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 5 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.

i) Hvilken type fordeling har summen $Y_1 + Y_2$?

ii) Regn ut forventningsverdien og standardavviket til summen $Y_1 + Y_2$.

Se for deg at du skal skjære til bredden av en hylleplate med kappsag. Når du har bestemt deg for å merke av en bredde B på hylleplata, vil du reelt sett sette et merke i bredden $M = B + \epsilon_1$, der $\epsilon_1 \sim N(0, \sigma_1)$. Når du har bestemt deg for å merke i bredde M på hylleplata, vil du reelt sett kutte plata så den får bredde $Z = M + \epsilon_2$, der $\epsilon_2 \sim N(0,\ \sigma_2)$. Anta at ϵ_1 og ϵ_2 er uavhengige.

1b

Anta at du har bestemt deg for å merke av en bredde $B = 80,00$ cm på hylleplata, og kutte plata langs merket etter beste evne. Anta at $\sigma_1 = \sigma_2 = 0,05$ cm. Regn ut standardavviket til faktisk realisert bredde Z av hylleplata etter at du har skåret den til.

(Fortsettes på side 2.)

1c

Anta at du skulle skjære hylleplata for å sette den i et skap med innvendig bredde 80,00 cm.

i) Regn ut sannsynligheten for at hylleplata ikke passer fordi den er for bred; det vil si har faktisk realisert bredde Z større enn 80,00 cm.

ii) For å øke sannsynligheten for at hylleplata passer: Regn ut største B du kan ta sikte på å merke av, og ha minst 95% sannsynlighet for at hylleplata passer i skapet (det vil si at hylleplata har faktisk realisert bredde $Z \leq 80,00$ cm).

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. Anta videre at det første elementet i hvert par er normalfordelt $X_j \sim N(\mu_X, \sigma)$, og det andre elementet i hvert par er normalfordelt $Y_j \sim N(\mu_Y, \sigma)$. Merk at alle observasjonene normalfordelte med samme standardavvik σ.

2a

I denne deloppgaven antar vi at $\mu_X = \mu_Y$, slik at X_j'ene og Y_j'ene har samme fordeling. For to uavhengige og kontinuerlige tilfeldige variabler X og Y fra samme sannsynlighetsfordeling, er det kjent at det er en 50% sannsynlighet for at X har større verdi enn Y.

i) Forklar hvorfor sannsynlighetsfordelinga til antall par der det første elementet er det største, er binomisk fordelt.

ii) Regn ut forventa antall par der det første elementet er det største.

iii) Regn ut sannsynligheten for at færre enn 40 av parene har det første elementet som det største.

2b

I denne og den neste deloppgaven, antar vi ikke lenger at $\mu_X = \mu_Y$. Det vil si at X_j'ene muligens har ulik forventningsverdi sammenlignet med Y_j'ene. Alle de 200 tilfeldige variablene er fremdeles uavhengige og normalfordelte med samme standardavvik.

i) Hva er fordelinga til den parvise differansen $X_j - Y_j$ i hvert par?

ii) Hva er fordelinga til differansen mellom utvalgsgjennomsnittene $\bar{X} - \bar{Y}$?

For begge spørsmålene om fordeling, oppgi type fordeling, forventningsverdi og standardavvik.

(Fortsettes på side 3.)

2c

Du vil gjennomføre en statistisk hypotesetest for matchede par for å undersøke om forventningsverdiene μ_X og μ_Y er ulike. Angi nullhypotese, alternativhypotese, statistisk signifikansnivå og standardisert testobservator. Gjennomfør den statistiske hypotesetesten når $\bar{X} = 10,02$, $\bar{Y} = 10,98$ og empirisk standardavvik for differansen $X_j - Y_j$ innenfor hvert av de 100 parene er $s_{X-Y} = 4,73$.

Hint: Legg merke til at empirisk gjennomsnitt av differansene $\overline{X-Y}$ er lik differansen mellom utvalgsgjennomsnittene $\bar{X}-\bar{Y}$, altså $\overline{X-Y} = \bar{X}-\bar{Y}$.

