STK1000 Innføring i anvendt statistikk, H25

Obligatorisk oppgave 2 av 2, biologiversjon

Innleveringsfrist: Mandag, 3. november 2025 kl 14:30 i Canvas (canvas.uio.no).

Instruksjoner

- For denne obligatoriske oppgave så vil vi fokusere på **læring** og ikke testing. For å få godkjent oppgave er det derfor nok å levere en **blank pdf** fil, men du må minst levere dette! Hvis du ønsker tilbakemelding på det du har gjort, bør du følge instruksjonene nedenfor. Innkluder også noen setninger i begynnelsen om hvilke deler du spesielt ønsker tilbakemelding på.
- 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 eller i R markdown/Quarto). Besvarelsen skal leveres som **én PDF-fil** og 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.
- Følg oppsettet i den foreslåtte malen for besvarelse av hver deloppgave.
- I oppgaver der du blir bedt om å programmere må du legge ved programkoden og levere den sammen med resten av besvarelsen. Det er viktig at programkoden du leverer inneholder et kjøreeksempel, slik at det er lett å se hvilket resultat programmet gir.
- Skriv kommentarer til hver `R`-kommando i programkoden du legger ved, dvs en kort forklaring av hva som skjer når kommandoen brukes. Etter et #-tegn kan man skrive kommentarer som ikke blir lest av `R`, noe som betyr at det ikke påvirker koden.
- 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 redegørelse.
- Merk at man har **ett forsøk** på å få oppgaven godkjent. Dette betyr at det ikke lenger gis andregangsforsøk.

Søknad om utsettelse av innleveringsfrist

Ved udokumenterte årsaker kan du benytte egenmelding for å få innvilget fire kalenderdager utsettelse. Send i så fall e-post til faglærer (thordist@math.uio.no) innen kl. 23:59 på innleveringsdagen.

Ved dokumenterte årsaker kan du søke om inntil en uke utsettelse totalt på en obligatorisk oppgave på bakgrunn av gyldig dokumentasjon. Se de fullstendige retningslinjene nedenfor. Retningslinjer

For å få adgang til avsluttende eksamen i dette emnet, må man bestå alle obligatoriske oppgaver i ett og samme semester. For fullstendige retningslinjer for innlevering av obligatoriske oppgaver, se www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html

LYKKE TIL!

Oppgave 1: Utvalgsfordelinger (Kapittel 5)

Alle biologiske organismer inneholder gener som blir avlest for å danne RNA under transkripsjonsprosessen i cellen. Gener forekommer i ulike lengder, og et gen blir ofte målt i antall nukleotider eller basepar (bp). Datasettet som brukes i denne oppgaven inneholder lengden til 1000 gener som har blitt tilfeldig trukket fra et datasett med alle genene i det menneskelige genomet[1]. Datafilen kan leses inn i `R` på følgende måte:

```
datapath = "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/oblig2/gene.txt"
genes = read.table(datapath, header=TRUE, sep=";")
```

Datafilen består av én linje for hver av de 1000 tilfeldig utvalgte genene og én kolonne, `Gene.Lengths`, genets lengde, oppgitt i antall nukleotider (basepar, bp). I oppgaven ser vi nærmere på fordelingen til gjennomsnittet av tilfeldige utvalg fra datasettet `genes`. Vi antar at de 1000 tilfeldig utvalgte genene i datasettet vårt er hele populasjonen, selv om den faktiske populasjonen er mye større.

a) **Fordelingen til populasjonen**: Regn ut gjennomsnittet og standardavviket til alle genlengdene. Bruk et histogram og andre måter å vurdere dataene til å vurder om genlengdene er tilnærmet normalfordelt.

Vi antar at gjennomsnittet vi har funnet i deloppgave a) er den sanne forventningsverdien i populasjonen, μ, og at standardavviket er det sanne standardavviket i populasjonen, σ.

b) **Utvalg av størrelse 100**: Trekk et tilfeldig utvalg av 100 genlengder fra datasettet ved å bruke følgende `R`-kode:

```
sample(genes$Gene.Lengths, 100)
```

Regn ut gjennomsnittlig størrelse $\bar{x}$ på genlengder i utvalget og sammenlign med forventningsverdien μ du har funnet i deloppgave a). Trekk så 100 tilfeldige utvalg, hvert av størrelse 100, og regn ut gjennomsnittet $\bar{x}$ for hvert av de 100 utvalgene. Finn forventningsverdi, standardavvik og form på fordelingen til $\bar{x}$. Sammenlign forventningsverdien og standardavviket med det du ville forvente ut fra at du kjenner μ og σ.

