NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET
INSTITUTT FOR MATEMATISKE FAG

Faglig kontakt under eksamen:
Oddgeir Samset, telefon 99 54 27 79

EKSAMEN I EMNE SIF5068 INDUSTRIELL STATISTIKK

Mandag 2. desember 2002
Tid: 09.00 – 14.00

Hjelpemidler: B - Alle trykte og håndskrevne hjelpemidler tillatt. Bestemt, enkel kalkulator tillatt.

Sensuren faller 6. januar 2003

Oppgave 1

Koaguleringstid (Y_{ij}) i sekund for blod fra dyr med tre forskjellige dietter A, B og C skal sammenlignes. I forsøket er det valgt ut 16 dyr, og resultatene er gitt under. Vi antar Y_{ij} er uavhengige, normalfordelte med samme ukjente varians σ^2 og at $E(Y_{ij}) = \mu_i$ for $i = 1,\ldots,3$.

```
Row   A   B   C    Analysis of Variance
  1  62  63  68
  2  60  67  66    Source     DF        SS        MS        F        P
  3  63  71  71    Factor      2    120.00     60.00    12.19    0.001
  4  59  64  67    Error      13     64.00      4.92
  5      65  68    Total      15    184.00
  6      66  68
                                            Individual 95% CIs For Mean
                                            Based on Pooled StDev
                   Level  N    Mean   StDev ---+---------+---------+---------+---
                   A      4    61.0   1.826 (-------*-------)
                   B      6    66.0   2.828                (------*-----)
                   C      6    68.0   1.673                      (-----*------)
                                            ---+---------+---------+---------+---
                                              59.5      63.0      66.5      70.0
                   Pooled StDev =     2.219
```

a)

Hva slags forsøk er dette, og hvordan bør randomiseringen være gjort her? Finn en estimator for σ^2 for dette tilfellet, og bruk utskriften over til å finne estimatet. Hvordan vil du teste om gruppene har forskjellig forventningsverdier? Skriv opp uttrykkene og forklar hvordan du kan finne p-verdien. Hva blir konklusjonen på testen dersom du bruker 5% nivå?

b)

Finn et 95% konfidensintervall for μ_1 til diett A. Finn deretter et 95% prediksjonsintervall for en måling av koaguleringstiden til et nytt dyr som har diett A.

Oppgave 2

For å undersøke 5 faktorer A, B, C, D og E ble det valgt å bruke et 2^{5-2} fraksjonelt faktorielt design.

Row	A	B	C	D	E	y
1	-1	-1	-1	1	-1	0.12
2	1	-1	-1	-1	-1	2.95
3	-1	1	-1	1	1	4.92
4	1	1	-1	-1	1	4.46
5	-1	-1	1	-1	1	-4.11
6	1	-1	1	1	1	-3.91
7	-1	1	1	-1	-1	-1.00
8	1	1	1	1	-1	-1.82

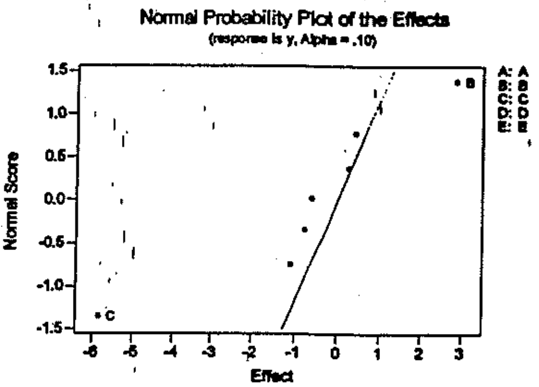
a)

Generatoren til designet er I = ACD = -BCE. Hva forteller generatoren oss om hvordan designet ble generert? Finn definerende relasjon til designet, og skriv opp alias-strukturen når man kan anta at alle tre-faktor og høyere ordens samspill kan neglisjeres. Hvilken resolusjon har designet?

Forklar hvorfor randomisering er viktig i gjennomføring av forsøk, og hvordan bør denne være gjort her?

```
Estimated Effects and Coefficients for y
(coded units)

Term          Effect      Coef
Constant                 0.201
A              0.440     0.220
B              2.873     1.437
C             -5.824    -2.912
D             -0.744    -0.372
E              0.274     0.137
A*B           -1.080    -0.540
A*E           -0.567    -0.283
```



```
Analysis of Variance for y (coded units)

Source              DF    Seq SS    Adj SS    Adj MS      F       P
Main Effects         5    85.987    85.987    17.197      *       *
2-Way Interactions   2     2.973     2.973     1.487      *       *
Residual Error       0     0.000     0.000     0.000
Total                7    88.960
```

b)

Skriv opp estimatorene for kontrastene B og AB, og gi estimatene for disse.

