Bokmålstekst.
Faglig kontakt under eksamen:
John Tyssedal, telefon 73593534

Eksamen i fag SIF 5068 Industriell statistikk

Onsdag 29. november
Tid: 09.00-13.00

Hjelpemidler: Godkjent lommekalkulator
Tapir: Statistiske tabeller og formler
Se også vedlagt ark.

Sensuren faller i uke 51.

Oppgave1.

Coca-Cola er hevdet å ha stimulerende effekt. For å undersøke om dette stemmer ble følgende forsøk utført:

12 studenter ble valgt ut tilfeldig og på forhånd trenet i fingertromming. 6 av studentene ble først trukket ut og satt til å drikke en flaske Cola. De andre 6 måtte drikke en flaske Solo. Studentene ble så satt til fingertromming og tallet på tromminger pr. minutt ble registrert.

Neste dag fikk de som drakk Coca-Cola første dag en flaske Solo mens de andre fikk en flaske Coca-Cola, og tallet på tromminger pr. minutt ble atter registrert. Resultatene ble:

Person nr. :	1	2	3	4	5	6	7	8	9	10	11	12
Cola :	252	245	252	243	240	242	248	249	242	238	220	243
Solo :	246	238	244	247	227	239	241	248	232	230	225	247
Diff :	6	7	8	-4	13	3	7	1	10	8	-5	-4

Datasettet ble først analysert ved hjelp av par t-testen og utskrift fra MINITAB er gitt nedenfor.

Paired T-Test and Confidence Interval

```
Paired T for cola - solo

                N     Mean     StDev   SE Mean
Cola           12   242.83      8.48      2.45
Solo           12   238.67      8.29      2.39
Difference     12     4.17      5.95      1.72

95% CI for mean difference: (0.39, 7.95)
T-Test of mean difference = 0 (vs > 0): T-Value = 2.43  P-
Value = 0.017
```

Formuler hypotesetesten. Hvilke forutsetninger om dataene bygger en slik test på? Sett opp utrykket for testobservator. Hva blir konklusjonen når signifikansnivået er 5%? Grunngi svaret.
Det er og mulig å utføre en test for å undersøke om Coca Cola har stimulerende effekt når en ikke forlanger at dataene skal være normalfordelte, men ha samme fordelings-egenskaper som ovenfor. Vis hvordan dette kan gjøres. Hva blir konklusjonen?

Oppgave 2

Nedenfor er det synt et 2^{4-1} forsøk utført for å undersøke hvordan bensinforbruket avheng av de 4 faktorene:

A: Dekkbredde	15 cm	17.5 cm
B: Mønsterdybde	5 mm	7.5 mm
C: Dekkmønster	Type 1	Type 2
D: Hastighet	70 km/t	95 km/t

Responsen er målt i hvor mange kilometer en kan kjøre pr liter.

A	B	C	D=ABC	Km/l
-1	-1	-1	-1	12.0
1	-1	-1	1	6.8
-1	1	-1	1	10.4
1	1	-1	-1	8.8
-1	-1	1	1	12.8
1	-1	1	-1	10.0
-1	1	1	-1	12.4
1	1	1	1	10.4

a) Hva blir definerende relasjon for dette forsøket? Hvilken resolusjon har det? Vis hvordan du kan estimere hovedeffekten for faktor A og samspillet mellom A og B og gi estimatene for disse. Hvorfor er det lettere å vurdere om det er en hovedeffekt av faktor A enn om det er et samspill mellom A og B?

I punkt 2b) og 2c) kan du anta at variablene er uavhengige og normalfordelte med lik varians.

b) Det ble bestemt at en skulle utføre den andre halvfraksjonen og. Estimerte effekter basert på alle de 16 forsøkene er gitt nedenfor. Du kan anta at alle tredje og høyere ordens samspill er 0, men du må regne med at det er en blokkeffekt mellom de to fraksjonene.
Finn ved hjelp av utskriften nedenfor et estimat for variansen til estimatorene til effektene og vurder hvem av disse som er signifikante på 5% nivå

Term	Effect	Coef
Constant		10.375
Block		-0.075
A	-2.950	-1.475
B	-0.050	-0.025
C	1.850	0.925
D	-1.050	-0.525
A*B	0.850	0.425
A*C	0.350	0.175
A*D	0.250	0.125
B*C	-0.350	-0.175
B*D	0.150	0.075
C*D	0.250	0.125
A*B*C	0.350	0.175
A*B*D	0.050	0.025
A*C*D	0.150	0.075
B*C*D	0.050	0.025

Analysis of Variance (coded units)

Source	DF	Seq SS	Adj MS
Blocks	1	0.0900	0.0900
Main Effects	4	52.9200	13.2300
2-Way Interactions	6	4.4600	0.7433
3-Way Interactions	4	0.6000	0.1500
Residual Error	0	0.0000	0.0000
Total	15	58.0700	

Nedenfor er synt estimerte forventede responsverdier for de 4 nivåkombinasjonene av A og B.

