

Bokmål tekst
 Faglig kontakt under eksamen:
 Jo Eidsvik, telefon 73591696

Eksamen i fag SIF 5066 Anvendt statistikk

Fredag 2. august 2002

Tid: 09.00-14.00

Hjelpemiddel: Alle skrevne og trykte. Enkel lommekalkulator
 Tapir: Formler og tabeller i statistikk/Statistiske tabeller og formler

Sensuren faller 23. august.

Oppgave 1.

En ny medisin som var tenkt brukt til midlertidig reduksjon av hjerterytme skulle sammenlignes med en standard medisin. 10 tilfeldig valgte pasienter ble plukket ut til å være med i forsøket. Siden en regnet med at effekten av medisinen var svært personavhengig, ble forsøket administrert slik at hver pasient fikk den ene medisinen en dag og den andre den påfølgende. Rekkefølgen de fikk medisinene i ble randomisert. Resultatene fra forsøket er gitt nedenfor og syner prosentvis reduksjon i hjerterytme.

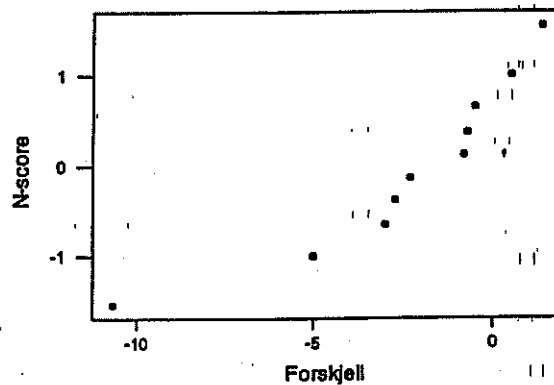
Pasient	1	2	3	4	5	6	7	8	9	10
Standard medisin	40.1	26.6	28.6	22.1	31.6	33.2	32.9	27.9	26.1	30.7
Ny medisin	40.8	37.3	31.3	24.4	30.2	34.0	33.4	27.4	29.1	35.7
Forskjell	-0.7	-10.7	-2.7	-2.3	1.4	-0.8	-0.5	0.5	-3.0	-5.0

- a) Gjør rede for hva slags forsøksopplegg som er brukt. Nedenfor er det synt utskrift av en analyse gjort med MINITAB på forskjellen mellom de to medisinene. Hvilke forutsetninger ligger til grunn for analysen? Formuler nullhypotesen og alternativ hypotese. Sett opp uttrykket for testobservator. Hva blir konklusjonen når signifikansnivået settes til 5%.

T-Test of the MeanTest of $\mu = 0.00$ vs $\mu \text{ not } = 0.00$

Variable	N	Mean	StDev	SE Mean	T	P
Forskjell	10	-2.38	3.46	1.10	-2.17	0.058

- b) Et normalplott av differansene mellom resultatene for de to medisinene er synt nedenfor. Kommenter dette.



Det ble foreslått at en også skulle utføre en Wilcoxon fortegns rang test på dataene. Hvilke forutsetninger bygger en slik test på? Formuler nullhypotesen og alternativ hypotese og utfør testen. Bruk 5% signifikansnivå. Hva blir konklusjonen? Sammenlign med resultatene i 1a) og kommenter.

Oppgave 2.

Et forsøk ble utført for å måle hvor fort en medisin løste seg opp i blodet på et dyr. Medisinen ble lagt i små beholdere sammen med vann eller enzym tatt fra blodet til dyret. Etter visse tidsrom ble det målt hvor mye av medisinen som hadde løst seg opp for hver av de 6 beholderne. Resultatene nedenfor er i %.

Tid i min.	Beholder					
	1	2	3	4	5	6
0	0	0	0	1	0	0
15	2	5	0	17	1	12
30	20	33	9	23	23	32
45	65	82	48	81	77	61
60	95	92	81	94	95	93
120	98	97	100	100	97	99

Det synes som om tiden har en opplagt effekt, men det var av særlig interesse å finne ut om beholderne påvirket oppløsningsraten. Deler av en regresjonsanalyse med MINITAB er gitt nedenfor.

Regression Analysis

The regression equation is

$$\text{Oppløs} = -1.56 + 0.52 \text{ Beholder} + 1.15 \text{ Tid}$$

Predictor	Coef	StDev	T	P
Constant	-1.558	*	-0.21	0.833
Beholder	0.521	1.641	*	0.753
Tid	1.15489	0.08706	13.27	0.000

S = * R-Sq = * R-Sq(adj) = 83.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	*	49805	24902	*	0.000
Residual Error	*	*	283		
Total	*	59138			

- a) Fyll inn de tallene som skal stå der det er merket med stjerner. Hvilke forutsetninger bygger en regresjonsanalyse på? Har regresjonen over signifikant forklaringsgrad? Grunngi svaret. Hvilken konklusjon vil du trekke vedrørende beholderens påvirkning av oppløsningsraten utifra analysen ovenfor?
- b) En statistiker hevdet at med rett randomisering ville en annen analyse egne seg bedre for å finne ut om beholderne påvirket oppløsingstiden. Foreslå en slik randomisering og gi en annen måte en kan analysere dataene på. Sett opp modell for analysen og spesifiser forutsetninger. Skriv også opp de kvadratsummene som inngår i analysen.

