TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Matematisk statistik
EE

TAMS 65 MATEMATISK STATISTIK I FORTSÄTTN.KURS
Tentamen måndagen den 2 juni 2008 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.

Obs! Skriv namn och personnummer på varje inlämnat papper.

1. Då man studerar den första siffran skild från 0 i olika sorters data har man funnit att siffrorna $1, 2, \ldots, 9$ inte är lika vanliga. I stället gäller Benfords lag:

$$P(X = k) =^{10} \log(k+1) -^{10} \log(k)$$

för $k = 1, \ldots, 9$ vilket ger sannolikhetsfunktionen

k:	1	2	3	4	5	6	7	8	9
$P(X=k)$:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

med avrundade siffror.

Detta kan användas för att t.ex. upptäcka bokföringsbrott. Förfalskade poster har troligen en annan sannolikhetsfördelning för den första siffran. Man har hämtat 355 poster från en budget för ett universitet och fått följande observerade frekvenser N_i för den första siffran skild från 0 för de olika posterna:

i:	1	2	3	4	5	6	7	8	9
N_i:	111	60	46	29	26	22	21	20	20

a) Betrakta budgetsiffrorna som slumpmässigt valda och pröva på nivån ungefär 0.05 hypotesen

$$H_0: \text{Universitetets budgetsiffror följer Benfords lag}$$

med hjälp av ett lämpligt test. (2p)

b) Konstruera utgående från det givna datamaterialet ett tvåsidigt konfidensintervall för $p_1 = P(X = 1)$ med approximativ konfidensgrad 95%. (1p)

2. Ett mobilt nätverk består av flera datorer (noder) som rör sig inom en nätverksarea. Ofta skickas meddelanden mellan två noder. Om den mottagande noden inte kan nås skickas meddelandet till en närliggande nod som sänder det vidare mot destinationen. Andelen meddelanden Y som går fram kallas "goodput" och den påverkas av genomsnittlig hastighet för noderna, x_1, och paustiden vid varje destination, x_2. Samhörande värden från en simuleringsstudie (från *Proceedings of the 2002 International Conference on Wireless Networks*):

Speed (m/s)	Pause Time (s)	Goodput (%)	Speed (m/s)	Pause Time (s)	Goodput (%)
5	10	95.111	20	40	87.800
5	20	94.577	20	50	89.941
5	30	94.734	30	10	62.963
5	40	94.317	30	20	76.126
5	50	94.644	30	30	84.855
10	10	90.800	30	40	87.694
10	20	90.183	30	50	90.556
10	30	91.341	40	10	55.298
10	40	91.321	40	20	78.262
10	50	92.104	40	30	84.624
20	10	72.422	40	40	87.078
20	20	82.089	40	50	90.101
20	30	84.937			

Data har först analyserats enligt

Modell 1: $Y = \beta_0' + \beta_1' x_1 + \beta_2' x_2 + \varepsilon'$

där $\varepsilon' \sim N(0, \sigma')$, se nedan. Efter att ha studerat residualplottarna som inte är alldeles lättolkade har en analys också gjorts enligt

Modell 2: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon$

där $\varepsilon \sim N(0, \sigma)$.

a) Pröva på nivån 0.01 för modell 1

H_0 : $\beta_1' = \beta_2' = 0$

mot

H_1 : Minst en av β_1' och β_2' är skild från 0

med hjälp av ett lämpligt test. (1p)

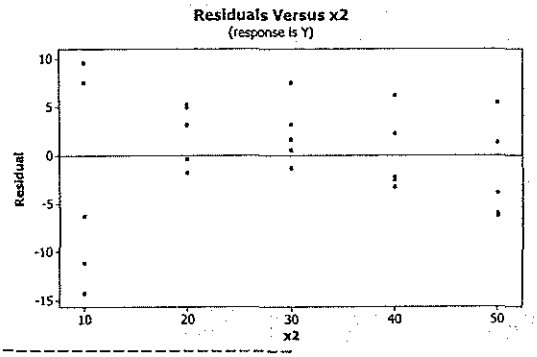
b) Hur beräknas residualerna för modell 1? Ange en formel. (0.5p)

c) Beskriver modell 2 data bättre än modell 1? Genomför ett lämpligt test på nivån 0.01. Både nollhypotes och mothypotes samt slutsatsen av testet ska redovisas. (1.5p)

