Avd. Matematisk statistik

KTH Matematik

TENTAMEN I 5B1504 MATEMATISK STATISTIK GRUNDKURS FÖR E (gamlingar)
TISDAGEN DEN 14 DECEMBER 2004 KL 8.00–13.00

Examinator: Gunnar Englund, 790 7416

Tillåtna hjälpmedel: Formel- och tabellsamling i matematisk statistik för E. Beta Mathematics Handbook. Kalkylator.

Införda beteckningar skall förklaras och definieras. Resonemang och uträkningar skall vara så utförliga och väl motiverade att de är lätta att följa. Numeriska svar skall anges med minst två siffrors noggrannhet.

Varje korrekt lösning ger 10 poäng. Gränsen för godkänt är preliminärt 20 poäng.

Resultatet anslås senast tisdagen den 11 januari på Matematisk statistiks anslagstavla i entréplanet, Lindstedtsvägen 25, rakt fram innanför porten och tentamen kommer att finnas tillgänglig på elevexpeditionen under sju veckor efter tentamensdagen.

Uppgift 1

a) Vi gör ett ”dubbelkast” med två tärningar, dvs kastar två vanliga symmetriska sexsidiga tärningar markerade med siffrorna 1,2,3,4,5,6. Beräkna sannolikheten att poängsumman blir 9. (3 p)

b) Man gör nu 500 sådana dubbelkast”. Beräkna med väl motiverade approximationer sannolikheten att man minst 50 gånger får poängsumman 9. (7 p)

Uppgift 2

I en undersökning av Sveriges älgstam var man intresserad av förekomsten av en viss gen, som antingen saknas, finns i enkel uppsättning eller finns i dubbel uppsättning hos de olika älgarna. Man kontrollerade 1000 på måfå valda älgar och konstaterade att 11 st hade genen i dubbel uppsättning, 187 st hade den i enkel uppsättning och att de övriga 802 älgarna saknade genen. Som teoretisk modell antog man att antalet gener av den aktuella typen hos en älg är $\mathrm{Bin}(2, p)$-fördelad och att de olika undersökta älgarnas gener var oberoende av varandra.

a) Beräkna Maximum Likelihood-skattningen av p baserad på ovanstående data. (5 p)

b) Teknologen Osquarulda känner inte till ML-metoden, men kom på intuitiva grunder fram till att p borde skattas med $p^* = \dfrac{x_1 + 2x_2}{2000}$ där x_1 = antalet älgar (av de 1000 undersökta) som hade genen i enkel uppsättning och x_2 = antalet älgar som hade genen i dubbel uppsättning. Undersök om p^* är en väntevärdesriktig skattning av p. (2 p)

c) Osquarulda vill dessutom räkna ut ett approximativt 95%-igt konfidensintervall för p baserad på sin skattning. Din uppgift är att utföra detta! <u>Ledning och varning:</u> x_1 och x_2 är ej utfall av oberoende stokastiska variabler. Fundera i stället över tolkningen av $x_1 + 2x_2$ när Du löser c-delen. (3 p)

Uppgift 3

För att jämföra två gödselmedel lät man 5 lantbrukare gödsla hälften av sin veteareal med medel A och den andra hälften med medel B. Man fick följande skördar per hektar:

Lantbrukare	1	2	3	4	5
Medel A	115.8	123.0	122.8	209.8	226.3
Medel B	121.2	122.1	128.0	214.9	229.2

För att få en enkel statistisk modell antog man att samtliga skördeutfall kan ses som utfall av oberoende normalfördelade stokastiska variabler med samma varians. Observera dock att lantbrukarna har gårdar med lite olika odlingsförutsättningar för vete.

a) Beräkna ett lämpligt 95%-igt konfidensintervall för skillnaden i förväntad skörd mellan arealer som gödslats med A respektive B. (7 p)

b) Testa hypotesen att gödselmedlen är lika bra mot alternativet att de skiljer sig åt. Nivå 5%. Slutsatsen skall klart framgå. (3 p)

Uppgift 4

$\big(X(t);\, t \in R\big)$ är en svagt stationär normalprocess med de okända väntevärdet m och med kovariansfunktionen

$$r_X(\tau) = 3\, e^{-2|\tau|}, \quad \tau \in R.$$

Låt

$$\overline{X}_T = \frac{1}{T}\int_0^T X(t)dt, \quad T > 0.$$

a) Visa att $\overline{X}_T$ är en väntevärdesriktig skattning av m för alla värden på $T > 0$. (3 p)

b) Beräkna $V(\overline{X}_T)$. (6 p)

c) Visa att skattningen $\overline{X}_T$ är konsistent (då $T \to \infty$). (1 p)

