MVE090 Matematisk statistik Z

Tentamen 30 augusti 2007 fm V

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

Examinator är Tommy Norberg, ankn 3528 eller 0730 79 42 09. Tommy går att nås per mobiltelefon under tentamen. Ingen jour således.

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. Härled formeln

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (3 \text{ p})$$

2. I en förorenad mark-tillämpning bedömmer man att risken att marken är förorenad är ca 70%. Mätapparaturen som används för att analysera jordprover ger tyvärr inte alltid rätt utslag. Känt är att den detekterar förorening i ca 95% av fallen om provet verkligen är förorenat och i ca 20% av dem då provet inte är det. Anta att du har ett prov i vilket förorening har detekterats. Hur stor är då sannolikheten att provet verkligen är förorenat. (4 p)

3. I en lagerbyggnad består brandlarmet av två st rökdetektorer och vi ska räkna på tillförlitligheten hos ett sådant larmsystem. Vid brand, eller snarare rökutveckling, räcker det att en detektor detekterar rök för att larmet ska gå. Låt A och B beteckna händelsen att den första resp andra rökdetektorn inte fungerar, som man säger, ”on demand”. Antag att $P(A) = P(B) \approx 10^{-2.5}$ samt att A och B är oberoende. Hur stor är då sannolikheten för ”system failure on demand” (d.v.s att larmet inte går fast rökutveckling sker). (3 p)

4. I nedanstående figur definieras en två-dimensionell diskret fördelning sånär som på en sannolikhet. (a) Varför är denna sannolikhet 0.009? (b) Är variablerna x och y oberoende? (c) Beräkna den betingade fördelningen för x givet att $y = 2$. (d) Beräkna fördelningen för variabeln $z = x + y$. (4 p)

	$y=$			
	0	1	2	3
$x = 0$	0.837	0.031	0.020	0.016
1	0.051	0.011	0.008	0.003
2	0.010	?	0.004	0.000

5. (a) Låt X vara en kontinuerligt fördelad stokastisk variabel med fördelningsfunktion $F(x)$ och täthet $f(x) = F'(x)$, låt U vara likformigt fördelad på det öppna enhetsintervallet $(0,1)$ och definiera

$$\widetilde{X} = F^{-1}(U)$$

där $F^{-1}(u)$ är inversen till $F(x)$. Visa att $\widetilde{X}$ och X har samma fördelning. (3 p)

(b) Paretofördelningens täthet är

$$f(x) = \frac{\alpha}{x_T}\left(\frac{x_T}{x}\right)^{\alpha+1} \qquad \text{för} \quad x > x_T$$

Här är $x_T > 0$ och $\alpha > 0$. Härled en algoritm (formel) för simulering av observationer med denna täthet. (2 p)

6. I livslängdssammanhang används ibland fördelningen definierad av tätheten

$$f(x) = 2\alpha x e^{-\alpha x^2} \qquad x > 0 \quad \alpha > 0$$

I en liten studie av lastbilars livstid erhölls följande tider i år:

14.23 4.20 13.82 14.75 6.46 12.69

Beräkna ML[1]-skattningen av α. Anta oberoende observationer. (4 p)

7. Visa att S^2 är en väntevärdesriktig skattning av σ^2. (3 p)

8. Man mätte en hållfasthet hos ett förstärkt material och erhöll i $n_1 = 5$ mätningar medelvärdet $\bar{x}_1 = 15.3$ och standardavvikelsen $s_1 = 3.37$. Man gjorde sedan $n_2 = 4$ kontrollmätningar på motsv icke-förstärkta material och erhöll då medelvärdet $\bar{x}_2 = 13.5$ och standardavvikelsen $s_2 = 2.95$. Stämmer det att dessa mätningar visar med en rimlig statistisk signifikans att det förstärkta materialet har en högre hållfasthet än det icke-förstärkta? Anta normalfördelade observationer. Hjälp: För F-fördelningens kvantiler gäller att $f_\alpha(\gamma_1, \gamma_2) = 1/f_{1-\alpha}(\gamma_2, \gamma_1)$. (4 p)

Lycka till!

