Kurskod: TAMS65
Provkod: TEN1

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Tentamen torsdagen den 15 mars 2018 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. Ur ett stickprov omfattande 9 observationer från $N(\mu, 1)$, har man beräknat stickprovsmedelvärdet $\bar{x} = 0.64$. Beräkna ett tvåsidigt 95% konfidensintervall för μ. (1p)

2. För att undersöka effekten av kalkning av försurade sjöar mättes pH i 8 sjöar, som var ungefär lika stora men hade olika grad av försurning. Mätningar gjordes dels före kalkning, dels efter kalkning:

Sjö	1	2	3	4	5	6	7	8
Före kalkning	5.2	5.8	4.3	5.2	4.6	4.7	5.8	5.5
Efter kalkning	5.6	6.3	4.9	5.8	5.5	5.7	6.1	5.4

   Undersök om kalkning gör skillnad, dvs. om det höjer pH-värdet i sjöarna. Motivera ditt svar med ett lämpligt 95% konfidensintervall. Normalfördelning kan antas. (3p)

3. Lönen är ofta kopplad till arbetslivserfarenhet. Låt y vara månadslönen för 50 civilingenjörer med arbetslivserfarenheten x år.

x	y	x	y	x	y
7	26075	21	43628	28	99139
⋮	⋮	⋮	⋮	⋮	⋮
18	49727	20	41721	11	38371
11	33233	26	82641		

   (a) Man analyserar först datamaterialet enligt modellen

   $$\textbf{Modell 1:} \quad Y = \gamma_0 + \gamma_1 x + \varepsilon'$$

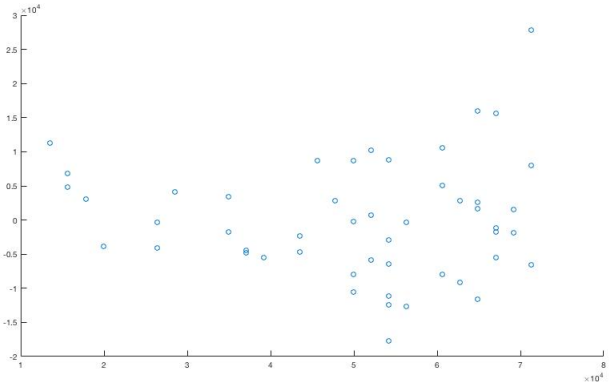
   och får residualplotten för residualerna mot de skattade väntevärdena. Förklara kortfattat vilken egenskap hos residualplotten som gör att man bör pröva att transformera datan. (0.5p)

   Resultat modell 1: $y = 11369 + 2141x$,

i	$\widehat{\gamma}_i$	$d(\widehat{\gamma}_i)$
0	11369	3160
1	2141.3	160.8

VARIANSANALYS

	Frihetsgrader	Kvadratsumma
REGR	1	$1.3239 \cdot 10^{10}$
RES	48	$3.5851 \cdot 10^{9}$
TOT	49	$1.6824 \cdot 10^{10}$



Figur 1: Residualplot modell 1

(b) Transformera datan enligt $z = \ln y$ och betrakta modellen

$$\textbf{Modell 2:} \quad Z = \beta_0 + \beta_1 x + \varepsilon$$

där ε-variablerna antas vara oberoende och $N(0, \sigma)$.
Resultat modell 2: $z = 9.84 + 0.05x$,

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	9.84133	0.05635
1	0.049978	0.002868

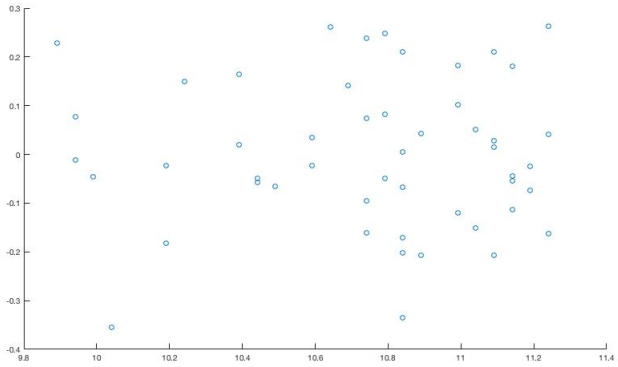
VARIANSANALYS

	Frihetsgrader	Kvadratsumma
REGR	1	7.2118
RES	48	1.1400
TOT	49	8.3519

och

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.13372 & -0.00628 \\ -0.00628 & 0.00035 \end{pmatrix}$$