Nedenfor er det synt en delvis utskrift fra en to-veis variansanalyse.

Two-way Analysis of Variance

Analysis of Variance for Oppløs				
Source	DF	SS	MS	F
Beholder	*	651.8	*	*
Tid	*	57468.5	*	*
Error	*	1017.4	*	*
Total	*	59137.6		

Beholder	Mean	Tid	Mean
1	46.7	1	0.2
2	51.5	2	6.2
3	39.7	3	23.3
4	52.7	4	69.0
5	49.2	5	91.7
6	49.5	6	98.8

- c) Fyll inn de tallene som skal stå der det står stjerner. Formuler nullhypotesen og alternativ hypotese og finn ut hvilken konklusjon en nå kan trekke når det gjelder om beholderne påvirker oppløsningsraten. Bruk 5% signifikansnivå.

En som hadde særlig tro på regresjonsanalyse, mente at en kunne forbedre analysen i 2a) ved å modellere tidseffekten på en annen måte. Han foreslo derfor å innføre 5 nye tidsvariable. Den første, Teff1 er gitt ved

$$\text{Teff1} = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0, 0, \dots, 0, 0, \dots, 0, 0, \dots, 0)^T$$

Det vil si at den inngår som en variabel som har verdien 1 når tiden er 15 minutt og verdien 0 ellers. På samme måte vil Teff2 ha verdien 1 når tiden er 30 minutt og 0 ellers og tilsvarende for Teff3, Teff4 og Teff5. Resultatet av å erstatte tidsvariabelen med 5 nye i regresjonsanalysen er gitt nedenfor.

Regression Analysis

The regression equation is
 Oppløs = - 1.85 + 0.576 Beholder + 6.00 Teff1 + 23.2 Teff2 + 68.8 Teff3
 + 91.5 Teff4 + 98.7 Teff5

Predictor	Coef	StDev	T	P
Constant	-1.850	3.996	-0.46	0.647
Beholder	0.5762	0.7326	0.79	0.438
Teff1	6.000	4.334	1.38	0.177
Teff2	23.167	4.334	5.35	0.000
Teff3	68.833	4.334	15.88	0.000
Teff4	91.500	4.334	21.11	0.000
Teff5	98.667	4.334	22.76	0.000

S = 7.507 R-Sq = 97.2% R-Sq(adj) = 96.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	57503.3	9583.9	170.06	0.000
Residual Error	29	1634.3	56.4		
Total	35	59137.6			

Source	DF	Seq SS
Beholder	1	34.9
Teff1	1	12717.6
Teff2	1	8300.0
Teff3	1	133.4
Teff4	1	7112.1
Teff5	1	29205.3

- d) Skriv opp den estimerte modellen. Hvorledes vil du tolke de estimerte parametrene? Sammenlign modellen med modellen i 2a) og finn ut hvilken du vil foretrekke. Hvilken konklusjon vil du nå trekke når det gjelder om beholderne påvirker oppløsningsraten? Forklar hvorfor konklusjonen i 2d) blir forskjellig fra konklusjonen i 2c).

Oppgave 3

- a) Det skal utføres et 2^{8-4} forsøk i faktorene A, B, C, D, E, F, G og H der generatorene er gitt ved: E = ABC, F = ABD, G = ACD og H = BCD. Finn definerende relasjoner for dette forsøket. Hvilken resolusjonen har det? Vis hvordan hovedeffekten av faktor A og tofaktorsamspillet mellom faktor A og faktor B er konfundert.
- b) Forklar hvorfor en ikke kan bruke en fold-over av forsøket til å løse opp konfunderingen mellom tofaktor-samspillene. Anta nå at alle samspill av orden 3 eller høyere kan neglisjeres. Et nytt 2^{8-4} forsøk ble utført der en snur fortegnet i kolonnen for faktor A, men holder fortegnet for de andre faktorkolonnene uforandret. Finn ut hvor mange to-faktor samspill der A er involvert en nå kan estimere ukonfundert. Vis hvordan du kommer fram til svaret.

Løysingsforslag SIFS066, August 2002

1a)

Det er brukt ein parplan

X_i - Prosentvis reduksjon med gammel medisin

Y_i - Prosentvis reduksjon med ny medisin

$D_i = X_i - Y_i \sim N(\delta, \sigma^2)$ og uavh.

