Föreläsning 15

Statistisk inlärning och dataanalys (Kungliga Tekniska Högskolan)

Föreläsning 15

15.1 Konvergens

Vi har sett många gånger nu att det kan vara hjälpsam om vi kan öka ett stickprovs storlek. I denna föreläsning ska vi undersöka vad händer när ett stickprovs storlek går mot oändligheten. Förutom att hjälpa med skattning kan det också förenkla beräkningar att arbeta med asymptotiska approximationer när stickprovets storleken är tillräcklig stor.

Vi ska betrakta två olika typer av konvergens. Den första är konvergens i sannolikhet: Vi säger att en sekvens av stockastiska variabler $X_1, X_2, \ldots$ *konvergerar i sannolikhet* mot X när vi har för alla $\varepsilon > 0$ att

$$P(|X_n - X| \geq \varepsilon) \longrightarrow 0$$

då $n \longrightarrow \infty$. Vi betecknar denna konvergens i sannolikhet som

$$X_n \xrightarrow{P} X \qquad \text{då} \qquad n \longrightarrow \infty.$$

Vi har redan sett denna version av konvergens i SF1918 där vi bevisade viktiga resultater som *Tjebysjovs olikhet*:

Sats 15.1 (Tjebysjovs olikhet)**.** *Låt X vara en stokastiska variabel med väntevärdet μ och standardavvikelse $\sigma > 0$. För varje $k > 0$ gäller*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Med hjälp av Tjebysjovs olikhet bevisade *den stora talens lagen*:

Sats 15.2 (Stora talens lag)**.** *Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade stokastiska variabler med väntevärdet μ och standardavvikelsen $\sigma > 0$. Låt*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Då har vi för varje $\varepsilon > 0$ att

$$P(|\bar{X}_n - \mu| < \varepsilon) \longrightarrow 1$$

då $n \longrightarrow \infty$. Med andra ord,

$$\bar{X}_n \xrightarrow{P} \mu \qquad \text{då} \qquad n \longrightarrow \infty.$$

I följande ska vi ofta ha en sekvens av stokastiska variabler $X_1, X_2, \ldots$ och betraka en statistika $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ på de första n variabler $X_1, \ldots, X_n$. För att beteckna att vi betrakta statistiken för $X_1, \ldots, X_n$ ska vi skriva $W_n = W_n(\boldsymbol{X}) = W_n(X_1, \ldots, X_n)$. Exempelvis, given $X_1, X_2, \ldots$ oberoende och likafördelade med väntevärdet μ beteckna vi stickprovsväntevärdet respektive stickprovsvariansen för $X_1, \ldots, X_n$ som

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{och} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Den andra typen av konvergens vi ska betrakta kallas för konvergens i fördelning: Vi säger att en sekvens av stokastiska variabler $X_1, X_2, \ldots$ *konvergerar i fördelning* mot X när

$$F_{X_n}(x) \longrightarrow F_X(x)$$

då $n \longrightarrow \infty$ för varje x. Vi betecknar konvergens i fördelning som

$$X_n \xrightarrow{D} X \qquad \text{då} \qquad n \longrightarrow \infty.$$

Vi har också sett (i SF1918) denna form av konvergens när vi pratade om *den centrala gränsvärdessatsen*:

Sats 15.3 (Den centrala gränsvärdessatsen)**.** *Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade med väntevärdet μ och varians $0 < \sigma^2 < \infty$. Då har vi att*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} Z \qquad \textit{då} \qquad n \longrightarrow \infty$$

där $Z \sim \mathrm{N}(0,1)$.

Eftersom vi skulle inte vilja specificiera Z varje gången skriver vi ofta denna kovergens som

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} \mathrm{N}(0,1) \qquad \text{då} \qquad n \longrightarrow \infty.$$

Motsvarande kan vi skriva

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2) \qquad \text{då} \qquad n \longrightarrow \infty.$$

Ett användbart resultat är att konvergens i sannolikhet är starkare än konvergens i fördelningen; dvs. om

$$X_n \xrightarrow{P} X \qquad \text{gäller} \qquad X_n \xrightarrow{D} X.$$

Generellt sett stämmer inte det omvända men när $X_n \xrightarrow{D} c$ för någon konstant c har vi att $X_n \xrightarrow{P} c$.

