UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse
Eksamensdag:	Fredag 3. desember 2021
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 5 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Alle hjelpemidler tillatt

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

Nedenfor er det gitt kritiske verdier $t_{\alpha,\nu}$ for t-fordelingen med ν frihetsgrader for noen verdier av α og ν. Du vil få bruk for tabellen i Oppgave 1 og 3.

α :	0.05	0.025	0.0125	0.01	0.005	0.001	0.0001	0.00001
$t_{\alpha,15}$	1.753	2.131	2.490	2.602	2.947	3.733	4.880	6.109
$t_{\alpha,16}$	1.746	2.120	2.473	2.583	2.921	3.686	4.791	5.959
$t_{\alpha,17}$	1.740	2.110	2.458	2.567	2.898	3.646	4.714	5.832
⋮				⋮				⋮
$t_{\alpha,69}$	1.667	1.995	2.291	2.382	2.649	3.213	3.929	4.580
$t_{\alpha,70}$	1.667	1.994	2.291	2.381	2.648	3.211	3.926	4.576
$t_{\alpha,71}$	1.667	1.994	2.290	2.380	2.647	3.209	3.923	4.571

Oppgave 1

17 unge kvinner blir lagt inn til behandling mot anorexia. De ble veid ved innleggelsestidspunktet, samt etter en periode med familieterapi. Et utdrag av før- og ettermålingene av vekt i lb (pund), samt differansen mellom de to, er vist under.

```
Før     Etter   Differanse
83.8    95.2    11.4
83.3    94.3    11.0
86.0    91.5     5.5
  .       .       .
  .       .       .
  .       .       .
89.9    93.8     3.9
86.0    91.7     5.7
87.3    98.0    10.7
```

La D_i differansen mellom vekten til kvinne nummer i etter terapien og før terapien, altså vektøkningen fra første til siste måling. Det antas at

(Fortsettes på side 2.)

$D_1, \ldots, D_{17} \overset{uif}{\sim} N(\mu_D, \sigma_D^2)$. En ønsker å undersøke om terapien virker ved å lage et konfidensintervall for μ_D, samt ved å test hypotesene

$$H_0: \ \mu_D \leq 0 \text{ mot } H_a: \ \mu_D > 0.$$

a

- Utled et 95% konfidensintervall for forventet vektøkning μ_D etter terapien.
- Beregn intervallet for dataene over, når du får vite at observert snitt og standardavvik for vektdifferensene er $\bar{d} = (1/17)\sum_{i=1}^{17} d_i = 7.26$ og $s_d = \sqrt{(1/16)\sum_{i=1}^{17}(d_i - \bar{d})^2} = 7.16$.

b

-Utled en hypotesetest med signifikansnivå 5% for hypotesene over vedrørende effekten av terapien på vektøkningen.
- Hva blir konklusjonen ut fra de observasjonene som ble gjort?

c

- Forklar hva en P-verdi generelt betyr, altså definér hva den er.
- Finn et uttrykk for P-verdien til testen i punkt c).
- Bruk tabellen over kritiske verdier for t-fordelingen til å si noe om størrelsesorden for P-verdien.

Oppgave 2

La $X_1, \ldots, X_n$ være uavhengige stokastiske variabler fra Weibull-fordelingen, dvs. med sannsynlighetstetthet

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}, \ x > 0,$$

med $0 < \alpha$ og $0 < \beta$. I hele oppgaven antas det at α er kjent og at $\alpha > 1$, mens β er ukjent. Videre får du oppgitt at for en Weibull-fordelt variabel er $\mathrm{E}(X^r) = \beta^r \Gamma\left(1 + \frac{r}{\alpha}\right)$ for en hvilken som helst $r > 0$ når $\alpha > 1$, der $\Gamma(\cdot)$ er gammafunksjonen.

a

Vis at momentestimatoren for β er $\tilde{\beta} = \frac{\bar{X}}{\Gamma\left(1+\frac{1}{\alpha}\right)}$, der $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

b

Bruk sentralgrenseteoremet til å vise at $\tilde{\beta}$ er tilnærmet $N(\beta, \sigma_{\tilde{\beta}}^2)$-fordelt for store n, med $\sigma_{\tilde{\beta}}^2 = \frac{\beta^2}{n}\left(\Gamma\left(1 + \frac{2}{\alpha}\right)/\Gamma\left(1 + \frac{1}{\alpha}\right)^2 - 1\right)$.

c

Sett opp likelihood- og log-likelihood-funksjonen, og finn et uttrykk for maksimum likelihood-estimatoren $\hat{\beta}$.

(Fortsettes på side 3.)

d

- Begrunn at $\hat{\beta}$ er tilnærmet $N(\beta, \sigma^2_{\hat{\beta}})$-fordelt for store n, og finn et uttrykk for $\sigma^2_{\hat{\beta}}$ når du får vite at Fisher-informasjonen i én observasjon er $I(\beta) = \frac{\alpha^2}{\beta^2}$.
- Hvilken av estimatorene $\tilde{\beta}$ og $\hat{\beta}$ vil du foretrekke? Begrunn svaret ditt (*Hint:* du kan bruke at $\alpha^2 \left(\Gamma\left(1 + \frac{2}{\alpha}\right) / \Gamma\left(1 + \frac{1}{\alpha}\right)^2 - 1 \right) > 1$ når $\alpha > 1$).

