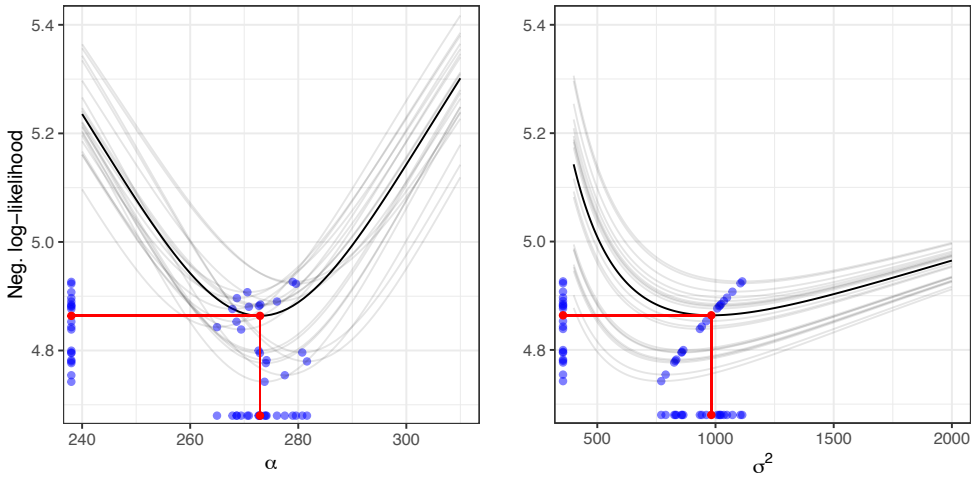
Vi ser, at det netop er entropien for normalfordelingen med varians $\hat{\sigma}^2$, jf. (1.4).

1.3.3 Fordelingen af log-likelihood og MLE

Lad os nu undersøge det hypotetiske spørgsmål: Hvilke observationer kunne vi ellers have fået? Hvis data vitterligt var en realisering af 50 uafhængige $\mathcal{N}(273, 982)$-fordelte variable, hvordan ville andre datasæt så have set ud? Hvordan ville (profil) log-likelihoodfunktionen have set ud? Hvordan ville estimaterne have set ud?

Data er blot *en* blandt mange mulige realiseringer og enhver realisering ville give anledning til sin egen log-likelihood. Det springende punkt er at indse, at en fordeling af data inducerer en fordeling af log-likelihoodfunktioner. Vi stiller det hypotetiske spørgsmål ovenfor for at forstå hvordan den fordeling ser ud. En væsentlig del af bogen går ud på at forstå det matematiske svar på spørgsmålet, men i dette afsnit ser vi på, hvad vi kan få ud af simulationer fra den estimerede model. En teknik som er kendt som *parametrisk bootstrapping*, og som også er nyttig i anvendt statistik.

Da vi har estimeret de ukendte parametre i den parametriske model har vi en fuldstændig karakterisering af fordelingen af data under modellen. Eller rettere, vi har vores bedste bud på en karakterisering – parametrene er jo blot estimater baseret på data. Men lad os lege at $X = (X_1, \ldots, X_{50})$ består af 50 uafhængige og $\mathcal{N}(273, 982)$-fordelte stokastiske variable. Så kan vi simulere et eller flere nye datasæt – hver med 50 nye variable. Figur 1.4 viser negative profil log-likelihoodfunktioner (skaleret med faktoren $1/N$) både for det oprindelige datasæt og for 20 nye datasæt. Som det fremgår er der variation fra datasæt til datasæt i, hvordan log-likelihoodfunktionen ser ud, hvad dens minimum er, og hvor det antages. Fordelingen af såvel minima som af



Figur 1.4: Negative profil log-likelihoodfunktioner skaleret med faktoren $1/N$ (profil krydsentropi) for reaktionstiderne (sort) og 20 simulerede datasæt (grå). De blå punkter angiver minima og maksimaliseringsestimatorer for de 20 simulerede datasæt. De røde punkter/linjer angiver minimum og maksimaliseringsestimator for reaktionstiderne.