Vi har också föjlande satsen:

Sats 15.4 (Satsen om kontinuerliga avbildningar)**.** *Om $g(x)$ är en kontinuerlig funktion har vi då att*

$$\begin{array}{lllll} \textit{När} & X_n \xrightarrow{P} X & \textit{gäller} & g(X_n) \xrightarrow{P} g(X), & \textit{och} \\ \textit{När} & X_n \xrightarrow{D} X & \textit{gäller} & g(X_n) \xrightarrow{D} g(X). & \end{array}$$

Det sista användbara resultatet kallas för *Slutskys sats*:

Sats 15.5 (Slutskys sats)**.** *När $X_1, X_2, \ldots$ konvergerar till X i fördelning och $Y_1, Y_2, \ldots$ konvergerar i sannolikhet till en konstant k gäller*

1. $X_n Y_n \xrightarrow{D} aX$, *och*
2. $X_n + Y_n \xrightarrow{D} X + a$.

15.2 Deltametoden

Observationerna hittills fokuserar på gränsvärdet för en viss sekvens av stokastiska variabler men vi är ofta intresserad av funktioner av sådana variabler. Man kan naturligtvis betrakta en transformation av stokastiska variabler men det kan bli komplicerad att hitta den rätta formen av täthetsfunktionen och då bevisa asymptotiska egenskaper. Vi introducerar nu ett sätt att bestämma asymptotiska fördelningar och/eller approximativa fördelningar statistiker $W_n(\boldsymbol{X})$ när n är stort. Metoden baseras på taylorpolynom.

Låt $g(x)$ vara en funktion för vilken $g^{(r)}(x) = \frac{d}{dx^r} g(x)$ existerar. Given en konstant c, den r : *te taylorpolynom* för $g(x)$ på c är

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(c)}{i!}(x-c)^i.$$

Taylorpolynom ger oss ett sätt att approximera funktionen $g(x)$ i en granne kring c. I denna kurs använder vi meste den första taylorpolynom för att approximera $g(x)$:

$$g(x) \approx T_1(x) = g(c) + g'(c)(x-c).$$

Vi kan också gör detsamma för fleravariabel fallet där $g(\boldsymbol{x}) : \mathbb{R}^k \longrightarrow \mathbb{R}$ dubbel derivierbar:

$$g(\boldsymbol{x}) \approx g(\boldsymbol{c}) + \sum_{i=0}^{k} \left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{c}} (x_i - c_i)$$

för $\boldsymbol{c} \in \mathbb{R}^k$.

Låt oss ta $\boldsymbol{c} = \boldsymbol{\mu}$ för någon k-dimensionella stokastisk vektorn $\boldsymbol{X} = (X_1, \ldots, X_k)$. Då har vi att

$$\begin{aligned}
g(\boldsymbol{x}) &\approx g(\boldsymbol{\mu}) + \sum_{i=1}^{k} \left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}} (X_i - \mu_i), \qquad \text{och} \\
\mathrm{E}[g(\boldsymbol{X})] &\approx \mathrm{E}[g(\boldsymbol{\mu})] + \sum_{i=1}^{k} \left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}} \mathrm{E}[x_i - \mu_i], \\
\mathrm{E}[g(\boldsymbol{X})] &\approx \mathrm{E}[g(\boldsymbol{\mu})] + \sum_{i=1}^{k} \left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}} 0, \\
\mathrm{E}[g(\boldsymbol{X})] &\approx \mathrm{E}[g(\boldsymbol{\mu})].
\end{aligned}$$

På samma sätt kan vi approximera variansen för $g(\boldsymbol{X})$ som

$$\begin{aligned}
\mathrm{Var}[g(\boldsymbol{X})] &\approx \mathrm{E}[(g(\boldsymbol{X}) - g(\boldsymbol{\mu}))^2], \\
&\approx \mathrm{E}\left[\left(\sum_{i=1}^{k} \left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}} (X_i - \mu_i)\right)^2\right], \\
&\approx \sum_{i=1}^{n} \left(\left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}}\right)^2 \mathrm{Var}[X_i] + 2\sum_{i>j} \left(\left.\frac{\partial g}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{\mu}} \left.\frac{\partial g}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{\mu}}\right) \mathrm{Cov}[X_i, X_j].
\end{aligned}$$

Medan dessa ger oss inte en approximation för hela fördelningen är det ofta tillräckligt att har en approximation för väntevärdet och variansen. Notisera att dessa approximationer stämmer bara om $\boldsymbol{X}$ varierar lite kring $\boldsymbol{\mu}$ eller när $g(\boldsymbol{X})$ är mycket smidig (nästan linjär). Vi kan inte vanligtvis kontrollera hur smidig funktionen är men vi kan kontrollera hur mycket den variera kring $\boldsymbol{\mu}$; exempelvis, när vi betrakta central gränsvärdessatsen tillämpad till X:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2),$$

där $\mu = \mathrm{E}[X]$ och $0 < \sigma^2 = \mathrm{Var}[X]$. När $n \longrightarrow \infty$ blir X mer och mer koncentrerad kring μ och approximationen $f_X(x) \approx T_1(x)$ blir rimligare. Här använder vi en lite annorlunda version av taylorsapproximation:

