Helle Sørensen, Anders Tolver (MATH, KU)
Efterår, 2021

Opgaver til Statistisk Dataanalyse 1

Opgave HS.5 (R Markdown)

Du opfordres kraftigt til at benytte R Markdown, når du arbejder i R. Denne opgave kan hjælpe dig igang. Du kan også se videoen med opsummering på kursusuge 1 (findes på Absalon). Der findes også et dokument om R Markdown under fanebladet R på kursushjemmesiden, hvor jeg løbende vil skrive om tekniske emner vedr. R Markdown, som I ønsker belyst.

RStudio har lavet et cheat sheet, som beskriver nogle af mulighederne med R Markdown. Du skal ikke bruge det i denne opgave, men hvis du googler *R markdown cheatsheet*, så popper det op. Der findes også en masse nyttig information på rmarkdown.rstudio.com.

Markdown kræver at man installerer diverse pakker, og det nemmeste er at installere dem når R beder om dem. Se punkt 1.

1. Klik *File* → *New File* → *R Markdown*.

   Hvis du allerede har alle de relevante pakker installeret, kommer der en boks frem der hedder *New R Markdown*. I så fald kan du gå videre til næste punkt.

   Hvis du ikke har de relevante pakker installeret, så kommer der i stedet en boks hvor RStudio spørger om du vil installere nogle pakker. Svar *Ja* til dette. Så installerer R pakkerne, og derefter kommer der en boks frem der hedder *New R Markdown*. Gå så videre til næste punkt.

2. I boksen *New R Markdown*:

   Vælg *Document* (default) i venstre side, udfyld *Title* med teksten „Hejsa“, og vælg *html* som output (default). Så skulle der gerne komme en ny fil i editoren. Gem filen på sædvanligvis, fx via *File* menuen. Angiv endelsen (file extension) `.Rmd`.

3. Klik nu på *Knit* (det lille garnnøgle øverst i vinduet). Så genererer RStudio en html-fil med tekst, R-kode og -output, og åbner html-filen (muligvis i et separet vindue). Filen bliver gemt i samme katalog som Rmd-filen.

   Gå tilbage til Rmd-filen, og identificér hvilke dele af koden der har genereret hvilke dele af output-filen.

4. Ret overskriften „Hejsa“ til „Jeg kan godt lide statistik“ (eller noget andet efter eget ønske). Skriv også en linje eller to om hvorfor du godt kan lide statistik. Slet al den nedenstående kode.

   Knit filen igen, og se om outputtet bliver som forventet.

5. Indsæt en *R chunk*, fx via *Code -¿ Insert* i topbjælken i editoren. Skriv noget kode, fx

   ```
   2 + 3
   x <- c(1,3,4,7)
   mean(x)
   ```

   Du skal ikke knitte nu!

   Klik derimod på den lille *Play* knap (grøn trekant) yderst til højre i R-chunken. Hvad sker der?

6. Lav en ny R-chunk, og skriv noget kode der laver noget grafik, fx:

   ```
   y <- c(4,8,1,3)
   plot(y ~ x)
   ```

   Klik på play-knappen. Hvad sker der?

7. Skriv lidt tekst nedenunder R-chunken hvor du kommenterer grafen, og knit dokumentet. Ser det ud som forventet?

8. Tilføj en linje i dokumentet der starter med to hashtags (##) og derefter har lidt tekst. Knit, og se hvad der sker.

9. Prøv at trykke på tandhjulet ud for en R-chunk, og check drop-down menuen i *Output*. Vælg en anden mulighed, knit, og se hvad der sker.

10. Hvis du vil lave Word- eller pdf-output i stedet for html-output: Gå til toppen af din markdownfil, og ret outputlinjen til en af nedenstående:

    ```
    output: pdf_document
    output: word_document
    ```

    Knit derefter.

    Du skal have MS Word installeret for at du kan lave word-output.

    For at lave pdf-output skal du have en version af programmet TeX installeret: MikTex til Windows, MacTeX til mac. Hvis du knitter uden at have TeX installeret, så kommer der en fejlmeddelelse hvor der er links til installation af TeX.

    Bemærk at der om MikTex står at det er vigtigt at vælge *Complete* snarere end *Basic* installation, men det kan man så vidt jeg kan se, ikke. Det kan heldigvis fikses: Når du bliver spurgt om du vil „Install packages on the fly“, så svar Yes, ikke „ask me first.“

    Du skal genstarte RStudio før du kan knitte til pdf.

