STK1000 H-2016 Løsningsforslag

Alle deloppgaver teller likt i vurderingen av besvarelsen.

Oppgave 1

a) De normalfordelte: $\bar{x}$ og sd for hver gruppe.
De skjevfordelte og de ekstremt skjevfordelte: Median og kvartiler for hver gruppe.
$\bar{x}$ og sd gir gode oppsummeringer av data som er tilnærmet normalfordelte (rimelig symmetriske og med lette haler), jfr regelen for $\bar{x} \pm 2 \cdot sd$ og $\bar{x} \pm 3 \cdot sd$
Median og kvartiler er robuste tall og egner seg for å beskrive data som ikke er symmetriske om midten.

b) Hypoteser og hypotesetester

Gjelder alle:	Alternativ formulering for de normalfordelte:	Alternativ formulering for de skjevfordelte (fordi n er stor og CLT trolig slår inn)	Alternativ formulering for de ekstremt skjevfordelte
H0: Gruppene er like	$\mu_0 = \mu_1$ Eller $\mu_0 - \mu_1 = 0$	$\mu_0 = \mu_1$ Eller $\mu_0 - \mu_1 = 0$	$\text{Rangsum}_0 = \text{Rangsum}_1$
H1: Gruppene er ikke like	$\mu_0 \neq \mu_1$ Eller $\mu_0 - \mu_1 \neq 0$	$\mu_0 \neq \mu_1$ Eller $\mu_0 - \mu_1 \neq 0$	$\text{Rangsum}_0 \neq \text{Rangsum}_1$
	Der μ_0 er forventningsverdien i den normalvektige gruppa (gruppe 0) og μ_1 er forventningsverdien i den overvektige gruppa (gruppe 1)		Der Rangsummene er summen av rangeringene til verdiene i de to gruppene.
	To-utvalgs t-test		Wilcoxon rank sum test

Fordi n er stor, antas det at man kan bruke to-utvalgs t-test både for de normalfordelte dataene og de skjevfordelte dataene, unntatt de 14 ekstremt skjeve.

Disse er trolig for skjeve til at CLT (Sentralgrenseteoremet) har slått inn nok ved denne utvalgsstørrelsen. Wilcoxon rank sum test er tryggest her.

c) Signifikansnivået er den (subjektivt vurdert) maksimalt akseptable sannsynligheten for Type I-feil. P(Type I-feil) = P(Forkaste H_0 | H_0) ≤ 0.05
Med 200 tester og P(Type I-feil) = P(Forkaste H_0 | H_0) = 0.05 i hver test:
Forventer 200*0.05=10 signifikante tester.
Ved tre tester:
P(Type I-feil) = P(Forkaste H_0 | H_0) = P(Minst én H_0 forkastes | Alle 3 H_0 er sanne) =
= 1 – P(Alle H_0 beholdes | H_0) = $1-0.95^3 = 0.1426$:

Oppgave 2

a) Regresjonsmodellen: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma)$, ε_i uavhengige.

Korrelasjonsanalyse for om det er en sammenheng mellom pris og nøyaktighet:
Korrelasjonen i «populasjonen» (den sanne korrelasjonen) er ρ.

H_0: Ingen sammenheng mellom pris og nøyaktighet, $\rho = 0$
H_1: Det er en sammenheng mellom pris og nøyaktighet, $\rho \neq 0$

(Alternativt ensidige hypoteser, der nøyaktigheten øker med prisen)

Regresjonsanalyse for pris og nøyaktighet: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

H_0: Ingen sammenheng mellom pris og nøyaktighet, $\beta_1 = 0$
H_1: Det er en sammenheng mellom pris og nøyaktighet, $\beta_1 \neq 0$

(Alternativt ensidige hypoteser, der nøyaktigheten øker med prisen)

Jeg har valgt tosidige hypoteser fordi det er mest konservativt, og jeg ikke vet noe om verken produksjon av meterstokker, deres nøyaktighet, eller prismodeller som blir brukt.

(Alternativt Jeg har valgt ensidige hypoteser fordi det er grunn til å undersøke om nøyaktigheten øker med pris.)

