Kurskod: TAMS65
Provkod: TEN1

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Tentamen fredagen den 27 oktober 2017 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, Matematisk statistik, MAI
Resultatet meddelas *normalt* via LADOK inom 12 arbetsdagar.
Tydliga svar och motiveringar krävs till varje uppgift.

1. Vid en intervjuundersökning ställdes bland annat två frågor, vilka man önskade studera kombinationen av svaren. På den första frågan fanns svarsalternativen A, B, C och D och på den andra frågan fanns svarsalternativen a, b och c. Man fick nedanstående tabell:

		Fråga 1			
		A	B	C	D
	a	26	29	25	39
Fråga 2	b	31	30	41	44
	c	16	33	43	29

   Kan man på 5%-nivån påvisa något samband mellan åsikterna vad det gäller de båda frågorna? (2p)

2. Man har noterat följande åldersjusterade årsvärden för antalet fall av malignt melanom per miljoner invånare för en viss region som valts slumpmässigt bland liknande regioner:

År						$\bar{x}_i$	s_i
1945-49:	15	15	20	25	27	20.4	5.55
1955-59:	29	25	26	32	38	30.0	5.24
1965-69:	39	41	38	47	44	41.8	3.70

   Modell: Vi har tre oberoende stickprov från $N(\mu_i, \sigma_i), i = 1, 2, 3$. Vi betraktar också observationerna inom varje stickprov som oberoende.

   a) Har de tre olika stickproven samma standardavvikelse? Genomför lämpligt test på nivån 5%. Det räcker att du gör ett test men motivera hur du drar dina slutsatser. (1p)

b) Antag att $\sigma_1 = \sigma_2 = \sigma_3$. Kan man med någon säkerhet hävda att väntevärdet för antalet fall av malignt melanom per miljoner invånare har ökat med mer än 50% från senare delen av 40-talet till senare delen av 60-talet, dvs. gäller det att $\mu_3 > 1.5\mu_1$? Motivera ditt svar med hjälp av ett lämpligt 95% konfidensintervall. (2p)

3. Man vill utnyttja en regressionsmodell för att beräkna energiförbrukningen för mindre likvärdiga tillverkningsindustrier. Som beroende variabel har man $y = 1000$-tals kWh per industri och månad och som förklaringsvariabler

$x_1 =$ kvalitetsindex som relaterar till kvaliteten på de tillverkade enheterna (högt värde = god kvalitet och noll innebär en standard kvalitet),

$x_2 =$ antalet sekunder per tillverkad enhet,

$x_3 =$ geografisk plats (1 för Sverige, och 0 för Centraleuropa).

Mätningar på 10 industrier gav följande resultat.

y	7	10	18.5	30	31	65	52.5	64	55	27.5
x_1	17.8	16.6	12.2	7.1	2.8	0.1	-2.9	-3.1	-0.7	4.4
x_2	170	210	150	190	110	250	140	155	180	130
x_3	0	1	0	1	1	1	1	0	0	1

Observationerna analyserades enligt två modeller

$$\textbf{Modell 1:} \quad Y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$\textbf{Modell 2:} \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

med olika parametrar och ε-variabler som är oberoende och fördelade $\varepsilon \sim N(0, \sigma)$.

Resultat **Modell 1**: $y = 50.1 - 2.59x_1$,

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	50.0930	3.3628
1	–2.5862	0.3675

VARIANSANALYS

	Frihetsgrader	Kvadratsumma
REGR	1	3628.6
RES	8	586.1
TOT	9	4214.7

och

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.154348 & -0.010009 \\ -0.010009 & 0.001843 \end{pmatrix}$$

Resultat **Modell 2**: $y = 24.6 - 2.79x_1 + 0.18x_2 - 6.77x_3$,

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	24.6117	3.7937
1	–2.7875	0.1167
2	0.1818	0.0220
3	–6.7677	1.7425

VARIANSANALYS

	Frihetsgrader	Kvadratsumma
REGR	3	4172.3
RES	6	42.4
TOT	9	4214.7

och

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 2.035428 & -0.002778 & -0.010792 & -0.169747 \\ -0.002778 & 0.001925 & -0.000060 & 0.004068 \\ -0.010792 & -0.000060 & 0.000068 & -0.000653 \\ -0.169747 & 0.004068 & -0.000653 & 0.429429 \end{pmatrix}$$

a) Beräkna R^2 för de två modellerna och diskutera vilken modell som verkar vara bäst. (1p)

b) Är Modell 2 signifikant bättre än Modell 1? Genomför ett lämpligt test på nivå 5%. (1p)

c) Hur mycket energi förbrukar en genomsnittsindustri i Sverige som tillverkar enheter på 120 sekunder med en standard kvalitet? Bilda ett lämpligt 95% intervall. (2p)

