

Kapitel 1

Bayesianska metoder

1.1 Översikt

Vi har en storhet (parameter) θ som vi vill t ex skatta, ge konfidensintervall för eller utföra en hypotesprövning om. Denna parameter har i tidigare kurser betraktats som fix men okänd. I Bayesiansk statistik så uppfattas θ i stället som ett utfall av en stokastisk variabel Θ som har en fördelning (den s k a-priori-fördelningen – a-priori=i förväg) beskriven av t ex en sannolikhetsfunktion eller en täthetsfunktion. Denna a-priori-fördelning kan t ex avspegla tidigare skattningar av θ , men kan också avspegla en subjektiv uppfattning om vilka värde på θ som är mest sannolika. A-priori-fördelningen kan också avspegla en genuin osäkerhet om parameterns värde genom att vi låter den ha en likformig fördelning.

Exempel 1.1 *Pumpars felintensitet*

Vi har ett förråd med pumpar som alla har konstant felintensitet. men där denna varierar mellan de olika pumparna. Vi väljer nu en pump på måfå och till denna pump hör en ”sann” felintensitet λ . (Vi kallar parametern λ i stället för θ av bekvämlighetsskäl.)

Denna felintensitet λ ser vi som ett framlottat värde från en stokastisk felintensitet Λ . A-priori-fördelningen för Λ beskriver då hur felintensiteten varierar mellan de olika pumparna i förrådet. Utfallet λ av denna stokastiska felintensitet beror alltså av vilken pump vi råkat få. När vi sedan tittar på tiden till ett fel på pumpen så har denna tid en $\text{Exp}(1/\lambda)$ -fördelning. Man kan uppfatta det som en lottning i två steg. Först lottas en felintensitet λ fram genom att vi råkar välja en viss pump, och sen blir det en slumpmässighet i tiden till första felet beskriven av exponentialfördelningen som har just denna felintensitet λ . \square

1.2 Likelihood-funktion

Traditionellt formuleras en statistisk modell för n observationer x_1, x_2, \dots, x_n som att x_i :na är ett utfall av (oftast oberoende) stokastiska variabler $X_1, X_2,$

\dots, X_n vars simultana fördelning beror av en parameter θ dvs t ex att man anger

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) \quad (x_1, x_2, \dots, x_n) \in Z^n$$

eller

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta) \quad (x_1, x_2, \dots, x_n) \in R^n$$

beroende på om vi har diskreta eller kontinuerliga mätdata.

Dessa sannolikheter (tätheter) att få de data vi fått har vi i tidigare kurser sett som en funktion av θ , den s k likelihoodfunktionen $L(x_1, x_2, \dots, x_n; \theta)$.

Maximum Likelihoodmetoden (ML-metoden) innebär att vi som skattning $\hat{\theta}$ av parametern θ tagit det värde på θ som maximerat likelihoodfunktionen $L(x_1, x_2, \dots, x_n; \theta)$.

En Bayesian uppfattar i stället denna likelihoodfunktion (dvs sannolikhet eller täthet) som en betingad sannolikhet (täthet) där man betingat på att $\Theta = \theta$. Man vill nu "vända på" denna betingning och i stället få fördelningen för Θ givet de observationer x_1, x_2, \dots, x_n vi fått. Man kanske minns från grundkursen att Bayes' sats var ett verktyg för att "vända på" betingningar. Man kan uppfatta detta som att man vill uppdatera a-priori-fördelningen för Θ med de observationer x_1, x_2, \dots, x_n vi fått och på så vis erhålla fördelningen för Θ givet de observationer x_1, x_2, \dots, x_n vi fått.

Om man skulle renodla skillnaden mellan en Bayesian och en icke-Bayesian (frekventist) så ser Bayesianen data (x -n) som fixa medan parametern är slumpmässig, medan icke-Bayesianen ser parametern som fix medan data ses som utfall av stokastiska variabler, dvs som slumpmässiga.

1.3 Bayes' sats

Sats 1.1 Bayes' sats

Om $\cup_{i=1}^{\infty} H_i = \Omega$ och $H_i \cap H_j = \emptyset$, $i \neq j$, dvs att H_i :na delar upp utfallsrummet Ω i disjunkta delar så gäller att

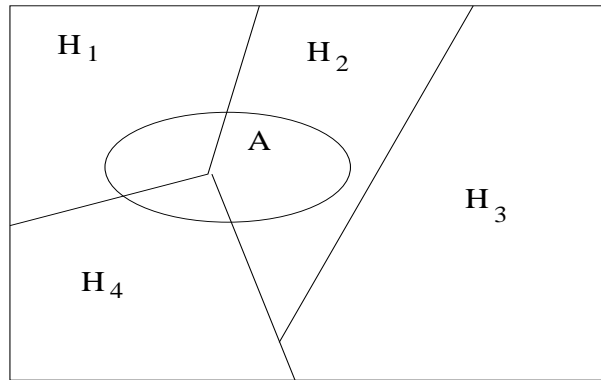
$$P(H_{i_0}|A) = \frac{P(H_{i_0} \cap A)}{P(A)} = \frac{P(A|H_{i_0})P(H_{i_0})}{P(A)} = \frac{P(A|H_{i_0})P(H_{i_0})}{\sum_{i=1}^{\infty} P(A|H_i)P(H_i)}$$

som kan anses som en formel som hjälper till att "vända på" betingningar. Jämför figur 1.1. \square

Denna sats visades redan tidigt i grundkursen men är här av avgörande betydelse. Bayesianisk statistik bygger helt och hållet på användningen av Bayes' sats.

