LINKÖPINGS UNIVERSITET
Institutionen för datavetenskap
Avdelningen för statistik och maskininlärning
Mattias Villani

2016-08-25
Sannolikhetslära och statistik, 6 hp
TDAB01

Tentamen i Sannolikhetslära och statistik (TDAB01), 6 hp

Tid:	08-12
Tillåtna hjälpmedel:	Miniräknare med tomt minne. Tabell- och formelsamling (delas ut tillsammans med tentamen)
Examinator:	Mattias Villani, tel. 070 – 0895205
Betyg:	Maximalt antal poäng: 20 poäng. Varje delfråga ger maximalt 5 poäng. Betyg 5 = 17-20 poäng Betyg 4 = 12.5-16.5 poäng Betyg 3 = 9-12 poäng

För full poäng krävs tydliga och väl motiverade svar.

1. Ett företag säljer två olika mobiler (A och B) som båda har garantitid 1 år. En av mobilernas funktioner kan gå sönder, om det sker under garantitiden kommer företaget att ersätta mobilen. Sannolikheten att ett fel sker under garantitiden är 0.002 för mobil A och 0.001 för mobil B. Det säljs 750 exemplar av mobil A och 500 exemplar av mobil B per år.

   (a) Vad är sannolikheten att antalet mobiler av typ A som går sönder under garantitiden överstiger 3?
   Lösning (2 poäng): Låt

   $$X_i = \begin{cases} 1 & \text{mobil av typ A har gått sönder} \\ 0 & \text{annars} \end{cases}$$

   för $1 \leq i \leq 750$. Då ges antalet mobiler av typ A som går sönder under garantitiden av $X = \sum_{i=1}^{750} X_i$. Vi vill bestämma $P(X > 3)$ där $X \sim Binomial(750, 0.002)$. Här använder vi Poisson approximationen och får att $P(X > 3) = 1 - P(X \leq 3) \approx 1 - e^{-1.5}(1 + 1.5 + \frac{1.5^2}{2} + \frac{1.5^3}{6}) = 0.0656$.

   (b) Givet att en mobil har gått sönder under garantitiden, vad är sannolikheten att mobilen är av typ A ?
   Lösning (3 poäng): Låt A vara händelsen att mobilen är av typ A, B händelsen att mobilen är av typ B och S händelsen att mobilen har gått sönder. Då gäller
   $P(A|S) = \frac{P(S|A)P(A)}{P(S)} = \frac{P(S|A)P(A)}{P(S|A)P(A)+P(S|B)P(B)} = \frac{0.002 \cdot \frac{750}{1250}}{0.002 \cdot \frac{750}{1250} + 0.001 \cdot \frac{500}{1250}} = \frac{3}{4}$.

2. Mattias har 100 uppgifter som han måste göra. Varje uppgift tar en exponentialfördelad tid med väntevärde 4 minuter att göra. Tiden det tar att göra en uppgift är oberoende av hur lång tid det tar att göra de andra uppgifterna.

   (a) Vad är sannolikheten att en uppgift tar mer än 10 minuter att göra?
   Lösning (1 poäng): Låt T vara tiden det tar att göra en uppgift. $P(T > 10) = e^{-\frac{10}{4}} \approx 0.082$.

(b) Mattias tycker att uppgifter som tar längre än 10 minuter att göra är tråkiga. Vad är sannolikheten att 10 eller fler av de 100 uppgifterna som måste göras är tråkiga?
Lösning(1 poäng): Låt

$$X_i = \begin{cases} 1 & \text{uppgiften är tråkig} \\ 0 & \text{annars} \end{cases}$$

för $1 \leq i \leq 100$. Då ges antalet tråkiga jobb av $S_n = \sum_{i=1}^{100} X_i$ där $S_n \sim Binomial(100, e^{-\frac{10}{4}})$. Enligt centrala gränsvärdesatsen konvergerar fördelningen för S_n mot $N(100 \cdot e^{-\frac{10}{4}}, \sqrt{100 \cdot e^{-\frac{10}{4}} \cdot (1 - e^{-\frac{10}{4}})}$. Vi får att $P(S_n \geq 10) = 1 - P(S_n \leq 9) = 1 - P(\frac{S_n - 100 \cdot 0.082}{\sqrt{100 \cdot 0.082 \cdot (1-0.082)}} \leq \frac{9 - 100 \cdot 0.082}{\sqrt{100 \cdot 0.082 \cdot (1-0.082)}}) \approx 1 - \Phi(0.29) = 0.386$.

