Föreläsning 9

Thomas Önskog

23/11 2017

Statistisk inferensteori

Nuförtiden skapas kontinuerligt mycket stora datamängder, det som populärt benämns Big Data, för exempelvis finansiella transaktioner, konsumtionsmönster och sociala nätverk osv. Statistikteorin ger oss matematiska metoder för att omsätta dessa datamängder i relevant kunskap. Inom statistikteorin arbetar vi ofta med följande problemställning. Vi har en mängd data som vi antar kommer från en sannolikhetsfördelning vars parametrar är okända. Vi vill använda datat för att uppskatta parametervärdena så bra som möjligt. Detta kallas **statistisk inferens** (= slutledning).

Låt $x_1, ..., x_n$ beteckna ett **stickprov**, dvs en samling observationer av de oberoende och likafördelade s.v. $X_1, ..., X_n$. Den gemensamma fördelningen fördelningen för X_i är i allmänhet inte helt känd, utan beror på en en- eller flerdimensionell **parameter** θ. De möjliga värdena på θ bildar ett **parameterrum** som betecknas Ω_θ.

Exempel. *Om $X_i \in N(\mu, \sigma)$, så är $\theta = (\mu, \sigma)$ och $\Omega_\theta = \mathbb{R} \times (0, \infty)$.*

Definition. *En **punktskattning** av en parameter θ är en funktion av stickprovet $x_1, ..., x_n$ som för varje val av data ger ett värde θ^*_{obs} på parametern θ, dvs $\theta^*_{obs} = \theta^*(x_1, ..., x_n)$.*

Exempel. *Fördelningen av löner antas ofta vara lognormalfördelade, dvs logaritmen av lönen är normalfördelad. Om vi antar att standardavvikelsen i lognormalfördelningen är känd ($\sigma = 1/4$), så ges månadslönen X hos en heltidsarbetande löntagare i Sverige av täthetsfunktionen*

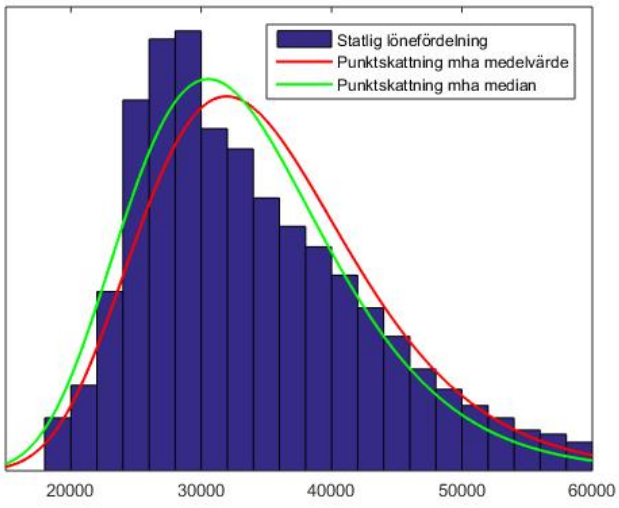
$$f_X(x) = \frac{4}{x\sqrt{2\pi}} e^{-8(\ln x - \mu)^2},$$

där μ är en okänd parameter som utgör ett lägesmått för $\ln X$.

Betrakta nu ett stickprov $x_1, ..., x_n$ som består av heltidslönerna hos samtliga statsanställda i Sverige under 2016. Två naturliga lägesmått för lönedatat är stickprovsmedelvärdet

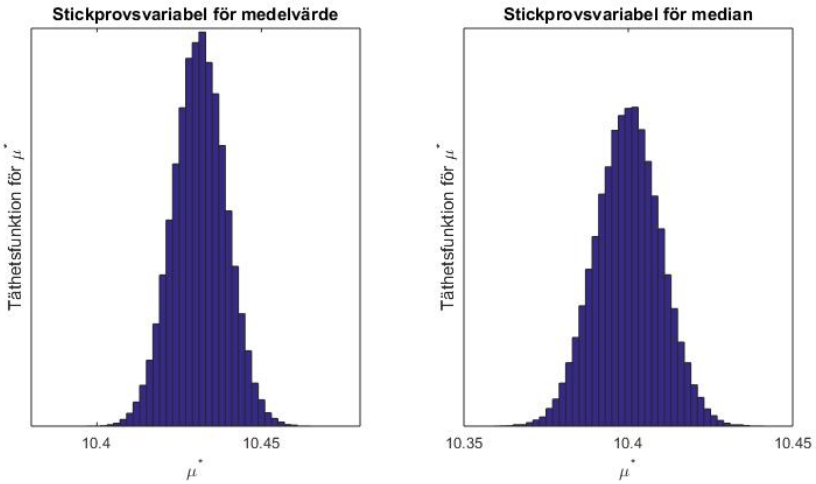
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 34\,000$$

av lönerna och stickprovsmedianen $\widetilde{x}_{0.5} = 32\,500$ av lönerna (den mittersta lönen i datat). Medellönen blir lite större än medianlönen eftersom de högsta lönerna drar upp medellönen mer än medianlönen. Motsvarande punktskattningar på parametern μ är $\mu^*_{obs} = \ln \overline{x} = 10.43$ respektive $\mu^*_{obs} = \ln \widetilde{x}_{0.5} = 10.39$. Figuren visar fördelningen för stickprovet och fördelningen för lognormalfördelningen med de båda punktskattningarna på parametern.