$H_0: \delta = 0$ $H_1: \delta \neq 0$

$$T = \frac{\bar{D}}{\frac{S_D}{\sqrt{10}}} \quad \text{der} \quad S_D^2 = \frac{1}{9} \sum_{i=1}^{10} (D_i - \bar{D})^2$$

p-verdi = 0,058 > 0,05 \rightarrow ikke forkast.

1b)

Dersom alle D_i kjennetegn for samme normalfordeling skulle plottet gitt ei tilnærma rett linje. Her er det iallfall minst ein observasjon som avviker fra resten.

Wilcoxon forbeholdt rang test bygger på uavh, kont, identisk og symmetriske fordelte variable.

$H_0: \tilde{\mu}_D = 0$ $H_1: \tilde{\mu}_D \neq 0$

Forkast:

Obs. verdi: 0.5 0.5 0.7 0.8 1.4 2.3 2.7 3.0 5.0 10.7

Rang: 1.5 1.5 3 4 5 6 7 8 9 10

$W_+ = 6.5$. Kritisk verdi = 8 \Rightarrow forkast H_0 .

enne test er ~~mer~~ mindre følsom for avvikende observasjoner

Turkey har verdien -10.7 forårsaka så stort standardavvik at vi ikke fikk forkasting med t-observatoren.

2)

a)
$$SE\text{ Dur } \hat{\beta}_0 = \frac{-1.558}{-0.21} = 7.335$$

T-verdi
$$\hat{\beta}_1 = \frac{0.521}{1.641} = 0.32$$

$$S = \sqrt{283} = 16.82$$

$$R^2 = \frac{49805}{59138} = 0.842$$

Variansanalyse tabellen.

Kilder	df	SS	MS	F
Regresjon	2			24902/283 = 88.05
Residual error	33	9333		
Total	35			

$H_0: \beta_1 = \beta_2 = 0$ $H_1: \text{ minst en forskjellig fra } 0$

$$F_{obs} = \frac{\frac{SSR}{2}}{\frac{SSE}{33}} = 88.05 \quad P(F_{2,33} > 88.05) = 0.000 = 7$$

Regresjon har signifikant forklaringsgrad.

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

gjør p-verdi 0.753. Dette betyr at beholder ikke har signifikant forklaringsgrad gitt at tid er med i modellen.

$$\hat{y} = -1.85 + 0.576 \text{ Behaldar} + 6.7 \text{ Test 1} + 23.2 \text{ Test 2} + 68.8 \text{ Test 3} \\ + 91.5 \text{ Test 4} + 98.7 \text{ Test 5}$$

$\hat{\beta}_1$ = estimert forandring i forventet respons når en aukar nummeret på behaldarane med 1 og dei andre tida er uforandra.

Tilsvarende viser dei andre parameterane ~~for~~ estimert forandring i forventet respons ved tilhøyrande hdsprang og for kvar behaldar.

2 a	2 d
$S = 16.8$	$S = 7.5$
$R^2 = 84.2$	$R^2 = 97.2$
$F = 88.05$	$F = 190.06$

} Dette indikerer at modellen i 2 d er \circ foretrekke.

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

p-verdien = 0.438 som betyr at behaldarar stige har tilgjeldig forklaringsgrad gitt at tidsvariablene er med i modellen.

Regressjonsanalyse modellen foreser ein linear samanking mellom responnen og nummereringa av behaldarane. Dette er ein nokre kunstig situasjon.

Oppgave 3

$$I = ABCE = ABDF = ACDG = BCDH$$

$$I^2 = CDEF = BDEG = ADEH = ACFG = ACFH = ABGH$$

$$I^3 = AETG = BEFH = CEGH = DFGA$$

$$I^4 = ABCDEFGH$$

Forsøkt har resulterert $R = \overline{IV}$

$$R = A + BCE + BDF + CDG + DEH + CFH + BGH + EFG + ABCDH + ACDEF + \dots + ADFGH + BCDEFGH$$

$$RB = AB + CE + DF + GH + BCDG + \dots + AEFH + ABCDEF + \dots + ABDFGH + CDEFGH$$

1. Dersom ein snur fortakna i kvar kolonne vil ein framleis ha $I = ABCE$ t.d. og ein får ikkje loyst opp i konfunderinga.

Snur fortakene i kolonne II og for.

$$- ABCE = - ABDF = - ACDG = BCDH$$

$$\Rightarrow ADEH = - ACFH = - ABGH = - AETG = - ABCDEFGH = \dots$$

Vi får at $RB = AB - CE - DF - GH$ som betyr at vi kan stå med

RB ukonfundert. Tilsvarende for alle de andre sannspelen

et faktor R ungar.