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

2016-10-28
Sannolikhetslära och statistik, 6 hp
TDAB01

Tentamen i Sannolikhetslära och statistik (TDAB01), 6 hp

Tid:	8-12
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. Totalpoängen avrundas till närmaste heltal. Betyg 5 = 18-20 poäng Betyg 4 = 14-17 poäng Betyg 3 = 10-13 poäng

För full poäng krävs tydliga och väl motiverade svar.

1. Låt X vara en kontinuerlig slumpvariabel med täthetsfunktion

$$f(x) = \begin{cases} \frac{c}{x^3} & \text{om x}\geq 1 \\ 0 & \text{annars} \end{cases}$$

   (a) Bestäm c.
   Lösning : $c = \frac{1}{\int_1^\infty \frac{1}{x^3} dx} = 2$

   (b) Ange fördelningsfunktionen för X.
   Lösning : Fördelningsfunktionen ges av

$$F(x) = \begin{cases} \int_1^x \frac{2}{t^3} dt = 1 - \frac{1}{x^2} & \text{om x}\geq 1 \\ 0 & \text{annars} \end{cases}$$

   (c) Bestäm sannolikheten att $x \leq 2$.
   Lösning:$\int_1^2 \frac{2}{x^3} dx = \frac{3}{4}$

   (d) Bestäm $E(X)$.
   Lösning : $E(X) = \int_1^\infty \frac{2}{x^2} dx = 2$.

2. Mejl kommer in enligt en Poissonprocess med intensiteten 6 mejl per minut.

   (a) Vad är sannolikheten att det inte kommer in några mejl de första 20 sekunderna?
   Lösning : Låt $N(t)$ beteckna antal mejl som kommit in vid tiden t. Sannolikheten att det har kommit in 0 mejl under de första 20 sekunderna är $P(N(\frac{1}{3}) = 0) = \exp(-\frac{6}{3})$.

   (b) Efter 30 sekunder har det kommit 3 mejl. Beräkna sannolikheten att det inte har kommit några mejl under de första 20 sekunderna.
   Lösning: $P(N(\frac{1}{3}) = 0 | N(\frac{1}{2}) = 3) = \frac{P(N(\frac{1}{2})=3|N(\frac{1}{3})=0)P(N(\frac{1}{3})=0)}{P(N(\frac{1}{2})=3)} = \frac{1}{27}$.

(c) Anta att ett mejl är spam med sannolikheten p och att mejlen är oberoende, bestäm fördelningen för antal spam vid tiden t givet att antal mejl är n vid tiden t.
Lösning: Antal spam $X(t)$ givet $N(t) = n$ är $Bin(n, p)$.

3. Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade $Bernoulli(p)$ med väntevärde p och varians $p(1-p)$.

(a) Härled maximumlikelihoodskattningen $\hat{p}$ för p.
Lösning : $\hat{p} = \frac{\sum_1^n X_i}{n}$

(b) Bestäm den approximativa fördelningen för $\sum_1^n X_i$ om n är stort. Bestäm även $E(\sum_1^n X_i)$ och $Var(\sum_1^n X_i)$.
Lösning : Enligt CGS är $\sum_1^n X_i$ approximativt normalfördelad med $E(\sum_1^n X_i) = np$ och $Var(\sum_1^n X_i) = np(1-p)$. Notera att här så konvergerar fördelningen för $\sum_1^n X_i$ mot normalfördelningen då $n \to \infty$.

(c) Använd approximationen från uppgift 3b för att beräkna ett 95% dubbelsidigt (approximativt) konfidensintervall för p då $\bar{X} = \frac{1}{2}$. Ge en korrekt tolkning av konfidensintervallet.
Lösning : Ett konfidensintervall för p med approximativ konfidensnivå 0.95 ges av $[\frac{1}{2} - 1.96 * \frac{1}{2\sqrt{n}}, \frac{1}{2} + 1.96 * \frac{1}{2\sqrt{n}}]$. Detta intervall kommer att täcka det sanna parametervärdet p i (approximativt) 95% av alla möjliga stickprov med n observationer från populationen.

4. Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade $Binomial(4, p)$ slumpvariabler och anta vidare att apriorifördelningen för p är en $Beta(\alpha, \beta)$.

(a) Härled aposteriorifördelningen.
Lösning : Vi har att likelihoodfunktionen är $f(\mathbf{x}|p) = \prod_{i=1}^n \binom{4}{x_i} p^{x_i}(1-p)^{(4-x_i)} \propto p^{\sum x_i}(1-p)^{(4n-\sum x_i)}$ och
$\pi(p) \sim p^{(\alpha-1)}(1-p)^{(\beta-1)}$ för $0 < p < 1$. Detta ger att aposteriorifördelningen är $\pi(p|\mathbf{x}) \propto p^{(\alpha+\sum x_i)-1}(1-p)^{(\beta+4n-\sum x_i)-1}$. Detta känner vi igen som täthetsfunktionen för en $Beta(\alpha + \sum x_i, \beta + 4n - \sum x_i)$

(b) Väntevärdet i aposteriorifördelningen används ibland som en (frekventistisk) estimator. Kalla denna estimator för $\hat{p}_B$. Bestäm $\hat{p}_B$ och beräkna dess samplingvarians.
Lösning : $\hat{p}_B = \frac{\alpha+\sum X_i}{\alpha+\beta+4n}$ och $Var(\hat{p}_B) = \frac{1}{(\alpha+\beta+4n)^2} Var\left(\sum X_i\right) = \frac{np(1-p)}{(\alpha+\beta+4n)^2}$.

(c) Ett 95% Highest Posterior Density (HPD) intervall är ett Bayesianskt osäkerhetsintervall för en parameter p som innehåller de värden på parametern p som har högst aposterioritäthet, och där sannolikheten att p tillhör intervallet är 0.95. Beräkna ett HPD intervall för p givet $n = 2$, $\alpha = \beta = 1$ och $\sum_{i=1}^n x_i = 8$.
Lösning : En mängd C är en$(1-\alpha)100\%$ credible region (i detta fall ett credible intervall) för parametern p om $\mathbf{P}(p \in C|\mathbf{X} = \mathbf{x}) = \int_{\mathbf{C}} \pi(p|\mathbf{x})dp = 1 - \alpha$ och C är en HPD om vi väljer den så att vi inkluderar alla punkter p med störst posteriortäthet $\pi(p|\mathbf{x})$.
Givet värdena $n = 2$, $\alpha = \beta = 1$ och $\sum x_i = 8$. så gäller att $\pi(p|\mathbf{x}) = 9p^8$ (täthetsfunktionen för $Beta(9, 1)$). Detta är en växande funktion vilket ger att C måste vara på formen $C = [a, 1]$, detta ger att vi behöver lösa ekvationen $0.95 = \int_a^1 9p^8 dp$ som ger att $C = [0.716871, 1]$

Lycka till!

Mattias