MVE090 Matematisk statistik Z, 7.5 hp

Tentamen 17 januari 2008 em M

Obs att detta är även omtentamen för TMS051, del B.

Tillåtna hjälpmedel är räknedosa utan lagrad information om kursen, Beta, kursens formel- och tabellsamling.

Examinator är Tommy Norberg, ankn 3528 eller 0730 79 42 09.

Assistent är Sofia Tapani, ankn 5336.

Jour: Sofia (eller Tommy) går att nås per telefon under tentamen.

Maximalt antal tentamenspoäng är 30, av dessa krävs normalt 12 för godkänt betyg och 18 resp 24 för 4:a och 5:a. Lösningar till tentamensproblemen går att ladda ner från kurshemsidan. Rättningsprotokoll anslås ej.

Svar och lösningar skall motiveras om ej annat sägs i uppgiften.

Uppgifter

1. Betrakta ett slumpmässigt försök i vilket händelserna A, B, C är ömsesidigt uteslutande och sådana att exakt en av dem måste inträffa. T.ex ett tärningskast och $A = \{1, 4\}$, $B = \{2\}$ och $C = \{3, 5, 6\}$. Låt F vara en annan händelse. Visa först allmänt att

$$P(F) = P(A)P(F|A) + P(B)P(F|B) + P(C)P(F|C) \tag{1}$$

och verifiera sedan (1) numeriskt för tärningskastfallet med F som händelsen att ett udda antal prickar erhålls. Antag därvid att tärningen är symmetrisk. Vad är $P(A|F)$? Ange först en formel för beräkning av $P(A|F)$. Räkna sedan ut (med hjälp av formeln) denna betingade sannolikhet i tärningskastexemplet. (5 p)

2. I en studie av risken för utbrott av vattenburen smitta (d.v.s smitta spriden genom dricksvattnet) på Europanivå redogjordes för 86 kända utbrott under åren 1990–2005. Antag en lämplig fördelning för antalet utbrott under ett år, och skatta sannolikheten för minst 2 utbrott under år 2008. (3 p)

3. Härled en formel för simulering av observationer ifrån Paretofördelningen. (3 p)

4. I denna uppgift ska vi studera en två-dimensionell stokastisk variabel N, X, där N är diskret och X är kontinuerlig. Täthet för N, X är funktionen

$$f(n, x) = e^{-\lambda}\frac{\lambda^n}{n!}\frac{1}{\sqrt{2\pi n}\sigma}e^{-\frac{1}{2}(x-n\mu)^2/(n\sigma^2)} \quad \text{för} \quad n = 0, 1, \ldots, \ -\infty < x < \infty$$

där $\lambda > 0$, $-\infty < \mu < \infty$ och $\sigma > 0$. Beräkna den betingade tätheten för X givet $N = n$. Motivera extra noga stegen i beräkningen. Beskriv motsvarande fördelning i ord. (4 p)

5. I en studie av ett nytt mätinstrument mätta man mot en s.k normal med det kända värdet 1. Därvid erhölls i $n = 25$ mätningar $\sum x = 27.71$ och $\sum x^2 = 35.984$. Bestäm ett uppåt begränsat 95% konfidensintervall för mätinstrumentets standardavvikelse. Antag normalfördelade observationer. (3 p)

6. I en simulering av en produktionsanläggning hann man under en dag göra 400 oberoende observationer av tiden det tar för en operatör att plocka ut ett bearbetat råämne ur en maskin och sätta in ett nytt, samt starta maskinen. Därvid erhölls $\sum x = 11\,028.6$ och $\sum x^2 = 314\,610.46$. Beräkna ett 99% konfidensintervall för förväntad operatörstid. Tidsenheten är sekunder, så det kan vara lämpligt att avrunda till en decimal. Ange förutsättningarna för dina beräkningar. (3 p)

