UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST110 — Statistiske metoder og dataanalyse
Eksamensdag:	Mandag 27. mai 2002.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 9 sider.	
Vedlegg:	Tabell over kumulative t-fordelinger.
Tillatte hjelpemidler:	Lommeregner, Formelsamling for ST100 og ST110, Haugens "Formler og tabeller", Jahren og Knutsens "Formelsamling i matematikk", Rottmanns "Mathematische Formelsammlung"

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

For rullestolbrukere kan trykkbelastninger være et problem som i verste fall kan føre til sårdannelser i setemusklene. Spesielle rullestolputer kan redusere trykkbelastningen. Vi skal i denne oppgaven se nærmere på en studie som sammenliknet to typer av rullestolputer, som vi for enkelhets skyld vil kalle pute I og pute II.

Ti voksne personer var med i studien. Hver av personene prøvde de to putetypene i tilfeldig rekkefølge, og trykket (i mmHg) mellom setemusklene og rullestolen ble målt. Resultatene ble som gitt i tabellen på neste side. Der er det også gitt noe beskrivende statistikk for differansene i tabellen.

(Fortsettes side 2.)

Person nr.	Pute I	Pute II	Differanse
1	47	59	12
2	43	47	4
3	42	43	1
4	76	84	8
5	49	70	21
6	62	86	24
7	53	68	15
8	75	61	−14
9	88	80	−8
10	77	76	−1

Variable	N	Mean	Median	StDev	SE Mean
Differanse	10	6,20	6,00	12,22	3,86

Variable	Minimum	Maximum	Q1	Q3
Differanse	-14,00	24,00	-2,75	16,50

(a) Vi tenker oss at differansene i tabellen er observerte verdier av stokastiske variable med forventningsverdi Δ. Bestem et 90% konfidensintervall for Δ og gi en fortolkning av intervallet. Hvilke forutsetninger bygger konfidensintervallet på?

(b) Studien av rullestolputene er gjort som et parforsøk, der hver person prøvde begge de to putene. Kan du angi en annen type forsøk som kunne ha vært benyttet for å sammenligne putene? Hvorfor tror du det ble valgt å gjøre et parforsøk?

Dataene i tabellen ovenfor er egentlig et utdrag av resultatene fra et blokkforsøk, der hver av de ti personene prøvde fem forskjellige slags puter. En analyse av hele datasettet ga variansanalysetabellen:

Source	DF	SS	MS	F	P
Pute	4	3847,3	961,8	10,42	0,000
Person	9	6489,1	721,0	7,81	0,000
Error	36	3324,3	92,3		
Total	49	13660,7			

(c) Vi tenker oss at trykkmålingen for pute nummer i og person nummer j er en observert verdi av en stokastisk variabel Y_{ij}. Skriv opp den modellen for Y_{ij}-ene som variansanalysen bygger på. I variansanalysetabellen er det gitt resultatet av to tester. Hvilke nullhypoteser er dette tester for, og hvilke konklusjoner kan du trekke av testene?

(Fortsettes side 3.)

Oppgave 2.

I industrien er det vanlig å gjøre forsøk for å undersøke hvor lenge ulike tekniske komponenter fungerer før de går i stykker. I et slikt forsøk ble en bestemt type springfjær testet ut. Ved hjelp av en spesiell testmaskin ble 10 springfjærer spent og spenningen sluppet opp om og om igjen. For hver springfjær ble det registrert hvor mange ganger den ble spent før den gikk i stykker. Resultatet ble som følger (gitt i 100 000 ganger)

2,25	1,71	1,98	1,89	1,89	1,35	1,62	1,35	1,17	1,62

Vi betrakter tallene i tabellen som observerte verdier av uavhengig og identisk fordelte stokastiske variable $T_1, T_2, \ldots, T_{10}$. En vanlig sannsynlighetsmodell for data av denne typen, er å anta T_i-ene Weibull-fordelt, dvs. at de har sannsynlighetstettheten

$$f(t) = \begin{cases} \lambda\alpha\, t^{\alpha-1}\, e^{-\lambda t^{\alpha}}, & \text{hvis} \quad t > 0; \\ 0, & \text{ellers.} \end{cases}$$

Vi vil først studere generelt hvordan vi kan bruke maksimum likelihood metoden til å estimere parametrene i Weibull-fordelingen på grunnlag av observerte verdier $t_1, t_2, \ldots, t_n$ av uavhengig og identisk Weibull-fordelte variable $T_1, T_2, \ldots, T_n$. I punktene e og f vender vi tilbake til tallene i tabellen.

