Kurskod: TAMS65

Provkod: TEN1

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Tentamen tisdagen den 8 juni 2010 kl 14-18.

Hjälpmedel: Formelsamling i matematisk statistik utgiven av matematiska institutionen samt miniräknare med tömda minnen. Inga anteckningar i formelsamlingen är tillåtet.
Betygsgränser: 8-11 poäng ger betyg 3, 11.5-14.5 ger betyg 4 och 15-18 poäng ger betyg 5.
Resultatet meddelas via LADOK.

1. I samband med nya miljöskyddsnormer skulle införas ville man undersöka hur många företag som redan upfyllde de nya normerna. Man undersökte därför 100 slumpvis valda företag med minst 100 anställda samt 200 slumpvis valda företag med högst 10 anställda.

Företagstyp	Uppfyller norm	Uppfyller ej norm
< 10 anställda	113	87
> 100 anställda	74	26

   Testa på 5%-nivån om det finns någon signifikant skillnad mellan små och stora företag. (2p)

2. Låt $x_1, \ldots, x_n$ vara ett stickprov av storleken $n > 2$ av oberoende observationer från en stokastiskt variabel X med väntevärdet $\mathrm{E}(X) = \mu$ och variansen $\mathrm{Var}(X) = \sigma^2$. Antag att vi har följande två skattningar av väntevärdet μ

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{och} \quad \hat{\mu}_2 = \frac{x_1 + x_n}{2}.$$

   a) Visa att båda skattningarna är väntevärdesriktiga. (1p)

   b) Vilken skattning är effektivast? Motivera ditt svar. (1p)

   c) Är någon av skattningarna en konsistent skattning av μ?
   Motivera ditt svar. (1p)

3. Under perioden 2001-01-01 till 2010-01-01 (alltså under 10 år) inträffade det 157 större dagsnedgångar (nedgångar större än 2.5%) för indexet *OMX Stockholm 30*. Antag att sådana nedgångar inträffar enligt en Poissonprocess med intensiteten λ nedgångar per år.

   a) Punktskatta λ. (1p)

   b) Konstruera ett tvåsidigt konfidensintervall för λ under den aktuella tidsperioden med konfidensgraden approximativt 0.95. (2p)

4. Att rökning kan vara skadligt för hälsan tror de flesta på. På 1970-talet gjordes flera studier för att påvisa skadligheten till följd av rökning. Nedanstående datamaterial kommer från en sådan undersökning.

Data Display

Row	Country	x	x1	y	x2
1	USA	3900	39,0	256,9	0
2	Canada	3350	33,5	211,6	0
3	Australia	3220	32,2	238,1	0
4	New Zealand	3220	32,2	211,8	0
5	United Kingdom	2790	27,9	194,1	0
6	Switzerland	2780	27,8	124,5	0
7	Ireland	2770	27,7	187,3	0
8	Iceland	2290	22,9	110,5	0
9	Finland	2160	21,6	233,1	0
10	West Germany	1890	18,9	150,3	0
11	Netherlands	1810	18,1	124,7	0
12	Greece	1800	18,0	41,2	1
13	Austria	1770	17,7	182,1	0
14	Belgium	1700	17,0	118,1	0
15	Mexico	1680	16,8	31,9	1
16	Italy	1510	15,1	114,3	1
17	Denmark	1500	15,0	144,9	0
18	France	1410	14,1	59,7	1
19	Sweden	1270	12,7	126,9	0
20	Spain	1200	12,0	43,9	1
21	Norway	1090	10,9	136,3	0

Vi har att

$x =$ genomsnittlig cigarettkonsumtion per vuxen och år,
$x_1 = x/100$,
$x_2 =$ 1 för medelhavsländer och Mexico, 0 annars,
$y =$ dödligheten i kranskärlssjukdomar per 100 000 invånare i åldern 35-64 år.

Data har analyserats enligt modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

där ε-variablerna är oberoende och $N(0, \sigma)$-fördelade. Nedan finns en Minitab analys för modellen.

