TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Matematisk statistik
EE

Kurskod: TAMS65

Provkod: TEN1

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Tentamen måndagen den 9 mars 2009 kl 14-18.

Hjälpmedel: Formelsamling i matematisk statistik utgiven av matematiska institutionen samt räknedosa med tömda minnen.

Betygsgränser: 8-11 poäng ger betyg 3, 11.5-14.5 poäng ger betyg 4, 15-18 poäng ger betyg 5.

Jourhavande lärare: Eva Enqvist, tel 281433.

Resultatet meddelas via LADOK.

Tydliga motiveringar krävs på varje uppgift.

Observera att uppgift 1 ger 2p och uppgift 4 ger 4p.

1. Med hänsyn till att bilmodeller lanseras tidigt under hösten kan man misstänka att valet av månad för inköp av ny bil inte sker slumpmässigt. Av 224 bilar sålda under ett år, räknat från 1 februari år 1 till och med 31 januari år 2, såldes 128 under perioden september - januari. Låt p vara sannolikheten att en person väljer att köpa bil under perioden september - januari.

   Konstruera ett tvåsidigt konfidensintervall för p med approximativ konfidensgrad 95%. Olika personer antas välja månad för bilinköp oberoende av varandra.

   Visar intervallet att bilköpen inte fördelar sig jämnt över årets tolv månader? Motivera ditt svar kortfattat. (2p)

2. Låt $x_1, \ldots, x_n$ vara observerade kölängder i ett kösystem. Observationerna har gjorts vid slumpmässiga tidpunkter med stora tidsavstånd, så vi kan anta att de s.v. $X_1, \ldots, X_n$ är oberoende. Vidare har vi sannolikhetsfunktionen

   $$p(x) = P(X_i = x) = \rho^x(1-\rho) \text{ för } x = 0, 1, 2, \ldots$$

   där ρ är den så kallade betjäningsfaktorn.

   a) Härled ML-skattningen av ρ och beräkna dess värde för följande observerade kölängder

   3 0 3 8 1 0 1 5 0 2

   (2p)

   b) Uppskatta med hjälp av resultatet i a) sannolikheten att kölängden är högst 1 vid en slumpmässig tidpunkt. (1p)

3. a) I en amerikansk undersökning ville man studera om elförbrukningen under tider med hög belastning minskade, om man fick rabatt på elpriset under tider med låg belastning. För tio hushåll mättes elkonsumtionen i maj månad under hög belastning året innan rabatten infördes och året efter. Resultat:

Hushåll	1	2	3	4	5	6	7	8	9	10
Före	200	180	240	425	120	333	418	380	340	516
Efter	160	175	210	370	110	298	368	250	305	477

Verkar det vara så att rabatten leder till att förbrukningen under tider med hög belastning minskar? Besvara frågan med hjälp av ett lämpligt konfidensintervall eller test. Nivå 0.05.

Du ska utnyttja en lämplig normalfördelningsmodell. Den valda modellen ska redovisas och motiveras. (2p)

b) Det händer ibland att datorer "låser sig" vilket bl.a. kan innebära att programvara behöver installeras på nytt. För sju nästan nya datorer har man registrerat tider mellan sådana låsningar och beräknat stickprovsstandardavvikelsen $s_1 = 14.17$. För sju datorer som varit i drift cirka ett år har man också registrerat sådana tider och fått $s_2 = 61.77$.

Pröva på nivån 0.01 hypotesen

$$H_0: \ \sigma_1 = \sigma_2 \quad \text{mot} \quad H_1: \ \sigma_1 \neq \sigma_2 .$$

Du får anta att tiderna mellan låsningar i de två fallen är oberoende och $N(\mu_1, \sigma_1)$ respektive $N(\mu_2, \sigma_2)$. (1p)

4. Vid industriell tillverkning studerar man så kallade produktivitetskvoter som mäter hur effektivt tillverkningen fungerar. Man har noterat (Schröder 1981) att det finns samband mellan produktivitetskvot och produktionsvolym. Ett företag har tre olika fabriker som producerar samma produkt. För var och en av dem har man observerat samhörande värden på produktivitetskvoten, Y, och antalet producerade enheter, x. Resultat

