Statistik för lärare, ht 2007, 7.5 hp

Tentamen 16 november 2007, kl 8:30–13:30 i V-huset

Tillåtna hjälpmedel är räknedosa utan lagrad information om kursen, formelsamling och bifogade tabellblad. (Formelsamling och tabellblad bifogas denna tes. Obs att tabellbladen skall återlämnas till tentamensvakt för vidarebefordran till institutionen.)

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

Jour är Anna Rudvik, ankn 3556.

Maximalt antal tentamenspoäng är 30, av dessa krävs normalt 12 för godkänt betyg och 21 för väl godkänt Lösningar till tentamensproblemen går att ladda ner från kurshemsidan. Rättningsprotokoll anslås ej. Rättningen kan granskas och ev överklagas på matematik-expeditionen.

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

Uppgifter

1. Om händelserna A och B vet man att $P(A) = 0.3$ samt att $P(B) = 0.5$.

   (a) Beräkna $P(A \cap B)$ om A och B vore oförenliga. (1 p)

   (b) Beräkna $P(A \cap B)$ om A och B vore oberoende. (1 p)

   (c) Hur stor kan $P(A \cap B)$ maximalt vara? (1 p)

   Gissa, om du inte kan svaret till (c).

2. Proportionen kvinnor i en viss åldersgrupp som har (icke diagnosticerad) bröstcancer är ca 0.8%. Om en kvinna har bröstcancer är sannolikheten att mammografi ger positivt utslag ca 90%, medan denna sannolikhet bara är ca 7% annars. Om en kvinna fått positivt svar i mammografi, hur stor är sannolikheten att hon verkligen har bröstcancer? Hur stor är denna sannolikhet om hon fått negativt svar? Kommentera gärna resultatet av dina beräkningar. (4 p)

3. Slumpvektorn (X, Y) är likformigt fördelad på mängden $\{(2,2), (2,3), (3,2), (3,3)\}$.

   (a) Beräkna $E(X)$. (1 p)

   (b) Beräkna $E(X+Y)$. (1 p)

   (c) Beräkna $E(XY)$. (1 p)

4. Antalet patogena (hälsofarliga) bakterier av en viss typ är i Göta Älv normalt ca 25 st per m^3 vatten. I ditt mikroskop betraktar du en droppe innehållande 1 ml älvvatten. Beräkna sannolikheten att din droppe innehåller minst en patogen bakterie. Hjälp för den som behöver: $1\ \mathrm{m}^3 = 1000$ l. (4 p)

5. Betrakta en urna med 3 röda, 5 blåa och 2 vita kulor. Man drar helt slumpmässigt och utan återläggning 4 st kulor. Låt X och Y vara antalet röda resp blåa kulor i urvalet.

(a) Ange genom att resonera kombinatoriskt sannolikhetsfunktionen $p_Y(l)$ för Y. (Denna deluppgift behöver ej motiveras.) (1 p)

(b) Hur stor är $E(Y)$ och $V(Y)$? (Denna deluppgift behöver ej motiveras.) (1 p)

(c) Skriv upp den betingade sannolikhetsfunktionen för Y, givet att $X = 2$. (Här krävs en motivering, som får lov att vara rent kombinatorisk, men inte behöver vara det.) (1 p)

(d) Är X och Y oberoende? (Även detta svar ska motiveras.) (1 p)

6. I två på varandra följande partisympatiundersökningar (gjorda utav ett visst partisympatiundersökningsinstitut) erhölls för partiet X andelarna 0.125 (i den första) samt 0.098. Journalisten som redogjorde för den senaste undersökningen påstod att förändringen är statistiskt säkerställd. Har han rätt? För att lösa uppgiften behöver du veta att urvalsstorlekarna var 1027 i den första resp 1233 i den andra undersökningen. Behöver du veta något mer? Skriv i såfall ned detta som en förutsättning för uppgiftens lösande. (4 p)

7. Antag att du har n st observationer $x_1, \ldots, x_n$ av en slumpvariabel $X \sim \mathrm{N}(\mu, \sigma)$. Ingen av parametrarna förutsätts vara känd.

