UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Tirsdag 30. november 2021
Tid for eksamen:	15:00 – 19:00
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Alle hjelpemidler er tillatt, men det er ikke tillatt å kommunisere eller samarbeide med andre.

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

Oppgavesettet har fire oppgaver som til sammen består av ti deloppgaver. Hver deloppgave teller likt.

Oppgave 1 Passe bred

1a

Anta at $Y_1 \sim N(\mu_1,\ \sigma_1)$ og $Y_2 \sim N(\mu_2,\ \sigma_2)$ er uavhengige og normalfordelte variabler. Da er

i) summen $Y_1 + Y_2$ normalfordelt, og

ii)
- forventningsverdien til summen $Y_1 + Y_2$ er $\mu_{Y_1+Y_2} = \mu_1 + \mu_2$
- variansen til summen $Y_1 + Y_2$ er $\sigma^2_{Y_1+Y_2} = \sigma_1^2 + \sigma_2^2$, og
- standardavviket til summen $Y_1 + Y_2$ er $\sigma_{Y_1+Y_2} = \sqrt{\sigma_1^2 + \sigma_2^2}$

1b

Det er gitt at $\sigma_1 = \sigma_2 = 0,05$. Standardavviket til realisert bredde Z er $\sigma_Z = \sigma_{M+\epsilon_2} = \sqrt{\sigma_M^2 + \sigma_{\epsilon_2}^2} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{2} \cdot 0,05 = 0,07$.

(Merk at Z ikke er standard normalfordelt selv om den heter 'Z'. Det er selvfølgelig like fint om en student velger å døpe om Z til feks R, dersom de foretrekker å reservere bokstaven Z til en standard normalfordelt variabel.)

1c

Du skulle skjære hylleplata for å sette den i et skap med innvendig bredde $80,00$ cm.

(Fortsettes på side 2.)

i) Realisert bredde Z er normalfordelt med forventning $\mu_Z = B$ og standardavvik $\sqrt{2} \cdot 0,05$, så når $B = 80,00$ er

$$P(Z < B) = P(Z < \mu_Z) = P(\tilde{Z} < 0) = 0,5 \tag{1}$$

for en standard normalfordelt variabel $\tilde{Z}$, eller rett og slett fordi alle normalfordelinger er symmetriske om forventningsverdien.

ii) Vi ønsker å finne B slik at 95% persentilen til Z er $80,0$. Vi vet at 95% persentilen til Z uttrykt ved B er

$$Z_{0,95} = B + z_{0,95} \cdot \sigma_Z \tag{2}$$

der $z_{0,95}$ er 95% persentilen til en standard normalfordelt variabel. Når vi setter uttrykket for 95% persentilen til Z (høyre side i ligning 1) lik $80,0$ og løser for B får vi

$$B = 80,00 - z_{0,95} \cdot \sigma_Z = 80,00 - 1.644854 \cdot \sqrt{2} \cdot 0.05 = 79.88. \tag{3}$$

Oppgave 2 Par av tilfeldige variabler

Anta at du har 100 par av variabler, (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_{100}, Y_{100}), der alle de 200 tilfeldige variablene er uavhengige, og $X_j \sim N(\mu_X, \sigma)$, $Y_j \sim N(\mu_Y, \sigma)$.

2a

I denne deloppgaven antar vi $P(X_j > Y_j) = 0.5$ for hvert par j

i) Definer for hvert par 'suksess' dersom det første elementet er det største.

Antall par N der det første elementet er det største, er da antall suksesser over 100 uavhengige forsøk, der hvert forsøk har suksess-sannsynlighet 0.5.

N er dermed binomisk fordelt $N \sim Bin(n = 100, p = 0.5)$.

ii) Forventa antall par der det første elementet er det største er $\mu_N = n \cdot p = 100 \cdot 0.5 = 50$

iii) A: R-kommandoen 'pbinom(39,size=100, prob=0,5)' angir kumulativ andel $0,01760$

B: Normalapproksimasjon med kontinuitetskorreksjon $P(N < 40) = P(Z < \frac{39,5-50}{\sqrt{100 \cdot 0,5 \cdot 0,5}}) = P(Z < -2,1) = 0,01786$

C. Det er mulig å bruke programvare for å beregne sannsynligheten ved bruk av det matematiske uttrykket

$$P(N = i) = \binom{100}{i} \cdot 0,5^{100}. \tag{4}$$

(Fortsettes på side 3.)