Om vi låter $A = \{X = x\}$ och $H_y = \{Y = y\}$ (notera att då är $\cup_{y=0}^{\infty} H_y = \Omega$ och att de är disjunkta) så får vi (jämför $P(A|B) = P(A \cap B)/P(B)$ eller $P(A \cap B) = P(A|B)P(B)$)

$$P(Y = y|X = x) = p_{Y|X=x}(y) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X|Y=y}(x)p_Y(y)}{p_X(x)} =$$

Figur 1.1: Uppdelning av utfallsrummet Ω i disjunkta H_i :n

$$= \frac{p_{X|Y=y}(x)p_Y(y)}{\sum_{k=0}^{\infty} p_{X|Y=k}(x)p_Y(k)}.$$

Denna ”diskreta” formel har ”kontinuerliga” motsvarigheter där man på lämpliga ställen byter sannolikhetsfunktion mot täthetsfunktion. Vi har t ex (om både X och Y är kontinuerliga)

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y=y}(x)f_Y(y)}{f_X(x)} = \frac{f_{X|Y=y}(x)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y=u}(x)f_Y(u)du}.$$

eller (om Y är kontinuerlig och X är diskret)

$$f_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)f_Y(y)}{p_X(x)} = \frac{p_{X|Y=y}(x)f_Y(y)}{\int_{-\infty}^{\infty} p_{X|Y=u}(x)f_Y(u)du}.$$

eller (om Y är diskret och X är kontinuerlig)

$$p_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)p_Y(y)}{f_X(x)} = \frac{f_{X|Y=y}(x)p_Y(y)}{\sum_{u=0}^{\infty} f_{X|Y=u}(x)p_Y(u)}.$$

Dessutom kan man göra flerdimensionella varianter av ovanstående uttryck där vi alltså betraktar X och/eller Y som flerdimensionella. Vi kommer här att låta Y stå för Θ och X stå för X_1, X_2, \dots, X_n (om vi har mer än en observation).

1.4 Några exempel på likelihoodfunktioner

Vi ger här ett antal elementära (och förhoppningsvis bekanta) exempel på likelihoodfunktioner.

Exempel 1.2 Binomialfördelning

Vi har ett utfall x av X som vi anser vara $\text{Bin}(n, \theta)$, dvs att

$$L(x; \theta) = p_{X|\Theta=\theta}(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, 2, \dots, n, \quad 0 \leq \theta \leq 1.$$

Här har vi alltså bara ett enda mätdata x . □

Exempel 1.3 *Poissonfördelning*

Vi har observationer x_1, x_2, \dots, x_n som är utfall av oberoende $\text{Po}(\theta)$ -fördelade variabler

X_1, X_2, \dots, X_n dvs

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta) &= p_{X_1, X_2, \dots, X_n | \Theta = \theta}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i; \theta) = \\ &= \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} \exp(-\theta) = \frac{\theta^{x_1 + x_2 + \dots + x_n}}{x_1! x_2! \dots x_n!} \exp(-n\theta) \end{aligned}$$

□

Exempel 1.4 *Normalfördelning*

Vi har observationer x_1, x_2, \dots, x_n som är utfall av oberoende stokastiska variabler X_1, X_2, \dots, X_n som alla är $N(m, \sigma^2)$, dvs med $\theta = (m, \sigma^2)$ har vi

$$\begin{aligned} L(x_1, x_2, \dots, x_n; m, \sigma^2) &= L(x_1, x_2, \dots, x_n; \theta) = \\ &= f_{X_1, X_2, \dots, X_n | \Theta = \theta}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i; m, \sigma) = \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\right) \end{aligned}$$

□

1.5 A-posteriori-fördelning

Vi ser nu likelihoodfunktionen som en betingad sannolikhet (täthet) för mätdata x_1, x_2, \dots, x_n givet $\Theta = \theta$. Med hjälp av Bayes' sats skaffar vi på betingningen och erhåller fördelningen för Θ givet observationerna. Denna fördelning för Θ kallas a-posteriori-fördelningen (a-posteriori = i efterhand). Man kan uppfatta det som om man med hjälp av Bayes' sats uppdaterar a-priori-fördelningen med den information som mätdata ger.

1.6 Några exempel

1.6.1 Binomial-fördelning

Exempel 1.5 *Binomialfördelning – tvåpunktsfördelning för Θ*

Låt X vara $\text{Bin}(5, \theta)$ där θ i sin tur är ett utfall av en stokastisk variabel Θ där $P(\Theta = 0.2) = 0.4$ och $P(\Theta = 0.5) = 0.6$. Vi har alltså en lottning i två steg – dels en lottning om vilken sannolikhet Θ vi har och sen en lottning om vad X blir. Vi kan se det som att X är $\text{Bin}(5, 0.2)$ med sannolikhet 0.4 och att X är $\text{Bin}(5, 0.5)$ med sannolikhet 0.6. Vi har a-priori-fördelningen som lägger

massan 0.4 respektive 0.6 på värdena 0.2 respektive 0.5 för Θ . Detta betyder att vi har den simultana fördelningen

$$p_{X,\Theta}(x, \theta) = P(X = x; \Theta = \theta) = P(X = x | \Theta = \theta)P(\Theta = \theta)$$

som

$$p_{X,\Theta}(x, 0.2) = \binom{5}{x} 0.2^x (1 - 0.2)^{5-x} \cdot 0.4, \quad x = 0, 1, 2, 3, 4, 5$$

och

$$p_{X,\Theta}(x, 0.5) = \binom{5}{x} 0.5^x (1 - 0.5)^{5-x} \cdot 0.6, \quad x = 0, 1, 2, 3, 4, 5$$