	A-	A+
B-	12.3	8.5
B+	11.4	9.3

Gi en tolkning av de estimerte effektene.

c) Dette forsøket ble egentlig utført i 4 blokker og blokkdelt etter de to samspillene ABCD og AC. Hvilket annet samspill blir nå konfundert med blokkeffekten? Vil denne blokkdelingen påvirke vurderingen av signifikans gjort i punkt 2b)? Grunngi svaret. Hvorfor må kvadratsummen for blokker i variansanalysetabellen bli summen av kvadratsummene for tre samspillsseffekter? Finn denne kvadratsummen og gjør de nødvendige forandringer i variansanalysetabellen gitt i punkt 2c)

Oppgave 3

Nedenfor er det gitt data for 31 tre av en viss type fra en nasjonalpark i USA. For hvert tre er det målt 3 variable. Disse er:

D: Diameter av treet målt i tommer 1.5 m over bakkenivå.
H: Høyde av treet målt i fot.
V: Volumet av treet målt i kubikkfot.

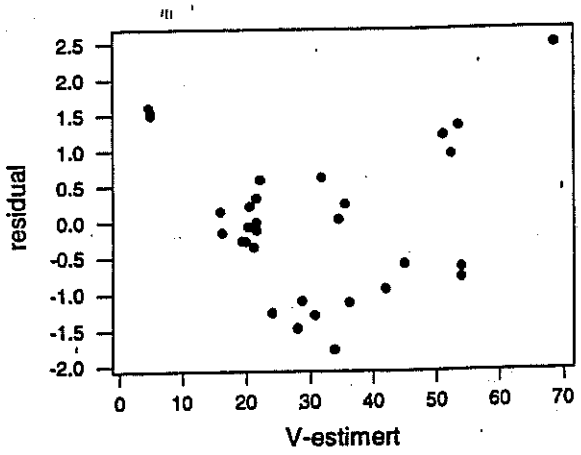
Nummer	D	H	V	Nummer	D	H	V
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Et problem med å måle volumet av et tre, er at en må sage det ned. Det er derfor av interesse å utvikle en modell som kan brukes til å estimere trevolumet utan at en trenger å felle treet. Nedenfor er det synt utskrift av en regresjonsanalyse med MINITAB der en har brukt V som responsvariabel og D og H som forklaringsvariable og der 4 av verdiene er erstattet med *.

```
The regression equation is
V = - 58.0 + 4.71 D + 0.339 H

Predictor       Coef      StDev       T       P
Constant     -57.988      8.638       *   0.000
D             4.7082          *   17.82   0.000
H                  *     0.1302    2.61   0.014

S = *        R-Sq = 94.8%     R-Sq(adj) = 94.4%

Analysis of Variance

Source          DF       SS       MS       F       P
Regression       2   7684.2   3842.1  254.97   0.000
Residual Error  28    421.9     15.1
Total           30   8106.1
```

a) Skriv opp tilpasset modell. Er denne like troverdig for alle verdier av D og H? Anta at feilleddene er uavhengige og $N(0,\sigma^2)$. Bruk utskriften til å finne de 4 verdiene som er erstattet med * og til å teste om modellen har signifikant forklaringsgrad. Bestem signifikansnivå selv.

b) Ei plott av (standardiserte) residual mot tilpassede volum-verdier er gitt nedenfor. Kommenter dette.



Det ble besluttet at en skulle prøve å innføre produktleddet mellom D og H, DH i modellen. En analyse med MINITAB ga følgende resultat:

V = 69.4 - 5.86 D - 1.30 H + 0.135 DH

Predictor	Coef	StDev	T	P
Constant	69.40	23.84	2.91	0.007
D	-5.856	1.921	-3.05	0.005
H	-1.2971	0.3098	-4.19	0.000
DH	0.13465	0.02438	5.52	0.000

S = 2.709 R-Sq = 97.6% R-Sq(adj) = 97.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	7908.0	2636.0	359.31	0.000
Residual Error	27	198.1	7.3		
Total	30	8106.1			

Tyder resultatene i utskriften på at alle leddene skal være med i modellen? Grunngi svaret. Hvorledes er den multiple determinasjonskoeffisienten definert? Hvordan tolker en denne? Gi noen grunner for å velge denne modellen framfor modellen i 3a).

c) Med tanke på formelen for sammenhengen mellom volum, diameter og høyde i en sylinder (evt. kjegle) ble det foreslått at en skulle innføre forklaringsvariabelen $X = D^2H$ og regressere V mot denne. Det synte seg at en da kunne tilpasse en modell uten konstantledd. Resultatet av denne regresjonen er synt nedenfor.

