Chalmers tekniska högskola
Matematisk statistik

TMS051 Matematisk statistik och simuleringsteknik, Z, del B

Tentamen 10 december 2005 em V

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

Examinator är Tommy Norberg, ankn 3528 eller 0730 794209.

Jour är Oskar Sandberg, ankn 5366.

Maximalt antal tentamenspoäng är 30, av dessa krävs 12 för godkänt betyg och 18 resp 24 för 4:a och 5:a. Lösningar går att ladda ner från kurshemsidan. Rättningsprotokoll anslås i första våningen nya Matematisk Centrum. Granskning av tentan kan under terminstid göras i mottagningsrummet i Matematisk Centrum må-fr 12^{30}–13.

Svar skall motiveras om ej annat sägs i uppgiften.

Fyll i enkäten när du är klar med uppgifterna nedan. Ju fler som svarar, desto bättre kan vi göra kursen!

Uppgifter

1. I en produktionsanläggning tillverkas varje dag 20 enheter. Av dessa väljs 4 slumpmässigt ut för kontroll. Om de 4 kontrollerade enheterna är felfria godkänns dagens produktion. Annars kontrolleras resterande 16 enheter och de med fel läggs åt sidan för korrigering. Antag att man under en dag tillverkar k felaktiga enheter och $20-k$ felfria. Beräkna sannolikheten att dagens produktion godkänns som funktion av k för $k=0,1,\ldots,20$. (3 p)

2. I det goda syftet att minska antalet som årligen dör i en viss form av cancer, tänker man varje år testa alla 50-åringar (detta kallas ofta för screening). Man vet att individer som har cancer-sjukdomen testar positivt i ca 4 fall av 5. Testet missar alltså att upptäcka ca 1 av 5 sjukdomsfall. Tyvärr kan testet även ge positivt utslag för individer som inte har sjukdomen. Detta sker i ca 1 fall av 10. Man bedömmer utifrån tidigare gjorda stora studier att ca 0.05% av alla 50-åringar har cancer-sjukdomen. Antag att man vid en teststation under 1 år klarar av att testa 10 000 50-åringar. (a) Ungefär hur många kommer testet att hitta som har cancer-sjukdomen? (b) Ungefär hur många kommer testet att sortera ut som inte har sjukdomen? (c) Antag att din mors väninna testar positivt. Hur stor är sannolikheten att hon verkligen är sjuk? (1+1+2 p)

3. Låt $X \sim \mathrm{U}(0,1)$ och $Y|x \sim \mathrm{U}(x,1)$. Bestäm $E[Y]$. (4 p)

4. Hur inser man enkelt att $\Phi(-z) = 1-\Phi(z)$? (3 p)

Vänd!

5. Antag att 0.375, 0.720, 0.068, 0.599, 0.099 är 5 oberoende observationer av en stokastisk variabel X med täthet

$$f(x) = \alpha x^{\alpha-1} \quad 0 < x < 1$$

för något $\alpha > 0$. Beräkna en punktskattning av α. (4 p)

6. Produktionssimulering. Tiden det tar att manuellt utföra moment i produktionen är ibland symmetriskt fördelad runt ett förväntat värde. I sådana fall kan normalfördelningen vara en bra approximation även om normalfördelade variabler i princip även kan anta negativa värden. Låt oss anta att så är fallet i en specifik applikation och att man i 16 oberoende observationer har erhållit medelvärdet 63.2 sekunder och standardavvikelsen 15.23 sekunder.

   (a) Beräkna ett symmetriskt konfidensintervall för väntevärdet μ med konfidensgraden 0.95. (2 p)

   (b) Beräkna, så gott det låter sig göras, P-värdet för testet av nollhypotesen $H_0 : \mu \leq 55.0$ mot alternativet $H_1 : \mu > 55.0$. (2 p)

7. Forts av uppgift 6.

   (a) Beräkna ett symmetriskt konfidensintervall för standardavvikelsen σ med konfidensgraden 0.95. (2 p)

   (b) Är det statistiskt säkerställt på nivån 5% att $\sigma < 20.0$. (2 p)

8. Forts av uppgifterna 6 och 7. Tiden det tar att utföra en manuell operation kan ju tänkas bero på vem som utför den. I uppgift 6 redovisades resultatet av 16 mätningar av operationstiden med en operatör som vi kan kalla A. Vid ett annat tillfälle gjordes 11 oberoende observationer av samma tid, men med en annan operatör som vi kan kalla B. Därvid erhölls medelvärdet 51.8 sekunder och standardavvikelsen 11.87 sekunder. Är det statistiskt säkerställt att operatörerna är olika snabba? (4 p)

Lycka till!