7. Tillverkare av allehanda produkter är (bör i alla fall vara) intresserade av sina produkters (förväntade) livstider. Statistik över livstider kan vara både dyr och tidskrävande att få fram. I ett fall studerades 5 enheter tills alla slutat fungera. Man erhöll

   3.48 2.41 1.43 2.75 6.59

   (tidsenhet: år). Man utgår ifrån att data är Weibullfördelade med parametrar α och β. I en verklig arbetssituation skulle din uppgift kanske vara att ML-skatta α och β. Men för att denna uppgift inte ska bli för komplicerad rent numeriskt vill jag att du endast ML-skattar α under antagandet att $\beta = 2$. Skatta sedan förväntad livslängd μ, samt det som i en del sammanhang kallas L20, nämligen den tid som i genomsnitt 80% av alla producerade enheter överlever. Klarar du inte att ML-skatta α, så går det ändå att skatta μ och L20. Gör i så fall det. (5 p)

8. Vid en jämförelse av två algoritmer för lösning av det s.k handelsresandeproblemet upprepade man en beräkning med algoritm A 5 gånger och med algoritm B 7 gånger. Men erhöll olika värden varje gång eftersom algoritmerna innehåller slumpmoment som ger varierande beräkningstider på samma problem. För algoritm A erhölls

   4.53 4.26 5.18 5.24 5.16

   och för algoritm B erhölls

   4.28 4.92 4.81 5.06 4.78 4.79 4.60

   Tidsenhet: minuter. Kan du m.h.a statistiska metoder du lärt i denna kurs påvisa att algoritmerna är olika i något avseende? Antag oberoende normalfördelade observationer.(4 p)

Kortfattade svar eller lösningar till MVE090 den 17/1-08

1. $P(F) = P(F \cap A) + P(F \cap B) + P(F \cap C) = P(A)P(F|A) + P(B)P(F|B) + P(C)P(F|C)$ (den 1a likheten följer ifrån additivitet och att utfallsrummet är $A \cup B \cup C$, den 2a ifrån definitionen av betingad sannolikhet). Formeln är $P(A|F) = P(A)P(F|A)/P(F)$. De numeriska bitarna överlåtes.

2. Poissonfördelning med intensiteten $\lambda \approx 86/16 = 5.375$. Sannolikheten för minst 2 utbrott under år 2008 är $\sum_{k=2}^{\infty} e^{-\lambda}\lambda^k/k! = 1 - (e^{-\lambda} + e^{-\lambda}\lambda) \approx 0.970$

3. Enl känd sats ska man lösa x ur $u = F(x)$ eller $u = 1 - F(x)$, där u är slumptalet. För Paretofördelningen gäller att $1 - F(x) = (x_T/x)^\alpha$ för $x > x_T > 0$, så vi får direkt ur den andra ekvationen att $x = x_T/u^{1/\alpha}$.

4. Det gäller här att tänka ut uppdelningen $f(x,n) = p(n)f(x|n)$ och känna igen Poi(λ)-tätheten

$$p(n) = e^{-\lambda}\frac{\lambda^n}{n!}$$

och inse att $N \sim \text{Poi}(\lambda)$. Den betingade tätheten för X givet $N = n$ är således

$$f(x|n) = \frac{1}{\sqrt{2\pi n}\sigma}e^{-\frac{1}{2}(x-n\mu)^2/(n\sigma^2)}$$

Detta är en tätheten i $\text{N}(n\mu, \sqrt{n}\sigma)$-fördelningen. Man kan tänka sig att X är en summa av N oberoende $\text{N}(\mu, \sigma)$-variabler.

