SANNOLIKHETSLÄRA OCH STATISTIK
FÖRELÄSNING 8

Mattias Villani

Avdelningen för Statistik och Maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Översikt

- Punktskattning
- Samplingfördelning
- Konfidensintervall
- Konfidensintervall för populationsväntevärden
- Konfidensintervall för proportioner

Punktskattning

- Grundproblem: sannolikhetsmodeller har **okända parametrar**, θ.
  - Ex medelinkomsten i Sveriges: populationens väntevärde μ
  - Ex andelen defekta komponenter i produktionen av en produkt p.
  - Ex spamfilter: β_0 och β_1 är parametrar

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

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

Momentmetoden

- Ex $X_1, ..., X_n | \mu \overset{iid}{\sim} \text{Poisson}(\mu)$. $E(X) = \mu$.
- Rimlig punktskattning av populationväntevärdet μ:

$$\hat{\mu} = \bar{x}$$

- **Moment** av ordningen $k = 1, 2, ...$

$$\mu_k = \mathbb{E}\left(X^k\right)$$

- **Samplemoment** av ordningen k

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

- Ex: $k = 1$: $\mu_1 = \mu = \mathbb{E}X$ och $m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$.

Momentmetoden

- Momentmetoden för att skatta k modellparametrar $\theta_1, ..., \theta_k$: Lös följande ekvationssystem m a p $\theta_1, ..., , \theta_k$:

$$\mu_1 = m_1$$
$$\mu_2 = m_2$$
$$\vdots$$
$$\mu_k = m_k$$

- Notera att $\mu_1, ..., \mu_k$ är funktioner av $\theta_1, ..., \theta_k$. Mer korrekt:

$$\mu_1(\theta_1, ..., \theta_k) = m_1$$
$$\mu_2(\theta_1, ..., \theta_k) = m_2$$
$$\vdots$$
$$\mu_k(\theta_1, ..., \theta_k) = m_k$$

Momentmetoden

- Ibland mer praktiskt att jobba med centralmoment.
- **Centralmoment** av ordningen $k = 2, 3...$

$$\mu_k^{'} = \mathbb{E}\left(X - \mu_1\right)^k$$

- **Samplecentralmoment** av ordningen k

$$m_k^{'} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k$$

- Notera att $\mu_2^{'} = Var(X)$ och

$$m_2^{'} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \neq \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Momentmetoden - Beta exempel

- Ex $X_1, ..., X_n|\alpha, \beta \overset{iid}{\sim} \text{Beta}(\alpha, \beta)$:

$$\mu_1 = \mathbb{E}X = \frac{\alpha}{\alpha+\beta}$$

$$\mu_2 = \mathbb{E}(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

- Momentskattningar: lös för α och β

$$\frac{\alpha}{\alpha+\beta} = m_1$$
$$\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = m_2$$

ger $\hat{\alpha} = \frac{m_1(m_2-m_1)}{m_1^2-m_2}$ och $\hat{\beta} = \frac{(m_1-m_2)(m_1-1)}{m_1^2-m_2}$.

Maximum likelihood-metoden

- **Maximum likelihood (ML) estimatorn**: Välj det θ som maximerar sannolikheten för data:

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

- Kontinuerliga fallet:

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

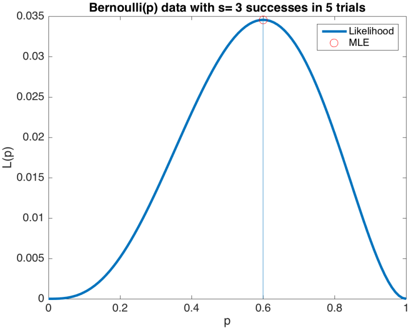
- **Likelihoodfunktionen** är sannolikheten för stickprovet sett **som en funktion av parametern**

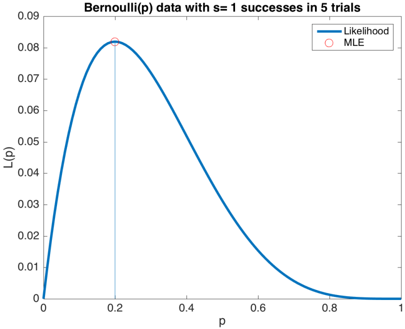
$$L(\theta) = P(x_1, ..., x_n | \theta)$$

- ML-estimatorn maximerar $L(\theta)$.
- Ex data från Bernoulli med sannolikhet p: $X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0, X_5 = 1$.

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

Maximum likelihood - Bernoulliexempel





Maximum likelihood-metoden

- Vi kan hitta **ML-skattningen** analytiskt: Lös m a p θ

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

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

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

- Ex Bernoulli:

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

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

- Kontrollera $\hat{p}$ är ett maximum - 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 $p = \hat{p}$.

ML-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

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

- Ex: $X_1, ..., X_n|\lambda \overset{iid}{\sim} Exp(\lambda)$ 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 $\hat{\lambda} = 1/\bar{x}$.