Datorutskrift till uppgift 2.

```
---------------------------------------------------------------------
ANALYS NR 1
MTB > Regress 'Y' 2 'x1' 'x2'

Regression Analysis: Y versus x1, x2

The regression equation is
Y = 84.0 - 0.457 x1 + 0.377 x2
```

```
Predictor       Coef  SE Coef(Stdev)
Constant      84.039          3.534
x1          -0.45681        0.09644
x2           0.37695        0.08733
S = 6.17514   R-Sq = 65.1%   R-Sq(adj) = 61.9%

Analysis of Variance

Source           DF       SS      MS
Regression        2  1566.03  783.02
Residual Error   22   838.91   38.13
Total            24  2404.95
```




```
ANALYS NR 2
MTB > Regress 'Y' 5 'x1' 'x2' 'x1^2'-'x2^2'

Regression Analysis: Y versus x1, x2, x1^2, x1x2, x2^2

The regression equation is
Y = 96.0 - 1.82 x1 + 0.565 x2 + 0.0140 x1^2 + 0.0247 x1x2 - 0.0118 x2^2
Predictor       Coef   SE Coef(Stdev)
Constant      96.024          3.946
x1           -1.8245         0.2376
x2            0.5652         0.2256
x1^2        0.014020       0.004745
x1x2        0.024731       0.003249
x2^2       -0.011793       0.003516
S = 2.94204   R-Sq = 93.2%   R-Sq(adj) = 91.4%

Analysis of Variance

Source           DF       SS      MS
Regression        5  2240.49  448.10
Residual Error   19   164.46    8.66
Total            24  2404.95
```

3. a) Låt $x_1, \ldots, x_n$ vara observationer av oberoende stokastiska variabler $X_1, \ldots, X_n$ sådana att $X_i \sim Po(\lambda t_i)$. Härled ML-skattningen av λ och undersök om den är väntevärdesriktig. (2p)

b) Låt $y_1, \ldots, y_n$ vara observationer av oberoende stokastiska variabler $Y_1, \ldots, Y_n$ sådana att $Y_i \sim Re(\theta, \theta+1)$. Bestäm momentskattningen av θ. (1p)

4. För sjutton olika kärnkraftverk har man under ett år noterat antalet säkerhetsrelaterade fel som uppstått då reaktorn inte varit i drift (motsvarande värden för reaktorer i drift finns också). Reaktorerna är av två typer BWR = boiling water reactor och PWR = pressurized water reactor. Data:

BWR:	1	5	4	7	10	8	7	10	4
PWR:	8	7	3	0	1	2	3	6	

Modell Vi har två oberoende stickprov från $Po(\mu_1)$ respektive $Po(\mu_2)$.

Finns det skillnad mellan de båda reaktortyperna i fråga om felintensiteten under ett år? Konstruera ett lämpligt konfidensintervall med konfidensgraden approximativt 95% och redovisa din slutsats. (3p)

5. Man har noterat följande åldersjusterade årsvärden för antalet fall av malignt melanom per miljon 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 $N(\mu_i, \sigma)$, $i = 1, 2, 3$. Vi betraktar också observationerna inom varje stickprov som oberoende.

a) Konstruera ett tvåsidigt konfidensintervall för σ med konfidensgraden 95%. (1p)

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

6. Ett företag tillverkar kolsyrade drycker. Man överväger att byta från en äldre typ av glasflaskor (typ 1) till en nyare sort (typ 2), som man hoppas är mera tåliga för inre tryck. Man har gjort 16 bestämningar av hållfastheten för inre tryck för var och en av flasktyperna och fått $\bar{x}_1 = 175.8$ respektive $\bar{x}_2 = 181.3$. De båda flasktillverkarna har har genom omfattande 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 0.05. Normalfördelning får förutsättas. (1.5p)

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

Lösningar till tentamen i TAMS65, Matematisk statistik I fk, 2008-06-02.

