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

2016-01-08
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 vara vikten på en honungsmelon och Y vikten på en vattenmelon. Antag att X och Y är normalfördelade med väntevärde 1 respektive 2 och varians 0.25 respektive 0.36. Du väljer slumpmässigt och oberoende en honungsmelon och en vattenmelon.

   (a) Uttryck den sammanlagda vikten i termer av X och Y. Vilken fördelning har den sammanlagda vikten?
   Lösning (poäng): $V = X+Y$ är normalfördelad med väntevärde $1+2 = 3$ och varians $0.25+0.36 = 0.61$.

   (b) Vad är sannolikheten att den sammanlagda vikten är mindre än 2 kg?
   Lösning (poäng): $P(V \leq 2) = \Phi(\frac{2-3}{\sqrt{0.61}}) \approx \Phi(-1.28) = 0.1$.

   (c) Kilopriset på honungsmelonen är 10 kr och vattenmelonen 20 kr, uttryck den totala kostnaden Z i termer av X och Y. Vilken fördelning har Z?
   Lösning (poäng): $Z = 10X + 20Y$ som är normalfördelad med väntevärde $10 * 1 + 20 * 2 = 50$ och varians $10^2 * 0.25 + 20^2 * 0.36 = 169$.

   (d) Vad är sannolikheten att priset Z blir mer än 20 kr?
   Lösning(poäng): $P(Z > 20) = 1 - P(Z \leq 20) = 1 - \Phi(\frac{20-50}{\sqrt{169}}) = 1 - \Phi(-2.3) = \Phi(2.3) = 0.989$.

2. Samtal inkommer till en telefonväxel enligt en Poisssonprocess med intensitet 1 samtal per minut.

   (a) Låt T vara tiden mellan två på varandra följande samtal. Vad är sannolikheten att T är större än 1.5 minuter?
   Lösning (poäng): T är exponentialfördelad med parametern 1, detta ger att $P(T > 1.5) = \int_{1.5}^{\infty} \exp(-t)dt = \exp(-1.5) \approx 0.22$.

   (b) Låt T_{min} vara den minsta tiden mellan två på varandra följande samtal fram till det tredje samtalet. Vad är sannolikheten att T_{min} är mindre än 0.5 minuter.
   Lösning (poäng): Låt T_1 vara tiden till det första samtalet, T_2 tiden mellan första och andra samtalet och T_3 tiden mellan det andra och tredje samtalet. Dessa tider är oberoende och

exponentialfördelade slumpvariabler med parameter 1 och $T_{min} = min(T_1, T_2, T_3)$. T_{min} är exponentialfördelad med parametern 3 och $P(T_{min} < 0.5) = \int_0^{0.5} 3\exp(-3t)dt = 1 - \exp(-1.5) \approx 0.78$.

(c) Låt S_3 vara tidpunkten då det tredje samtalet inkommer, vilken fördelning har S_3?
Lösning (poäng): $S_3 = T_1 + T_2 + T_3$ är summan av tre exponentialfördelade variabler med parametern 1 och är alltså gammafördelad med parametrarna $(3, 1)$.

3. Ett flygbolag uppskattar sannolikheten att en person med biljett inte dyker upp till flyget är 0.06 och är oberoende av andra personers benägenhet att dyka upp till flyget. Man väljer därför att sälja fler biljetter än de 200 platser som finns på planet för att minimera risken att få tomma platser. Hur många biljetter kan flygbolaget som mest sälja för att sannolikheten att alla passagerare får plats ska vara större än 0.99?
Lösning (poäng): Låt n vara antalet biljetter som säljs och låt

$$X_i = \begin{cases} 1 & \text{om personen som köper den i:te biljetten kommer till flyget} \\ 0 & \text{annars} \end{cases}$$

för $1 \leq i \leq n$. Då har vi att X_i är Bernoullifördelade slumpvariabler med $P(X_i = 0) = 0.06$ och $P(X_i = 1) = 0.94$. Det totala antalet passagerare som kommer till flyget ges av $S_n = \sum_{i=1}^{n} X_i$. Vi har att $\mathbb{E}[S_n] = n\mathbb{E}[X_i]$ och $Var(S_n) = nVar(X_i)$. Vi har dessutom att

$$P(S_n > 200) < 0.01$$

måste vara uppfylld. Enligt gränsvärdesatsen följer att

$$P(S_n > 200) = P(\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} > \frac{200 - E(S_n)}{\sqrt{Var(S_n)}}) < 0.01$$

kan uttryckas som

$$P(Z > \frac{200 - 0.94n}{\sqrt{n * 0.94 * 0.06}}) < 0.01,$$

där Z är normalfördelad variabel med väntevärde 0 och varians 1. Detta ger ekvationen

$$\frac{200 - 0.94n}{\sqrt{n * 0.94 * 0.06}} \geq 2.33.$$

Detta är en andragradsekvation i variabeln $\sqrt{n}$ som har lösningen $(n \geq 0)$ $n \leq 204.35$. Antalet biljetter de kan sälja som mest är 204.

