

TDAB01 Sannolikhetslära och Statistik

Jose M. Peña
IDA, Linköpings Universitet

Föreläsning 8

- ▶ Punktskattning
- ▶ Maximum likelihood metoden
- ▶ Samplingfördelning
- ▶ Konfidensintervall
- ▶ Konfidensintervall för populationsväntevärden
- ▶ Konfidensintervall för proportioner

- ▶ Grundproblem: Sannolikhetsmodeller har **okända** parametrar, θ .
 - ▶ T ex medelinkomsten i Sverige. Populationens väntevärde μ är okänd.
 - ▶ T ex andelen defekta komponenter i produktionen av en produkt. Sannolikheten p är okänd.
 - ▶ T ex spamfilter. Parametrarna β_0 och β_1 är okända i

$$p(\text{spam}|\text{antal\$}) = \frac{\exp(\beta_0 + \beta_1 \cdot \text{antal\$})}{1 + \exp(\beta_0 + \beta_1 \cdot \text{antal\$})}$$

- ▶ Vi vill använda (tränings)data för att bestämma värden för dessa parametrar.
- ▶ **Punktskattning**: Vår bästa **gissning** utifrån data.

Maximum likelihood metoden

- ▶ **Maximum likelihood (ML) estimatorn:** Välj värdet för θ som maximerar sannolikheten för data, dvs

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} P(x_1, \dots, x_n | \theta)$$

- ▶ Kontinuerliga fallet:

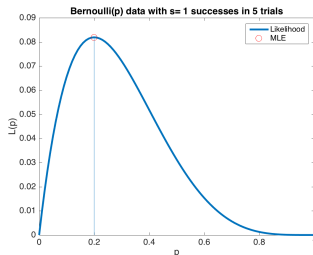
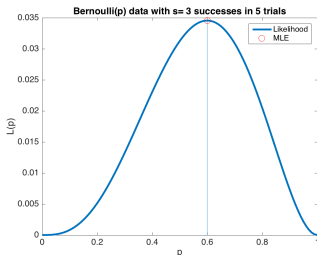
$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} f(x_1, \dots, x_n | \theta)$$

- ▶ **Likelihoodfunktionen** är sannolikheten för stickprovet sett **som en funktion av parametern**

$$L(\theta) = P(x_1, \dots, x_n | \theta)$$

- ▶ ML-estimatorn maximerar $L(\theta)$.
- ▶ Exempel: Data $X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1$ från $Bernoulli(p)$.

$$L(p) = (1-p)pp(1-p)p = p^3(1-p)^2$$



Maximum likelihood metoden

- ▶ Vi kan hitta **ML-skattningen** analytiskt: Lös med avseende på θ

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

- ▶ Oftast enklare att **maximera log-likelihoodfunktionen**

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0$$

- ▶ Exempel: Bernoulli data.

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\partial}{\partial p} (s \ln p + f \ln(1 - p)) = \frac{s}{p} + f \frac{-1}{1 - p} = \frac{s}{p} - \frac{f}{1 - p} = 0$$

vilket ger lösningen $\hat{p} = \frac{s}{s+f} = \frac{s}{n}$.

- ▶ Kontrollera \hat{p} är ett maximum, dvs andraderivatan är negativ i $p = \hat{p}$.

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{s}{p^2} - \frac{f}{(1 - p)^2} < 0$$

för alla $p \in [0, 1]$, inklusive $p = \hat{p}$.

Maximum likelihood metoden

- ▶ Notera att oberoende data är praktiskt

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

så log-likelihooden blir en summa som är lättare att derivera

$$\ln L(\theta) = \sum_{i=1}^n \ln f(x_i|\theta).$$

- ▶ Exempel: $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$ ger

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

och

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = \frac{n}{\lambda} - n\bar{x},$$

och därmed $\hat{\lambda} = 1/\bar{x}$.

