STK1110 høsten 2019

Obligatorisk oppgave 2 av to

Innleveringsfrist

Torsdag 7. november 2019, klokken 14:30 i Canvas (canvas.uio.no).

Instruksjoner

Du velger selv om du skriver besvarelsen for hånd og scanner besvarelsen eller om du skriver løsningen direkte inn på datamaskin (for eksempel ved bruk av LaTeX). Besvarelsen skal leveres som én PDF-fil. Scannede ark må være godt lesbare. Besvarelsen skal inneholde navn, emne og oblignummer.

Det forventes at man har en klar og ryddig besvarelse med tydelige begrunnelser. Husk å inkludere alle relevante plott og figurer. Studenter som ikke får sin opprinnelige besvarelse godkjent, men som har gjort et reelt forsøk på å løse oppgavene, vil få én mulighet til å levere en revidert besvarelse. Samarbeid og alle slags hjelpemidler er tillatt, men den innleverte besvarelsen skal være skrevet av deg og reflektere din forståelse av stoffet. Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

I oppgaver der du blir bedt om å programmere må du legge ved programkoden og levere den sammen med resten av besvarelsen.

Søknad om utsettelse av innleveringsfrist

Hvis du blir syk eller av andre grunner trenger å søke om utsettelse av innleveringsfristen, må du ta kontakt med studieadministrasjonen ved Matematisk institutt (e-post: studieinfo@math.uio.no) i god tid før innleveringsfristen.

For å få adgang til avsluttende eksamen i dette emnet, må man bestå alle obligatoriske oppgaver i ett og samme semester.

Spesielt om det obligatoriske oppgavesettet i STK1110

Du skal bruke programpakken R til å gjøre beregningene i oppgavene 1-3, og du må angi hvilke kommandoer du har brukt for å komme fram til svarene dine. For å få godkjent besvarelsen, må du ha minst 65 % riktig *på hver av de fire oppgavene.*

For fullstendige retningslinjer for innlevering av obligatoriske oppgaver, se her:

www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html

LYKKE TIL!

Du kan få hjelp om `R`-kommandoer ved å se på scriptene på timeplanen for perioden 27/9–18/10:
`https://www.uio.no/studier/emner/matnat/math/STK1110/h19/timeplan/index.html#FOR`

Oppgave 1. Tabellen nedenfor viser målinger av kroppstemperaturen for ti for friske menn $(x_1, \ldots, x_{10})$ og ti friske kvinner $(y_1, \ldots, y_{10})$.

	Kroppstemperatur									
Menn	36.1	36.3	36.4	36.6	36.6	36.7	36.7	37.0	36.5	37.1
Kvinner	36.6	36.7	36.8	36.8	36.7	37.0	37.1	37.3	36.9	37.4

Du kan lese tallene inn i R ved kommandoene:

```
path="https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt"
temp=read.table(path,header=T)
```

Hensikten med oppgaven er å undersøke om det er forskjell på kroppstemperaturen for menn og kvinner.

a) Lag boksplott og normalfordelingsplott for observasjonene.
Kommenter hva plottene sier deg.

I resten av oppgaven antar vi at observasjonene er realisasjoner av normalfordelte variabler. I punkt b antar vi at variansene er like for menn og kvinner, en forutsetning vi ser nærmere på i punkt c.

b) Utled en test med signifikansnivå 5% for nullhypotesen at forventet kroppstemperatur er den samme for menn og kvinner mot alternativet at det er forskjell mellom kjønnene. Utfør testen ved å sette inn i de formlene du har utledet. Hva blir konklusjonen din?

   Beregn også P-verdien og gi en fortolkning av den. Kontroller svarene dine ved å bruke `R`-prosedyren `t.test`. (Det kan være lurt å sjekke hjelpe-siden for `t.test` for å se hvordan du skiller mellom antagelsen om like eller ulike varianser for de to kjønn.)

c) Utled og gjennomfør en F-test for å undersøke om det er noen grunn til å påstå at variansene er forskjellige for de to kjønn. Kontroller svarene dine ved å bruke `R`-prosedyren `var.test`.

Oppgave 2. Siden eneggede tvillinger har de samme genene, brukes ofte tvillingstudier til å kartlegge hvordan miljøet virker inn på ulike egenskaper. I en bok av den amerikanske forskeren Susan Faber finner vi data for $n = 31$ tvillingpar, der den ene tvillingen vokste opp hos biologiske foreldre (Twin A) og den andre vokste opp hos andre familiemedlemmer, foster- eller adoptiv-foreldre (Twin B). Nedenfor finner du en oppsummering av IQ-målinger for disse personene. Spørsmålet vi ønsker å belyse er om det er forskjell i IQ hos eneggede tvillinger der den ene tvillingen har vokst opp hos sine biologiske foreldre, og den andre ikke.

```
              N   Mean  StDev  SE Mean
Twin A       31  93.32  15.41     2.77
Twin B       31  96.58  13.84     2.49
Difference   31  -3.26   8.81     1.58
```