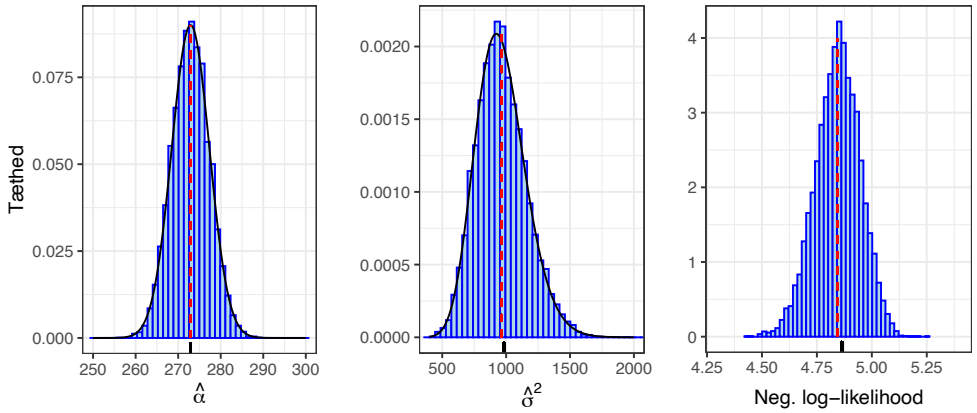
maksimaliseringsestimatorerne er antydet på figuren, og de fordelinger fortæller os noget om usikkerheden af estimatet baseret på det oprindelige datasæt.

Vi kan undersøge fordelingerne lidt nøjere ved at skrue op for antallet af nye datasæt. Figur 1.5 viser histogrammer for fordelingerne baseret på 10.000 datasæt. Histogrammerne giver os et billede af usikkerheden på vores parameterestimater, og de kan bruges til at konstruere konfidensintervaller for parametrene, som udtrykker den usikkerhed. Hvordan man normalt gør det på en præcis og systematisk måde udskyder vi; den intuitive kommunikation af usikkerhed, som histogrammerne er udtryk for, er tilstrækkelig i første omgang.

Hvad der imidlertid er ret interessant er, at det faktisk er muligt at finde de teoretiske fordelinger:

$$\hat{\alpha} \sim \mathcal{N}(273, 19,6)$$
$$\hat{\sigma}^2 \sim 19,6\,\chi^2_{49}$$

Her betegner $19,6\,\chi^2_{49}$ en χ^2-fordeling med 49 frihedsgrader og en skalaparame-



Figur 1.5: Histogrammer for fordelingerne af $\hat{\alpha}$ og $\hat{\sigma}^2$ samt den (skalerede) negative log-likelihood, dvs. $-\ell_X(\hat{\alpha}, \hat{\sigma}^2)/N = \frac{1}{2}\log(2\pi e\hat{\sigma}^2)$, for 10.000 simulerede datasæt. Den stiplede røde linje viser gennemsnittet for de simulerede fordelinger. De estimerede værdier for reaktionstiderne, som simulationerne er baseret på, er angivet (sort mærke) og figuren viser også tæthederne for de teoretiske fordelinger af $\hat{\alpha}$ og $\hat{\sigma}^2$ (sort kurve).

ter på $19,6$. Se opgave 1.1 og opgave 1.2 for detaljerne. Figur 1.5 viser tætheden for disse teoretiske fordelinger, og de er selvfølgelig i god overensstemmelse med histogrammerne. Det er dog værd at bemærke at gennemsnittet for $\hat{\sigma}^2$ er ca. $19,6 \times 49 = 960 < 982$, og det fremgår også af figuren. Maksimaliseringsestimatoren for variansen underestimerer altså i middel variansen en lille smule i forhold til variansen på 982, som er brugt til at simulere data.

Figur 1.5 viser endvidere fordelingen af den (skalerede) negative log-likelihoodfunktion i MLE, $-\ell_X(\hat{\alpha}, \hat{\sigma}^2)/N$, som her er det samme som den estimerede krydsentropi af normalfordelingen relativt til μ_0 og essentielt en logaritmetransformation af $\hat{\sigma}^2$. For reaktionstiderne er denne størrelse $4,86$. Fordelingen fortæller os bl.a. at denne størrelse – som alle andre størrelser afledt af data – faktisk har en tilknyttet fordeling, dvs. en usikkerhed. Hvis vi vil sammenligne normalfordelingens fit til data med andre fordelingers fit bør vi inddrage den usikkerhed i vores overvejelser. F.eks. har en gammafordeling en estimeret krydsentropi på $4,84$ og en log-normalfordeling en estimeret krydsentropi på $4,83$. I begge tilfælde lidt mindre end de $4,86$, og fra vores generelle betragtninger om krydsentropien fitter disse fordelinger data en lille smule bedre end normalfordelingen. Men i lyset af

usikkerheden er forskellene meget små, og vi kan næppe konkludere at disse to fordelinger på nogen afgørende måde er et bedre fit end normalfordelingen.