[1]ML är en förkortning av engelskans Maximum Likelihood

Svar eller lösningar till MVE090 den 30/8-07

1. $P(A \cup B) = P(A \cup B \setminus A) = P(A) + P(B \setminus A) = P(A) + P(B) - P(A \cap B)$. Den första likheten följer av att $A \cup B = A \cup B \setminus A$, den andra av additionsaxiomet (händelserna A och $B \setminus A$ utesluter ju varandra) och den tredje av $P(B) = P(B \setminus A \cup B \cap A) = P(B \setminus A) + P(B \cap A) \Rightarrow P(B \setminus A) = P(B) - P(A \cap B)$. Även här har jag använt additionsaxiomet. Händelserna $B \setminus A$ och $B \cap A$ är ju ömsesidigt uteslutande.

2. Bayes formel ger direkt att

$$P(F|D) = \frac{P(F)P(D|F)}{P(F)P(D|F) + P(F')P(D|F')} = \frac{0.70 \cdot 0.95}{0.70 \cdot 0.95 + 0.30 \cdot 0.20} = 0.917$$

Här är F händelsen att marken är förorenad och D händelsen att förorening detekteras.

3. Failure F inträffar om, och endast om, både detektorerna ej larmar, d.v.s precis då $A \cap B$ inträffar. Så $P(F) = P(A \cap B) = P(A)P(B) = 10^{-5.0}$.

4. (a) Annars skulle inte summan av alla sannolikheter vara 1.000.

   (b) Nej, eftersom $P(x=0, y=0) \neq P(x=0)P(y=0)$. Vi har ju att $P(x=0, y=0) = 0.837$, $P(x=0) = 0.837 + 0.031 + 0.020 + 0.016 = 0.904$, $P(y=0) = 0.837 + 0.051 + 0.010 = 0.898$ och $0.904 \cdot 0.898 = 0.811792 \neq 0.837$.

   (c) $P(y=2) = 0.020 + 0.008 + 0.004 = 0.032$, så $P(x=0|y=2) = \frac{0.020}{0.032} = 0.625$, $P(x=1|y=2) = \frac{0.008}{0.032} = 0.250$ och $P(x=2|y=2) = \frac{0.004}{0.032} = 0.125$. (Kontroll: $0.625 + 0.250 + 0.125 = 1.000$)

   (d) $P(z=0) = 0.837$, $P(z=1) = 0.051 + 0.031 = 0.082$, $P(z=2) = 0.010 + 0.011 + 0.016$, $P(z=3) = 0.009 + 0.008 + 0.016 = 0.033$, $P(z=4) = 0.004 + 0.003 = 0.007$, $P(z=5) = 0.000$. (Kontroll: $0.837 + 0.082 + 0.041 + 0.033 + 0.007 + 0.000 = 1.000$)

5. (a) Det räcker att visa att $\widetilde{X}$ har samma fördelningsfunktion som X. Då måste ju även deras tätheter sammanfalla. Så

$$\widetilde{F}(x) = P(\widetilde{X} \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

   ty av det faktum att $F(x)$ är växande följer att $F^{-1}(u) \leq x \leftrightarrow u \leq F(x)$.

   (b) Om $f(x) = \frac{\alpha}{x_T}\left(\frac{x_T}{x}\right)^{\alpha+1}$ för $x > x_T$, så är $F(x) = \int_{x_T}^{x} f(y)\,dy = 1 - \left(\frac{x_T}{x}\right)^{\alpha}$. Låt u vara ett slumptal. Genom att lösa ut x ur likheten $u = 1 - F(x)$ erhåller vi en simulerad observation x. Detta är en omedelbar konsekvens av satsen som visades i (a). Vi får

$$u = 1 - F(x) \quad \Leftrightarrow \quad u = \left(\frac{x_T}{x}\right)^{\alpha} \quad \Leftrightarrow \quad u^{1/\alpha} = \frac{x_T}{x} \quad \Leftrightarrow \quad x = x_T\, u^{-1/\alpha}$$

