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

2015-10-29
Sannolikhetslära och statistik, 6 hp
TDAB01

Tentamen i Sannolikhetslära och statistik (TDAB01), 6 hp

Tid:	14-18
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. Låt $x_1, x_2, ..., x_n$ vara oberoende observationer från en slumpvariabel X som har sannolikhetsfördelningen

$$P(X=x) = \begin{cases} \frac{1-\theta}{3} & \text{för } x=-1 \\ \frac{1}{3} & \text{för } x=0 \\ \frac{1+\theta}{3} & \text{för } x=1 \end{cases}$$

där $-1 \leq \theta \leq 1$.

(a) Visa att sannolikheterna ovan faktiskt definierar en sannolikhetsfördelning.
Lösning (1 poäng): Vi måste visa följande:
att för en godtycklig händelse A

$$\mathbb{P}(A) \geq 0$$

och att sannolikheten för hela utfallsrummet är 1, dvs

$$\mathbb{P}(\Omega) = 1$$

Att sannolikheterna inte är negativa följer av att $-1 \leq \theta \leq 1$. Det andra följer av att

$$\mathbb{P}(\Omega) = \mathbb{P}(X=-1 \cup X=0 \cup X=1) = \mathbb{P}(X=-1) + \mathbb{P}(X=0) + \mathbb{P}(X=1) = 1$$

(b) Beräkna $\mathbb{E}X$ och $Var(X)$.
Lösning (1 poäng): Väntevärdet är

$$\mathbb{E}X = (-1) \cdot P(X=-1) + 0 \cdot P(X=0) + 1 \cdot P(X=1) = \frac{2\theta}{3}$$

och variansen är

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = (-1)^2 \cdot P(X=-1) + 0^2 \cdot P(X=0) + 1^2 \cdot P(X=1) - (\frac{2\theta)}{3})^2 = \frac{2}{3} - (\frac{2\theta)}{3})^2$$

(c) Beräkna momentskattningen av θ om $n = 2$ och du har observerat $x_1 = 0$ och $x_2 = 1$.
Lösning (1 poäng): Momentskattningen ges av $\mathbb{E}X = \bar{x}$. Här är

$$\mathbb{E}X = (-1) \cdot P(X = -1) + 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = \frac{2\theta}{3}$$

och

$$\bar{x} = \frac{1}{2}$$

så vi får att

$$\hat{\theta} = \frac{3}{4}$$

(d) Beräkna maximum likelihood-skattningen av θ om $n = 2$ och du har observerat $x_1 = 0$ och $x_2 = 1$.
Lösning (1 poäng): Likelihoodfunktionen är

$$L = \frac{1}{3} \cdot \frac{1+\theta}{3}$$

detta är en rätlinje och maximum antas i

$$\theta = 1$$

Kom ihåg: När man ska hitta extrempunkterna måste man även titta på intervallets ändpunkter, i detta fall $\theta = -1$ och $\theta = 1$.

(e) Antag nu att du istället har observerat $x_1 = 1$ och $x_2 = 1$. Visa att momentskattningen för θ är problematisk i det här fallet. Beräkna maximum likelihood-skattningen.
Lösning (1 poäng): Anledningen till att momentskattningen inte är definierad är för att θ kommer att vara större än 1 vilket medför att de definierade "sannolikheterna" blir negativa. Likelihoodfunktionen är

$$L = (\frac{1+\theta}{3}) \cdot (\frac{1+\theta}{3})$$

detta är en andragradfunktion med en minimipunkt i $\theta = -1$ och maximum antas i intervallets ändpunkt $\theta = 1$. Kom ihåg: När man ska hitta extrempunkterna måste man även titta på intervallets ändpunkter, i detta fall är ändpunkterna $\theta = -1$ och $\theta = 1$.

2. Kalle har 2 mynt i fickan, ett symmetrisk där sannolikheten för krona är 0.5 och ett skevt där sannolikheten för krona är 0.55. Kalle tar på måfå ett av mynten ur fickan.

(a) Vad är sannolikheten att Kalle får krona när han singlar (kastar) myntet en gång?
Lösning (1 poäng) : Låt S vara händelen att det dragna myntet är det symmetriska. Lagen om total sannolikhet ger att:

$$P(Krona) = P(Krona|S)P(S) + P(Krona|S^c)P(S^c) = 0.5 * 0.5 + 0.55 * 0.5 = 0.525$$

- Vad är sannolikheten att få exakt 1 krona om Kalle singlar myntet 4 ggr?
Lösning (1 poäng) : Låt

$$X_i = \begin{cases} 1 & \text{om det blir krona i kast } i \\ 0 & \text{annars} \end{cases}$$

Det totala antalet kronor ges av $T = \sum_{i=1}^{4} X_i$, där $T|S \sim Binomial(4, 0.5)$ och $T|S^c \sim Binomial(4, 0.55)$, detta ger att

$$P(T = 1) = P(T = 1|S)P(S)+P(T = 1|S^c)P(S^c) = \binom{4}{1}0.5^1*0.5^3*0.5+\binom{4}{1}0.55^1*0.45^3*0.5 = 0.225238$$

En del har gjort misstaget att använda $P(Krona)$ från den tidigare uppgiften, dvs att sätta $P(Krona) = 0.525$ och $P(Klave) = 0.475$. Detta är korrekt ifall man tillåter att lämna tillbaka myntet efter varje kast, men det är inte det man gör, det man gör är att man tar upp ett mynt på måfå och kastar 4 gånger, utan återläggning.