(a) Vilken fördelning har $\bar{X}$? (Obs att i deluppgiften ingår att beräkna $E(\bar{X})$ och $D(\bar{X})$.) (2 p)

(b) Motivera uttrycket $\mu = \bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$ för beräkning av ett konfidensintervall för μ med konfidensgrad $100(1-\alpha)\%$. (2 p)

8. Man klassificerade 150 slumpvis utvalda företagsledare i kön och huruvida de var förstfödda eller ej:

	Förstfödd?		
	Ja	Nej	S:a
Kvinnligt kön	20	22	42
Manligt kön	34	74	108
S:a	54	96	150

Är dessa två variabler oberoende? Testa på nivån 5%. (Här är det viktigt att du noga formulerar ditt svar.) (4 p)

Obs att data i samtliga uppgifter ovan är påhittade och inte speglar verkliga förhållanden. Lycka till med lösandet av uppgifterna och i din karriär som lärare.

Kortfattade lösningar eller svar till tentamen 16 nov 2007

1. (a) 0
   (b) 0.15
   (c) 0.3

2. $P(\text{Cancer}|\text{Pos testres}) = \frac{0.008 \cdot 0.90}{0.008 \cdot 0.90 + (1-0.008) \cdot 0.07} = 0.0939$ resp
   $P(\text{Cancer}|\text{Neg testres}) = \frac{0.008 \cdot 0.10}{0.008 \cdot 0.10 + (1-0.008) \cdot 0.93} = 0.000866$
   Tänk själv igenom vad dessa sannolikheter säger.

3. (a) $E(X) = 2 \cdot 0.5 + 3 \cdot 0.5 = 2.5$
   (b) $E(X+Y) = 4 \cdot 0.25 + 5 \cdot 0.5 + 6 \cdot 0.25 = 5$ eller $E(X+Y) = (4+5+5+6)/4 = 5$ eller $E(X+Y) = E(X) + E(Y) = 2.5 + 2.5 = 5$
   (c) $E(XY) = 4 \cdot 0.25 + 6 \cdot 0.5 + 9 \cdot 0.25 = 6.25$ eller $E(XY) = (4+6+6+9)/4 = 6.25$ (Svaret $E(XY) = E(X)E(Y) = 2.5 \cdot 2.5 = 6.25$ ger 0 p. Varför?)

4. $X \sim \text{Poi}(\lambda)$, där $\lambda = 25 \cdot 10^{-6}$ st per ml, och sökt är $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda} \approx \lambda = 25 \cdot 10^{-6}$ (Varför är $e^{-\lambda} \approx 1 - \lambda$?)

5. (a) $p_Y(l) = \frac{\binom{5}{l}\binom{5}{4-l}}{\binom{10}{4}}$ för $l = 0, 1, 2, 3, 4$
   (b) $E(Y) = 4 \cdot \frac{5}{10} = 2$
   $V(Y) = 4 \cdot \frac{5}{10} \cdot (1 - \frac{5}{10}) \frac{10-4}{10-1} = \frac{2}{3}$
   (c) $p_{Y|X=2}(l) = \frac{P(X=2, Y=l)}{P(X=2)} = \frac{\binom{5}{l}\binom{2}{2-l}}{\binom{7}{2}}$, ty…
   (d) Nej, ty… (Obs att motiveringen ska handla om X och Y, inte om vad som händer med t.ex sannolikheten att dra en röd i 1:a resp 2:a dragningen.)

   Skälet till att (a) och (b) ej behövde motiveras är att svaren finns att hämta i formelsamlingen. Uppgift (a) borde lydit (a) ”Ange, t.ex genom att resonera kombinatoriskt,…”.