$$g(x) \approx g(c) + g'(d)(x - c),$$

där d ligger mellan x och c. Denna approximation stämmer när $g(x)$ är kontinuerlig. Anta nu att vi har en sekvens av stokastiska variabler $W_1, W_2, \ldots$ för vilken

$$\sqrt{n}(W_n - c) \xrightarrow{D} \mathrm{N}(0, v)$$

stämmer för någon $c \in \mathbb{R}$ och $v > 0$. Då har vi att

$$g(W_n) \approx g(c) + g'(d)(W_n - c).$$

Notisera att $W_n \xrightarrow{P} c$ då $n \longrightarrow \infty$ enligt Slutskys sats tillämpad till $\sqrt{n}(W_n - c)\frac{1}{\sqrt{n}}$. Eftersom vi antar att $|d - c| < |W_n - c|$ följer det att $d \to c$ och därmed $g'(d) \to g'(c)$. Vi kan ordna om den ovanstående approximationen och multplicera med $\sqrt{n}$ för att få

$$\begin{aligned}
\sqrt{n}(g(W_n) - g(c)) &\approx \sqrt{n}(g'(d)(W_n - c)), \\
&\approx g'(d)(\sqrt{n}(W_n - c)).
\end{aligned}$$

Eftersom $g(x)$ är kontinuerlig har vi att $g'(d) \xrightarrow{P} g'(c)$ då $n \longrightarrow \infty$ enligt satsen om kontinuerliga avbildningar. Vi har också antagit att $\sqrt{n}(W_n - c) \xrightarrow{D} \mathrm{N}(0, v)$. Det föjler då från Slutskys sats att

$$\sqrt{n}(g(W_n) - g(c)) \xrightarrow{D} g'(c)\,\mathrm{N}(0, v) = \mathrm{N}(0, g'(c)^2 v).$$

Detta resultat kallas för *deltametoden*:

Sats 15.6 (Deltametoden)**.** *Låt $W_1, W_2, \ldots$ vara en sekvens av stokastiska variabler som uppfyller*

$$\sqrt{n}(W_n - c) \xrightarrow{D} \mathrm{N}(0, v)$$

för någon $c \in \mathbb{R}$ och $v > 0$. För $g(x)$ kontinuerlig antar att $g'(x)$ existerar och $g'(c) \neq 0$. Då det gäller

$$\sqrt{n}(g(W_n) - g(c)) \xrightarrow{D} \mathrm{N}(0, g'(c)v).$$

Det följer att för tillräcklig stor n kan vi approximera fördelningen för $g(W_n)$ som

$$g(W_n) \approx \mathrm{N}\left(g(c), \frac{g'(c)v}{n}\right).$$

Exempel 15.1. Låt $X_1, X_2, \ldots$ vara oberoende och Ber(p)-fördelade. Vi har att $\mathrm{E}[X_i] = p$ och $\mathrm{Var}[X_i] = p(1-p)$. Det följer då enligt central gränsvärdessatsen att

$$\sqrt{n}(\bar{X}_n - \mathrm{E}[X_i]) \xrightarrow{D} \mathrm{N}(0, \mathrm{Var}[X_i]),$$

eller

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow{D} \mathrm{N}(0, p(1-p)),$$

Om vi skulle vilja skatta oddset $\frac{p}{1-p}$ istället för bara p kan vi ta $g(x) = \frac{x}{1-x}$ och använda deltametoden för att få

$$\sqrt{n}(g(\bar{X}_n) - g(p)) \xrightarrow{D} \mathrm{N}(0, g'(p)^2 p(1-p)),$$

eller

$$\sqrt{n}(g(\bar{X}_n) - p/(1-p)) \xrightarrow{D} \mathrm{N}\left(0, \frac{1}{(1-p)^4}(p(1-p))\right) = \mathrm{N}\left(0, \frac{p}{(1-p)^3}\right),$$

För stort n kan vi approximera fördelningen för $g(\bar{X}_n)$ som

$$\mathrm{N}\left(\frac{p}{1-p}, \frac{p}{n(1-p)^3}\right).$$

Skattningen blir mer exakt när $n \longrightarrow \infty$ men konvergenshastigheten ges av $p/(1-p)^3$. För liten p konvergens går snabbt men för stort p blir det ganska långsamt eftersom nämnaren blir stor.

Deltametoden är ganska användbar eftersom det kan vara tillämpad så länge som vi har ett gränsvärde

$$\sqrt{n}(W_n - c) \xrightarrow{D} \mathrm{N}(0, v),$$

exempelvis, när som helst vi har ett resultat från den centrala gränsvärdessatsen.