Vi ser også af figur 1.5 at gennemsnittet af $-\ell_X(\hat{\alpha}, \hat{\sigma}^2)/N$ er $4,84 < 4.86$, så ligesom for variansen underestimeres krydsentropien også en lille smule i middel. Hvad betyder det? Det betyder at $-\ell_X(\hat{\alpha}, \hat{\sigma}^2)/N$ i middel giver et lidt for optimistisk bud på, hvor godt $\mathcal{N}(\hat{\alpha}, \hat{\sigma}^2)$ fitter μ_0. Denne form for *overfit* skyldes, at maksimaliseringsestimatoren er optimeret til at fitte $\hat{\mu}_0$ (data) og ikke μ_0.

Det er faktisk muligt at korrigere teoretisk for dette overfit. Lad X og $\tilde{X}$ betegner to uafhængige datasæt hver med $N = 50$ uafhængige observationer fra en $\mu_0 = \mathcal{N}(273, 982)$-fordeling, og lad $\tilde{f} = f^{\otimes N}_{\tilde{\alpha},\tilde{\sigma}^2}$, hvor $(\tilde{\alpha}, \tilde{\sigma}^2)$ er MLE baseret på $\tilde{X}$. Så vil den forventede krydsentropi

$$E(H(\mu_0, \tilde{f})) = -\frac{1}{N} E(\ell_X(\tilde{\alpha}, \tilde{\sigma}^2)) > H(f_0)$$

udtrykke hvor godt den estimerede fordeling $\tilde{f}$ i middel fitter μ_0. Ved en Taylorudvikling er det muligt at vise, at

$$E(H(\mu_0, \tilde{f})) = -\frac{1}{N} E(\ell_X(\hat{\alpha}, \hat{\sigma}^2)) + R,$$

hvor restleddet kan approksimeres på en overraskende simpel måde;

$$R \simeq \frac{k}{N}$$

for en model parametriseret med en k-dimensional parameter. I det konkrete eksempel er $k = 2$, hvilket giver at $R \simeq 2/50 = 0.04$. Det er også muligt at vise at

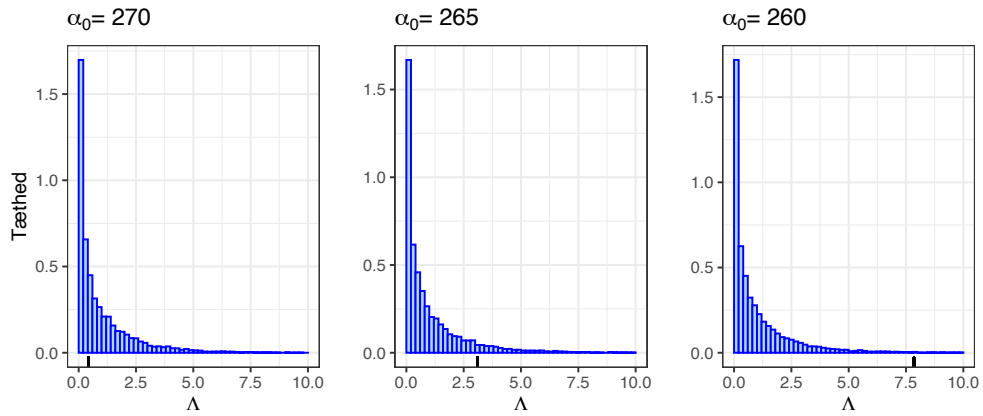
$$R \simeq 2(H(f_0) + \frac{1}{N} E(\ell_X(\hat{\alpha}, \hat{\sigma}^2))),$$

som på basis af simulationerne giver $R \simeq 2 \times (4.86 - 4,84) = 2 \times 0.02 = 0.04$ i det konkrete eksempel, hvilket stemmer med det generelle resultat.