(c) Vad är sannolikheten att han hinner göra alla uppgifter under 8 timmar?
Lösning (3 poäng): Låt $T_1, T_2, \ldots T_{100}$ vara oberoende och exponentialfördelade variabler med väntevärde 4, där T_i är tiden det tar att göra uppgift i. Vi vill bestämma $P(\sum_{i=1}^{100} T_i \leq 8 \cdot 60)$. Enligt centrala gränsvärdesatsen konvergerar fördelningen för $\sum_{i=1}^{100} T_i$ mot $N(100 \cdot 4, 100 \cdot 16)$ och vi får att $P(\sum_{i=1}^{100} T_i \leq 8 \cdot 60) = P(\frac{\sum_{i=1}^{100} T_i - 100 \cdot 4}{\sqrt{100 \cdot 16}} \leq \frac{480 - 100 \cdot 4}{\sqrt{100 \cdot 16}}) \approx \Phi(2) = 0.977$.
OBS: $\sum_{i=1}^{100} T_i \sim Gamma(100, \frac{1}{4})$.

3. Varje dag tar Norah samma väg från universitetet till träningshallen. Det finns 4 stoppsignaler på vägen och hon noterar följande: Om en stoppsignal visar grönt, kommer nästa stoppsignal att visa grönt med sannolikheten 0.5 och rött med sannolikheten 0.5. Om stoppsignalen däremot visar rött kommer nästa stoppsignal att visa rött med sannolikheten 0.6 och grönt med sannolikheten 0.4.

(a) Ange övergångsmatrisen som tillhör Markovkedjan.
Lösning (2 poäng): Låt $X(n) = 1$ om den n:te stoppsignalen är grön och $X(n) = 2$ om den är röd. Då gäller att övergångsmatrisen ges av: $P = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{pmatrix}$.

(b) Ange 2-stegs övergångsmatrisen och förklara vad den innebär.
Lösning (1 poäng): Se boken och lösningen på c)-delen nedan.

(c) Om det första stoppet visar grön, vad är sannolikheten att tredje stoppet visar röd?
Lösning (2 poäng): $P_{12}^{(2)} = 0.5 \cdot 0.5 + 0.5 \cdot 0.6 = 0.55$, se kapitel 6.

4. Antag att $X_1, X_2, ..., X_n$ är oberoende och fördelade enligt $Ge(\theta)$ (geometrisk fördelning).

(a) Beräkna momentskattningen $\hat{\theta}_{mom}$ av θ.
Lösning (0.5 poäng): Från formelsamlingen vet vi att $\mathbb{E}(X) = 1/\theta$ om $X \sim Ge(\theta)$. Momentmetoden sätter stickprovsmoment (sample moment) lika med populationsmoment (population moment)

$$\bar{x} = \mathbb{E}(X)$$

och löser med avseende på parametern θ. För geometrisk fördelning får vi alltså ekvationen $\bar{x} = 1/\theta$, och momentskattningen är därför $\hat{\theta}_{mom} = 1/\bar{x}$.

(b) Härled maximum likelihood-skattningen $\hat{\theta}$ av θ.
Lösning (1.5 poäng): Likelihoodfunktionen har formen (pga oberoende)

$$L(\theta) = p(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} (1-\theta)^{x_i - 1}\theta = (1-\theta)^{\sum_{i=1}^{n} x_i - n}\theta^n.$$

Så log-likelihooden är

$$l(\theta) = \ln L(\theta) = \left(\sum_{i=1}^{n} x_i - n\right) \ln(1-\theta) + n \ln \theta.$$

För att hitta ML-skattningen löser vi $\frac{d}{d\theta}l(\theta) = 0$ med avseende på θ:

$$\frac{d}{dp}l(\theta) = \left(\sum_{i=1}^{n} x_i - n\right)\frac{-1}{1-\theta} + \frac{n}{\theta} = 0$$

$$\frac{n}{\theta} = \left(\sum_{i=1}^{n} x_i - n\right)\frac{1}{1-\theta}$$

$$n = \left(\sum_{i=1}^{n} x_i\right)\theta$$

så

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}.$$

Vi kontrollerar att det verkligen är ett maximum:

$$\frac{d^2}{d\theta^2}l(\theta) = -\left(\sum_{i=1}^{n} x_i - n\right)\frac{1}{(1-\theta)^2} - \frac{n}{\theta^2} = -\left(\sum_{i=1}^{n} x_i - n\right)\frac{1}{(1-\theta)^2} - \frac{n}{\theta^2}$$

vilket visar att $\frac{d^2}{d\theta^2}l(\theta) < 0$ för alla $\theta \in (0,1)$ och att $\hat{\theta}$ därför är maximum likehood-skattningen (notera att $\sum_{i=1}^{n} x_i \geq n$).