15.3 Konsistens

Det händer ofta att vi är inte intresserad i en sekvens av stokastiska variabler men istället en sekvens av skattningar $W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \ldots$; exempelvis, se Exempel 15.1. Vi en sådan sekvens är vi intresserad av egenskaper för skattningarna då stickprovets storlek går till oändlighet. Vi säger att en sekvens av skattningar $W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \ldots$ för en parameter θ är *konsistent* när

$$W_n(\boldsymbol{x}) \xrightarrow{P} \theta \qquad \text{för alla } \theta \in \Omega \qquad \text{då} \qquad n \longrightarrow \infty.$$

Skattningarna $W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \ldots$ är ofta detsamma på meningen att de har samma ”form” och skatta med samma ”metod” men tillämpa till stickprov av annorlunda storlek. Det är därmed vanligtvis att säga att skattningen $W_n(\boldsymbol{X})$ är en konsistent skattning för θ när vi menar att *sekvensen* $W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \ldots$ är konsistent för θ. Exempelvis, om vi tar $W_n(\boldsymbol{X}) = \bar{X}_n$ är $W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), \ldots$ konsistent för $\mu = \mathrm{E}[X_i]$ enligt stora talens lagen. Vi säger bara att $\bar{X}_n$ är en konsistent skattning för μ.

Generellt sett, om vi anta att $\mathrm{E}[W_n] \xrightarrow{P} \theta$ då $n \longrightarrow \infty$, dvs. vi antar att W_n är *asymptotiskt väntevärdesriktig*, kan vi hitta N för varje $\varepsilon > 0$ så att

$$|\,\mathrm{E}[W_n] - \theta| < \frac{\varepsilon}{2} \qquad \text{för alla } n > N.$$

I detta fall har vi att

$$\begin{aligned}|W_n - \theta| &= |W_n - \mathrm{E}[W_n] + \mathrm{E}[W_n] - \theta|,\\ &\leq |W_n - \mathrm{E}[W_n]| + |\,\mathrm{E}[W_n] - \theta|, \qquad \text{(trekant olikhet)}\\ &\leq |W_n - \mathrm{E}[W_n]| + \frac{\varepsilon}{2}.\end{aligned}$$

Om $\mathrm{Var}[W_n] \xrightarrow{P} 0$ följer det från Tjebysjovs olikhet att $W_n \xrightarrow{P} \mathrm{E}[W_n]$. Så det finns N så att för alla $n > N$ $|W_n - \mathrm{E}[W_n]| < \varepsilon/2$. Det följer att $|W_n - \theta| < \varepsilon$ och

$$P(|W_n - \theta| > \varepsilon) \leq P(|W_n - \mathrm{E}[W_n]| > \varepsilon/2).$$

Enligt Tjebysjovs olikhet har vi att

$$P(|W_n - \mathrm{E}[W_n]| > \varepsilon/2) \leq \frac{\mathrm{Var}[W_n]}{(\varepsilon/2)^2},$$

och därför

$$P(|W_n - \theta| > \varepsilon) \xrightarrow{P} 0$$

om $\mathrm{Var}[W_n] \longrightarrow 0$ då $n \longrightarrow \infty$.

Det följer att ett nödvändigt villkor för konsistens är att

$$\mathrm{E}[W_n] \xrightarrow{P} \theta \qquad \text{och} \qquad \mathrm{Var}[W_n] \xrightarrow{P} 0$$

för alla $\theta \in \Omega$ då $n \longrightarrow \infty$. Detta är egentligen både nödvändigvist och tillräckligt.

Exempelvis, $\bar{X}_n$ är konsistent för $\mu = \mathrm{E}[X_i]$ eftersom det är väntevärdesriktig och $\mathrm{Var}[\bar{X}_n] = \mathrm{Var}[X_i]/n \longrightarrow 0$ när $n \longrightarrow \infty$.

Vi har också att ML-skattningar är konsistent:

Sats 15.7. *Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade med täthetsfunktionen (eller sannolikhetsfunktion) $f_{X_i}(x_i|\theta)$. Låt $\hat{\theta}_n$ vara ML-skattningen för θ given $X_1, \ldots, X_n$. Anta att*

1. *$f_{X_i}(x_i|\theta) \neq f_{X_i}(x_i|\theta')$ när $\theta \neq \theta'$ (dvs. θ är* identifierbar*),*
2. *stöd av $f_{X_i}(x_i|\theta)$ beror inte på θ,*
3. *$f_{X_i}(x_i|\theta)$ är deriverbar, och*
4. *parameter mängden Ω är öppen.*

Då har vi att

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \qquad \text{då} \qquad n \longrightarrow \infty$$

för alla $\theta \in \Omega$; dvs. $\hat{\theta}_n$ är konsistent.

