Föreläsning 8: Konfidensintervall

Matematisk statistik

David Bolin
Chalmers University of Technology
September 23, 2015

Statistik

Stickprov

Ett stickprov av storlek n är n oberoende observationer av en slumpvariabel X. Vi kan alltså skriva stickprovet som observationer av n slumpvariabler $X_1, \ldots, X_n$ där alla X_i är oberoende och likafördelade.

I praktiken kan vi ofta resonera oss fram till vilken typ av fördelning som bör användas för X, men vi vet typiskt inte vilka värden vi ska anta för fördelningens parametrar.

Många frågor vi har när vi tillämpar statistikteori kan reduceras till följande matematiska fråga: Givet observationer $x_1, \ldots, x_n$, vad kan vi säga om parametrarna i fördelningen som genererade stickprovet?

Punktskattningar

Skattning

En skattning av en parameter θ är en funktion $\hat{\theta}(X_1, \ldots, X_n)$ av observationerna.

En skattning kan avse både en slumpvariabel och ett tal:

- $\hat{\theta}(X_1, \ldots, X_n)$ är en slumpvariabel som har en viss fördelning som ger information om hur bra skattningen är.
- $\hat{\theta}(x_1, \ldots, x_n)$ är ett tal beräknat från data, detta sägs vara vår punktskattning av parametern.

Två egenskaper vi vill att vår skattare ska ha:

- *väntevärdesriktig*: $\mathsf{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$.
- Den ska ha låg varians om n är stort, helst vill vi att $\mathsf{V}(\hat{\theta}(X_1, \ldots, X_n)) \to 0$ då $n \to \infty$.

Att en skattare är väntevärdesriktig betyder att värdena i medel kommer ligga kring det sanna värdet om vi gör flera upprepade stickprov av storlek n. För ett givet stickprov behöver värdet inte ligga i närheten av μ, och variansen hos skattaren ger en indikation på hur långt ifrån det sanna värdet det kan ligga.

Ofta anger man det så kallade standardfelet för skattaren, vilket ges av standardavvikelsen för skattaren.

Exempel

Antag att $\mathsf{E}(X_i) = \mu$ och $\mathsf{V}(X_i) = \sigma^2$. Stickprovsmedelvärdet $\bar{X}$ är en väntevärdesriktig skattare av väntevärdet μ, dvs $\mathsf{E}(\bar{X}) = \mu$. Variansen hos skattaren ges av $\mathsf{V}(\bar{X}) = \frac{\sigma^2}{n}$. En väntevärdesriktig skattare av variansen σ^2 ges av stickprovsvariansen

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Moment

I allmänhet kan det vara svårt att hitta på en bra skattare, vi behöver därför någon systematisk metod för att ta fram skattare. En vanlig metod är momentmetoden.

Moment

För en slumpvariabel X defineras det p:te momentet som $\mathsf{E}(X^p)$.

Stickprovsmoment

För ett stickprov $x_1, \ldots, x_n$ definieras det p:te stickprovsmomentet som

$$M_p = \frac{1}{n}\sum_{i=1}^{n} x_i^p$$

Momentmetoden bygger på att vi ofta kan uttrycka momenten för X som funktioner av parametrarna.

Momentmetoden

Antag att vi har K parametrar vi vill skatta.

1. Beräkna momenten $\mathsf{E}(X), \dots, \mathsf{E}(X^K)$ och motsvarande sampelmoment $M_1, \dots, M_K$.
2. Lös ekvationssystemet

$$\begin{aligned} \mathsf{E}(X^1) &= M_1 \\ \mathsf{E}(X^2) &= M_2 \\ &\vdots \\ \mathsf{E}(X^K) &= M_K \end{aligned}$$

   med avseende på parametrarna.
3. De uttryck för parametrarna som löser ekvationssystemet är våra parameterskattningar.