(a) För att bestämma vilket av mynten han fått tag i kastar han det dragna myntet 1000 gånger. Om han får 525 kronor eller mer drar han slutsatsen att det är det skeva myntet, får han färre än 525 drar han slutsatsen att det är det symmetriska myntet. Vad är sannolikheten att Kalle drar fel slutsats? Eventuella approximationer ska motiveras.

Lösning (3 poäng) : Låt S vara händelen att det dragna myntet är det symmetriska och F händelsen att Kalle drar fel slutsats efter sitt test. Lagen om total sannolikhet ger att:

$$P(F) = P(F|S)P(S) + P(F|S^c)P(S^c)$$

där $P(S) = P(S^c) = \frac{1}{2}$. Låt nu

$$X_i = \begin{cases} 1 & \text{om det blir krona i kast } i \\ 0 & \text{annars} \end{cases}$$

Det totala antalet kronor ges av $T = \sum_{i=1}^{1000} X_i$. Om det dragna myntet är symmetriskt har vi att $\mathbb{E}[X_i] = \frac{1}{2}$ och Var(X_i)= $\frac{1}{4}$, vilket ger att $\mathbb{E}[T|S] = 1000 * \frac{1}{2}$ och Var$(T|S)$= $1000 * \frac{1}{4}$ och om det dragna myntet inte är symmetriskt har vi att $\mathbb{E}[X_i] = \frac{11}{20}$ och Var(X_i)= $\frac{99}{400}$ vilket ger att $\mathbb{E}[T|S] = 1000 * \frac{11}{20}$ och Var$(T|S)$= $1000 * \frac{99}{400}$. Med hjälp av centrala gränsvärdessatsen får vi att:

$$P(F|S) = P(T \geq 525|S) = P(\frac{T - 1000 \cdot 0.5}{\sqrt{1000 \cdot 0.25}} \geq \frac{525 - 1000 \cdot 0.5}{\sqrt{1000 \cdot 0.25}}|S) \approx 1 - \Phi(1.) = 0.0571$$

Vi har också att :

$$P(F|S^c) = P(T \leq 525|S^c) = P(\frac{T - 1000 \cdot 0.55}{\sqrt{1000 \cdot \frac{99}{400}}} \leq \frac{525 - 1000 \cdot 0.55}{\sqrt{1000 \cdot \frac{99}{400}}}|S^c) \approx \Phi(-1.59) = 0.0559$$

Totalt får vi att $P(F) \approx 0.0571 \cdot 0.5 + 0.0559 \cdot 0.5 = 0.0565$.

3. Två telefonväxlar A och B öppnar båda kl 8:00 på morgonen och samtal inkommer därefter till växlarna enligt två oberoende Poissonprocesser med intensiteter λ_A respektive λ_B samtal per timme.

   (a) Vad är sannolikheten att det kl 9:00 har inkommit exakt tre samtal till växel A?
   Lösning (1 poäng): Antalet samtal $A(t)$ till växel A är Poissonfördelat med parametern $\lambda_A t$, där $t = 1$ i detta fall, så sannolikheten är $P(A(1) = 3) = e^{-\lambda_A} \frac{\lambda_A^3}{3!}$.

   (b) Vad är sannolikheten att tiden mellan två på varandra följande samtal till växel B överstiger en timme?
   Lösning (1 poäng): Tiden T mellan två samtal är exponentialfördelad så $P(T > 1) = e^{-\lambda_B}$.

   (c) Chefen kommer in kl 8:15 och får veta att det inte har kommit något samtal till växel A ännu. Vad är sannolikheten att tiden till det första samtalet till växel A överstiger en timme?
   Lösning (1.5 poäng): Låt T_A vara tiden till första samtalet som inkommer växel A. Vi vill ha $P(T_A > 1|T_A > 0.25)$, denna sannolikhet kan beräknas med hjälp av den minneslösa egenskapen som exponentialfördelningen har eller så härleder man egenskapen.

$$P(T_A > 1|T_A > 0.25) = \frac{P(T_A > 1 \cap T_A > 0.25)}{P(T_A > 0.25)} = \frac{P(T_A > 1)}{P(T_A > 0.25)} = e^{-\lambda_A \cdot 0.75}$$

   (d) Låt T vara tiden (räknat från 8:00) tills det inkommer ett samtal till någon av växlarna. Vilken fördelning har T?
   Lösning (1.5 poäng): Tiden tills det kommer ett samtal till någon av växlarna ges av $min(T_A, T_B)$. Eftersom processerna är oberoende är $min(T_A, T_B)$ exponential fördelad med parametern $\lambda_A + \lambda_B$. Jämför denna uppgift med uppgift 4.11b i kursboken.