6. Anta att vi har n oberoende observationer $x_1, \ldots, x_n$. Troligheten är då

$$L(\alpha) = \prod_i f(x_i) = \prod_i 2\alpha x_i e^{-\alpha x_i^2} = 2^n\, \alpha^n \left(\prod_i x_i\right) e^{-\sum_i \alpha x_i^2}$$

ML-skattningen $\widehat{\alpha}$ av α är det α som maximerar troligheten. Det beräknas så här:

$$\mathcal{L}(\alpha) = \ln L(\alpha) = n \ln 2 + n \ln \alpha + \sum_i \ln x_i - \alpha \sum_i x_i^2$$

$$\frac{d\mathcal{L}(\alpha)}{d\alpha} = 0 \quad \Rightarrow \quad \frac{n}{\alpha} - \sum_i x_i^2 = 0 \quad \Rightarrow \quad \alpha = \frac{n}{\sum_i x_i^2}$$

Att detta är en maximipunkt fås ur faktumet att $\frac{d^2\mathcal{L}(\alpha)}{d\alpha^2} < 0$ för alla $\alpha > 0$ (d.v.s att troligheten är konkav), vilket ej behöver visas. Så i föreliggande studie är alltså ML-skattningen

$$\widehat{\alpha} = \frac{n}{\sum_i x_i^2} = \frac{6}{14.23^2 + 4.20^2 + 13.82^2 + 14.75^2 + 6.46^2 + 12.69^2} = \frac{6}{831.4555} \approx 0.007216$$

7. $(n-1)S^2 = \sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - \sum_i 2X_i\bar{X} + \sum_i \bar{X}^2 = \sum_i X_i^2 - n\bar{X}^2$, så

$$(n-1)ES^2 = \sum_i EX_i^2 - nE\bar{X}^2$$

Notera att $\sigma^2 = \mathrm{Var}[X_i] = EX_i^2 - (EX_i)^2 = EX_i^2 - \mu^2$, så $EX_i^2 = \sigma^2 + \mu^2$. Analogt fås $E\bar{X}^2 = \mathrm{Var}[\bar{X}] + (E\bar{X})^2 = \sigma^2/n + \mu^2$. Så

$$(n-1)ES^2 = n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2$$

vilket skulle visas.

8. Bäst är att först avgöra om vi kan räkna som om varianserna är lika eller ej. Under nollhypotesen att de är lika är varianskvoten $F(4,3)$-fördelad. Vi jämför därför $s_1^2/s_2^2 = 3.37^2/2.95^2 = 1.305$ med t.ex 10%- resp 90%-kvantilerna, som är $f_{0.1}(4,3) = 5.343$ och $f_{0.9}(4,3) = 1/f_{0.1}(3,4) = 1/4.191 = 0.239$. Vi kan alltså inte förkasta att $\sigma_1^2 = \sigma_2^2$ ens på 20%-nivån, så det torde vara helt säkert att räkna som om $\sigma_1^2 = \sigma_2^2$. Vi får då den sammanvägda variansskattningen

$$s^2 = \frac{4 \cdot 3.37^2 + 3 \cdot 2.95^2}{4+3} = 10.2193 = 3.197^2$$

och antalet frihetsgrader i denna är $\gamma = 4 + 3 = 7$. Teststatistikan, som har värdet,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1.8}{3.197\sqrt{\frac{1}{5} + \frac{1}{4}}} = 0.839$$

är inte särskilt signifikant större än 0. Vi har ju att $t_{0.1}(7) = 1.415 > 0.839$. Svaret på den ställda frågan är alltså: nej, det stämmer inte, för inte ens på 10%-nivån kan vi förkasta nollhypotesen $\mu_1 \leq \mu_2$ till förmån för $\mu_1 > \mu_2$.