1a) Om H_0 är sann är de förväntade frekvenserna $np_i = 355\,p_i$

i	1	2	3	4	5
np_i	106.86	62.48	44.38	34.44	28.05

6	7	8	9
23.79	20.59	18.11	16.33

Teststorhet: $Q = \dfrac{\sum (N_i - np_i)^2}{np_i} = 2.49$

H_0 förkastas om $Q > c$. Den s.v. Q är appr $\chi^2(8)$ om H_0 är sann. Tabell ger $c = 15.51$.
$Q = 2.49 < 15.51$. H_0 kan inte förkastas. Resultatet tyder på god anpassning till H_0.

b) $x = 111$ är observation av $X \sim Bin(n, p_1)$ där $n = 355$.
Vi har $\hat{p}_1 = \frac{x}{n}$ och den s.v. $\hat{P}_1$ är appr $N(p_1, \sqrt{p_1(1-p_1)/n})$ då $n\hat{p}_1(1-\hat{p}_1) > 10$.

Hjälpvariabeln $\dfrac{\hat{P}_1 - p_1}{\sqrt{\hat{P}_1(1-\hat{P}_1)/n}}$ är appr $N(0,1)$

och den ger

$$I_{p_1} = (\hat{p}_1 \mp 1.96\sqrt{\hat{p}_1(1-\hat{p}_1)/355}) = (0.3127 \mp 0.0482) =$$
$$= (0.2645, 0.3609) \simeq (0.26, 0.36)$$

Vi ser att $0.301 \in I_{p_1}$.

Anm. Här tänker vi oss ett slumpmässigt urval ur en stor population med siffror.

2a) Vi gör det övergripande F-testet.

Teststorhet: $V = \dfrac{Q_{REGR}/2}{Q_{RES}/22} = 20.53$

H_0 förkastas om $V > c$. Den s.v. $V \sim F(2,22)$ om H_0 är sann. Tabell ger $c \approx 5.76$.

$20.53 > 5.76$. H_0 förkastas. Minst en av x_1 och x_2 gör nytta som förklaringsvariabel.

b) Residualen $e_j = y_j - \hat{\beta}_0' - \hat{\beta}_1' x_{j1} - \hat{\beta}_2' x_{j2}$.

c) Vi prövar

$\tilde{H}_0$: $\beta_{11} = \beta_{12} = \beta_{22} = 0$

mot

$\tilde{H}_1$: minst en av β_{11}, β_{12} och β_{22} är $\neq 0$

Teststorhet: $W = \dfrac{(Q_{RES}^{(1)} - Q_{RES}^{(2)})/3}{Q_{RES}^{(2)}/19} = 25.97$

$\tilde{H}_0$ förkastas om $W > \tilde{c}$. Den s.v. $W \sim F(3,19)$ om $\tilde{H}_0$ är sann. Tabell ger $\tilde{c} \approx 5.015$.

$25.97 > 5.015$; $\tilde{H}_0$ förkastas. Minst en av de tre nya förklaringsvariablerna gör nytta.

3a) $L(\lambda) = \left[\dfrac{(\lambda t_1)^{x_1}}{x_1!} e^{-\lambda t_1}\right] \cdot \ldots \cdot \left[\dfrac{(\lambda t_n)^{x_n}}{x_n!} e^{-\lambda t_n}\right] =$

$= \left\{\prod_{i=1}^{n} \dfrac{t_i^{x_i}}{x_i!}\right\} \cdot \lambda^{\sum_1^n x_i} \cdot e^{-\lambda \sum_1^n t_i}$

$\ell(\lambda) = \ln L(\lambda) = \text{konst} + \sum_1^n x_i \ln \lambda - \lambda \sum_1^n t_i$

$\ell'(\lambda) = \sum_1^n x_i \cdot \dfrac{1}{\lambda} - \sum_1^n t_i = \dfrac{\sum_1^n x_i - \lambda \sum_1^n t_i}{\lambda}$

$\ell'(\lambda) = 0$ för $\hat{\lambda} = \dfrac{\sum_1^n x_i}{\sum_1^n t_i}$

λ	0	$\hat{\lambda}$
$\ell'(\lambda)$	$+$	0 $-$
$L(\lambda)$	↗	↘

Alltså maximum, så $\hat{\lambda}$ är ML-skattningen.