Stickprovet $x_1, ..., x_n$ kan ses som ett utfall av de s.v. $X_1, ..., X_n$. Därmed blir $\theta^*_{\text{obs}} = \theta^*(x_1, ..., x_n)$ ett utfall av den s.v. $\theta^*(X_1, ..., X_n)$ som kallas för **stickprovsvariabeln**. Stickprovsvariabelns fördelning beror på fördelningarna för $X_1, ..., X_n$ som i sin tur beror av θ. Vi skriver ofta θ^* istället för $\theta^*(X_1, ..., X_n)$.

Exempel. *I exemplet ovan är stickprovsvariabeln antingen logaritmen av medelvärdet $\overline{X}$ av n oberoende lognormalfördelade s.v. med parameter μ eller logaritmen av medianen $\widetilde{X}_{0.5}$ av n oberoende lognormalfördelade s.v. med parameter μ. Fördelningarna för båda dessa s.v. är relativt komplicerade, men figuren nedan visar fördelningarna för dessa båda stickprovsvariabler för $n = 1000$ givet att $\mu = 10.4$. Om vi har noterat lönen för $n = 1000$ slumpmässigt valda personer ur en population där lönerna är lognornalfördelade med parameter $\mu = 10.4$, så kommer μ^*_{obs} att vara utfall från nedanstående fördelningar.*



Vi ska nu undersöka tre kriterier för vad som är en god punktskattning och sedan belysa dem med ett exempel.

Definition. *En punktskattning θ^*_{obs} är **väntevärdesriktig** om $\underline{E(\theta^*) = \theta}$ för alla $\theta \in \Omega_\theta$.*

En väntevärdesriktig punktskattning har inget systematiskt fel. Vi önskar också att skattningarna ska bli bättre ju mer data som vi har tillgängligt. För att kunna kvantifiera detta låter vi nedan θ_n^* beteckna en stickprovsvariabel baserad på ett stickprov med n element.

Definition. *En punktskattning θ_{obs}^* är **konsistent** om det för varje fixt $\theta \in \Omega_\theta$ och varje fixt $\epsilon > 0$ gäller att $\mathbb{P}(|\theta_n^* - \theta| > \epsilon) \to 0$, då $n \to \infty$.*

Det är också önskvärt att välja en punktskattning så att stickprovsvariabeln har så liten spridning som möjligt.

Definition. *Om θ_{obs}^* och $\widehat{\theta}_{obs}$ är två väntevärdesriktiga punktskattningar och $V(\theta^*) \leq V(\widehat{\theta})$ för alla $\theta \in \Omega_\theta$ (med strikt olikhet för något θ), så sägs θ^* vara mer **effektiv** än $\widehat{\theta}$.*

Vi illustrerar kriterierna med följande exempel. Låt $x_1, ..., x_n$ vara mätningar av en fysikalisk konstant gjorda med ett instrument som har ett mätfel med känd standardavvikelse σ. Stickprovet kan antas vara utfall av oberoende s.v. $X_1, ..., X_n$ sådana att $X_i \in \mathrm{N}(\theta, \sigma)$, där θ är det sanna värdet på konstanten. Som punktskattning av θ kan vi välja **stickprovsmedelvärdet** $\theta_{\text{obs}}^* = \overline{x}$ och motsvarande stickprovsvariabel blir då $\theta^* = \overline{X} \in \mathrm{N}(\theta, \sigma/\sqrt{n})$. Punktskattningen är väntevärdesriktig eftersom

$$E(\theta^*) = E(\overline{X}) = \theta.$$

Att punktskattningen är konsistent visas på samma sätt som stora talens lag. Vidare är $V(\theta^*) = \sigma/\sqrt{n}$, vilket kan användas för att jämföra denna punktskattningar med andra punktskattningar av θ med avseende på effektivitet. Stickprovsmedelvärdet är inte den mest effektiva skattningen av väntevärdet i alla möjliga situationer, mer ger ofta en god skattning.

Om det istället är variansen σ^2 som är okänd, så används ofta **stickprovsvariansen**

$$\theta_{\text{obs}}^* = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

som punktskattning. Även denna punktskattning är väntevärdesriktig och konsistent. Kravet på väntevärdesriktighet är en anledning till att vi dividerar med $n-1$ istället för med n i definitionen av stickprovsvariansen. Liksom för stickprovsmedelvärdet, ger stickprovsvariansen ofta goda skattningar, men den är inte den mest effektiva skattningen i alla situationer.