Oppgave 3

Vi går tilbake til vektdataene fra Oppgave 1. Det fulle datasettet inkluderer vektmålinger av 72 unge kvinner på to forskjellige tidspunkter. Av de 72 kvinnene fikk 17 familieterapi (dette er dataene fra Oppgave 1), 29 fikk kognitiv behandling og 26 fikk ingen behandling, og utgjør en kontrollgruppe.

Vi velger først å analysere dataene ved hjelp av enveis variansanalyse, der målingene av interesse er differensen mellom siste og første vektmåling, dvs. vektøkningen fra første til siste måling (disse ble kalt D_i i Oppgave 1), og grupperingen er gjort med type behandling som faktor. Resultatene fra analysen er vist under.

```
                   Df Sum Sq Mean Sq F value Pr(>F)
factor(behandling)  2    615  307.32   5.422 0.0065 **
Residuals          69   3911   56.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a

- Forklar oppsettet for å utføre en slik analyse på dette datasettet, altså skriv ned modellen og forklar hvilke antakelser som gjøres og hvilke hypoteser som testes.
- Hva kan du si om effekten av behandling ut fra resultatene?

Et alternativ til enveis variansanalyse er en lineær regresjonsmodell, med målt vektøkning som responsvariabel Y . Videre velger vi i denne omgang å gruppere kvinnene etter om de fikk behandling eller ikke, uten å spesifisere hvilken behandling de eventuelt fikk. Vår ene forklaringsvariabel x er da gitt ved:

$$x_i = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling} \end{cases}$$

og modellen er:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

der vi antar at $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$. Resultatet av denne analysen er gitt i R-utskriften nedenfor.

```
Call:
lm(formula = vekt.diff ~ behandling.gruppert)

Residuals:
```

(Fortsettes på side 4.)

```
     Min       1Q   Median       3Q      Max
-13.6804  -5.3054  -0.8804   6.1946  16.9196

Coefficients:
                    Estimate Std. Error  t value  Pr(>|t|)
(Intercept)           -0.450      1.502   -0.300  0.76535
behandling.gruppert    5.030      1.879    2.677  0.00924 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.658 on 70 degrees of freedom
Multiple R-squared:  0.09289,Adjusted R-squared:  0.07993
F-statistic: 7.168 on 1 and 70 DF,  p-value: 0.009239
```

b

- Bruk utskriften fra R til å angi hva estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$ er.
- Vis at

$$\mathrm{E}(Y_i|x_i = 0) = \beta_0 \qquad \text{og} \qquad \mathrm{E}(Y_i|x_i = 1) = \beta_0 + \beta_1,$$

og bruk dette til å gi en fortolkning av $\hat{\beta}_0$ og $\hat{\beta}_1$.
- Lag også et 95% konfidensintervall for β_1 (du trenger ikke å utlede det, men skriv opp formelen du bruker).
- Hva kan du si om effekten av behandling på forventet vektøkning ut fra intervallet?

Vi ønsker nå å undersøke effekten av de to ulike behandlingene. Vi definerer derfor de to forklaringsvariablene x_1 og x_2 ved:

$$x_{1i} = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling 1} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling 1} \end{cases} \qquad x_{2i} = \begin{cases} 1, & \text{kvinne } i \text{ fikk behandling 2} \\ 0, & \text{kvinne } i \text{ fikk ikke behandling 2} \end{cases}$$

der behandling 1 er familieterapi og behandling 2 er kognitiv behandling. Vår nye modell er da gitt ved:

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

der vi antar at $\varepsilon_i^* \overset{uif}{\sim} N(0, (\sigma^*)^2)$. Resultatet av denne analysen er gitt i R-utskriften nedenfor.

```
Call:
lm(formula = vekt.diff ~ factor(behandling), x = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-12.565  -4.543  -1.007   3.846  17.893
```

(Fortsettes på side 5.)

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.450      1.476  -0.305   0.7614
factor(behandling)1    7.715      2.348   3.285   0.0016 **
factor(behandling)2    3.457      2.033   1.700   0.0936 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.528 on 69 degrees of freedom
Multiple R-squared:  0.1358,Adjusted R-squared:  0.1108
F-statistic: 5.422 on 2 and 69 DF,  p-value: 0.006499
```

c

- Finn $\mathrm{E}(Y_i|x_{1i} = j, x_{2i} = k)$ for de tre mulige kombinasjonene $(0, 0)$, $(1, 0)$ og $(0, 1)$ for j, k, og bruk disse til å gi en fortolkning av $\hat{\gamma}_0$, $\hat{\gamma}_1$ og $\hat{\gamma}_2$.
- Hvorfor er $\gamma_0 = \beta_0$, der β_0 er konstantleddet fra Oppgave a)?

Vi ønsker å finne ut om hver av de to behandlingsformene har en effekt på forventet vektøkning eller ikke. Vi vil derfor teste hypotesene

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

for $j = 1, 2$.

d

- Forklar hvilke testobservatorer og tilhørende forkastningsområder du kan bruke for å teste hypotesene over med 5% signifikansnivå.
- Utfør så testene for $j = 1$ og 2 basert på R-utskriftene.
- Hva kan du si om effekten av familieterapi (behandling 1) og kognitiv behandling (behandling 2), og hvordan samsvarer disse resultatene med resultatene fra Oppgave a) og b)?

END