LØSNINGSFORSLAG TIL STK1000-EKSAMEN H09

Oppgave 1

(1a). Man randomiserte for å få gruppetilordningen mest mulig objektiv. Mennene i forsøket meldte seg frivillig, så dette er ikke en SRS selv om gruppetilordningen var tilfeldig. Dette gjør at man må være forsiktig med å komme til veldig bastante konklusjoner om fiskedietts innvirkning på *alle* menn med høyt blodtrykk. Strengt talt kan man bare bruke undersøkelsen til å komme frem til konklusjoner om disse 14 mennene, og ikke si at fiskeoljedietter *forårsaker* blodtrykkreduksjon. Her kan man muligens dra inn s 178 i boken hvor man diskuterer når det er rimelig å kunne komme med påstander om kausalitet selv når man ikke har et riktig designet forsøk.

Gjennomsnittsreduksjon for diett med fiskeolje er $\bar{x} = 6.57$, mens for vanlig olje er gjennomsnittet $\bar{y} = -1.14$. Standardavvikene er $s_x = 5.85$ for fiskeolje og $s_y = 3.19$ for vanlig olje.

Det virker som at fiskeoljen har en positiv effekt på reduksjon, men at det ikke er noe særlig forskjell på blodtrykk før og etter diett med vanlig olje.

(1b). I begge konfidensintervallene har man $7 - 1 = 6$ frihetsgrader, som gir en t^* verdi på 2.447 fra tabell D i boken. Altså får man et 95% konfidensintervall på

$$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}} = 6.57 \pm 2.447 \frac{5.85}{\sqrt{7}} = [1.16, 11.99] \tag{1}$$

for forventet reduksjon for diett med fiskeolje og

$$\bar{y} \pm t^* \frac{s_y}{\sqrt{n}} = -1.14 \pm 2.447 \frac{3.19}{\sqrt{7}} = [-4.09, 1.80] \tag{2}$$

for forventet reduksjon for diett med vanlig olje. Tosidige hypotesetester med fastsatt forkastningsnivå på 5% og 95% konfidensintervaller er det samme: Man kan derfor teste

$$H_0 : \mu_X = 0 \qquad \text{mot } H_A : \mu_X \neq 0$$

hvor X er reduksjon med fiskeoljediett, ved å sjekke om 0 er med i konfidensintervallet i likning (1). Det er det ikke, altså forkastes H_0 på et 5%s nivå. Tilsvarende testes

$$H_0 : \mu_Y = 0 \qquad \text{mot } H_A : \mu_Y \neq 0$$

hvor Y er reduksjon for diett med vanlig olje, ved å sjekke om 0 er med i konfidensintervallet i likning (2). Siden null er med forkastes ikke H_0 på et 5%s nivå.

Hver hypotestetest antar at observasjonene er uavhengige og identisk fordelte med normalfordelingen. Siden antall observasjoner ikke er stor, trenger man å anta normalitet, og kan ikke lene seg på sentralgrenseteoremet.

(1c). Man bruker en two-sample t-test, og den tilnærmingen til frihetsgrader man blir bedt om å bruke for manuell regning er $n - 1 = 6$ (her her n_1 og n_2 like, så man får tilfeldigvis likt antall frihetsgrader som i b).

Man tester

$$H_0 : \mu_X = \mu_Y$$

og det kan rimeliggjøres med både tosidige eller ensidige alternativer. Det er klart man egentlig er interessert i om man får mer reduksjon med fiskeolje enn med vanlig olje, men så lenge man argumenter fornuftig for hvilken alternativ hypotese man bruker er begge greit. Hvis man er interessert i en ensidig alternativ hypotese vil det si

$$H_A : \mu_X > \mu_Y,$$

som vil si $\mu_X - \mu_Y > 0$, så det er da viktig at man bruker $\bar{x} - \bar{y}$ i teststatistikken.

Teststatistikken er

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{n}}} = \frac{6.57 - (-1.14)}{\sqrt{\frac{34.29}{7} + \frac{10.14}{7}}} = 3.06.$$

Tabell D viser at $1\% < P(T > 3.06) < 2\%$, så man man får en P-verdi mellom 2% og 4% for et tosidig alternativ, og mellom 1% og 2% for et ensidig alternativ. Man forkaster H_0 på et 5% nivå for begge valgene: Det virker som at det *er* en forskjell mellom forventet reduksjon for de to diettene.

Man antar at observasjonene er uavhengige og med samme fordeling, og siden n er såpass liten må man anta at observasjonene er normalfordelte. Igjen er ikke antall observasjoner stort nok til at man kan lene seg på sentralgrenseeffekten, og trenger normalitetsantagelsen.

