Eksamen i Stat100 - Statistikk
Mandag 11. desember 2006 9.00 - 12.30.
Forslag til løsning

Oppgave 1 Aggressivitet.

For enkelhets skyld lar vi A og A^c betegne hjemmetegnene at en tilfeldig valgt student er henholdsvis "Aggressiv" og "Ikke aggressiv" og F og F^c betegne hjemmetegnene at en tilfeldig valgt student er henholdsvis "Førstefødt" og "Ikke førstefødt".

a) $P(F) = P(F \text{ og } (\underbrace{A \text{ eller } A^c}_{\text{sikker hendelse}})) =$

A og A^c er disjunkte hendelser

$P((\underbrace{F \text{ og } A}_{\text{Disjunkte hendelser}}) \text{ eller } (\underbrace{F \text{ og } A^c}_{\text{Disjunkte hendelser}})) = P(F \text{ og } A) + P(F \text{ og } A^c) =$

$\frac{15}{100} + \frac{25}{100} = 0.15 + 0.25 = \underline{0.40}$ (eller 40%)

b) $P(A|F) = \frac{P(A \text{ og } F)}{P(F)} = \frac{\frac{15}{100}}{0.40} = \frac{0.15}{0.40} =$

$\underline{0.375}$ (eller 37,5%)

c) Vi har (som i punkt a): $P(A) = P(A \text{ og } F) + P(A \text{ og } F^c)$

$= 0.15 + 0.15 = 0.30.$

Dessuten (fra punkt b): $P(A|F) = 0.375$.

Altså: $P(A) \neq P(A|F)$, hvilket betyr at A og F er ikke uavhengige hjemmelagn.

Merknad: Dette kan vises på flere (ekvivalente) måter (for eksempel ved å vise at $0.15 = P(A \text{ og } F) \neq P(A) \cdot P(F) = 0.30 \cdot 0.40 = 0.12$ eller at $0.50 = P(F|A) \neq P(F) = 0.40$).

d) Marginalfordelingene er:

For Aggressivitet: $P(A) = 0.30$ og $P(A^c) = 0.70$.
For Fødselsnummer: $P(F) = 0.40$ og $P(F^c) = 0.60$.

Dersom Aggressivitet og Fødselsnummer hadde vært uavhengige faktorer så ville:

$P(A \text{ og } F) = P(A) \cdot P(F) = 0.30 \cdot 0.40 = 0.12$ (= 12 %)
$P(A \text{ og } F^c) = P(A) \cdot P(F^c) = 0.30 \cdot 0.60 = 0.18$ (= 18 %)
$P(A^c \text{ og } F) = P(A^c) \cdot P(F) = 0.70 \cdot 0.40 = 0.28$ (= 28 %)
$P(A^c \text{ og } F^c) = P(A^c) \cdot P(F^c) = 0.70 \cdot 0.60 = 0.42$ (= 42 %)

Tabellen ville inneholdt de fire prosentene ovenfor.

Oppgave 2 Barns mentale ferdighet

a) $H_0: \mu = 0$ mot $H_1: \mu \neq 0$

Definer $D_i = X_i - Y_i$, $i = 1, 2, \dots, n$ ($n = 15$ her)

Det gir $ED_i = EX_i - EY_i = (\lambda_i + \mu) - \lambda_i = \mu$

Dessuten antar vi at $D_1, D_2, \dots, D_n$ er uavhengige og at D_i er normalfordelt med forventning μ og ukjent varians σ^2.

Testobservator:

$$T = \frac{\bar{D} - 0}{\frac{S}{\sqrt{n}}} = \frac{\bar{D}}{\sqrt{S^2}}\sqrt{n}$$

der $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ og $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$.

Innsatt det tilgjengelige datamaterialet gir dette (idet vi merker oss at $\bar{D} = \bar{X} - \bar{Y} = 39.4 - 18.1 = 21.3$):

$$T = \frac{21.3}{32.82}\cdot\sqrt{15} = 2.51$$

Vi forkaster H_0 hvis $|T| > t_{0.01/2}^{(n-1)} = t_{0.005}^{(14)} = 2.977$,

der $t_\alpha^{(k)}$ er øvre α-fraktil i t fordelingen med k frihetsgrader.

