Föreläsning 06

Statistisk inlärning och dataanalys (Kungliga Tekniska Högskolan)

Föreläsning 6

6.1 A priori och a posteriori

Som vi diskuterade i föregående kurs finns det två huvudsakliga ramverk för att skatta parametrar inom statistiken: den frekventistiska (eller klassiska) och den bayesianska. Hittills har vi diskuterat den frekventistiska metoden, vilken låter oss skatta parametrar med hjälp av exempelvis momentmetoden och ML-metoden. Dessa punktskattningar kan sedan utvärderas genom att jämföra deras riskfunktioner för olika parametervärden.

Det som sammanbinder de frekventistiska metoderna är att parametern $\boldsymbol{\theta}$ anses vara fixt, men okänd (inom en viss mängd). Med andra ord så antar vi att det finns en underliggande parameter som använts för att generera den data som observeras. Målet är då att försöka bestämma denna okända parameter så väl som möjligt.

Även om de frekventistiska metoderna fungerar bra i många sammanhang är de mindre användbara när vi anser att $\boldsymbol{\theta}$ är ett utfall av en stokastisk variabel $\boldsymbol{\Theta}$ med en viss fördelning. Detta kan ske i olika situationer, men det vanligaste är att vi vill kvantifiera information om vilka värden på parametern som är mer sannolika än andra. Denna information är relevant vid två tillfällen: innan vi observerar vår data (a priori) och efter vi observerar datat (a posteriori).

Informationen a priori kan bestå av subjektiva uppfattningar om vilka värden vi tror parametern borde ta. Det kan också grunda sig i tidigare experiment som visat att parametern tar vissa värden oftare än andra. Med andra ord så behöver denna information nödvändigtvis vara subjektiv.

Informationen om parametern a posteriori kombinerar informationen som vi har innan observation (a priori) samt observationerna i sig. Detta sker med hjälp av Bayes sats.

6.1.1 Apriorifördelning och aposteriorifördelning

I denna föreläsning antar vi att vi endast har en parameter att skatta, dvs. θ är en skalär. Senare kommer vi att generalisera detta till vektorvärda parametrar $\boldsymbol{\theta}$. Vi har som regel en parameter Θ med täthetsfunktion (eller sannolikhetsfunktion) givet av $f_\Theta(\theta)$. Denna fördelning kallas för *apriorifördelningen*. Givet att Θ tar ett visst värde θ, dvs. att $\Theta = \theta$, så har vi en *datafördelning* som ges av täthetsfunktionen (eller sannolikhetsfunktionen) $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta)$. Dessa är de enda ingredienserna i en bayesiansk modell. Utifrån apriorifördelningen och datafördelningen kan vi använda observationer $\boldsymbol{x}$ för att dra olika slutsatser om parametern Θ. Här utmärker sig det bayesianska synsättet i och med att det egentligen bara finns *en* metod för att göra inferens, till skillnad från i det frekventistiska ramverket där olika val (momentmetoden, ML-metoden, osv.) ger olika resultat och vi inte alltid vet vilket som bäst speglar verkligheten. Denna entydighet kan vara en fördel, men i vissa fall också en nackdel då denna enda metod kan leda till svåra beräkningar. I de situationer som vi betraktar i denna kurs kommer dock beräkningarna visa sig hanterbara.

Metoden som vi syftar på här är att ta fram *aposterioriförderlningen*, dvs. den betingade fördelningen för Θ givet observationen $\boldsymbol{X} = \boldsymbol{x}$. Bayes regel säger att aposteriorifördelningens täthetsfunktion (eller sannolikhetsfunktion) ges av

$$f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) = \frac{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})}.$$

Marginalfördelningen av $\boldsymbol{X}$ (som ges av $f_{\boldsymbol{X}}(\boldsymbol{x})$) kallas ibland för *aprioriprediktiva fördelningen* då den motsvarar hur vi förväntar oss att datan är fördelad innan vi observerar den. Den beräknas oftast genom lagen om total sannolikhet, vilket ger

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_\Omega f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta) d\theta$$

i det kontinuerliga fallet.

Vi repeterar först ett exempel från föregående kurs.

Exempel 6.1. Låt $X \mid \Theta = \theta \sim \text{Bin}(n, \theta)$ och $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$. Då har vi

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) &= \left(\binom{n}{x} \theta^x (1-\theta)^{n-x} \right) \left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \right) \\ &= \binom{n}{x} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0+x-1}(1-\theta)^{\beta_0+n-x-1}. \end{aligned}$$

