Statistik för Fysiker m. fl., MSN 260
Tentamenskrivning den 17 januari 2005
Lösningar

1. Visa att två oberoende diskreta stokastiska variabler också är okorrelerade. Ange ett exempel på beroende men okorrelerade stokastiska variabler. (4p)

a. Kalla de två oberoenden stokastiska variablerna X och Y och antag att de är heltalsvärda. Eftersom korrelationen är kovariansen $\mathbf{C}(X,Y)$ dividerad med produkten av standarddeviationerna, räcker det att visa att $\mathbf{C}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = 0$. Men

$$\mathbf{E}[XY] = \sum_{i,j} \mathbf{P}(X=i, Y=j) = \sum_{i,j} ij\mathbf{P}(X=i)\mathbf{P}(Y=j) =$$

$$= \sum_i i\mathbf{P}(X=i) \sum_j j\mathbf{P}(Y=j) = \mathbf{E}[X]\mathbf{E}[Y],$$

där den andra likheten gäller just pga oberoendet.

b. Bokens exempel är $X \in N(0,1)$ och $Y = |X|$. Mer allmänt kunde man ta X med vilken som helst fördelning som är symmetrisk kring origo och antar minst fyra värden. (Varför duger det inte med bara två, t. ex. $X \pm 1$, bägge med sannolikheten $\frac{1}{2}$?)

2. Utan någon riktig anledning har du fått för dig att testa om du har en mycket ovanlig sjukdom. Den förekommer i själva verket endast i cirka ett fall på tiotusen individer. Det blodprov som tas med efterföljande analys anses ge en säker diagnos: Av hundra sjuka individer ger provet i snitt positivt resultat för 99, medan risken för att en frisk individ skall uppvisa ett positivt provresultat är så liten som 0,01. Om du nu får höra att ditt prov var positivt, hur stor är då sannolikheten att du verkligen är sjuk? (4 p)

Använd Bayes sats, S = sjuk, R = resultatet positivt:

$$\mathbf{P}(S|R) = \frac{\mathbf{P}(S \cap R)}{\mathbf{P}(R)} =$$

$$\frac{\mathbf{P}(R|S)\mathbf{P}(S)}{\mathbf{P}(R|S)\mathbf{P}(S) + \mathbf{P}(R|S^c)\mathbf{P}(S^c)} \approx 0.1$$

Tröstrikt att veta!

3. En forskningsrapport innehåller 90% -iga konfidensintervall för tio olika konstanter. De olika intervallen baserar sig på oberoende mätserier.

a. Vilken är sannolikheten att de alla innehåller "sin" konstant? (1p)

b. Låt ν beteckna antalet intervall som missar sin konstant. Vilket är ν:s mest sannolika värde? (2p)

a. Vart och ett täcker sin konstant med sannolikheten 0.9 Eftersom intervallen är oberoende täcker alla sin konstant med sannolikheten $0.9^{10} \approx 0.35$.

b. Sannolikheten att k konstanter ska vara täckta är

$$\binom{10}{k} 0.9^k 0.1^{10-k}, k = 0, 1, 2 \ldots 10.$$

Kvoten mellan två på varandra följande sådana uttryck (för k och $k+1$) är

$$\frac{k+1}{10-k}\frac{0.9}{0.1} = 9\frac{k+1}{10-k}.$$

Detta är mindre än 1 för $k = 0$, större annars. Alltså är det mest sannolikt att ett konfidensintervall missar sitt mål.

4. Mottagarna av 2004 års Riksbankspris i ekonomisk vetenskap till Alfred Nobels minne, E. Prescott och F. Kydland, har i ett berömt arbete antagit att arbetslösheten u_t år t beror linjärt av skillnaden mellan inflation detta år π_t och dess väntevärde,

$$u_t = u^* - \alpha(\pi_t - \mathbb{E}[\pi_t]).$$

Storheten u^* kallar de "jämviktsarbetslöshet" och α tänker de sig positivt, ju större inflation, desto mindre arbetslöshet. Som framgår av nedanstående data för Sverige (www.scb.se), kan detta inte vara sant i strikt mening. Gör därför en linjär regressionsansats

$$u_t = u^* - \alpha x_t + e_t,$$

där den oberoende variabeln $x_t = \pi_t - \mathbb{E}[\pi_t]$ är avvikelsen från förväntad inflation och $e_t, t = 1998, 1999, \ldots, 2004$ är oberoende fel som är $N(0, \sigma^2)$ för något σ.

