UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Torsdag 5. juni 2014.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 4 sider.	
Vedlegg:	Tabell over normal-, t-, χ^2- og F-fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120

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

Oppgave 1.

I et eksperiment vil man studere hvordan to ulike former for jern (Fe^{2+} og Fe^{3+}) lagres i kroppen. Den formen som lagres best, vil være best egnet som kosttilskudd. 108 mus ble tilfeldig delt i 6 grupper med 18 i hver; 3 grupper fikk Fe^{2+} i tre forskjellige konsentrasjoner, og 3 grupper fikk Fe^{3+} i de samme tre konsentrasjonene. Jernet ble gitt oralt, og var radioaktivt merket, slik at det kunne detekteres. Etter en gitt tid ble prosenten av jernet som fremdeles var lagret i kroppen, målt. Disse målingene ble videre log-transformert, slik at fordelingene er mer symmetriske og variansen kan antas konstant.

Gjennomsnittet av log-transformert prosent lagret jern i de seks gruppene var

	0.3 mM	1.2 mM	10.2 mM
Fe^{2+}	2.40	2.09	1.68
Fe^{3+}	2.28	1.90	1.16

Resultatet av en analyse av de log-transformerte observasjonene i R gav ANOVA-tabellen nedenfor.

```
                    Df     Sum Sq      Mean Sq     F value
Iron_type            1      2.074       2.074           ?
Doseage              2     15.588           ?       22.53
Iron_type:Dosage     ?      0.810       0.405        1.17
Residuals          102     35.296       0.346
```

(Fortsettes på side 2.)

a) Lag en figur som viser hvordan gjennomsnittlig log data avhenger av konsentrasjon og jern-type. Hva indikerer figuren om forskjellen på Fe^{2+} og Fe^{3+}? Forklar hva interaksjon er i denne sammenheng, og hvordan en slik figur kan gi indikasjoner på om det er interaksjon mellom type jern og konsentrasjon eller ikke.

b) Sett opp modellen som ligger til grunn for ANOVA-analysen som er blitt utført. Fyll inn de manglende tallene i tabellen, markert med ?. Forklar hvordan du har funnet dem.

c) Test nullhypotesen om at det ikke er noen interaksjon mellom type jern og konsentrasjon. Bruk nivå $\alpha = 0.05$. Spesifiser hypotesene du tester, testobservator, fordeling og resulterende forkastningsområde. Skriv en konklusjon.

d) Er det noen effekt av type jern på log-transformert prosent lagret jern? Bruk $\alpha = 0.05$ også her.

e) Utled et 95% konfidensintervall for forskjellen mellom (log) prosent lagret jern for Fe^{2+} og Fe^{3+}.

Oppgave 2. Vi skal se på en meget enkel modell, der $X_1, X_2, ..., X_{100}$ er uavhengige observasjoner fra den samme normalfordelingen $N(\theta, 1)$, og man vil teste om forventningen θ er null eller ikke, dvs. $H_0 : \theta = 0$ mot $H_a : \theta \neq 0$.

a) Skriv opp likelihoodfunksjonen for denne modellen, og vis at maximum likelihood estimator MLE for θ blir $\hat{\theta} = \bar{X}_{100}$, dvs. gjennomsnittet av de 100 observasjonene. Sett opp uttrykket for LR_{100}, likelihood ratioen, som er utgangspunktet for likelihood-ratio-testen for H_0, og finn en enkel formel for

$$-2 \log LR_{100}$$

i denne situasjonen.

b) Hva sier den generelle teorien om fordelingen til $-2 \log LR_n$ når n vokser? Hvordan fungerer dette approksimasjonsresultatet for situasjonen i denne oppgaven? Begrunn svaret ditt. Utled likelihood-ratio-testen (finn forkastningsområde) med signifikansnivå α for $H_0 : \theta = 0$ mot $H_a : \theta \neq 0$.

Oppgave 3. La $X_1, ..., X_n$ være et tilfeldig utvalg fra en kontinuerlig fordeling med tetthet $f(x; \theta) = \theta x^{\theta - 1}$, der $0 < x < 1$ og $\theta > 0$.

a) Vis at Fisher-informasjonen i utvalget er n/θ^2. Finn maximum likelihood estimator MLE for θ og kall denne $\hat{\theta}$. Hva blir asymptotisk fordeling for $\hat{\theta}$ når n vokser?

