UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Mandag 30. mai 2005.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 6 sider.	
Vedlegg:	Tabeller over normal-, t-, F- og χ^2-fordelingene.
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og for STK1120.

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

Oppgave 1.

Frank and Althoen (Statistics – concepts and applications, 1994) skriver om et forsøk på å undersøke effekten av ulike signaler på reaksjonstider hos menn og kvinner. Det er totalt 15 menn og 15 kvinner med i undersøkelsen og disse blir tilfeldig fordelt i 3 grupper med 5 menn og 5 kvinner i hver. Signalet for den første gruppen er en hørlig tone (Tone), signalet for den andre gruppen er et kort lyssignal (Light) og signalet for den tredje gruppen er et mild elektrisk puls fra et metall-bånd rundt håndlettet (Pulse). Alle personene er instruert til å presse [ENTER] knappen på en PC så snart signalet er oppfanget, og tiden mellom signalet og responsen blir automatisk lagret. Hver person får 12 forsøk og summen av responstidene (i sekunder) er gitt i tabellen nedenfor.

(Fortsettes side 2.)

	Tone	Light	Pulse
	10.0	6.0	9.1
	7.2	3.7	5.8
Men	6.8	5.1	6.0
	6.0	4.0	4.0
	5.0	3.2	5.1
	10.5	6.6	7.3
	8.8	4.9	6.1
Women	9.2	2.5	5.2
	8.1	4.2	2.5
	13.4	1.8	3.9

(a) Tabellen nedenfor viser et (delvis) resultat for variansanalyse på disse data.

```
Source       df   SS     MS    F
Kjonn         1   2.133   *  0.6410
Signaltype    *  97.267   * 14.6119
Interaksjon   *    *      *  3.4952
Error         *  79.880   *
Total        29 202.547
```

Sett opp den modell som ligger bak denne analysen. Hvilke antagelser bygger denne modellen på?

Fyll ut de manglende tallene (markert med *).

(b) Forklar hvordan du kan teste om `Interaksjon` er en signifikant faktor.

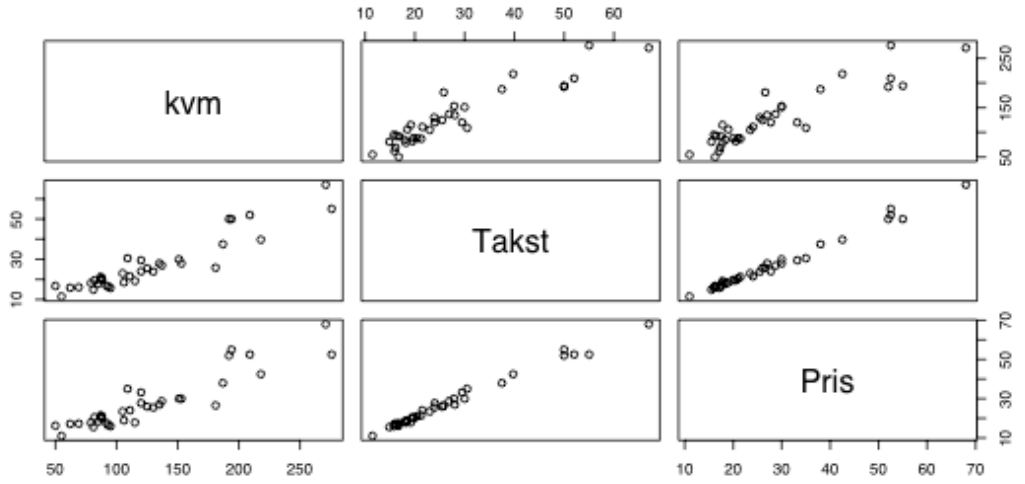
Når du utfører denne testen med signifikansnivå $\alpha = 0.05$, hva blir konklusjonen?

Test også på de andre faktorene i modellen og kommentér resulatene.

Oppgave 2.

Torsdag 28. april 2005 har Asker og Bærum Budstikka en artikkel (med overskrift **Åtte av ti over takst**) om priser av salg av boliger i forhold til boligtakst. I artikkelen er det også en tabell over 35 boliger med takst (Takst), antall kvadratmeter (kvm) og hvilke priser boligene er solgt for (Pris). (Denne tabellen er gjengitt bakerst i oppgavesettet for kompletthetsskyld men vil ikke være nødvendig å bruke for å løse denne oppgaven.) Vi vil i denne oppgaven se på disse dataene og prøve å se på sammenhengen mellom pris, takst og antall kvadratmeter. Dataene er vist i kryssplott nedenfor.