år	arbetslöshet (tusental)	inflation (procent)
1998	280	1,5
1999	240	1,5
2000	200	1,5
2001	175	3
2002	175	2,5
2003	220	2
2004	250	1

(Om du föredrar att ha även arbetslösheten i procentenheter, kan du sätta Sveriges arbetskraft till 4,5 miljoner människor.) Antag att $\mathbb{E}[\pi_t] = 2$ under dessa år.

a. Skatta u^* och α. (2p)

Som kursboken föreslår, skriv om regressionsekvationen till

$$u_t = u - \alpha(x_t - \bar{x}) + e_t,$$

där $u = u^* - \alpha\bar{x}$ Eftersom felen är normalfördelade, sammanfaller minsta-kvadrat och trolighets (=ML)skattningarna och blir (circumflex betecknar skattning)

$$\hat{u} = \bar{u} = 215,7,$$

$$\hat{\alpha} = -\frac{\sum(x_t - \bar{x})\sum(u_t - \bar{u})}{\sum(x_t - \bar{x})^2} \approx 44,7,$$

$$\hat{u^*} \approx 222,$$

där SCB:s data för perioden ger siffervärdena.

b. Skatta σ^2 väntevärdesriktigt. (2p)

Variansen skattas väntevärdesriktigt av residualkvadratsumman genom antalet observationer n minus två:

$$\sum(u_t - \hat{u^*} - \hat{\alpha}x_t)^2/(n-2) = s^2 \approx 1832.$$

c. Ange ett 95% -igt konfidensintervall för α. (2p)

Ett sådant ges av

$$\hat{\alpha} \pm ts/\sqrt{\sum(x_r - \bar{x})^2},$$

där t är 97,5-procentsfraktilen av av t-fördelningen med fem frihetsgrader.Intervallet blir ungefär (-100, 190) (om inte jag har räknat fel).

d. Kan man (inom ramen för denna enkla modell) på grundval av denna datamängd förkasta nollhypotesen om oberoende, $H_0 : \alpha = 0$? Resonera mer allmänt om modellens rimlighet som beskrivning av sambandet mellan arbetslöshet och inflation i Sverige de senaste decennierna. (2p)

Nej, det kan man inte eftersom konfidensintervallet innehåller noll. (Å andra sidan är ju detta en mycket kort dataserie.)

Arbetslösheten beror rimligen av en rad olika faktorer, lagstiftning, sjukskrivnings- och pensioneringsbestämmelser, utöver de renodlat ekonomiska som inflationen. Givetvis kan man inte utesluta att totalrelationen ändå skulle kunna sammanfattas i ett linjärt samband, men det finns å andra sidan kanppast något som tala för detta snarare än t ex ett polynomiellt samband, eller något helt annat funktionssamband – eller inget alls. De data som uppgiften ger kan inte rimligen anses styrka teorien.

5. Två kommunicerande kärl innehåller tillsammans M molekyler. Varje tidpunkt flyttar sig en slumnpmässigt vald molkeyl från det ena kärlet till det andra. (Att molekylen är slumpmässigt vald betyder att varje molekyl väljs med samma sannolikhet, $1/M$.) Observationsserien startar med alla molekyler i den ena behållaren, så att om denna vid tidpunkten t innehåller X_t molekyler, $t = 0, 1, 2, 3, \ldots$ så är $X_0 = M$.

a. Visa att $X_0, X_1, X_2, X_3 \ldots$ utgör en Markovkedja och ange dennas övergångssannolikheter. (2p)

Om $X_n = x$, så blir

$$X_{n+1} == \begin{cases} x-1 & \text{med sannolikheten } x/M \\ x+1 & \text{med sannolikheten } 1 - x/M \end{cases}$$

oavsett vad $X_0, X_1, \ldots X_{n-1}$ är.

b. I det långa loppet stabiliserar sig molekylfördelningen i den meningen att för alla k gäller att $\mathbb{P}(X_n = k) \to v_k$, som är den stationära molekylfördelningen. Kan du bevisa – eller åtminstone ange argument för – att

$$v_k = \binom{M}{k} 2^{-M}?$$

(4p)

Det är lätt att visa att allmänt är

$$v_k = \sum_j v_j p_{jk},$$

om p_{jk} betecknar övergångssannolikheten från j till k. Sätt nu in vad övergångssannolikheterna är i detta speciella fall, enligt a., och kolla att det angivna uttrycket är lösningen.

6. Den stokastiska variabeln X har frekvensfunktionen

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{för } x > \theta \\ 0 & \text{annars.} \end{cases}.$$

Låt $(X_1, X_2, \ldots, X_n)$ vara ett stickprov på X och sätt $Y = \min(X_1, X_2, \ldots, X_n)$.

(a) Vad har Y för fördelningsfunktion? Frekvensfunktion? (2 p)

$$\mathbf{P}(Y \leq y) = 1 - \mathbf{P}(Y > y) = 1 - (\mathbf{P}(X > y))^n = 1 - (1 - \mathbf{P}(X \leq y))^n$$

ger fördelningsfunktionen. Derivera, så får du frekvensfunktionen.

(b) Bestäm a och b så att $aY + b$ blir en väntevärdesriktig skattning av θ. (2 p)

Bestäm $\mathbf{E}[Y]$ och utnyttja

$$\theta = a\mathbf{E}[Y] + b.$$

(c) Hur stort stickprov krävs för att $V(aY + b)$ ska bli mindre än 10^{-4}? (2 p)

Här får man räkna ut variansen, men det går ju eftersom man har frekvensfunktionen!