Nogle råd hvis du arbejder med R Markdown:

- Lad være med at knitte hele tiden. Kør i stedet de enkelte kommandoer linjevis (som sædvanlig) eller hele chunks. Når du er tilfreds med R-koden til en opgave eller en del af en opgave, så knitter du og checker resultatet.
- Når du knitter, kan du ikke bruge data eller andet som du har indlæst "udenfor" filen. Hvis du for eksempel skal bruge data fra *isdals*-pakken, så skal der være datalinjer à la følgende i Rmd-filen:

  ```
  library(isdals)
  data(antibio)
  ```

- Hvis der ikke kommer noget output, så læs fejlmeddelelsen. Det er ikke altid helt nemt at læse, men i det mindste får man at vide hvor i filen der er problemer.
- Som altid: Sørg for at organisere din fil fornuftigt. Lav overskrifter, skriv tekst passende steder. Lad være med at skrive al R-kode i en enkelt chunk. Lav i stedet nye R-chunks til nye opgaver/delopgaver.
- Vær disciplineret, og sørg for kun at gemme de relevante kommandoer, ikke alle de kommandoer der viste sig at være forkerte eller irrelevante. Dette gælder selvfølgelig også hvis du ikke arbejder med Markdown, men i et almindeligt R-program.
- Du kan åbne outputfilerne i andre programmer end RStudio. Brug fx din yndlingsbrowser til html-filerne når du skal kigge på dem efterfølgende.

Opgave HS.6 (Marginalskat)

Marginalskatteprocenten angiver hvor meget indkomstskat man betaler af „den sidst tjente krone“.

Vi betragter i denne opgave populationen bestående af 18-64 årige fuldt beskæftigede lønmodtagere og selvstændigt erhvervsdrivende der *ikke* betaler topskat. Vi antager at marginalskatteprocenten for denne population er normalfordelt med middelværdi 41.5% og spredning 1.4%.

1. Lav en skitse af tætheden for den givne normalfordeling. *Vink:* Se side 79 i bogen.
2. Bestem sandsynligheden for at en tilfældig person fra den givne population har en marginalskatteprocent der er under 38.7%? Illustrer sandsynligheden på din skitse fra spørgsmål 1 (eller lav en ny skitse).
3. Bestem sandsynligheden for at en tilfældig person fra den givne population har en marginalskatteprocent der er over 42%? Illustrer sandsynligheden på din skitse fra spørgsmål 1 (eller lav en ny skitse).
4. Bestem sandsynligheden for at en tilfældig person fra den givne population har en marginalskatteprocent der er mellem 40% og 42.2%. Illustrer sandsynligheden på din skitse fra spørgsmål 1 (eller lav en ny skitse).
5. Betragt 10 personer fra populationen og antag at deres marginalskatteprocenter er uafhængige af hinanden (jordbrugsøkonomerne kan overveje hvornår dette er en rimelig antagelse).

   Angiv fordelingen af *gennemsnittet* af de ti personers marginalskatteprocent. *Vink:* Du kan bruge Infobox 4.3.

   Bestem derefter sandsynligheden for at gennemsnittet for de ti personer er mindre end 42%. Sammenlign med sandsynligheden fra spørgsmål 3, og forklar forskellen.
6. Overvej hvorfor det er vigtigt at vi kun ser på personer der ikke betaler topskat.

Opgave HS.7 (A-vitaminindtag)

Læs eksempel 4.6 i bogen (side 84–85) pm indtaget af A-vitamin. Konklusionen er at det kan antages at *logaritmen til A-vitaminindtaget* for mænd er normalfordelt med middelværdi 7.48 og spredning 0.44.

1. Brug normalfordelingsantagelsen ovenfor til at bestemme sandsynligheden for at en tilfældig mand har et A-vitaminindtag der er mindre end 1000. Husk at log er den naturlige logaritmen.

   Sammenlign med histogrammet på side 85: Virker tallet fornuftigt?

2. Hvilken sandsynlighed får du hvis du i stedet antager at selve A-vitaminindtaget — altså uden log-transformation — er normalfordelt?

3. Hvilken af sandsynlighederne stoler du mest på? Hvorfor?

Opgave HS.8 (Nettoindtjening ved forpagtning)

Man har indhentet data fra 52 landmænd der har forpagtede arealer. Data består af nettoindtjeningen fra det forpagtede areal, dvs. det inkluderer dækningsbidrag og EU-støtte og omkostninger er fratrukket.

Gennemsnittet for de 52 værdier af nettoindtjening er 4482 kr/ha, mens stikprøvespredningen er 696 kr/ha.

