TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Matematisk statistik
EE

# TAMS 65 MATEMATISK STATISTIK I FORTSÄTTN.KURS
## Tentamen lördagen den 16 augusti 2008 kl 8-12.

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.

Obs! Skriv namn och personnummer på varje inlämnat papper.

1. Medieingenjörer har utvecklat en teknik för att korta ner sändningstider för TV-reklam vilket minskar kostnaderna för dem som annonserar. Det man undrar är om reklamen är lika effektiv efter sådan tidskompression. En grupp på 200 studenter delades någorlunda slumpmässigt in i tre grupper. Grupp ett (57 st) fick se normalversionen på 30 sekunder av reklamen, grupp två (74 st) fick en komprimerad version på 24 sekunder och grupp tre (69 st) fick se en komprimerad version på 20 sekunder. Två dagar senare frågade man studenterna om namnet på det märke som hade annonserats. Antalen som kom ihåg märket framgår av nedanstående tabell.

| | Kom ihåg märket | |
|---|---|---|
| Grupp | Ja | Nej |
| 1 | 15 | 42 |
| 2 | 32 | 42 |
| 3 | 10 | 59 |

Undersök med ett lämpligt $\chi^2$-test på nivån 5% om de tre grupperna har samma benägenhet att komma ihåg märket som reklamen handlade om. Ange både nollhypotes och mothypotes. Slutsatsen av testet ska framgå tydligt. (3p)

2. Töjbarheten hos provbitar av stål som tillverkats på två olika sätt, A och B har mätts:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4.45 | 4.54 | 3.87 | 5.76 | 4.26 | 4.65 | 4.23 | 5.03 | 4.85 | 5.45 | 4.54 | 3.95 |
| B | 4.02 | 3.45 | 4.21 | 4.65 | 4.32 | 3.76 | 3.54 | 4.08 | 4.31 | 3.87 | 4.07 | 3.59 |

a) Modell: Vi har oberoende observationer $x_1, \ldots, x_{12}$ från $N(\mu_1, \sigma_1)$ för A och $y_1, \ldots, y_{12}$ från $N(\mu_2, \sigma_2)$ för B.

Vidare är medelvärdena och stickprovsstandardavvikelserna $\bar{x} = 4.6317$ och $s_x = 0.5676$ respektive $\bar{y} = 3.9892$ och $s_y = 0.3608$.

a) Har man utgående från detta datamaterial någon anledning att tro att de båda standardavvikelserna $\sigma_1$ och $\sigma_2$ är olika? Genomför ett lämpligt test på nivån 0.10 och redovisa din slutsats. (1p)

b) Anta nu att $\sigma_1 = \sigma_2$. Redan innan man gjorde mätningarna misstänkte man att behandling B skulle ge mindre töjbarhet. Bekräftas denna hypotes? Konstruera ett lämpligt 95% konfidensintervall och redovisa din slutsats. (2p)

3. Data i tabellen nedan är lön, $y$, i dollar och arbetslivserfarenhet, $x$, mätt i år för ett sticpprov på 50 civilingenjörer.

| Years of Experience | Salary | Years of Experience | Salary | Years of Experience | Salary |
|---|---|---|---|---|---|
| $x$ | $y$ | $x$ | $y$ | $x$ | $y$ |
| 7 | $26,075 | 21 | $43,628 | 28 | $99,139 |
| 28 | 79,370 | 4 | 16,105 | 23 | 52,624 |
| 23 | 65,726 | 24 | 65,644 | 17 | 50,594 |
| 18 | 41,983 | 20 | 63,022 | 25 | 53,272 |
| 19 | 62,308 | 20 | 47,780 | 26 | 65,343 |
| 15 | 41,154 | 15 | 38,853 | 19 | 46,216 |
| 24 | 53,610 | 25 | 66,537 | 16 | 54,288 |
| 13 | 33,697 | 25 | 67,447 | 3 | 20,844 |
| 2 | 22,444 | 28 | 64,785 | 12 | 32,586 |
| 8 | 32,562 | 26 | 61,581 | 23 | 71,235 |
| 20 | 43,076 | 27 | 70,678 | 20 | 36,530 |
| 21 | 56,000 | 20 | 51,301 | 19 | 52,745 |
| 18 | 58,667 | 18 | 39,346 | 27 | 67,282 |
| 7 | 22,210 | 1 | 24,833 | 25 | 80,931 |
| 2 | 20,521 | 26 | 65,929 | 12 | 32,303 |
| 18 | 49,727 | 20 | 41,721 | 11 | 38,371 |
| 11 | 33,233 | 26 | 82,641 | | |