Detta ger aprioriprediktiva fördelningen

$$\begin{aligned} f_X(x) &= \int_0^1 f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta \\ &= \binom{n}{x} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \int_0^1 \theta^{\alpha_0+x-1}(1-\theta)^{\beta_0+n-x-1} d\theta \\ &= \binom{n}{x} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0+x)\Gamma(\beta_0+n-x)}{\Gamma(\alpha_0+\beta_0+n)} \int_0^1 \frac{\Gamma(\alpha_0+\beta_0+n)}{\Gamma(\alpha_0+x)\Gamma(\beta_0+n-x)} \theta^{\alpha_0+x-1}(1-\theta)^{\beta_0+n-x-1} d\theta \\ &= \binom{n}{x} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0+x)\Gamma(\beta_0+n-x)}{\Gamma(\alpha_0+\beta_0+n)} \cdot 1 \\ &= \binom{n}{x} \frac{\Gamma(\alpha_0+\beta_0)\Gamma(\alpha_0+x)\Gamma(\beta_0+n-x)}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(\alpha_0+\beta_0+n)}, \end{aligned}$$

där vi har använt faktumet att täthetsfunktionen hos en $\text{Beta}(\alpha_0+x, \beta_0+n-x)$ integrerar till ett. Denna fördelning kallas för *beta-binomialfördelningen* med parametrar n, α_0, β_0.

Vi bildar nu aposteriorifördelningen

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &= \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{f_X(x)} \\ &= \frac{\Gamma(\alpha_0+\beta_0+n)}{\Gamma(\alpha_0+x)\Gamma(\beta_0+n-x)} \theta^{\alpha_0+x-1}(1-\theta)^{\beta_0+n-x-1}. \end{aligned}$$

Med andra ord har vi $\Theta \mid X = x \sim \text{Beta}(\alpha_1, \beta_1)$ där $\alpha_1 = \alpha_0 + x$ och $\beta_1 = \beta_0 + n - x$.

Denna typ av beräkning kan tillämpas på andra fördelningar.

Exempel 6.2. Låt $X \mid \Theta = \theta \sim \text{Po}(\theta)$ och $\Theta \sim \text{Gamma}(\alpha_0, \lambda_0)$. Då har vi

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) &= \frac{\theta^x}{x!} e^{-\theta} \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\lambda_0\theta} \\ &= \frac{\lambda_0^{\alpha_0}}{x!\Gamma(\alpha_0)} \theta^{\alpha_0+x-1} e^{-(\lambda_0+1)\theta}, \end{aligned}$$

samt

$$\begin{aligned} f_X(x) &= \int_0^1 f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta \\ &= \frac{\lambda_0^{\alpha_0}}{x!\Gamma(\alpha_0)} \int_0^{+\infty} \theta^{\alpha_0+x-1} e^{-(\lambda_0+1)\theta} d\theta \\ &= \frac{\lambda_0^{\alpha_0}}{x!\Gamma(\alpha_0)} \frac{\Gamma(\alpha_0+x)}{(\lambda_0+1)^{\alpha_0+x}} \int_0^{+\infty} \frac{(\lambda_0+1)^{\alpha_0+x}}{\Gamma(\alpha_0+x)} \theta^{\alpha_0+x-1} e^{-(\lambda_0+1)\theta} d\theta \\ &= \frac{\lambda_0^{\alpha_0}}{x!\Gamma(\alpha_0)} \frac{\Gamma(\alpha_0+x)}{(\lambda_0+1)^{\alpha_0+x}} \cdot 1 \\ &= \frac{\Gamma(\alpha_0+x)}{x!\Gamma(\alpha_0)} \lambda_0^{\alpha_0} (\lambda_0+1)^{-\alpha_0-x}, \end{aligned}$$

där vi har använt att täthetsfunktionen hos en $\text{Gamma}(\alpha_0+x, \lambda_0+1)$-fördelning integrerar till ett.

Vi får då aposteriorifördelningen

$$\begin{aligned} f_{\Theta|X}\theta \mid x &= \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{f_X(x)} \\ &= \frac{\frac{\lambda_0^{\alpha_0}}{x!\Gamma(\alpha_0)}\theta^{\alpha_0+x-1}e^{-(\lambda_0+1)\theta}}{\frac{\Gamma(\alpha_0+x)}{x!\Gamma(\alpha_0)}\lambda_0^{\alpha_0}(\lambda_0+1)^{-\alpha_0-x}} \\ &= \frac{(\lambda_0+1)^{\alpha_0+x}}{\Gamma(\alpha_0+x)}\theta^{\alpha_0+x-1}e^{-(\lambda_0+1)\theta}, \end{aligned}$$

dvs. att $\Theta \mid X = x \sim \text{Gamma}(\alpha_1, \lambda_1)$ där $\alpha_1 = \alpha_0 + x$ och $\lambda_1 = \lambda_0 + 1$.

Det finns ett mycket användbart trick för att ta fram aposteriorifördelningar som baserar sig på att alla täthetsfunktioner integrerar till ett (och att alla sannolikhetsfunktioner summerar till ett). Detta betyder att alla täthetsfunktioner och sannolikhetsfunktioner endast behöver bestämmas upp till en konstant faktor (dvs. faktorer som inte beror på utfallsvariabeln). Denna konstant kan sedan bestämmas med hjälp av integration (eller summering). I princip kan vi nästan alltid hoppa över detta steg genom att identifiera vilken fördelning det gäller och sätta in rätt konstant. Vi arbetar alltså i praktiken med *onormaliserade täthetsfunktioner* och *onormaliserade sannolikhetsfunktioner*.