1. Låt $p(k)$ vara den sökta sannolikheten. Produktionen godkännes om ingen felaktig hittas. Sannolikheten för detta är

$$p(k) = \frac{\binom{20-k}{4}\binom{k}{0}}{\binom{20}{4}} = \frac{20-k}{20}\,\frac{19-k}{19}\,\frac{18-k}{18}\,\frac{17-k}{17}$$

för $k = 0, \ldots, 16$. Då $k = 17, \ldots, 20$ är $p(k) = 0$. (Obs att uttrycket längst till höger för $p(k)$ faktiskt ger rätt resultat även då $k = 17, \ldots, 20$.)

2. (a) $10\,000 \cdot 0.0005 \cdot 4/5 = 4$

   (b) $10\,000 \cdot 0.9995 \cdot 1/10 = 999.5 \approx 1000$ (obs att det är de misstänkta sjukdomsfallen, inalles $4 + 999.5 = 1003.5$, som testet sorterar ut)

   (c) $\dfrac{4}{4+999.5} \approx 0.00399$

   Det blev en del onödiga missförstånd i (b). Men jag menar att de som sorteras ut av test av den här typen är de som testar positivt. Jag borde varit tydligare i problemformuleringen.

3. $EY = \displaystyle\int_0^1\int_x^1 yf(x,y)\,dydx = \int_0^1\int_x^1 yf(x)f(y|x)\,dydx = \int_0^1\int_x^1 \frac{y}{1-x}\,dydx$
$= \displaystyle\int_0^1 \frac{1-x^2}{2(1-x)}\,dx = \int_0^1 \frac{1+x}{2}\,dx = \frac{1+\frac{1}{2}}{2} = \frac{3}{4}$. Flera har, utan matematisk motivering, använt likheten $EY = \int_0^1 E[Y|x]f(x)\,dx$, vilket är helt korrekt. Men obs att bristen på motivering naturligtvis ger poängavdrag. Och det räcker inte att säga att man har hittat den i Beta!

4. Resultatet följer av att tätheten $\varphi(z) = \Phi'(z)$ är en jämn funktion. Därför gäller att $\Phi(-z) = \int_{-\infty}^{-z}\varphi(u)\,du = \int_z^{\infty}\varphi(u)\,du = 1 - \Phi(z)$.

5. Trolighetsskattningen är $\hat{\alpha}_{\mathrm{L}} = \dfrac{n}{-\sum_i \ln x_i} \approx \dfrac{5}{6.823} \approx 0.73$. Momentskattningen är $\hat{\alpha}_{\mathrm{M}} = \dfrac{\bar{x}}{1-\bar{x}} = \dfrac{0.3722}{1-0.3722} \approx 0.59$. Båda svaren, korrekt motiverade, ger full poäng på uppgiften.

6. Alla bör känna till att $(\bar{X} - \mu)/(S/\sqrt{n})$ är t-fördelad med $n-1$ frihetsgrader. (Se formelsamlingen eller Beta annars.) Ur t-tabell, $n - 1 = 15$ frihetsgrader, läser vi av kvantilen $t_{0.025} = 2.131$. Således gäller (a) att

$$\mu = \bar{x} \pm t_{0.025}s/\sqrt{n} = 63.2 \pm 2.131 \cdot 15.23/\sqrt{16} = 63.2 \pm 8.1$$

definierar ett symmetriskt konfidensintervall för μ med konfidensgraden 0.95. (b) Obs utfall av $(\bar{X} - \mu)/(S/\sqrt{n})$ då $\mu = 55.0$ är 2.154, så vi ser direkt att P-värdet < 0.025. Genom att även jämföra med 0.01-kvantilen, som är 2.602 ser vi att P är $\not< 0.01$. Svaret är alltså att P-värdet ligger mellan 1% och 2.5%. (Den som har utrustning för exakt beräkning bör erhålla P-värdet 0.0240.)

