Helle Sørensen, Anders Tolver (MATH, KU)
Efterår, 2020

Opgaver til Statistisk Dataanalyse 1

Opgave HS.13 (sudoku)

På StatData1 blev der i 2018 udført et forsøg, hvor de studerende blev bedt om at løse en af fire forskellige sudokuer. De fire sudokuer var ens i struktur, men var enten med tal, latinske bogstaver, græske bogstaver eller symboler.

Data ligger i filerne `sudoku-sd1.xlsx` og `sudoku-sd1.txt`. I denne opgave skal vi kun bruge variablene **type**, **min** og **sek**.

1. Indlæs data og lav en ny variabel i datasættet vha. følgende kommando hvor det antages at datasættet er kaldt **sudokuData**:

   ```
   sudokuData <- transform(sudokuData, tid=60*min+sek)
   ```

   Forklar hvad den nye variabel måler.

   Hvis du har indlæst data fra Excel, så brug desuden følgende kommando, så variablen **type** bliver kodet som en faktor (dette sker automatisk hvis du indlæser fra `txt`-filen):

   ```
   sudokuData <- transform(sudokuData, type=factor(type))
   ```

2. Vi er interesseret i at sammenligne de fire sudokutyper med hensyn til hvor lang tid det tager at løse dem.

   - Hvad vil du bruge som responsvariabel? Som forklarende variabel?
   - Hvilken type analyse vil du bruge: Lineær regression, ensidet ANOVA, en enkelt stikprøve, eller noget helt andet?
   - Lav en figur der illustrerer eventuelle forskelle på de fire typer.

3. Fit modellen med R, dvs. brug funktionen `lm`. Brug i første omgang kommandoen

   ```
   lm(tid ~ type-1, data=sudokuData)
   ```

   Lav et summary af modellen, og angiv et estimat for den forventede tid det tager at løse hver af de fire typer sudoker.

   Angiv også standard errors hørende til estimaterne. Hvorfor er de fire standard errors ikke ens?

   Angiv desuden residualspredningen, dvs. den estimerede spredning på enkeltobservationer.

4. Bestem et estimat for *forskellen* i forventet tid som det tager at løse sudokuen med latinske og græske bogstaver. Angiv også den tilhørende standard error.

   Vink: Det er nyttigt at ændre `lm`-kommandoen en smule.

5. Bestem et konfidensinterval for *forskellen* i forventet tid som det tager at løse sudokuen med latinske og græske bogstaver.

6. R bruger som default græske bogstaver som referencegruppe, fordi navnet på denne gruppe kommer først i alfabetet, men det er umiddelbart mere naturligt at bruge *Tal* som referencegruppe.

   Fit modellen igen, denne gang med tal som referencegruppe, og bestem konfidensintervaller for forskellen i forventet tid mellem *Tal* og hver af de andre grupper.

   Vink: Det er nyttigt at bruge `relevel`, se evt. side 137.

7. Hvad kan du (foreløbig) konkludere om forskellene mellem de fire sudokutyper?

Opgave HS.14 (Højde og vægt)

Data i filerne `stud2017-v2.txt` og `stud2017-v2.xls` indeholder blandt andet selvrapporteret højde og vægt for 154 studerende fra kurset i 2017. Disse variable hedder **hojde** og **vaegt**, mens køn er angivet i variablen **kon**. Hvis du anvender data fra `.txt-filen` kan du fx. bruge følgende R-kommando til indlæsning:

```
data <- read.table(file = "stud2017-v2.txt"
                   , header  = T, sep = "\t", dec = ",")
```

Vi er interesseret i at undersøge om der er en sammenhæng mellem højde og vægt og at kunne prædiktere vægten udfra højden — i det (begrænsede) omfang det kan lade sig gøre. Vi vil analyse data for mænd og kvinder hver for sig.

1. Hvad vil du bruge som responsvariabel? Som forklarende variabel?

   Hvilken type analyse vil du bruge: Lineær regression, ensidet ANOVA, en enkelt stikprøve, eller noget helt andet?

2. Indlæs data, og lav separate del-datasæt for mænd og kvinder. Du kan fx bruge funktionen `subset`.

Betragt først kvinderne:

3. Lav en passende figur der viser data.

4. Fit en lineær regressionsmodel til data, dvs. brug `lm`.

   Betragt en kvinde på 165 cm, og angiv et estimat for hendes forventede vægt.

5. Angiv estimatet for hældningen og den tilhørende standard error. Bestem også 95% konfidensintervallet for hældningen.

6. Tyder data på at der er en sammenhæng mellem højde og vægt for kvinder?

   Vink: Hvilken værdi af hældningen svarer til at der *ikke* er en sammenhæng og er dermed særligt interessant?