Detta trick kan användas i olika sammanhang men är extra användbart när man beräknar aposteriorifördelningar på grund av följande observation: i aposteriorifördelningen kan den aprioriprediktiva fördelningen $f_{\boldsymbol{X}}(\boldsymbol{x})$ betraktas som konstant (då denna inte beror på θ). Vi har alltså

$$f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) \propto f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta).$$

För att hitta aposteriorifördelningen behöver vi därför bara multiplicera aprioriförde lningen med datafördelningen och sedan hitta den fördelning som överenstämmer med denna form upp till en konstant. (Vi kommer se senare i kursen att denna formulering också är användbar även då vi inte kan identifiera en känd fördelning då den onormaliserade formen kan användas för att simulera från aposteriorifördelningen.) Vi slipper därmed beräkna den aprioriprediktiva fördelningen för att ta fram aposteriorifördelningen, vilket förenklar processen avsevärt. Utöver detta kan vi också ta bort och lägga till konstanta faktorer (konstanter med avseende på θ) vid behov.

Exempel 6.3. I binomialfallet ovan har vi då

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\ &\propto \theta^{\alpha_0+x-1}(1-\theta)^{\beta_0+n-x-1}, \end{aligned}$$

där vi har tagit bort alla faktorer som inte beror på θ. Från denna onormaliserade täthetsfunktion ser vi att $\Theta \mid X = x$ måste vara fördelad enligt $\text{Beta}(\alpha_0 + x, \beta_0 + n - x)$.

För Poissonfördelningen har vi, på samma sätt,

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &\propto \frac{\theta^x}{x!}e^{-\theta} \cdot \frac{\lambda_0^{\alpha_0}}{\Gamma(\alpha_0)}\theta^{\alpha_0-1}e^{-\lambda_0\theta} \\ &\propto \theta^{\alpha_0+x}e^{-(\lambda_0+1)\theta}, \end{aligned}$$

dvs. att $\Theta \mid X = x \sim \text{Gamma}(\alpha_1, \lambda_1)$ med $\alpha_1 = \alpha_0 + x$, $\lambda_1 = \lambda_0 + 1$ som vi såg tidigare.

Detta trick kan avsevärt förenkla beräkningarna när integralerna för den aprioriprediktiva fördelningen blir komplicerade, som i fallet med normalfördelad data (se övning 6).

6.1.2 Flera observationer

När man går från en aprioriförde lning till en aposteriorifördelning brukar man säga att man *uppdaterar* fördelningen för parametern med avseende på vissa data. Ibland kallas man detta för en *bayesiansk uppdatering*. Idén här är att vi har någon uppfattning om hur en viss parameter beter sig (dess aprioriförde lning) och att vi uppdaterar denna uppfattning (för att få dess aposteriorifördelning) när vi observerar vissa data.

En konsekvens av detta är att aposteriorifördelningarna ovan enkelt generaliseras till flera datapunkter i ett stickprov. Om vi har $X_1, X_2, \ldots, X_n$ stokastiska variabler, betingat oberoende givet $\Theta = \theta$, har vi enligt kedjeregeln

$$\begin{aligned} f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n \mid \theta) &= f_{X_n|X_1,\ldots,X_{n-1},\Theta}(x_n \mid x_1,\ldots,x_{n-1},\theta) f_{X_1,\ldots,X_{n-1}|\Theta}(x_1,\ldots,x_{n-1} \mid \theta) \\ &= f_{X_n|\Theta}(x_n \mid \theta) f_{X_1,\ldots,X_{n-1}|\Theta}(x_1,\ldots,x_{n-1} \mid \theta), \end{aligned}$$

där vi i andra raden utnyttjat att $X_1 \perp\!\!\!\perp X_1,\ldots,X_{n-1} \mid \Theta$ och tillämpat Sats 3.2. Med tanke på att

$$f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) \propto f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n \mid \theta) f_\Theta(\theta)$$

har vi då

$$\begin{aligned} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) &\propto f_{X_n|\Theta}(x_n \mid \theta) f_{X_1,\ldots,X_{n-1}|\Theta}(x_1,\ldots,x_{n-1} \mid \theta) f_\Theta(\theta) \\ &\propto f_{X_n|\Theta}(x_n \mid \theta) f_{\Theta|X_1,\ldots,X_{n-1}}(\theta \mid x_1,\ldots,x_{n-1}). \end{aligned}$$

Med andra ord kan vi uppdatera fördelningen för Θ först med avseende på $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$ och sedan ta detta som apriorifördelning och uppdatera med avseende på $X_n = x_n$. Induktion ger att detta kan delas upp i n stycken separata uppdateringar, en för varje $X_i = x_i$. Notera att ordningen inte spelar någon roll här – det viktiga är att vi uppdaterar med avseende på olika stokastiska variabler som alla är betingat oberoende givet parametern.