(Fortsettes på side 3.)

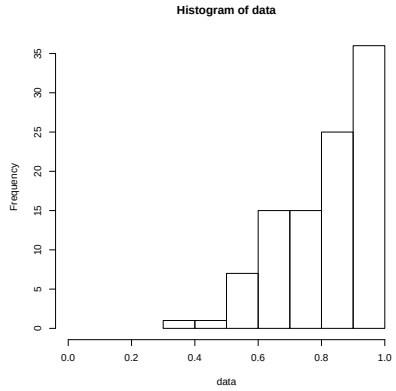

Figure 1: Histogram for 100 observasjoner i oppgave 3

b) Finn et tilnærmet 99% konfidensintervall for θ basert på datasettet gjengitt i histogrammet. De 100 observasjonene fra $f(x;\theta)$ gir

$$\sum_{i=1}^{100} x_i = 81.38, \quad \sum_{i=1}^{100} x_i^2 = 68.35, \quad \sum_{i=1}^{100} \log(x_i) = -22.47,$$

der log betegner den naturlige logaritmen (som i resten av oppgavesettet).

Oppgave 4. New York by er full av italienske restauranter. De varierer i pris og kvalitet. Vi skal studere hvordan pris betalt for en middag på italiensk restaurant i NYC kan forklares av hvor god maten er, kombinert med grad av service og eleganse (dekor). Alle disse tre forklaringsvariablene er fastsatt på en kontinuerlig skala av restaurantkritikere, der lav score er dårlig og høy score er bra. Vi har data fra $n = 168$ restaurantbesøk, og tilpasser en multippel regresjonsmodell i R med følgende resultat:

```
Call:
lm(Price ~ Food + Service + Decor, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8440  -3.7039  -0.1525   3.6218  19.0576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07
Food          1.5556     0.3731   4.170 4.93e-05
Service       0.1350     0.3957   0.341    0.733
Decor         1.8473     0.2176   8.491 1.17e-14
---

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared: 0.617,Adjusted R-squared:  0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

(Fortsettes på side 4.)

Her betegner 'Food' score for kvaliteten på maten.

a) Sett opp modellen som ligger til grunn for analysen. Bruk matriseform. Benytt notasjonen $\mathbf{Y}$ for responsvektor, $\mathbf{X}$ for designmatrisen og $\boldsymbol{\beta}$ for parametervektoren. Angi dimensjonen til alle vektorer/matriser. Formuler de vanlige forutsetningene vi gjør for en slik modell. Forklar kort hva de estimerte koeffisientene sier om sammenhengen mellom responsen og de tilhørende forklaringsvariablene.

b) Vis at minste kvadraters estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ er forventningsrett for $\boldsymbol{\beta}$. Begrunn kort hvorfor $\hat{\beta}_j$-ene blir normalfordelte. Utled et uttrykk for kovariansmatrisen til $\hat{\boldsymbol{\beta}}$, $\text{Cov}(\hat{\boldsymbol{\beta}})$. Du kan bruke at dersom $\mathbf{A}$ er en matrise med konstanter og $\mathbf{V} = \mathbf{AU}$, så er $\text{Cov}(\mathbf{V}) = \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{A}^T$.

c) Utfør, trinn for trinn, en hypotesetest for å teste om kvaliteten på maten (Food), under ellers like omstendigheter (service og dekor), har en signifikant positiv effekt på prisen. Du kan bruke tall fra R-utskriften. Skriv en konklusjon.

d) Multippel R^2 er definert ved

$$R^2 = 1 - \frac{SSE}{SST}$$

der $SSE = \sum_i (Y_i - \hat{Y}_i)^2$ og SST= $\sum_i (Y_i - \bar{Y})^2$. Forklar hvordan denne skal tolkes. Justert R^2, R^2_{adj}, er definert som

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)},$$

der p er antall kovariater i modellen. Diskuter hvorfor denne kan være å foretrekke fremfor R^2 i mange (hvilke?) situasjoner.

e) Hvilke hypoteser testes i siste linje i R-utskriften (F-test), og hvorledes blir konklusjonen?

Takk for innsatsen!