5. $s^2 = (35.985 - 27.71^2/25)/24 = 0.2196 = 0.4687^2$. Ur $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$ konstaterar vi att $P((n-1)s^2/\sigma^2 \geq 13.8484) = 0.95$ (se lämplig tabell), så $\sigma^2 \leq (n-1)s^2/13.8484 = 0.3806 = 0.617^2$ eller $\sigma \leq 0.617$ gäller med konfidensen 95%. Använder man den förenklade metoden i Beta, så ska man få $\sigma < 1.316s = 0.617$

6. Har man 400 observationer vore det mycket ovanligt om inte centrala gränsvärdessatsen kan tillämpas. Den samt faktumet $\sigma \approx s$ ger att approximativt gäller $(\bar{x} - \mu)/(s/\sqrt{n}) \sim \text{N}(0,1)$. Så med ca 99% konfidens gäller $\mu = \bar{x} \pm 2.576s/\sqrt{n}$. Återstår att beräkna $\bar{x} = 11028.6/400 = 27.572$ och $s^2 = (314610.46 - 11028.6^2/400)/399 = 26.4045 = 5.1385^2$ och, därför, $\mu = 27.6 \pm 0.7$

7. Weibulltätheten är $f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta}$, så troligheten är

$$L(\alpha, \beta) = \textstyle\prod_i \alpha\beta x_i^{\beta-1}e^{-\alpha x_i^\beta} = \alpha^n\beta^n\left(\prod_i x_i\right)^{\beta-1}e^{-\alpha\sum_i x_i^\beta}$$

Den logaritmeras och maximeras m.a.p α för fixt $\beta = 2$, enl

$$\log L(\alpha,\beta) = n\log\alpha + n\log\beta + (\beta - 1)\textstyle\sum_i x_i - \alpha\sum_i x_i^\beta$$
$$0 = \tfrac{\partial \log L(\alpha,\beta)}{\partial\alpha} = \tfrac{n}{\alpha} - \textstyle\sum_i x_i^\beta \quad \Rightarrow \quad \alpha = \frac{n}{\sum_i x_i^\beta}$$

ML-skattningen av α är således $\widehat{\alpha} = \frac{n}{\sum_i x_i^\beta} = \frac{5}{70.9540} = 0.0705$, ty $\beta = 2$. I formelsamlingen ses att $\mu = \alpha^{-1/\beta}\Gamma(1 + 1/\beta) = \alpha^{-1/2}\Gamma(1.5)$. I den ser vi också att $\Gamma(1.5) = 0.5\Gamma(0.5) = 0.5\sqrt{\pi}$, så vi får till sist skattningen $\widehat{\mu} = 0.0705^{-1/2}0.5\sqrt{\pi} = 3.338$. Den som inte kan ML-skatta α, beräknar istället $\widehat{\mu} = \bar{x} = 3.332$ och sedan $\widehat{\alpha}$ ur sambandet $\mu = \alpha^{-1/2}\Gamma(1.5)$. Han/hon bör få $\widehat{\alpha} = 0.0707$. L20 är estimatet av 0.20-kvantilen $x_{0.2}$. Den fås teoretiskt genom att lösa $F(x_{0.2}) = 0.20$. Ur $F(x_{0.2}) = 1 - e^{-\alpha x_{0.2}^\beta}$ fås att $x_{0.2} = ((-\log 0.8)/\alpha)^{1/\beta}$, så L20 $= ((-\log 0.8)/0.0705)^{1/2} = 1.779 \approx 1.78$. En sista titt på data ger att resultatet är rimligt (en observation är ju mindre och de övriga fyra större).

8. Man börjar med att notera (räkna ut) att $n_A = 5$, $\bar{x}_A = 4.874$, $s_A = 0.449$ och $n_B = 7$, $\bar{x}_B = 4.749$, $s_B = 0.250$. Använd därvid gärna formlerna $\bar{x} = \left(\sum x\right)/n$ och $s^2 = \left(\sum x^2 - \left(\sum x\right)^2/n\right)/(n-1)$. Ur den hemska formlen på sida 348 får vi att antalet frihetsgrader i jämförelsen ska vara 5. Räkna sedan ut

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_A^2/n_a + s_B^2/n_B}} = 0.506$$

och jämför med $t_{0.10}(5) = 1.476$. Så inte ens med 20% felrisk kan vi påstå att $\mu_a \neq \mu_B$.