Kapitel 1

Likelihood metoden

I dette kapitel konfronteres læseren brat med det teoretiske grundlag for statistikken, sådan som det ser ud når man holder sig til *likelihood metoden,* som er langt den mest udbredte metode inden for faget statistik. Den følgende gennemgang af likelihood metoden er meget teoretisk og kortfattet, og man skal ikke vente at fatte ret meget af hvad der foregår første gang man læser det. Men når metoden er blevet brugt nogle gange i de følgende kapitler kan man vende tilbage, og forhåbentlig forstå lidt mere.

1.1. Observationen.

Givet er et *observationsrum* (det samme som i sandsynlighedsregningen kaldes et *udfaldsrum*) E og en *observation* $x \in E$. En observation, i denne jargon, er ikke bare et enkelt eller nogle få tal, men oftest et helt datamateriale, svarende til at E er $\mathbf{N}_0^n$ eller $\mathbf{R}^n$ for et eller andet (undertiden ret stort) tal n.

1.2. Modellen.

Til forklaring af hvordan observationen er fremkommet benyttes en *statistisk model,* hvorved forstås en parametriseret mængde $\{P_\vartheta \mid \vartheta \in \Theta\}$ af sandsynlighedsfordelinger på E. Ideen er, at observationen x antages fremkommet som værdien af en stokastisk variabel $X \in E$ med en af disse fordelinger P_ϑ, vi ved bare ikke hvilken. ϑ er den ukendte *parameter*, og Θ kaldes *parameterrummet.* Oftest er Θ en åben delmængde af et reelt talrum $\mathbf{R}^d$. Vi skriver så $\vartheta = (\vartheta_1, \dots, \vartheta_d)$, og siger at modellen har d parametre $\vartheta_1, \dots, \vartheta_d$. Hvis parametriseringen er injektiv (altså $\vartheta' \neq \vartheta'' \Longrightarrow P_{\vartheta'} \neq P_{\vartheta''}$) siger vi at modellen er af *dimension* d.

1.3. Likelihoodfunktionen.

Ved likelihoodfunktionen, eller kort *likelihood'en*, forstås funktionen

$$L : \Theta \to \mathbf{R}$$

givet ved

$$L(\vartheta) = p_\vartheta(x)$$

hvor p_ϑ betegner sandsynlighedsfunktionen for P_ϑ. Likelihoodfunktionen er altså sandsynligheden for det udfald vi har observeret, som funktion af den ukendte parameter.

Denne definition kan naturligvis kun bruges for diskrete fordelinger, altså typisk når observationsrummet er $\mathbf{N}_0^n$ eller lignende. Hvis fordelingerne P_ϑ er kontinuerte (typisk $E = \mathbf{R}^n$) erstatter man sandsynlighedsfunktionen med tætheden.

Logaritmen til likelihood'en — kaldet *log–likelihooden* — betegnes

$$l(\vartheta) = \log L(\vartheta).$$

For $L(\vartheta) = 0$ sættes $l(\vartheta)$ naturligt lig med $-\infty$. Men for de fleste af de statistiske modeller man bruger i praksis gælder, at likelihoodfunktionen altid er positiv, svarende til at de sandsynlighedsfordelinger der indgår i modellen alle har strengt positiv sandsynlighedsfunktion eller tæthed.

For de statistiske metoder der er knyttet til likelihood og log–likelihood funktionerne gælder, at de alene er baserede på *forholdet* mellem likelihood'ens værdier i forskellige punkter, eller *differensen* mellem log–likelihood'ens værdier i forskellige punkter. De er således upåvirkede af om man ganger likelihood'en med en positiv konstant, eller tilsvarende adderer en konstant til log–likelihood'en. Ved opskrivning af en likelihood (log–likelihood) er det derfor tilladt at ignorere multiplikative (additive) led, der ikke afhænger af parameteren.