Normalt formuleres den teoretiske korrektion for optimismen i termer af *Akaikes informationskriterie*:

$$\text{AIC} = -2\ell_x(\hat{\theta}) + 2k,$$

hvor $\hat{\theta}$ er MLE for en k-dimensional parameter. Man kan, jf. argumenterne ovenfor, tænke på $\text{AIC}/(2N)$ som et estimat for $E(H(\mu_0, \tilde{f}))$. Skalaen for AIC er som den er



Figur 1.6: Histogrammer for fordelingerne af likelihood ratio teststørrelsen Λ for tre forskellige hypotetiske værdier af middelværdien α_0. Værdien af Λ for reaktionstiderne er indikeret på hvert histogram.

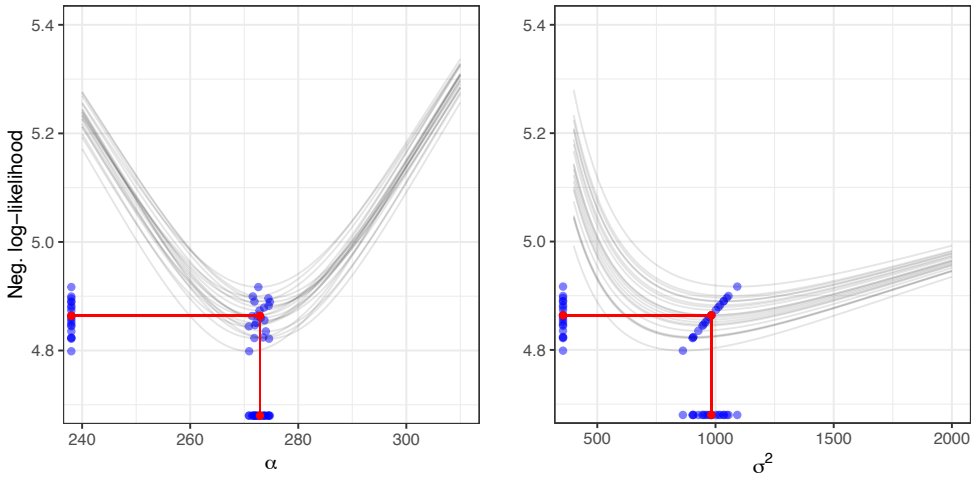
af historiske grunde, men vi kan konvertere til skalaen for krydsentropi ved division med $2N$. Størrelsen $-\ell_x(\hat{\theta})/N$ er, som vi har argumenteret for, et optimistisk estimat for den forventede krydsentropi, og korrektionsleddet k/N justerer netop for denne optimisme. I det konkrete eksempel er optimismen, og dermed korrektionen, meget beskeden i forhold til variationen, som figur 1.5 viser. For modeller med flere parametre i forhold til N kan korrektionen have større betydning, men man må ikke glemme at AIC$/(2N)$ er et *estimat* for $E(H(\mu_0, \tilde{f}))$, og at der altid er knyttet en usikkerhed til sådan en størrelse.

Hvor AIC er et direkte forsøg på at gøre estimeret krydsentropi sammenlignelig på tværs af forskellige modeller, så kan log-likelihoodfunktionen bruges på en noget mere direkte og præcis måde, hvis en model er en simplificering af en anden. F.eks. hvis vi vil undersøge en specifik hypotese om en parameter. For reaktionstiderne kunne vi spørge om normalfordelingen med $\alpha = 270$ eller $\alpha = 260$ fitter data. Hvis vi holder α fast på en værdi, α_0, følger det af (1.6) at MLE for σ^2 er

$$\hat{\sigma}^2(\alpha_0) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \alpha_0)^2.$$

Vi kan nu se på forskellen

$$\Lambda = 2(\ell_X(\hat{\alpha}, \hat{\sigma}^2) - \ell_X(\alpha_0, \hat{\sigma}^2(\alpha_0))) = N \log\left(\frac{\hat{\sigma}^2(\alpha_0)}{\hat{\sigma}^2}\right) \tag{1.7}$$



Figur 1.7: Negative profil log-likelihoodfunktioner skaleret med faktoren $1/N$ (profil krydsentropi) for 20 simulerede datasæt med $N = 500$. De blå punkter angiver minima og maksimaliseringsestimatorer for de 20 simulerede datasæt. De røde punkter/linjer angiver minimum og maksimaliseringsestimator for reaktionstiderne.