Regression Analysis: y versus x1; x2

The regression equation is
y = 76,2 + 4,08 x1 - 80,1 x2

Predictor	Coef	SE Coef	T
Constant	76,22	26,71	2,85
x1	4,083	1,077	3,79
x2	-80,09	19,94	-4,02

S = 34,8508 R-Sq = 75,3% R-Sq(adj) = 72,6%

Analysis of Variance

Source	DF	SS	MS	F
Regression	2	66746	33373	27,48
Residual Error	18	21862	1215	
Total	20	88608		

a) Tyder analysen på att cigarettkonsumtionen kan ha betydelse för dödligheten i kranskärlssjukdomar? Motivera ditt svar med ett lämpligt konfidensintervall eller test på nivån 0.05. (1p)

b) Är det skillnad på dödligheten för medelhavsländer och andra länder? Motivera ditt svar med ett lämpligt konfidensintervall eller test på nivån 0.05. (1p)

c) Konstruera ett 95% prediktionsintervall för dödligheten i kranskärlssjukdomar för ett medelhavsland med cigarrettkonsumtionen 3500 cigaretter per vuxen och år.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.587558 & -0.0223965 & -0.247131 \\ -0.022397 & 0.0009553 & 0.007875 \\ -0.247131 & 0.0078755 & 0.327424 \end{pmatrix}.$$

(2p)

5. Låt $x_1, \ldots, x_n$ vara ett stickprov från en stokastisk variabel X med täthetsfunktionen

$$f(x) = \lambda e^{-\lambda(x-x_t)}, \quad \text{för } x > x_t > 0,$$

där λ är en okänd parameter och x_t en känd konstant.

a) Skissa täthetsfunktionen och förklara vad det är för någon. (1p)

b) Skatta λ med maximum-likelihood-metoden. (2p)

6. En *autoregressiv modell av 1:a ordningen* (en AR(1)-modell) är en enkel modell som ger möjlighet att utföra prognoser av framtida variablers utfall. Modellen ges av uttrycket

$$X_t - \mu = \theta(X_{t-1} - \mu) + \varepsilon_t \quad \text{för } t = 1, 2, \ldots,$$

där $\mu = \mathrm{E}(X_t)$ och ε_t är oberoende med $\varepsilon_t \sim N(0, \sigma)$ för alla t.

Man kan visa att kovariansen mellan variabler i följden beräknas enligt

$$\mathrm{Kov}(X_t, X_{t+k}) = \frac{\sigma^2}{1-\theta^2}\theta^{|k|} \quad \text{för } k = 0, \pm 1, \pm 2, \ldots$$

Antag att $\theta = \frac{1}{2}$ och $\sigma^2 = \frac{3}{4}$. Beräkna värdet på konstanten c så att

$$\mathrm{P}(X_t - c < X_{t+1} < X_t + c) = 0.95.$$

(3p)

Kurskod: TAMS65

Lösningar

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Lösningar till tentamen tisdagen den 8 juni 2010 kl 14-18.

1. Homogenitetstest med följande tabell.

Företagstyp	Uppfyller norm	Uppfyller ej norm	n_i
< 10 anställda	$N_{11} = 113$	$N_{12} = 87$	200
	$200\hat{p}_1 = 124.67$	$200\hat{p}_2 = 75.33$	
> 100 anställda	$N_{21} = 74$	$N_{22} = 26$	100
	$100\hat{p}_1 = 62.33$	$100\hat{p}_2 = 37.67$	
N_i	187	113	$n = 300$
	$\hat{p}_1 = \frac{187}{300}$	$\hat{p}_2 = \frac{113}{300}$	

Detta ger nu teststorheten

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(N_{ij} - n_i\hat{p}_j)^2}{n_i\hat{p}_j}$$
$$= \frac{(113-124.67)^2}{124.67} + \frac{(87-75.33)^2}{75.33} + \frac{(74-62.33)^2}{62.33} + \frac{(26-37.67)^2}{37.67} \approx 8.7.$$

Förkasta hypotesen att det skulle vara lika om $T > c$ där c fås ur en $\chi^2((2-1)(2-1))$-tabell, dvs. $c = \chi^2_{0.95}(1) = 3.84$. Vi har alltså $\underline{\underline{T = 8.7 > 3.84 = c}}$, dvs. förkasta likhet.

Med stor sannolikhet så verkar det finnas skillnad.

2. Låt $\widehat{M}_1$ och $\widehat{M}_2$ vara

$$\widehat{M}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{och} \quad \widehat{M}_2 = \frac{X_1 + X_n}{2}.$$

a)

$$\mathrm{E}\left(\widehat{M}_1\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{E}\left(X_i\right)}_{=\mu} = \frac{1}{n}n\mu = \mu \quad \underline{\underline{\text{vvr.}}}$$

$$\mathrm{E}\left(\widehat{M}_2\right) = \mathrm{E}\left(\frac{X_1 + X_n}{2}\right) = \frac{\mathrm{E}(X_1) + \mathrm{E}(X_n)}{2} = \frac{\mu + \mu}{2} = \mu \quad \underline{\underline{\text{vvr.}}}$$

b)

$$\mathrm{Var}\left(\widehat{M}_1\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}\left(X_i\right)}_{=\sigma^2} = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$\mathrm{Var}\left(\widehat{M}_2\right) = \mathrm{Var}\left(\frac{X_1 + X_n}{2}\right) = \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_n)}{4} = \frac{\sigma^2 + \sigma^2}{4} = \frac{\sigma^2}{2}$$

