06. september 2018

STK1000 – Innføring i anvendt statistikk

Obligatorisk oppgave 1 av 2. Datasettene til de to første oppgavene er hentet fra biologi.

Innleveringsfrist

Torsdag 20. september 2018, klokken 14:30 i Devilry (devilry.ifi.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. Det er viktig at programkoden du leverer inneholder et kjøreeksempel, slik at det er lett å se hvilket resultat programmet gir.

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.

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

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

LYKKE TIL!

For alle oppgavene

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.

Oppgave 1

I denne oppgaven skal vi se på en studie [1] hvor det ble gjennomført målinger av basal metabolsk hastighet (BMR) hos sebrafinker (*Taeniopygia guttata*). Basalmetabolisme er den mengden energi et endotermt dyr bruker per tidsenhet i hvile, og måles i antall mL O_2 per min. Datasettet som benyttes i denne oppgaven inneholder informasjon om basal metabolsk hastighet for 150 sebrafinker som er tilfeldig valgt fra det opprinnelige datasettet.

Datafilen kan leses inn i R på følgende måte:

```
data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data/zebrafinch.txt"
zebrafinch <- read.table(data,header=TRUE)
```

Datafilen består av én linje for hver av de 150 sebrafinkene. I første kolonne er følgende variabel gitt:

- `BMR`: Målt basal metabolsk hastighet til individet (mL O_2 per min).

a) Plott et histogram av `BMR` ved å bruke `R`. Beskriv hva du ser fra histogrammet. `R`-hint: `hist()`.

b) Finn gjennomsnitt $\bar{x}$ (x betegner variabelen `BMR`) og median for `BMR` ved bruk av `R`. Hva er disse tallene uttrykk for?

c) Finn standardavvvik s og inter-kvartil avstand til `BMR` ved bruk av `R`. Forklar hvordan disse tallene oppsummerer spredningen i basal metabolsk hastighet.

d) Vurder om det er rimelig å anta at `BMR` er normalfordelt. Forklar hvordan du vurderer dette. `R`-hint: `qqnorm()` og `qqline()`.

I resten av oppgaven: Vi antar at `BMR` er normalfordelt. Vi kjenner ikke populasjonens μ og σ, men vi later som vi kjenner disse ved å la $\mu = \bar{x}$ og $\sigma = s$.

e) Gi et uttrykk for den standardiserte verdien av `BMR`. Regn ut den standardiserte verdien av `BMI` $= 0.8$ mL O_2 per min . Forklar hvordan svaret skal tolkes.

f) Hva er sannnsynligheten for at sebrafink har `BMI` lavere enn 0.6 mL O_2 per min? `R`-hint: `pnorm()`.

g) Hva er sannsynligheten for at en sebrafink har `BMI` høyere enn 1.0 mL O_2 per min?

Oppgave 2

Menneskelig aktivitet kan bidra til endret atferd hos dyrelivet. I en studie [2] ble det undersøkt om ulv (*Canis lupus*) som var utsatt for høyere grad av jakt, opplevde en økning i produksjon av steroidehormoner. Målinger av konsentrasjonen av steroidehormoner ble gjennomført ved å analysere hårprøver fra 148 individuelle ulver fra ulike regioner i nordlige deler av Canada, fra enten et lett jaktet område eller et tungt jaktet område. Stereoidehormonene ble målt i antall picogram (pg) per miligram (mg) hår. I denne oppgaven har hormonene testosteron og kortison blitt valgt ut fra det opprinnelige datasettet.

Datafilen kan leses inn i R på følgende måte:

```
data ="http://www.uio.no/studier/emner/matnat/math/STK1000/data/wolves.txt"
wolf <- read.table(data,header=TRUE)
```

Datafilen består av én linje for hver av de 148 hårprøvene. I fire kolonner er følgende variable gitt:

- `sex`: Kjønnet til hvert individ (`F` for hunn, `M` for hann).
- `population`: Indikator for om individet kommer fra en lett eller tungt jaktet populasjon (1 = lett jaktet, 2 = tungt jaktet).
- `cpmg`: Konsentrasjonen av kortisol i hårprøven (pg/mg).
- `tpmg`: Konsentrasjonen av testosteron i hårprøven (pg/mg).

a) Forklar forskjellen på en kategorisk og en kvantitativ variabel. Hvilke av variablene i datasettet er kvalitative, og hvilke variable er kvantitative?

b) Lag oppsummeringer av den/de kategoriske variablene i filen ved å bruke følgende R-koder:

```
table(variabelnavn)
pie(table(variabelnavn))
```

Kommenter.

c) Du skal nå dele opp dataene dine i to datasett: Ett som inneholder data for ulvene som er lett jaktet (`wolf.lett`) og ett for ulvene som er tungt jaktet (`wolf.tungt`). Dette gjør du ved følgende `R`-kommandoer:

```
wolf.lett <- wolf[wolf[,"population"]==1,]
wolf.tungt <- wolf[wolf[,"population"]==2,]
```

d) For hvert av datasettene `wolf.lett` og `wolf.tungt`, plott histogram og boxplott av kortisolkonsentrasjon i hårprøvene ved å bruke `R`-kommandoene:

```
par(mfrow=c(2,2))
hist(wolf.lett)
boxplot(wolf.lett)
hist(wolf.tungt)
boxplot(wolf.tungt)
```