Exempel 6.4. Givet $X_1, X_2, \ldots, X_n$ betingat oberoende givet Θ med $X_i \mid \Theta = \theta \sim \text{Po}(\theta)$ och $\Theta \sim \text{Gamma}(\alpha_0, \lambda_0)$ gör vi ansatsen

$$f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) \propto \theta^{\alpha_0 + \sum_{i=1}^n x_i - 1} e^{-(\lambda_0 + n)\theta}$$

utifrån det tidigare resultatet om Poissonfördelade data. Med andra ord vill vi visa att

$$\Theta \mid X_1 = x_1, \ldots, X_n = x_n \sim \text{Gamma}(\alpha_n, \lambda_n)$$

där $\alpha_n = \alpha_0 + \sum_{i=1}^n x_i$ och $\lambda_n = \lambda_0 + n$. Vi ser att för $n = 1$ återfår vi detta resultat och om det håller för $n - 1$ har vi

$$\begin{aligned} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) &\propto f_{X_n|\Theta}(x_n \mid \theta) f_{\Theta|X_1,\ldots,X_{n-1}}(\theta \mid x_1,\ldots,x_{n-1}) \\ &\propto \theta^{x_n} e^{-\theta} \cdot \theta^{\alpha_0 + \sum_{i=1}^{n-1} x_i - 1} e^{-(\lambda_0 + (n-1))\theta} \\ &= \theta^{\alpha_0 + \sum_{i=1}^n x_i - 1} e^{-(\lambda_0 + n)\theta}, \end{aligned}$$

vilket visar att ansatsen håller genom induktion.

På liknande sätt kan vi ta fram aposteriorifördelningar för andra populära datafördelningar, som exponentialfördelad och normalfördelad data (se övning 6). Bland annat har vi att för $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$ där alla element är betingat oberoende givet Θ och $X_i \mid \Theta = \theta \sim \text{N}(\theta, \sigma^2)$ för $i = 1, 2, \ldots, n$ med $\Theta \sim \text{N}(\mu_0, \tau_0^2)$ så har vi

$$\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{N}(\mu_n, \tau_n^2)$$

där

$$\begin{aligned} \mu_n &= \left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) \Big/ \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \right) \\ \tau_n^2 &= \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}. \end{aligned}$$

Ett annat resultat är att för $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$ med alla element betingat oberoende givet Θ och $X_i \mid \Theta = \theta \sim \text{Exp}(\theta)$ för $i = 1, 2, \ldots, n$ och $\Theta \sim \text{Gamma}(\alpha_0, \lambda_0)$ så har vi

$$\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(\alpha_n, \lambda_n)$$

där

$$\begin{aligned} \alpha_n &= \alpha_0 + n \\ \lambda_n &= \lambda_0 + \sum_{i=1}^n x_i. \end{aligned}$$

Nu när vi har ett par formler kan vi börja genomföra några beräkningar.

Exempel 6.5. På väg till jobbet behöver du korsa en hårt trafikerad cykelväg. Det finns ett övergångställe, men cyklisterna verkar inte speciellt intresserade av att stanna, så du väntar på ett mellanrum så att du kan komma över. Under tiden räknar du antalet cyklister som passerar under varje 10-sekundersintervall:

Intervall (sek.)	0–10	10–20	20–30	30–40	40–50	50–60
Antal cyklister	4	4	1	3	2	6

Antag att antalet cyklister X_i under varje intervall har den betingade fördelningen $X_i \mid \Theta = \theta \sim \text{Po}(\theta)$ där $X_1, X_2, \ldots, X_6$ är betingat oberoende givet Θ. För att använda vår uppdateringsregel ovan antar vi att $\Theta \sim \text{Gamma}(\alpha, \lambda)$ för några α, λ. Från formelsamlingen har vi att $\text{E}[\Theta] = \alpha/\lambda$ samt $\text{Var}[\Theta] = \alpha/\lambda^2$. Ett vettigt antagande kan vara att vi förväntar oss circa en cykel per tio sekunder (dvs. sex stycken per minut) med en osäkerhet (standardavvikelse) på ± 1 cyklist. Vi har då $\text{E}[\Theta] = 1$ och $\text{Var}[\Theta] = 1^2 = 1$, vilket ger $\alpha = 1$, $\lambda = 1$.

Datan ovan ger $n = 6$ och $\sum_{i=1}^{n} x_i = 20$. Enligt vår uppdateringsformel har vi då att $\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(1 + 20, 1 + 6) = \text{Gamma}(21, 7)$. Denna aposteriorifördelning har väntevärde $\text{E}[\Theta \mid \boldsymbol{X} = \boldsymbol{x}] = 21/7 = 3$ och standardavvikelse $\text{Std}[\Theta \mid \boldsymbol{X} = \boldsymbol{x}] = \sqrt{21/7^2} = \sqrt{21}/7 \approx 0.655$. Vår parameter Θ är därför sannolikt mycket större än vi tidigare hade trott, närmare tre än ett.