North Plant		South Plant		West Plant	
Productivity Ratio	No. of Units Produced	Productivity Ratio	No. of Units Produced	Productivity Ratio	No. of Units Produced
1.30	1000	1.43	1015	1.61	501
0.90	400	1.50	925	0.74	140
1.21	650	0.91	150	1.19	303
0.75	200	0.99	222	1.88	930
1.32	850	1.33	545	1.72	776
1.29	600	1.15	402	1.39	400
1.18	756	1.51	709	1.86	810
1.10	500	1.01	176	0.99	220
1.26	925	1.24	392	0.79	160
0.93	300	1.49	699	1.59	626
0.81	258	1.37	800	1.82	640
1.12	590	1.39	660	0.91	190

Man hoppades kunna hitta ett gemensamt samband mellan Y och x för de tre fabrikerna och gjorde först en analys enligt

Modell 1: $Y = \beta_0' + \beta_1' x + \beta_2' x^2/1000 + \varepsilon'$

För att undersöka om det verkade finnas systematiska skillnader mellan fabrikerna gjorde man också en analys enligt

Modell 2: $Y = \beta_0 + \beta_1 x + \beta_2 x^2/1000 + \gamma_2 z_2 + \gamma_3 z_3 + \varepsilon$

där

$$z_2 = \begin{cases} 1 & \text{för "South Plant"} \\ 0 & \text{annars} \end{cases} \qquad z_3 = \begin{cases} 1 & \text{för "West Plant"} \\ 0 & \text{annars.} \end{cases}$$

Samtliga ε-variabler antas vara oberoende och normalfördelade med väntevärde 0.

a) Titta först på analysen enligt modell 1.
Gör $x^2/1000$ nytta som förklaringsvariabel? Motivera ditt svar med hjälp av ett lämpligt 95% konfidensintervall. (1p)

b) Beskriver modell 2 data bättre än modell 1? Genomför ett lämpligt F-test på nivån 0.05. Både nollhypotes och mothypotes ska anges. (1.5p)

c) Tyder analysen enligt modell 2 på att det finns en systematisk skillnad mellan den västra och den södra fabriken? Punktskatta den parameter som beskriver skillnaden, konstruera ett 95% konfidensintervall för den och redovisa din slutsats. (1.5p)

Datorutskrifter från Minitab finns på nästa sida.

5. Ett företag som säljer mjukvara har tre olika kontor som hjälper kunderna med de problem som uppstår. För att undersöka om kontoren fungerar lika bra har man slumpmässigt valt ut samtal för varje kontor och undersökt om kundens problem har blivit löst. Samtalen har valts ur en mycket stor mängd samtal. Resultat:

	Problem	
Kontor	Löst	Ej löst
1	257	43
2	264	86
3	283	97

a) Är de tre kontoren lika framgångsrika i fråga om support? Genomför ett lämpligt χ^2-test på nivån 0.01. (2p)

b) Låt p_i för $i = 1, 2, 3$ vara sannolikheten att kontor nr i **misslyckas** med att lösa ett kundproblem. Man har länge misstänkt att kontor nr 1 fungerar bättre än de övriga. Undersök detta med ett lämpligt konfidensintervall för $\frac{1}{2}(p_2 + p_3) - p_1$ med approximativ konfidensgrad 95%. (1p)

Datorutskrift till uppgift 4.

Regression Analysis: Y versus x, x^2/1000 (enligt modell 1)

```
The regression equation is
Y = 0.482 + 0.00233 x - 0.00134 x^2/1000


Predictor        Coef    SE Coef (Stdev)
Constant       0.4818     0.1391
x           0.0023310  0.0005812
x^2/1000   -0.0013447  0.0005173


S = 0.197585   R-Sq = 62.9%   R-Sq(adj) = 60.7%


Analysis of Variance

Source          DF      SS      MS
Regression       2  2.1847  1.0923
Residual Error  33  1.2883  0.0390
Total           35  3.4730
```

Regression Analysis: Y versus x, x^2/1000, z2, z3 (enligt modell 2)

```
The regression equation is
Y = 0.176 + 0.00264 x - 0.00154 x^2/1000 + 0.225 z2 + 0.400 z3


Predictor        Coef    SE Coef (Stdev)
Constant      0.17597     0.08352
x           0.0026429  0.0003172
x^2/1000   -0.0015385  0.0002816
z2            0.22464     0.04392
z3            0.39977     0.04448


S = 0.107134   R-Sq = 89.8%   R-Sq(adj) = 88.4%


Analysis of Variance

Source          DF       SS       MS
Regression       4  3.11718  0.77929
Residual Error  31  0.35581  0.01148
Total           35  3.47299


MTB > print m1
```