som et mål for, hvor meget bedre modellen med $\hat{\alpha}$ fitter end modellen med α_0. Da $\sigma^2(\alpha_0) \geq \hat{\sigma}^2$ er $\Lambda \geq 0$, og jo større Λ er, jo dårligere fitter modellen med $\alpha = \alpha_0$. Størrelsen Λ kaldes ofte *likelihood ratio teststørrelsen*. Figur 1.6 viser de simulerede fordelinger af Λ for $\alpha_0 = 270$, 265 og 260 sammenholdt med værdien af Λ for data. Fordelingerne ligner hinanden utroligt meget, og er faktisk alle tre approksimativt χ^2_1-fordelinger. På basis af hvor ekstrem den observerede værdi af Λ er i fordelingerne, konkluderer vi, at $\alpha = 270$ fitter data fint, $\alpha = 265$ er et dårligere fit, men ikke så ekstremt, at vi ville afvise fittet helt, mens $\alpha = 260$ fitter så dårligt, at vi vil afvise 260 som en rimelig middelværdi. Den konklusion er helt i overensstemmelse med det indtryk man får af fordelingen af $\hat{\alpha}$ fra figur 1.5.

I alle overvejelserne ovenfor om fordelinger knyttet til log-likelihoodfunktionen og MLE har størrelsen af datasættet, N, været fast. I den konkrete eksempel er $N = 50$. Hvad ville der ske, hvis N var større? Figur 1.7 viser profil log-likelihoodfunktioner for simulerede datasæt for $N = 500$. Det fremgår klart ved sammenligning med figur 1.4 at variationen af log-likelihoodfunktionerne er mindre for $N = 500$ end for $N = 50$. Denne formindskede variation af kurverne oversætter til en formindsket variation af $\hat{\alpha}$ og $\hat{\sigma}^2$ såvel som $-\ell_X(\hat{\alpha}, \hat{\sigma}^2)/N$.

Det er helt forventeligt at et større N giver mindre usikkerhed og dermed mere præcise estimater. Vi vil endvidere forvente at

$$-\frac{1}{N}\ell_X(\alpha,\sigma^2) = -\frac{1}{N}\sum_{i=1}^{N}\log f_{\alpha,\sigma^2}(X_i) \to H(\mu_0, f_{\alpha,\sigma^2})$$

for $N \to \infty$; og det er faktisk præcis det som store tals lov fra sandsynlighedsteorien fortæller os – for $X_1, X_2, \ldots$ uafhængige og identisk μ_0-fordelte. Den matematiske statistik handler i vid udstrækning om at forstå, hvordan denne konvergens giver anledning til brugbare kvantificeringer af usikkerhed for endelige værdier af N. Usikkerhed som udtrykkes i termer af fordelinger for f.eks. MLE eller likelihood ratio teststørrelsen. Normalfordelingsmodellerne er specielle i den henseende, fordi vi i vid udstrækning kan opnå matematisk eksakte resultater for alle værdier af N, mens vi for stort set alle andre modeller må stille os tilfredse med asymptotiske resultater for $N \to \infty$.

1.4 Normalfordelinger og transformationer

For at finde eksakte fordelingsresultater i normalfordelingsmodellerne har vi brug for en række standard transformationsresultater knyttet til normalfordelinger. I dette afsnit opsummeres de vigtigste resultater uden beviser, og der henvises generelt til Ernst Hansens bog "Measure Theory", især afsnittene 18.5 og 20.5.

Husk at en regulær normalfordeling $\mathcal{N}(\xi,\Sigma)$ på $\mathbb{R}^N$ har tæthed

$$f_{\xi,\Sigma}(x) = \left(\frac{1}{(2\pi)^N \det(\Sigma)}\right)^{1/2} e^{-\frac{1}{2}(x-\xi)^T\Sigma^{-1}(x-\xi)}$$

mht. lebesguemålet. Her er $\xi \in \mathbb{R}^N$ og Σ er en positiv definit $N \times N$-matrix.

Sætning 1.4.1 (Egenskaber for den regulære normalfordeling)**.** *Hvis $X \sim \mathcal{N}(\xi,\Sigma)$ gælder følgende:*

1. *$E(X_i) = \xi_i$ og $\operatorname{cov}(X_i, X_j) = \Sigma_{i,j}$.*
2. *X_i og X_j er uafhængige netop hvis $\Sigma_{i,j} = 0$.*
3. *Hvis B er en $k \times N$ matrix af rang k vil $BX \sim \mathcal{N}(B\xi, B\Sigma B^T)$.*

Resultaterne er en opsummering af EH Measure Theory, eksempel 19.15, sætning 18.27 og korollar 18.29.