Maximum likelihood-metoden

Tanken bakom maximum likelihood metoden är att vi vill hitta det parametervärde som mest troligast producerade det stickprov vi har. Metoden är baserad på den så kallade likelihood-funktionen, som för ett stickprov $x_1 \ldots, x_n$ är

$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

där $f(x)$ är täthetsfunktionen för fördelningen och θ är parametrarna vi vill skatta. I det diskreta fallet byter vi täthetsfunktionen mot sannolikhetsfunktionen.

ML skattare

Maximum likelihood-skattaren av en parameter θ ges av

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta)$$

Maximum likelihood-metoden

Eftersom logaritmen är en monoton och ökande funktion kan ML skattningen även beräknas som

$$\hat{\theta}_{ML} = \arg\max_{\theta} l(\theta)$$

där $l(\theta) = logL(\theta)$ är log-likelihood funktionen.

Det är oftare enklare att maximera denna ekvation, eftersom vi har att

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i).$$

Vi maximerar $l(\theta)$ genom att derivera funktionen med avseende på θ, sätta derivatan lika med noll, och lösa ut θ.

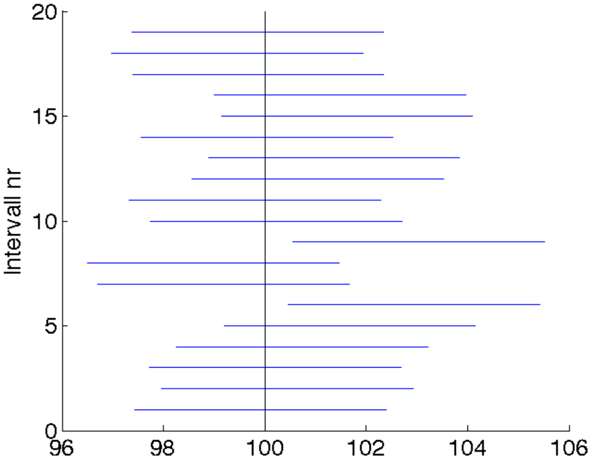
Konfidensintervall

Konfidensintervall

Låt $X_1, \ldots, X_n$ vara slumpvariabler med en fördelning som har θ som en parameter med θ_0 som sant okänt värde. Ett $100(1-\alpha)\%$ konfidensintervall för θ med konfidensgraden $1-\alpha$ är ett intervall $[a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ sådant att

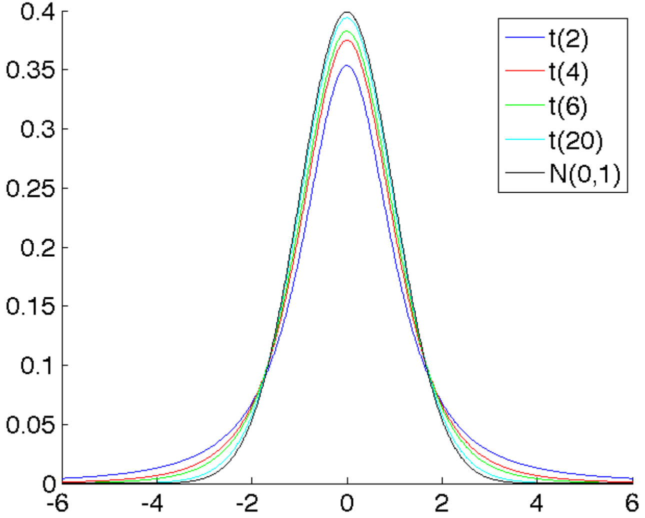
$$\mathsf{P}(a \leq \theta_0 \leq b) = 1 - \alpha.$$

- (a, b) är ett slumpmässigt intervall, eftersom a och b är slumpvariabler som beror av $X_1, \ldots, X_n$.
- Konfidensintervallet ska alltså tolkas som att om vi gör upprepade mätningar av $X_1, \ldots, X_n$ och bildar konfidensintervallet för alla dessa mätningar så kommer $100(1-\alpha)\%$ av dessa intervall täcka det sanna värdet θ_0.



20 konfidensintervall för μ, där vart och ett är bildat av 10 N$(100, 16)$-observationer.

$t(n)$-fördelningen



$\chi^2(n)$-fördelningen