1. Benyt formel (5.21) til at bestemme et 95% konfidensinterval.

2. Hvilket af nedenstående fortolkninger er korrekt:

   - Der er 95% sandsynlighed for at nettoindtjeningen for en tilfældig forpagter vil havne i intervallet.
   - Intervallet indeholder de værdier af populationsgennemsnittet der med rimelighed passer med data, hvor populationsgennemsnittet er den gennemsnitlige nettoindtjening for samtlige forpagtere.

Opgave HS.9 (Indlæsning af data)

Data er oftest tilgængelige i eksterne filer, så man har brug for at indlæse data fra disse filer til R. Denne opgave viser indlæsning af de samme data fra to filformater: Excel (fordi det er sådan de fleste gemmer deres data) og tekstfiler (fordi det virker uden brug af R-pakker).

Data består af kropsmålinger fra 243 mænd. For hver mand har man målt omkredsen ved hofte og mave, begge dele i cm. Desuden har man bestemt mændenes fedtprocent med en præcis målemetode baseret på opdriften ved undervandsvejning.

Data er tilgængelige på kursushjemmesiden i filerne `johnson-fatpct.xlsx` og `johnson-fatpct.txt`. Start med at gemme filerne i den mappe som du bruger til dit R-arbejde på kurset.

Indlæsning fra Excel

Indlæsning af data fra Excelfiler er blevet ganske let i nye versioner af RStudio. Hvis nedenstående ikke virker, er det formentlig fordi du har en gammel version af RStudio.

1. Åbn filen `johnson-fatpct.xlsx`, og forstå strukturen: Passer antallet af datalinjer, og giver variabelnavnene mening i forhold til beskrivelsen ovenfor?
2. Klik på *Import Data* i øverste høje vindue i RStudio, og vælg *From Excel*. Klik *Browse* og klik dig frem til den ønskede Excel-fil. Check at datasættet ser fint ud i *Preview*.

   I *Import options, Name* kan du ændre navnet, som R vil give datasættet. Skriv (fx) `dat1` her. Lad være med at ændre andre ting med mindre du ved hvad du gør. I *Code Preview* kan du se hvilken kode der vil blive udført. Klik *Import*.
3. Læg mærke til at R nu kørte koden og lavede et datasæt, som nu vises. Kopiér den kode som blev genereret, over i dit R-program eller Markdown fil. Hvis du bruger Markdown, så slet kodelinen med `View`.
4. Prøv fx kommandoen `mean(dat1$hip)`.
5. En anden og lettere (?) metode til indlæsning af Excelfiler er at installere og loade R-pakken `readxl` og dernæst indlæse datafilen med kommandoen `read_excel("johnson-fatpct.xlsx")`

Indlæsning fra tekstfil

Indlæsning af tekstfiler foregår med funktionen `read.table` som beskrevet i bogens appendix B.2.3.

6. Brug *Session* → *Set Working directory* → *Choose directory* til at skifte working directory til den folder hvor du har gemt tekstfilen.
7. Brug så kommandoen

   ```
   dat2 <- read.table('johnson-fatpct.txt', header=TRUE)
   ```

   Nu skulle `dat2` gerne optræde i listen over datasæt i *Environment* i øverste højre vindue. Har det samme størrelse som `dat1`?
8. Prøv fx kommandoen `mean(dat2$hip1)`. Får du samme resultat som før?

Opgave HS.10 (Konfidensinterval for en enkelt stikprøve på flere måder)

Denne opgave bruger hoftemålingerne fra opgave HS.9, altså variablen **hip** fra datafilerne `johnson-fatpct.xlsx` og `johnson-fatpct.txt`.

1. Indlæs data som beskrevet i opgave HS.9, enten fra Excelfilen eller fra tekstfilen, hvis du ikke allerede har gjort det.
2. Undersøg grafisk om hoftemålingerne kan antages at være normalfordelte.

I de næste spørgsmål skal du antage at målingerne er uafhængige og normalfordelte. Hvis y_i er hoftemålingen for person i, antager vi altså at $y_i \sim N(\mu, \sigma^2)$ hvor μ og σ er middelværdi og spredning for populationen.