Om man nu observerar ett utfall x av X , t ex $X = 4$ så är man intresserad av fördelningen för Θ givet denna observation av $X = 4$. Denna fördelning för parametern Θ givet observationen är a-posteriori-fördelningen för Θ givet $X = 4$. Notera att likelihoodfunktionen anger sannolikheten för observationer givet utfallen av parametern Θ , medan a-posteriorifördelningen anger fördelningen för parametern givet observationen. Denna ”vändning” av betingningen är precis vad Bayes’ sats tillhandahåller. Vi får då (med $X = 4$) att

$$\begin{aligned} P(\Theta = 0.2 | X = 4) &= \\ &= \frac{P(X = 4 | \Theta = 0.2)P(\Theta = 0.2)}{P(X = 4 | \Theta = 0.2)P(\Theta = 0.2) + P(X = 4 | \Theta = 0.5)P(\Theta = 0.5)} = \\ &= \frac{\binom{5}{4} 0.2^4 (1 - 0.2)^{5-4} \cdot 0.4}{\binom{5}{4} 0.2^4 (1 - 0.2)^{5-4} \cdot 0.4 + \binom{5}{4} 0.5^4 (1 - 0.5)^{5-4} \cdot 0.6} = 0.0266 \end{aligned}$$

och $P(\Theta = 0.5 | X = 4) = 0.9734$, dvs vår a-prioriuppfattning om sannolikheterna för 0.2 och 0.5 som parametervärden förändras från 0.4 resp 0.6 till att bli 0.0266 resp 0.9734 då vi observerar $X = 4$, som ju tyder på att Θ nog är ganska stort. \square

Exempel 1.6 *Binomialfördelning – allmän diskret fördelning för Θ*

På motsvarande sätt skulle det fungera om vi hade en (diskret) a-priorifördelning för Θ på värden $\theta_1, \theta_2, \theta_3, \dots$ beskriven av sannolikhetsfördelningen $P(\Theta = \theta_i)$, $i = 1, 2, \dots$. Om vi sen har att X är $\text{Bin}(n, \theta_i)$ så betyder det alltså att X med sannolikhet $P(\Theta = \theta_i)$ är $\text{Bin}(n, \theta_i)$, dvs att $P(X = x | \Theta = \theta_i) = \binom{n}{x} \theta_i^x (1 - \theta_i)^{n-x}$ - notera detta är helt enkelt likelihoodfunktionen då $X = x$ för $\Theta = \theta_i$. Om vi nu observerar $X = x$ så ger detta för Θ a-posteriorifördelningen

$$P(\Theta = \theta_{i_0} | X = x) = \frac{P(X = x | \Theta = \theta_{i_0})P(\Theta = \theta_{i_0})}{\sum_{i=1}^{\infty} P(X = x | \Theta = \theta_i)P(\Theta = \theta_i)}.$$

Detta går naturligtvis utmärkt att räkna ut i en konkret situation med t ex Matlab. Man genererar en vektor `theta` som innehåller de olika θ_i -värdena och en vektor `ptheta` som innehåller $P(\Theta = \theta_i)$ på gittret definierat av `theta` samt en vektor `pxtheta` som innehåller värdet av $\binom{n}{x} \theta_i^x (1 - \theta_i)^{n-x}$ för de olika θ_i :na (och det observerade värdet x) samt skaffar sig vektorn

post=pxtheta.*ptheta/sum(pxtheta.*ptheta)

som då kommer att innehålla a-posteriori-sannolikheterna för de olika θ_i -na i **theta**. Notera att operationen `.*` utför multiplikationen av de två vektorna koordinatvis. Vektorn **ptheta**:s i -te komponent behöver bara vara proportionell mot $P(\Theta = \theta_i)$ och vektorn **ptheta** alltså inte behöver summera sig till 1 i den numeriska kalkylen eftersom de förekommer i både täljare och nämnare!□

Exempel 1.7 *Binomialfördelning – kontinuerlig fördelning för Θ*

På samma sätt skulle en täthet $f_\Theta(\theta)$, $0 \leq \theta \leq 1$ som a-prioritäthet för Θ ge upphov till a-posteriori-tätheten

$$\begin{aligned} f_{\Theta|X=x}(\theta) &= \frac{P(X = x|\Theta = \theta)f_\Theta(\theta)}{p_X(x)} = \frac{P(X = x|\Theta = \theta)f_\Theta(\theta)}{\int_0^1 P(X = x|\Theta = u)f_\Theta(u)du} = \\ &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}f_\Theta(\theta)}{\int_0^1 \binom{n}{x}u^x(1-u)^{n-x}f_\Theta(u)du} \end{aligned}$$

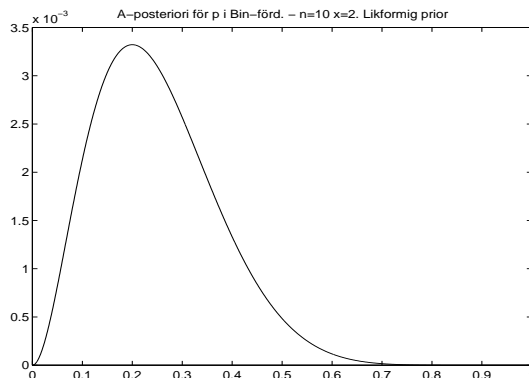
Nämnamren i detta uttryck, alltså

$$p_X(x) = P(X = x) = \int_0^1 P(X = x|\Theta = u)f_\Theta(u)du$$

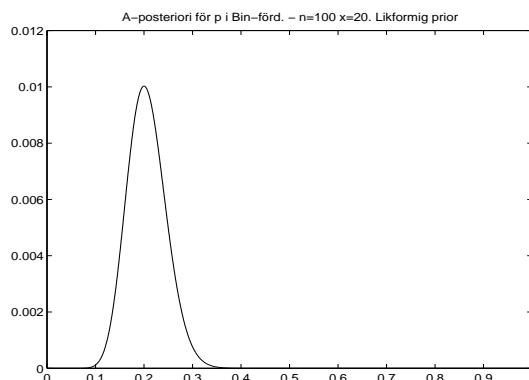
är ofta rätt besvärlig att räkna ut för en given täthet, dvs för en given f_Θ -funktion. Därför väljer man gärna denna täthet på ett sådant sätt att integralen går lätt att räkna ut. Speciellt går det bra om man låter Θ ha en k Beta-fördelning, som gör att a-posteriori-fördelningen också blir en (annan) Beta-fördelning. Mer om detta i samband med s k konjugerade fördelningar som beskrivs i avsnitt 1.8 på sid 10.