Oppgave 2

(2a). Gjennomsnittene kommer fra randomiserte forsøk som ble utført hver for seg. Det er derfor rimelig å anse hvert gjennomsnitt som uavhengig av de andre.

Det er i utgangspunktet ingen grunn til at antall blomster per plante skal ha en fordeling nær normalfordelingen. Derimot vil gjennomsnitt alltid være nærmere normalfordelingen enn variablene selv, takket være sentralgrenseteoremet.

Siden dette er et randomisert forsøk er det rimelig å kunne fastsette kausale koblinger, så man kan si hvilken lyskonfigurasjon (av de man prøvde) som *forårsaker* flest blomster per planter.

(2b). Standardavviket til et gjennomsnitt av variable som har standardavvik σ er $\sigma/\sqrt{n}$. Dette står som en formel i boken.

For å halvere standardavviket til et gjennomsnitt må man ha fire ganger flere så mange observasjoner. Dette står i boken, og man kan også lett regne dette ut.

(2c). Regresjomsmodellen er gitt ved

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$

der ϵ_i er uavhengige og normalfordelte tilfeldige variable forventning null og standardavvik σ. Modellens parametere er $\beta_0, \beta_1, \beta_2$ og σ, og har estimater $b_0 = 71.306$, $b_1 = -0.040471$, $b_2 = 12.158$ og $S = 6.44107$.

(2d). Det er ingenting som tyder på at modellantagelsene er tydelig brutt. Histogrammet av resudalene ser kanskje ikke veldig klokkeformet ut, men kvantilplottet viser at det ikke er et et stort avvik fra normalitet. Kryssplottet på figur 2 viser dessuten at antagelsen om lineæritet ser tydelig ut til å være opprettholdt, og at vi får forklart systemet i datasettet godt med regresjonsmodellen.

R^2 er andelen forklart variasjon av regresjonsmodellen. Jo høyere R^2 er, jo bedre tydet det på at regresjon gikk hvis modellantagelsene ikke er brudt. En R^2-verdi på 79.9 er i denne sammenhengen høyt, og tyder på at den tilpassede modellen forklarer mye av variasjonen i datasettet.

(2e). Hypotesetestene som utføres i Minitab-tabellen er om regresjonskoefisientene β_0, β_1 og β_2 er null mot tosidige alternativer. Altså testes

$$H_0 : \beta_j = 0 \text{ mot } H_A : \beta_j \neq 0$$

for $j = 0, 1, 2$.

Alle resulterende P-verdier er tilnærmet null, så både lysintensitet og starttid har en statistisk signifikant innflytelse på gjennomsnittlig antall blomster per plante.

P-verdiberegningene baserer seg på t-fordelingen med $n-p-1 = n-3 = 24-3 = 21$ frihetsgrader.

For å regne ut et 95% konfidensintervall for β_2 bruker man en t^*-verdi på 2.080 fra tabell D, og får

$$b_2 \pm t^* \text{SE}_{b_2} = 12.158 \pm 2.080 \times 2.630 = (6.69, 17.63).$$

(2f). Vår tilpassede modell er

$$\hat{Y}_i = 71.306 - 0.040471 \times \text{Lysintensitet} + 12.158 \times \text{ Starttid},$$

som vil si at sterk lysintensitet minker antall blomster per plante og at man venter å få rundt 12 flere blomster per plante hvis man starter lysbehandling 24 dager før FFI i forhold til hvis man hadde startet lysbehandling ved FFI. Plantene får flest blomster når man har lav lysintensitet og starter lysbehandlingen 24 dager før FFI.

Fra kryssplottet i figur 2 og regresjonstilpassingen diskutert over burde man i hvertfall prøve lavere lysintensiteter ved et nytt forsøk. Siden plantene hvor lysbehandling ble startet 24 timer før FFI gjennomgående fikk gjennomsnittlig flere blomster per plante virker det rimelig å starte behandlingen i et nytt forsøk 24 dager før FFI, og kanskje også andre tidskonfigurasjoner enda lenger før FFI.

Den laveste lysintensiteten man prøvde var 150, og man burde undersøke hvor langt nedover man fortsetter å få bedre og bedre effekt. Men det er en klar grense: hvis man har null lysintensitet så blomster nok ingen av plantene. Det virker derfor urimelig å tro at det lineære systemet fortsetter så mye lenger enn i det området av lysintensiteter vi har sett på. Dessuten er det ikke opplagt at det må være slik at plantene som fikk behandling startet ved FFI må fortsette å være systematisk lavere enn de som fikk

behandling startet 24 timer før FFI også ved lave lysintensiteter etter ikke-lineæritet starter. Hvis man har nok ressurser kan det virke rimelig å prøve lave lysintensiteter, og samtidig også prøve forskjellige starttidene.