Vvr? $E(\hat{\Lambda}) = E\left(\frac{\sum_1^n X_i}{\sum_1^n t_i}\right) = \frac{1}{\sum_1^n t_i} E\left(\sum_1^n X_i\right) =$

$= \frac{1}{\sum_1^n t_i} \sum_1^n E(X_i) = \frac{1}{\sum_1^n t_i} \sum_1^n \lambda t_i = \lambda$

Alltså är $\hat{\Lambda}$ väntevärdesriktig.

b) $E(X_i) \overset{①}{=} \frac{\theta + (\theta+1)}{2} = \theta + \frac{1}{2}$ ① se F-S.

Ekv. $\bar{y} = \hat{\theta} + \frac{1}{2}$ ger $\hat{\theta} = \bar{y} - \frac{1}{2}$ som är momentskattningen.

4) $x_1, \ldots, x_9$ är observationer från $Po(\mu_1)$

$y_1, \ldots, y_8$ -" - -" - -" - $Po(\mu_2)$.

Vi söker ett konfidensintervall för $\mu_1 - \mu_2$.

$\hat{\mu}_1 = \bar{x} = 6.222$; $\hat{\mu}_2 = \bar{y} = 3.750$.

$\sum_1^9 X_i$ är $Po(9\mu_1)$ och $\sum_1^8 Y_j$ är $Po(8\mu_1)$. Eftersom $9\hat{\mu}_1 > 15$ och $8\hat{\mu}_2 > 15$ är normalappr. tillåten. Då följer att $\bar{X} - \bar{Y}$ är appr. normalfördelad. Parametrar

$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$

$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\mu_1}{9} + \frac{\mu_2}{8}$

Hjälpvariabeln $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\bar{X}}{9} + \frac{\bar{Y}}{8}}}$ appr $N(0,1)$.

Instängning ger

$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} \mp 1.96\sqrt{\frac{\bar{x}}{9} + \frac{\bar{y}}{8}}\right) = (0.36, 4.58)$

Bara positiva värden. Felintensiteten verkar vara större för BWR.

5a) σ^2- skattning $s^2 = \frac{4s_1^2 + 4s_2^2 + 4s_3^2}{12} =$

$= 23.98$; $s = 4.90$; frihetsgrad: 12.

Den s.v. $\frac{12S^2}{\sigma^2} \sim \chi^2(12)$

$P(4.40 \leq \frac{12S^2}{\sigma^2} \leq 23.35) = 0.95$

$P(\frac{12S^2}{23.35} \leq \sigma^2 \leq \frac{12S^2}{4.40}) = 0.95$

Ger $I_\sigma = (3.51, 8.09)$.

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

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

Den s.v. $\bar{X}_3 - 1.5\bar{X}_1 \sim N(\mu_3 - 1.5\mu_1, \sqrt{\frac{\sigma^2}{5} + 1.5^2 \cdot \frac{\sigma^2}{5}})$

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

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

Ger

$I_{\mu_3 - 1.5\mu_1} = (\bar{X}_3 - 1.5\bar{X}_1 - 1.78 s \cdot \sqrt{\frac{3.25}{5}}, \infty) \simeq$

$\simeq (11.2 - 7.0, \infty) = (4.2, \infty)$

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%.

6a) Vi har observationer från $N(\mu_1, \sigma_1)$ och $N(\mu_2, \sigma_2)$, där σ_1 och σ_2 är kända.

$\hat{\mu}_2 - \hat{\mu}_1 = \bar{X}_2 - \bar{X}_1$

Den s.v. $\bar{X}_2 - \bar{X}_1 \sim N(\mu_2 - \mu_1, \sqrt{\frac{\sigma_2^2}{16} + \frac{\sigma_1^2}{16}}) =$

$= N(\mu_2 - \mu_1,\ 1.0643)$

Teststorhet: $u = \dfrac{\bar{x}_2 - \bar{x}_1 - 4}{1.0643} = 1,409$

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

H_0 förkastas om $u > 1.645$.

$1.409 < 1.645$; H_0 kan inte förkastas. Vi kan inte bevisa att de nya flaskorna är bättre.

b) Styrkan då $\mu_2 - \mu_1 = 7$:

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

$$= P(\bar{X}_2 - \bar{X}_1 > 5.751 \text{ om } \mu_2 - \mu_1 = 7) =$$

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

$$= 1 - \Phi(-1.174) = \Phi(1.174) \simeq 0.88$$