Observera dock det skulle gå bra numeriskt att utföra denna integration med hjälp av Matlab genom att generera en ”diskretiserad” variant av täthetsfunktionen f_Θ . Om man alltså inte är ute efter ett slutet analytiskt uttryck för a-posteriori-fördelningen utan nöjer sig med en numeriskt given funktion är intergrationsbekymret inte så stort. Man approximerar tätheten med en diskret fördelning och applicerar sedan (t ex i Matlab) Bayes’ sats för diskret fördelning hos Θ . Ett annat alternativ är att utföra en numerisk integration.

Som exempel kan man ta situationen att x är ett utfall av X som är $\text{Bin}(n, \theta)$, dvs $p_{X|\Theta=\theta}(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. Figurerna 1.2 och 1.3 visar a-posteriori-fördelningarna för Θ om $n = 10$, $x = 2$ respektive $n = 100$, $x = 20$ då man haft likformig fördelning på $[0, 1]$ för Θ , dvs $f_\Theta(\theta) = 1$, $0 \leq \theta \leq 1$. Som jämförelse kan man även i figur 1.4 se hur liten skillnaden blir om man som a-priori-fördelning för Θ haft en linjärt växande täthet, dvs $f_\Theta(\theta) = 2\theta$, $0 \leq \theta \leq 1$. □



Figur 1.2: A-posteriori-fördelning för p i binomialfördelning $\text{Bin}(10, p)$ när $x = 2$ observerats. Likformig a-priori-fördelning.



Figur 1.3: A-posteriori-fördelning för p i binomialfördelning $\text{Bin}(100, p)$ när $x = 20$ observerats. Likformig a-priori-fördelning.

1.7 Binomialfördelning

Om man vill räkna exakt i Binomialfördelningssituationen blir det lite besvärligare. Med rätt val av a-priori-fördelning kan det hela bli lite enklare.

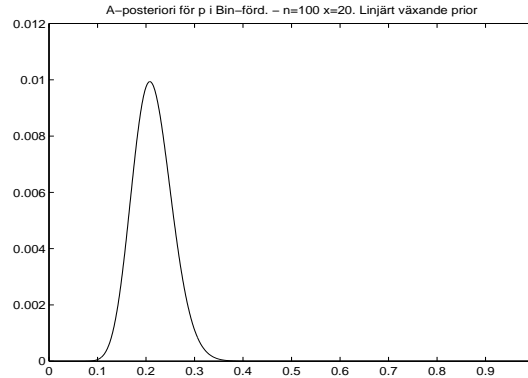
Exempel 1.8 Likformig fördelning

Om man tar $f_{\Theta}(\theta) = 1$, $0 \leq \theta \leq 1$ får man (efter enkla kalkyler) att

$$f_{\Theta|X=x}(\theta) = \frac{(n+1)!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \text{ för } 0 \leq \theta \leq 1$$

en s k Beta($x+1, n-x+1$)-fördelning. \square

Detta utgör ett exempel på en lämpligt vald a-priori-fördelning som gör a-posteriori-fördelningen enkel. Vi inför en klass av fördelningar som visar sig vara lämplig klass av a-priori-fördelningar för Binomialfördelningssituationen.



Figur 1.4: A-posteriori-fördelning för p i binomialfördelning $\text{Bin}(100, p)$ när $x = 20$ observerats. Linjärt växande a-priori-fördelning.

1.7.1 Beta-fördelning

Definition 1.1 En stokastisk variabel Θ säges ha $\text{Beta}(r, s)$ -fördelning om

$$f_{\Theta}(u) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} u^{r-1}(1-u)^{s-1} \text{ för } 0 \leq u \leq 1$$

där $r, s > 0$. $\text{Beta}(r, s)$ -fördelningen har väntevärdet $r/(r+s)$.

Här är $\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$ – speciellt är $\Gamma(n) = (n-1)!$ (som erhålls med hjälp av partiell integration). Γ -funktionen kan ses som en generalisering av ”fakultet” till även icke-heltal. Notera dock att Γ -funktionen är ”förskjuten” en enhet i förhållande till fakultet. Den beräknas i Matlab med funktionen `gamma`.

Ex är $\text{Beta}(1,1)$ -fördelningen samma sak som likformig fördelning på intervallet $[0, 1]$.

Eftersom man har totalmassan 1 i de olika Beta -fördelningarna ser man att

$$\int_0^1 u^{r-1}(1-u)^{s-1} du = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

eller mer egentligt att koefficienten i Beta -fördelningens täthet valts så att totalmassan är 1. Av detta följer också att $\text{Beta}(r, s)$ -fördelningen har väntevärdet $r/(r+s)$.

1.7.2 Beta-fördelning som a-priori-fördelning

Om nu vi för $\text{Bin}(n, \theta)$ -fördelade mätdata låter θ vara ett utfall av Θ som är $\text{Beta}(a, b)$, dvs vi låter Θ ha a-priori-fördelningen $\text{Beta}(a, b)$ så får vi a-posteriori-fördelningen

$$f_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(x)f_{\Theta}(\theta)}{p_X(x)} = \frac{p_{X|\Theta=\theta}(x)f_{\Theta}(\theta)}{\int_0^1 p_{X|\Theta=u}(x)f_{\Theta}(u)du}$$

Vi ser att

$$\begin{aligned} p_X(x) &= \int_0^1 p_{X|\Theta=u}(x) f_{\Theta}(u) du = \\ &= \int_0^1 \binom{n}{x} u^x (1-u)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1} du = \\ &= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 u^{x+a-1} (1-u)^{n-x+b-1} du \end{aligned}$$

Den sista integralen ser man är

$$\frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(x+a+n-x+b)} = \frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(a+n+b)}$$

som ger att

$$p_X(x) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(a+n+b)}$$

Detta ger då att

$$f_{\Theta|X=x}(\theta) = \frac{\Gamma(a+n+b)}{\Gamma(x+a)\Gamma(n-x+b)} \cdot \theta^{x+a-1} (1-\theta)^{n-x+b-1}$$

dvs att a-posteriori-fördelningen för Θ (givet $X=x$) är $\text{Beta}(x+a, n-x+b)$. Speciellt ser vi att om $a=b=1$ (dvs att vi har en likformig a-priori-fördelning på $[0, 1]$) så blir a-posteriori-fördelningen en $\text{Beta}(x+1, n-x+1)$ -fördelning.