(a) La $\boldsymbol{\theta} = (\alpha, \lambda)^T$. Sett opp likelihood funksjonen, og vis at log-likelihood funksjonen blir

$$l(\boldsymbol{\theta}) = n\log\lambda + n\log\alpha + (\alpha - 1)\sum_{i=1}^{n}\log t_i - \lambda\sum_{i=1}^{n} t_i^{\alpha}$$

(b) Vis at score funksjonene blir

$$\begin{aligned} s_1(\boldsymbol{\theta}) = \frac{\partial}{\partial\alpha} l(\alpha,\lambda) &= \frac{n}{\alpha} + \sum_{i=1}^{n}\log t_i - \lambda\sum_{i=1}^{n} t_i^{\alpha}\log t_i \\ s_2(\boldsymbol{\theta}) = \frac{\partial}{\partial\lambda} l(\alpha,\lambda) &= \frac{n}{\lambda} - \sum_{i=1}^{n} t_i^{\alpha} \end{aligned}$$

(*Vink:* Husk at $\frac{\partial}{\partial\alpha}(t^{\alpha}) = t^{\alpha}\log t$.)

(c) Bestem de andre ordens partielle deriverte

$$\frac{\partial^2}{\partial\alpha^2} l(\alpha,\lambda)\ , \quad \frac{\partial^2}{\partial\lambda^2} l(\alpha,\lambda) \quad \text{og} \quad \frac{\partial^2}{\partial\alpha\partial\lambda} l(\alpha,\lambda)$$

og forklar hva vi mener med den observerte informasjonsmatrisen $\bar{\boldsymbol{J}}(\boldsymbol{\theta})$.

(Fortsettes side 4.)

(d) Forklar hvordan vi kan bruke Newton-Raphsons metode til å finne maksimum likelihood estimatet $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\lambda})^T$.

Et alternativ til Newton-Raphsons metode er Fishers scoringsalgoritme. Hvorfor er det ikke så hensiktsmessig å bruke scoringsalgoritmen for Weibull-fordelte data?

Fins det andre alternativ til Newton-Raphsons metode vi kunne ha benyttet til å finne maksimum likelihood estimatet?

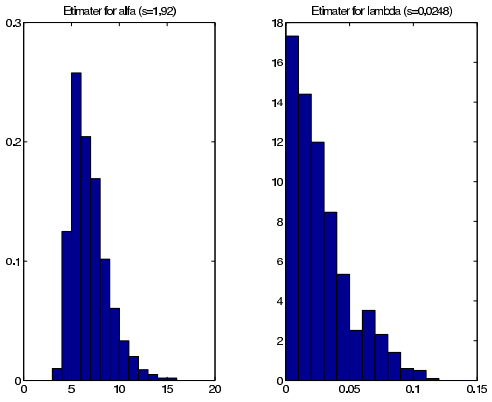
(e) For dataene i tabellen gir Newton-Raphsons metode maksimum likelihood estimatene $\hat{\alpha} = 5,98$ og $\hat{\lambda} = 0,0284$. Den inverse av den observerte informasjonsmatrisen, beregnet for de estimerte parametrene, blir

$$\bar{\boldsymbol{J}}^{-1}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} 2,13 & -0,0406 \\ -0,0406 & 0,000855 \end{bmatrix}$$

Angi estimat for standardfeilene til $\hat{\alpha}$ og $\hat{\lambda}$. Hvilket generelt resultat bruker du for å finne disse estimatene?

(f) Resultatene i forrige punkt benytter at $\hat{\alpha}$ og $\hat{\lambda}$ er tilnærmet normalfordelte. Siden vi bare har $n = 10$ observasjoner, kan vi frykte at normaltilnærmingen ikke er så god her. Et alternativ til normaltilnærmingen er å bruke den parametriske bootstrap metoden. Forklar tankegangen bak den parametriske bootstrap metoden, og gjør rede for hvordan den kan brukes til å bestemme standardfeilen til estimatene.

