Løsningsforslag eksamen STAT100 Høst 2010

Oppgave 1

a) To-utvalg, parvise data. La Y være tilfeldig variabel som angir antall drepte i periode 1 og tilsvarende X for periode 2. Vi antar parvise avhengigheter mellom samme måned i de to periodene. La $D_i = X_i - Y_i$ angi differansen for måned *i* (*i=1,2,…,5=n).* Videre antar vi modellen $D_i \sim N(\mu_D, \sigma_D)$ for differansene. Her er μ_D forventet differanse i antall trafikkdrepte i vintermånedene mellom de to periodene, og σ_D er standardavviket som beskriver variasjonen i slike differanser over vintermånedene. Fra R-utskrift for parvis test finner vi estimater: $\hat{\mu}_D = \bar{D} = 55.2$ og $\hat{\sigma}_D = 14.46$.

b) Vi kan teste med én-utvalgs test på differansene om veisalting har ført til flere trafikkdrepte i P2 enn i P1. Vi ønsker da å teste hypotesene:

$H_0 : \mu_D = 0$ mot $H_1 : \mu_D > 0$

Testobservator for testen er

$T = \frac{\hat{\mu}_D}{s_D/\sqrt{n}} = \frac{55.2}{14.46/\sqrt{5}} = 8.53$

som også kan leses direkte ut fra R-utskriften. Vi forkaster nullhypotesen dersom testobservatoren er større enn $t_{0.05,4} = 2.13$. Her kan vi forkaste og hevde at det har skjedd en økning i forventet antall drepte fra periode P1 til P2.

c) Hypotesene som vi ønsker å teste:

H_0 : Periode og årstid er uavhengige, mot H_1 : De er avhengighet mellom periode og årstid, noe som betyr at endring fra periode P1 til P2 er forskjellig i de to årstidene.

Her utfører vi en kontingenstabell-analyse på

	P1	P2	Sum
vinter	875 (=X_{11})	1151 (=X_{12})	2026 (=R_1)
sommer	1544 (=X_{21})	1407 (=X_{22})	2951 (=R_2)
Sum	2419 (=K_1)	2558 (=K_2)	4977 (=n)

Beregningene kan gjøres manuelt ved at man finner testobservator:

$Q = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$ der $E_{ij} = \frac{R_i K_i}{n}$

Dette gir som i utskriften nedenfor $Q = 40.11$. Denne er nå kji-kvadratfordelt med 1 frihetsgrad. Nullhypotesen forkastes dersom *Q* er større enn $\chi^2_{0.05,1}$ = 3.84. Dette gir

forkastning av nullhypotesen og vi kan hevde at endringen fra P1 til P2 i antall trafikkdrepte er forskjellig i vintermånedene og i sommermånedene. Siden man utfra tallene ser at økningen er større i vintermånedene enn i sommermånedene (som har nedgang…) støtter dette påstanden om at vei-salting har ført til flere drepte i trafikken vinterstid.

R-utskrift fra kontingensanalyse:

```
Pearson's Chi-squared test

data:  .Table

X-squared = 40.1087, df = 1, p-value = 2.402e-10

> .Test$expected # Expected Counts

          1        2

1  984.7085 1041.292

2 1434.2915 1516.708

> round(.Test$residuals^2, 2) # Chi-square Components

      1     2

1 12.22 11.56

2  8.39  7.94
```

Oppgave 2

a) La Y_{ij} være koaguleringstiden for blod for måling nr *j* ved diett nr *i*, der $i = 1, \ldots, 4$ og $j = 1, \ldots, n_i$ og $n_1 = 4, n_2 = 6, n_3 = 6$ og $n_4 = 8$. Vi antar en ANOVA modell for dataene:

$Y_{ij} = \mu_i + \epsilon_{ij}$, der $\epsilon_{ij} \sim N(0, \sigma)$ og uavhengige.

Eventuelt kan vi skrive:

$Y_{ij} \sim N(\mu_i, \sigma)$

Gruppeforventningene estimeres ved gruppegjennomsnittene, dvs

$\hat{\mu}_1 = \bar{y}_1 = 61$, $\hat{\mu}_2 = \bar{y}_2 = 66$, $\hat{\mu}_3 = \bar{y}_3 = 68$ og $\hat{\mu}_4 = \bar{y}_4 = 66$

Mens standardavviket σ som beskriver variasjonen mellom koaguleringstider innen diett estimeres med $\hat{\sigma} = \sqrt{MS_E} = \sqrt{5.6} = 2.37$. Vi finner MS_E som tall nummer (6) i tabellen.