Data Display

$$
\begin{matrix}
0.607735 & -0.0020367 & 0.0016417 & -0.110053 & -0.127659 \\
-0.002037 & 0.0000088 & -0.0000076 & 0.000105 & 0.000131 \\
0.001642 & -0.0000076 & 0.0000069 & -0.000086 & -0.000079 \\
-0.110053 & 0.0001051 & -0.0000856 & 0.168030 & 0.085527 \\
-0.127659 & 0.0001309 & -0.0000794 & 0.085527 & 0.172407
\end{matrix}
= (\mathbf{X}^T\mathbf{X})^{-1}
$$

6. Med hjälp av en ny inflationsmodell har man ett halvår i förväg förutspått att det förväntade priset på en matkorg med ett visst innehåll i ett område i en amerikansk delstat kommer att vara 145.75 dollar. Priset i en slumpmässigt vald affär för en matkorg med det aktuella innehållet anses vara en stokastisk variabel som är $N(\mu, \sigma)$, där omfattande mätningar under ganska lång tid visat att $\sigma = 9.5$ är ett rimligt värde. Man misstänker att inflationsmodellen underskattar prisökningarna. Matkorgar från tio slumpmässigt valda affärer i det aktuella området visar sig ha ett genomsnittspris $\bar{x} = 149.80$ dollar.

a) Pröva på nivån 0.05 hypotesen

$$H_0;\ \mu = 145.75 \qquad \text{mot} \qquad H_1;\ \mu > 145.75\,. \qquad \text{(1p)}$$

b) För hur många affärer borde man ha tagit reda på priset för matkorgen om man vill pröva H_0 mot H_1 med hjälp av ett test på nivån 0.05 och vill att H_0 ska förkastas med minst sannolikheten 0.80 om $\mu = 149.75$? (2p)

Lösningar till tentamen i TAMS65, 2009-03-09.

1. $x = 128$ är observation av $X \sim Bin(n,p)$ där $n = 224$.

$\hat{p} = \frac{x}{n} = 0.5714$

Eftersom $n\hat{p}\hat{q} = 54.86 > 10$ är normalapproximation för X tillåten. Då gäller också att

$\hat{P} = \frac{X}{n}$ appr $N(p, \sqrt{\frac{pq}{n}})$ eftersom

$E(\hat{P}) = E(\frac{X}{n}) = \frac{1}{n} \cdot np = p$

$Var(\hat{P}) = Var(\frac{X}{n}) = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$

Hjälpvariabeln $\frac{\hat{P} - p}{\sqrt{\frac{\hat{P}\hat{Q}}{n}}}$ är appr $N(0,1)$ och ger

$I_p = (\hat{p} \mp 1.96\sqrt{\frac{\hat{p}\hat{q}}{224}}) = (0.507, 0.636)$

Om köparna väljer månad slumpmässigt skulle $p = \frac{5}{12} = 0.417 < 0.507$. De som köper bil verkar föredra perioden september–januari.

2. a) $L(\vartheta) = [\vartheta^{x_1}(1-\vartheta)] \cdot \ldots \cdot [\vartheta^{x_n}(1-\vartheta)] =$

$= \vartheta^{\sum_1^n x_i}(1-\vartheta)^n$

$l(\vartheta) = \ln L(\vartheta) = \sum_1^n x_i \ln\vartheta + n\ln(1-\vartheta)$

$l'(\vartheta) = \sum_1^n x_i \cdot \frac{1}{\vartheta} + n \cdot \frac{1}{1-\vartheta} \cdot (-1) = \frac{\sum_1^n x_i - \vartheta\sum_1^n x_i - n\vartheta}{\vartheta(1-\vartheta)} =$

$= \frac{n[\bar{x} - \vartheta(1+\bar{x})]}{\vartheta(1-\vartheta)}$

$l'(\vartheta) = 0$ för $\hat{\vartheta} = \frac{\bar{x}}{1+\bar{x}}$;

ϑ	0	$\hat{\vartheta}$	1
$l'(\vartheta)$		+ 0 −	
$l(\vartheta)$		↗	↘

Alltså maximum d.v.s. ML-skattn. $\hat{\vartheta} = \frac{\bar{x}}{1+\bar{x}} = 0.697$

b) $P(X \le 1) = 1 - \vartheta + (1-\vartheta)\vartheta = 1 - \vartheta^2$ som skattas med $1 - \hat{\vartheta}^2 \simeq 0.514$

3. a) Eftersom vi har parvisa mätningar bildar vi differenserna $d_i = x_i - y_i$, där x_i är mätvärdet före rabatt och y_i mätvärdet efter rabatt för hushåll nr i.

d_i : 40 5 30 55 10 35 50 130 35 39

Modell: De s.v. $D_i \sim N(\mu_D, \sigma_D)$, där μ_D är ett mått på den systematiska skillnaden.