Nedenfor er det gitt histogrammer basert på 1000 bootstrap estimater av α og λ. Over hvert histogram er det også gitt empiriske standardavvik for de 1000 bootstrap estimatene. Hva blir estimatene for standardfeilen til $\hat{\alpha}$ og $\hat{\lambda}$ basert på bootstrap metoden? Diskuter ut fra bootstrap estimatene hvor god/dårlig normaltilnærmingen er.



(Fortsettes side 5.)

Oppgave 3.

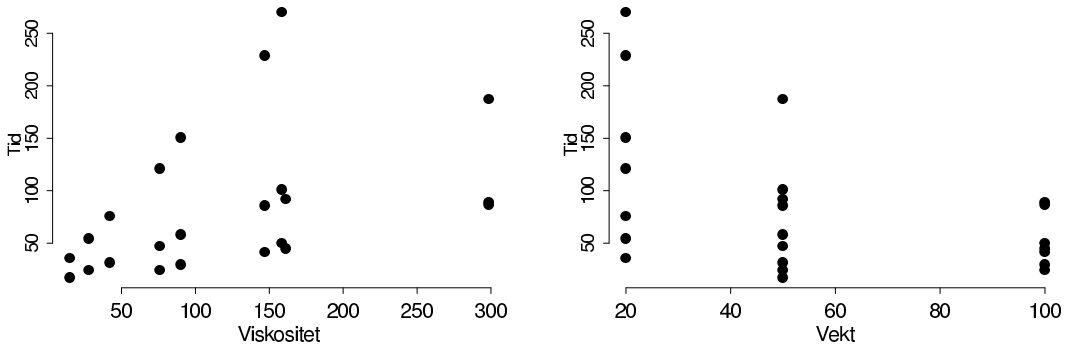
Viskositet (tykkhetsgrad) av en væske blir bestemt ved å måle tiden en indre sylinder av et såkalt viskositetsmeter bruker for å utføre et gitt antall rotasjoner. Viskositetsmeteret kan bli gitt ulike vektpåvirkninger som influerer på rotasjonstidene.

Måleinstrumentet blir kalibrert ved å måle tiden med varierende vekter på væsker med kjente viskositeter. Datasettet vi skal studere kommer fra en slik kalibrering.

Datasettet er gitt i tabellen nedenfor:

v	w	Y	v	w	Y	v	w	Y
14,7	20	35,6	27,5	50	24,3	75,7	100	24,6
27,5	20	54,3	42,0	50	31,4	89,7	100	30,0
42,0	20	75,6	75,7	50	47,2	146,6	100	41,7
75,7	20	121,2	89,7	50	58,3	158,3	100	50,3
89,7	20	150,8	146,6	50	85,6	161,1	100	45,1
146,6	20	229,0	158,3	50	101,1	298,3	100	89,0
158,3	20	270,0	161,1	50	92,2	298,3	100	86,5
14,7	50	17,6	298,3	50	187,2	158,3	20	270,0

Figuren nedenfor viser plott av tid mot viskositet (venstre) og tid mot vekt (høyre).



En mulig modell for sammenhengen mellom responsen Y og forklaringsvariablene vekt w og viskositet v er

$$Y = \beta_0 + \beta_1 v + \beta_2 w + \beta_3 z + \varepsilon$$

der $z = v * w/100$. En kjøring av denne modellen i MINITAB ga følgende utskrift:

(Fortsettes side 6.)

```
Regression Analysis: Y versus V; W; Z

The regression equation is
Y = 46,6 + 1,16 V - 0,639 W - 0,835 Z

Predictor        Coef     SE Coef          T        P
Constant        46,63       23,92       1,95    0,066
V              1,1589      0,1996       5,81    0,000
W             -0,6389      0,4232      -1,51    0,148
Z             -0,8345      0,2789      -2,99    0,007

S = 31,73       R-Sq = 80,8%     R-Sq(adj) = 77,7%
```

(a) Forklar innholdet i MINITAB-utskriften ovenfor.