Anta at observasjonene er uavhengige og identisk normalfordelte med konstant varians σ^2. Hva er variansen til estimatorene for B og AB over? Hvordan kan du finne et estimat for denne variansen dersom man kan anta at alle to-faktor og høyere ordens samspill kan neglisjeres?

Dersom vi ikke kan anta at to-faktor samspillene kan neglisjeres, hvilke nye forsøk ville du da argumentere for å gjennomføre? Forklar.

c)

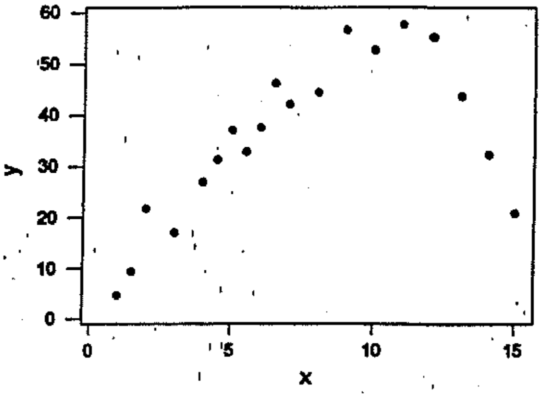
Ved hjelp av et 2^6 faktorielt design kan vi undersøke 6 faktorer A, B, C, D, E og F. Et slikt design krever imidlertid mange enkeltforsøk. I stedet skal du lage ¼ fraksjon av et 2^6 design, dvs et 2^{6-2} design med resolusjon IV.

Én av de definerende relasjonene i dette designet skal være I=ABDE. Hva blir da de andre? Er det mulig å kjøre dette forsøket i 2 eller 4 blokker slik at alle hovedeffektene og to-faktor samspillene er ukonfunderte med blokk-effektene? Forklar.

Oppgave 3

Prosentandelen (x) av en tilsetning til en bestemt metallegering er blitt registrert for å undersøke nærmere betydning av tilsetningen på hardheten (Y) til metallegeringen. Dataene som ble samlet inn er gitt under.

```
Row       y      x
  1    4.71    1.0
  2    9.34    1.5
  3   21.73    2.0
  4   17.03    3.0
  5   26.87    4.0
  6   31.27    4.5
  7   37.10    5.0
  8   32.92    5.5
  9   37.67    6.0
 10   46.38    6.5
 11   42.22    7.0
 12   44.55    8.0
 13   56.55    9.0
 14   52.73   10.0
 15   57.72   11.0
 16   55.23   12.0
 17   43.86   13.0
 18   32.52   14.0
 19   20.82   15.0
```



a)

Plott av dataene over og kjennskap til prosessen indikerer at vi bør prøve en regresjonsmodell av typen $E(Y)=\beta_0+\beta_1x+\beta_2x^2$. Resultat fra regresjonsanalysen kjørt i MINITAB er gitt under. Vi har her benyttet symbolene `'x^2'` for x^2 og `'XPXInverse'` for matrisen $(X'X)^{-1}$.

Angi hvilke antagelser som ligger til grunn for regresjonsanalysen. Forklar hvordan disse antagelsene kan sjekkes og hvordan vi kan oppdage at noen av antagelsene ikke holder.

Lokaliser de 5 verdiene som har blitt merket med #. Finn de verdiene som egentlig skal stå der, og la det komme klart frem hvordan du finner de nevnte verdiene.
I utskriften under er det gitt fire p-verdier. Hva er det som testes, hvordan er p-verdiene fremkommet, og hva slags informasjon gir de?

```
The regression equation is
y = - 10.2 + 12.9 x - 0.684 x^2

Predictor        Coef    SE Coef        T       P
Constant      -10.202      4.510    -2.26   0.038
x              12.881      1.330        #   0.000
x^2          -0.68357    0.08197    -8.34   0.000

S = #            R-Sq = #          R-Sq(adj) = 86.0%

Analysis of Variance

Source           DF        SS        MS        F       P
Regression        2    3881.6    1940.8    56.43   0.000
Residual Error   16     550.2      34.4
Total            18    4431.9

Source       DF     Seq SS
x             1     1490.1
x^2           1          #

Matrix XPXInverse
 0.591508 -0.157583  0.008620
-0.157583  #        -0.003077
 0.008620 -0.003077  0.000195
```

b)

Bruk den estimerte modellen til å predikere hardheten av metallegeringen når prosentandelen av tilsetningen er 7.0.