Minsta kvadrat- och maximum likelihoodskattningar

Det finns analytiska metoder för att härleda den bästa punktskattningen i en given situation. De båda vanligaste sådana metoderna är **minsta kvadratmetoden** (MK) och **maximum likelihoodmetoden** (ML). I definitionerna av båda metoderna låter vi $x_1, ..., x_n$ vara ett stickprov av de oberoende s.v. $X_1, ..., X_n$.

Definition. ***MK-skattningen*** *θ_{MK}^* av parametern θ är det värde i Ω_θ som minimerar **kvadratsumman***

$$Q(\theta) = \sum_{i=1}^{n}(x_i - E(X_i))^2,$$

där $E(X_i)$ beror på den okända parametern θ.

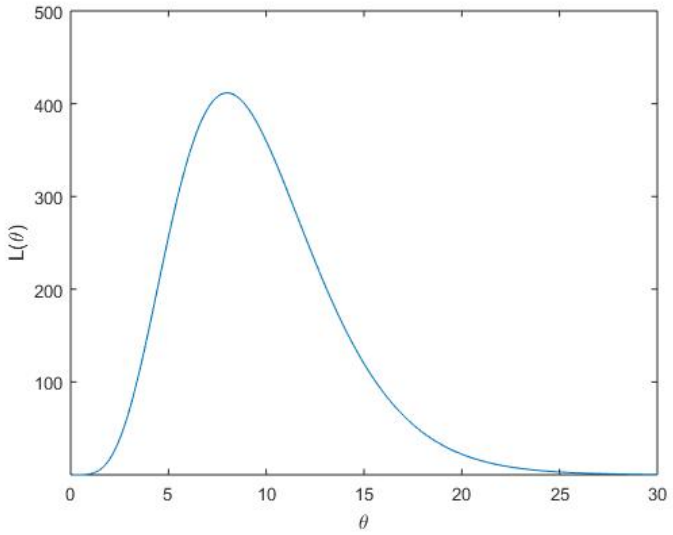
Om exempelvis $X_i \in \text{Exp}(\theta)$, så är $E(X_i) = 1/\theta$. Läs mer om MK-skattningar på egen hand i boken.

Definition. ***ML-skattningen*** θ^*_{ML} *av parametern* θ *är det värde i* Ω_θ *som maximerar* ***likelihoodfunktionen***

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta),$$

om X_i *är kontinuerlig. Om* X_i *är diskret, så byts* f_{X_i} *mot* p_{X_i}.

Likelihoodfunktionen ger ett mått på sannolikheten att stickprovet $x_1, ..., x_n$ uppkommer för olika värden på parametern θ. ML-skattningen θ^*_{ML} är det parametervärde för vilket det är som mest troligt att vi får just det stickprov som vi har fått. Bilden nedan visar ett exempel på hur en likelihoodfunktionen kan se ut.



Produkten i definitionen av L gör ofta räkningarna komplicerade, så vanligen räknar vi med

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f_{X_i}(x_i; \theta),$$

istället för $L(\theta)$. Detta går lika bra eftersom $L(\theta)$ och $\ln L(\theta)$ antar maximum för samma värde på θ.

Exempel. *Vi har stickprovet* $x_1 = 0.77$, $x_2 = 0.82$, $x_3 = 0.92$, $x_4 = 0.94$ *och* $x_5 = 0.98$ *som antas komma från fördelningen* $f_X(x) = \theta x^{\theta-1}$, *för* $x \in [0, 1]$ *där* $\theta > 0$ *är en okänd parameter. Vi vill bestämma ML-skattningen av* θ *och bestämmer därför först likelihoodfunktionen och dess logaritm.*

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n (x_1 \cdots x_n)^{\theta-1}.$$

$$\ln L(\theta) = \ln(\theta^n (x_1 \cdots x_n)^{\theta-1}) = n \ln \theta + (\theta - 1) \ln(x_1 \cdots x_n).$$

Derivering med avseende på den okända parametern θ ger nu

$$\frac{d}{d\theta}\ln L(\theta) = \frac{n}{\theta} + \ln(x_1 \cdots x_n) = 0 \qquad \Rightarrow \qquad \frac{n}{\theta^*_{ML}} = -\ln(x_1 \cdots x_n)$$
$$\Rightarrow \qquad \theta^*_{ML} = -\frac{n}{\ln(x_1 \cdots x_n)},$$

så för det givna stickprovet så är $\theta^*_{ML} = -5/\ln(0.77 \cdot 0.82 \cdot 0.92 \cdot 0.94 \cdot 0.98) = 8.0$. *Notera att figuren ovan visar likelihoodfunktionen i detta exempel och vi ser från figuren att likelihoodfunktionen maximeras för* $\theta \approx 8$. *Toppen på likelihoodfunktionen är dock relativt flack, så värden inom intervallet* $\theta \in (6, 10)$ *ger ungefär lika stora värden på* $L(\theta)$.