- Hur skattas parametrarna β_0 och β_1? Visa att dessa skattningar är väntevärdesriktiga. (1p)
- Verkar arbetslivserfarenhet ha betydelse för lönen och i så fall på vilket sätt? Motivera ditt svar med ett lämpligt 95% konfidensintervall. (1p)
- Modell 2 innebär att lönen $Y = e^{\beta_0+\beta_1 x+\varepsilon}$ vilket ger att $\mathrm{E}(Y) \approx e^{\beta_0+\beta_1 x}$ om $\mathrm{var}(\varepsilon)$ är liten. Konstruera med hjälp av modell 2 ett 95% konfidensintervall för detta approximativa värde på $\mathrm{E}(Y)$ för en civilingenjör utan arbetslivserfarenhet. (1.5p)



Figur 2: Residualplot modell 2

4. Betrakta en stokastisk variabel med följande täthetsfunktion

$$f_X(x) = \frac{1}{x\sqrt{\pi}} e^{-(\ln x - \mu)^2}, \quad x > 0.$$

   (a) Härled maximum-likelihood-skattningen av μ baserat på ett stickprov om n oberoende observationer av X. (2p)

   (b) Bestäm täthetsfunktionen för $\ln X$. (1p)
   Ledning: Om $Y = g(X)$ så gäller att $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$.

   (c) Avgör om maximum-likelihood-skattningen av μ är väntevärdesriktig samt bestäm dess varians. (1p)

5. För att studera effekten av en viss typ av antirökpropaganda genomförs ett försök på följande sätt. Innan propagandan sätts in skattar man andelen rökare i den aktuella populationen till 24% med hjälp av ett stickprov omfattande 158 individer. Sedan propagandan genomförts konstaterar man för ett annat urval omfattande 212 individer att 44 är rökare.

   Skatta effekten av antirökpropagandan med ett 95% konfidensintervall. (3p)

6. En mindre industri har noterat en uppgång i antalet *större* beställningar och funderar på att eventuellt anställa mer personal. Under senaste månaden har man fått 32 beställningar.

   Låt antalet beställningar under t månader vara en stokastisk variabel $X \sim Po(\lambda t)$.

   Om $\lambda > 25$ anser man sig behöva mer personal.

   (a) Pröva på nivån 5% hypotesen $H_0 : \lambda = 25 \quad \text{mot} \quad H_1 : \lambda > 25$. (1.5p)

   (b) Under hur många månader framåt i tiden behöver man räkna beställningar om man vill genomföra hypotesprövningen i (a) och man vill att styrkan för testet ska vara minst 90% om $\lambda = 35$. (1.5p)

   Lämpliga approximationer får användas i både (a) och (b).

Kurskod: TAMS65
Provkod: TEN2

MATEMATISK STATISTIK I FORTSÄTTNINGSKURS

Lösningsförslag till tentamen torsdagen den 15 mars 2018 kl 8–12

1. (a) $I_\mu = \left(\bar{x} \mp z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = \left(\bar{x} \mp \frac{z_{0.975}}{3}\right) = \left(0.64 \mp \frac{1.96}{3}\right) = \underline{\underline{(-0.01\ ;\ 1.29)}}$

   (b) $|I_\mu| = 2 \cdot z_{0.975}\frac{\sigma}{\sqrt{n}} = \frac{3.92}{\sqrt{n}} < 1 \quad \Rightarrow \quad n > 3.92^2 = 15.37$, dvs välj $n \geq 16$.