Ved å utnytte at $P(N=j)=P(N=100-j)$, kan man redusere antall ledd, og feks beregne

$$P(N<40)=\frac{1-P(N=50)}{2}-\sum_{j=40}^{49}P(N=j)=0,01760 \quad (5)$$

2b

I denne og den neste deloppgaven, antar vi ikke lenger at $\mu_X=\mu_Y$

i) Den parvise differansen X_j-Y_j i hvert par er normalfordelt, slik at $X_j-Y_j\sim N(\mu_X-\mu_Y,\sqrt{2}\cdot\sigma)$.

ii) Differansen mellom utvalgsgjennomsnittene $\bar{X}-\bar{Y}$ er normalfordelt $N(\mu_X-\mu_Y,\sqrt{2}\cdot\frac{\sigma}{10})$ fordi variansen

$$\sigma^2_{\bar{X}-\bar{Y}}=\sigma^2_{\bar{X}}+\sigma^2_{\bar{Y}}=\sigma^2_X/n_X+\sigma^2_Y/n_Y=\frac{\sigma^2+\sigma^2}{100}=2\cdot\left(\frac{\sigma}{10}\right)^2 \quad (6)$$

2c

Vi velger å gjennomføre en statistisk hypotesetest på $\alpha=5\%$ signifikansnivå. Nullhypotese: $H_0:\ \mu_X=\mu_Y$, (tosidig) alternativ-hypotese $H_0:\ \mu_X\neq\mu_Y$.

Under H_0 er $\mu_X=\mu_Y$ og standardisert testobservator er

$$T=\frac{\overline{X-Y}-(\mu_X-\mu_Y)}{s_{\overline{X-Y}}}=\frac{\overline{X-Y}-0}{s_{X-Y}/\sqrt{100}}=\frac{\overline{X-Y}}{s_{X-Y}/10} \quad (7)$$

Setter inn $\bar{X}=10,02$, $\bar{Y}=10,98$ og $s_{X-Y}=4,73$, og får

$$T=\frac{10,02-10,98}{4,73/10}=-2,029598. \quad (8)$$

T-fordelinga med 99 frihetsgrader gir da tosida P-verdi $2\cdot 0,0225=0,045$.

Med statistisk signifikansnivå $\alpha=5\%$, har vi observert en p-verdi $0,045$ som er mindre enn signifikansnivået α.

Vi forkaster nullhypotesen til fordel for alternativhypotesen, og konkluderer at forventningsverdiene μ_X og μ_Y er ulike.

Oppgave 3 Forhold mellom kroppsmål

3a

i) En enkel lineær regresjonsmodell for responsvariabel kroppslengde (y_i) og forklaringsvariabelen fot.navle (x_i), er

$$y_i=\beta_0+\beta_1\cdot x_i+\epsilon_i,\quad i=1,2,\cdots,223 \quad (9)$$

der individuell variasjon $\epsilon_i\sim N(0,\sigma)$ er uavhengig og normalfordelt for hvert av individene.

Modellantagelsene er dermed:

(Fortsettes på side 4.)

1) Forventa kroppslengde $\mu_{y_i} = \beta_0 + \beta_1 \cdot x_i$ er **lineært** avhengig av verdien på forklaringsvariabelen fot.navle x_i.
2) Spredninga i kroppslengde y_i er gitt ved standardavvik σ i underpopulasjonen der fot.navle har verdi x_i, for enhver verdi av x_i; med andre ord, **spredninga** om forventningsverdien **varierer ikke med forklaringsvariabelen** x_i.
3) Leddene for individuell variasjon, ϵ_i, $i = 1, 2, \cdots, 223$ er **uavhengige**. Ekvivalent: gitt verdiene til fot.navle-målene x_i for hvert individ, er kroppslengdene y_i uavhengige, og
4) Leddene for individuell variasjon, ϵ_i, $i = 1, 2, \cdots, 223$ er **normalfordelte**. Ekvivalent: Gitt verdien til fot.navle-målet x_i for individ i, er kroppslengden y_i normalfordelt.

ii)
- Parameteren β_0 er konstantleddet i lineærmodellen, og beskriver forventa kroppslengde for et individ med fot.navle-mål $x_i = 0$.
- Parameteren β_1 er stigningstallet i lineærmodellen, og beskriver forventa antall enheter (cm) økning i kroppslengde y_i når fot.navle-målet øker med én enhet (cm).
- Parameteren σ beskriver spredninga i kroppslengde y_i om forventningsverdien μ_{y_i}, der μ_{yi} bestemmes av forklaringsvariabelen x_i.

iii) Estimatet til parameteren β_0 er $b_0 = 38.9$. Estimatet til parameteren β_1 er $b_1 = 1.27$. Estimatet til parameteren σ er $s = 3.6$.