Faktum är att dessa (rätt besvärliga) integrationer kan undvikas om man noterar att i

$$f_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(x) f_{\Theta}(\theta)}{p_X(x)}$$

utgör nämnaren bara är en normering så att $f_{\Theta|X=x}(\theta)$ verkligen blir en sannolikhetstäthet, dvs har totalmassa (totalintegral)=1.

Vi noterar att

$$\begin{aligned} f_{\Theta|X=x}(\theta) &= \frac{p_{X|\Theta=\theta}(x) f_{\Theta}(\theta)}{p_X(x)} = \\ &= \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x} \Gamma(a+b) \theta^{a-1} (1-\theta)^{b-1}}{\Gamma(a)\Gamma(b) p_X(x)} \propto \theta^{x+a-1} (1-\theta)^{n-x+b-1}. \end{aligned}$$

där vi använder symbolen \propto för "proportionell mot". Som funktion av θ är detta samma uttryck som för $\text{Beta}(x+a, n-x+b)$ -fördelningen och alltså måste normeringen vara sådan att vi får just $\text{Beta}(x+a, n-x+b)$ -fördelningen.

Denna typ av resonemang är mycket praktiskt och fungerar ju eftersom vi bara är intresserade av beroendet av θ eftersom vi är ute efter a-posteriori-fördelningen för Θ dvs att få ett funktionellt uttryck i θ . Allt övrigt i uttrycket är enbart att betrakta som en normering eftersom det är en sannolikhetsfördelning (sedd som funktion av θ) vi är ute efter!

1.8 Allmänt om konjugerade fördelningar

Vad man ofta vill göra är att välja a-priori-fördelningen som en lämpligt vald parametrisk familj som gör att a-posteriori-fördelningen tillhör samma parametriska familj.

Ovan såg vi att Beta-fördelningarna var sådana att om man hade en $\text{Beta}(a, b)$ -fördelning som a-priori-fördelning för Θ i Binomialfördelningssituationen, där X var $\text{Bin}(n, \Theta)$, så blev a-posteriori-fördelningen, då vi observerat $X = x$ en $\text{Beta}(x + a, n - x + b)$ -fördelning.

1.8.1 Definition av konjugerad fördelning

Definition 1.2 *Konjugerad fördelning*

Låt familjen av a-priori-fördelningar $f_{\Theta}(\theta) = f_{\Theta}(\theta; \boldsymbol{\alpha})$ bero av en parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)$. Antag att a-posteriori-fördelningen $f_{\Theta|X=\mathbf{x}}$ tillhör samma parametriska familj, dvs att

$$f_{\Theta|X=\mathbf{x}}(\theta) = f_{\Theta}(\theta; \boldsymbol{\beta})$$

för ett annat värde på parametern

$$\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\alpha}, \mathbf{x}) = (\beta_1, \beta_2, \dots, \beta_p).$$

Då säges familjen $f_{\Theta}(\theta, \boldsymbol{\alpha})$ av tänkbara a-priori-fördelningar vara konjugerad till familjen av fördelningar $F_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\theta})$ för vektorn \mathbf{X} □

Det fiffiga med konjugerade fördelningar är alltså att a-priori-fördelningen och a-posteriori-fördelningen kan sammanfattas med hjälp av $\boldsymbol{\alpha}$ respektive $\boldsymbol{\beta}$ där alltså $\boldsymbol{\beta}$ typiskt beror av våra observationer \mathbf{x} och $\boldsymbol{\alpha}$.

1.8.2 Exponentialfördelning

Om vi låter Λ ha en $\Gamma(a, b)$ -fördelning och låter $X|\Lambda = \lambda$ vara $\text{Exp}(1/\lambda)$ dvs att X har "felintensiteten" λ där λ är ett utfall av Λ som är $\Gamma(a, b)$ så kommer vi nedan att se att $\Lambda|X = x$ kommer att ha en $\Gamma(a + 1, b + x)$ -fördelning. Detta betyder alltså att *Gamma*-fördelningarna är konjugerade till familjen av exponentialfördelningar. Vi har nämligen när Λ är $\Gamma(a, b)$

$$f_{\Lambda}(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

och (eftersom $X|\Lambda = \lambda$ var $\text{Exp}(1/\lambda)$)

$$f_{X|\Lambda=\lambda}(x) = \lambda e^{-\lambda x}.$$

Detta ger att

$$f_{\Lambda|X=x}(\lambda) = \frac{f_{X|\Lambda=\lambda}(x) f_{\Lambda}(\lambda)}{f_X(x)}$$

(vi håller bara reda på λ -beroendet!)

$$\begin{aligned} f_{\Lambda|X=x}(\lambda) &\propto \lambda e^{-\lambda x} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \propto \\ &\propto \lambda^a e^{-(b+x)\lambda} = \lambda^{(a+1)-1} e^{-(b+x)\lambda} \end{aligned}$$

som (sett som funktion av λ) alltså säger att $\Lambda|X = x$ måste vara $\Gamma(a+1, b+x)$.

Om vi alltså har a-priori-fördelningen $\Gamma(a, b)$ för Λ och observerar $X = x$ så blir $\Lambda|X = x$ en $\Gamma(a+1, b+x)$ -fördelning.