4. Låt $X_1, ..., X_n$ vara ett stickprov på en kontinuerlig stokastisk variabel X med täthetsfunktionen

$$f_X(x) = \frac{2}{\theta^2}(\theta - x), \quad 0 \leq x \leq \theta,$$

där θ är en okänd parameter.

a) Beräkna momentskattningen av θ. (2p)

b) Bestäm variansen för momentskattningen av θ. Vad har skattningen för egenskaper? (2p)

5. Ett företag tillverkar gasol för industrin. Man överväger att byta från en äldre typ av behållare (typ 1) till en nyare sort (typ 2), som man hoppas är mer tålig för inre tryck. Man har gjort 16 mätningar av hållfastheten för inre tryck för var och en av behållarna och fått $\bar{x}_1 = 175.8$ respektive $\bar{x}_2 = 181.3$. De båda tillverkarna av behållarna har genom omfattade mätningar funnit att $\sigma_1 = 3.25$ och $\sigma_2 = 2.75$.

a) Pröva

$$H_0 : \mu_2 - \mu_1 = 4 \quad \text{mot} \quad H_1 : \mu_2 - \mu_1 > 4$$

på nivån 5%. Normalfördelning får förutsättas. (1.5p)

b) Beräkna styrkan för testet i a), då $\mu_2 - \mu_1 = 7$. (1.5p)

6. Lår X vara en diskret stokastisk variabel med sannolikhetsfunktionen

$$p_X(x) = (1-p)p^x, \quad x = 0, 1, 2, ...$$

Man testar hypotesen $H_0 : p = 0.9$ mot $H_1 : p > 0.9$ genom att förkasta H_0 för $X \geq 30$. Bestäm testets signifikansnivå samt dess styrka för $p = 0.99$. (2p)

Kurskod: TAMS65
Lösningar

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Lösningsförslag till tentamen onsdagen den 27 oktober kl 8–12.

1) H_0 : inget samband föreligger, mot H_1 : samband föreligger. Homogenitetstest ger

$$T = \sum_{j=1}^{3}\sum_{i=1}^{4} \frac{(N_{ij} - n_i\hat{p}_j)^2}{n_i\hat{p}_j} = 10.49.$$

Förkasta H_0 om $T > c = \chi^2_{0.95}((4-1)(3-1)) = \chi^2_{0.95}(6) = 12.60$. Alltså, vi kan inte förkasta H_0, och vi kan inte påvisa att samband mellan frågorna föreligger.

2a) Störst skillnad mellan s_1 och s_3, test hypotesen $H_0 : \sigma_1^2 = \sigma_3^2$, mot $H_1 : \sigma_1^2 \neq \sigma_3^2$ på nivån 5%. Teststorhet

$$v = \frac{s_1^2}{s_3^2} = 2.25$$

och förkasta H_0 om $v < a = F_{0.025}(4,4) < 1$ eller $v > b = F_{0.975}(4,4) = 9.60 > 1$. Eftersom $1 < v < b$ så kan vi inte förkasta H_0 och vi kan inte påvisa någon skillnad mellan σ_1^2 och σ_3^2.

b) Vi konstruerar ett nedåt begränsat intervall för $\mu_3 - 1.5\mu_1$.

$\hat{\mu}_3 - 1.5\hat{\mu}_1 = \bar{x}_3 - 1.5\bar{x}_1 = 11.2$

Vi har den sv. $\bar{X}_3 - 1.5\bar{X}_1 \sim N\left(\mu_3 - 1.5\mu_1, \sqrt{\frac{\sigma^2}{5} + 1.5^2\frac{\sigma^2}{5}}\right)$ och hjälpvariabeln

$$\frac{\bar{X}_3 - 1.5\bar{X}_1 - (\mu_3 - 1.5\mu_1)}{S\sqrt{\frac{3.25}{5}}} \sim t(12),$$

där $S^2 = \frac{4S_1^2 + 4S_2^2 + 4S_3^2}{12}$ och $s^2 = 23.98$.

$$P\left(\frac{\bar{X}_3 - 1.5\bar{X}_1 - (\mu_3 - 1.5\mu_1)}{S\sqrt{3.25/5}} < 1.78\right) = 0.95$$

ger

$$I_{\mu_3 - 1.5\mu_1} = \left(\bar{x}_3 - 1.5\bar{x}_1 - 1.78s\sqrt{3.25/5}\ ;\ \infty\right) = (4.2\ ;\ \infty).$$

Bara positiva värden. Alltså $\mu_3 > 1.5\mu_1$ med stor sannolikhet, vilket tyder på att fallen med malignt melanom ökat med minst 50%.