Alltså, för $n > 2$ så har vi $\underline{\underline{\text{Var}\left(\widehat{M_1}\right) < \text{Var}\left(\widehat{M_2}\right)}}$ vilket ger att skattning 1 ($\widehat{M_1}$) $\underline{\underline{\text{är effektivast}}}$.

c) Båda skattningar är vvr, se a). Vidare gäller att

$$\text{Var}\left(\widehat{M_1}\right) \to 0 \quad \text{när } n \to \infty \qquad \text{och} \quad \text{Var}\left(\widehat{M_2}\right) = \frac{\sigma^2}{2} \quad \text{för alla } n.$$

Alltså, $\underline{\underline{\widehat{M_1} \text{ är en konsistent skattning av } \mu \text{ men } \widehat{M_2} \text{ är det inte}}}$.

3. Låt X_{10} = antalet nedgångar på tio år. Vi har då att $X_{10} \sim Po(10\lambda)$, där λ är det förväntade antalet nedgångar per år. Vidare så har vi en observation $x_{10} = 157$.
a) En punktskattning av λ är $\hat{\lambda} = \dfrac{x_{10}}{10} = \dfrac{157}{10} = \underline{\underline{15.7}}$.

b) $X_{10} \sim Po(10\lambda) \approx N(10\lambda, \sqrt{10\lambda})$ eftersom $10\hat{\lambda} > 15$. Detta ger hjälpvariabeln

$$\frac{X_{10} - 10\lambda}{\sqrt{10\lambda}} \approx N(0,1),$$

vilket ger approximationen

$$\frac{X_{10} - 10\lambda}{\sqrt{10\hat{\lambda}}} = \frac{X_{10} - 10\lambda}{\sqrt{X_{10}}} \approx N(0,1).$$

Stäng in hjälpvariabeln enligt

$$\begin{aligned} 0.95 &= \text{P}\left(-a < \frac{X_{10} - 10\lambda}{\sqrt{X_{10}}} < a\right) = \ldots \\ &= \text{P}\left(\frac{1}{10}\left(X_{10} - a\sqrt{X_{10}}\right) < \lambda < \frac{1}{10}\left(X_{10} + a\sqrt{X_{10}}\right)\right), \end{aligned}$$

där $a = 1.96$ fås ur en Normalfördelningstabell på 97.5%. Vi har nu konfidensintervallet för λ som

$$I_\lambda = \left(\frac{1}{10}\left(x_{10} \mp a\sqrt{x_{10}}\right)\right) = \underline{\underline{(13.2, 18.2)}}.$$

4. a) Bilda I_{β_1}. Hjälpvariabel $\dfrac{\hat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(18)$.
Vi har att $d(\hat{\beta}_1) = s\sqrt{h_{11}} = 1.077$, dvs. konfidensintervallet blir

$$I_{\beta_1} = \Big(\underbrace{\hat{\beta}_1}_{=4.083} \mp \underbrace{t_{0.975}(18)}_{=2.10} \underbrace{s\sqrt{h_{11}}}_{=1.077} \Big) = \underline{\underline{(4.083 \mp 2.262) > 0}}.$$

$\underline{\underline{\text{Ja, cigarettkonsumtionen verkar ha betydelse}}}$.

b) Bilda I_{β_2}. Hjälpvariabel $\dfrac{\hat{\beta}_2 - \beta_2}{S\sqrt{h_{22}}} \sim t(18)$.

Vi har att $d(\hat{\beta}_2) = s\sqrt{h_{22}} = 19.94$, dvs. konfidensintervallet blir

$$I_{\beta_2} = \big(\underbrace{\hat{\beta}_2}_{=-80.09} \mp 2.10 \underbrace{s\sqrt{h_{22}}}_{=19.94} \big) = \underline{\underline{(-80.09 \mp 41.87) < 0}}.$$

Ja, medelhavsländerna (och Mexico) har färre döda i kranskärlssjukdomar.

c) $x = 3500$ ger $x_1 = 35$ och medelhavsland ger $x_2 = 1$. Låt den ännu ej observerade stokastiska variabeln Y_0 vara

$$Y_0 = \beta_0 + 35\beta_1 + \beta_2 + \varepsilon_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0,$$