Bemærk at $X_1, \ldots, X_N$ er uafhængige og identisk $\mathcal{N}(\alpha, \sigma^2)$-fordelte netop hvis

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \right).$$

Dvs. netop hvis $X \sim \mathcal{N}(\alpha \mathbf{1}, \sigma^2 I_N)$, hvor $\mathbf{1}$ er en søjlevektor af lutter 1-taller og I_N er $N \times N$-identitetsmatricen. Vi har altså, at

$$f_{\alpha,\sigma^2}^{\otimes N} = f_{\alpha \mathbf{1}, \sigma^2 I_N}.$$

Sætning 1.4.2 (Sum af kvadrerede standard normalfordelinger)**.** *Hvis $X_1, \ldots, X_k$ er* uafhængige $\mathcal{N}(0,1)$*-fordelte er summen af kvadraterne χ^2-fordelt med k frihedsgrader;*

$$X_1^2 + \ldots + X_k^2 \sim \chi_k^2.$$

Beviset er i EH Measure Theory, Example 20.11.

OBS: χ_k^2 er også en Γ-fordeling med formparameter $k/2$ og skalaparameter 2. Den har middelværdi k og varians $2k$.

Ved en skalatransformation følger:

Korollar 1.4.3 (Sum af kvadrerede normalfordelinger)**.** *Hvis $X_1, \ldots, X_k$ er* uafhængige $\mathcal{N}(0, \sigma^2)$*-fordelte er*

$$X_1^2 + \ldots + X_k^2 \sim \sigma^2 \chi_k^2.$$

Lad $X_1, X_2, \ldots, X_m$ være uafhængige med $X_i \sim \sigma^2 \chi_{k_i}^2$ og definer de partielle summer

$$S_i = X_1 + \ldots + X_i \sim \sigma^2 \chi_{k_1 + \ldots + k_i}^2, \quad i = 1, \ldots, m.$$

Lad endvidere $B(\alpha, \beta)$ betegne en betafordeling med formparametre α og β.

Sætning 1.4.4 (Brøker af χ^2-fordelte variable)**.** *Variablene*

$$\frac{S_1}{S_2} = \frac{X_1}{X_1 + X_2},\ \frac{S_2}{S_3} = \frac{X_1 + X_2}{X_1 + X_2 + X_3}, \ldots, \frac{S_{m-1}}{S_m} = \frac{X_1 + \ldots + X_{m-1}}{X_1 + \ldots + X_m},\ S_m = X_1 + \ldots + X_m$$

er uafhængige og for $i = 1, \ldots, m-1$ *er*

$$\frac{S_i}{S_{i+1}} \sim B\left(\frac{k_1 + \ldots + k_i}{2}, \frac{k_{i+1}}{2}\right).$$

Beviset er i EH Measure Theory, Example 20.24.

Sætning 1.4.5 (Brøk af en normalfordeling og en χ-fordeling)**.** *Lad* $X \sim \mathcal{N}(0, \sigma^2)$ *og* $S \sim \frac{\sigma^2}{k}\chi^2_k$ *være uafhængige, så er*

$$\frac{X}{\sqrt{S}} \sim t_k$$

hvor t_k *betegner t-fordelingen med k frihedsgrader (formparameter* $k/2$*).*

Beviset er i EH Measure Theory, Example 20.27. Fordelingen af $\sqrt{S}$ kaldes naturligt en χ-fordeling med k frihedsgrader og skalaparameter $\sigma/\sqrt{k}$.

Sætning 1.4.6 (Brøk af to χ^2-fordelinger)**.** *Hvis* $X_1 \sim \frac{\sigma^2}{k_1}\chi^2_{k_1}$ *og* $X_2 \sim \frac{\sigma^2}{k_2}\chi^2_{k_2}$ *er uafhængige, så er*

$$\frac{X_1}{X_2} \sim F(k_1, k_2),$$

dvs. F-fordelt med frihedsgrader (k_1, k_2)*.*

Beviset er i EH Measure Theory, Example 20.28.