3a) $R_1^2 = 0.86$ och $R_2^2 = 0.99$. Modell två har ett betydligt bättre R^2-värde och verkar ha en bättring anpassning till datan.

b) $H_0 : \beta_2 = \beta_3 = 0$ mot $H_1 : \beta_2 \neq 0$ och/eller $\beta_3 \neq 0$, på nivån 5%. Teststorhet

$$v = \frac{(SS_E^{(1)} - SS_E^{(2)})/2}{SS_E^{(2)}/6} = 38.4$$

och förskasta H_0 om $v > c = F_{0.95}(2, 6) = 5.14$. Alltså, förkasta H_0, dvs. modell 2 är signifikant bättre än modell 1.

c) Låt $Y_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0$ där $\boldsymbol{u} = \begin{pmatrix} 1 & 0 & 120 & 1 \end{pmatrix}'$. Väntevärdet $\boldsymbol{u}'\boldsymbol{\beta}$ skattas med $\boldsymbol{u}'\widehat{\boldsymbol{\beta}} = 39.6613$. Konfidensintervallet ges av

$$I_{Y_0} = \left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}} \mp t_{0.975}(6)\, s \sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}} \right) = (35.76\ ;\ 43.56),$$

där $t_{0.975}(6) = 2.45$, $s = \sqrt{\dfrac{SS_{RES}}{6}} = 2.6591$ och $\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u} = 0.3578$.

4a) Väntevärdet beräknas som

$$\mathrm{E}(X) = \int_0^\theta x \frac{2}{\theta^2}(\theta - x)dx = \left[\frac{x^2}{\theta} - \frac{2x^3}{3\theta^2} \right]_0^\theta = \theta - \frac{2}{3}\theta = \frac{\theta}{3}$$

och moment-metoden ger $\mathrm{E}(X) = \bar{x} \quad \Rightarrow \quad \widehat{\theta} = 3\bar{x}$.

b) Variansen för $\widehat{\Theta}$ ges av

$$\mathrm{var}(\widehat{\Theta}) = \mathrm{var}(3\bar{X}) = 9\,\mathrm{var}(\bar{X}) = \frac{9}{n}\,\mathrm{var}(X).$$

Vidare har vi

$$\mathrm{E}(X^2) = \int_0^\theta x^2 \frac{2}{\theta^2}(\theta - x)dx = \left[\frac{2x^3}{3\theta} - \frac{x^4}{2\theta^2} \right]_0^\theta = \frac{2}{3}\theta^2 - \frac{\theta^2}{2} = \frac{\theta^2}{6}$$

med $\mathrm{var}(X) = \dfrac{\theta^2}{6} - \left(\dfrac{\theta}{3}\right)^2 = \dfrac{\theta^2}{18}$ och slutligen

$$\mathrm{var}(\widehat{\Theta}) = \frac{\theta^2}{2n}.$$

Eftersom skattningen är vvr och $\mathrm{var}(\widehat{\Theta}) \to 0$ när $n \to \infty$ så är det också en konsistent skattning av θ.

5a) Vi har observationer från $N(\mu_1, \sigma_1)$ och $N(\mu_2, \sigma_2)$, där σ_1 och σ_2 är kända.

Skattningen $\hat{\mu}_2 - \hat{\mu}_1 = \bar{x}_2 - \bar{x}_1$ är en observation från

$$\bar{X}_2 - \bar{X}_1 \sim N\left(\mu_2 - \mu_1, \sqrt{\frac{\sigma_2^2}{16} + \frac{\sigma_1^2}{16}}\right) = N(\mu_2 - \mu_1, 1.0643)$$

Teststorhet

$$u = \frac{\bar{x}_2 - \bar{x}_1 - 4}{1.0643} = 1.409$$

Den sv. $U \sim N(0,1)$ om H_0 är sann. H_0 förkastas om $u > 1.645$. Eftersom $1.409 < 1.645$ så kan vi inte förkasta H_0. Vi kan alltså inte bevisa att de nya behållarna är bättre.

b) Styrkan för $\mu_2 - \mu_1 = 7$ ges av

$$P\left(\frac{\bar{X}_2 - \bar{X}_1 - 4}{1.0643} > 1.645 \text{ då } \mu_2 - \mu_1 = 7\right) = P\left(\bar{X}_2 - \bar{X}_1 > 5.751 \text{ då } \mu_2 - \mu_1 = 7\right)$$

$$P\left(\frac{\bar{X}_2 - \bar{X}_1 - 7}{1.0643} > \frac{5.751 - 7}{1.0643} \text{ då } \mu_2 - \mu_1 = 7\right) = 1 - \Phi(-1.174) = \Phi(1.174) \approx 0.88.$$

6) Signifikansnivån ges av

$$P(\text{förkasta } H_0, \text{ då } H_0 \text{ sann}) = P(X \geq 30, \text{ då } p = 0.9) = \sum_{x=30}^{\infty} 0.1 \cdot 0.9^x = 0.9^{30} \approx 0.04$$

och styrkan för $p = 0.99$ ges av

$$h(0.99) = P(X \geq 30, \text{ då } p = 0.99) = 0.99^{30} \approx 0.74.$$