Både korrelasjonsanalysen basert på Pearson's r, og regresjonsanalysen gir en p-verdi for (H_0: $\hat{\beta}_1 = 0$) på 0.397, som vil gi konklusjonen «Behold H_0» på alle signifikansnivåer under 0.397. Vi beholder derfor H_0, og konkluderer med at det er ingen signifikant sammenheng mellom nøyaktighet og pris.

```
        Pearson's product-moment correlation  p-value = 0.3969

Coefficients:             Estimate Std. Error t value Pr(>|t|)
               pris      0.001752   0.002021   0.867    0.397
```

Oppgave 3

a) Effektmål: Et tall som oppsummerer effekten av (variasjon i) blodsukker på (variasjon i) fødselsvekt. I en regresjonsanalyse er det regresjonskoeffisienten som viser stigningstallet til regresjonslinja (β_1 i regresjonsligningen $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$).
β_1 er stigningstallet til regresjonslinja. Det viser hvor mange enheters forskjell i fødselsvekt (y) som forventes når blodsukkeret (x) øker med en enhet.

Estimat: $\hat{\beta}_1 = 172$

95% konfidensintervall:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S.E.(\hat{\beta}_1)} \sim T_{n-2} \leftarrow T_{200-2} = T_{198}$$

$$P\left(-t_{198,0.025} < \frac{\hat{\beta}_1 - \beta_1}{S.E.(\hat{\beta}_1)} < t_{198,0.025}\right) = 0.95$$

95% KI for β_1: $\hat{\beta}_1 \pm \begin{matrix} 1.984 \\ 1.962 \end{matrix} \cdot S.E.(\hat{\beta}_1)$ (t_{100} / t_{1000}) → [9,335] eller [11,333]

Evt også med $z = 1.96$, hvis det begrunnes.

b) En konfunderende variabel er en variabel som både påvirker responsvariabelen og forklaringsvariabelen i en regresjonsanalyse (common cause), og dermed også påvirker sammenhengen (estimatet for effektmålet) mellom de to variablene. Vi må ha ekspertkunnskap om problemet for å avgjøre om en variabel er en konfounder.

Ja, mors body mass index (bmi) kan sies å være en konfunderende variabel for sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt, fordi:

Her har vi nok opplysninger i oppgaveteksten til å kunne anta at bmi kan påvirke (det målte) blodsukkeret, altså forklaringsvariabelen, og at bmi også kan fødselsvekta (responsvariabelen) gjennom andre mekanismer enn blodsukkeret. I så fall vil estimatet for sammenhengen mellom blodsukker og fødselsvekt være biased/feilaktig, hvis vi ikke tar hensyn til bmi i analysen.

c) Nytt estimat for sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt:
$\hat{\beta}_1 = 94$
Nytt 95% konfidensintervall :

$$t = \frac{\hat{\beta}_1 - \beta_1}{S.E.(\hat{\beta}_1)} \sim T_{n-3} = T_{197}$$

95% KI for β_1: $93.8 \pm \begin{matrix} 1.984 \\ 1.962 \end{matrix} \cdot 87.4$ → [-80, 267] eller [-78, 265]

Nei, det er ikke en signifikant sammenheng mellom mors blodsukkernivå og barnets fødselsvekt.

Begrunnelse 1: Dette tilsvarer en hypotesetest for
H_0: Ingen sammenheng mellom mors blodsukkernivå og barnets fødselsvekt, $\beta_1 = 0$, mot
H_1: Det er en sammenheng mellom mors blodsukkernivå og barnets fødselsvekt, $\beta_1 \neq 0$
i en regresjonsmodell med tre parametere, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, der x_{1i} er blodsukker og x_{2i} er bmi, som vist i den siste utskriften i oppgaven. Der ser vi at p-verdien er 0.28, hvilket betyr at H0 beholdes på nivå 0.05.