1.4. Maksimum–likelihood estimation.

Estimation (dvs. skønsmæssig bestemmelse) af parameteren ϑ foretages normalt ved *maksimum likelihood* metoden, som går ud på at estimere ϑ med den værdi $\hat{\vartheta}$ der maksimerer likelihood'en. Altså den (forhåbentlig entydige) værdi af parameteren, som opfylder

$$L(\hat{\vartheta}) \geq L(\vartheta) \text{ for alle } \vartheta \in \Theta.$$

Denne størrelse kaldes (hvis den eksisterer) *maksimum–likelihood estimatoren,* eller kort *ML–estimatoren.* Bestemmelse af den kræver i princippet en funktionsundersøgelse af L eller $l = \log L$. Oftest er log–likelihood'en nemmere at have med at gøre end selve likelihood'en, og i pæne tilfælde, hvor der er et entydigt maksimum (og ikke andre lokale maksima, minima, saddelpunkter etc.) beregnes $\hat{\vartheta}$ som løsning til *likelihoodligningerne*

$$\frac{d}{d\vartheta_j} l(\vartheta) = 0 \ , \quad j = 1, \ldots, d.$$

Når man angiver $\hat{\vartheta}$ bør man sædvanligvis også oplyse noget om estimationens *usikkerhed.* Det gør man normalt ved at angive et (eksakt eller approksimativt) *konfidens–* eller *sikkerhedsinterval* for den enkelte parameter. For eksempel defineres et 95% sikkerhedsinterval for en reel parameter (eller parameterfunktion) $\eta = \eta(\vartheta)$ som et interval $I(x) = [\eta_{\text{nedre}}(x), \eta_{\text{øvre}}(x)]$ med den egenskab, at hændelsen $\{\eta \in I(X)\}$ har sandsynlighed (eksakt eller approksimativt) 0.95.

1.5. Testning.

Testning eller *hypoteseprøvning* går ud på følgende. Antag at vi har en formodning om, at observationen lige så godt eller næsten lige så godt kan beskrives ved en *reduceret model* $\{P_\vartheta \mid \vartheta \in \Theta_0\}$, bestemt ved en delmængde Θ_0 af Θ. For at undersøge om dette kan være rigtigt går vi frem på følgende måde. Betragt *kvotientteststørrelsen* eller *likelihood–ratio teststørrelsen*

$$q = \frac{L(\hat{\vartheta}_0)}{L(\hat{\vartheta})} = \frac{\max_{\vartheta\in\Theta_0} L(\vartheta)}{\max_{\vartheta\in\Theta} L(\vartheta)}$$

hvor $\hat{\vartheta}_0$ betegner maksimaliseringsestimatoren i den reducerede model givet ved Θ_0. Denne størrelse vil åbenbart ligge mellem 0 og 1, da tælleren (som maksimum af samme ikke–negative funktion over en mindre mængde) er mindre end nævneren. Princippet er nu, at man forkaster hypotesen om at modellen kan reduceres på denne måde, hvis denne størrelse er ekstremt lille i sin fordeling under hypotesen. Ligesom alt andet, der er udregnet på basis af observationen x, er q jo den observerede værdi af en stokastisk variabel Q, der følger en fordeling, som i princippet kan regnes ud ved transformation af den oprindelige fordeling af X. At q er meget lille betyder jo — hvis vi tænker på likelihood'ens oprindelige definition — at den oprindelige model for visse parameterværdier tillægger betydeligt større sandsynlighed til den hændelse vi har observeret, end den reducerede model er i stand til.

1.6. Signifikansniveau.

Hvis der (for eksempel) for alle $\vartheta \in \Theta_0$ gælder

$$P_\vartheta(Q \leq q) \leq 0.01$$

ræsonnerer vi som så: Hvis vi vil tro på at den ukendte parameter ϑ tilhører delmængden Θ_0 må vi samtidig acceptere, at der her er sket noget som normalt kun indtræffer højst 1 ud af 100 gange. Hvis derimod vi holder os til den store model, har vi ikke dette forklaringsproblem, for under denne model kan Q have andre fordelinger, der tillægger stor sandsynlighed til værdier af Q som er $\leq q$.

