

MVE051/MSG810 2016 Föreläsning 8

Petter Mostad

Chalmers

November 30, 2016

- ▶ Ett $100(1 - \alpha)\%$ konfidensintervall för en parameter θ är en formel som från ett stickprov $X_1, \dots, X_n$ beräknar stokastiska variabler L_1 och L_2 så att, för alla θ ,

$$\Pr[\theta \in [L_1, L_2]] = 1 - \alpha$$

- ▶ Korrekt tolkning av konfidensintervall: Om man N gånger genererar *nya* data från fördelningen och från dessa beräknar *nya* konfidensintervall för parametern θ enligt formeln, så vill $100(1 - \alpha)\%$ av dessa innehålla θ , när $N \rightarrow \infty$.
- ▶ FELAKTIG tolkning av konfidensintervall: Om man beräknar ett konfidensintervall från givna data så är det $100(1 - \alpha)\%$ sannolikt att θ är i intervallet.
- ▶ Dock: Med vissa extra antaganden kan den felaktiga tolkningen bli korrekt, och då ger konfidensintervallet nyttig information om θ i din modell. De extra antaganden är oftast rimliga. Därmed har konfidensintervallet blivit populärt.

Exempel 1: Konfidensintervall för μ i Normal(μ, σ)-fördelningen

- ▶ Om $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma)$ är ett stickprov, om $z_{\alpha/2}$ är så att $[-z_{\alpha/2}, z_{\alpha/2}]$ innehåller 100(1 - α)% av sannolikheten i standard normalfördelning, och om vi definerar

$$L_1 = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$$

$$L_2 = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$$

så är $[L_1, L_2]$ ett 100(1 - α)% konfidensintervall för μ .

- ▶ Bevis: Baseras på att vi har visat att $\bar{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.
- ▶ Märk hur vi hittar $z_{\alpha/2}$ i tabellen för standard normalfördelning. Traditionellt används $\alpha = 0.05$, som ger $z_{0.05/2} = 1.96$.
- ▶ Märk: Formlerna för L_1 och L_2 innehåller σ , så detta intervall kan bara användas om *fördelningens* varians σ^2 är känd (det räcker inte med stickprovsvariansen beräknat från data!)

Fördelningen för variansestimatorn $S^2 = \hat{\sigma}^2$

- ▶ Vi såg att konfidensintervallet för μ konstruerades ut i från att vi kände fördelningen till en estimator för μ . På samma sätt skall vi konstruera ett konfidensintervall för σ^2 ut i från fördelningen till estimatorn $\hat{\sigma}^2$.
- ▶ Om $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma)$ är ett stickprov och vi definerar estimatorn

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

så har estimatorn en fördelning så att

$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1),$$

alltså $(n-1)S^2/\sigma^2$ har en χ^2 fördelning med $n-1$ frihetsgrader.

- ▶ Bevis: Ett bevis kan sättas i hop genom att
 - ▶ Först visa att S^2 och $\bar{X}$ är oberoende stokastiska variabler (kan visas med momentgenererande funktioner).
 - ▶ Sen använder man detta för att beräkna den moment-genererande funktionen till $(n-1)S^2/\sigma^2$ och visa att dens form motsvarar $\chi^2(n-1)$ fördelningen.

Exempel 2: Konfidensintervall för σ^2

- ▶ Om $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma)$ är ett stickprov, om $\chi_{1-\alpha/2}^2$ och $\chi_{\alpha/2}^2$ är så att $[\chi_{1-\alpha/2}^2, \chi_{\alpha/2}^2]$ innehåller 100(1 - α)% av sannolikheten i en $\chi^2(n-1)$ fördelning, och om vi definierar

$$\begin{aligned}L_1 &= (n-1)S^2/\chi_{\alpha/2}^2 \\L_2 &= (n-1)S^2/\chi_{1-\alpha/2}^2\end{aligned}$$

så är $[L_1, L_2]$ ett 100(1 - α)% konfidensintervall för σ^2 .

- ▶ Bevis: Baseras på att vi har visat att $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.
- ▶ Märk hur vi hittar $\chi_{1-\alpha/2}^2$ och $\chi_{\alpha/2}^2$ i tabellen för $\chi^2(k)$ -fördelningar!
- ▶ Märk: Formlerna för L_1 och L_2 innehåller inte μ så detta intervallet kan användas även när μ är okänd!

T fördelningen

- ▶ En variabel X har en (Student) T fördelning med γ frihetsgrader, vi skriver $X \sim T(\gamma)$, om täthetsfunktionen är

$$f(x) = \frac{\Gamma(\gamma + 1)/2}{\Gamma(\gamma/2)\sqrt{\pi\gamma}} \left(1 + \frac{x^2}{\gamma}\right)^{-(\gamma+1)/2}$$

- ▶ När $\gamma \rightarrow \infty$ så går T fördelningen mot en standard normalfördelning. När T är mindre, så blir fördelningen mera spetsig i mitten, och får "tyngre svanser" än standard normalfördelning.
- ▶ Vi har $\mathbb{E}[X] = 0$ och $\text{Var}[X] = \gamma/(\gamma - 2)$ om $\gamma > 2$ (om $\gamma \leq 2$ så har T-fördelningen ingen ändlig varians).
- ▶ En viktig egenskap: Om $Z \sim \text{Normal}(0, 1)$ och $X \sim \chi^2(\gamma)$ är oberoende, så har vi

$$\frac{Z}{\sqrt{X/\gamma}} \sim T(\gamma).$$

- ▶ Tabeller för värden av fördelningsfunktionen till $T(\gamma)$ för olika γ finns i Milton.