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

(a) Härled maximum likelihoodskattningen $\hat{\theta}$ för θ. Visa att $\frac{1}{\hat{\theta}}$ är väntevärdesriktig (unbiased) för $\frac{1}{\theta}$.
Lösning (2 poäng): Likelihoodfunktionen är

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

och log-likelihooden är därför

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

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 löser för θ

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

vilket ger lösningen

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

Vi kontrollerar att detta verkligen är ett maximum genom att undersöka om andra-derivatan är negativ i $\theta = 1/\bar{x}$

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

för alla θ och därmed även för $\theta = \frac{1}{\bar{x}}$. $\hat{\theta} = \frac{1}{\bar{x}}$ är den unika maximum-likelihoodskattningen. Att $\frac{1}{\hat{\theta}}$ är väntevärdesriktig för $\frac{1}{\theta}$ kan visas så här:

$$\mathbb{E}\left(\frac{1}{\hat{\theta}}\right) = \mathbb{E}\left(\bar{x}\right) = \mathbb{E}\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) = \frac{\mathbb{E}x_1 + \mathbb{E}x_2 + ... + \mathbb{E}x_n}{n} = \frac{\frac{1}{\theta} + \frac{1}{\theta} + ... + \frac{1}{\theta}}{n} = \frac{\frac{n}{\theta}}{n} = \frac{1}{\theta}.$$

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

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

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

(c) Antag att har observerat n oberoende observationer $x_1, ..., x_n$ från modellen i a) med medelvärdet $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Beräkna den prediktiva fördelningen för en ny observation $\tilde{x}$ från samma population, dvs beräkna

$$p(\tilde{x}|x_1, ..., x_n).$$

[Ledtråd: Den prediktiva fördelningen får man genom att integrera ut den okända parameter θ med avseende på aposteriorifördelningen $p(\theta|x_1, ..., x_n)$].
Lösning (1.5 poäng):

$$\begin{aligned}
p(\tilde{x}|x_1, ..., x_n) &= \int p(\tilde{x}|x_1, ..., x_n, \theta) p(\theta|x_1, ..., x_n) d\theta \\
&= \int p(\tilde{x}|\theta) p(\theta|x_1, ..., x_n) d\theta \\
&= \int \theta e^{-\theta\tilde{x}} \frac{\lambda^{\alpha+n}}{\Gamma(\alpha+n)} \theta^{\alpha+n-1} e^{-(\lambda+n\bar{x})\theta} d\theta \\
&= \frac{\lambda^{\alpha+n}}{\Gamma(\alpha+n)} \int \theta^{\alpha+n+1-1} e^{-(\tilde{x}+\lambda+n\bar{x})\theta} d\theta \\
&= \frac{\lambda^{\alpha+n}}{\Gamma(\alpha+n)} \frac{\Gamma(\alpha+n+1)}{(\tilde{x}+\lambda+n\bar{x})^{\alpha+n+1}} \\
&= \frac{\lambda^{\alpha+n}(\alpha+n)}{(\tilde{x}+\lambda+n\bar{x})^{\alpha+n+1}},
\end{aligned}$$

för $\tilde{x} > 0$. Jag har använt att $\theta^{\alpha+n+1-1} e^{-(\tilde{x}+\lambda+n\bar{x})\theta}$ är proportionellt mot $Gamma(\alpha+n+1, \lambda+n\bar{x})$ fördelningen (som vi ju vet integrerar till ett eftersom det är en p.d.f.). Jag har också använt att $\Gamma(a+1) = a\Gamma(a)$ för alla $a > 0$. Fördelningen ovan kallas ibland för en Gamma-Gamma fördelning och är relaterad till Paretofördelningen. Men det räcker att få fram täthetens form ovan för att få full poäng.

Lycka till!

Mattias