Kurskod: TAMS65
Provkod: TEN1

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Tentamen måndagen den 15 augusti 2016 kl 8–12

Hjälpmedel: Formelsamling i matematisk statistik utgiven av matematiska institutionen och/eller formelsamling "Formel- och tabellsamling i matematisk statistik TAMS65 (Martin Singull)". Inga anteckningar i formelsamlingarna är tillåtet. Miniräknare med tömda minnen.
Betygsgränser: 8-11 poäng ger betyg 3, 11.5-14.5 ger betyg 4 och 15-18 poäng ger betyg 5.
Examinator: Martin Singull, 013–281447
Resultatet meddelas *normalt* via LADOK inom 12 arbetsdagar.
Tydliga svar och motiveringar krävs till varje uppgift.

1. En leverantör påstår att mängden mjöl i en 2000g-påse kan betraktas som en s.v. $X \sim N(2000, 10)$. Vid en undersökning av 200 påsar fick man följande resultat:

Mjölets vikt	<1990g	1990-2010g	>2010g
Antal påsar	48	109	43

   Mjölmängden i olika påsar är oberoende av varandra.

   a) Pröva med hjälp av ett χ^2-test på nivån 1% hypotesen att mjölmängderna är normalfördelade precis som leverantören påstår. (2p)

   b) Låt p vara sannolikheten att en mjölpåse innehåller mindre än 1990g mjöl. Konstruera ett tvåsidigt konfidensintervall för p med konfidensgraden approximativt 95%. (1p)

2. Två grupper om vardera 90 patienter deltog i ett experiment i vilket den ena gruppen fick medicin mot allergi medan den andra gruppen fick placebo (verkningslöst preparat). I den första gruppen uppvisade 32 personer allergiska symptom och i placebogruppen 51 personer sådana symptom. Är det tillräckligt bevis för att man på nivån 5% ska kunna dra slutsatsen att allergimedicinen minskar risken för allergiska reaktioner? (3p)

3. Y_1 och Y_2 är oberoende stokastiska variabler $Y_1 \sim N(3, 1)$ och $Y_2 \sim N(5, 2)$.

   a) Bestäm fördelningen för den stokastiska vektorn $\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ där $U_1 = 2Y_1 - Y_2$ och $U_2 = Y_1 + Y_2$. (1p)

   b) Låt istället $U_1 = aY_1 - Y_2$. För vilket värde på a är U_1 och U_2 oberoende? (1p)

4. Man vill utnyttja en regressionsmodell för att beräkna energiförbrukningen i villor. Som beroendevariabel har man y = energiförbrukning per villa (enhet: 1000-tal kWh) och som förklaringsvariabler: x_1 medeltemperatur ($^\circ C$), x_2 bostadsyta (m^2) samt x_3 isolering som är 1 för ja och 0 för nej. Observerade värden:

x_1	17.8	16.6	12.2	7.1	2.8	0.1	-2.9	-3.1	-0.7	4.4
x_2	130	190	150	190	210	250	190	155	180	160
x_3	0	1	0	1	1	1	1	0	0	1
y	7	10	15.5	7	25	31	29	30	28.5	21

Man analyserar observationerna enligt modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

där $\varepsilon \sim N(0, \sigma)$ (oberoende). I analysen har vi fått den skattad regressionslinje och variansanalys:

$$y = 16.2 - 0.99x_1 + 0.069x_2 - 2.99x_3,$$

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	16.2196	3.0243
1	-0.9925	0.0591
2	0.0685	0.0180
3	-2.9897	1.0919

VARIANSANALYS

	Frihetsgrader	Kvadratsumma
REGR	?	670.0211
RES	?	8.8789
TOT	9	678.9000

(observera att SS_{REGR} och SS_{RES} hade bytt plats på original tentan) och

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 6.180779 & -0.065514 & -0.035783 & 1.223042 \\ -0.065514 & 0.002357 & 0.000326 & -0.010121 \\ -0.035783 & 0.000326 & 0.000219 & -0.009150 \\ 1.223042 & -0.010121 & -0.009150 & 0.805691 \end{pmatrix}$$

a) Hur många frihetsgrader har kvadratsummorna REGR och RES? (1p)

b) Verkar ytan på husen vara av betydelse för energiförbrukningen? Genomför ett lämpligt test på nivån 1%. (1p)

c) Vilken energiförbrukning har ett speciellt hus i Linköping (medeltemperatur $7.9°C$) med isolering och en yta på 170 m^2? Bilda ett lämpligt 95% intervall för att besvara frågan. (2p)