Samplingfördelningen

- ▶ Hur bra är en estimator $\hat{\theta}$?
- ▶ Väntevärdesriktig ? $\mathbb{E}(\hat{\theta}) = \theta$.
- ▶ $Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.
- ▶ **Samplingfördelningen** beskriver variationen i $\hat{\theta}$ **över alla stickprov** av en viss storlek n .
- ▶ **Standardfelet** för $\hat{\theta}$ är $\sqrt{Var(\hat{\theta})}$, dvs standardavvikelsen för $\hat{\theta}$ **över alla stickprov** av storleken n .
- ▶ **Mean Squared Error:**

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

Samplingfördelningen

- ▶ Exempel: Poisson data. ML-estimator för λ : \bar{X} . Se Example 9.7 i Baron.
- ▶ Väntevärdesriktig: $\mathbb{E}(\hat{\lambda}) = \lambda$ och $\text{Var}(\hat{\lambda}) = \frac{\sigma^2}{n} = \frac{\lambda}{n}$.
- ▶ Notera att $\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}$ beror på den okända parametern λ . Lösning: Sätt $\lambda = \hat{\lambda} = \bar{x}$ eller sätt $\sigma^2 = s^2$.
- ▶ Tekniker för att härleda samplingfördelningen för en estimator $\hat{\theta}$:
 - ▶ Om X_1, \dots, X_n är iid från $N(\mu, \sigma^2)$ så är $\hat{\theta} = \bar{X} \sim N(\mu, \sigma^2/n)$ **exakt**.
 - ▶ **CLT med väntevärdesriktighet**: $\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$.
 - ▶ **Bootstrapsimulering** approximativt.
- ▶ **Bootstrap**:
 - ▶ Skapa N **bootstrapstickprov** $x^{(1)}, \dots, x^{(N)}$ av samma storlek som det ursprungliga stickprovet genom dragning **med återläggning**.
 - ▶ Beräkna estimatet $\hat{\theta}(x^{(1)}), \dots, \hat{\theta}(x^{(N)})$ för var och ett av dessa N stickprov.
 - ▶ Den empiriska fördelningen för $\hat{\theta}(x^{(1)}), \dots, \hat{\theta}(x^{(N)})$ (tänk histogram) är en approximation av samplingfördelningen för $\hat{\theta}$.

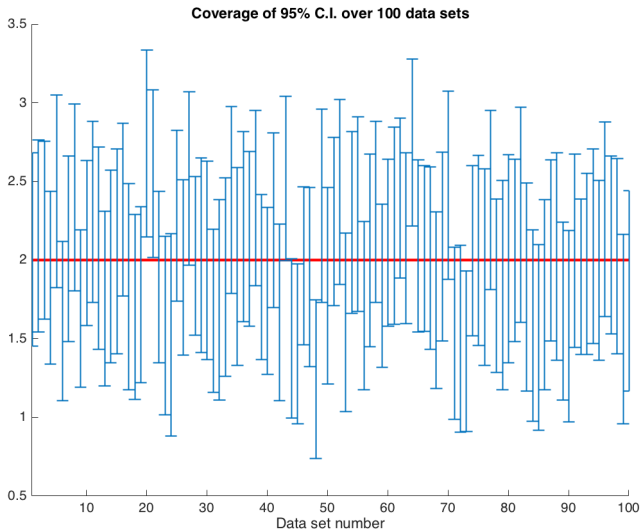
Konfidensintervall

- ▶ Punktskattning ger bara en bästa gissning för θ . Konfidensintervall är ett försök att beskriva osäkerheten om θ .
- ▶ **95%-igt konfidensintervall** för θ är ett intervall $[a, b]$ sådant att

$$P\{a \leq \theta \leq b\} = 0.95.$$

- ▶ Viktigt: Parametern θ är en fix konstant. Det är **intervallet** som är **slumpmässigt**, dvs a och b är funktioner av stickprovet.
- ▶ **Tolkning:** Ett 95%-igt konfidensintervall $[a, b]$ kommer att **täcka** parametervärdet θ , dvs $\theta \in [a, b]$, i 95% av alla möjliga stickprov. Alltså om vi räknar a och b från alla stickprov, täcker intervallet θ i 95% av fallen. Denna konfidens säger mer om metoden för att räkna intervallet än om det specifika intervallet vi fick från stickprovet. Tänk på sannolikheten att intervallet täcker θ snarare än på sannolikheten att θ ligger i intervallet.
- ▶ Man kan naturligtvis ha andra **konfidensnivåer** än 95%, men 90%, 95% och 99% är vanligast.

Konfidenzintervall

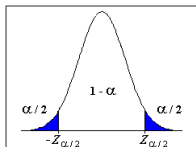


Konfidensintervall - standardprocedur

- Antag **normalfördelad väntevärdesriktig estimator** $\hat{\theta}$, t ex \bar{X} vid normalfördelade data (eller CLT argument). Då gäller

$$Z = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \sim N(0,1)$$

- Låt z_{α} vara $(1 - \alpha)\%$ percentilen i $N(0,1)$ fördelningen, dvs värdet som klipper av ytan α till **höger**. Tabell A4 i Baron ger att $z_{0.025} = 1.96$.