3b

i) Et 95% konfidensintervall for β_1 er gitt ved $b_1 \pm t^* SE_{b_1}$, der t^* er 97.5%-persentilen til t-fordelinga med 221 frihetsgrader; $t^* = 1,970756$.

$b_1 - t^* SE_{b_1} = 1,197651$ og $b_1 + t^* SE_{b_1} = 1,347389$.

Et 95% konfidensintervall for β_1 er dermed $[1.20,\ 1.35]$.

ii) predict(fit, newdata = data.frame(fot.navle=104.0), interval= 'confidence', level= 0.95)

iii) predict(fit, newdata = data.frame(fot.navle=104.0), interval= 'predict', level= 0.95)

iv)
- Et 95% konfidensintervall for β_1 er et intervall av mulige verdier, i samsvar med dataene, for stigningstallet (forventa antall enheter (cm) økning i kroppslengde y_i per enhet økning i fot.navle-målet). Videre er konfidensintervallet er konstruert med en metode som i 95% av tilfellene metoden blir brukt, vil konstruere et intervall som inneholder den sanne verdien av populasjonsparameteren β_1.
- Et 95% konfidensintervall for forventa kroppslengde μ_y av et individ med fot.navle lik 104, er et intervall av mulige verdier, i samsvar med dataene, for forventa kroppslengde for et tilfeldig

(Fortsettes på side 5.)

individ med angitt verdi for forklaringsvariabelen fot.navle lik $104,0$. Videre er konfidensintervallet konstruert med en metode som i 95% av tilfellene metoden blir brukt, vil konstruere et intervall som inneholder den sanne forventningsverdien μ_y i underpopulasjonen av individer med fot.navle-mål $x_i = 104,0$.

– Et 95% prediksjonsintervall for kroppshøydemål y_i når fot.navle-målet er 104.0, er et intervall av mulige verdier av responsvariabelen kroppshøydemål for individer med fot.navle-mål lik $104,0$cm. Videre, er et 95% prediksjonsintervall konstruert med en metode som i 95% av tilfellene metoden blir brukt, vil konstruere et intervall som inneholder responsverdien y for en tilleggsobservasjon med kjent verdi av forklaringsvariabelen fot.navle-mål x. Prediksjonsintervallet inkluderer variabiliteten i en fremtidig observasjon om forventningsverdien i underpopulasjonen av individer med fot.navle-mål lik $x_i = 104,0$, og prediksjonsintervallet for kroppslengde blir derfor bredere enn konfidensintervallet for forventa kroppslengde.

Oppgave 4 Snø til jul

Sjansen for snø til jul for Frank varierer med kalenderår og om han er på hytta.

4a

En logistisk regresjonsmodell for sammenhengen mellom responsvariabelen 'snø til jul' (y) og forklaringsvariablene kalenderår (X_1) og 'jul på hytta' (X_2), er

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} \tag{10}$$

der p_i er sannsynligheten $P(y_i = 1)$ for snø til jul for individ i med kjent verdi x_1, x_2 for forklaringsvariablene. Videre er

- β_1 er den naturlige logaritmen av odds-ratioen for snø til jul for et års økning i kalenderår
- β_2 er den naturlige logaritmen av odds-ratioen for snø til jul for jul på hytta sammenligna med i hovedstanden.

4b

Et 95% kondensintervall for parameteren β_2 er gitt ved formelen

$$b_0 \pm z_{0,975} \cdot SE_{b_0} \tag{11}$$

(Fortsettes på side 6.)

der $z_{0,975}$ er 97.5% persentilen til en standard normalfordelt variabel. Det gir at

$$[0.76152 - 1.96 \cdot 0.84862,\ 0.76152 + 1.96 \cdot 0.84862] = [-0.9017752,\ 2.424815] \tag{12}$$

er et 95% kondensintervall for parameteren β_2.

Siden β_2 er den naturlige logaritmen av odds-ratioen for snø til jul for jul på hytta sammenligna med i hovedstanden, er et 95% kondensintervall for odds-ratioen for snø til jul på hytta sammenligna med hjemme for Frank for et gitt kalenderår bestemt av 95% konfidensintervallet til parameteren β_2, som

$$e^{b_0 \pm z_{0.975} \cdot SE_{b_0}}. \tag{13}$$

Et 95% kondensintervall for odds-ratioen for snø til jul på hytta sammenligna med hjemme, er

$$[e^{-0.9017752},\ e^{2.424815}] = [0.41,\ 11.30]. \tag{14}$$

Vi konkluderer med at Frank har en høyst usikker effekt av å reise på hytta, med tanke på sannsynligheten for snø til jul.