Eksamen i STK1110 høsten 2021 - løsningsforslag

Oppgave 1

a
Vi vet at $\bar{D} = 1/n \sum_{i=1}^{17} D_i \sim N(\mu_D, \sigma_D^2/n)$, slik at

$$\frac{\bar{D} - \mu_D}{S_D/\sqrt{17}} \sim t_{17-1},$$

der $S_D = \sqrt{1/(17-1)\sum_{i=1}^{17}(D_i - \bar{D})^2}$ er det empiriske standardavviket. Vi får

$$\begin{aligned}
&\mathrm{P}\left(-t_{0.025,16} \leq \frac{\bar{D} - \mu_D}{S_D/\sqrt{17}} \leq t_{0.025,16}\right) = 0.95 \\
\to &\mathrm{P}\left(\bar{D} - t_{0.025,16}\frac{S_D}{\sqrt{17}} \leq \mu_D \leq \bar{D} + t_{0.025,16}\frac{S_D}{\sqrt{17}}\right) = 0.95.
\end{aligned}$$

Et 95% konfidensintervall for μ_D er dermed gitt ved

$$\bar{d} \pm t_{0.025,16}\frac{s_D}{\sqrt{17}}.$$

Vi setter inn og får $(3.9, 10.9)$. Vi legger merke til at hele konfidensintervallet for forventet vektøking ligger over 0, hvilket er en indikasjon på at på at terapien bidrar til vektøkning.

b
Vi bruker

$$T = \frac{\bar{D} - 0}{S_D/\sqrt{17}} = \frac{\bar{D}}{S_D/\sqrt{17}}.$$

Testbservatoren T er da t_{17-1}-fordelt dersom $\mu_D = 0$. Da den alternative hypotesen er at $\mu_D > 0$, er det naturlig å forkaste H_0 dersom $T \geq c$, der c er slik at signifikansnivået på testen er 5%. Med $c = t_{0.05,16}$ får vi signifikansnivået

$$\begin{aligned}
\mathrm{P}(\text{Forkaste } H_0 \mid H_0 \text{ er sann}) =&\mathrm{P}\left(\frac{\bar{D}}{S_D/\sqrt{17}} \geq t_{0.05,16} \;\middle|\; \mu_D \leq 0\right) \\
\leq&\mathrm{P}\left(\frac{\bar{D}}{S_D/\sqrt{17}} \geq t_{0.05,16} \;\middle|\; \mu_D = 0\right) \\
=&0.05,
\end{aligned}$$

som ønsket. Vi setter inn og får $t_{obs} = \frac{7.26}{/7.16\sqrt{17}} = 4.18$.Da $t_{obs} gleq - t_{0.05,16} = 1.746$, forkaster vi H_0 ved 5% signifikansnivå og konkluderer med at forventet vektøking etter familieterapi er større enn 0, ltså at terapien fungerer. Dette samsvarer godt med resultatene fra Oppgave b).

c

Generelt er P-verdien lik sannsynligheten, beregnet under forutsetning av at nullhypotesen er sann, for at vi vil få en verdi av testobservatoren som er minst like mye i motsetning til nullhypotesen som den verdien vi faktisk fikk. Så her er P-verdien lik sannsynligheten, når $\mu_D = 0$, for at vi vil få en verdi av testobsetvatorene T som er lik $t_{obs} = 4.18$ eller større, dvs

$$\mathrm{P}(T \geq t_{obs} \mid \mu_D = 0).$$

Fra tabellen over kritiske verdier for t-fordelingen ser vi at

$$t_{0.001,16} \leq t_{obs} \leq t_{0.0001,16},$$

hvilket betyr at

$$0.0001 \leq \mathrm{P}(T \geq t_{obs} \mid \mu_D = 0) \leq 0.001,$$

altså at P-verdien ligger mellom 0.0001 og 0.001, og er betydelig mindre enn signifikansnivået 5% fra b).

Oppgave 2

a

Vi finner momentestimatoren $\tilde{\beta}$ ved å løse ligningen vi får når vi setter første teoretiske moment $\mathrm{E}(X_i)$ lik første empiriske moment $\bar{X} = 1/n \sum_{i=1}^{n} X_i$. Forventningen $\mathrm{E}(X_i)$ får vi ved å bruke opplysningen om $\mathrm{E}(X_i^r)$ med $r = 1$. Vi får:

$$\tilde{\beta}\Gamma\left(1 + \frac{1}{\alpha}\right) = \bar{X}.$$

Det gir

$$\tilde{\beta} = \frac{\bar{X}}{\Gamma\left(1 + \frac{1}{\alpha}\right)}.$$

b

I henhold til sentralgrenseteoremet er $\bar{X}$ tilnærmet $N(\mathrm{E}(X_i), \mathrm{V}(X_i)/n)$-fordelt for store n. Da må $\tilde{\beta}$, som er en konstant ganget med $\bar{X}$, også være tilnærmet normalfordelt. Videre er:

$$
\begin{aligned}
\mathrm{E}(\tilde{\beta}) =& \frac{1}{\Gamma\left(1+\frac{1}{\alpha}\right)}\mathrm{E}(\bar{X}) = \frac{1}{\Gamma\left(1+\frac{1}{\alpha}\right)}\mathrm{E}(X_i) = \beta \\
\sigma^2_{\tilde{\beta}} = \mathrm{V}(\tilde{\beta}) =& \frac{1}{\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2}\mathrm{V}(\bar{X}) \\
=& \frac{1}{n\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2}\mathrm{V}(X_i) \\
=& \frac{1}{n\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2}\left(\mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2)\right) \\
=& \frac{1}{n\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2}\left(\beta^2\Gamma\left(1+\frac{2}{\alpha}\right) - \beta^2\left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2\right) \\
=& \frac{\beta^2}{n}\left(\Gamma\left(1+\frac{2}{\alpha}\right)/\Gamma\left(1+\frac{1}{\alpha}\right)^2 - 1\right).
\end{aligned}
$$

c

La $x_1, \ldots, x_n$ være de observerte verdiene av $X_1, \ldots, X_n$. Da er likelihood funksjonen gitt ved

$$
L(\beta) \overset{uif}{=} \prod_{i=1}^{n} f(x_i; \alpha, \beta) = \prod_{i=1}^{n} \frac{\alpha}{\beta^\alpha} x_i^{\alpha-1} e^{-\left(\frac{x_i}{\beta}\right)^\alpha} = \frac{\alpha^n}{\beta^{n\alpha}} e^{-\frac{1}{\beta^\alpha}\sum_{i=1}^n x_i^\alpha} \prod_{i=1}^{n} x_i^{\alpha-1}.
$$

Det gir log-likelihood-funksjonen

$$
\log L(\beta) = n\log(\alpha) - n\alpha\log(\beta) + (\alpha-1)\sum_{i=1}^{n}\log(x_i) - \frac{1}{\beta^\alpha}\sum_{i=1}^{n} x_i^\alpha.
$$

Vi får maksimum likelihood-estimatet $\hat{\beta}$ ved å løse ligningen $\frac{\partial \log L(\beta)}{\partial \beta} = 0$. Altså er $\hat{\beta}$ løsningen av ligningen

$$
-\frac{n\alpha}{\hat{\beta}} + \frac{\alpha}{\hat{\beta}^{\alpha+1}}\sum_{i=1}^{n} x_i^\alpha = 0,
$$

som gir maksimum likelihood-estimatet

$$
\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^\alpha\right)^{1/\alpha},
$$

og tilsvarende maksimum likelihood-estimator

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^{\alpha} \right)^{1/\alpha}.$$

d

Det er kjent at maksimum likelihood-estimatoren $\hat{\beta}$ (under visse regularitetsbetingelser som er oppfylt her) er tilnærmet normalfordelt $N(\beta, \sigma^2_{\hat{\beta}})$ med $\sigma^2_{\hat{\beta}} = 1/(nI(\beta))$. Her får vi

$$\sigma^2_{\hat{\beta}} = 1/(nI(\beta)) = \frac{\beta^2}{n\alpha^2}.$$

Dermed er

$$\frac{\mathrm{V}(\tilde{\beta})}{\mathrm{V}(\hat{\beta})} = \alpha^2 \left(\Gamma\left(1 + \frac{2}{\alpha}\right) / \Gamma\left(1 + \frac{1}{\alpha}\right)^2 - 1 \right) > 1.$$

Da begge estimatorene er tilnærmet normalfordelt med forventning β for store n og variansen til maksimum likelihood-estimatoren $\hat{\beta}$ er minst, vil vi foretrekke denne framfor momentestimatoren $\tilde{\beta}$.

Oppgave 3

a

Enveis variansanalyse dreier seg om sammenligning av forventningsverdien i flere populasjoner, der populasjonene er definert ut fra én bestemt faktor. I dette tilfellet er det type behandling som er faktoren gruppene blir delt inn etter, og vi vil teste om forventet vektøkning i de tre gruppene er den samme eller ikke.

La X_{ij} være vektøkningen til kvinne i i behandlingsgruppe j, der $j = 1$ er kontrollgruppa, $j = 2$ tilsvarer famileterapi og $j = 3$ tilsvarer kognitiv behandling. Vi antar da at

$$X_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2),$$

og at alle X_{ij}-ene er uavhengige. I enveis variansanalyse tester vi hypotesene

$$H_0: \ \mu_1 = \mu_2 = \mu_3 \text{ mot } H_a: \text{minst én } \mu_i \text{ er forskjellig fra de andre.}$$

Vi ser at P-verdien fra F-testen som er brukt er veldig liten (0.0065). Vi forkaster dermed H_0, og konkluderer med at minst én av behandlingene gir en annen forventet vektøkning enn kontrollgruppa.

b

Vi ser av R-utskriften at $\hat{\beta}_0 = -0.45$ og $\hat{\beta}_1 = 5.03$. Vi har:

$$\begin{aligned} \mathrm{E}(Y_i|x_i = 0) =& \mathrm{E}(\beta_0 + \beta_1 \cdot 0 + \varepsilon_i) = \beta_0 + \mathrm{E}(\varepsilon_i) = \beta_0 \\ \mathrm{E}(Y_i|x_i = 1) =& \mathrm{E}(\beta_0 + \beta_1 \cdot 1 + \varepsilon_i) = \beta_0 + \beta_1 + \mathrm{E}(\varepsilon_i) = \beta_0 + \beta_1. \end{aligned}$$

Det betyr at (estimert) forventet vektøkning for kvinnene i kontrollgruppa er $\hat{\beta}_0 = -0.45$ lb, mens (estimert) ekstra forventet vektøkning for kvinnene som har fått behandling, i forhold til de som ikke har fått behandling er $\hat{\beta}_1 = 5.03$ lb.