Notera att ett annat val av apriorifördelning skulle ge ett annat svar här. Låt oss säga att vi fortfarande förväntar oss att Θ befinner sig runt ett, men att vi har mycket mindre osäkerhet, dvs. vi tror starkt på detta påstående (vi kanske går här varje dag och observerar sällan eller aldrig fler än en eller två cyklister per tio sekunder). Vi kanske har $\text{Std}[\Theta] = 1/4$, vilket ger $\text{Var}[\Theta] = 1/16$. Tillsammans med $\text{E}[\Theta] = 1$ får vi då $\alpha = 16$, $\lambda = 16$. I detta fall får vi aposteriorifördelningen $\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(16 + 20, 16 + 6) = \text{Gamma}(36, 22)$, vilken är koncentrerad närmare ett än den tidigare aposteriorifördelningen. Vi har, till exempel $\text{E}[\Theta \mid \boldsymbol{X} = \boldsymbol{x}] = 36/22 = 18/11 \approx 1.636$ och $\text{Std}[\Theta \mid \boldsymbol{X} = \boldsymbol{x}] = \sqrt{36/22^2} = 3/11 \approx 0.273$.

I detta fall så är aposteriorifördelningen ganska starkt beroende på vilken aprioriförde lning som väljs. Detta beror på att vi har relativt få datapunkter ($n = 6$), vilket gör att aprioriförde lningen får större betydelse än vi annars skulle se.

6.1.3 Bayesianska punktskattningar

Även om det centrala objektet inom bayesiansk inferens är aposteriorifördelningen (eftersom denna innehåller *all* om parametern som vi får från datat) är det ofta användbart att sammanfatta denna aposteriorifördelning på olika sätt. Till exempel kan vi, som ovan, använda väntevärdet och variansen (eller standardavvikelsen) hos aposteriorifördelningen för att bilda en uppfattning om hur parametern beter sig. Detta är så pass vanligt att vi benämner dessa som *aposterioriväntevärdet*, *aposteriorivariansen* och *aposterioristandardavvikelsen*. (Ibland kallas de också helt enkelt för det *betingade väntevärdet*, den *betingade variansen* och den *betingade standardavvikelsen*.) För aprioriförde lningen har vi på samma sätt *aprioriväntevärdet*, *apriorivariansen* samt *aprioristandardavvikelsen*.

För binomialfördelad data $X \mid \Theta = \theta \sim \text{Bin}(n, \theta)$ där $\Theta \sim \text{Beta}(\alpha, \beta)$ har vi då

$$\text{E}[\Theta \mid X = x] = \frac{\alpha + x}{\alpha + \beta + n}$$

vilket kan jämföras med väntevärdet hos aprioriförde lningen

$$\text{E}[\Theta] = \frac{\alpha}{\alpha + \beta}.$$

En annan skattning är ML-skattningen för θ (dvs. inom den frekventistiska ramverket), vilken ges av x/n. Intressant nog kan vi se aposterioriväntevärdet som en interpolation mellan aprioriväntevärdet $\alpha/(\alpha+\beta)$ och ML-skattningen x/n:

$$\text{E}[\Theta \mid X = x] = \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{x}{n}.$$

Ju mindre n är, desto närmare ett ligger $(\alpha + \beta)/(\alpha + \beta + n)$ och desto mer vikt lägger vi på aprioväntevärdet $\alpha/(\alpha + \beta)$. Å andra sidan, för stora n så närmar vi oss istället ML-skattningen x/n. Den bayesianska metoden ger oss alltså ett sätt att (automatiskt) väga av inflytandet av tidigare information (dvs. a priori) och den observerade datan.

En användbar olikhet för aposteriorivariansen fås genom att betraka lagen om total sannolikhet, som säger att

$$\text{Var}[\Theta] = \text{E}[\text{Var}[\Theta \mid \boldsymbol{X}]] + \text{Var}[\text{E}[\Theta \mid \boldsymbol{X}]].$$

Eftersom den andra termen är större än eller lika med noll får vi

$$\mathrm{E}[\mathrm{Var}[\Theta \mid \boldsymbol{X}]] \leq \mathrm{Var}[\Theta].$$

Med andra ord så är aposteriorivariansen i genomsnitt mindre än apriorivariansen (där medelvärdet tagits över alla möjliga observationer $\boldsymbol{X}$). Ofta så håller också denna olikhet för enskilda observationer $\boldsymbol{X} = \boldsymbol{x}$. Detta är ganska naturligt, då tillförandet av information (dvs. observationen $\boldsymbol{X} = \boldsymbol{x}$) rimligtvis borde minska osäkerheten hos parametern Θ och därför ge oss en mindre varians. Notera att undantag finns till denn tumregel (se övning 6).

Andra vanliga sätt att sammanfatta aposteriorifördelningen är att använda kvantiler, eller mer specifikt medianen av aposteriorifördelningen. Denna kallas helt enkelt för *aposteriorimedianen*. Ett annat alternativ är att beräkna typvärdet, dvs. det värde på θ som maximerar $f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x})$. Detta kallas för *maximum a posteriori* (MAP) för Θ givet $\boldsymbol{X} = \boldsymbol{x}$.