$\hat{\mu}_D = \bar{d} = 42.9$; $\hat{\sigma}_D = s_D = 34.35$.

Vi vill visa att $\mu_D > 0$ och konstruerar därför ett nedåt begränsat konfidensintervall för μ_D.

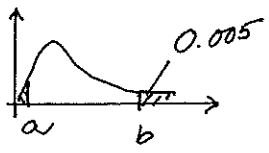
Hjälpvariabeln $\dfrac{\bar{D} - \mu_D}{S_D / \sqrt{10}} \sim t(9)$ och ger

$$I_{\mu_D} = \left(\bar{d} - t \cdot \frac{s_D}{\sqrt{10}}, \infty \right) = (42.9 - 19.88, \infty)$$

$$= (23.0, \infty) \text{ där } t = 1.83.$$

Bara positiva värden i I_{μ_D}, vilket tyder på att rabatten ger en sytematisk minskning av elkonsumtionen under tid med hög belastning, men skillnaden skulle också kunna bero på skillnad i vädret mellan år 1 och år 2.

b) Teststorhet: $v = \dfrac{s_2^2}{s_1^2} = 19.00$



Den s.v. $V \sim F(6,6)$ om H_0 är sann.

H_0 förkastas om $v < a$ el. $v > b$.

$F(b) = 0.995$; tabell ger $b = 11.07$

$P(V \leq a) = 0.005$ d.v.s. $P\left(\frac{1}{V} \geq \frac{1}{a}\right) = 0.005$

Då den s.v. $\frac{1}{V} \sim F(6,6)$ får vi $\frac{1}{a} = 11.07$ och $a = 0.0903$.

$19.00 > 11.07$; H_0 förkastas; $\sigma_2 > \sigma_1$ med stor sth.

4. a) $I_{\beta_2'} = (\hat{\beta}_2' \mp t \cdot s\sqrt{h_{22}}) = (-0.0013447 \mp t \cdot 0.0005173) =$

$= (-0.0013447 \mp 0.0010522) = (-0.00240, -0.00029)$

där $t = 2.034$ ges i $t(33)$-tabell.

0 ∉ $I_{\beta_2'}$. Alltså verkar $x^2/1000$ göra nytta som förkl. variabel.

b) Vi prövar $H_0: \gamma_2 = \gamma_3 = 0$ mot H_1: minst en av γ_2 och γ_3 är skild från 0, där H_0 innebär att modell 1 duger och H_1 att modell 2 är bättre.

Teststorhet: $W = \dfrac{(Q_{RES}^{(1)} - Q_{RES}^{(2)})/2}{Q_{RES}^{(2)}/31} = 40.62$

H_0 förkastas om $W > c$. Den s.v. $W \sim F(2, 31)$ om H_0 är sann. Tabell ger $c = 3.31$.

$40.62 > 3.31$; H_0 förkastas; modell 2 beskriver data bättre.

c) Parametern $\gamma_3 - \gamma_2$ beskriver den systematiska skillnaden mellan de två fabrikerna.

Vi konstruerar $I_{\gamma_3 - \gamma_2}$:

$$\hat{\gamma}_3 - \hat{\gamma}_2 = (0\ 0\ 0\ -1\ 1) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = w^T\hat{\beta} = 0.1751$$

Den s.v. $\hat{\gamma}_3 - \hat{\gamma}_2 \sim N(\gamma_3 - \gamma_2, \sigma\sqrt{0.1694})$

eftersom $w^T (X^T X)^{-1} w = 0.1694$.

σ^2 skattas med $s^2 = \dfrac{Q_{RES}}{31}$; $s = 0.1071$; fr.gr.: 31.

Hjälpariabeln $\dfrac{\hat{\gamma}_3 - \gamma_2 - (\gamma_3 - \gamma_2)}{s\sqrt{0.1694}} \sim t(31)$ och ger

$I_{\gamma_3 - \gamma_2} = (\hat{\gamma}_3 - \hat{\gamma}_2 \mp 2.038\, s\sqrt{0.1694}) = (0.082, 0.262)$

Bara positiva värden. Västra fabriken bättre.

5. a) H_0: Sannolikheten p att ett problem inte blir löst är lika stor för de tre kontoren

mot

H_1: H_0 ej sann

prövas med ett homogenitetstest.

H_0 sann: $\hat{p} = \frac{43+86+97}{1030} = 0.2194$; $1-\hat{p} = 0.7806$.