I den teoretiske fremstilling af testteorien forestiller man sig, at der på forhånd er fastlagt en størrelse α, kaldet *signifikansniveauet*, ofte $\alpha = 0.05$. Man forkaster så hypotesen "$\vartheta \in \Theta_0$" *på niveau* α, eller siger at testet udviser *signifikans på niveau* α, hvis der for alle $\vartheta \in \Theta_0$ gælder

$$P_\vartheta(Q \leq q) \leq \alpha.$$

Standardværdien $\alpha = 0.05$ svarer til, at man ikke accepterer en simplere model, hvis det indebærer at der i så fald er sket noget som sker højst én ud af 20 gange. Det betyder så til gengæld, at man i 1 ud af 20 tilfælde vil komme til at forkaste den reducerede model, selv om den er rigtig. Så i mange tilfælde — især hvis man har vægtige grunde til at tro på den reducerede model — bør man benytte et mindre signifikansniveau, f.eks. $\alpha = 0.01$.

1.7. P–værdi.

Sandsynligheden $P_\vartheta(Q \leq q)$, den såkaldte *P–værdi* eller *testsandsynlighed*, er i mange tilfælde approksimativt uafhængig af $\vartheta \in \Theta_0$. I praksis benytter man ikke en så firkantet fremgangsmåde som beskrevet ovenfor, da den jo fører til det rene hasard ("knald eller fald") i situationer hvor P–værdien falder tæt ved α. Desuden indeholder selve P–værdien vigtig information om den grad af *signifikans* hvormed hypotesen kan afvises. Der er meget stor forskel på en P–værdi lige under 0.05 og en P–værdi på 0.000001. Så man rapporterer hellere selve P–værdien, og overlader i et vist omfang til andre, hvad de har lyst til at tro på. Fortolkningen af en P–værdi på f.eks. 0.000001 er groft sagt, at hvis man vil insistere på at den reducerede model er rigtig, så må man også acceptere, at der netop i dette tilfælde er sket noget som normalt kun sker én ud af en million gange.

OPGAVE 1.7.1. Du spiller bridge sammen med A, B og C. A er en slyngel, som siges at være god til at mingelere det sådan at han selv får en masse esser når han giver. Efter to spil, hvor han har været giver, har han selv fået alle fire esser begge gange. Du har den største trang til at stikke ham et par på snuden. I hvor høj grad kan denne trang retfærdiggøres kvantitativt?

1.8. Approksimation af kvotientteststørrelsens fordeling.

Delmængden Θ_0 er i mange tilfælde en pæn mængde af lavere dimension end Θ, f.eks. en hyperplan eller flade som kan parametriseres injektivt med d_0 parametre, $d_0 < d$. I så fald — og under en række yderligere regularitetsantagelser, som det vil føre alt for vidt at komme ind på her — gælder, at fordelingen af Q for $\vartheta \in \Theta_0$ med god approksimation kan beskrives ved, at

$$-2 \log Q = 2 \left(l(\hat{\vartheta}) - l(\hat{\vartheta}_0) \right)$$

(der ofte kaldes kvotientteststørrelsen, ligesom Q selv) følger en χ^2–fordeling med $d - d_0$ frihedsgrader. Bemærk at $-2 \log Q$ er ikke–negativ og afhænger monotont aftagende af Q, så det er for *store* værdier af denne størrelse man skal forkaste den reducerede model. I praksis foregår testning derfor ofte ved, at man udregner $-2 \log q$ og slår op i en χ^2–tabel med det antal parametre man har "fjernet" fra modellen som antal frihedsgrader. P–værdien er så approksimativt lig med halesandsynligheden

$$P\left(\chi^2_{d-d_0} \geq -2\log q\right)$$

i denne fordeling. I mangel af bedre må man, da tabeller kun angiver visse fraktiler, nøjes med at angive en øvre grænse for P–værdien, svarende til den nærmeste fraktil i tabellen der er $\leq -2\log q$.

OPGAVE 1.8.1. $x = (41, 74, 42, 97, 38, 82, 195, 37, 2, 148) \in \mathbf{N}_0^{10}$ antages at være fremkommet ved observation af 10 uafhængige geometrisk fordelte (Ssr. side 37) variable med samme parameter $p > 0$.

(a) Opskriv likelihood og log–likelihood, og udregn maksimaliseringsestimatoren for p.

(b) Foretag ved hjælp af den nævnte χ^2–approksimation kvotienttestet for hypotesen $p = 0.01$.