5. Låt $X_1, X_2, ..., X_n$ vara ett stickprov från en fördelning med sannolikhetsfunktionen

$$p(x) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, \quad \text{för } x = -1, 0, 1 \quad \text{och } 0 \leq \theta \leq 1.$$

a) Beräkna maximum-likelihood-skattningen av θ. (2p)

b) Undersök om ML-skattningen är väntevärdesriktig. (1p)

6. Man har sex observationer från $N(\mu_1, \sigma)$ och åtta observationer från $N(\mu_2, 2\sigma)$.

1)	10.4	8.7	9.9	9.0	10.5	8.6		
2)	11.5	13.0	12.0	9.5	13.1	9.7	10.6	12.4

a) Pröva $H_0 : \sigma = 1$ mot $H_1 : \sigma < 1$ på nivån 10%. Hela datamaterialet måste utnyttjas. (1.5p)

b) Beräkna testets styrka för $\sigma = 0.63$. (1.5p)

Kurskod: TAMS65
Lösningar

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Lösningsförslag till tentamen måndagen den 15 augusti 2016 kl 8–12.

1a) $P(X < 1990) = \Phi\left(\dfrac{1990 - 2000}{10}\right) = \Phi(-1) = 1 - \Phi(1) = 0.1537$

$P(1990 < X < 2010) = \Phi(1) - \Phi(-1) = 0.6826$

$P(2010 < X) = 1 - \Phi(1) = 0.1537$

Teststorhet:

$$Q = \frac{(48 - 31.74)^2}{31.74} + \frac{(109 - 136.52)^2}{136.52} + \frac{(43 - 31.74)^2}{31.74} = 17.87 > \chi^2_{0.99}(2) = 9.92.$$

Alltså, leverantörens normalfördelning kan förkastas på nivån 1%

b) $\hat{p} = \dfrac{48}{200} = 0.24$ och $200\hat{p}(1-\hat{p}) = 36.48 > 10$

Hjälpvariabeln $\dfrac{\widehat{P} - p}{\sqrt{\dfrac{\widehat{P}(1-\widehat{P})}{200}}} \approx N(0,1)$ ger intervallet

$$I_p = \left(\hat{p} \mp 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{200}}\right) = (0.24 \mp 0.06) = \underline{(0.18; 0.30)}$$

2) $x = 32$ obs av $X \sim Bin(90, p_1)$ och $y = 51$ obs av $Y \sim Bin(90, p_2)$. Testa hypotesen $H_0 : p_1 = p_2$ mot $H_1 : p_1 < p_2$ på nivån 5%.

$$I_{p_2 - p_1} = \left(\hat{p}_2 - \hat{p}_1 - 1.645\sqrt{\frac{\hat{p}_1\hat{q}_1}{90} + \frac{\hat{p}_2\hat{q}_2}{90}}; 1\right) = (0.0916; 1)$$

På nivån 5% kan vi alltså dra slutsatsen att allergimedicinen minskar risken för allergiska reaktioner.

3a) Transformationen kan skrivas som $\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ och det gäller att

$E(\mathbf{U}) = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \end{pmatrix}$ och $\mathbf{C_U} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 8 & -2 \\ -2 & 5 \end{pmatrix}$.

$\mathbf{U}$ är normalfördelad eftersom det är en transformation av en normalfördelad vektor. Alltså gäller att

$$\mathbf{U} \sim N_2\left(\begin{pmatrix} 1 \\ 8 \end{pmatrix}, \begin{pmatrix} 8 & -2 \\ -2 & 5 \end{pmatrix}\right).$$

b) $\mathbf{C_U} = \begin{pmatrix} a & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\begin{pmatrix} a & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} a^2+4 & a-4 \\ a-4 & 5 \end{pmatrix}$

U_1 och U_2 är oberoende om $\mathbf{C_U}$ är en diagonalmatris, alltså om $a = 4$.