`R`-hint: Bruk følgende kode til å trekke 100 tilfeldige utvalg og beregne gjennomsnitt:

```
meanvec = rep(NA, 100)
for(i in 1:100) {
  sample.now = sample(genes$Gene.Lengths, 100)
  meanvec[i] = mean(sample.now)
}
```

c) **Utvalg av størrelse 10**: Gjenta deloppgave b) for utvalgsstørreslse 10. Sammenlign resultatene her med resultatene i b).

d) **Sannynligheter**: Hva er sannsynligheten for at du trekker et utvalg gener med gjennomsnittlig genlengde $\bar{x}$ større enn 3000? Beregn sannsynligheten for henholdsvis $n = 10$ og $n = 100$, og sammenlign.

[1] M. Whitlock and D. Schluter. Data for Chapter 4 — estimating with uncertainty. https://whitlockschluter3e.zoology.ubc.ca/chapter04.html.

R-hint: Bruk `pnorm()`.

e) **Bias og varians:** Hva er betydningen av begrepene bias og varians av en observator? Hva vet du om bias og varians av observatoren gjennomsnitt (beregnet fra et utvalg) som et estimat for forventningsverdien i populasjonen?

Oppgave 2: Signifikanstester og konfidensintervaller for å trekke slutninger om forventninger (Kapittel 6 og 7)

Ozon (O_3) er en gass man finner i atmosfæren. Ozongass oppstår i reaksjonen mellom vanlig oksygengass (O_2) og ultrafiolett stråling. Den høyeste konsentrasjonen av ozongass finnes i stratosfæren; det meste av UV-strålingen fra sola absorberes i ozonlaget. Ozon forekommer imidlertid også i mindre konsentrasjoner på bakkenivå. Her dannes gassen i en kjemisk reaksjon mellom nitrogenoksid (NO_2) og O_2. På grunn av sin oksiderende virkning, er ozongass skadelig for mennesker; stort inntak kan føre til redusert lungefunksjon og diverse luftveissykdommer.

I denne oppgaven skal vi analysere ozonnivået i New York på syttitallet. Datasettet består av 108 målinger av ozonkonsentrasjon i tidsrommet mai - september 1973[2]. Datafilen kan leses inn i `R` på følgende måte:

```
datapath = "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/oblig2/ozone.txt"
newyork = read.table(datapath, header=TRUE)
head(newyork, n=3)
```

```
  Ozone Temp Month Day
1    41   67     5   1
2    36   72     5   2
3    12   74     5   3
```

Datafilen består av én linje for hver av de observasjonene fra sommeren 1973 (108 linjer totalt) og fire kolonner: gjennomsnittlig ozon-nivå (`Ozone`, målt i ppb), gjennomsnittlig temperatur (`Temp`, målt i Fahrenheit), indikator for måned og dag når målingen er tatt (`Month` og `Day`).

Oppgaven handler om signifikanstester og konfidensintervaller for å trekke slutninger om forventninger.

a) **Antakelser**: Gi en oppsummering av ozonnivået i New York 1973 med relevante tall og figurer. Vurder spesielt form på fordeling, eventuelle uteliggere, og om det er rimelig å anta en normalfordeling. For å trekke slutninger om forventet ozonnivå i New York om sommeren skal vi bruke t-prosedyrer. Oppgi antakelsene som ligger til grunn for å bruke disse. Forklar også hvorfor vi bør bruke t-prosedyrer og ikke z-prosedyrer.

b) **Test 1**: Normalt forventet ozon-nivå på årsbasis i en Skandinavisk by er 30 ppb. Var forventet ozon-nivå i New York på syttitallet i perioden mai-september mer enn 30 ppb? Begynn med å formulere nullhypotese og alternativ hypotese. Utfør deretter en hypotesetest og formuler en passende konklusjon. Lag også 90%- og 99%-konfidensintervaller for forventet ozon-nivå i New York på syttitallet i perioden mai-september. Gi en intuitiv forklaring på hva forskjellene mellom

[2] J. M. Chambers, W. S. Cleveland, B. Kleiner and P. A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth & Brooks. Cole Statistics/Probability Series. 1983.

de beregnede intervallene er.

