

Föreläsning 8: Konfidensintervall

Matematisk statistik

David Bolin
Chalmers University of Technology
September 20, 2017



Stickprov

Ett stickprov av storlek n är n oberoende observationer av en slumpvariabel X . Vi kan alltså representera stickprovet som n slumpvariabler $X_1, \dots, X_n$ 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 inte vilka värden vi ska använda för fördelningens parametrar.

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

Punktskattningar

Skattning

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

En skattning kan avse både en slumpvariabel och ett tal:

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

Två egenskaper vi typiskt vill att vår skattning ska ha:

- *väntevärdesriktig*: $E(\hat{\theta}(X_1, \dots, X_n)) = \theta$.
- Låg varians, helst $V(\hat{\theta}(X_1, \dots, X_n)) \rightarrow 0$ då $n \rightarrow \infty$.

Att en skattning är väntevärdesriktig betyder att den 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 det sanna värdet.
- Variansen hos skattningen ger en indikation på hur långt ifrån det sanna värdet det kan ligga.
- Ofta anger man det så kallade standardfelet för skattningen, vilket ges av standardavvikelsen för skattningen.

Exempel

Antag att $E(X_i) = \mu$ och $V(X_i) = \sigma^2$. Då är stickprovsmedelvärdet $\hat{\mu} = \bar{X}$ en väntevärdesriktig skattare av μ , dvs $E(\hat{\mu}) = \mu$. Variansen hos skattaren ges av $V(\hat{\mu}) = \frac{\sigma^2}{n}$. En väntevärdesriktig skattare av variansen σ^2 ges av stickprovsvariansen $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Momentmetoden

Antag att vi har K parametrar vi vill skatta.

- 1 Beräkna momenten $E(X), \dots, E(X^K)$ och motsvarande stickprovsmoment $M_p = \frac{1}{n} \sum_{i=1}^n x_i^p$, för $p = 1, \dots, K$
- 2 Lös följande ekvationssystem med avseende på parametrarna

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

- 3 De uttryck för parametrarna som löser ekvationssystemet är våra parameterskattningar.

Maximum likelihood-metoden

Tanken bakom ML-metoden är att vi vill hitta det parametervärde som gör att det stickprov vi har blir så troligt som möjligt.

Metoden är baserad på den så kallade likelihood-funktionen, som för ett stickprov $x_1 \dots, x_n$ är

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

där $f_{\theta}(x)$ är täthetsfunktionen eller sannolikhetsfunktionen för fördelningen och θ är parametrarna vi vill skatta.

ML skattning

Maximum likelihood-skattningen av en parameter θ ges av

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

Maximum likelihood-metoden

ML-skattningen även beräknas som

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

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

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

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

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

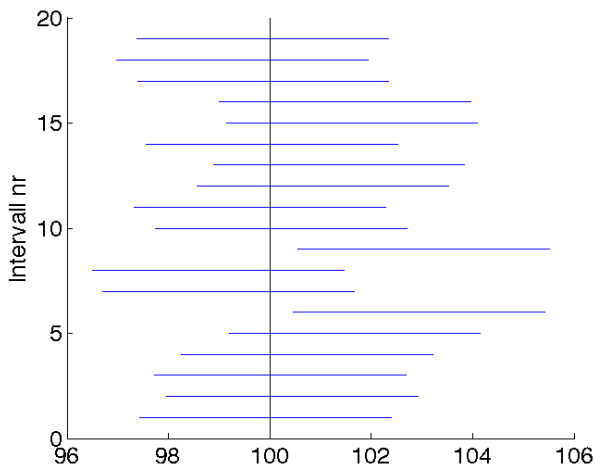
Konfidsensintervall

Konfidsensintervall

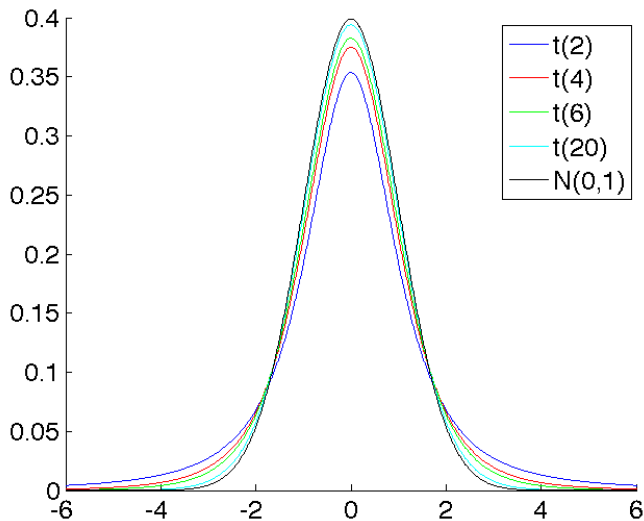
Låt $X_1, \dots, X_n$ vara slumpvariabler med en fördelning som har θ som en parameter med θ_0 som sant okänt värde. Ett $100(1 - \alpha)\%$ konfidsensintervall för θ med konfidsensgraden $1 - \alpha$ är ett intervall $[a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$ sådant att

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

- (a, b) är ett slumpmässigt intervall, eftersom a och b är slumpvariabler som beror av $X_1, \dots, X_n$.
- Konfidsensintervallet ska alltså tolkas som att om vi gör upprepade mätningar av $X_1, \dots, X_n$ och bildar konfidsensintervallet 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