(Fortsettes side 3.)



Nedenfor er resultatet av en lineær regresjonsanalyse med Pris (i 100.000 kr) som responsvariabel og antall kvm. og Takst (også i 100.000 kr) som forklaringsvariable:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72634    0.65168   1.115    0.273
kvm         -0.01828    0.01278  -1.430    0.162
Takst        1.09548    0.05506  19.895    0.000

Residual standard error: 1.562 on 32 degrees of freedom
Multiple R-Squared: 0.9877
```

(a) Sett opp regresjonsmodellen som denne analysen bygger på og de vanlige antagelser som ligger til grunn for analyse basert på denne modellen.

Forklar hva de ulike delene av utskriften står for.

(b) Per og er interessert i en annonsert bolig på 105 kvm med Takst på 2.300.000 kr. Basert på modellen ovenfor, beregn en prediksjon for Pris.

Utled et generelt uttrykk for variansen til prediksjonsfeilen som funksjon av varianser og kovarianser for estimatene til regresjonskoeffisientene.

For de ovenfor angitte verdier av kvm og Takst blir variansen til prediksjonsfeilen lik 0.0821 (dette behøver du ikke å regne ut). Bruk dette til å lage et prediksjonsintervall for Pris.

(c) Resultatene ovenfor viser at kvm ikke er en signifikant variabel i modellen (og selve estimatet på koeffisienten blir faktisk negativ!). Samtidig ville en forvente en klar positiv sammenheng mellom pris og størrelse på boliger.

Hvordan kan man forklare dette resultatet?

(Fortsettes side 4.)

Oppgave 3.

Dataene nedenfor angir antall forekomster av lever-svulst i rotter etter eksponering av ulike mengder DDT (Tomatis, et al.,1972):

Dose (100ppm)	Antall dyr testet	Svulst forekomster
0	111	4
0.02	105	4
0.10	124	11
0.50	104	13
2.50	90	60

Vi vil anta forekomster av svulst er uavhengige mellom dyrene.
Anta sannsynligheten for svulst følger en såkalt Armitage-Doll model, som sier at

$$\Pr(\text{forekomst av svulst}|x) = 1 - e^{-(\beta_0+\beta_1 x+\beta_2 x^2)}$$

der x er dose av DDT (i 100ppm). Vi er interessert i å estimere parametervektoren $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.
Vi vil i det videre la n_c være antall testede dyr på dose nivå x_c og y_c tilhørende antall dyr med forekomster av svulst. Videre vil vi la C være antall dosenivåer.

(a) Sett opp likelihood funksjonen for $\boldsymbol{\beta}$ og vis at

$$\frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta}) = \sum_{c=1}^{C} x_c^j [y_c \frac{e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}}{1 - e^{-(\beta_0+\beta_1 x_c+\beta_2 x_c^2)}} - (n_c - y_c)], \quad j = 0, 1, 2$$

Regn ut forventningen til $\frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta})$ og kommentér hvordan det stemmer med den generelle teorien.

(b) Numerisk optimering av likelihooden ga

$$\hat{\boldsymbol{\beta}} = (0.0459, 0.1627, 0.1019)$$