6. Ett standard konfidensintervall för förändringen är

$$0.125 - 0.098 \pm 1.96 \cdot \sqrt{\frac{0.125 \cdot 0.875}{1027} + \frac{0.098 \cdot 0.902}{1233}} = 0.027 \pm 0.026$$

Man kan även lösa uppgiften genom ett homogenitetstest så här: Man listar först ut att kontingenstabellen är

$$\begin{array}{cc|c} 128 & 121 & 249 \\ 899 & 1112 & 2011 \\ \hline 1027 & 1233 & 2260 \end{array}$$

och beräknar sedan χ^2-avståndet (som har 1 frihetsgrad) till

$$\frac{(128 - 1027 \cdot 249/2260)^2}{1027 \cdot 249/2260} + \frac{(249 - 1233 \cdot 249/2260)^2}{1233 \cdot 249/2260} + \frac{(899 - 1027 \cdot 2011/2260)^2}{1027 \cdot 2011/2260} + \frac{(1112 - 1233 \cdot 2011/2260)^2}{1233 \cdot 2011/2260} = 4.014$$

vilket $> 3.84 = \chi^2_{0.05}(1)$ och $< 5.02 = \chi^2_{0.025}(1)$. P-värdet ligger således i intervallet 0.025, 0.05 och vi kan därför förkasta hypotesen om homogenitet (d.v.s att inget har ändrats) på nivån 5%.
Båda beräkningarna styrker journalistens påstående. (När man säger att något är statistiskt säkerställt är det underförstått att konfidensen är 95% om något annat ej sägs.)
För att våra beräkningar ska vara korrekta, behöver man även veta att.... (Många har trott att man behöver veta att populationen man drar ur är väldigt stor, praktiskt taget oändlig, men det är något annat mycket viktigare som efterfrågades. Att populationen är mäst intill oändlig hör så självklart till allmänbildningen att det behöver man inte ange i uppgiftstexten.)

7. (a) $\bar{X}$ är normalfördelad med väntevärde μ och standardavvikelse $\sigma/\sqrt{n}$. För beviset av $E(\bar{X}) = \mu$ och $V(\bar{X}) = \sigma^2/n$, se kursen.
   (b) Se kursen.

   I uppgiftstexten borde funnits med att $x_1, \ldots, x_n$ är oberoende.

8. Vi räknar ut att χ^2-avståndet mellan uppmätta och estimerade förväntade frekvenser under nollhypotesen om oberoende är $\frac{(20-54\cdot42/150)^2}{54\cdot42/150} + \frac{(22-96\cdot42/150)^2}{96\cdot42/150} + \frac{(34-54\cdot108/150)^2}{54\cdot108/150} + \frac{(74-96\cdot108/150)^2}{96\cdot108/150} = 1.5750 + 0.8860 + 0.6125 + 0.3445 = 3.418$. Under H_0 är detta en obs som är $\chi^2(1)$-fördelad, och vi läser ur tabell att $\chi^2_{0.10}(1) = 2.71 < 3.418$ medan $\chi^2_{0.05}(1) = 3.84 > 3.418$. P-värdet ligger således i intervallet (0.05, 0.10), så ett test på nivån 5% av H_0 : oberoende mot H_1 : beroende förkastar ej. (Men observera att data ändå ganska starkt tyder på att egenskaperna är beroende.)

Det är viktigt att du själv reflekterar över de utelämnade motiveringarna i lösningsförslagen ovan.

Här följer ett problem som kunde varit med på tentamen den 16 nov 2007.

1. Låt $X_1, X_2, \ldots$ vara oberoende Bernoulli(p)-variabler. Vilken fördelning har
   (a) $S = \sum_{i=1}^n X_i$ (2 p)
   (b) $M = \min\{i \geq 1 : X_i = 1\}$ (2 p)

   Obs att i båda deluppgifterna ingår inte bara att tala om vad fördelningen heter, utan även att härleda resp uttryck för sannolikhetsfunktionen.