Vilken av dessa sammanfattningar som används beror på tillämpningen. Vi kommer att se i senare föreläsningar hur man kan använda beslutsteori för att bestämma vilken sammanfattning som passar bäst i en viss situation.

6.2 Aposterioriprediktiva fördelningen

En viktig aspekt av bayesiansk inferens är att den erbjuder ett naturligt ramverk för förutsägelse av nya datapunkter utifrån de observerade data, så kallad *prediktion*. Detta är också möjligt inom frekventistisk inferens, men där krävs att vi binder oss vid en specifik skattning av parametern θ för att generera nya datapunkter. Osäkerheten i själva skattningen av θ kan inte enkelt tas i beaktning.

Aposteriorifördelningen gör det däremot enkelt att beräkna fördelningen för nya (icke observerade) data genom att tillämpa kedjeregeln och marginalisera bort parametern Θ. Om våra nya datapunkter har datafördelning $f_{\boldsymbol{Y}|\Theta}(\boldsymbol{y} \mid \theta)$ där $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \Theta$ så får vi att $\boldsymbol{Y}$ givet $\boldsymbol{X}$ har fördelningen

$$\begin{aligned}
f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x}) &= \int_\Omega f_{\boldsymbol{Y},\Theta|\boldsymbol{X}}(\boldsymbol{y}, \theta \mid \boldsymbol{x}) d\theta \\
&= \int_\Omega f_{\boldsymbol{Y}|\Theta,\boldsymbol{X}}(\boldsymbol{y} \mid \theta, \boldsymbol{x}) f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) d\theta \\
&= \int_\Omega f_{\boldsymbol{Y}|\Theta}(\boldsymbol{y} \mid \theta) f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) d\theta,
\end{aligned}$$

där den sista likheten fås genom att tillämpa det betingade oberoendet mellan $\boldsymbol{X}$ och $\boldsymbol{Y}$ givet Θ.

Fördelningen för $\boldsymbol{Y}$ givet $\boldsymbol{X}$ kallas för den *aposterioriprediktiva fördelningen* av $\boldsymbol{Y}$ givet $\boldsymbol{X}$. Det är alltså den (prediktiva) fördelning som vi anser nya datapunkter borde ha givet en viss observation (dvs. a posteriori).

Exempel 6.6. Låt oss återgå till Poissonfördelad data med gammafördelad parameter. Vi har alltså $X_1, X_2, \ldots, X_n$ betingat oberoende givet $\Theta = \theta$ där $X_i \mid \Theta = \theta \sim \mathrm{Po}(\theta)$ för $i = 1, 2, \ldots, n$ samt $\Theta \sim \mathrm{Gamma}(\alpha_0, \lambda_0)$. Detta gav $\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{Gamma}(\alpha_n, \lambda_n)$ där $\alpha_n = \alpha_0 + \sum_{i=1}^n x_i$ och $\lambda_n = \lambda_0 + n$.

Om vi nu har $Y \perp\!\!\!\perp \boldsymbol{X} \mid \Theta$ så att $Y \mid \Theta = \theta \sim \mathrm{Po}(\theta)$ så får vi

$$\begin{aligned}
f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) &= \int_0^{+\infty} f_{Y|\Theta}(y \mid \theta) f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) d\theta \\
&= \int_0^{+\infty} \frac{\theta^y}{y!} e^{-\theta} \cdot \frac{\lambda_n^{\alpha_n}}{\Gamma(\alpha_n)} \theta^{\alpha_n - 1} e^{-\lambda_n \theta} d\theta \\
&= \frac{\lambda_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{1}{y!} \int_0^{+\infty} \theta^{\alpha_n + y - 1} e^{-(\lambda_n + 1)\theta} d\theta \\
&\propto \frac{1}{y!} \int_0^{+\infty} \theta^{\alpha_n + y - 1} e^{-(\lambda_n + 1)\theta} d\theta \\
&= \frac{1}{y!} \frac{\Gamma(\alpha_n + y)}{(\lambda_n + 1)^{\alpha_n + y}} \\
&\propto \frac{\Gamma(\alpha_n + y)}{y!} (\lambda_n + 1)^y
\end{aligned}$$

där vi i sista likheten har använt att täthetsfunktionen för en gammafördelning integrerar till ett. För att identifiera fördelningen antar vi att $\alpha_0 \in \mathbb{Z}_{>0}$, vilket ger att $\alpha_n + y \in \mathbb{Z}_{>0}$ och $\Gamma(\alpha_n + y) = (\alpha_n + y - 1)!$. Vi kan då introducera

ett par konstanter (med avseende på y) och skriva

$$\begin{aligned} f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) &\propto \frac{(\alpha_n + y - 1)!}{y!} \left(\frac{1}{\lambda_n + 1}\right)^{\alpha_n + y} \\ &\propto \frac{(\alpha_n + y - 1)!}{(\alpha_n - 1)! y!} \left(\frac{\lambda_n}{\lambda_n + 1}\right)^{\alpha_n} \left(\frac{1}{\lambda_n + 1}\right)^{y}. \end{aligned}$$

Alltså har vi att Y givet $\boldsymbol{X}$ är negativt binomialfördelad, dvs. att $Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{NegBin}\left(\alpha_n, \frac{\lambda_n}{\lambda_n + 1}\right)$.