Fördelningen för $(\bar{X} - \mu)/(S/\sqrt{n})$

- ▶ Vårt tidigare konfidensintervall för μ berodde på σ . Vi vill nu konstruera ett konfidensintervall som beror på S^2 i stället. Vi gör detta genom att studera en funktion av $\bar{X}$ och S^2 .
- ▶ Om $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma)$ är ett stickprov så har vi

$$(\bar{X} - \mu)/(S/\sqrt{n}) \sim T(n - 1)$$

Statistikan har alltså en T fördelning med $n - 1$ frihetsgrader.

- ▶ Ett bevis kan baseras på att
 - ▶ $\bar{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.
 - ▶ $(n - 1)S^2/\sigma^2 \sim \chi^2(n - 1)$
 - ▶ $\bar{X}$ och S^2 är oberoende.
 - ▶ Egenskapen till T fördelningen på förra sidan.

Exempel 3: Konfidensintervall för μ som inte beror på σ

- ▶ Om $X_1, \dots, X_n \sim \text{Normal}(\mu, \sigma)$ är ett stickprov, om $t_{\alpha/2}$ är så att $[-t_{\alpha/2}, t_{\alpha/2}]$ innehåller 100(1 - α)% av sannolikheten i en T-fördelning med $n - 1$ frihetsgrader, och om vi definierar

$$L_1 = \bar{X} - t_{\alpha/2}S/\sqrt{n}$$

$$L_2 = \bar{X} + t_{\alpha/2}S/\sqrt{n}$$

så är $[L_1, L_2]$ ett 100(1 - α)% konfidensintervall för μ .

- ▶ Bevis: Baseras på att vi har visat att $(\bar{X} - \mu)/(S/\sqrt{n}) \sim T(n - 1)$.
- ▶ Märk hur vi hittar $t_{\alpha/2}$ i tabellen för T-fördelningen.
- ▶ Märk: Formlerna för L_1 och L_2 innehåller INTE σ , så detta intervall KAN användas om *fördelningens* varians σ^2 är okänd.

Centralgränsteoremet (Central Limit Theorem, CLT)

- ▶ Antag $X_1, \dots, X_n$ är ett stickprov från en fördelning med väntevärde μ och varians σ^2 . Då gäller som gräns när $n \rightarrow \infty$, att

$$\bar{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$$

- ▶ För ändliga n kan detta användas som en approximation. Hur stor n behöver vara för att approximationen skall vara OK beror på vilken noggrannhet man behöver, och egenskaperna för fördelningen X kommer ifrån.
- ▶ Vissa fördelningar har ingen varians; då gäller heller inte CLT!

Exempel 4: Approximativt konfidensintervall för μ baserad på CLT

- ▶ Om $X_1, \dots, X_n$ är ett stickprov från en fördelning med väntevärde μ och varians σ^2 , om $z_{\alpha/2}$ är så att $[-z_{\alpha/2}, z_{\alpha/2}]$ innehåller $100(1 - \alpha)\%$ av sannolikheten i standard normalfördelning, och om vi definerar

$$\begin{aligned}L_1 &= \bar{X} - z_{\alpha/2}S/\sqrt{n} \\L_2 &= \bar{X} + z_{\alpha/2}S/\sqrt{n}\end{aligned}$$

så är $[L_1, L_2]$ ett *approximativt* $100(1 - \alpha)\%$ konfidensintervall för μ om n är stor.

- ▶ Bevis: Baseras på att för stora n har vi approximativt $\bar{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$ och för stora n har vi approximativt $S \approx \sigma$.
- ▶ Märk: För att detta skall gälla måste variansen σ^2 existera. Man kan alltid beräkna S ; det betyder inte att σ existerar!

Normalfördelningen som approximation

- ▶ Det är *medelvärdet* $\bar{X}$ som blir normalfördelad när antalet variabler man tar medelvärdet över ökar. Ett stickprov blir INTE mera normalfördelad bara därför att stickprovet blir större!
- ▶ Dock är det så att, för vissa fördelningar och vissa parametervärden, så kan själva fördelningen skrivas som en summa (eller medelvärde) av ett antal oberoende variabler. I dessa fall kan själva fördelningen approximeras som normal.
- ▶ Exempler på detta är:
 - ▶ Binomialfördelningen när n är stor och p inte för nära 0 eller 1.
 - ▶ Poissonfördelningen när intensiteten λ är stor.
 - ▶ Gamma-fördelningen när α är stor.
 - ▶ Därmed också χ^2 fördelningen med många frihetsgrader.