Uppgift 5

Låt $\big(X(t);\, t \in R\big)$ vara en svagt stationär normalprocess med $E\big(X(t)\big) = 1$ och spektraltäthet

$$S_X(\omega) = \frac{\omega^2}{(1+\omega^2)^2}, \quad -\infty < \omega < \infty.$$

Beräkna $P\big(X(3) \le X(5) + 2\big)$. (10 p)

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG TENTAMEN I 5B1504 TISDAGEN DEN 14 DECEMBER 2004 KL 8.00–13.00

Uppgift 1

a) Det finns 36 möjligheter av resultat på de två tärningarna som vi betraktar som lika sannolika. Vi kan t ex beteckna dessa med (i, j), $i, j = 1, 2, \cdots, 6$. Av dessa innebär (3,6), (4,5), (5,4) och (6,3) att poängsumman blir 9. Sannolikheten att få poängsumma 9 är alltså 4/36=$\underline{1/9}$.

b) Låt X='antalet gånger poängsumman är 9'. Vi har här typsituation för binomialfördelning och inser att X är Bin(500,1/9). Vi söker $P(X \geq 50)$.

Eftersom $500 \cdot \frac{1}{9}(1 - \frac{1}{9}) = 4000/81 \approx 49.3827 \geq 10$ kan vi göra normalapproximation av binomialfördelningen och har alltså

$$\mathrm{Bin}(500, 1/9) \approx N\big(500/9, \sqrt{500 \cdot \frac{1}{9}(1 - \frac{1}{9})}\big)$$

dvs $\mathrm{Bin}(500, 1/9) \approx N(500/9, \sqrt{49.3827} \approx N(55.56, 7.0273)$. Vi får

$$P(X \geq 50) = P\left(\frac{X - 55.56}{7.0273} \geq \frac{50 - 55.56}{7.0273}\right) \approx 1 - \Phi\left(\frac{-5.56}{7.0273}\right) =$$

$$= \Phi(5.56/7.0273) \approx \Phi(0.7906) \approx \underline{0.7854}$$

Uppgift 2

Låt U_i = antalet av genen som älg nr i har, $i = 1, 2, \cdots, 1000$. $U_1, U_2, \cdots, U_{1000}$ är oberoende och är alla $\mathrm{Bin}(2, p)$-fördelade, dvs

$$P(U_i = 0) = (1 - p)^2, \;\; P(U_i = 1) = 2p(1 - p), \;\; P(U_i = 2) = p^2.$$

Vi observerar att x_2 st (=11) U_i:n är 2, x_1 st (=187) U_i:n är 1 och att $1000 - x_1 - x_2$ st (=802) st U_i:n är 0. Likelihoodfunktionen blir då

$$L(p) = \prod_{i=1}^{1000} P(U_i = u_i) = (p^2)^{x_2} \left(2p(1 - p)\right)^{x_1} \left((1 - p)^2\right)^{1000 - x_1 - x_2} =$$

$$= p^{2x_2 + x_1}(1 - p)^{x_1 + 2000 - 2x_2 - 2x_1} 2^{x_1} = p^{x_1 + 2x_2}(1 - p)^{2000 - 2x_2 - x_1} 2^{x_1}$$

som ger

$$\ln L(p) = (x_1 + 2x_2) \ln p + (2000 - 2x_2 - x_1) \ln(1 - p) + x_1 \ln 2.$$

Vi får

$$\frac{d \ln L(p)}{dp} = \frac{x_1 + 2x_2}{p} - \frac{2000 - 2x_2 - x_1}{1 - p}$$

$\frac{d \ln L(p)}{dp} = 0$ ger

$$p^* = \frac{x_1 + 2x_2}{2000} = \frac{187 + 2 \cdot 11}{2000} = \underline{0.1045}.$$

b) x_1 är ett utfall av X_1 som är $\text{Bin}(1000, 2p(1-p))$ och vidare är x_2 ett utfall av X_2 som är $\text{Bin}(1000, p^2)$. Vi får

$$E(p^*) = E\left(\frac{X_1 + 2X_2}{2000}\right) = \frac{1}{2000}\left(E(X_1) + 2E(X_2)\right) = \frac{1}{2000}\left(1000 \cdot 2p(1-p) + 2 \cdot 1000p^2\right) = p$$

dvs p^* <u>är</u> en väntevärdesriktig skattning av p.