Vi ska inte bevisa denna sats men idéen är att

$$\frac{1}{n} \log L(\hat{\theta}_n|\boldsymbol{x}_n)$$

konvergerar till $\mathrm{E}[\log f_{X_i}(x_i|\theta)]$ för alla $\theta \in \Omega$. Med hjälp av detta faktum kan vi bevisa att $\hat{\theta}_n \xrightarrow{P} \theta$ då $n \longrightarrow \infty$.

15.3.1 Asymptotisk normalitet

Vi är ibland inte intresserad av att skatta θ men en funktion $\tau(\theta)$ istället. (Se Exempel 15.1.) Från Sats 5.2 har vi att $\tau(\hat{\theta})$ är ML-skattningen för $\tau(\theta)$ om $\hat{\theta}$ är ML-skattningen för θ. Eftersom $\hat{\theta}_n \xrightarrow{P} \theta$ i Sats 15.7 har vi också att

$$\tau(\hat{\theta}_n) \xrightarrow{P} \tau(\theta) \qquad \text{då} \qquad n \longrightarrow \infty$$

för alla $\tau(\theta)$.

Medan konsistens är en bra egenskap säger det ingenting om konvergenshastigheten. Det skulle vara bra att ha en metod för att jämföra två olika konsistenta skattningar. Ett sätt att göra det är att betrakta konvergens i fördelning. Eftersom $\tau(\hat{\theta}_n)$ kovergerar till $\tau(\theta)$ i sannolikheten har vi att $\tau(\hat{\theta}_n)$ konvergerar till $\tau(\theta)$ i fördelning. Vi har dock att $\mathrm{Var}[\tau(\hat{\theta}_n)] \longrightarrow 0$ då $n \longrightarrow \infty$. Därför är gränsfördelningen för $\tau(\hat{\theta}_n)$ bara en punkt. För att få en icke-trivial gräns betraktar vi istället

$$\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{D} W \qquad \text{då} \qquad n \longrightarrow \infty.$$

Det händer ofta att denna gräns är en normalfördelning. I detta fall säger vi att W_n är *asymptotiskt normal*; dvs.

$$\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{D} \mathrm{N}(0, \sigma^2(\theta)).$$

Variansen $\sigma^2(\theta)$ kallas för den *asymptotiska variansen*. När θ är förstod skriver vi ofta bara σ^2. För att jämföra olika konsistenta skattningar kan vi jämföra deras asymptotiska varianser. Medan σ^2 kallas för en ”varians” är det inte variansen för någon W_n. Istället är det konvergenshastigheten för sekvensen $W_1, W_2, \ldots$.

Vi såg i Sats 5.3 att den bästa möjliga variansen för en väntevärdesriktig skattning W för θ ges av Cramér-Rao nedre gränsen:

$$\frac{1}{I(\theta)}.$$

Resultatet generaliserar till oväntevärdesriktiga skattningar W som

$$\mathrm{Var}[W|\theta] \geq \frac{\tau'(\theta)^2}{I(\theta)}.$$

Låt oss betraktar en konsistent skattning W_n för $\tau(\theta)$. Eftersom W_n är konsistent har vi att det är asymptotiskt väntevärdesriktig och därmed förväntar vi oss att en asymptotisk version av Cramér-Rao nedre gränsen kommer att stämmer. Om W_n är också asymptotiskt normal och uppfyller denna asymptotisk version av Crámer-Rao kommer vi ha att

$$\sqrt{n}(W_n - \tau(\theta) \xrightarrow{D} \mathrm{N}\left(0, \frac{\tau'(\theta)^2}{J(\theta)}\right)$$

där

$$J(\theta) = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f_X(X|\theta)\right]$$

är den *normaliserade Fisher informationen*. Vi har därmed att den bästa möjliga asymptotiska variansen är

$$\frac{\tau'(\theta)^2}{J(\theta)}.$$

Vi säger att en sekvens $W_1, W_2, \ldots$ är *asymptotiskt effektiv* för $\tau(\theta)$ när vi har att

$$\sqrt{n}(W_n - \tau(\theta)) \xrightarrow{D} \mathrm{N}\left(0, \frac{\tau'(\theta)^2}{J(\theta)}\right).$$

Det händer ofta att ML-skattningar är asymptotiskt effektiv:

Sats 15.8. *Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade med täthetsfunktion (eller sannolikhetsfunktion) $f_{X_i}(x_i|\theta)$ och låt $\hat{\theta}_n$ vara ML-skattningen för θ given $X_1, \ldots, X_n$. När villkoren i Sats 15.7 och ytterligare två villkorer (som alltid håller för exemplar/uppgifter i denna kurs) har vi att $\tau(\hat{\theta}_n)$ är asymptotiskt effektiv för $\tau(\theta)$.*