Fordi $|T| = 2.51 \leq 2.977$ kan vi ikke forkaste H_0 med 1% nivå. Det betyr at det tilgjengelige datamaterialet ikke gir oss grunnlag for å påstå at det er en annen tid på en slik bygging i andre forsøk enn i første forsøk.

b) Et 99% konfidensintervall for μ er gitt ved:

$$\left[\bar{D} - t_{0.01/2}^{(n-1)} \cdot \frac{s}{\sqrt{n}},\ \bar{D} + t_{0.01/2}^{(n-1)} \cdot \frac{s}{\sqrt{n}}\right] =$$

$$\left[\bar{D} - t_{0.005}^{(14)} \cdot \frac{s}{\sqrt{15}},\ \bar{D} + t_{0.005}^{(14)} \cdot \frac{s}{\sqrt{15}}\right] =$$

$$\left[14.6 - 2.977 \cdot \frac{15.73}{\sqrt{15}},\ 14.6 + 2.977 \cdot \frac{15.73}{\sqrt{15}}\right] =$$

$$[14.6 - 12.09,\ 14.6 + 12.09] = [2.51,\ 26.69]$$

Fordi intervallet $[2.51, 26.69]$ ikke inneholder 0 må vi nå forkaste H_0 med 1% nivå. Det betyr at det synes å være en annen tid på en slik bygging i andre forsøk enn i første forsøk.

Merknad: Å forandre den galt registrerte tiden 178 til den riktige tiden 78 medfører at konklusjonen på den aktuelle problemstillingen blir annerledes.

c) Fra utlykkene for forventning og varians i kjikvadratfordelingen får vi:

- $E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = n-1 \Leftrightarrow \frac{1}{\sigma^2}E\left(\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = n-1$

  $\Leftrightarrow E\left(\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = (n-1)\sigma^2$

- $\mathrm{Var}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = 2(n-1) \Leftrightarrow \frac{1}{\sigma^4}\mathrm{Var}\left(\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = 2(n-1)$

  $\Leftrightarrow \mathrm{Var}\left(\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = 2(n-1)\sigma^4$

Dette gir:

$$\underline{E\,\hat{\sigma}^2_{(1)}} = E\left(\frac{1}{n}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = \frac{1}{n}(n-1)\sigma^2 = \frac{n-1}{n}\sigma^2 = \underline{\left(1-\frac{1}{n}\right)\sigma^2}$$

$$\underline{\mathrm{Var}\,\hat{\sigma}^2_{(1)}} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = \frac{1}{n^2}2(n-1)\sigma^4 = \underline{2\frac{n-1}{n^2}\sigma^4}$$

$$\underline{E\,\hat{\sigma}^2_{(2)}} = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \underline{\sigma^2}$$

$$\underline{\mathrm{Var}\,\hat{\sigma}^2_{(2)}} = \mathrm{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\bar{D})^2\right) = \frac{1}{(n-1)^2}2(n-1)\sigma^4 = \underline{\frac{2}{n-1}\sigma^4}$$

$\hat{\sigma}^2_{(2)}$ er forventningsrett, i motsetning til $\hat{\sigma}^2_{(1)}$.

Vi ser dog at $E\,\hat{\sigma}^2_{(1)} = \left(1-\frac{1}{n}\right)\sigma^2 \xrightarrow{n\to\infty} \sigma^2$.

$$\frac{\mathrm{Var}\,\hat{\sigma}^2_{(1)}}{\mathrm{Var}\,\hat{\sigma}^2_{(2)}} = \frac{\frac{2(n-1)}{n^2}\sigma^4}{\frac{2}{n-1}\sigma^4} = \frac{(n-1)^2}{n^2} = \left(\frac{n-1}{n}\right)^2 = \left(1-\frac{1}{n}\right)^2$$

Vi ser at

- $\frac{\text{Var } \hat{\sigma}^2_{(1)}}{\text{Var } \hat{\sigma}^2_{(2)}} = \left(1 - \frac{1}{n}\right)^2 < 1$, som gir $\text{Var } \hat{\sigma}^2_{(1)} < \text{Var } \hat{\sigma}^2_{(2)}$