Et 95% kondidensintervall for β_1 er gitt ved

$$\hat{\beta}_1 \pm t_{0.025,70} s_{\hat{\beta}_1},$$

der $S^2_{\hat{\beta}_1}$ er den forventningsrette estimatoren for $\mathrm{Var}(\hat{\beta}_1)$. Når vi setter tall fra R-utskriften, får vi

$$5.03 \pm 1.994 \cdot 1.879 = (1.28, 8.78).$$

Vi ser at hele konfidensintervallet ligger over verdien 0. Det betyr ar vi med stor sikkerhet vet at forventet vektøkning er større blant kvinnene som får behandling enn de som ikke får det, og altså at behandlingen har effekt på forventet vektøkning.

c

Med den nye modellen

$$Y_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \varepsilon_i^*,$$

får vi

$$\begin{aligned} \mathrm{E}(Y_i|x_{1i} = 0, x_{2i} = 0) =& \mathrm{E}(\gamma_0 + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 + \varepsilon_i^*) = \gamma_0 + \mathrm{E}(\varepsilon_i^*) = \gamma_0 \\ \mathrm{E}(Y_i|x_{1i} = 1, x_{2i} = 0) =& \mathrm{E}(\gamma_0 + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 + \varepsilon_i^*) = \gamma_0 + \gamma_1 + \mathrm{E}(\varepsilon_i^*) = \gamma_0 + \gamma_1 \\ \mathrm{E}(Y_i|x_{1i} = 0, x_{2i} = 1) =& \mathrm{E}(\gamma_0 + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 + \varepsilon_i^*) = \gamma_0 + \gamma_2 + \mathrm{E}(\varepsilon_i^*) = \gamma_0 + \gamma_2. \end{aligned}$$

Det betyr at (estimert) forventet vektøkning for kvinnene i kontrollgruppa er $\hat{\gamma}_0 = -0.45$ lb, mens (estimert) ekstra forventet vektøkning for kvinnene som har fått familieterapi, i forhold til de som ikke har fått behandling er $\hat{\gamma}_1 = 7.72$ lb og (estimert) ekstra forventet vektøkning for kvinnene som har fått kognitiv behandling, i forhold til de som ikke har fått behandling er $\hat{\gamma}_2 = 3.46$ lb. Videre er både β_0 og γ_0 lik E(Y|kvinnen har ikke fått behandling), altså er de like.

d

For å teste hypotesene

$$H_0: \ \gamma_j = 0 \text{ mot } H_a: \ \gamma_j \neq 0,$$

$j = 1, 2$ kan vi bruke testobservatorene $T_j = \frac{\hat{\gamma}_j}{S_{\hat{\gamma}_j}}$, der $S^2_{\hat{\gamma}_j}$ er den forventningsrette estimatoren for $\mathrm{V}(\hat{\gamma}_j)$. Disse testobservatorene er t_{72-3}-fordelt under H_0. Vi har et tosidig alternativ, og vil derfor forkaste H_0 for både store positive og store negative verdier av T_j-ene. Vi forkaster da H_0 dersom $T_j \leq -t_{0.025,69}$ eller $T_j \geq t_{0.025,69}$ med $t_{0.025,69} = 1.995$. Alternativt (men dette er ikke påkrevd her) kan en velge å bruke Bonferroni-korreksjonen med 2.5% signifikansnivå i hver test for å få et samlet signifikansnivå på 5% for de to testene bruker vi, og forkaster da H_0 dersom $T_j \leq -t_{0.0125,69}$ eller $T_j \geq t_{0.0125,69}$ med $t_{0.0125,69} = 2.291$.

Fra R-utskriften har vi $t_{1,obs} = 3.285$, som er i forkastningsområdet, og $t_{2,obs} = 1.700$, som ikke er i forkstningsområdet. Vi forkaster altså nullhyotesen om at $\gamma_1 = 0$, men vi kan ikke forkaste nullhypotesen om at $\gamma_2 = 0$. Med andre ord kan vi konkludere med at familieterapi har en signifikant effekt på vektøkningen til kvinnene, mens vi ikke kan gjøre det samme for den kognitive behandlingen. Hvis vi sammenligner med resultatene fra Oppgave a) og b), ser det ut til at vektøkningen til de som har fått familieterapi skiller seg fra vektøkningen til de andre kvinnene, og at det er det som slår ut i hypotesetestene i a) og b).