Vi ønsker å optimalisere hardheten. Finn en estimator for den prosentandelen av tilsetning som gir størst hardhet i metallegeringen, og finn et tilnærmet uttrykk for variansen til denne estimatoren. Bestem numerisk verdi for estimatoren ved å benytte utskriften fra regresjonsanalysen i a).

Hvorfor er det viktig å alltid etterprøve resultatene fra en slik analyse?

c)

Anta at vi i stedet hadde prøvd modellen $E(Y) = \beta_0 + \beta_1 x$. Bruk resultatene fra utskriften fra regresjonsanalysen i a) til å teste om β_1 er signifikant forskjellig fra 0 på 1% nivå i denne modellen.

Eksamen SIF5068 Mandag 2. desember 2002

Løsningsforslag

Oppgave 1:

a) Enveisgruppering

Alle 16 dyrene bør tilordnes diett tilfeldig (gitt n_i), blodprøvene tas og analyseres i tilfeldig rekkefølge.

$$\hat{\sigma}^2 = MSE = \frac{\sum\sum(Y_{ij}-\bar{Y}_{i\cdot})^2}{13}$$

Innsatt: $$\hat{\sigma}^2 = \frac{\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2}{13} = \underline{4.92}$$

Modell: $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

$H_0: \mu_1 = \mu_2 = \mu_3$ H_1: Minst én forskjellig

$F = \frac{MS_A}{MS_E} \sim F_{2,13}$ Forkast H_0 dersom $F_{observert}$ blir "stor"

p-verdi $= P(F > F_{observert} \mid H_0) = \underline{0.001}$

$\Rightarrow$ Forkaster H_0 på 5% nivå, og påstår at gruppene er forskjellige.

b) 95% konf. int: Utgangspunkt er $\bar{Y}_{1\cdot} - \mu_1 \sim N(0, \frac{\sigma^2}{n_1})$

$\bar{Y}_{1\cdot} \pm t_{\alpha/2,13} \cdot S/\sqrt{4}$ gir $(58.60, 63.40)$

95% pred. int: Utgangspunkt er $Y^* - \bar{Y}_{1\cdot} \sim N(0, \sigma^2 + \frac{\sigma^2}{n_1})$

$\bar{Y}_{1\cdot} \pm t_{\alpha/2,13} \cdot S\sqrt{1+\frac{1}{4}}$ gir $(55.63, 66.37)$

Oppgave 2:

a) 2^3 $D = AC$ $E = -BC$

Def. rel.: $I = ACD = -BCE = -ABDE$

Alias-struktur:

$A \to A + CD$
$B \to B - CE$
$C \to C + AD - BE$
$D \to D + AC$
$E \to E - BC$
$AB \to AB - DE$
$AE \to AE - BD$

Resolusjon $R = III$

Alle 8 forsøkene gjennomføres i tilfeldig rekkefølge.

b) $\hat{B} = \frac{(Y_3 + Y_4 + Y_7 + Y_8) - (Y_1 + Y_2 + Y_5 + Y_6)}{4}$ Innsatt: $\hat{B} = 2{,}87$

$\hat{AB} = \frac{(Y_1 + Y_4 + Y_5 + Y_8) - (Y_2 + Y_3 + Y_6 + Y_7)}{4}$ $\hat{AB} = -1{,}08$

$\text{Var}(\hat{B}) = \frac{4}{n}\sigma^2 = \underline{\frac{1}{2}\sigma^2}$

$\hat{\sigma}^2 = 1.49 \Rightarrow \text{Var}(\hat{B}) = \frac{1}{2} \cdot 1.49$

Nye forsøk: Folding vil gjøre at vi kan estimere hovedeffektene ukonfundert med 2fi.

c) Starter med et 2^4 design

A B C D AB AC AD BC BD CD $\underset{\substack{\uparrow \\ F}}{ABC}$ $\underset{\substack{\uparrow \\ E}}{ABD}$ ACD BCD --

E = ABD
F = ABC er ett mulig valg

Generator: I = ABDE = ABCF
Def. rel. : I = ABDE = ABCF = BCEF

Korteste ord er 4 ⇒ $R = \underline{IV}$

<u>2 blokker</u>:

Vi kan blokkdele etter ACD eller BCD

ACD → ACD + BCE + BDF + ABDEF
BCD → BCD + ACE + ADF + DEF

Blokk blir da <u>ikke</u> konfundert med hovedeffekt eller 2f

<u>4 blokker</u>:

Må da bruke begge samspillene ACD og BCD

Samspill mellom blokker blir da <u>AB</u> og vi
være konfundert med AB-samspillet.