The regression equation is
V = 0.00211 X

Predictor	Coef	StDev	T	P
Noconstant				
X	0.00210810	0.00002722	77.44	0.000

S = 2.455
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36144	36144	5996.41	0.000
Residual Error	30	181	6		
Total	31	36325			

Tyder denne utskriften på at denne modellen gir betre tilpasning til dataene enn modellen gitt i 3b)? Utled et teoretisk utrykk for minste kvadratsumsestimatoren for stigningstallet og finn og et utrykk for variansen til denne. Anta at feilleddene er uavhengige og $N(0,\sigma^2)$ og finn et 95% prediksjonsintervall for volumet når D=15 og H = 80.

d La $\mathbf{X}$ være en designmatrise i en regresjonsmodell, vektoren $\hat{\boldsymbol{\beta}}$ en estimator for koeffisientene og $\mathbf{x}_0$ en vektor der 1. element er 1 og resten av elementene er innsatte verdier for forklaringsvariablene i den samme regresjonsmodellen. Er det noen sammenheng mellom utrykkene $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ og $\sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$? Forklar.

En annen statistiker foreslo med bakgrunn i den samme sammenhengen mellom volum, diameter og høyde gitt i 3c) at en skulle modellere en sammenheng mellom logaritmene til de tre variablene. Venter du å finne et konstantledd i en slik modell? Anta at konstantleddet og begge forklaringsvariablene ln(D) og ln(H) blir signifikante i en slik modell og at feilleddene er uavhengige $N(0,\sigma^2)$. Sett opp uttrykket for et 95% prediksjonsintervall for ln(V) for gitte verdier av D og H. Vis at det med en slik modell og vil være mulig å finne et prediksjonsintervall for V.

Løsningsforslag SIF 5068 Industriell Statistikk

Oppgave 1

X_i : Cola

Y_i : Solo

$D_i = X_i - Y_i$ er uavh, identisk normalfordelte, $i = 1, 2, \ldots, n$

$H_0: \mu_0 = 0 \qquad H_1: \mu_D > 0$

$$T = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} \quad \text{der} \quad S_D = \sqrt{\sum_{i=1}^{12} \frac{(D_i - \bar{D})^2}{11}}$$

P-verdien = 0.017 $\Rightarrow$ forkast H_0 på 5% nivå.

Wilcoxons rang sum test (uavh. identisk, kont. og symm. fordelt

	+	+	−	−	−	+	+	+	+	+	+	+
	1	3	4	4	5	6	7	7	8	8	10	13
R_i:	1	2	3.5	3.5	5	6	7.5	7.5	9.5	9.5	11	12

$W^- = 12 < 17$ = kritisk verdi $\Rightarrow$ forkast H_0

Oppgave 2

a) $D = ABC \Rightarrow I = ABCD$ og forsøket var resolusjon IV

$$\hat{\gamma} = \frac{6.8 + 6.8 + 10 + 10.4 - (12 + 10.4 + 12.8 + 12.4)}{4} = \frac{-11.6}{4} = -2.9$$

$$\hat{AB} = \frac{12 + 8.8 + 12.8 + 10.4 - (6.8 + 10.4 + 10 + 12.4)}{4} = \frac{4.4}{4} = 1.1$$

$\hat{A} = \hat{A} + \hat{BCD}$ og ventelig lik $\hat{A}$ siden trefaktorsamspel ofte er neglesjerbare

$\hat{AB} = \hat{AB} + \hat{CD}$. Kan ikkje skilje mellom $\hat{AB}$, $\hat{CD}$ evt. $\hat{AB} + \hat{CD}$

b) $$\hat{\sigma}^2_{\text{eff}} = \frac{(0.35)^2 + (0.05)^2 + (0.15)^2 + (0.05)^2}{4} = \frac{0.1225 + 0.0025 + 0.0225 + 0.0025}{4} = \underline{0.037}$$

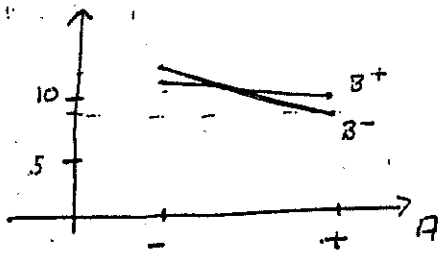
Kritisk verdi for signifikans av effektar er $t_{0.025,4}\sqrt{0.0375} = \underline{0.538}$

som syner at A, C, D og AB er signifikante.