Beskriv hva du ser fra hvert plott. Sammenlign plottene for de to datasettene.

e) For hvert av datasettene `wolf.lett` og `wolf.tungt`, finn gjennomsnitt og median av kortisolkonsentrasjon i hårprøvene, og sammenlign disse størrelsene. Kommenter og forklar. Sammenlign også resultatene for de to datasettene.

f) Læreboken nevner to ulike måter å oppsummere data numerisk på: femtallsoppsummering og gjennomsnitt med standardavvik. Hvilken av disse måtene oppsummerer kortisolkonsentrasjonen best for hvert av datasettene `wolf.lett` og `wolf.tungt`? Begrunn svaret.

g) Vurder om konsentrasjonen av kortisol i hårprøvene (`cpmg`) er tilnærmet normalfordelt for hvert av datasettene `wolf.lett` og `wolf.tungt`.

Oppgave 3

I denne oppgaven skal vi se på kroppsmål gjennomført på 223 studenter som tok emnet BIO2150 ved UiO mellom 2012 og 2016. Hensikten med øvelsen var å samle inn data som kunne brukes til statistisk analyse. I datasettet `vitruvisk.txt`[1] finner man resultatene for de 223 studentene som

[1]Datasettet er laget og behandlet av Halvor Aarnes og Tom Andersen fra IBV, UiO.

deltok på øvelsen. Navnet på datasettet refererer til Leonardo da Vincis berømte Den Vitruviske Mann", som er en illustrasjon av menneskets proposjoner.

Datafilen kan leses inn i R på følgende måte:

```
data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data/vitruvisk.txt"
vitruvisk <- read.table(data,header=TRUE)
```

Datafilen består av én linje for hver av de 223 studentene. I fem kolonner er følgende variable gitt:

- `kjonn`: Kjønnet til personen (`K` for kvinne, `M` for mann).
- `kroppslengde`: Total kroppslengde (cm).
- `fot.navle`: Avstand fra gulv til navle (cm).
- `navle.isse`: Avstand fra navle til isse (cm).
- `favn`: Avstand fra venstre langfingerspiss til høyre langfingerspiss med utstrakte armer (cm).

a) Bruk `R`-funksjonen `summary()` til å se et sammendrag av data. Hvor mange av målingene ble gjennomført på kvinner, og hvor mange av målingene ble gjennomført på menn? Hva er fem-punkts-oppsummeringene for kroppslengde og fot.navle?

b) Bruk `plot()`-funksjonen i `R` til å lage et spredningsplott med `fot.navle` på x-aksen og `kroppslengde` på y-aksen. Bruk følgende `R`-kode:

```
plot(vitruvisk$fot.navle,vitruvisk$kroppslengde,xlab="navlehøyde",
ylab="kroppslengde")
```

Hva kan du slutte om sammenhengen mellom navlehøyde og kroppslengde ut fra dette plottet?

c) Bruk `cor()`-funksjonen i `R` til å regne ut korrelasjonen mellom `fot.navle` og `kroppslengde`. Forklar hvordan svaret skal tolkes.

d) Tilpass en lineær modell for sammenhengen mellom navlehøyde og kroppslengde. Plott regresjonslinjen i spredningsplottet fra oppgave b). Bruk følgende R-kode:

```
fit <- lm(vitruvisk$kroppslengde vitruvisk$fot.navle)
abline(fit)
```

e) Finn koeffisientene til den lineære modellen ved `summary(fit)`. Ifølge modellen, hvor mye øker kroppslengde hvis navlehøyde øker med én cm?

f) Bruk modellen i d) til å predikere kroppslengden til en person med en navle som er 121 cm over bakken.

g) Bruk også `summary()`-kommandoen til å finne ut hvor stor andel av variasjonen i kroppslengede som forklares av navlehøyde. Sammenlign med svaret i oppgave c).

h) Lag et plott av residualene i modellen. Identifiser eventuelle uteliggere. Hva kan du si om hvor godt modellen passer til og beskriver dataene ut fra dette plottet? Bruk følgende R-kode:

```
plot(vitruvisk$fot.navle,residuals(fit))
abline(h=0)
```

Referanser

[1] K. J. Mathot, K. Martin, B. Kempenaers, and W. Forstmeier. Basal metabolic rate can evolve independently of morphological and behavioural traits. *Heredity*, 111:175, 2013.

[2] M. Bryan Heather, E. G. Smits Judit, Lee Koren, C. Paquet Paul, E. Wynne-Edwards Katherine, and Marco Musiani. Heavily hunted wolves have higher stress and reproductive steroids than wolves with lower hunting pressure. *Functional Ecology*, 29(3):347–356, 2014.