Vi kan använda denna sats för att skatta variansen för $\tau(\hat{\theta}_n)$ när n är stort. Vi har att

$$\mathrm{Var}[\tau(\hat{\theta}_n)|\theta] \approx \frac{\tau'(\theta)^2}{nJ(\theta)}.$$

Vi kan inte använda denna för att få en approximation för $\mathrm{Var}[\tau(\hat{\theta}_n)|\theta]$ direkt eftersom täljaren och numeraren i kvoten på höger sidan beror på θ (som är okänt). Vi kan approximerar dessa värde i följande sätt: Först kan vi anta att $\hat{\theta}_n$ är tillräckligt nära till θ så att vi kan approximerar $\tau(\theta) \approx \tau(\hat{\theta}_n)$ respektive $J(\theta) \approx J(\hat{\theta}_n)$.

Eftersom $J(\theta)$ kräver ett väntevärde med avseende på X behöver kanske vi också en approximation för denna funktion. Det kan vi göra som

$$\begin{aligned} J(\theta) &= -\operatorname{E}\left[\frac{d^2}{d\theta^2}\log f_X(x|\theta)\right], \\ &\approx -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f_X(x_i|\theta), \\ &= -\frac{1}{n}\frac{d^2}{d\theta^2}L(\theta|\boldsymbol{x}). \end{aligned}$$

Nämnaren approximeras därför som

$$nJ(\theta) \approx n\hat{J}(\hat{\theta}) = -\frac{d^2}{d\theta^2}\log L(\hat{\theta}|\boldsymbol{x})$$

som kallas för den *observerade Fisher informationen* och betecknas $\hat{I}_n(\hat{\theta}_n)$. Vi har därför en approximation för $\operatorname{Var}[\tau(\hat{\theta}_n)|\theta]$

$$\hat{\operatorname{Var}}[\tau(\hat{\theta})n)|\theta] \approx \frac{\tau'(\hat{\theta}_n)^2}{\hat{I}_n(\hat{\theta}_n)} = \frac{\tau'(\hat{\theta}_n)^2}{-\frac{d^2}{d\theta^2}\log L(\hat{\theta}_n|\boldsymbol{x})}.$$

Exempel 15.2. Låt $X_1, X_2, \dots$ vara oberoende och $\operatorname{Ber}(\theta)$-fördelade. Då har vi att $\hat{\theta}_n = \bar{X}_n$ given $X_1, \dots, X_n$. Med hjälp av invarians egenskapen för ML-skattningar har vi då att ML-skattningen för oddset $\theta/(1-\theta)$ är

$$\frac{\hat{\theta}_n}{1-\hat{\theta}_n}.$$

Här $\tau(x)$ ges av

$$\tau(x) = \frac{x}{1-x}$$

och

$$\tau'(x) = \frac{1}{(1-x)^2}.$$

Vi har också att

$$f_{X_i}(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

och vi har därmed att

$$\frac{d^2}{d\theta^2}\log f_{X_i}(x_i|\theta) = -\frac{n\hat{\theta}_n}{\theta^2} - \frac{n(1-\hat{\theta}_n)}{(1-\theta)^2}.$$

Det följer att

$$\hat{I}_n(\hat{\theta}_n) = -\left(-\frac{n\hat{\theta}_n}{\hat{\theta}_n^2} - \frac{n(1-\hat{\theta}_n)}{(1-\hat{\theta}_n)^2}\right) = \frac{n}{\hat{\theta}_n(1-\hat{\theta}_n)}.$$

Därför har vi att

$$\hat{\operatorname{Var}}\left[\frac{\hat{\theta}_n}{1-\hat{\theta}_n}\middle|\theta\right] = \frac{\frac{1}{(1-\hat{\theta}_n)^4}}{\frac{n}{\hat{\theta}_n(1-\hat{\theta}_n)}} = \frac{\hat{\theta}_n}{n(1-\hat{\theta}_n)}.$$

Notisera att denna varians är samma som den som vi få med centrala gränsvärdessatsen och deltametoden.