1.8.3 Normalfördelning

Antag att vi har X_i som är $N(m, \sigma_0^2)$ med σ_0 känd. Om m är ett utfall av M som har en a-priori-fördelning som är $N(\mu, s^2)$ så får vi

$$\begin{aligned} f_{M|\mathbf{X}=\mathbf{x}}(m) &\propto L(m; \mathbf{x}) f_M(m) = \left(\prod_{i=1}^n f_{X_i}(x_i|M=m) \right) f_M(m) \propto \\ &\propto \exp\left(-\frac{\sum_{i=1}^n (x_i - m)^2}{2\sigma_0^2}\right) \exp\left(-\frac{(m - \mu)^2}{2s^2}\right) \\ &\propto \exp\left(-\frac{1}{2} \left(\frac{n}{\sigma_0^2} + \frac{1}{s^2}\right) \left(m - \frac{n\bar{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}\right)^2\right) \end{aligned}$$

genom triviala med tidsödande kvadratkompletteringar. Detta uttryck kan, som funktion av m , identifieras med tätheten för en normalfördelning bortsett från proportionalitetsfaktorer (som inte innehåller m) nämligen

$$N\left(\frac{n\bar{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right)$$

och detta måste alltså vara a-posteriori-fördelningen för M dvs fördelningen för

$$(M|\mathbf{X} = \mathbf{x}) = (M|X_1 = x_1; X_2 = x_2; \dots; X_n = x_n)$$

Slutsatsen är alltså att i denna situation är normalfördelningarna sin egen konjugerade fördelning.

Detta harvande med konjugerade fördelningar kan ha ett visst intresse för sig, men utgör också en motivering för hur praktisk Markov Chain Monte Carlo-simulering är. I sådan simulering slipper man alla dessa komplicerade kalkyler rörande konjugerade fördelningar utan man simulerar fram a-posteriori-fördelningen med förvånansvärd enkelhet. För att få tillbörlig respekt för hur praktiskt detta är har vi därför studerat hur komplicerat det är att reda ut detta med konjugerade fördelningar.

1.8.4 A-posteriori-fördelning som ny a-priori

När man fått en observation och kombinerar denna med a-priori-fördelningen och därvid erhåller en a-posteriori-fördelning så kan ju tänkas göra ytterligare en observation. Om man då som a-priori-fördelning för denna andra observation tar a-posteriori-fördelningen efter första observationen så skulle man erhålla en ny a-posteriori-fördelning. Denna blir precis samma som om man med den ursprungliga a-priori-fördelningen gjort två oberoende observationer och då beräknat a-posteriori-fördelningen givet båda observationerna. I denna mening kan man alltså se a-posteriori-fördelningen som en successiv uppdatering av a-priori-fördelningen med de erhållna observationerna.

Sats 1.2 Låt observationerna x_1, x_2, \dots, x_n vara oberoende likafördelade givet $\Theta = \theta$. Vi har a-priori-fördelningen $f_\Theta(\theta)$ för Θ . Man får samma a-posteriori-fördelning med följande två betraktelsesätt:

1) Vi beräknar a-posteriori-fördelningen för Θ givet första observationen x_1 och använder denna a-posteriori-fördelning som a-priori-fördelning för Θ då data nr 2 x_2 insamlas. A-posteriori-fördelningen givet detta x_2 används sedan som a-priori-fördelningen vid analys med hjälp av x_3 , osv ända till sista observationen x_n .

2) Vi beräknar a-posteriori-fördelningen givet hela observationsvektorn x_1, x_2, \dots, x_n .

Bevis:

Låt $f_1(\theta|x_1)$ vara a-posteriori-fördelningen efter att vi observerat x_1 och $f_2(\theta|x_2)$ vara a-posteriori-fördelningen efter man observerat x_2 och har $f_1(\theta|x_1)$ som a-priori-fördelning. På samma sätt låter vi $f_i(\theta|x_i)$ vara a-posteriori-fördelningen efter att x_i observerats för $i = 1, 2, \dots, n$ och vi har $f_{i-1}(\theta|x_{i-1})$ som a-priori-fördelning. Vi får då

$$f_1(\theta|x_1) \propto f_\Theta(\theta)f_{X_1}(x_1|\theta)$$

och

$$f_2(\theta|x_2) \propto f_1(\theta|x_1)f_{X_2}(x_2|\theta) \propto f_\Theta(\theta)f_{X_1}(x_1|\theta)f_{X_2}(x_2|\theta)$$

och på samma sätt erhålls

$$f_n(\theta|x_n) \propto f_{n-1}(\theta|x_{n-1})f_{X_n}(x_n|\theta) \propto \dots \propto f_\Theta(\theta)f_{X_1}(x_1|\theta)f_{X_2}(x_2|\theta) \dots f_{X_n}(x_n|\theta).$$

Å andra sidan gäller att

$$\begin{aligned} f_\Theta(\theta|x_1, x_2, \dots, x_n) &\propto f_\Theta(\theta)f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n|\theta) = \\ &= \text{oberoendet} = f_\Theta(\theta) \prod_{i=1}^n f_{X_i}(x_i|\theta) \end{aligned}$$

och alltså ger båda metoderna samma resultat. Faktum är att i beviset behövs inte ens antagandet att X_i är likafördelade, men däremot måste de vara betingat oberoende givet θ . \square

1.9 Bayesianisk inferens

För en Bayesian uttrycker a-posteriori-fördelningen all information om vår uppfattning om Θ när vi tagit hänsyn till våra mätdata och a-priori-fördelningen. Man kan även sammanfatta denna fördelning med t ex väntevärdet, dvs $E(\Theta|X = x)$ eller medianen eller det värde (moden) som ger maximum i tätheten för a-posteriori-fördelningen. Oftast används just det betingade väntevärdet $E(\Theta|X = x)$, dvs väntevärdet i a-posteriori-fördelningen och denna skattning brukar kallas Bayes-skattningen $\hat{\theta}_{\text{Bayes}}$. Man kan följande notera att om man har en likformig fördelning (dvs $f_{\Theta}(\theta) = \text{konstant}$) som a-priori-fördelning så är det just Maximum-Likelihood-skattningen som maximerar a-posteriori-tätheten! Den vanliga skattningsteorin som bygger på Maximum-Likelihoodmetoden är specialfallet att ha likformig a-priori-fördelning och skatta med det värde som maximerar a-posteriori-tätheten. Klassisk teori kan därför ses som ett specialfall av Bayesianisk metodik.