(c) Beskriv i detalj hur man kan använda Bootstrap för att approximera variansen för $\hat{\theta}$ utifrån ett givet stickprov $x_1, ..., x_n$.
Lösning (1 poäng): Se Kapitel 10.3 i Barons bok. Bootstrap är en simuleringsmetod för att approximera samplingfördelningar och moment (t ex varians) för samplingfördelningar. Kortfattat så simulerar man $j = 1, ..., N$ stycken *bootstrapreplikat* från data. Varje bootstrapreplikat består av n observationer som dras slumpmässigt *med återläggning* från det datamaterial man har $(x_1, ..., x_n)$. I varje bootstrapreplikat beräknar man den estimator som man är intresserad av, i vårt fall $\hat{\theta}_{(j)} = 1/\bar{x}_{(j)}$, där $\bar{x}_{(j)}$ är medelvärdet i det j:te bootstrapreplikatet. Dessa $\hat{\theta}_{(1)}, ..., \hat{\theta}_{(N)}$ värden approximerar samplingfördelningen för $\hat{\theta}$ (vi kan t ex rita upp ett histogram). Vi kan nu approximera $Var(\hat{\theta})$ med stickprovsvariansen av $\hat{\theta}_{(1)}, ..., \hat{\theta}_{(N)}$:

$$Var(\hat{\theta}) \approx \frac{\sum_{j=1}^{N}\left(\hat{\theta}_{(j)} - \bar{\hat{\theta}}\right)^2}{N-1},$$

där $\bar{\hat{\theta}}$ är medelvärdet av $\hat{\theta}_{(1)}, ..., \hat{\theta}_{(N)}$.

(d) Härled aposteriorifördelningen för θ baserat på observationerna $x_1, ..., x_n$ från $Ge(\theta)$. Använd den konjugerade aprioriförderlningen. Ge ett uttryck för $\mathbb{E}(\theta|x_1, ..., x_n)$ och jämför denna bayesianska estimator med maximum likelihood-estimatorn.
Lösning (2 poäng): Från 4b) vet vi att likelihoodfunktionen är

$$p(x_1, ..., x_n|\theta) = (1-\theta)^{\sum_{i=1}^{n} x_i - n}\theta^n$$

vilket har formen $\theta^c(1-\theta)^d$ där $c = n$ och $d = \sum_{i=1}^{n} x_i - n$ är konstanter som kan beräknas från stickprovet. Detta är samma form på likelihooden som när man har observerat Bernoulliförsök. Så även för Geometriskt fördelade data är den konjugerade aprioriförderlningen $\theta \sim Beta(a, b)$. Aposteriorifördelningen är då

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta)p(\theta) \\ &\propto \theta^n(1-\theta)^{\sum_{i=1}^{n} x_i - n}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{n+a-1}(1-\theta)^{\sum_{i=1}^{n} x_i - n + b - 1} \end{aligned}$$

vilket är proportionellt mot täthetsfunktionen för en $Beta\,(n+a, \sum_{i=1}^{n} x_i - n + b)$ fördelning. Väntevärdet aposteriori är (se formel för väntevärde i Beta-fördelningen i formelsamlingen)

$$\mathbb{E}(\theta|x_1, ..., x_n) = \frac{n+a}{(n+a) + (\sum_{i=1}^{n} x_i - n + b)} = \frac{1 + a/n}{(a+b)/n + \bar{x}}.$$

$\mathbb{E}(\theta|x_1, ..., x_n) \to 1/\bar{x} = \hat{\theta}$ när $n \to \infty$ (större och större stickprov) eller när $a \to 0$ och $b \to 0$ (svagare och svagare apriorikunskap).

Lycka till!

Mattias