Veiledning eksamen STK1000 7. juni 2004

Oppg. 1

a) Ja: Ved å bruke studentregisteret sikrer hun et (representativt) tilfeldig utvalg. Det er også fornuftig å stratifisere på kjønn når det gjelder holdninger til EU. (Kan også nevne at størrelsen 400+400 på utvalget er OK).

b) Ja:

- n=400 uavhengige enkeltforsøk (studenter)
- registrerer JA eller NEI for hver (kun to muligheter)
- P(JA)=p for hver student
- Y=antall JA av de 400

Da er Y binomisk fordelt (400, p). Samme for X med P(JA)=q. Studentene kan gjerne problematisere f.eks. uavhengighetsantagelsen.

c)

$$\hat{\mathrm{p}} = \frac{\mathrm{Y}}{400}$$

$$\mu_{\hat{\mathrm{p}}} = \frac{1}{400} \cdot \mu_{\mathrm{Y}} = \frac{1}{400} \cdot 400 \cdot \mathrm{p} = \mathrm{p},$$

dvs. estimatoren er forventningsrett.

d) p=0.5. Tilnærming til normalfordeling OK fordi n(=400) er stor. Sjekker np=200 > 10 og n(1-p)=200 > 10 (kan evt. kommentere at binomisk fordeling er symmetrisk når p=0.5).

Så

$$\mathrm{Y} \sim N(\mathrm{np}, \sqrt{\mathrm{np(1\text{-}p)}}),$$

dvs.

$$\mathrm{Y} \sim N(200, 10)$$

og

$$P(Y \geq 220) \approx P(Z \geq 2) = 0.0228$$

e)

$$\widehat{\text{p-q}} = \frac{Y}{400} - \frac{X}{400}$$

$$\mu_{\widehat{\text{p-q}}} = \frac{1}{400} \cdot \mu_Y - \frac{1}{400} \cdot \mu_X = \frac{1}{400} \cdot 400 \cdot p - \frac{1}{400} \cdot 400 \cdot q = \text{p-q},$$

dvs. estimatoren er forventningsrett.

Bruker at X og Y er uavhengige:

$$\sigma^2_{\widehat{\text{p-q}}} = \frac{1}{400^2} \cdot \sigma^2_Y + \frac{1}{400^2} \cdot \sigma^2_X = \frac{1}{400^2}(400 \cdot p(1\text{-}p) + 400 \cdot q(1\text{-}q)) = \frac{1}{400}(p(1\text{-}p) + q(1\text{-}q))$$

og

$$\sigma_{\widehat{\text{p-q}}} = \frac{1}{20}\sqrt{p(1\text{-}p) + q(1\text{-}q)}.$$

Innsatt verdiene får vi $\widehat{\text{p-q}} = -0.15$, dvs. kvinnene er mindre positive enn mennene.

Oppg.2

a) Observasjoner (x_i, y_i) av vekt og pulsendring for student i, $i = 1, ..., 34$. Enkel lineær regresjonsmodell:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

der β_0 er skjæringspunktet med y-aksen og β_1 er stigningskoeffisienten. Vi antar at ϵ_i er uavhengige og normalfordelte $N(0, \sigma)$.

b) 95% konfidensintervall for β_1:

$$-0.34194 \pm t^* 0.0993$$

der $t^* \approx 2.042$ (bruker 30 frihetsgrader i tabellen, egentlig har vi 32 frihetsgrader). Intervallet blir (-0.54, -0.14). Da 0 ikke er med i konfidensintervallet, kan vi konkludere med at β_1 er signifikant forskjellig fra 0 (et 95% konfidensintervall tilsvarer en tosidig test med nivå 0.05).

c) R-Sq er andelen av variasjonen i y som forklares av den lineære regresjonen av y på x. En slik andel på 26.4% sier at modellen ikke er spesielt god.

d) Observasjoner $(x_{i1}, x_{i2}, x_{i3}, y_i)$ av vekt, høyde, kjønn og pulsendring for student i, $i = 1, ..., 34$. Multippel regresjonsmodell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

Vi antar også her at ϵ_i er uavhengige og normalfordelte $N(0, \sigma)$, $i = 1, ..., 34$.

t-fordelingen i dette tilfellet har 34-3-1=30 frihetsgrader (n-p-1). I b) har vi egentlig en t-fordeling med 34-2=32 frihetsgrader (n-2).

e) Kjønn er her den eneste forklaringsvariabelen som er signifikant forskjellig fra 0. Når kjønn er med, kan vi droppe vekt og høyde.

Hvis høyde og vekt er likt, får kvinnene en ekstra pulsendring på 14.5 slag per minutt i forhold til menn (ikke påkrevet å kommentere dette).

I den enkle modellen fikk vi det overraskende resultatet at høyere vekt ga mindre pulsendring. Dette kan forklares med at menn har en tendens til å veie mer enn kvinner, og at mennene i undersøkelsen hadde lavere pulsendring enn kvinnene.