7. Angiv residualspredningen, dvs. estimatet for spredningen omkrng regressionslinien.

8. Betragt to kvinder med en højdeforskel på 1 cm. Hvor meget vil du forvente at den højeste kvinde vejer mere end den laveste kvinde? Angiv også et 95% konfidensinterval for den forventede vægtforskel.

9. Betragt to kvinder med en højdeforskel på 10 cm. Hvor meget vil du forvente at den højeste kvinde vejer mere end den laveste kvinde?

   Bestem også et 95% konfidensinterval for den forventede vægtforskel.

Betragt så mændene:

10. Fit den tilsvarende model for mændene, og angiv estimat, standard error og konfidensinterval for hældningen i modellen.
11. Sammenlign standard error for hældningen mellem mænd og kvinder. Hvorfor er den meget større for mændene end for kvinderne?
12. Tyder det umiddelbart på at „vægtforskellen ved en ekstra cm i højden" er forskellig for mænd og kvinder?

Opgave HS.15 (Gæt på figur 2 og 3)

Husk data fra opgave HS.11 vedr. studerende gæt på antal punkter i tre punktplot. Data er tilgængelige i filerne `punktplot2017.xlsx` og `punktplot2017.txt`.

I videoen som opsummerer på indholdet kursusuge 2 (9/9-2020) analyserede vi gættene på antallet af punkter i figur 1 og konkluderede at man generelt gætter for lavt på antallet af punkter i denne figur. Analysen er tilgængelig i R-programmet hørende til denne undervisningsdag/video.

1. Forklar hvorfor der er tale en *en enkelt stikprøve* når du kigger på data fra hver figur for sig.
2. Udfør samme analyse for figur 2 og figur 3 som vi udførte for figur 1 onsdag eftermiddag i uge 2.

   Specielt: Bestem estimat og 95% konfidensinterval for det typiske gæt (medianen) for hver af figurerne.
3. Er konklusionen den samme for alle tre figurer i forhold til om det typiske gæt er korrekt, for højt eller for lavt?

Opgave HS.16 (Parrede vs uparrede stikprøver)

Nedenfor beskrives tre situationer hvor data består af to stikprøver. Afgør for hver situation om der er tale om parrede eller uparrede stikprøver, og forklar hvorfor.

1. 129 ægtepar indgår i et kostforsøg hvor både manden og kvinden registererer deres kalorieindtag en enkelt dag. Data består således af i alt 258 tal, og formålet er at estimere forskellen mellem mænds og kvinders kalorieindtag.
2. Nogle forskere ønsker at undersøge forskellen i bakterieflora mellem to bønnesorter, og har derfor opsat et potteforsøg med 32 potter. I hver potte er der plantet en bønneplante af hver slags, og ved slutningen af forsøget måles bakteriediversiteten for hver plante (64 planter i alt).
3. Data består af overskuddet i 2017 per ansat fra 48 tilfældigt udvalgte virksomheder med 1–5 ansatte og 48 tilfældigt udvalgte virksomheder med 6–10 ansatte. Data er indsamlet med henblik på at undersøge forskellen i overskud mellem de to typer virksomheder.

Opgave HS.17 (Forhold mellem hjerte og kropsvægt)

Datasættet **cats** fra R-pakken *MASS* indeholder kropsvægt i kg og hjertevægt i gram for 97 hankatte og 47 hunkatte. Du kan få adgang til data med følgende kommandoer:

```
library(MASS)
data(cats)
```

Vi har kigget på disse data nogle gange ved forelæsningerne til at illustrere lineær regression, men denne opgave bruger data på en lidt anden måde: Vi skal sammenligne forholdet mellem hjerte- og kropsvægt mellem han-og kunkatte.

1. Lav en ny variabel i datasættet, der angiver forholdet mellem hjerte- og kropsvægt, altså `Forhold = Hwt/Bwt`.
2. Vi vil som sagt undersøge om forholdet mellem hjerte- og kropsvægt er forskelligt for hankatte og hunkatte. Er der tale om parrede eller uparrede data?
3. Lav „parallelle boxplots“ , dvs. boxplots for hanner og hunner ved siden af hinanden. Bestem også stikprøvespredningen for hannerne og hunnerne hver for sig. Kan vi med rimelighed antage at spredningen er ens i de to grupper?
4. Bestem et estimat for forskellen i forholdet mellem han- og hunkatte. Angiv også den tilhørende standard error og et 95% konfidensinterval for forskellen.
5. Udfør et hypotesetest for hypotesen om at forholdet mellem hjerte- og kropsvægt i gennemsnit er det samme for hankatte og hunkatte. Hvad er konklusionen?