15.3.2 Asymptotisk relativ effektivitet

Asymptotisk normalitet och asymptotisk effektivitet kan vara användbara när vi skulle vilja jämföra två olika konsistenta skattningar V_n och W_n. Ett sätt att göra det är att beräkna det asymptotiska relativa effektivitetet för V_n relativ till W_n. När

$$\begin{aligned} \sqrt{n}(V_n - \tau(\theta)) &\xrightarrow{D} \operatorname{N}(0, \sigma_V^2), \\ \sqrt{n}(W_n - \tau(\theta)) &\xrightarrow{D} \operatorname{N}(0, \sigma_W^2) \end{aligned}$$

är det *asymptotiska relativa effektivitetet* för V_n relativ till W_n definieras som

$$ARE(V_n, W_n|\theta) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Ett stort värde för ARE föreslår att asymptotisk variansen för V_n är relativt liten till variansen för W_n. Värdet för ARE beror på värdet för θ så vi måste vara försiktigt när vi använder ARE för att utvärdera den relativa prestanden. Eftersom ML-skattningar är asymptotiska effektiva har vi att $ARE \leq 1$ när W_n är ML-skattningen. När värdet för ARE är nära till 1 i detta fall kan V_n vara användbar om ML-skattningen är svårt att beräkna, exempelvis.

Exempel 15.3. Låt $X_1, X_2, \dots$ vara oberoende och Po(θ)-fördelade. Vi skulle vilja skatta $P(X_1 = 0) = e^{-\theta}$. Ett sätt att göra det är att räkna antalet 0:er i stickprovet $X_1, \dots, X_n$ och delar det med n; dvs. vi definierar stokastiska variablen

$$Y_i = \mathbf{1}_n(X_i),$$

och betraktar skattningen $V_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Eftersom $Y_1, \dots, Y_n$ är oberoende och Ber$(e^{-\theta})$-fördelade har vi att

$$\mathrm{E}[V_n] = e^{-\theta} \qquad \text{och}\, \mathrm{Var}[V_n] = \frac{e^{-\theta}(1 - e^{-\theta})}{n}.$$

Vi har också att den asymptotiska variansen för V_n är $e^{-\theta}(1 - e^{-\theta})$. (Detta följer från den centrala gränsvärdessatsen.)

För ML-skattningen har vi $W_n = e^{-\hat{\theta}_n}$ där $\hat{\theta}_n = \bar{X}_n$ är ML-skattningen för θ given $X_1, \dots, X_n$. Enligt den centrala gränsvärdessatsen har vi att

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathrm{N}(0, \mathrm{Var}[X_i]) = \mathrm{N}(0, \lambda).$$

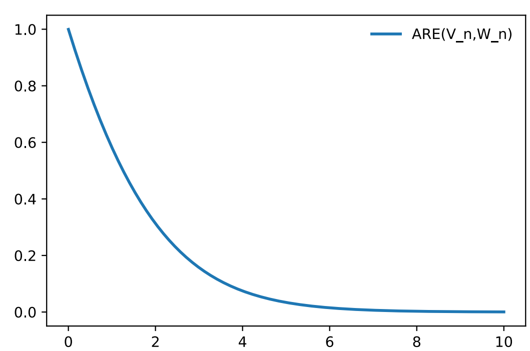
Med hjälp av deltametoden där $g(x) = e^{-x}$ och $g'(x) = -e^{-x}$ får vi

$$\sqrt{n}(V_n - e^{-\theta}) \xrightarrow{D} \mathrm{N}(0, \theta e^{-2\theta}).$$

Det följer att

$$ARE(V_n, W_n) = \frac{\theta e^{-2\theta}}{e^{-\theta}(1 - e^{-\theta})} = \frac{\theta}{e^{\theta}(1 - e^{-\theta})} = \frac{\theta}{e^{\theta} - 1}.$$

När vi plotta $ARE(V_n, W_n)$ som en funktion på $\theta > 0$ ser vi att det är nära till 1 för liten λ men det blir mindre och mindre då $\theta \longrightarrow \infty$. Det stämmer eftersom skattningen W_n kastar bort alla $X_i \neq 0$ och dessa blir många när θ blir stort:



15.4 Asymptotiska egenskaper av posteriori fördelningen

Vi kan också betrakta asymptotiska överväganden för Bayesiansk modeller. Det kan visas att våra två vanliga bayesianska punktskattningar, aposteriorivändevärdet och aposterioritypvärdet

$$\mathrm{E}[\Theta|\boldsymbol{X} = \boldsymbol{x}] \qquad \text{och} \qquad \hat{\theta} = \mathrm{argmax}_{\theta \in \Omega}\, f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}),$$

är konsistenta för θ_0 givet att data dras från $f_{X|\Theta}(x|\theta_0)$. Beviset går på samma sätt som för ML-skattningar.