Oppgave 3 Forhold mellom kroppsmål

I denne oppgaven skal vi ta et gjensyn med data av kroppsmål gjennomført på 223 studenter som tok emnet BIO2150 ved UiO mellom 2012 og 2016. Vi gjorde eksplorativ dataanalyse av disse dataene i obligatorisk oppgave 1.

Vi vil undersøke sammenhengen mellom de to variablene 'kroppslengde' (total kroppslengde, målt i cm) og 'fot.navle' (avstand fra gulv til navle, målt i cm). I de følgende to deloppgavene, ta utgangspunkt i den modifiserte R-utskriften ([...] og ? er ikke faktisk kode, men kode som er skjult eller fjernet):

```
> data <- read.table(file = [...], header = TRUE)
>

> fit = lm(kroppslengde~fot.navle, data=data)
> summary(fit)

Call:
lm(formula = kroppslengde ~ fot.navle, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7843  -1.9667  -0.0568   2.1695  11.8981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.89675    3.98888   9.751   <2e-16 ***
fot.navle    1.27252    0.03799  33.499   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.605 on 221 degrees of freedom
Multiple R-squared:  0.8355,Adjusted R-squared:  0.8347
F-statistic:  1122 on 1 and 221 DF,  p-value: < 2.2e-16

>
```

(Fortsettes på side 4.)

```
> predict(?, ?, interval= 'confidence', ?)
       fit      lwr      upr
1 171.2393 170.7596 171.7189
>
> predict(?, ?, interval= 'predict', ?)
       fit      lwr      upr
1 171.2393 164.1189 178.3596
```

3a

i) Skriv opp en enkel lineær regresjonsmodell mellom responsvariabelen kroppslengde og forklaringsvariabelen fot.navle. Husk å inkludere antagelsene for modellen.

ii) Forklar hva hver av de tre parameterne β_0, β_1 og σ beskriver.

iii) Les av estimatene til de tre parameterne β_0, β_1 og σ fra R-utskriften.

3b

i) Regn ut et 95% konfidensintervall for β_1 basert på informasjon fra R-utskriften. De ulike stegene i utregningen må vises.

ii) Skriv ned R-kommandoen som bruker predict-funksjonen til å estimere et 95% konfidensintervall for forventa kroppslengde for en person med fot.navle lik 104 cm.

iii) Skriv også ned R-kommandoen som bruker predict-funksjonen til å estimere et 95% prediksjonsintervall for kroppslengde for en person med fot.navle lik 104 cm.

iv) Skriv opp definisjonene av intervallet for hver av disse tre intervallene du har regnet ut over.

Oppgave 4 Snø til jul

Sjansen for snø til jul for Frank Ogfri varierer med kalenderår og om han er i hovedstaden eller på hytta på fjellet på julaften.

4a

Skriv opp en logistisk regresjonsmodell for sammenhengen mellom responsvariabelen 'snø til jul' (y) og forklaringsvariablene kalenderår (X_1) og 'jul på hytta' (X_2).

(Fortsettes på side 5.)

4b

En R-utskrift for den tilpassede logistiske regresjonsmodellen er gitt under. Benytt informasjon fra utskriften til å konstruere et 95% kondensintervall for odds-ratioen for snø til jul for på hytta sammenlignet med hjemme for Frank for et gitt kalenderår.

```
Deviance Residuals:
  Min       1Q   Median       3Q      Max
-1.7404  -0.7460   0.2602   0.5617   2.1632

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept) 199.25505   48.45225   4.112 3.92e-05 ***
  cal_year   -0.10013       0.02437   -4.109 3.97e-05 ***
  hytta       0.76152        0.84862     0.897     0.37
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 86.616  on 64  degrees of freedom
Residual deviance: 56.250  on 62  degrees of freedom
AIC: 62.25

Number of Fisher Scoring iterations: 5
```