1.5 Opsummering

Kapitlet berører flere fundamentale problemstillinger indenfor statistik – og er selvfølgelig slet ikke nået i dybden med dem. Via det gennemgående normalfordelingseksempel skulle de centrale pointer være trådt frem, men den primære intention med kapitlet er at motivere en række matematiske konstruktioner og resultater, som bliver

behandlet i dybden i andre kapitler, og det er nok værd at læse dette kapitel igen når matematikken er faldet mere på plads.

Kapitlet stiller også en lang række spørgsmål, og vi skal her på listeform forsøge at give korte svar på nogle af de vanskelige spørgsmål, der blev stillet undervejs. Ikke mindst for at have nogle pejlemærker for, hvad det er matematikken skal hjælpe os med at få svar på.

- Vi indsamler data for at svare på spørgsmål af generel karakter. Data skal via induktion tillade os at generalisere.
- Statistiske modeller giver en matematiske ramme for generaliseringen.
- Systematisk variation og kendte afhængigheder bygges ind i vores modeller. Øvrig variation opfattes som bestående af uafhængige og usystematiske komponenter, der kan beskrives sandsynlighedsteoretisk.
- Parametriserede statistiske modeller kan tilpasses data via maksimering af likelihoodfunktionen.
- Modeller kan sammenlignes indbyrdes og en model kan sammenlignes med data via log-likelihoodfunktionen.
- Visualisering og andre *ad hoc* procedurer kan også anvendes til at undersøge overensstemmelse mellem model og data.
- Usikkerhed kvantificeres ved at svare på spørgsmålet: Hvad kunne vi ellers have observeret (givet vores model)?

Vi skal i de følgende kapitler i stor detaljegrad gennemgå alle ovenstående punkter for en række normalfordelingsmodeller. Resultatet vil være et fleksibelt værktøj, som vi kan bruge til praktisk dataanalyse. Det er en Schweizerkniv – et multiværktøj som kan mange ting, og som finder mange anvendelser. Men multiværktøjet kan ikke alt! Det er vigtigt at erkende, at hvis værktøjet ikke egner sig til opgaven, så skal vi lade være med at bruge det.

Det skal også understreges af ovenstående punktliste ikke skal ses som en normativ programerklæring for, hvordan vi *bør* lave statistik. Det er en konstatering af, at sådan *kan* vi lave statistik, og når vi laver statistik på den måde, udfolder bogen

hvordan vi gør det korrekt i praksis. Bogens matematiske udledninger må altså ikke fortolkes som en deduktion af den universelt korrekte måde at lave statistik på. Alle deduktionerne er kvalificerede af en række forudsætninger såsom "hvis vi ønsker at bruge den her model", "hvis vi ønsker at udregne den her størrelse", "hvis vi ønsker at kvantificere usikkerhed på den her måde", og "hvis modellen er rimelig". Bogen giver således svar på, hvordan vi laver de statistiske analyser korrekt, hvis vi gør det indenfor de rammer, som bogen opstiller. Og forhåbentlig opleves bogen ikke som en opskrift, som læseren kan følge blindt, men som en dybdegående og meningsfyldt uddybelse af hvordan ovenstående punktliste kan realiseres i praksis. Når den erkendelse er opnået tilskyndes læseren på det kraftigste til også at undersøge andre rammer for statistik.

1.6 Opgaver

Opgave 1.1. Lad $X_1, \ldots, X_N$ være uafhængige og identisk $\mathcal{N}(\alpha, \sigma^2)$-fordelte og lad $X = (X_1, \ldots, X_N)^T$.

1. Vis at
$$\hat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} X_i = BX$$
hvor $B = \frac{1}{N}\mathbf{1}^T$.
2. Vis at $\hat{\alpha} \sim \mathcal{N}(\alpha, \sigma^2/N)$.
3. Undersøg ovenstående fordelingsresultat ved simulation. Fortolk betydningen af faktoren $1/N$ på variansen for $\hat{\alpha}$.