4a) $df_{REGR} = 3$ och $df_{RES} = 6$

b) Testa hypotesen $H_0 : \beta_2 = 0$ mot $H_1 : \beta_2 \neq 0$ på nivån 1%

Teststorhet: $T = \dfrac{\hat{\beta}_2}{d(\hat{\beta}_2)} \sim t(6)$ då H_0 är sann. $t = \dfrac{\hat{\beta}_2}{d(\hat{\beta}_2)} = 3.81 > t_{0.995}(6) = 3.71$. Alltså, förkasta H_0, bostadens yta har med stor sannolikhet betydelse för energiförbrukningen.

Man kan också dra samma slutsats från konfidensintervallet $I_{\beta_2} = (\hat{\beta}_2 \mp t_{0.995}(6) d(\hat{\beta}_2)) = (0.0017; 0.1353)$.

c) Bilda prediktionsintervall för $Y_0 = \beta_0 + 7.9\beta_1 + 170\beta_2 + \beta_3 + \varepsilon_0$. Låt $\boldsymbol{u} = \begin{pmatrix} 1 & 7.9 & 170 & 1 \end{pmatrix}'$. Prediktionsintervallet ges av

$$I_{Y_0} = \left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}} \mp t\, s \sqrt{1 + \boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}} \right) = (13.6; 20.5),$$

där $\boldsymbol{u}'\widehat{\boldsymbol{\beta}} = 17.03$, $t = t_{0.975}(6) = 2.45$, $s = 1.2165$ och $\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u} = 0.3121$. Ganska brett intervall.

(Självklart ger även $s^2 = 670.0211/6$ poäng på uppgiften eftersom det var vad som var givet på original tentan. Intervallet ges då av $I_{Y_0} = (0; 46.7)$.)

5a) $L(\theta) = \left(\dfrac{\theta}{2}\right)^{\sum_{i=1}^n |x_i|} (1-\theta)^{n - \sum_{i=1}^n |x_i|}$

$\ln L(\theta) = \ln \dfrac{\theta}{2} \sum_{i=1}^n |x_i| + (n - \sum_{i=1}^n |x_i|) \ln(1-\theta)$

$\dfrac{\partial \ln L}{\partial \theta} = \dfrac{\sum_{i=1}^n |x_i|}{\theta} - \dfrac{n - \sum_{i=1}^n |x_i|}{1-\theta} = 0 \quad$ ger $\quad \widehat{\theta} = \dfrac{1}{n}\sum_{i=1}^n |x_i|$ samt $\widehat{\Theta} = \dfrac{1}{n}\sum_{i=1}^n |X_i|$

b) $E(\widehat{\Theta}) = E\left(\dfrac{1}{n}\sum_{i=1}^n |X_i|\right) = \dfrac{1}{n}\sum_{i=1}^n E\left(|X_i|\right)$

$= \dfrac{1}{n} n \left(\dfrac{\theta}{2} \cdot 1 + 0 \cdot (1-\theta) + \dfrac{\theta}{2} \cdot 1\right) = \theta \Rightarrow$ Skattningen är väntevärdesriktig.

6a) Vi har att $x_1, ..., x_6$ obs från $N(\mu_1, \sigma)$ och $y_1, ..., y_8$ obs från $N(\mu_2, 2\sigma)$.
Låt $z_i = \dfrac{y_i}{2}$. Då gäller att $z_1, ..., z_8$ obs från $N(\mu_2/2, \sigma)$.

Variansen σ^2 skattas nu med $s^2 = \dfrac{5s_x^2 + 7s_z^2}{12} = 0.596$.

Teststorhet: $\dfrac{12s^2}{1} = 7.149$. Den s.v. $\dfrac{12s^2}{1} \sim \chi^2(12)$ om H_0 är sann.

H_0 förkastas om $12s^2 < 6.30$ men $7.149 > 6.30$. Alltså vi kan inte förkasta H_0.

b) Styrkan för $\sigma = 0.63$ ges av

$$P(12S^2 < 6.30 \text{ om } \sigma = 0.63) = P\left(\frac{12S^2}{0.63^2} < 15.83 \text{ om } \sigma = 0.63\right) \approx 80\%,$$

från tabell för $\dfrac{12S^2}{0.63^2} \sim \chi^2(12)$.