UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2120 — Skisse til løsning/fasit.
Eksamensdag:	Torsdag 7. juni 2012.
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.

Dette er en toveis kontingenstabell med $I = 3$ aldersgrupper og $J = 3$ diagnosekategorier. Homogenitetstest:

$$\mathrm{H}_0 : p_{1j} = \cdots = p_{Ij}, \qquad \forall\ j = 1, \cdots, J.$$

Under H_0 er

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

tilnærmet kji-kvadratfordelt med $(I-1)(J-1)$ frihetsgrader.

Her får vi

$$\mathrm{H}_0 : p_{1j} = p_{2j} = p_{3j} \qquad \text{for } j = 1, 2, 3.$$

Med nivå $\alpha = 0.05$ og 4 frihetsgrader skal vi forkaste H_0 når $\chi^2 > 9.488$. Får opplyst at $\chi^2 = 1.5232$, dvs. ingen grunn til forkastning, kan ikke påstå at det er signifikant forskjell mellom aldersgruppene.

Til info (ikke del av løsning, sjekker f.eks. at det er forventet minst 5 i hver celle):

```
> d=c(7, 11, 22, 8, 17, 36, 9, 24, 56)
> n=matrix(d,ncol=3)
> show(chisq.test(n))
        Pearson's Chi-squared test
data:  n
X-squared = 1.5232, df = 4, p-value = 0.8225
```

(Fortsettes på side 2.)

```
> ni = rowSums(n)
> nj = colSums(n)
> N = sum(n)
> e = ni %o% nj/N
> e
          [,1]      [,2]     [,3]
[1,]  5.052632  7.705263 11.24211
[2,] 10.947368 16.694737 24.35789
[3,] 24.000000 36.600000 53.40000
> n
     [,1] [,2] [,3]
[1,]    7    8    9
[2,]   11   17   24
[3,]   22   36   56
```

Oppgave 2.

a) Modell $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, der $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, eller evt. $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Her er $\mathbf{Y} = (Y_1, \ldots, Y_{84})^T$ 84-dimensjonal responsvektor, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_4)^T$ 5-dimensjonal parametervektor, $\mathbf{X} = \{x_{ij}\}$ er en 84×5-dimensjonal designmatrise, støyleddene $\boldsymbol{\epsilon}$ en 84-dimensjonal vektor, og $\mathbf{I}$ er en 84×84-dimensjonal identitetsmatrise.

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Høy genekspresjon av gen 1 og gen 2 gir lavere bentetthet, mens høy genekspresjon av gen 3 og gen 4 gir høyere bentetthet.

b) Y_i-ene er uavhengige og normalfordelte, $\hat{\beta}_j$-ene er lineærkombinasjoner av slike og blir derfor også normalfordelte.

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

95% konfidensintervall for β_2, koeffisienten for gen 2: t-fordeling med 79 frihetsgrader, bruker konservativt fra tabell $t_{0.025,60} = 2.00$, estimert standardfeil 0.5931 fra utskrift, gir intervall (-3.5137, -1.1413).

(Fortsettes på side 3.)

c) Under H_0 er alle $Y_i \sim N(\beta_0, \sigma^2)$, så med $S^2 = \sum_i (Y_i - \bar{Y})^2/(n-1)$ vet vi at $(n-1)S^2/\sigma^2 \sim \chi^2$-fordelt med $n-1$ frhetsgrader. Men $(n-1)S^2/\sigma^2 = \text{SST}/\sigma^2$ så da har vi $\text{SST}/\sigma^2 \sim \chi^2_{n-1}$.
Bruker $\text{SST}/\sigma^2 = \text{SSE}/\sigma^2 + \text{SSR}/\sigma^2$. Vet at $\text{SST}/\sigma^2 \sim \chi^2_{n-1}$ og $\text{SSE}/\sigma^2 \sim \chi^2_{n-(k+1)}$ (fra formelsamling). Bruker resultatet om summen av to uavhengige χ^2-fordelinger (i formelsamling!) til å vise at SSR/σ^2 må være $\chi^2_{n-1-(n-(k+1))} = \chi^2_k$ under H_0.

d)

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-(k+1))} = \frac{\frac{\text{SSR}}{\sigma^2}/k}{\frac{\text{SSE}}{\sigma^2}/(n-(k+1))}.$$