En direkt tillämpning av den aposterioriprediktiva fördelningen är helt enkelt att göra en prediktion, dvs. att förutsäga att någonting kommer att hända med en viss sannolikhet som beror på den datan vi har observerat.

Exempel 6.7. Låt oss återgå till cykelvägsexemplet. Där hade vi $\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(\alpha_n, \lambda_n)$ med $\alpha_n = 21$ och $\lambda_n = 7$ (för den första aprioriförde lningen). Vad är nu sannolikheten att antalet cyklister under nästa 10-sekundersintervall är noll (och vi kan korsa cykelvägen säkert)? Enligt formeln ovan har vi att om $Y \mid \Theta = \theta \sim \text{Po}(\theta)$ och $Y \perp\!\!\!\perp \boldsymbol{X} \mid \Theta$ så har vi att

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{NegBin}\left(\alpha_n, \frac{\lambda_n}{\lambda_n + 1}\right) = \text{NegBin}(21, 7/8).$$

Den sökta sannolikheten är alltså

$$\text{P}(Y = 0 \mid \boldsymbol{X} = \boldsymbol{x}) = \binom{21 - 1}{0} (7/8)^{21} (1/8)^0 \approx 0.0606,$$

dvs. relativt liten (kanske är värt att ta en omväg).

En annan tillämpning är för att göra en rimlighetsbedömning av modellen. Vi kommer att gå in mer i detalj i detta under senare föreläsningar, men grundidén är att generera nya datapunkter $\boldsymbol{Y}$ enligt samma datafördelning som originaldatat. Om den denna aposteriorifördelning överenstämmer bra med de observerade datat kan vi konstatera att modellen är någorlunda korrekt.

Exempel 6.8. Vi repeterar dragning av sex stycken utfall från $\text{NegBin}(21, 7/8)$ och får

$$\begin{array}{cccccc} 8 & 5 & 3 & 3 & 1 & 1 \\ 3 & 1 & 2 & 3 & 1 & 6 \\ 4 & 3 & 1 & 2 & 0 & 4 \\ 1 & 2 & 5 & 4 & 5 & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

Om vi jämför dessa med våra observerade data

$$\begin{array}{cccccc} 4 & 4 & 1 & 3 & 2 & 6 \end{array}$$

så förefaller de inte vara särskilt avvikande (noll är relativt sällsynt, medan vi har få värden över fem). För att kvantifiera denna avvikelse behöver vi mer avancerade verktyg som kommer att introduceras i en senare föreläsning.

Notera att den aposterioriprediktiva fördelningen inte nödvändigtvis behöver baseras på den ursprungliga datafördelningen. Ibland vill vi prediktera relaterade storheter som beror på samma parameter.

Exempel 6.9. I cykelvägsexemplet var vi intresserade av när antalet cyklister i ett 10-sekundersintervall är noll. Som vi såg är sannolikheten att detta inträffar för ett givet intervall ganska låg. En relaterad fråga är under hur många intervall vi måste vänta innan vi får noll cyklister per intervall. Om vi har $Y \mid \Theta = \theta \sim \text{Po}(\theta)$ är sannolikheten att ha noll cyklister lika med

$$\text{P}(Y = 0 \mid \Theta = \theta) = \frac{\theta^0}{0!} e^{-\theta} = e^{-\theta}.$$

Antalet intervall K vi måste vänta innan detta inträffar har då en geometrisk fördelning med parameter $e^{-\theta}$, dvs. vi har $K \mid \Theta = \theta \sim \text{Geom}(e^{-\theta})$. Detta ger i sin tur aposterioriprediktiva fördelningen

$$\begin{aligned} f_{K|\boldsymbol{X}}(k \mid \boldsymbol{x}) &= \int_0^{+\infty} f_{K|\Theta}(k \mid \theta) f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) d\theta \\ &= \int_0^{+\infty} e^{-\theta}(1-e^{-\theta})^k \cdot \frac{\theta^{\alpha_n}}{\Gamma(\alpha_n)} e^{-\lambda_n \theta} d\theta \\ &= \frac{1}{\Gamma(\alpha_n)} \int_0^{+\infty} (1-e^{-\theta})^k \theta^{\alpha_n} e^{-(\lambda_n+1)\theta} d\theta. \end{aligned}$$

Integralen ovan går att beräkna (binomialutveckling av $(1-e^{-\theta})^k$, vilket ger en summa av gammaintegraler), men det kräver ganska mycket arbete.