I tabellen er `STDev` empirisk standardavvik regnet fra enkeltobservasjoner mens `SE Mean` er standardfeilen til gjennomsnittet.

a) Begrunn hvorfor en paret sammenligning er best egnet i denne situasjonen. Beskriv kort hvilke antakelser du må legge til grunn for den videre analysen.

b) Kall forventet forskjell mellom Twin A og Twin B for μ_D. Sett opp nullhypotese og alternativ hypotese for å besvare spørsmålet om forskjell i IQ. Finn en egnet testobservator, og beregn dennes numeriske verdi. Beregn så tilhørende P-verdi. Spesifiser antall frihetsgrader i fordelingen du bruker. Formuler din konklusjon på testen.

c) Finn et 95% konfidensintervall for μ_D. Hva betyr det at dette intervallet bare dekker negative verdier? Forklar kort sammenhengen mellom tosidig testing og konfidensintervaller.

Oppgave 3. Tabellen nedenfor viser observasjoner av styrken til ti plastikkbeholdere (`Strength`), og to mulige forklaringsvariabler: temperaturen under produksjonen av beholderene (`Temperature`, målt i Celsius grader) og trykket under produksjonen (`Pressure`, målt i pund per kvadrattomme).

Strength	Temperature	Pressure
40.20	290	12
8.30	270	12
30.20	240	12
18.50	210	20
28.20	250	16
35.30	260	12
27.30	220	12
28.10	210	10
35.00	250	12
16.70	210	20

Du kan lese dataene inn i `R` ved kommandoene:

```
path="https://www.uio.no/studier/emner/matnat/math/STK1110/data/plastic.txt"
plast=read.table(path,header=T)
```

Vi er spesielt interessert i variabelen `Strength`, som vi vil se på som responsvariabel i en regresjons-setting. I første omgang konsentrere oss om forklaringsvariabelen `Temperature`.

a) Utfør en enkel lineær regresjon med `Strength` som responsvariabel og `Temperature` som forklaringsvariabel. Plott dataene sammen med den tilpassende regresjonslinjen og kommentér resultatene.

b) Lag konfidensintervall for regresjonskoeffsienten for `Temperature`. Gir intervallet indikasjon på om temperatur er en viktig forklaringsvariabel?

c) Lag så konfidensintervall for forventet verdi av `Strength` og dessuten prediksjonsintervall for nye verdier av `Strength` ved de følgende tre verdiene av `Temperature`: $210, 240$ og 270. Sammenlign intervallene og diskuter forskjellene mellom konfidens- og prediksjonsintervall.

Vi ser så på forklaringsvariabelen `Pressure`.

d) Gjenta beregningene i punktene a og b, men nå med `Pressure` som forklaringsvariabel i stedet for `Temperature`. Kommenter resultatene.

   Hvis du skulle velge mellom én av de to forklaringsvariablene `Temperature` og `Pressure`, hvilken ville du valgt? Begrunn svaret.

Oppgave 4. Når vi bruker en lineær regresjonsmodell, vil vi vanligvis ha med et konstantledd i modellen. Men noen få ganger kan det være aktuelt å bruke en modell uten konstantledd.

Vi vil i denne oppgaven se nærmere på den enkle lineære regresjonsmodellen uten konstantledd. Mer presist vil vi se på følgende situasjon. Vi antar at

$$Y_i = \beta x_i + \epsilon_i\,; \quad i = 1, 2, \ldots, n; \tag{1}$$

hvor x_i-ene er gitte størrelser, ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte, og β og σ^2 er ukjente parametere.

a) Vis at minste kvadraters estimator for β er

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

b) Vis at $\hat{\beta}$ er forventningsrett og bestem variansen til $\hat{\beta}$. Forklar at $\hat{\beta}$ er normalfordelt.

En estimator for σ^2 er

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \hat{\beta} x_i \right)^2 .$$

Det kan vises at $\hat{\beta}$ og S^2 er uavhengige og at $(n-1)S^2/\sigma^2$ er kji-kvadrat fordelt med $n-1$ frihetsgrader. (Du skal ikke vise det.)

c) Forklar at

$$\frac{\hat{\beta} - \beta}{S} \sqrt{\sum_{i=1}^{n} x_i^2}$$

er t-fordelt med $n-1$ frihetsgrader.

d) Vi ønsker å teste nullhypotesen $\mathrm{H}_0 : \beta = \beta_0$ mot den alternative hypotesen $\mathrm{H}_\mathrm{a} : \beta \neq \beta_0$, der β_0 er et gitt tall (for eksempel $\beta_0 = 0$). Bestem en test med signifikansnivå α for dette hypoteseprøvingsproblemet.

e) Vis at

$$\sum_{i=1}^{n} (Y_i - \hat{\beta} x_i)^2 = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2,$$

og bruk det til å vise at S^2 er en forventningsrett estimator for σ^2. (Du får altså her ikke benytte resultatet gitt rett før punkt c.)