c) Vi noterar att $x_1 + 2x_2$ är antalet "genpositioner" som är besatta av genen. Eftersom U_i var $\text{Bin}(2, p)$ och de är oberoende, är det "oberoende lottningar" vid varje genposition. Alltså gäller att $X_1 + 2X_2$ är $\text{Bin}(2000, p)$.
Eftersom vidare $np(1-p) \approx 2000p^*(1-p^*)$=$2000 \cdot 0.1045(1-0.1045)$=$187.1595 \geq 10$ är normalapproximation tillåten, dvs X_1+$2X_2$ är approximativt $N\left(2000p, \sqrt{2000p(1-p)}\right)$ och p^* är alltså approximativt $N\left(p, \sqrt{\frac{p(1-p)}{2000}}\right)$. Ett approximativt 95%-igt konfidensintervall för p blir alltså med approximativa metoden (FS 11.3)

$$p^* \pm \lambda_{0.025}\sqrt{\frac{p^*(1-p^*)}{2000}} = 0.1045 \pm 1.9600\sqrt{\frac{0.1045(1-0.1045)}{2000}} = \underline{0.1045 \pm 0.0134}.$$

Uppgift 3

a) Observationer i par! Bilda nya data som medel A – medel B. Nya data blir $z_1, z_2, \ldots, z_5$ dvs. nya data blir $-5.4, 0.9, -5.2, -5.1, -2.9$ och vi får

$$\bar{z} = \frac{1}{5}(-5.4 + 0.9 - 5.2 - 5.1 - 2.9) = -3.54$$

$$s_z^2 = \frac{1}{5-1}\left(\sum_{i=1}^{5} z_i^2 - 5\bar{z}^2\right) = \frac{1}{4}(5.4^2 + 0.9^2 + 5.2^2 + 5.1^2 + 2.9^2 - 5 \cdot 3.54^2) = 7.192, \;\; s_z = 2.68.$$

Vi får med t-metoden konfidensintervallet

$$\bar{z} \pm t_{0.025}(4)\frac{s_z}{\sqrt{5}} = -3.54 \pm 2.78\frac{2.68}{\sqrt{5}} = \underline{-3.54 \pm 3.33}.$$

b) Eftersom 0 ej tillhör konfidensintervallet förkastas hypotesen att medlen är likvärdiga. Statistisk säkerställd skillnad alltså (nivå 5%).

Uppgift 4

a)

$$E(\overline{X}_T) = E\Big(\frac{1}{T}\int_0^T X(t)dt\Big) = \frac{1}{T}\int_0^T E(X(t))dt = \frac{1}{T}\int_0^T m\,dt = m$$

dvs $\overline{X}_T$ är en väntevärdesriktig skattning av m.

b)

$$V(\overline{X}_T) = C(\overline{X}_T, \overline{X}_T) = C\big(\frac{1}{T}\int_0^T X(t)dt, \frac{1}{T}\int_0^T X(s)ds\big) = \frac{1}{T^2}\int_0^T \int_0^T C\big(X(t), X(s)\big)dsdt =$$

$$= \frac{1}{T^2}\int_0^T \int_0^T r_X(s-t)dsdt = \frac{1}{T^2}\int_0^T \int_0^T 3\,e^{-2|s-t|}dsdt = \frac{2}{T^2}\int_0^T \left(\int_0^s 3\,e^{-2(s-t)}dt\right)ds =$$

$$= \frac{6}{T^2}\int_0^T e^{-2s}\left(\int_0^s e^{2t}dt)\right)ds = \frac{6}{T^2}\int_0^T e^{-2s}\left(\frac{1}{2}(1-e^{2s})\right)ds = \underline{\frac{3}{T^2}\left(T + \frac{1}{2}e^{-2T} - \frac{1}{2}\right)}.$$

c) Eftersom $V(\overline{X}_T) \to 0$ då $T \to \infty$ så är $\overline{X}_T$ en konsistent skattning ty den är väntevärdesriktig.

Uppgift 5

Vi får med formelsamlingens hjälp att $r_X(\tau) = \frac{1}{4}e^{-|\tau|}(1- \mid \tau \mid)$ vilket ger att $\big(X(3),\, X(5)\big)^T$ är tvådimensionellt normalfördelad med $E(X(3)) = 1$ och $E(X(5)) = 1$. Vidare är $V(X(3)) = V((X(5)) = r_X(0) = 1/4$, och $C(X(3), X(5)) = r_X(2) = -\frac{1}{4}\,e^{-2}$ som ger att $X(3) - X(5)$ är

$$N\left(0, \sqrt{r_X(0) + r_X(0) - 2r_X(2)}\right) = N\left(0, \sqrt{\frac{1}{2}\big(1 + e^{-2}\big)}\right),$$

och detta ger

$$P(X(3) - X(5) \leq 2) = \Phi\left(\frac{2}{\sqrt{\frac{1}{2}(1+e^{-2})}}\right) \approx 0.99598$$