Datamaterialet har analyserats med hjälp av Minitab, se nedan.
a) I analys nr 1 har data analyserats enligt

Modell 1: $Y = \gamma_0 + \gamma_1 x + \varepsilon'$

och residualerna har plottats mot de skattade väntevärdena (fitted value). Förklara kortfattat vilken egenskap hos residualplotten som gör att man bör pröva att transformera data. (0.5p)

b) I analys nr 2 används modellen

Modell 2: $Z = \beta_0 + \beta_1 x + \varepsilon$ där $Z = \ln Y$

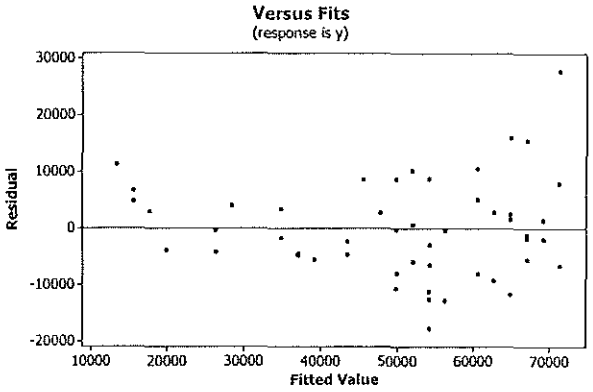
och där $\varepsilon$-variablerna antas vara oberoende och $N(0, \sigma)$.

b1) Verkar arbetslivserfarenheten ha betydelse för lönen och i så fall på vilket sätt? Motivera ditt svar med hjälp av ett lämpligt 95% konfidensintervall. (1p)

b2) Modell 2 innebär att lönen $Y = e^{\beta_0 + \beta_1 x + \varepsilon}$ vilket ger att $E(Y) \approx e^{\beta_0 + \beta_1 x}$ om $\mathrm{Var}(\varepsilon)$ är liten. Konstruera med hjälp av modell 2 ett 95% konfidensintervall för detta approximativa värde på $E(Y)$ för en ny civilingenjör utan arbetslivserfarenhet. (1p)

c) I den tredje analysen har modell 2 också använts men observation nr 19 med $x = 4$ har tagits bort, eftersom man misstänkt att den är en avvikande observation en så kallad outlier. Bekräftar analysen att observation nr 19 avviker? Motivera ditt svar med hjälp av lämplig information från analys nr 3. Det ska tydligt framgå vilken information du utnyttjar. (0.5p)

*Datorutskrift till uppgift 3 finns på nästa sida.*




ANALYS NR 1
Regression Analysis: y versus x

The regression equation is
y = 11369 + 2141 x

| Predictor | Coef | SE Coef |
|---|---|---|
| Constant | 11369 | 3160 |
| x | 2141.3 | 160.8 |

S = 8642.26  R-Sq = 78.7%

---

ANALYS NR 2
MTB > let c3=loge(c2)

Regression Analysis: lny versus x

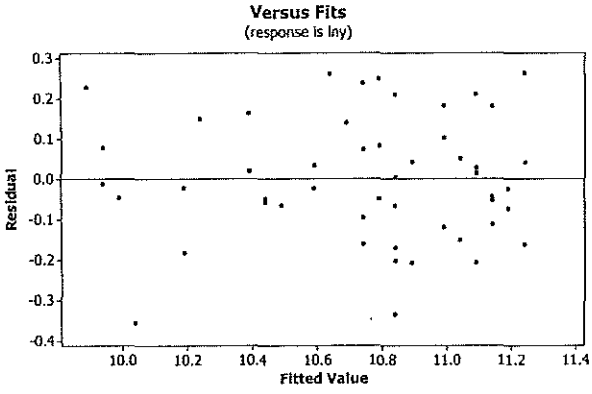
The regression equation is
lny = 9.84 + 0.0500 x

| Predictor | Coef | SE Coef (Stdev) |
|---|---|---|
| Constant | 9.84133 | 0.05635 |
| x | 0.049978 | 0.002868 |

S = 0.154113  R-Sq = 86.3%

Analysis of Variance

| Source | DF | SS |
|---|---|---|
| Regression | 1 | 7.2118 |
| Residual Error | 48 | 1.1400 |
| Total | 49 | 8.3519 |




---

ANALYS NR 3
MTB > copy c1 c3 c4 c5;
SUBC> omit 19.
Regression Analysis: lny3 versus x3