Opgave 1.2. Lad $X_1, \ldots, X_N$ være uafhængige og identisk $\mathcal{N}(0, \sigma^2)$-fordelte og lad $X = (X_1, \ldots, X_N)^T$. Definer endvidere
$$\tilde{X}_i = X_i - \frac{1}{N} \sum_{j=1}^{N} X_j$$
og $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_N)^T$ således at
$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \tilde{X}_i^2.$$

1. Vis at
$$\tilde{X} = \left(I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T \right) X.$$
2. Vis at den symmetriske matrix
$$B = \left(I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T \right)$$
har egenværdierne 1 med multiplicitet $N - 1$ og 0 med multiplicitet 1.
Vink: Matrixdeterminantlemmaet giver at
$$\det(cI_N + b\mathbf{1}\mathbf{1}^T) = (1 + bN/c)c^N.$$
Find det karakteristiske polynomium for B.

3. Vis ved diagonalisering at $B = QQ^T$ for en ortogonal $N \times (N-1)$-matrix Q. (At Q er ortogonal betyder at $Q^TQ = I_{N-1}$.)

4. Vis at $Q^TX \sim \mathcal{N}(0, \sigma^2 I_{N-1})$.

5. Vis at
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N-1}(Q^TX)_j^2$$
og konkluder at $\hat{\sigma}^2 \sim \frac{\sigma^2}{N}\chi^2_{N-1}$.

Opgave 1.3. Lad $X \sim \mu_0$ være en reel stokastisk variabel med endelig middelværdi og varians og lad $\ell_X(\alpha, \sigma^2)$ være log-likelihoodfunktionen for $\mathcal{N}(\alpha, \sigma^2)$-fordelingen.

1. Vis at
$$E(\ell_X(\alpha, \sigma^2)) = -\frac{1}{2\sigma^2}E\left((X-\alpha)^2\right) - \frac{1}{2}\log(2\pi\sigma^2).$$

2. Vis at $E(\ell_X(\alpha, \sigma^2))$ maksimeres for $\alpha = E(X)$ og $\sigma^2 = V(X)$.

Opgave 1.4. Lad $X \sim \mu_0$ på $\mathbb{R}^N$, og lad $f_{\xi,\Sigma}$ betegne tætheden for den regulære normalfordeling på $\mathbb{R}^N$ med middelværdivektor ξ og kovariansmatrix Σ. Vis at krydsentropien af $\mathcal{N}(\xi, \Sigma)$ relativt til μ_0 er

$$H(\mu_0, f_{\xi,\Sigma}) = \frac{1}{2}\mathrm{tr}(\Sigma^{-1}E((X-\xi)(X-\xi)^T)) + \frac{N}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma).$$

Konkluder at entopien for den regulære normalfordeling er

$$H(f_{\xi,\Sigma}) = \frac{N}{2} + \frac{N}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma).$$

Opgave 1.5. Lad $\mu = f \cdot \nu$ og $\mu_0 = f_0 \cdot \nu$ være sandsynlighedsmål på målrummet $(\mathcal{X}, \mathbb{F}, \nu)$. Såfremt krydsentropien af μ relativt til μ_0 er veldefineret og entropien af μ_0 er endelig, $H(f_0) < \infty$, defineres *Kullback-Leibler divergensen* fra μ til μ_0 som

$$D(\mu_0 \,\|\, \mu) = H(\mu_0, f) - H(f_0).$$

1. Vis at $D(\mu_0 \,\|\, \mu) \geq 0$.

2. Vis at
$$D(\mu_0 \,\|\, \mu) = \int f_0(x)\log\left(\frac{f_0(x)}{f(x)}\right)\mathrm{d}\nu(x).$$

3. Vis at $D(\mu_0 \,\|\, \mu)$ ikke afhænger af valget af grundmål ν.

Den absolutte værdi af såvel entropi som krydsentropi afhænger af valget af ν, mens ovenstående resultat viser, at $D(\mu_0 \,\|\, \mu)$ er uafhængig af dette valg. Dermed kan KL-divergensen vitterligt fortolkes som et absolut mål for hvor meget fordelingen μ afviger fra μ_0, mens krydsentropien kun kan fortolkes som et relativt mål.

Litteratur

Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, New York, 2016.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014. ISBN 978-1-4398-4095-5.

Ernst Hansen. *Introduktion til Matematisk Statistik, Bind 1*. Afdeling for Anvendt Matematik og Statistik, KU, 2012. ISBN 978-87-70782-08-1.

Steffen Lauritzen. *Basic Mathematical Statistics*. Department of Mathematical Sciences, UCPH, 2021. ISBN 978-87-7125-075-6.

David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, New York, 2003. ISBN 0-521-64298-1.