Det bayesianska perspektivet föredrar att inte anta sådana saker. Så är det bättre att studera den asymptotiska formen för hela aposteriorifördelningen. För att göra det betrakar vi taylorpolynomet $T_2(x)$ för log-aposteriorfördelningen runt aposterioritypvärdet $\hat{\theta}$:

$$\begin{aligned}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) &\approx \log f_{\Theta|\boldsymbol{X}}(\hat{\theta}|\boldsymbol{x}) + \frac{d}{d\theta}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + \frac{1}{2}\frac{d^2}{d\theta^2}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2,\\ &\approx \log f_{\Theta|\boldsymbol{X}}(\hat{\theta}|\boldsymbol{x}) + \frac{1}{2}\frac{d^2}{d\theta^2}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2,\end{aligned}$$

där vi använde

$$\frac{d}{d\theta}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}} = 0$$

eftersom $\hat{\theta}$ är aposterioritypvärdet. Det följer att

$$f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) \propto\approx e^{-\frac{\hat{I}(\hat{\theta})}{2}(\theta-\hat{\theta})^2}$$

där

$$\hat{I}(\hat{\theta}) = -\frac{d^2}{d\theta^2}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}}$$

är den *bayesianska observerade informationen*. Det föler att

$$\Theta|\boldsymbol{X}=\boldsymbol{x} \approx \mathrm{N}(\hat{\theta}, (\hat{I}(\hat{\theta}))^{-1}).$$

Exempel 15.4. Låt $X_1, X_2, \ldots$ vara oberoende och Ber(θ)-fördelade där $\Theta \sim \text{Beta}(\alpha_0, \beta_0)$. Vi har att

$$\Theta|\boldsymbol{X}=\boldsymbol{x} \sim \text{Beta}\left(\alpha_0 + \sum_{i=1}^n x_i, \beta_0 + n - \sum_{i=1}^n x_i\right) = \text{Beta}(\alpha_1, \beta_1).$$

Så har vi att

$$\begin{aligned}\hat{I}(\hat{\theta}) &= -\frac{d^2}{d\theta^2}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\bigg|_{\theta=\hat{\theta}},\\ &= (\alpha_1-1)\frac{1}{\hat{\theta}^2} + (\beta_1-1)\frac{1}{(1-\hat{\theta})^2}\end{aligned}$$

Vi har också att aposterioritypvärdet för Beta(α_1, β_1) är

$$\hat{\theta} = \frac{\alpha_1 - 1}{\alpha_1+\beta_1-2}.$$

Därför är

$$\hat{I}(\hat{\theta}) = \frac{(\alpha_0+\beta_0+n-2)^3}{(\alpha_0+\sum_{i=1}^n x_i - 1)(\beta_0+n-\sum_{i=1}^n x_i - 1)}.$$

Asymptotiskt har vi då att

$$\Theta|\boldsymbol{X}=\boldsymbol{x} \sim \mathrm{N}\left(\frac{\alpha_0+\sum_{i=1}^n x_i - 1}{\alpha_0+\beta_0+n-2}, \frac{(\alpha_0+\sum_{i=1}^n x_i - 1)(\beta_0+n-\sum_{i=1}^n x_i - 1)}{(\alpha_0+\beta_0+n-2)^3}\right).$$

Notisera att för den likformiga apriorifördelningen $\Theta \sim \text{Beta}(1,1)$ förenklar detta till

$$\Theta|\boldsymbol{X}=\boldsymbol{x} \sim \mathrm{N}\left(\bar{x}_n, \frac{1}{n}\bar{x}_n(1-\bar{x}_n)\right).$$

Generellt sett ser vi att $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ går mot

$$\mathrm{N}\left(\theta_0, \frac{1}{nJ(\theta_0)}\right)$$

då $n \longrightarrow \infty$ givet att $\boldsymbol{x}$ dras från $f_{X|\Theta}(x|\theta_0)$. Detta stämmer eftersom $\hat{\theta}$ är konsistent för θ_0 och därför $\hat{\theta}_n \xrightarrow{P} \theta_0$ då $n \longrightarrow \infty$. Det gäller också att

$$\begin{aligned}\hat{I}(\theta) &= -\left.\frac{d^2}{d\theta^2}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\right|_{\theta=\hat{\theta}}, \\ &= -\frac{d^2}{d\theta^2}\log f_{\Theta}(\theta) - \sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f_{X_i|\Theta}(x_i|\theta)\end{aligned}$$

Medan den första terminen är konstant med avseende på n ökar den andra terminen då $n \longrightarrow \infty$. Eftersom varje termin i summan har väntevärdet $J(\theta_0)$ går summan mot $nJ(\theta_0)$ enligt stora talens lagen. Aprioriterminen domineras av summan så får vi resultatet. Notisera att denna asymptotisk varians $1/J(\theta_0)$ är samma som vi får med ML-skattningen. Med andra ord är aposteriorifördelningen asymptotiskt normal men det är också asymptotiskt effektiv (på frekventiskt meningen).