Tallene som mangler i ANOVA-tabellen er gitt nedenfor:

Analysis of Variance Table

Response: coagulation

	Df	Sum Sq	Mean Sq	F value
Diet	**3(k-1)**	228	**76.0 (SSG/3)**	**13.571 (MSG(MSE)**
Residuals	**20(N-k)**	**112 (SST-SSG)**	**5.6 (SSE/20)**	
Total	**23(N-1)**	340		

Mao:

(1) = 3

(2) = 20

(3) = 23

(4) = 112

(5) = 76

(6) = 5.6

(7) = 13.57

b) Testing av diett-effekt: Hypoteser:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, mot H_1 : Minst to diett-forventninger er ulike

Vi forkaster nullhypotesen dersom testobservatoren F=13.57 er stor og for testnivå 5% vil vi forkaste dersom den er større enn $F_{0.05,3,20}$ = 3.10. Vi fokaster nullhypotesen og hevder at minst to av diettene har forskjellig forventet koaguleringstid for blod.

c) Vi definerer en kontrast, θ, for å teste forskjellen mellom diett B og D:

$\theta = \mu_2 - \mu_4 = 0 \cdot \mu_1 + 1 \cdot \mu_2 + 0 \cdot \mu_3 + (-1) \cdot \mu_4$

Denne estimerer vi med

$\hat{\theta} = \bar{y}_2 - \bar{y}_4 = 66 - 61 = 5$

Standardfeilen til denne er

$SE(\hat{\theta}) = \sqrt{MS_E(\frac{1^2}{6} + \frac{(-1)^2}{8})} = 1.28$

Vi tester hypotesen ved å konstruere testobservatoren

$T = \frac{\hat{\theta}}{SE(\hat{\theta})} = \frac{5}{1.28} = 3.91$

Vi forkaster nullhypotesen på 1% nivå dersom |T| > $t_{0.005,20} = 2.845$, hvilket den er, og vi hevder at det er forskjell på diett B og kontroll-dietten D. Se for øvrig R-utskrift:

```
B-D

      Estimate Std. Error  t value     Pr(>|t|) DF lowerCI upperCI

diet         5   1.278019 3.912304 0.0008635834 20 2.334098 7.665902
```

Oppgave 3

a) Dersom vi antar at forskerne antok modellen

$Z_i = \alpha + \beta X_i + \epsilon_i \qquad$ der $\qquad \epsilon_i \sim N(0, \sigma)$

vil et residual defineres som avviket mellom den enkelte observerte verdi og den tilpassede verdien ifølge den estimerte modellen:

$\hat{Z}_i = \hat{\alpha} + \hat{\beta} X_i$

Da blir et residual for observasjon nr *i* lik: $e_i = Z_i - \hat{Z}_i$.

Dersom modellantagelsen om linearitet, og konstant varians for dataene holder, vil et plott av residualer mot tilpassede verdier vise at disse har ingen sammenheng, og at variasjonen er tilsynelatende konstant for residualene for alle verdier av den tilpassede verdien. Residualplottet som er gitt i oppgaven antyder en klar kurvesammenheng samt

økende variasjon med økende verdi av den tilpassede verdien. Dermed har vi brudd på to av de ovenfornevnte modellantagelser og modellen kan ikke sies å være tilfredsstilende.

b) Modell:

$Y_i = \alpha + \beta X_i + \epsilon_i$

Tolkning av modellparametre:

α : Forventet antall skilpadder (på logskala) når X=0 (dvs en øy med areal lik e^0 km^2 = 1 km^2)

β : Forventet endring i antall skilpadder (på logskala) dersom man går fra øy med et gitt areal til en annen øy med areal som er én enhet større (målt i log(km^2)).

σ : Angir standardavviket for antall skilpadder (på log-skala) for øyer av samme størrelse.

Modellantagelsen som vi gjør er:

1) Vi antar en lineær sammenheng mellom Y og X som angitt i modellen
2) Vi antar at $\epsilon_i \sim N(0, \sigma)$, dvs normalfordelte støyledd og med konstant varians (st.avvik)
3) Vi antar at alle støyleddene er uavhengige av hverandre.

c) Fra utskriften finner vi at $\hat{\alpha} = 2.9037$ og at $\hat{\beta} = 0.3886$. Videre estimeres σ som s (=Residual standard error) = 0.7842.

Vi tester om det er sammenheng mellom Y og X ved å formulere hypotesene:

$H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$

Vi forkaster nullhypotesen dersom $|T| > t_{\alpha/2,n-2}$, der

$T = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.3886}{0.0416} = 9.34$

og for testnivå 5% finner vi $t_{0.025,28} = 2.048$

Vi kan dermed hevde en signifikant sammenheng mellom log-areal av øyene og log-antall av skilpadder.

d) For en øy med X=1.55 finner vi en prediksjon på log-antall skilpadder som

$\hat{Y} = 2.9037 + 0.3886 * 1.55 = 3.506$

Et prediksjonsintervall for en øy med areal lik gjennomsnittet (på log-skala) forenkles til

$$\hat{Y} \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n}}$$

Som gir

$$3.506 \pm 2.048 \cdot 0.7842 \sqrt{1 + \frac{1}{30}} = [1.873, 5.139]$$

Se for øvrig R-utskriften:

```
fit      lwr      upr

3.506028 1.873158 5.138899
```

Vi har gitt en prediksjon på log-skal av antall arter på en øy med størrelse X=1.55. Vi kan transformere oss tilbake til antall arter ved å bruke at $Z = e^Y = 2.718^Y$ og plugge inn vår prediksjon som Y. Dette gir oss et predikert antall arter lik:

$\hat{Z} = e^{3.506} = 33.3$