2. Före kalkning: x_i är en observation från $X_i \sim N(\mu_i, \sigma_1)$, $i = 1, ..., 8$, samt
   Efter kalkning: y_i är en observation från $Y_i \sim N(\mu_i + \Delta, \sigma_2)$, $i = 1, ..., 8$.
   Bilda differenserna $Z_i = Y_i - X_i \sim N(\Delta, \sigma)$, $i = 1, ..., 8$. Vi vill testa hypotesen

   $$H_0 : \Delta = 0 \quad \text{mot} \quad H_1 : \Delta > 0, \quad \text{på nivån } 5\%,$$

   genom att bilda ett 95% konfidensintervall för Δ på formen $I_\Delta = (a\ ;\ \infty)$.
   $\widehat{\Delta} = \bar{z} = 0.525$ som är en observation från $\bar{Z} \sim N\left(\Delta, \frac{\sigma}{\sqrt{8}}\right)$, där vi skattar σ^2 med
   $s^2 = \frac{1}{8-1}\sum_{i=1}^{8}(z_i - \bar{z})^2 = 0.119$.
   Stäng in hjälpvariabeln $\frac{\bar{Z} - \Delta}{s/\sqrt{8}} \sim t(7)$ uppåt enligt $95\% = P\left(\frac{\bar{Z} - \Delta}{S/\sqrt{8}} < t_{0.95}(7)\right)$
   vilket ger intervallet

   $$I_\Delta = \left(\bar{z} - t_{0.95}(7)\frac{s}{\sqrt{8}}\ ;\ \infty\right) = \underline{\underline{(0.2945\ ;\ \infty)}}.$$

   Alltså, kalkningen höjer pH i sjöarna med stor sannolikhet.

3. (a) Vi har följande

   $$L(\mu) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{x_i\sqrt{\pi}} e^{-(\ln x_i - \mu)^2} = \pi^{-n/2}\left(\prod_{i=1}^{n}\frac{1}{x_i}\right) e^{-\sum_{i=1}^{n}(\ln x_i - \mu)^2}$$

   $$l(\mu) = \ln L(\mu) = -\frac{n}{2}\ln\pi + \ln\left(\prod_{i=1}^{n}\frac{1}{x_i}\right) - \sum_{i=1}^{n}(\ln x_i - \mu)^2$$

   $$l'(\mu) = 0 + 0 + \sum_{i=1}^{n} 2(\ln x_i - \mu) = 0 \quad \text{ger } \underline{\underline{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i}},$$

   vilket är maximum-likelihood-skattningen eftersom $l''(\hat{\mu}) = -2n < 0$.

(b) Vi har att $F_Y(y) = P(Y \le y) = P(\ln X \le y) = P(X \le e^y) = F_X(e^y)$ som ger

$$f_Y(y) = \frac{d}{dy}F_Y(y) = e^y f_X(e^y) = e^y \frac{1}{e^y\sqrt{\pi}} e^{-(\ln e^y - \mu)^2} = \frac{1}{\sqrt{\pi}} e^{-(y-\mu)^2}, \quad -\infty < y < \infty.$$

Vi ser direkt att detta är täthetsfunktionen för en normalfördelning med väntevärde μ och standardavvikelse $1/\sqrt{2}$, dsv. $\underline{\underline{Y = \ln X \sim N\left(\mu, \frac{1}{\sqrt{2}}\right)}}$.

(c) Från (b) har vi att $\mathrm{E}(\ln X_i) = \mu$ och $\mathrm{var}(\ln X_i) = \frac{1}{2}$. Skattningens väntevärde och varians blir

$$\mathrm{E}(\hat{\mu}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \ln X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{E}\left(\ln X_i\right)}_{=\mu} = \frac{1}{n} \cdot n \cdot \mu = \underline{\underline{\mu \quad \text{(väntevärdesriktig)}}}$$

$$\mathrm{var}(\hat{\mu}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} \ln X_i\right) = /\textit{ober.}/ = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{var}\left(\ln X_i\right)}_{=\frac{1}{2}} = \frac{1}{n^2} \cdot n \cdot \frac{1}{2} = \underline{\underline{\frac{1}{2n}}}.$$

(Man kan också lösa uppgiften direkt genom att beräkna väntevärde och varians för $\ln X$ från täthetsfunktionen $f_X(x)$.)