4. Låt $X_1, X_2, ..., X_n$ vara oberoende observationer från en Poissonfördelning med parameter θ.

   (a) Härled maximum likelihoodskattningen $\hat{\theta}$ för θ och beräkna dess samplingvarians.
   Lösning (1.5 poäng): Likelihoodfunktionen är

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta} = \frac{\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} e^{-n\theta}$$

   och log-likelihooden är därför

$$\log L(\theta) = \left(\sum_{i=1}^{n} x_i \right) \log \theta - n\theta - c,$$

   där $c = \log \prod_{i=1}^{n} x_i!$ som inte beror på θ. Maximum-likelihood skattningen är det värde på θ som maximerar $L(\theta)$, eller ekvivalent $\log L(\theta)$. Vi beräknar första-derivatan, sätter den till noll och läser för θ

$$\frac{d \log L(\theta)}{d\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - n = 0$$

   vilket ger lösningen

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

   Vi kontrollerar att detta verkligen är ett maximum genom att undersöka om andra-derivatan är negativ i $\theta = \bar{x}$

$$\frac{d^2 \log L(\theta)}{d\theta^2} = -\frac{\sum_{i=1}^{n} x_i}{\theta^2} < 0$$

   för alla θ och därmed även för $\theta = \bar{x}$. $\hat{\theta} = \bar{x}$ är den unika maximum-likelihoodskattningen.

(b) Antag att din apriorifördelning för θ är en $Gamma(\alpha, \lambda)$ fördelning. Visa att aposteriorifördelningen för θ är en Gamma-fördelning med parametrar $\alpha_x = \alpha + \sum_{i=1}^n x_i$ och $\lambda_x = \lambda + n$.
Lösning (1.5 poäng): Enligt Bayes sats är aposteriorifördelningen

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta)\\ &= \frac{\theta^{\sum x_i}}{\prod_{i=1}^n x_i!}e^{-n\theta}\cdot\frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}\\ &\propto \theta^{\sum \alpha+x_i-1}e^{-(\lambda+n)\theta},\end{aligned}$$

som är proportionell mot tätheten för en $Gamma(\alpha + \sum_{i=1}^n x_i, \lambda + n)$ fördelning.

(c) Antag att du bedömmer din apriorikunskap sådan att $E(\theta) = 1$ och $Std(\theta) = 2$ (standardavvikelse). Ge en uttryck (formel) för aposteriorifördelningen för θ baserat på observationerna $x_1 = 2, x_2 = 1$ och $x_3 = 0$. Du ska alltså fortfarande anta att $\theta \sim Gamma(\alpha, \lambda)$ apriori.
Lösning (1 poäng): Eftersom $\theta \sim Gamma(\alpha, \lambda)$ apriori så gäller att

$$\mathbb{E}\theta = \frac{\alpha}{\lambda}, \qquad Std(\theta) = \sqrt{Var(\theta)} = \frac{\sqrt{\alpha}}{\lambda}.$$

Alltså kan vi lösa följande ekvationer med avseende på α och λ

$$\frac{\alpha}{\lambda} = 1, \qquad \frac{\sqrt{\alpha}}{\lambda} = 2,$$

vilket ger $\alpha = \lambda = 1/4$. Data ger att $n = 3$ och $\sum_{i=1}^n x_i = 3$. Aposteriorifördelningen är alltså: $\theta|\mathbf{x} \sim Gamma(3.25, 3.25)$ vilket har tätheten (se tabell- och formelsamling):

$$\pi(\theta|\mathbf{x}) = \frac{3.25^{3.25}}{\Gamma(3.25)}\theta^{3.25-1}e^{-3.25\theta} = 18.08\cdot\theta^{2.25}e^{-3.25\theta}.$$

(d) Förklara vad som menas med påståendet: 'Intervallet $[a, b]$ är ett 95%-igt konfidensintervall för θ'.
Lösning (0.5 poäng): Påståendet innebär att att det **slumpmässiga intervallet** $[a, b]$ kommer att innefatta parametervärdet θ i **95% av alla stickprov (samples) från populationen**. Här är alltså parameterns värde θ fixt (icke slumpmässigt) och intervallet är slumpmässigt (varierar från ett stickprov till det nästa). Det är **fel** att tolka ett konfidensintervall som ett interval som innehåller parametervärdet θ med sannolikheten 95%.

(e) Förklara vad som menas med påståendet: 'Intervallet $[a, b]$ är ett 95%-igt bayesianskt sannolikhetsintervall för θ'. (Sannolikhetsintervall kallas ibland också för kredibilitetsintervall. På engelska säger man ofta credibility interval).
Lösning (0.5 poäng): Den bayesianska tolkningen av sannolikhetsintervall är att den som skapat intervallet anser att θ befinner sig i intervallet $[a, b]$ **med sannolikheten** 0.95. Det bayesianska intervallet bygger alltså på ett **subjektivt sannolikhetsbegrepp** och uttalar sig **inte** om hur ofta intervallet kommer att täcka det sanna värden i ett antal upprepade stickprov från populationen.

Lycka till!

Mattias