Samlingfördelningen

- Hur bra är en **estimator** $\hat{\theta}$?
- **Väntevärdesriktig**? $\mathbb{E}\hat{\theta} = \theta$.
- $\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.
- **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):

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

Samlingfördelningen

- Ex. Poisson. ML-estimator för μ: $\bar{X}$.
- Väntevärdesriktig: $\mathbb{E}(\hat{\mu}) = \mu$ och $Var(\hat{\mu}) = \frac{\sigma^2}{n} = \frac{\mu}{n}$.
- Notera att $Var(\hat{\mu}) = \frac{\mu}{n}$ beror på den okända parametern μ. Lösning: sätt $\mu = \hat{\mu} = \bar{x}$.
- Två vanliga tekniker för att räkna ut samplingfördelningen för en estimator $\hat{\theta}$:
  - **Centrala gränsvärdessatsen**: $\hat{\theta} \overset{approx}{\sim} N\left[\theta, Var(\hat{\theta})\right]$.
  - **Bootstrap**simulering.
- Bootstrap:
  - Skapa N **bootstrapstickprov** $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$ av samma storlek som det ursprungliga stickprovet genom dragning **med återläggning**.
  - Beräkna estimatet $\hat{\theta}(\mathbf{x}^{(1)}), ..., \hat{\theta}(\mathbf{x}^{(N)})$ för var och ett av dessa N stickprov.
  - Den empiriska fördelningen för $\hat{\theta}(\mathbf{x}^{(1)}), ..., \hat{\theta}(\mathbf{x}^{(N)})$ (tänk histogram) är en approximation av samplingfördelningen för $\hat{\theta}$.

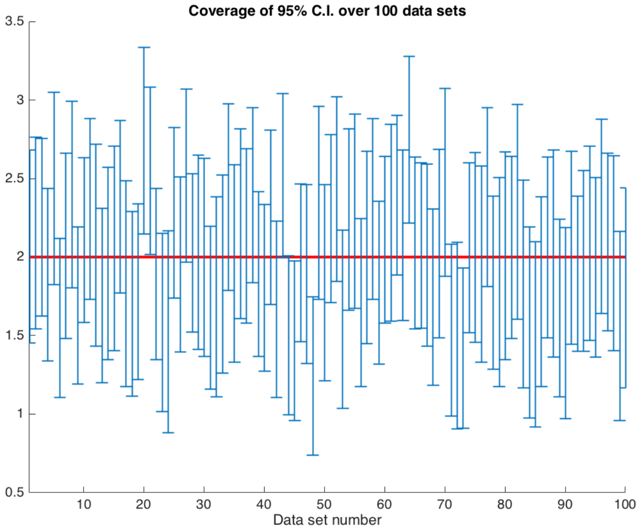
Konfidensintervall

- 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

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

- Viktigt: det är **intervallet** som är **slumpmässigt**. Parametern θ är en fix konstant.
- **Tolkning**: ett 95%-igt konfidensintervall $[a, b]$ kommer att **täcka** ($\theta \in [a, b]$) parametervärdet θ i 95% av alla möjliga stickprov.
- Man kan naturligtvis ha andra **konfidensnivåer** än 95%. 90%, 95% och 99% är vanligast. Se den lite klumpiga allmänna notationen $(1 - \alpha) \cdot 100\%$-igt konfidensintervall i Baron.

Konfidensintervall



Konfidensintervall - standardprocedur

- Standardprocedur för att skapa ett 95%-igt konfidensintervall.
- Antag normalfördelad väntevärdesriktig estimator $\hat{\theta}$. Då gäller

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

- Låt $z_{0.975}$ vara 97.5% percentilen i $N(0, 1)$ fördelningen. Tabell A4 i Baron ger att $z_{0.975} = 1.96$.
- Då gäller att

$$\mathbf{P}\left\{-z_{0.975} \leq \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \leq z_{0.975}\right\} = 0.95$$

  vilket kan skrivas om som

$$\mathbf{P}\left\{\hat{\theta} - z_{0.975} \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{0.975} \cdot \sigma(\hat{\theta})\right\} = 0.95$$

- Alltså: $[\hat{\theta} - z_{0.975} \cdot \sigma(\hat{\theta}), \hat{\theta} + z_{0.975} \cdot \sigma(\hat{\theta})]$ är ett 95%-igt konfidensintervall för θ.

Konfidensintervall för populationsväntevärdet

- $\theta = \mu$. $\hat{\theta} = \bar{X}$. $\sigma(\hat{\theta}) = Std(\bar{X}) = \sigma/\sqrt{n}$. 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: $\bar{X} \pm z_{0.975} \cdot \frac{\sigma}{\sqrt{n}}$ är ett (approximativt) 95%-igt konfidensintervall för θ.
- **Bestämning av stickprovsstorlek** n. Vi kan bestämma n så att vi får ett konfidensintervall av given bredd.

Konfidensintervall - okänt standardfel

- I praktiken är $\sigma(\hat{\theta})$ inte känd utan måste skattas (estimeras) från data. Ex: $Std(\bar{X}) = \sigma/\sqrt{n}$ och σ är ofta okänd.
- Vid stora stickprov (stort n) får vi ett bra approximativt konfidensintervall genom att ersätta $\sigma(\hat{\theta})$ med en skattning. T ex $s/\sqrt{n}$ istället för $\sigma/\sqrt{n}$.
- Konfidensintervall för populationsväntevärdet μ vid **små stickprov en normalfördel**ad **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 2.5% percentilen i t-**fördelningen** med $\nu = n-1$ frihetsgrader. Läses av från Tabell A5 i Baron.

Konfidensintervall för en andel

- Ex. 196 av 2000 utfrågade svarar att de röstar på centerpartiet. Hur stor andel p röstar på centerpartiet i hela populationen?
- $\hat{p} = 0.098$ är ML-skattningen. Med 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 $X_i \overset{iid}{\sim} Bernoulli(p)$. Då gäller $\mathbb{E}X_i = p$ och $Var(X_i) = p(1-p)$. Alltså

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

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

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