Om man i binomialfördelningssituationen har en $Beta(r, s)$ -fördelning som a-priori-fördelning för Θ och observerar utfallet $X = x$ av $Bin(n, \theta)$ så såg vi att a-posteriori-fördelningen för Θ blev en $Beta(x + r, n - x + s)$ -fördelning. Denna a-posteriori-fördelning har ett väntevärde som är $(r + x)/(r + s + n)$ och detta utgör då Bayes-skattningen $\hat{\theta}_{\text{Bayes}}$ av Θ . Man kan för övrigt notera att

$$\hat{\theta}_{\text{Bayes}} = \frac{r + x}{r + s + n} = \frac{n}{r + s + n} \cdot \frac{x}{n} + \frac{r + s}{r + s + n} \cdot \frac{r}{r + s}$$

där x/n är den traditionella ML-skattningen av θ och $r/(r + s)$ är väntevärdet av a-priori-fördelningen $Beta(r, s)$. Bayes-skattningen av θ är alltså ett viktat medelvärde av ML-skattningen och a-priori-skattningen (dvs väntevärdet) där vikterna är $n/(r + s + n)$ respektive $(r + s)/(r + s + n)$. Man ser att om bara n är stort läggs nästan all vikt på skattningen som härrör sig ur data (dvs x/n).

1.9.1 Bayesianiska konfidensintervall

Man kan också göra Bayesianiska konfidensintervall (credibility intervals) för parametern genom att ta punkter i a-posteriori-fördelningen mellan vilka det ligger t ex 95% av massan. Det finns även andra alternativ som att välja de regioner där a-posteriori-tätheten är stor och att avpassa området så att totala sannolikhetsmassan blir t ex 95%. Detta kan innebära att "konfidensintervallet" blir uppdelat i flera disjunkta delar om a-posteriori-fördelningen inte är unimodal (dvs har mer än en topp).

Det numeriska intervallet $(a(\mathbf{x}), b(\mathbf{x})) = (a(x_1, x_2, \dots, x_n), b(x_1, x_2, \dots, x_n))$ är alltså ett konfidensintervall för Θ med konfidensgrad $1 - \alpha$ om

$$P(a(\mathbf{x}) < \Theta < b(\mathbf{x}) | \mathbf{X} = \mathbf{x}) = \int_{a(\mathbf{x})}^{b(\mathbf{x})} f_{\Theta}(\theta | \mathbf{X} = \mathbf{x}) d\theta = 1 - \alpha$$

dvs i ord just två punkter $a(\mathbf{x})$ och $b(\mathbf{x})$ mellan vilka det finns massan $1 - \alpha$ i a-posteriori-fördelningen för Θ .

Exempel 1.9 *Bayesianskt konfidensintervall för normalfördelning*

Vi såg tidigare att om X_i är $N(m, \sigma_0^2)$ där σ_0 är känd och m var ett utfall av M och vi har $N(\mu, s^2)$ som a-priori-fördelning för M så blev a-posteriori-fördelningen för M givet observationerna x_1, x_2, \dots, x_n

$$N\left(\frac{n\bar{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right).$$

Om vi låter $s \rightarrow \infty$ vilket svarar mot att vi väljer likformig fördelning för m som a-priori-fördelning ser vi att a-posteriori-fördelningen för M blir $N(\bar{x}, \sigma_0^2/n)$ och det 95%-iga Bayesianska konfidensintervallet för m blir $\bar{x} \pm \lambda_{0.025}\sigma_0/\sqrt{n}$ dvs det helt sedvanliga intervallet. Tolkningen av detta intervall är alltså att mellan de två gränserna finns 95% av massan i a-posteriori-fördelningen, dvs i överensstämmelse med den "naiva" (och felaktiga) tolkningen av konfidensintervall som ett intervall som med sannolikheten 95% innehåller parametern. \square

1.9.2 Prediktiv fördelning

A-posteriori-fördelningen är ju ett sätt att beskriva fördelningen för parametern Θ då man gjort observationerna \mathbf{x} . Ett praktiskt sätt att utnyttja denna a-posteriori-fördelning är att studera fördelningen för en ny observation X_0 . Man definierar den prediktiva fördelningen som

$$f_{X_0}(x_0|\mathbf{X} = \mathbf{x}) = \int_{\Omega} f_{X_0}(x_0|\Theta = \theta) f_{\Theta}(\theta|\mathbf{X} = \mathbf{x}) d\theta.$$

Denna integral innebär alltså att vi som fördelning för X_0 tar en hopviktning av tätheten för X_0 för olika θ -värden med hänsyn till hur sannolika dessa θ -värden är enligt a-posteriori-fördelningen.

Exempel 1.10 Låt N vara $\text{Po}(\lambda)$ då $\Lambda = \lambda$ där Λ är $\Gamma(a, b)$. Vi observerar $N = n$. Detta betyder alltså att

$$f_{\Lambda}(\lambda) = \frac{\lambda^{a-1} b^a}{\Gamma(a)} e^{-b\lambda}, \quad \lambda > 0$$

och

$$P(N = n|\Lambda = \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, 2, \dots$$

Man får då att a-posteriori-fördelningen för Λ (dvs givet $N = n$) blir (där vi bara håller reda på λ -beroendet)

$$\begin{aligned} f_{\Lambda}(\lambda|N = n) &\propto f_{\Lambda}(\lambda) P(N = n|\Lambda = \lambda) \propto \\ &\propto \lambda^{a-1} e^{-b\lambda} \cdot \lambda^n e^{-\lambda} = \lambda^{n+a-1} e^{-\lambda(b+1)} \end{aligned}$$

vilket alltså innebär att $\Lambda|N = n$ är $\Gamma(n + a, b + 1)$ -fördelad dvs att

$$f_{\Lambda}(\lambda|N = n) = \frac{\lambda^{n+a-1} (b+1)^{n+a}}{\Gamma(n+a)} e^{-(b+1)\lambda}, \quad \lambda > 0.$$

Vi har förresten därigenom i förbigående visat att Γ -fördelningarna utgör konjugerade fördelningar till Poisson-fördelningen.