R-hint: Bruk `t.test()`. Du kan bruke argumentet `conf.level = <sett inn konfidensnivå>` i `t.test()` for å angi konfidensnivået.

c) **Test 2**: Er det forskjell på ozonnivået i perioden juli/august og mai/juni/september? Formuler nullhypotese og alternativ hypotese. Utfør den passende testen og formuler en konklusjon.

R-hint: Bruk `t.test()` med to variabler som input. Del opp dataene ved følgende R-kommandoer:

```
oz.juli.august = newyork[newyork$Month %in% c(7,8),"Ozone"]
oz.mai.juni.sept = newyork[newyork$Month %in% c(5,6,9),"Ozone"]
```

Oppgave 3: Lineær regresjon (Kapittel 10)

R har mange pakker som kan inneholde nyttige funksjoner for statistisk analyse. Noen av disse pakkene inneholder også datasett. I denne oppgaven skal vi bruke et datasett som heter `cats`, som kommer fra R-pakken `MASS`[3]. Datasettet `cats` inneholder målinger av 144 voksne katters kroppsvekt og hjertevekt, og kan leses inn på følgende måte:

```
library(MASS)
head(cats, n=3)
```

```
  Sex Bwt Hwt
1   F   2 7.0
2   F   2 7.4
3   F   2 9.5
```

Datafilen består av én linje for hver av de 144 kattene og tre kolonner: kjønnet til katten (`F` for hunnkatter, `M` for hannkatter), kroppsvekten til katten (`Bwt`, målt i kg) og hjertevekten til katten (`Hwt`, målt i g).

Oppgaven går ut på å tilpasse en lineær modell for hjertevekten av katter.

a) **Regresjonsmodell:** Lag et spredningsplott med `Bwt` på x-aksen og `Hwt` på y-aksen. Sett navn på aksene, i tillegg til å gi plottet en passende tittel. Er det grunn til å tro at det er et lineært forhold mellom variablene på bakgrunn av plottet? Begrunn svaret. Lag en lineær modell med `Hwt` som responsvariabel og tegn regresjonslinjen i spredningsplottet. Hvilke antagelser ligger til grunn for en enkel lineær regresjonsmodell? Vurder om disse antagelsene er oppfylt for den tilpassede modellen.

b) **Skjæringspunkt og stigningstall**: Bruk `summary()`-funksjonen til å finne modellens skjæringspunkt og stigningstall. Ifølge modellen, hvor mye øker katters forventede hjertevekt når kroppsvekten øker med 1 kg? `summary()`-utskriften til den lineære modellen viser at modellens stigningstall har en veldig lav p-verdi. Hvordan skal dette tolkes? Hva er nullhypotesen i denne sammenhengen? Finn et 95%-konfidensintervall for stigningstallet i modellen og gi en tolkning av resultatet.

[3] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. 2002.

R-hint: Bruk følgende kode til å beregne konfidensintervallet:

```
bwt = cats$Bwt
hwt = cats$Hwt
fit = lm(hwt ~ bwt)
b1 = summary(fit)$coefficients[2, 1]
se.b1 = summary(fit)$coefficients[2, 2]
df = fit$df.residual
lower = b1 + qt(0.025, df) * se.b1
upper = b1 + qt(0.975, df) * se.b1
```

c) **Prediksjoner**: Plott modellens prediksjonsintervall og konfidensintervall for forventet respons. Forklar hvordan disse to intervallene skal tolkes. Forklar også hvorfor det ene intervallet er bredere enn det andre. Man måler kroppsvekten av en ny katt. Kroppsvekten til denne katten er 3.0 kg. Gi et 95%-prediksjonsintervall for hjertevekten til denne katten ved å lese av plottet.

R-hint: Bruk følgende kode til å lage plottet:

```
par(mex=0.75, cex=0.7)
plot(bwt, hwt, xlim = c(1.6, 4), ylim = c(3, 22),
     xlab = "navn x-akse", ylab = "navn y-akse", main = "tittel")
abline(fit)
xval = seq(1, 4.5, by = 0.01)
new = data.frame(bwt = xval)
pred.int = predict(fit, newdata = new, interval = "prediction")
mean.int = predict(fit, newdata = new, interval = "confidence")
matlines(xval, cbind(pred.int[, 2], pred.int[, 3]), lty = 2, col = "steelblue")
matlines(xval, cbind(mean.int[, 2], mean.int[, 3]), lty = 2, col = "tomato")
```