Vi får de skattade förväntade frekvenserna $n_i(1-\hat{p})$ resp. $n_i\hat{p}$:

Kontor	Löst	Ej löst
1	234.18	65.82
2	273.21	76.19
3	296.63	83.37

Teststorhet: $Q = \frac{(257-234.18)^2}{234.18} + \frac{(43-65.82)^2}{65.82} +$

$+ \ldots + \frac{(97-83.37)^2}{83.37} = 14.41$

H_0 förkastas om $Q > c$. Den s.v. Q är appr $\chi^2((3-1)(2-1)) = \chi^2(2)$ om H_0 är sann.

Tabell ger $c = 9.22$.

$14.41 > 9.22$; H_0 förkastas. De tre kontoren fungerar inte lika bra.

b) Vi vill visa att $\theta = \frac{p_2+p_3}{2} - p_1 > 0$ och konstruerar därför I_θ, nedåt begränsat.

$\hat{p}_1 = \frac{x_1}{n_1} = 0.1433$; $\hat{p}_2 = \frac{x_2}{n_2} = 0.2457$; $\hat{p}_3 = \frac{x_3}{n_3} = 0.2553$

$\hat{\theta} = \frac{1}{2}(\hat{p}_2+\hat{p}_3) - \hat{p}_1 = 0.1072$

Eftersom $n_i\hat{p}_i(1-\hat{p}_i) > 10$ är de s.v. $\hat{P}_i$ appr. normalfördelade och då är även $\hat{\Theta}$ appr. normalfördelad med parametrar

$E(\hat{\Theta}) = E\left(\frac{1}{2}\cdot\frac{X_2}{n_2} + \frac{1}{2}\cdot\frac{X_3}{n_3} - \frac{X_1}{n_1}\right) = \frac{1}{2}p_2 + \frac{1}{2}p_3 - p_1 =$

$= \theta$

$$\mathrm{Var}(\hat{\Theta}) = \frac{1}{4}\cdot\frac{n_2 p_2 q_2}{n_2^2} + \frac{1}{4}\cdot\frac{n_3 p_3 q_3}{n_3^2} + \frac{n_1 p_1 q_1}{n_1^2}$$

Hjälpvariabeln $\dfrac{\hat{\Theta} - \Theta}{\sqrt{\frac{\hat{p}_2\hat{q}_2}{4n_2} + \frac{\hat{p}_3\hat{q}_3}{4n_3} + \frac{\hat{p}_1\hat{q}_1}{n_1}}}$ appr $N(0,1)$.

Ger $I_\theta = (\hat{\theta} - 1.645\, d(\hat{\theta}), 1) = (0.065, 1)$

Alltså är $\theta \geq 0.065 > 0$. Kontor 1 är bättre.

6. a) $\hat{\mu} = \bar{x}$; den s.v. $\bar{X} \sim N(\mu, \frac{9.5}{\sqrt{n}})$.

Teststorhet: $u = \dfrac{\bar{x} - 145.75}{9.5/\sqrt{n}} = 1.348 \quad (n = 10)$

Den s.v. $u \sim N(0,1)$ om H_0 är sann.

H_0 förkastas om $u > 1.645$ för $\alpha = 0.05$.

$1.348 < 1.645$. Alltså kan vi inte förkasta H_0.

b) H_0 förkastas om $\dfrac{\bar{x} - 145.75}{9.5/\sqrt{n}} > 1.645$ (jfr. a))

$0.80 \leq P(H_0 \text{ förkastas om } \mu = 149.75) =$

$= P\left(\dfrac{\bar{X} - 145.75}{9.5/\sqrt{n}} > 1.645 \text{ om } \mu = 149.75\right) =$

$= P(\bar{X} > 145.75 + 1.645 \cdot 9.5/\sqrt{n} \text{ om } \mu = 149.75) =$

$= P\left(\dfrac{\bar{X} - 149.75}{9.5/\sqrt{n}} > \dfrac{-4 + 1.645 \cdot 9.5/\sqrt{n}}{9.5/\sqrt{n}} \text{ om } \mu = 149.75\right) =$

$= 1 - \Phi\left(-\dfrac{4\sqrt{n}}{9.5} + 1.645\right) = \Phi\left(\dfrac{4\sqrt{n}}{9.5} - 1.645\right)$

$\dfrac{4\sqrt{n}}{9.5} - 1.645 \geq 0.842$; $n \geq \left[\dfrac{9.5}{4}(1.645 + 0.842)\right]^2$

Alltså $n \geq 35$.