Istället kan vi ta ett par genvägar. Den första är att använda lagen om total förväntan, vilket ger oss

$$\begin{aligned} \text{E}[K \mid \boldsymbol{X} = \boldsymbol{x}] &= \text{E}[\text{E}[K \mid \Theta = \theta, \boldsymbol{X} = \boldsymbol{x}] \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \text{E}[\text{E}[K \mid \Theta = \theta] \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \text{E}\left[\frac{1-e^{-\theta}}{e^{-\theta}} \,\middle|\, \boldsymbol{X} = \boldsymbol{x} \right] \\ &= \text{E}[e^{\theta} - 1 \mid \boldsymbol{X} = \boldsymbol{x}] = \text{E}[e^{\theta} \mid \boldsymbol{X} = \boldsymbol{x}] - 1, \end{aligned}$$

där vi bland annat har utnyttjat att $K \perp\!\!\!\perp \boldsymbol{X} \mid \Theta$. Detta väntevärde kan beräknas genom en enkel gammaintegral

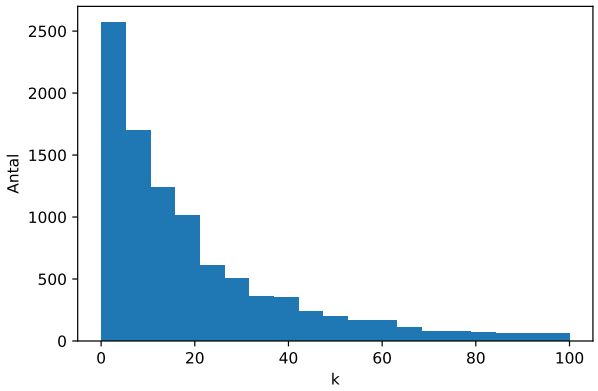
$$\begin{aligned} \text{E}[e^{\theta} \mid \boldsymbol{X} = \boldsymbol{x}] &= \int_0^{+\infty} e^{\theta} f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) d\theta \\ &= \int_0^{+\infty} e^{\theta} \frac{\theta^{\alpha_n}}{\Gamma(\alpha_n)} \theta^{\alpha_n - 1} e^{-\lambda_n \theta} d\theta \\ &= \int_0^{+\infty} \frac{\theta^{\alpha_n}}{\Gamma(\alpha_n)} \theta^{\alpha_n - 1} e^{-(\lambda_n - 1)\theta} d\theta \\ &= \frac{\lambda_n^{\alpha_n}}{\Gamma(\alpha_n)} \frac{\Gamma(\alpha_n)}{(\lambda_n - 1)_n^{\alpha}} \\ &= \left(\frac{\lambda_n}{\lambda_n - 1} \right)^{\alpha_n} \end{aligned}$$

I vårt fall har vi alltså

$$\text{E}[K \mid \boldsymbol{X} = \boldsymbol{x}] = \left(\frac{7}{7-1} \right)^{21} - 1 = \left(\frac{7}{6} \right)^{21} - 1 \approx 24.5.$$

Vi måste alltså vänta i genomsnitt 24.5 intervall, dvs. runt 245 sekunder. Alltså är det definitivt värt att ta en annan väg (eller gå ut i cykelvägen och ta sina chanser).

Ett annat alternativ är att använda simulering för numeriskt approximera den aposterioriprediktiva fördelningen och beräkna de sökta väntevärdena eller sannolikheterna empiriskt. Vi gör detta genom att generera ett stort antal utfall n_{samp} av $\Theta \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Gamma}(21, 7)$, vilket ger $\theta_1, \ldots, \theta_{n_{\text{samp}}}$. Dessa kan sedan användas för att generera utfall av $K \mid \boldsymbol{X} = \boldsymbol{x}$ genom att för varje $i = 1, \ldots, n_{\text{samp}}$ generera k_i utifrån fördelningen $\text{Geom}(\theta_i)$. Vi får då n_{samp} utfall av $K \mid \boldsymbol{X} = \boldsymbol{x}$ och kan plotta deras histogram (här har vi tagit $n_{\text{samp}} = 10000$):



Om vi beräknar det aritmetiska medelvärdet får vi 25.2, vilket ligger ganska nära väntevärdet vi fick ovan.

Histogrammet ovan vittnar dock om att det är ett fåtal större utfall av K som drar upp medelvärdet. Merparten av utfallen ser ut att ligga under 25. En viktig fördel med simuleringar är att vi kan enkelt skatta andra viktiga storheter, som till exempel aposteriorimedianen, genom att göra motsvarande beräkningar på de numeriska värden vi erhållet. Till exempel får vi att aposteriorimedianen är approximativt 13, vilket betyder att i hälften av fallen behöver vi bara vänta $13 \cdot 10 = 130$ sekunder, dvs. strax över två minuter.

Vi kan också approximera sannolikheter som $\mathrm{P}(K <= 6 \mid \boldsymbol{X} = \boldsymbol{x})$, dvs. sannolikheten att vi måste vänta högst en minut. Detta fås genom att räkna andelen av våra genererade utfall k som är mindre än eller lika med 6. För datat ovan får vi att denna sannolikhet är ungefär 0.29.