Hvilke forutsetninger bygger utskriften på?

Argumenter også hvorfor en fornuftig alternativ modell er

$$Y = \beta_0 + \beta_1 v + \beta_2 z + \varepsilon. \qquad (*)$$

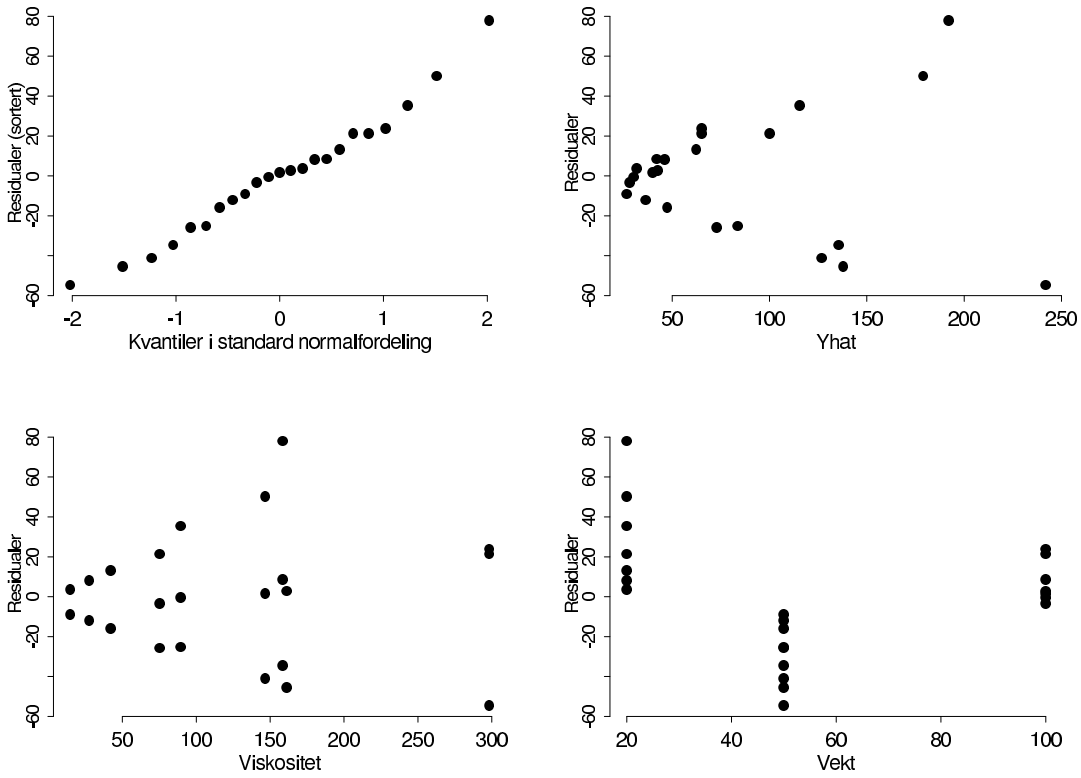
I resten av oppgaven vil vi holde oss til den reduserte modellen $(*)$. En MINITAB-utskrift basert på denne modellen er gitt nedenfor.

```
Regression Analysis: Y versus V; Z

The regression equation is
Y = 15,5 + 1,35 V - 1,18 Z

Predictor        Coef     SE Coef          T        P
Constant        15,50       12,52       1,24    0,230
V              1,3515      0,1584       8,53    0,000
Z             -1,1849      0,1595      -7,43    0,000

S = 32,73       R-Sq = 78,4%     R-Sq(adj) = 76,3%
```

(b) Plottene på neste side viser QQ-plott av residualene (øverst til venstre), residualer mot $\hat{Y}$ (øverst til høyre), residualer mot v (nederst til venstre) og residualer mot w (nederst til høyre).

Forklar hva de ulike plott kan brukes til og diskuter disse plottene.

(Fortsettes side 7.)



(c) Uansett dine konklusjoner i forrige oppgave, vil vi i denne oppgaven basere oss på modellen $(*)$.

Hvis $v = 50,0$ og $w = 60,0$, beregn et estimat for forventningen til Y.

