Løsningsforslag til eksamen 11. desember 2008 stat100

Oppgave 1
Dette er parvis sammenligning, da et tvillingpar er et naturlig par.
La D_i være $X_i - Y_i$ der X_i er IQ hos twin A_i og Y_i er IQ hos twin B_i.
i = 1, 2, ...31

Vi antar at D_i-ene er uavhengige og Normalfordelte med forventning μ_d og standardavvik σ_d.

Vi tester
H_0: $\mu_d = 0$ (ingen forskjell på IQ med hensyn på hvor de har vokst opp)
H_0: $\mu_d \neq 0$ (IQ er forskjellig avhengig av om de vokser opp hos biologiske foreldre eller adoptivforeldre)
P-verdi: $P(|\overline{\mathbf{D}}| > 3{,}26)$ dersom H_0 er sann.
Dette er det samme som
P(det observerte eller noe enda mer ekstremt dersom H0 er sann)

$$P(|D| > 3{,}26) = P\left(\left|\frac{\overline{D}}{s_D/\sqrt{n}}\right| > \frac{\mathbf{3{,}26}}{\mathbf{8{,}81/\sqrt{31}}}\right) = P(|T| \geq 2{,}06) \text{ der } T = \frac{\overline{D}}{s_D/\sqrt{n}}$$

Vi ser at dette har en sannsynlighet på 0,048. Altså kan vi forkaste H0 på 5 % nivå, og påstå at oppvekstforhold har betydning for IQ.

Oppgave 2
a) Modell: $Y_i = \alpha - \beta x_i + e_i$ der e_i-ene er uavhengige og $N(0, \sigma)$.
Y_i er avlingnr. i, og x_i er såtid nr i. i = 1, 2, ...14.
Estimater (fra utskrift)
$\hat{\alpha} = 554.5$

$\hat{\beta} = -2{,}82$

$\hat{\sigma} = 21.8$
Dersom vi sår 1, april estimerer vi gjennomsnittsavling til 554,5 kg
Foreventet tap i avling pr sådag utsatt estimeres til 2,82 kg.
Spredning (standardavvik) for avling med samme såtid estimeres til 21,8 kg.
Modellen passer relativt bra, selv om vi kan se en svak krumming på residualplottet. R^2 er relativt stor. Andel variasjon i avling forklart av såtid er 68,9 %.

H_0: $\beta = 0$. (Tidlig) Såtid har ingen betydning
H_1: $\beta < 0$ Tidlig såtid har betydning.

Merk nå x øker så betyr dette at du sår seinere på våren.

$T = \dfrac{\hat{\beta}}{S.E.(\hat{\beta})}$ er t- fordelt med 12 frihetsgrader under H0.

Dermed forkast H0 hvis T < $-t_{\alpha,12}$, er α er valgt nivå

Vi får at T = -5.13, dermed kan vi forkaste på 1 % nivå, (der t $t_{0,01,\,12}$ = 2,681).

Vi påståprat tidligere du sår, jo større vil forventet avling bli.

Vi estimerer denne til 450,26 kg (se utskrift). Et 95 % KI for denne forventningen er (427, 473)

Bredden er gitt ved $2t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Bredde avtar med økende n gjennom frihetsgrader på t.-en, 1/n og kvadratsummen i nevner innenfor rottegnet. Denne kan vi kontrollere

Bredde avtar dersom konfidensgraden avtar

Bredde øker med økende $\hat{\sigma}$. Denne kan vi indirekte kontrollere ved valg av x,

Men ved valgt x er denne låst.

Bredde øker dersom den x-verdien vi skal estimere Y for ligger langt fra gjennomsnittet for de data vi bruket i parameterestimeringen.

En enkelt avling ligger med 95 % sikkerhet innen PI (397, 503)

Oppgave 3

a) Modell: $Y_{ij} \sim N(\mu_i, \sigma)$ der Y_{ij} er kolesterolnivå for prøve j på slankemiddel i.

i = 1, 2, 3.

j = 1, 2 . .n_i

$n_1 = n_3 = 4$, $n_2 = 3$.

$\hat{\mu}_1 = 3{,}3$

$\hat{\mu}_2 = 2{,}76$

$\hat{\mu}_3 = 3{,}25$

$\hat{\sigma} = 0{,}29$

b)

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: minst to er forskjellig.

F = 3,36. Forkast H_0 hvis F > 4,46 (fra tabell over f-fordeling)

F er ikke sto nok. Vi kan ikke forkaste H_0.

Vi har ikke nok bevis for å påstå at det er forskjell i kolesterolnivå på disse tre slankepulverne.

One-way ANOVA: A; B; C

```
Source  DF      SS      MS     F
Factor   2  0,5688  0,2844  3,36
```

```
Error    8  0,6767  0,0846
Total   10  1,2455

Level  N    Mean   StDev
A      4  3,3000  0,2160
B      3  2,7667  0,2082
C      4  3,2500  0,3873

Pooled StDev = 0,2908
```

4a)

La A være variabel røyking/ikke røyking, B er variabel kjønn.

H0: Uavhengighet mellom kjønn og røyking d.v.s.

$P(A_i \cap B_j) = P(A_i)P(B_j)$ for alle 4 par av i og j.

H1: Avhengighet mellom kjønn og røyking

Dette er en Kji-kvadrat-test (se utskrift)

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

       jente   gutt  Total
    1     40     20     60
       30,00  30,00
       3,333  3,333

    2     60     80    140
       70,00  70,00
       1,429  1,429

Total    100    100    200
```

Q = Chi-Sq = 9,524; DF = 1;

Denne må sammenlignes med 3,84 (fra tabell i boka over kji-kvadratfordeling på 5 % med 1 frihetsgrad). Side utregnet Q er større enn 3,84, forkaster vi H0 på 5 % nivå.

Og påstår at det er sammenheng mellom kjønn og andelen røykere.

Merk at testen ga: P-Value = 0,002

4b)

Two-Sample T-Test and CI

```
Sample   N   Mean  StDev  SE Mean
1       20  13,00   4,00     0,89
2       40  11,00   3,00     0,47


Difference = mu (1) - mu (2)
Estimate for difference:  2,00000
95% CI for difference:  (0,15779; 3,84221)
T-Test of difference = 0 (vs not =): T-Value = 2,17  P-Value = 0,034
DF = 58
Both use Pooled StDev = 3,3605
```

La μ_1 være forventet antall sigaretter pr dag for gutter som røyker og La μ_2 være forventet antall sigaretter pr dag for jenter som røyker

Et 95 % KI for $\mu_1 - \mu_2$ er gitt ved `(0,15779; 3,84221)`. Siden 0 ikke er inne i intervallet, kan vi forkaste H0: $\mu_1 - \mu_2 = 0$ og tro på et tosidig alternativ: H1: $\mu_1 - \mu_2 \neq 0$. Vi har altså funnet forskjell mellom kjønn.

c) Siden vi bare har positive verdier i KI, vil vi estimere at forskjell mellom kjønn er positiv. Altså blant røykere vil gutter i gjennomsnitt mer røyke mer enn jenter. Samtidig er det en overvekt av jenter som røyker.
Oppsummert. Hvis man først røyker, har gutter det største problemet, mens dersom vi ser på alle er det blant jenter en finner røykere.