The regression equation is
lny3 = 9.88 + 0.0481 x3

| Predictor | Coef | SE Coef |
|---|---|---|
| Constant | 9.88358 | 0.05592 |
| x3 | 0.048076 | 0.002819 |

S = 0.146025  R-Sq = 86.1%

Analysis of Variance

| Source | DF | SS |
|---|---|---|
| Regression | 1 | 6.2031 |
| Residual Error | 47 | 1.0022 |
| Total | 48 | 7.2053 |

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|
| 1 | 10.0759 | 0.0457 | (9.9840, 10.1677) | (9.7681, 10.3837) |

Values of Predictors for New Observations

| New Obs | x3 |
|---|---|
| 1 | 4.00 |

---

4. a) I Sverige finns 450 småföretag inom en viss bransch. För att kartlägga försäljningen via internet skickar man ut en enkät till 150 slumpmässigt valda företag i den aktuella branschen. Den första frågan är : Har ni försäljning via internet? Bland de 150 företagen uppger 60 att de har försäljning via internet. Låt

$p =$ andelen företag inom branschen med försäljning via internet.

Konstruera ett tvåsidigt konfidensintervall för $p$ med approximativ konfidensgrad 95%. (2p)

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

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

5. I kvalitetsarbetet inom tillverkningsindustrin genomförs en mängd mätningar med vars hjälp man kan avgöra om olika kvalitetskrav är uppfyllda. För en viss kvalitetsvariabel samlar man under en dag in tjugofem oberoende mätvärden $x_1, \ldots, x_{25}$. Man har undersökt att det är rimligt att anta att de stokastiska variablerna $X_1, \ldots, X_{25}$ är oberoende och att $X_i \sim N(\mu, 1.2)$, där målvärdet är $\mu = 30$. Man prövar på nivån 0.05

$$H_0: \ \mu = 30 \qquad \text{mot} \qquad H_1: \ \mu > 30.$$

a) Man har fått $\bar{x} = 30.35$. Genomför hypotesprövningen. (1.5p)

b) För vilka $\mu$-värden är testets styrka minst 0.75? (1.5p)

6. a) Ofta då man studerar mätnoggrannhet gör man dubbelmätningar på föremål med olika egenskaper. Sedan bildar man för varje par av mätningar differensen mellan mätvärdena och får då observationer $y_i$ av oberoende stokastiska variabler $Y_i \sim N(0, \sigma\sqrt{2})$, där $\sigma$ är standardavvikelsen för ett mätfel. Låt $y_1, \ldots, y_n$ vara sådana observerade värden. Härled ML-skattningen av $\sigma$ baserad på $y_1, \ldots, y_n$ och undersök om motsvarande $\sigma^2$-skattning är väntevärdesriktig. (2p)

b) Antalet registrerade partiklar från ett radioaktivt prov beskrivs av en Poissonprocess med intensiteten $\lambda$. Detta innebär att antalet registrerade partiklar under t sekunder är $Po(\lambda t)$ och att registreringar under olika tidsintervall är oberoende. Mätningar genomfördes under 100 olika tidsintervall av längden 10 sekunder och det visade sig att för 52 tidsintervall saknades registreringar helt. Punktskatta $\lambda$ med hjälp av enbart denna information. (1p)

# Lösningar till tentamen i TAMS65 Matematisk statistik I, fk, 2008-08-16.

1. Vi gör ett homogenitetstest.

$H_0$: de tre grupperna har samma sannolikhet $p$ att komma ihåg märket

$H_1$: sannolikheterna att komma ihåg är inte lika stora.

Om $H_0$ är sann har vi $\hat{p} = \frac{57}{200} = 0.285$ och vi får de skattade förväntade frekvenserna $n_i \hat{p}$ respektive $n_i(1-\hat{p})$ enl. följande tabell:

| Grupp | Ja | Nej |
|---|---|---|
| 1 | 16.245 | 40.755 |
| 2 | 21.090 | 52.910 |
| 3 | 19.665 | 49.335 |

Teststorhet: $Q = \sum_i \sum_j \frac{(N_{ij} - n_i \hat{p}_{j\cdot})^2}{n_i \hat{p}_{j\cdot}} =$

$= \frac{(15-16.245)^2}{16.245} + \frac{(42-40.755)^2}{40.755} + \ldots + \frac{(59-49.335)^2}{49.335} =$

$= 14.67$

$H_0$ förkastas om $Q > c$. Den s.v. $Q$ är appr. $\chi^2((2-1)(3-1)) = \chi^2(2)$ om $H_0$ är sann. Tabell ger $c = 5.99$.