Men, dog $\frac{\text{Var } \hat{\sigma}^2_{(1)}}{\text{Var } \hat{\sigma}^2_{(2)}} = \left(1 - \frac{1}{n}\right)^2 \xrightarrow[n \to \infty]{} 1$

Jeg vil bruke $\hat{\sigma}^2_{(2)}$ fordi den er forventningsrett (i motsetning til $\hat{\sigma}^2_{(1)}$), selv om variansen til $\hat{\sigma}^2_{(1)}$ er litt mindre enn variansen til $\hat{\sigma}^2_{(2)}$. Når antall observasjoner (n) øker vil forskjellene avta, både mellom forventningene til $\hat{\sigma}^2_{(1)}$ og $\hat{\sigma}^2_{(2)}$ og mellom variansene til $\hat{\sigma}^2_{(1)}$ og $\hat{\sigma}^2_{(2)}$.

Oppgave 3 Protein i melk

a) Modell

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad , i = 1, 2, \dots, n \quad (n = 14 \text{ kyr})$$

Estimater (se utskrift fra Minitab)

$\hat{\alpha} = 0.17558$ (Constant)

$\hat{\beta} = 0.024576$ (x)

$\hat{\sigma}^2 = 0.00177$ (MS_E)

Her er $\sigma^2 = \text{Var}(\varepsilon_i)$, idet vi antar at $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ er uavhengige med $\varepsilon_i \sim N(0, \sigma^2)$.

Estimert forventet melkeprotein når melkeproduksjonen er x_0 kaller vi $E\hat{Y}_0$. Da har vi $E\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$.

Med $x_0 = 35$ får vi $E\hat{Y}_0 = 0.17558 + 0.024576 \cdot 35 = 1.03574$
Med $x_0 = 25$ får vi $E\hat{Y}_0 = 0.17558 + 0.024576 \cdot 25 = 0.78998$

Et estimat for den forandringen som forventes i innhold av melkeprotein dersom melkeproduksjonen for en ku øker fra 25 til 35 kg pr. dag er derfor lik $1.03574 - 0.78998 = 0.24576$
$(= (35-25) \cdot \hat{\beta} = 10 \cdot 0.024576)$.

b) Predikert innhold av melkeprotein dersom melkeproduksjonen for en ku er x_0 kaller vi $\hat{Y}_0$.
Da har vi $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$. Med $x_0 = 30$ kg pr. dag:

$$\hat{Y}_0 = 0.17558 + 0.024576 \cdot 30 = 0.91286$$

Et 95% predikasjonsintervall for $\hat{Y}_0$ er gitt ved

$$\hat{\alpha} + \hat{\beta} \cdot 30 \pm t^{(n-2)}_{0.05/2} \cdot SD \sqrt{1 + \frac{1}{n} + \left(\frac{30 - \bar{x}}{SD / SE(\hat{\beta})} \right)^2} =$$

$$0.17558 + 0.024576 \cdot 30 \pm t^{(12)}_{0.025} \cdot \sqrt{0.00177} \cdot \sqrt{1 + \frac{1}{14} + \left(\frac{30 - 29.56}{\sqrt{0.00177} / 0.001523} \right)^2} =$$

$$0.91286 \pm 2.179 \cdot 0.04207 \cdot \sqrt{1 + \frac{1}{14} + 0.0002537} =$$

$$0.91286 \pm 2.179 \cdot 0.04207 \cdot 1.03522 =$$

$$0.91286 \pm 0.09490 = [0.81796, 1.00776]$$

c) Tilpasset modell gir:

$$\hat{EY} = \hat{\alpha} + \hat{\beta} x$$

Fra dette har vi:

$$\frac{\hat{EY}}{x} 100 = \frac{\hat{\alpha}}{x} 100 + \hat{\beta} 100$$

der $\frac{\hat{EY}}{x} 100$ uttrykker (estimert) forventet prosent protein i melk.

Fordi (estimert) forventet prosent protein i melk avhenger av x (avtar når x = melkeproduksjonen øker) synes det ikke riktig uten videre å påstå at melk inneholder en fast forventet prosent protein.