Faktor C: Skifte av dekkmønster frå Type 1 til Type 2 gjev ein estimert auke i km/l på 1.85

Faktor D: Auke av fart frå 70 km/t til 95 km/t gjev ein estimert reduksjon i km/l på 1.05

Det er samspel mellom dekkbredde, A, og mønsterdjupn B



Det er større auke i bensinforbruket ved å auke dekkbredda når mønsterdjupn er 5 mm enn når den er 7.5.
Best økonomi får vi ved smalast dekkbredde og minst mønsterdjupn.

Smale dekk vil ha minst mønsterdjupn, medan breie vil ha den største mønsterdjupna.

c) B_1 = ABCD, B_2 = AC $\Rightarrow$ B_1, B_2 = BD ikke ortogonal og konfundert med blokkeffekten.

Alle dei resterande kolonnene er ortogonale på eventuelle blokkeffektar $\Rightarrow$ estimering av effektar og standardavvik til blokkeffektar uforandra slik at det ikke vil gje forandring i vurdering av signifikans for desse.

Vi har: $SS_T = \sum_{i=1}^{16} (y_i - \bar{y})^2 = SS_{ABCD} + SS_{AC} + SS_{BD} + SS_{REST}$

og $SS_T = SS_{Blokk} + SS_{REST}$

$\Rightarrow$ $SS_B = SS_{ABCD} + SS_{AC} + SS_{BD} = 0.09 + 0.49 + 0.09 = 0.67$

Variansanalyse tabellen

Kjelde	DF	SS
Blokk	3	$0.09 + 0.49 + 0.09 = 0.67$
N-effektar	4	$= 52.92$
To-faktor samspel	4	$4.46 - 0.49 - 0.09 = 3.88$
Tre-faktor -"-	4	$= 0.60$

a) $\hat{V} = -58 + 4.71D + 0.339H$

För ökad D och H leder $\hat{V}$ högre

$H_0: \beta_D = \beta_H = 0$ $\quad H_1$ minst ett $\neq 0$

$P(F_{2,28} \geq 254.97) = 0.000 \Rightarrow$ förkastar H_0 och påstår att modellen har signifikant förklaringsgrad.

$T = \frac{-57.988}{8.638} = -6.71$, $\quad S_D = \frac{4.7082}{17.82} = 0.2643$

$\hat{\beta}_H = 0.1302 \cdot 2.61 = 0.339$, $\quad s = \sqrt{\frac{431.9}{28}} = 3.882$

b) Alla ledda är signifikanta på att låta för ändra ledda + konstantledet är med på nivå minst 0.007.

$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ förklarar hur stor del av variationen som ges i förhål med modellen. (Minitab ger denna i %.

$R_b^2 > R_a^2$, $\quad S_b < S_a$ Funktion har ovan ej alla ledda är ledart signifikanta.

c) Vi ser att $S_c < S_b$ medan som tyder på bättre tillpassning.

$G = \sum_{i=1}^{n}(y_i - bx_i)^2$, $\quad \frac{dG}{db} = -2\sum_{i=1}^{n}(y_i - bx_i)x_i = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum x_i^2}$

$Var\ \hat{\beta} = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

$D = 15$, $\quad H = 80$ ger $\hat{y}_0 = 15^2 \cdot 80 \cdot 0,00\,2108 = 37.98$

$SD(y - \hat{y}_0) = \sqrt{3.455^2 + (15^2 \cdot 80)^2 \cdot (0.0000\,2122)} = 2.504$

$t_{0.025,\ 30} = 2.042$

$\Rightarrow$ 95% prediktionsintervall $37.98 \pm 2.504 \cdot 2.042$

$= 37.98 \pm 5.11 = (32.87, 43.09]$

d) $\text{Var}(x_0'\hat{\beta}) = x_0' \text{Var}(\hat{\beta}) x_0 = \sigma^2 x_0'(X'X)^{-1} x_0$

$V = k \cdot D^2 H \Rightarrow \ln V = \ln k + 2\ln D + \ln H \Rightarrow$ formeln

konstantled.

La U vara ny design matris och W: transformerat respons.

$W = \ln V$. V, fås:

$$P\left(\hat{W}_0 - t_{\frac{\alpha}{2},27} s \sqrt{1 + u_0'(U'U)^{-1}u_0} < W < \hat{W}_0 + t_{\frac{\alpha}{2},27} s \sqrt{1 + u_0'(U'U)^{-1}u_0}\right)$$

$$\Rightarrow P\left(e^{\hat{W}_0 - t_{\frac{\alpha}{2},27} s \sqrt{1 + u_0'(U'U)^{-1}u_0}} < V < e^{\hat{W}_0 + t_{\frac{\alpha}{2},27} s \sqrt{1 + u_0'(U'U)^{-1}u_0}}\right)$$