$14.67 > 5.99$. Alltså kan $H_0$ förkastas. De tre grupperna har med stor sannolikhet inte samma benägenhet att komma ihåg märket.

2a) $H_0: \sigma_1 = \sigma_2$ mot $H_1: \sigma_1 \neq \sigma_2$.

Teststorhet: $V = \frac{s_x^2}{s_y^2} = 2.475$

$H_0$ förkastas om $V < a$ eller $V > b$.

Den s.v. $V \sim F(11, 11)$ om $H_0$ är sann.

Tabell ger $b = \frac{2.98 + 2.91 + 2.75 + 2.69}{4} \simeq 2.83$

$P(V < a) = P(\frac{1}{V} > \frac{1}{a})$. Eftersom $\frac{1}{V} \sim F(11,11)$

får vi att $\frac{1}{a} = b$ dvs $a = \frac{1}{b} = 0.35$

$0.35 < 2.475 < 2.83$

$H_0$ kan inte förkastas. Vi har inte kunnat visa att $\sigma_1 \neq \sigma_2$.

b) Vi vill visa att $\mu_2 < \mu_1 \Leftrightarrow \mu_1 - \mu_2 > 0$ och gör därför ett nedåt begränsat konfidensintervall för $\mu_1 - \mu_2$.

$\hat{\mu}_1 - \hat{\mu}_2 = \bar{x} - \bar{y} = 0.6425$

Den s.v. $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma^2}{12} + \frac{\sigma^2}{12}})$

$\sigma^2$ skattas med $s^2 = \frac{11 s_x^2 + 11 s_y^2}{22} = 0.2262$;

$s = 0.4756$; fr. gr.: 22.

Hjälpvariabeln $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S/\sqrt{6}} \sim t(22)$ och ger

$I_{\mu_1 - \mu_2} = (\bar{x} - \bar{y} - t s/\sqrt{6}, \infty) = (0.6425 - 0.3339, \infty) =$

$= (0.309, \infty)$ där $t = 1.72$.

Alltså är $\mu_1 - \mu_2 > 0.309 > 0$.

Metod B ger med stor sannolikhet sämre genomsnittlig töjbarhet.

3a) För stora skattade väntevärden har vi större spridning hos residualerna. Då är det ofta bra att logaritmera Y-variabeln.

b1) $I_{\beta_1} = (\hat{\beta}_1 \mp t s \sqrt{h_{11}}) = (0.049978 \mp t \cdot 0.002868) =$

$= (0.049978 \mp 0.005770) \simeq (0.0442, 0.0557)$

$t = 2.012$ ges i $t(48)$-tabell

Bara positiva värden i $I_{\beta_1}$. Lönen ökar med ökad erfarenhet.

b2) En ny civilingenjör har $E(Y) \simeq e^{\beta_0}$.

$I_{\beta_0} = (\hat{\beta}_0 \mp t \cdot s\sqrt{h_{00}}) = (9.84133 \mp 2.012 \cdot 0.05635) =$

$= (9.84133 \mp 0.11338) = (9.72795,\ 9.95471)$

$I_{E(Y)} \simeq (e^{9.72795},\ e^{9.95471}) \simeq (16780,\ 21051)$

c) Obs. nr 19 har $y_0 = 16105$ och $\ln y_0 = 9.687$

Enligt datorutskriften har vi prediktionsintervallet $I_{\ln Y_0} = (9.7681,\ 10.3837)$

$9.687 \notin I_{\ln Y_0}$. Alltså är observation nr 19 avvikande. Lönen för den här personen är extremt låg.

4a) $x = 60$ är observation av $X \sim Hyp(N, n, p)$, där $N = 450$ och $n = 150$.

$\hat{p} = \frac{x}{n} = 0.4$

Den s.v. $X$ är appr $N\left(np, \sqrt{\frac{N-n}{N-1} np(1-p)}\right)$, eftersom $\frac{N-n}{N-1} n\hat{p}(1-\hat{p}) > 10$.

Då följer att den s.v. $\hat{P}$ appr $N\left(p, \sqrt{\frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}}\right)$.