4. Låt p_1 = andelen rökare före propagandan och p_2 = andelen rökare efter propagandan. Vi har då att $\hat{p}_1 = 0.24$ och $\hat{p}_2 = \frac{44}{212} \approx 0.21$. Vidare är $x_1 = n_1\hat{p}_1 = 38$ en observation från $X_1 \sim Bin(n_1, p_1) \approx N(n_1p_1, \sqrt{n_1p_1(1-p_1)})$ eftersom $n_1\hat{p}_1(1-\hat{p}_1) = 28.8 > 10$, och på samma sätt $x_2 = 44$ en observation från $X_2 \sim Bin(n_2, p_2) \approx N(n_2p_2, \sqrt{n_2p_2(1-p_2)})$, där $n_1 = 158$ och $n_2 = 212$.

$\hat{p}_1 - \hat{p}_2 = 0.03$ är en observation från $\widehat{P}_1 - \widehat{P}_2 \approx N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$ och hjälpvariabeln blir

$$\frac{\widehat{P}_1 - \widehat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{\widehat{P}_1(1-\widehat{P}_1)}{n1} + \frac{\widehat{P}_2(1-\widehat{P}_2)}{n_2}}} \approx N(0,1)$$

vilket ger intervallet

$$I_{p_1-p_2} = \left(\hat{p}_1 - \hat{p}_2 \mp z_{0.975}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) = \underline{\underline{(-0.06\ ;\ 0.12)}},$$

där $z_{0.975} = 1.96$. Alltså, vi kan inte säga att antirökpropagandan gör någon skillnad.

5. (a) $x = 32$ är en observation av $X \sim Po(\lambda) \approx N(\lambda, \sqrt{\lambda})$ om $\lambda > 15$. Vi skattar λ med $\hat{\lambda} = x$.

Teststorhet: $x = 32$, och förkasta H_0 om $x > c$ (pga. H_1), dvs.

$$5\% = P(X > c \text{ om } H_0 \text{ är sann}) = P(X > c \text{ om } \lambda = 25) \approx 1 - \Phi\left(\frac{c-25}{\sqrt{25}}\right),$$

vilket ger $\quad 95\% = \Phi\left(\dfrac{c-25}{5}\right) \quad$ ger $\quad \dfrac{c-25}{5} = 1.645 \quad$ och $\quad c = 33.2.$

Men, $\underline{\underline{x = 32 < 33.2 = c, \; H_0 \text{ kan inte förkastas.}}}$

(b) Under de t närmaste månaderna får man y beställningar. Den sv. $Y \sim Po(\lambda t) \approx N(\lambda t, \sqrt{\lambda t})$

Teststorhet: y, och förkasta H_0 om $y > b$ (pga. H_1), dvs.

$$5\% = P(Y > b \text{ om } \lambda = 25) \approx 1 - \Phi\left(\frac{b-25t}{\sqrt{25t}}\right)$$

och $\quad 95\% = \Phi\left(\dfrac{b-25t}{\sqrt{25t}}\right) \quad$ med $\quad \dfrac{b-25t}{\sqrt{25t}} = 1.645. \quad (*)$

Styrkan ger också att

$$90\% = P(Y > b \text{ om } \lambda = 35) \approx 1 - \Phi\left(\frac{b-35t}{\sqrt{35t}}\right) = \Phi\left(-\frac{b-35t}{\sqrt{35t}}\right),$$

med $\quad -\dfrac{b-35t}{\sqrt{35t}} = 1.282. \quad (**)$

Ekvation $(*)$ och $(**)$ ger $\underline{\underline{t = 2.5, \text{ dvs. man ska mäta under minst 2.5 månader.}}}$