Bruker tipset til å vise $F \sim F_{k,n-(k+1)}$. Når H_0 er sann, er variasjonen som forklares av modellen liten, dvs. SSR liten. Når H_0 er usann, og modellen predikerer godt, blir SSR stor. Forkaster på nivå α når $F > F_{\alpha,k,n-(k+1)}$. I forhold til data i oppgaven, med $\alpha = 0.05$ finner vi $F_{\alpha,k,n-(k+1)} = F_{0.05,4,79} \approx F_{0.05,4,60} = 2.53$. Fra utskrift har vi F=33.53, og P-verdi≈ 0, forkaster H_0 og konkluderer med at minst en av genene har noe å si for bentetthet, modellen er nyttig.

e) Multippel kvadrert korrelasjon R^2 er andelen av totalvariasjonen i $\mathbf{Y}$ som forklares av modellen. Tall mellom 0 og 1. Kan skrive R^2 =SSR/SST=(SST-SSE)/SST=1-SSE/SST og vise at $F = (n-(k+1))R^2/k(1-R^2)$, dvs. vi kan finne F fra R^2 og omvendt (ekvivalente observatorer).

Oppgave 3.

a) Standard enveis variansanalyse, se bok eller formelsamling.

b) Gitt at $\max_{i_1,i_2} |(\bar{Y}_{i_1\cdot} - \mu_{i_1}) - (\bar{Y}_{i_2\cdot} - \mu_{i_2})|/\sqrt{MSE/J} \sim Q_{I,I(J-1)}$, kan vi sette opp

$$1-\alpha = P(\max_{i_1,i_2} |(\bar{Y}_{i_1\cdot} - \mu_{i_1}) - (\bar{Y}_{i_2\cdot} - \mu_{i_2})|/\sqrt{MSE/J} \le Q_{\alpha,I,I(J-1)}).$$

Hvis den største differansen er mindre, må alle være mindre, så

$$1-\alpha = P(|(\bar{Y}_{i_1\cdot} - \mu_{i_1}) - (\bar{Y}_{i_2\cdot} - \mu_{i_2})|/\sqrt{MSE/J} \le Q_{\alpha,I,I(J-1)} \ \forall\, i_1, i_2).$$

For alle mulige par i_1, i_2 får vi derfor simultane konfidensintervaller for $\mu_{i_1} - \mu_{i_2}$ (med konfidenskoeffisient $(1-\alpha)100\%$ simultant) på formen

$$\bar{y}_{i_1\cdot} - \bar{y}_{i_2\cdot} \pm Q_{\alpha,I,I(J-1)}\sqrt{MSE/J}.$$

Sier at μ_{i_1} og μ_{i_2} er signifikant forskjellige dersom intervallet ikke inneholder 0. Bør nevne underskåringsprosedyren kort (s. 554 i boken).

(Fortsettes på side 4.)

Oppgave 4.

a) Lokasjonsparameter x_m gitt, kun en ukjent parameter, θ.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i|x_m,\theta) \\ &= \theta^n x_m^{n\theta} \prod_{i=1}^{n} x_i^{-\theta-1} \end{aligned} \tag{1}$$

.

$$l(\theta) \propto n\log(\theta) + \theta(n\log(x_m) - \sum_{i=1}^{n}\log(x_i)).$$

Deriverer mhp θ og setter lik 0, får

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n}\log(x_i) - n\log(x_m)}.$$

$$\bar{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right] = \frac{n}{\theta^2}.$$

Når n er stor har vi

$$\hat{\theta} \sim N(\theta, \frac{\theta^2}{n}).$$

b) Har $x_m = 1$. $n = 30$ observasjoner gir MLE $\hat{\theta} = 2.1$. Estimerer $\mathrm{Var}(\hat{\theta})$ med $\frac{\hat{\theta}^2}{n} = 0.147$. Tilnærmet 95% konfidensintervall for θ: $\hat{\theta} \pm 1.96\sqrt{0.147}$, som gir intervallet (1.349, 2.851).

c) Vi skal ha $(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$, der $\bar{\delta}$ og $\underline{\delta}$ er øvre og nedre 2.5% kvantiler i fordelingen til $\hat{\theta}^* - \hat{\theta}$. $\bar{\delta} = 3.172 - 2.1 = 1.072$ og $\underline{\delta} = 1.531 - 2.1 = -0.569$ gir bootstrapintervallet (1.028, 2.669). Vi ser at normaltilnærmingen ikke er spesielt god, da $n = 30$ er for liten.

d) Med lokasjonsparameteren x_m som ukjent parameter får vi

$$l(x_m,\theta) \propto n\log(\theta) + \theta(n\log(x_m) - \sum_{i=1}^{n}\log(x_i)).$$

Denne skal maksimeres m.h.p. x_m. Dette oppnår vi når x_m er størst mulig. Må derfor ha at MLE for x_m er $\hat{x}_m = \min_i x_i$.