Vi får då den prediktiva fördelningen för en ny observation N_0 till

$$\begin{aligned} P(N_0 = n_0 | N = n) &= \int_0^\infty P(N_0 = n_0 | \Lambda = \lambda) f_\Lambda(\lambda | N = n) d\lambda = \\ &= \int_0^\infty \frac{\lambda^{n_0}}{n_0!} e^{-\lambda} \cdot \frac{\lambda^{n+a-1} (b+1)^{n+a}}{\Gamma(n+a)} e^{-(b+1)\lambda} d\lambda. \end{aligned}$$

Detta ger efter en del enkla räkningar att

$$P(N_0 = n_0 | N = n) = \frac{(b+1)^{n+a}}{\Gamma(n+a)} \cdot \frac{1}{n_0!} \cdot \frac{\Gamma(n_0 + n + a)}{(b+2)^{n_0+n+a}}, \quad n_0 = 0, 1, 2, \dots$$

Man kan notera att detta fungerar även om $N = 0$, dvs vi kan få fram den prediktiva fördelningen även då ingen händelse observerats. Vi får för $n = 0$ att

$$P(N_0 = n_0 | N = 0) = \frac{(b+1)^a}{\Gamma(a)} \cdot \frac{1}{n_0!} \cdot \frac{\Gamma(n_0 + a)}{(b+2)^{n_0+a}}$$

som för $a = 1$ och $b = 1$ ger

$$P(N_0 = n_0 | N = 0) = \frac{2}{3^{n_0+1}}, \quad n_0 = 0, 1, 2, \dots$$

dvs t ex $P(N_0 = 0 | N = 0) = 2/3$ och $P(N_0 = 1 | N = 0) = 2/9$. Detta betyder att $N_0 | N = 0$ i detta fall är Geo(2/3)-fördelningen.

Man kan jämföra detta med de problem en icke-Bayesian skulle råka i om han observerar $N = 0$ och alltså skattar λ med $\hat{\lambda} = 0$. Ingen tror väl att andra värden än 0 är omöjliga?!

1.10 Hur väljs a-priori-fördelning?

Detta att man som Bayesian måste välja en a-priori-fördelning är både en svaghet och en styrka för de Bayesianiska metoderna. Oftast känns valet av a-priori-fördelning ganska godtyckligt, men, glädjande nog, spelar valet förvånansvärt liten roll i de flesta situationer. Det kan också uppfattas som en styrka att man kan väga in "förutfattade meningar" om vilka parametervärden som är troligast i sin a-priori-fördelning.

Oftast har valet av a-priori-fördelning dock mindre styrts av vad man haft anledning att tro om parametern än av att kalkylerna skall bli hanterliga – speciellt har Bayesianer känt sig bundna av att hålla sig till konjugerade fördelningar.

1.10.1 Icke-informativa a-priori-fördelningar

En icke-informativ a-priori-fördelning för Θ är i princip en likformig fördelning för Θ . Om de tänkbara värdena för Θ är ett ändligt intervall uppstår inga matematiska problem. I binomialfördelningssituationen är alltså likformig fördelning på intervallet $(0, 1)$ en sådan icke-informativ a-priori-fördelning. Skulle de tänkbara värdena för Θ vara ett oändligt intervall kan matematiska problem uppstå eftersom man inte kan definiera en likformig sannolikhetsfördelning på ett oändligt intervall. I och för sig behöver inte a-priori-fördelningen vara normerad så ibland fungerar matematiskt en likformig "fördelning" (oändlig totalmassa) på ett oändligt intervall men man kan stöta på problem i form av att även a-posteriori-fördelningen inte går att normera.

Ibland kan man "fuska" och i stället ta likformig fördelning på ett stort (men ändligt) intervall $(0, M)$ där M valts så stor att alla "realistiska" värden på Θ finns i intervallet. Om man t ex tror att realistiska värden på felintensiteten Λ är ca 1 kan man välja a-priori-fördelning som likformig fördelning på intervallet $(0, 1000)$. Om man inte räknar exakt utan har tänkt sig att utföra en numerisk kalkyl i t ex Matlab så kommer man ändå att tvingas inskränka sig till ett ändligt intervall. En annan vanlig teknik är att välja a-priori-fördelningen som en sannolikhetsfördelning som är "nästan konstant" – t ex kan man välja $N(0, 1000)$ -fördelningen eller $\Gamma(10^{-3}, 10^{-3})$ -fördelningen beroende på om man vill ha "nästan" likformig fördelning på hela x -axeln eller på positiva x -axeln.

Hela begreppet icke-informativ a-priori-fördelning är dock rätt skumt eftersom de inte är invarianta under omparametriseringar. Om man t ex har likformig fördelning för standardavvikelsen σ eller för variansen σ^2 så kommer detta att spela en viss roll. På samma sätt spelar det roll om vi använder oss av felintensitet λ eller medellivslängd $m = 1/\lambda$ som den parameter vi studerar. Båda dessa val av parametrering kan ju anses vara "naturliga".

En annan vanlig invändning från Bayesianer om godtyckligheten i a-priori-fördelningar är att en icke-Bayesian som baserar sin inferens på t ex Maximum Likelihood-metoden "egentligen" är en Bayesian som har likformig fördelning på parameter-rummet som a-priori-fördelning. Att detta i någon mån är sant inser man av att om man har konstant a-priori-fördelning så blir a-posteriori-fördelningen direkt proportionell mot likelihood-funktionen. Att använda sig av ML-skattningen, dvs att maximera likelihoodfunktionen är inget annat än att som Bayesian använda moden (dvs maximum) i a-posteriori-fördelningen som skattning.