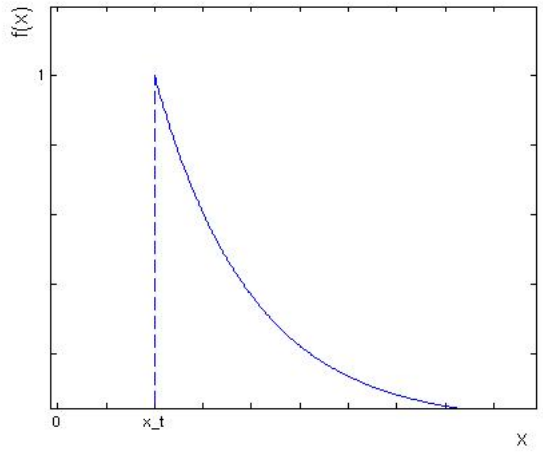
där $\varepsilon_0 \sim N(0, \sigma)$ och $\boldsymbol{u} = (1,\ 35,\ 1)'$.

Bilda nu I_{Y_0}. Hjälpvariabel $\dfrac{Y_0 - \boldsymbol{u}'\widehat{\boldsymbol{\beta}}}{S\sqrt{1 + \boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}} \sim t(18)$, där $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0,\ \widehat{\beta}_1,\ \widehat{\beta}_2)$.

Vi har nu att $\boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u} = 0.5745$ och $s = 34.8508$ vilket ger efter instängning av Y_0 intervallet

$$I_{Y_0} = \Big(\underbrace{\boldsymbol{u}'\widehat{\boldsymbol{\beta}}}_{=139.04} \mp \underbrace{t_{0.975}(18)}_{=2.10} \underbrace{s\sqrt{1 + \boldsymbol{u}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{u}}}_{=43.73} \Big) = \underline{\underline{(47.2, 230.9)}} \quad \text{(Väldigt långt...)}$$

5. a) $f(x) = \lambda e^{-\lambda(x - x_t)}$, för $x > x_t > 0$



$f(x)$ är täthetsfunktionen för en trunkerad exponentialfördelning (trunkerad vid x_t).

b) Likelihoodfunktionen ges av

$$L(\lambda) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda(x_i - x_t)} = \lambda^n e^{-\lambda \sum_{i=1}^{n}(x_i - x_t)}$$

med log-likelihoodfunktionen

$$l(\lambda) = \ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} (x_i - x_t).$$

Derivera map. λ och lös ekvationen

$$\frac{dl(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}(x_i - x_t) = 0$$

vilket ger

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n}(x_i - x_t) \quad \text{och} \quad \hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(x_i - x_t)}.$$

Alltså, vi har att

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}(x_i - x_t)} = \frac{n}{\sum_{i=1}^{n} x_i - nx_t)} = \underline{\underline{\frac{1}{\bar{x} - x_t}}}.$$

Vidare har vi att

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{n}{\lambda^2} < 0$$

vilket ger att $\hat{\lambda}$ är maximum-likelihood-skattningen.

6. Med $\theta = \dfrac{1}{2}$ och $\sigma^2 = \dfrac{3}{4}$ har vi att

$$\mathrm{Kov}(X_t, X_{t+k}) = \frac{\sigma^2}{1-\theta^2}\theta^{|k|} = \frac{\frac{3}{4}}{1-\left(\frac{1}{2}\right)^2}\left(\frac{1}{2}\right)^{|k|} = \left(\frac{1}{2}\right)^{|k|} \quad \text{för } k = 0, \pm 1, \ldots$$

Vi vill nu beräkna c så att

$$0.95 = \mathrm{P}(X_t - c < X_{t+1} < X_t + c) = \mathrm{P}(-c < \underbrace{X_{t+1} - X_t}_{=Y} < c)$$

$$= \mathrm{P}(-c < Y < c),$$

där Y är en linjärkombination av normalfördelningar enligt

$$Y = X_{t+1} - X_t = (-1,\ 1)\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix}.$$

Kovariansmatrisen för vektorn $\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix}$ ges av

$$\mathrm{Kov}\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} = \begin{pmatrix} \left(\frac{1}{2}\right)^0 & \left(\frac{1}{2}\right)^1 \\ \left(\frac{1}{2}\right)^1 & \left(\frac{1}{2}\right)^0 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

och alltså har vi att $\mathrm{E}(Y) = 0$ och

$$\mathrm{Var}(Y) = (-1,\ 1)\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = 1.$$

Konstanten c bestäms nu lätt genom

$$\mathrm{P}(-c < Y < c) = 0.95, \quad \text{där } Y \sim N(0,1),$$

dvs. $\underline{\underline{c = 1.96}}$