Hvis $\boldsymbol{X}$ er matrisen med 1. kolonne bestående av 1-ere, 2. kolonne av v_i-ene og 3. kolonne av z_i-ene, så er

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = 10^{-3}\begin{bmatrix} 146,3045 & -1,2181 & 0,5641 \\ -1,2181 & 0,0234 & -0,0203 \\ 0,5641 & -0,0203 & 0,0237 \end{bmatrix}$$

Bruk dette til å beregne et estimat for standardfeilen til ditt estimat ovenfor.

Oppgave 4.

I forrige oppgave prøvde vi å tilpasse lineære modeller til tiden Y. Teoretiske betraktninger tilsier imidlertid at følgende modell for den (ikke-lineære) sammenhengen mellom tid Y, vekt w og viskositet v er rimelig:

$$Y = \frac{\beta_1 v}{w - \beta_2} + \varepsilon. \qquad (**)$$

(Fortsettes side 8.)

I denne oppgaven vil vi analysere denne modellen. Vi vil anta at alle støyleddene er uavhengige og normalfordelte med forventning 0 og varians σ^2. La $\boldsymbol{\theta} = [\beta_1, \beta_2, \sigma^2]^T$. Numerisk optimering av log-likelihood funksjonen for modell $(**)$ ga følgende estimater:

$\hat{\beta}_1 = 29,40, \hat{\beta}_2 = 2,218, \hat{\sigma}^2 = 35,87$

Videre ble den inverse Fisher informasjonsmatrisen

$$\bar{I}(\widehat{\boldsymbol{\theta}})^{-1} = \begin{bmatrix} 0,765 & -0,512 & 0 \\ -0,512 & 0,404 & 0 \\ 0 & 0 & 55,95 \end{bmatrix}$$

(a) Sammenlikn variansestimatet for støyleddet basert på den ikke-lineære modellen $(**)$ med variansestimatet basert på den lineære modellen $(*)$ i den forrige oppgaven. Kommenter!

Angi et estimat på standardfeilen til $\hat{\beta}_2$ og bruk dette til å konstruere et tilnærmet 95% konfidensintervall for β_2.

Hva blir konklusjonen på å teste hypotesen $H_0 : \beta_2 = 0$ mot $H_1 : \beta_2 \neq 0$. Forklar hvorfor dette svarer til å teste om modellen er lineær.

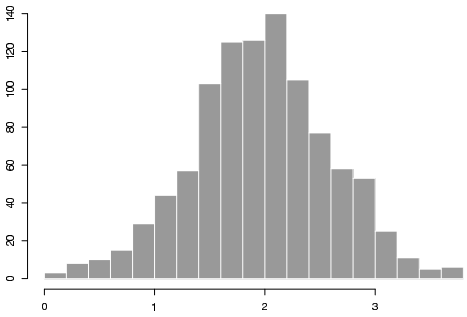
Ikke-parametrisk bootstrapping ble også utført basert på den ikke-lineære modellen $(**)$.

Bootstrap simuleringene ga

$$\sum_{b=1}^{1000} (\hat{\beta}^*_{2,b} - \bar{\hat{\beta}}_2)^2 = 387,82.$$

der $\hat{\beta}^*_{2,i}$ er estimatet for β_2 basert på bootstrap simulering nr. i og $\bar{\hat{\beta}}_2$ er gjennomsnittet av $\hat{\beta}^*_{2,i}$-ene.

Et histogram av $\hat{\beta}^*_{2,i}$-ene er vist nedenfor.



Histogram av $\hat{\beta}^*_2$

(Fortsettes side 9.)

(b) Forklar hvordan de nødvendige bootstrap simuleringer utføres.

Lag et alternativt estimat for standardfeilen til $\hat{\beta}_2$ basert på bootstrap simuleringene.

Diskuter forskjeller/likheter mellom denne standardfeilen og den du fikk i (c).

(c) Vi er igjen interessert i å estimere forventningsverdien for Y når $v = 50,0$ og $w = 60,0$. Hva blir estimatet i dette tilfellet?

Forklar hvordan du kan bruke bootstrapping til å lage et estimat for standardfeilen for dette estimatet.

SLUTT