- Då gäller att

$$P\left(-z_{0.025} \leq \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \leq z_{0.025}\right) = 0.95$$

vilket kan skrivas om som

$$P(\hat{\theta} - z_{0.025} \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{0.025} \cdot \sigma(\hat{\theta})) = 0.95$$

- Alltså, $[\hat{\theta} - z_{\alpha/2} \cdot \sigma(\hat{\theta}), \hat{\theta} + z_{\alpha/2} \cdot \sigma(\hat{\theta})]$ är ett $(1 - \alpha)\%$ -igt konfidensintervall för θ .

Konfidsensintervall för populationsväntevärdet

- ▶ Antag $\theta = \mu$, $\hat{\theta} = \bar{X}$, $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) = \mu$, och $\sigma(\hat{\theta}) = \text{Std}(\bar{X}) = \sigma/\sqrt{n}$. Dessutom, σ **antas känd**.
- ▶ Centrala gränsvärdessatsen ger att $\hat{\theta} = \bar{X}$ är approximativt normalfördelad när n är stort ($n \geq 30$). Oavsett hur data är fördelade. Alltså, $\bar{X} \pm z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$ är ett (approximativt) 95%-igt konfidsensintervall för θ . Om data är normalfördelade är intervallet exakt.
- ▶ **Bestämning av stickprovsstorlek n :** Vi kan bestämma n så att vi får ett konfidsensintervall av given bredd.
- ▶ Se Examples 9.13 och 9.15 i Baron.

Konfidensintervall för populationsväntevärdet

- ▶ I praktiken är $\sigma(\hat{\theta})$ inte känd utan måste skattas (estimeras) från data, t ex $\sigma(\hat{\theta}) = \text{Std}(\bar{X}) = \sigma/\sqrt{n}$ och σ är ofta okänd.
- ▶ **Vid stort n** får vi en bra skattning av $\sigma(\hat{\theta})$ genom att ersätta den med t ex $s(\hat{\theta}) = s/\sqrt{n}$.
- ▶ **Vid stort n** får vi ett bra approximativt konfidensintervall genom att ersätta $\sigma(\hat{\theta})$ med $s(\hat{\theta})$, och anropa centralagränsvärdessatsen som tidigare.
- ▶ **Vid litet n** funkar $s(\hat{\theta})$ inte lika bra. Då, konfidensintervall för populationsväntevärdet μ vid **små stickprov** från en **normalfördelad population**:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}(0, 1)$$

- ▶ Så ett exakt 95%-igt konfidensintervall för μ ges då av

$$\bar{X} \pm t_{0.025}(n-1) \frac{s}{\sqrt{n}}$$

där $t_{0.025}(n-1)$ är 97.5% percentilen i **t -fördelningen** med $\nu = n - 1$ frihetsgrader. Läses av från Tabell A5 i Baron.

- ▶ Se Example 9.19 i Baron.
- ▶ För **små stickprov** från en **icke-normalfördelad population** använd bootstrap för att approximera samplingfördelningen (se slide 8 eller sida 337 i Baron).

Konfidensintervall för en andel

- ▶ Exempel: 196 av 2000 utfrågade svarar att de röstar på centerpartiet. Hur stor andel p röstar på centerpartiet i hela populationen ?
- ▶ $\hat{p} = 196/2000$ är ML-skattningen. Men hur säkra är vi ?
- ▶ $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$ där $X_i = 1$ om den i :te utfrågade person röstar på centerpartiet och $X_i = 0$ annars. Så \hat{p} är också ett medelvärde !
- ▶ Antag att $X_i \stackrel{iid}{\sim} \text{Bernoulli}(p)$. Då gäller $\mathbb{E}(X_i) = p$ och $\text{Var}(X_i) = p(1 - p)$. Alltså,

$$\mathbb{E}(\hat{p}) = p \text{ och } \text{Var}(\hat{p}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} np(1 - p) = \frac{p(1 - p)}{n}$$

- ▶ $\sigma(\hat{p})$ beror på p , som vi ersätter med en skattning: $s(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$.
- ▶ Centralagränsvärdessatsen ger ett approximativt $(1 - \alpha)100\%$ -igt intervall

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = (0.085, 0.111)$$

- ▶ Punktskattning
- ▶ Maximum likelihood metoden
- ▶ Samplingfördelning
- ▶ Konfidensintervall
- ▶ Konfidensintervall för populationsväntevärden
- ▶ Konfidensintervall för proportioner