$$\bar{J}^{-1}(\hat{\boldsymbol{\beta}}) = \frac{1}{1000} \begin{pmatrix} 0.2207 & -0.9849 & 0.3705 \\ -0.9849 & 13.2389 & -5.4734 \\ 0.3705 & -5.4734 & 2.8201 \end{pmatrix}$$

Bruk dette til å konstruere et 95% konfidensintervall for β_1. Spesifisér hvilke antagelser/resultater du baserer deg på ved konstruksjon av konfidensintervallet.

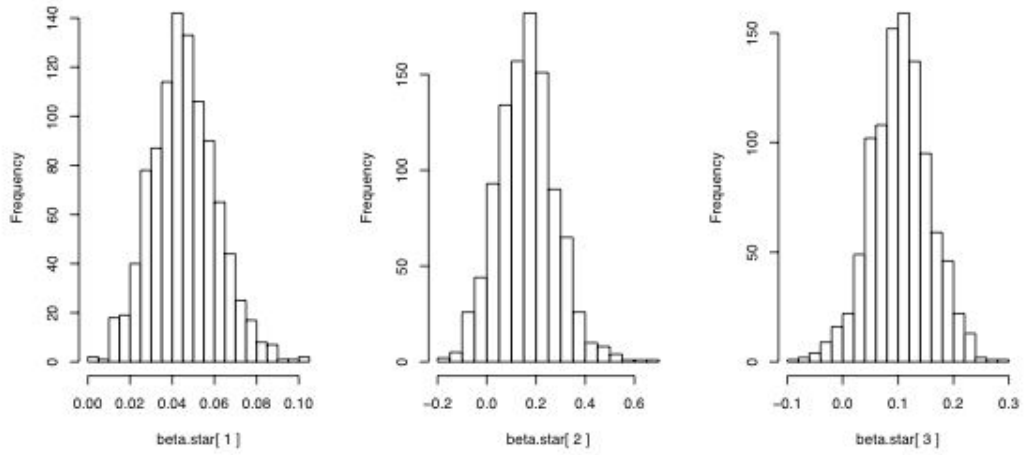
(Fortsettes side 5.)

(c) En parametrisk bootstrapping ble utført på problemet, som ga følgende resultater:

	Gj.snitt	Std.avvik	0.025 kv.	0.975 kv.
β_0^*	0.0456	0.0154	0.0157	0.0772
β_1^*	0.1619	0.1169	-0.0594	0.3952
β_2^*	0.1040	0.0538	-0.0077	0.2107

Her er "Gj.snitt" gjennomsnitt, "Std.avvik" er empirisk standard avvik, "0.025 kv."er 2.5% kvantilen og "0.975" er 97.5% kvantilene til bootstrap samplene av de tilhørende β-estimatene.

Figuren nedenfor viser også histogrammer av de simulerte β_i^*-ene.



Bruk dette til å konstruere et standard bootstrap intervall (basic bootstrap interval) på 95% nivå for β_1.

Kommentér eventuelle forskjeller med det intervallet du fikk i (b).

Oppgave 4.

14. mai 2001 stemte høyesterett i USA ned statlige lover som lovliggjorde marijuana for medisinsk bruk. Den amerikanske Gallup organisasjonen utførte senere en undersøkelse av tilfeldig valgte amerikanere (18 år eller eldre) der de ble spurt om de støttet den begrensede bruk av marijuana når foreskrevet av leger for å lindre smerter og lidelser. (Kilde: Sullivan, Statistics - Informed Decisions using data, 2004.) Resulatet av undersøkelsen, oppdelt etter aldersgrupper, er følgende:

	Alder		
Mening	18-29 år	30-49 år	50 år eller eldre
For	172	313	258
Imot	52	103	119

(Fortsettes side 6.)

En ønsker å teste om det er sammenheng mellom alder og mening. Sett opp en passende null-hypotese for å teste dette.
Du får oppgitt at den tilhørende testobservator er $\chi^2 = 6.6814$ i dette tilfellet (dette behøver du ikke å regne ut). Bruk dette til å utføre testen. Angi også øvre og nedre grenser for p-verdien til testen.

Tabell over boligdata for oppgave 2

	kvm	Pris	Takst		kvm	Pris	Takst
1	194	55.00	50.00	19	105	23.50	23.00
2	109	35.00	30.50	20	209	52.50	52.00
3	120	27.75	24.00	21	88	21.00	20.50
4	120	33.20	29.50	22	106	19.00	18.50
5	218	42.50	39.75	23	187	38.00	37.50
6	111	24.10	21.50	24	85	18.40	17.90
7	153	30.00	27.90	25	88	20.00	19.70
8	192	52.00	50.00	26	87	21.50	21.30
9	137	28.75	26.90	27	95	16.00	15.80
10	130	25.50	23.90	28	93	16.50	16.30
11	62	17.10	15.90	29	151	30.00	30.00
12	69	17.30	16.10	30	79	17.80	18.20
13	271	68.00	67.00	31	50	16.25	16.75
14	82	20.50	19.50	32	55	11.00	11.50
15	181	26.60	25.80	33	135	27.00	28.00
16	92	17.50	16.80	34	115	17.80	19.20
17	81	15.50	14.90	35	276	52.50	55.00
18	125	26.00	25.50				

SLUTT