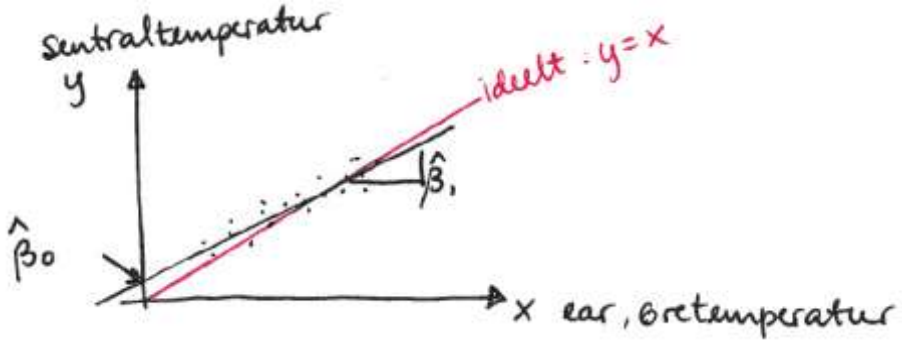
Begrunnelse 2: 95% KI for β_1 fra c) inneholder H_0-verdien β_1=0, og det forteller oss det samme som hypotesetesten, nemlig at vi beholder H_0 (på nivå 0.05).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2964.105    336.907   8.798 7.03e-16 ***
blodsukker    93.763     87.426   1.072   0.2848
bmi           19.883      8.397   2.368   0.0189 *
```

Den signifikante sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt som vi så i oppgave a) forsvinner altså når vi korrigerer for den konfunderende variabelen bmi, og det er derfor grunn til å tro at sammenhengen mellom mors blodsukkernivå og barnets fødselsvekt ikke var reell, men skyldtes konfundering.

Oppgave 4

a) Fra utskriften ser vi at estimatet for parameterne β_0 og β_1 er $\hat{\beta}_0 = 3.7$, og $\hat{\beta}_1 = 0.92$.
Tolkning: $\hat{\beta}_0$ viser hvor regresjonsligninga skjærer y-aksen. Hvis øretermometer og sentraltemperaturen viste det samme (som de ideelt sett burde gjøre), ville denne vært 0. At $\hat{\beta}_0$>0, viser at sentraltemperaturen er litt høyere enn øretemperaturen.
Tilsvarende viser $\hat{\beta}_1$ stigningstallet til regresjonslinja. Igjen, hvis øretermometer og sentraltemperaturen viste det samme (som de ideelt sett burde gjøre), ville denne vært 1.



Hypotesetestene som reflekteres i de to første p-verdiene i utskriften:
H_0: $\beta_0 = 0$, mot
H_1: $\beta_0 \neq 0$ (p-verdi 0.014)
og

H_0: $\beta_1 = 0$, mot
H_1: $\beta_1 \neq 0$ (p-verdi <0.001)

b) 95% prediksjonsintervall for sentraltemperaturen når øretemperaturen viser 38 ºC:

Prediksjonsintervall:

$\hat{y} \pm t_{n-2,\,0.025} \cdot S.E.(\hat{y})$

$\hat{\beta}_0 + \hat{\beta}_1 x^* = 3.75 + 0.92 \cdot 38 = 38.71$

$S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$, $\sum_i (x_i - \bar{x})^2 = (n-1)\cdot(sd_x)^2$

$0.5172 \cdot \sqrt{1 + \frac{1}{237} + \frac{(38 - 37.11)^2}{236 \cdot 0.83^2}} = 0.5195$

$t_{237} \rightarrow 1.984$ hvis $n = 100$
$\rightarrow 1.962$ $n = 1000$

Prediksjonsintervall:

$38.71 \pm \begin{matrix} 1.984 \\ 1.962 \end{matrix} \cdot 0.5195 \rightarrow [37.7, 39.7]$
$\rightarrow [37.7, 39.7]$

c) 95% konfidensintervall for forventet forskjell på de to målemetodene:

$\overline{diff} = 0.78$

$sd_{diff} = 0.52$

95 % KI for δ = differansen på de to metodene

$\frac{\overline{diff} - \delta}{sd_{diff}/\sqrt{n}} \sim T_{237-1} \rightarrow 0.78 \pm \begin{matrix} 1.984 \\ 1.962 \end{matrix} \cdot 0.0338 \rightarrow [0.71, 0.85]$
$\rightarrow [0.71, 0.85]$

Kommentar: Både regresjonsanalysen tidligere i oppgaven og konfidensintervallet viser at det er en statistisk signifikant forskjell på øretemperaturen og sentraltemperaturen, mer spesifikt at sentraltemperaturen (den riktige temperaturen) er høyere enn det øretemperaturen viser. Det er derfor grunn til å være forsiktig med å bruke øretemperatur, spesielt hvis man har med kritisk syke pasienter å gjøre, eller pasienter som ikke tåler å ha høy feber.