<u>Uheldig</u>.

Oppgave 3 :

a) Modell : $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon_i$

Antagelser: ε_i uavh. og $N(0, \sigma^2)$

Antagelsene sjekkes med diverse plott.

i) $T = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}}$ $(T = 9.68)$

ii) $S^2 = MSE$ $(S = 5.864)$

iii) $R^2 = \frac{SS_R}{SS_T} \cdot 100\%$ $(R^2 = 87.6\%)$

iv) $SS(\beta_2 | \beta_0, \beta_1) = SS(\beta_1, \beta_2 | \beta_0) - SS(\beta_1 | \beta_0)$ $(2391,6)$

v) $Var(\hat{\beta}) = (X'X)^{-1} \hat{\sigma}^2$ (0.05146)

$H_0: \beta_0 = 0 \quad H_1: \beta_0 \neq 0$

$p\text{-verdi} = P\left(T_{16} > \left| \frac{\hat{\beta}_0}{\sqrt{\hat{Var}(\hat{\beta}_0)}} \right| \,\Big|\, H_0\right)$

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

$H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$

partielle t-tester tester β_i gitt at de andre er med i modellen

$H_0: \beta_1 = \beta_2 = 0 \quad H_1:$ minst én forskjellig

Tester om X har betydning i regresjonen

$p\text{-verdi} = P(F_{2,16} > F_{observert} | H_0)$

(5/6)

b) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$

$\hat{y} = -10.2 + 12.9\,x - 0.684\,x^2$

$x = 7$: $\hat{y} = \underline{46.58}$

$$\frac{\partial \hat{y}}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2 x = 0 \Rightarrow \underline{\hat{x} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}}$$

Innsatt : $\hat{x} = \frac{12.9}{2 \cdot 0.684} = \underline{9.43}$

$$\hat{x} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} = g(\hat{\beta}_1, \hat{\beta}_2)$$

Taylorutvikler om $E(\hat{\beta}_1) = \beta_1$ og $E(\hat{\beta}_2) = \beta_2$

$$\hat{x} \approx g(\hat{\beta}_1, \hat{\beta}_2)\big|_{(\beta_1 \beta_2)} + \Sigma_{i=1}^{2} \frac{\partial g(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_i}\bigg|_{(\beta_1, \beta_2)} (\hat{\beta}_i - \beta_i) + R$$

$$= -\frac{\beta_1}{2\beta_2} - \frac{1}{2\beta_2}(\hat{\beta}_1 - \beta_1) + \frac{\beta_1}{2\beta_2^2}(\hat{\beta}_2 - \hat{\beta}_2) + R$$

$$\mathrm{Var}(\hat{x}) \approx \left(\frac{1}{2\beta_2}\right)^2 \mathrm{Var}(\hat{\beta}_1) + \left(\frac{\beta_1}{2\beta_2^2}\right)^2 \mathrm{Var}(\hat{\beta}_2) - \frac{2\beta_1}{4\beta_2^3} \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2$$

Har $\mathrm{Var}(\underset{\sim}{\hat{\beta}}) = (X'X)^{-1}\sigma^2$

Vi kjenner ikke β_1, β_2 og σ^2 så vi setter inn estimatene for disse.

$\mathrm{Var}(\hat{x}) \approx 0.0024 \times 34.4 = \underline{0.083}$

Dataene er riktige. Modellen er kun riktig i den grad den tilpasser dataene.

Man bør gjøre nye forsøk rundt $x = 9.43$ for å fin-tune

c) Modell: $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$

Ønsker å teste $H_0: \beta_1 = 0$. $H_1: \beta_1 \neq 0$

La * indikere verdier fra modellen under a)

$$SS(\beta_1 | \beta_0) = SS(\beta_1^* | \beta_0^*) = 1490.1$$

$$SS_E = SS_T - SS_R = 4431.9 - 1490.1 = 2941.8$$

$$MS_E = \frac{SS_E}{17}$$

$$F = \frac{SS_R/1}{SS_E/17} = \frac{1490.1}{2941.8/17} = \underline{8.61}$$

Under H_0 vil $F \sim F_{1,17}$

$$P(F_{1,17} > k_1 | H_0) \leq \alpha \Rightarrow k_1 = f_{0.01,1,17} = \underline{8.40}$$

<u>Forkast H_0</u>