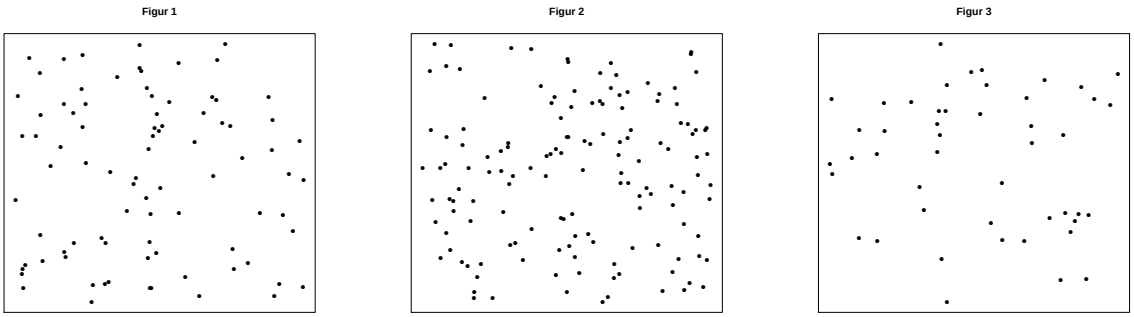
3. Beregn gennemsnit og stikprøvespredning for variablen `hip`.
4. Brug formel (5.21) til at bestemme et 95% konfidensinterval for middelværdien μ. Husk at du kan bruge `qt` til at finde den relevante t-fordelingsfraktil (se slides eller side 140 i bogen).

   Overvej nøje hvad fortolkningen af konfidensintervallet er.
5. Brug funktionen `t.test` til at bestemme det samme konfidensinterval (se slides eller side 138 i bogen).
6. Brug funktionen `lm` til at bestemme det samme konfidensinterval (se slides eller side 138 i bogen).

Opgave HS.11 (Gæt på antal punkter)

På *Statistisk Datanalyse 1* i 2017 viste Helle Sørensen de studerende tre punktplot og bad dem gætte på hvor mange punkter der var i hvert plot. De fik cirka 5 sekunder til at se på hver figur inden de skulle gætte.

Figurerne er vist nedenfor. Det sande antal punkter i figurerne er 86, 142 og 47.



De studerendes gæt er tilgængelige i filerne `punktplot2017.xlsx` og `punktplot2017.txt`. Variablene hedder **figur1**, **figur2** og **figur3**.

1. Indlæs data i R. Nedenfor kaldes datasættet **punktplot2017**.
2. Tegn histogrammer og QQ-plots for variablen **figur1** og derefter for den log-transformerede variabel **log(figur1)**. Er det mest rimeligt at antage at **figur1** eller **log(figur1)** er normalfordelt?

Den såkaldte *variationskoefficient* (coefficient of variation, CV) for en variabel y er defineret som

$$\text{CV}(y) = \frac{\text{sd}(y)}{\text{mean}(y)} = \frac{s}{\bar{y}}$$

Efter indlæsning af data til datasættet **punktplot2017**, kan vi for variablen **figur1** beregne gennemsnit, stikprøvespredning, variarionskoefficient og spredning af log-data på følgende måde:

```
> mean(punktplot2017$figur1)
[1] 70.69231
> sd(punktplot2017$figur1)
[1] 23.5597
> sd(punktplot2017$figur1) / mean(punktplot2017$figur1)
[1] 0.3332711
> sd(log(punktplot2017$figur1))
[1] 0.3285933
```

3. De beregnede tal er indsat i første linje i nedenstående tabel. Udfyld resten af tabellen.

Figur	Korrekt antal punkter	$\bar{y}$	s	CV	sd for log(figur)
Figur 1	86	70.7	23.6	0.333	0.329
Figur 2	142				
Figur 3	47				

4. Kommentér tallene i tabellen: (a) Hvad sker der med s når $\bar{y}$ vokser. Kan du forklare hvorfor? (b) Hvad betyder det at CV er stort set konstant? (c) Hvis du skulle analysere data fra alle tre figurer samtidig; hvilken skala ville det så være hensigtsmæssigt at arbejde på? (d) Hvornår kan det hjælpe at log-transformere data?

5. Brug kommandoerne

   ```
   plot(punktplot2017)
   cor(punktplot2017)
   ```

   Kommentér resultaterne.

Opgave HS.12 (Ekstra spørgsmål til opgave 4.5)

Dette er ekstra spørgsmål til opgave 4.5 i lærebogen.

1. Bestem et interval som vi vil forvente at 95% af alle gestationstider ligger i. Du kan antage at gestationstiden er normalfordelt med middelværdi 341.08 og spredning 3.07.

2. Brug `t.test` til at bestemme et 95% konfidensinterval for den gennemsnitlige gestationstid.

3. Forklar forskellen på de to intervaller. Hvilket er det bredeste? Hvorfor?