7. I t.ex kursens formelsamling ser vi att $(n-1)S^2/\sigma^2$ är $\chi^2(n-1)$-fördelad. Här är $n = 16$ och observerat värde av variansen S^2 är $s^2 = 15.23^2$. Ur χ^2-tabell, $n-1 = 15$ frihetsgrader, avläser vi kvantilerna $\chi^2_{0.975} = 6.262$, $\chi^2_{0.95} = 7.261$ och $\chi^2_{0.025} = 27.488$. Således gäller att händelsen $6.262 \le (n-1)S^2/\sigma^2 \le 27.488$ har sannolikheten 0.95. Stoppar vi in $n = 16$ och den observerade variansen $s^2 = 15.23^2$, samt stuvar om olikheten erhålles

$$\frac{15 \cdot 15.23^2}{27.488} \le \sigma^2 \le \frac{15 \cdot 15.23^2}{6.262}$$

Medelst uträkning av ytterleden och rotutdragning av alla leden får vi att det i (a) sökta konfidensintervallet är $11.25 \le \sigma \le 23.57$. (b) Även händelsen $7.261 \le (n-1)S^2/\sigma^2$ inträffar med sannolikheten 0.95. Stoppar vi in antalet observationer n och observerad varians s^2, stuvar om olikheten och drar roten ur båda leden erhålles analogt att påståendet $\sigma \le \sqrt{15 \cdot 15.23^2/7.261} = 21.89$ har konfidensen 0.95. Det är således ej statistiskt säkerställt på nivån 5% att $\sigma < 20.0$.

8. Vi börjar med att avgöra om de bägge operatörernas standardavvikelser $\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}$ kan anses lika eller ej. Kvoten $S^2_{\mathrm{A}}/S^2_{\mathrm{B}}$ är (om $\sigma_{\mathrm{A}} = \sigma_{\mathrm{B}}$) F-fördelad med $n_{\mathrm{A}}-1 = 15$ resp $n_{\mathrm{B}} - 1 = 10$ frihetsgrader i täljare och nämnare. I F-tabell avläser vi kvantilerna $f_{0.90} \approx 0.486$ och $f_{0.10} \approx 2.244$. Med sannolikheten 0.80 gäller såleds att $0.486 \le S^2_{\mathrm{A}}/S^2_{\mathrm{B}} \le 2.244$. Den observerade varianskvoten är $s^2_{\mathrm{A}}/s^2_{\mathrm{B}} = 15.23^2/11.87^2 = 1.646$. Eftersom $0.486 \le 1.646 \le 2.244$ kan vi inte förkasta $\sigma_{\mathrm{A}} = \sigma_{\mathrm{B}}$ ens på nivån 20%. Data tyder alltså inte det minsta på att standardavvikelserna skulle kunna vara olika, så vi antar i den fortsatta analysen att de är lika. Vi väger därför samman våra två variansskattningar $s^2_{\mathrm{A}} = 15.23^2$ och $s^2_{\mathrm{B}} = 11.87^2$ till en med frihetsgradtalet $n_{\mathrm{A}} + n_{\mathrm{B}} - 2 = 25$ enligt

$$s^2 = \frac{(n_{\mathrm{A}}-1)s^2_{\mathrm{A}} + (n_{\mathrm{B}}-1)s^2_{\mathrm{B}}}{n_{\mathrm{A}} + n_{\mathrm{B}} - 2}$$

$$= \frac{15 \cdot 15.23^2 + 10 \cdot 11.87^2}{25} = 195.5305 = 13.98^2$$

Observerat värde på teststatistikan $(\bar{X}_{\mathrm{A}} - \bar{X}_{\mathrm{B}})/(S\sqrt{\frac{1}{n_{\mathrm{A}}} + \frac{1}{n_{\mathrm{B}}}})$ är 2.082. I t-tabell med 25 frihetsgrader avläses $t_{0.025} = 2.060$. P-värdet, då man testar nollhypotesen att operatörerna är lika snabba (har samma väntevärde) mot att de är olika snabba, är därför något lägre än $2 \cdot 0.025 = 0.05$. Således gäller att vi kan förkasta nollhypotesen att operatörerna är lika snabba på nivån 5%. Det är m.a.o statistiskt säkerställt att operatörerna är olika snabba. De som inte bryr sig om att, som ovan, undersöka om varianserna är lika mister 1p. Smith-Satterthwaites frihetsgradtal fick jag till 24.51, vilket efter bortkastandet av decimalerna blir $\gamma = 24$. Observerat värde på teststatistikan $(\bar{X}_{\mathrm{A}} - \bar{X}_{\mathrm{B}})/\sqrt{\frac{s^2_{\mathrm{A}}}{n_{\mathrm{A}}} + \frac{s^2_{\mathrm{B}}}{n_{\mathrm{B}}}}$ är 2.182 som ska jämföras med $t_{0.025} = 2.064$. Vi ser att vi kan förkasta nollhypotesen att operatörerna är lika snabba på nivån 5%. Det är m.a.o även med denna metod statistiskt säkerställt att operatörerna är olika snabba.

Statistik:	5:or	4:or	3:or	U:n	Σ
	6	6	2	2	16