Hjälpvariabeln $\dfrac{\hat{P} - p}{\sqrt{\frac{N-n}{N-1} \cdot \frac{\hat{P}(1-\hat{P})}{n}}}$ appr $N(0,1)$

och den ger $I_p = \left(\hat{p} \mp 1.96\sqrt{\frac{N-n}{N-1} \cdot \frac{\hat{p}(1-\hat{p})}{n}}\right) =$

$= (0.4 \mp 0.064) = (0.336,\ 0.464)$

Den sanna andelen företag med internetförsäljning ligger med stor slh mellan 33.6 % och 46.4 %.

b) Den stok. vektorn $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\right)$

eftersom komponenterna är två oberoende normalvariabler.

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = A\,\mathbb{X}$$

Då är $\mathbb{U}$ normalfördelad eftersom den är en linjär transformation av en normalfördelad vektor. Parametrar

$$E(\mathbb{U}) = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \end{pmatrix}$$

$$C_{\mathbb{U}} = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -4 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 8 & -2 \\ -2 & 5 \end{pmatrix}$$

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 8 \end{pmatrix}, \begin{pmatrix} 8 & -2 \\ -2 & 5 \end{pmatrix}\right).$$

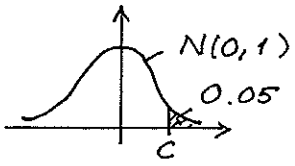
5a) $\hat{\mu} = \bar{x}$; den s.v. $\bar{X} \sim N(\mu, \frac{1.2}{\sqrt{25}}) = N(\mu, 0.24)$

Teststorhet: $u = \dfrac{\bar{x} - 30}{0.24} = 1.458$

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

$H_0$ förkastas om $u > c$.

$\Phi(c) = 0.95$. Tabell ger $c = 1.645$.




$1.458 < 1.645$; $H_0$ kan inte förkastas. Vi kan inte påvisa avvikelse från målvärdet.

b) Styrkefunktionen

$h(\mu) = P(H_0$ förkastas om $\mu$ är det sanna värdet$) =$

$= P\left(\dfrac{\bar{X} - 30}{0.24} > 1.645 \text{ om } \mu \text{ sanna värdet}\right) =$

$= P(\bar{X} > 30.395 \text{ om } \mu \text{ sanna värdet}) =$ ↖ $\bar{X} \sim N(\mu, 0.24)$

$= 1 - \Phi\left(\dfrac{30.395 - \mu}{0.24}\right) = \Phi\left(\dfrac{\mu - 30.395}{0.24}\right)$

$h(\mu) \geq 0.75$ om $\dfrac{\mu - 30.395}{0.24} \geq 0.6745$ dvs $\mu \geq 30.557$

6a) $L(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2}\sqrt{2\pi}} e^{-y_i^2/4\sigma^2} =$

$= \text{konst} \cdot \sigma^{-n} \cdot e^{-\frac{1}{4\sigma^2}\sum_1^n y_i^2}$

$l(\sigma) = \ln L(\sigma) = \ln \text{konst} - n \ln \sigma - \frac{1}{4\sigma^2} \sum_1^n y_i^2$

$l'(\sigma) = -\frac{n}{\sigma} + \frac{1}{2\sigma^3} \sum_1^n y_i^2 = \frac{\sum_1^n y_i^2 - 2n\sigma^2}{2\sigma^3}$

$l'(\sigma) = 0$ för $\hat{\sigma}^2 = \frac{1}{2n} \sum_1^n y_i^2$

| $\sigma$ | 0 | $\hat{\sigma}^2$ |
|---|---|---|
| $l'(\sigma)$ | + | 0 − |
| $l(\sigma)$ | ↗ | ↘ |

Alltså maximum. ML-skattn. $\hat{\sigma} = \sqrt{\frac{1}{2n} \sum_1^n y_i^2}$.

$E\left(\frac{1}{2n} \sum_1^n Y_i^2\right) = \frac{1}{2n} \sum_1^n E(Y_i^2) = \frac{1}{2n} \cdot n E(Y_i^2) =$

$\overset{①}{=} \frac{1}{2} Var(Y_i) = \frac{1}{2} \cdot 2\sigma^2 = \sigma^2$ ; ① $E(Y_i) = 0$.

Alltså är $\hat{\sigma}^2$ väntevärdesriktig.

b) Låt $X$ vara antalet registreringar under ett 10s-intervall.

Då gäller att $X \sim Po(10\lambda)$ och $P(X=0) = e^{-10\lambda}$.

$z = 52$ är observation av $Z \sim Bin(100, p)$

där $p = e^{-10\lambda}$.

$\hat{p} = 0.52$ ger $e^{-10\hat{\lambda}} = 0.52$ dvs

$-10\hat{\lambda} = \ln 0.52 \Leftrightarrow \hat{\lambda} = 0.0654$