Introduktion till R

Pär Leijonhufvud

(September 14, 2022)

Översikt

- Alla har – eller kommer efter idag(?) – att ha följande installerat på sin dator:
  - R, helst 4.x
  - RStudio
  - ett antal bibliotek (libraries) som gör livet enklare
- All kod kommer att finnas dels i
  - Teams, och dels på
  - https://leijonhufvud.org/R-seminarier2022/

Kommandorad eller GUI?

Grunden

Kommandorad, med kommandon eller skript som exekveras

GUI

Det finns ett antal, t.ex. R Commander eller Blue Sky.

RStudio

Texteditor: VIM, EMACS, Notepad++, ed. . .

RStudio: En utvecklingsmiljö för R, med fönster för kommandon, skript, output, osv. Ger också hjälp med "tab-completion"

Skriptspråk

De kommandon du skriver – t.ex. `median(acovid$Mätvärde)` – kan även sättas ihop till ett skript.

```
> summary(ToothGrowth)
      len        supp         dose
 Min.   : 4.20   OJ:30   Min.   :0.500
 1st Qu.:13.07   VC:30   1st Qu.:0.500
 Median :19.25           Median :1.000
 Mean   :18.81           Mean   :1.167
 3rd Qu.:25.27           3rd Qu.:2.000
 Max.   :33.90           Max.   :2.000
> long.tooth <- ToothGrowth[ ToothGrowth$len > 20, ]
> summary(long.tooth)
      len        supp         dose
 Min.   :21.20   OJ:18   Min.   :0.500
 1st Qu.:23.30   VC:10   1st Qu.:1.000
 Median :25.50           Median :2.000
 Mean   :25.73           Mean   :1.661
 3rd Qu.:26.85           3rd Qu.:2.000
 Max.   :33.90           Max.   :2.000
>
```

```
> summary(ToothGrowth)
  long.tooth <- ToothGrowth[ ToothGrowth$len > 20, ]
  summary(long.tooth)
```

.

```
source("skript1.R", encoding="UTF-8")
```

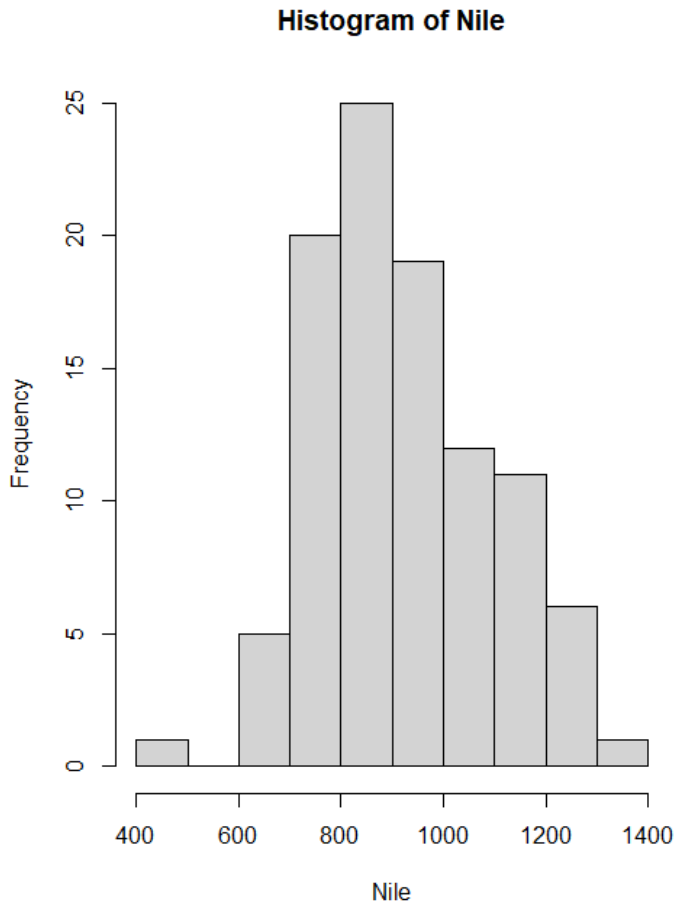
Att installera paket

```
# Installera
> install.packages("tidyverse")
# Ladda in
> library(tidyverse)
```

Lite basala statistik-kommandon

```
> min(Nile)
456
> max(Nile)
1370
> summary(Nile)
Min.  1st Qu.  Median Mean 3rd Qu.  Max.
456.0 798.5 893.5 919.4 1032.5 1370.0
```

Ett enkelt histogram



Ta ut delar av ett objekt

- `Nile[2]` Det andra värdet i Nile
- `Nile[c(1,3,5)]` Det första, tredje och femte värdet
- `mean(Nile[1:50]` medelvärdet av de första 50 värdena
- `nilen.1.50 <- Nile[1:50]` medelvärdet av de första 50 värdena sparas i ett nytt objekt

Värden som matchar ett krav

```
> antal_över_1200 <- sum(Nile>1200)
> antal_över-1200
[1] 7
```

Värden som matchar ett krav

```
> antal_över_1200 <- sum(Nile>1200)
> antal_över-1200
[1] 7
```

```
> nilen1200 <- which(Nile > 1200)
> nilen1200
[1]  4  8  9 22 24 25 26
> length(nilen1200)
[1] 7
```

Eller så kan vi ta en genväg, och ber R att visa oss alla värden i Nile som matcha

```
> Nile[Nile > 1200]
[1] 1210 1230 1370 1210 1250 1260 1220
```

Läsa in en tabell med data

Enklast: CSV

`read_csv2("väg/till/filen.csv", fileEncoding="UTF-8-BOM")` (read_csv2: en genväg till en fil med semikolon, decimalkomma, osv.

Excel-filer

Går bra, men tar mer minne

```
read_xlsx("filen.xlsx", sheet=1)
```

Läsa in data från CSV eller XLSX

```
> acovid <- read.csv2("rjh-acovid.csv", fileEncoding = "UTF-8-BOM")
> str(acovid)
'data.frame':   2044 obs. of  9 variables:
 $ Kön                 : chr  "K" "K" "K" "M" ...
 $ Ålder..År.          : chr  "6,73" "3,32" "5,8" "47,48" ...
 $ Beställare          : chr  "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BA
 $ Analys              : chr  "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS
 $ Registreringstid    : chr  "2021-12-17 16:20" "2021-08-12 11:49" "2021-06-03 19:28" "2021-03-26 17:5
 $ Tekniskt.godkännande: chr  "2021-12-17 18:14" "2021-08-12 13:58" "2021-06-03 21:55" "2021-03-26 20:1
 $ Ankomsttid          : chr  "2021-12-17 17:27" "2021-08-12 12:00" "2021-06-03 19:52" "2021-03-26 18:0
 $ Ledtid              : chr  "00:47:00" "01:58:00" "02:03:00" "02:14:00" ...
 $ Mätvärde            : num  0.07 0.08 0.07 0.06 0.06 0.06 0.06 9.32 0.06 0.06 ...
> acovid.xlsx <- readxl::read_xlsx("rjh-acovid.xlsx",sheet = 1)
> str(acovid.xlsx)
tibble [2,044 × 9] (S3: tbl_df/tbl/data.frame)
 $ Kön                 : chr [1:2044] "K" "K" "K" "M" ...
 $ Ålder (År)          : chr [1:2044] "6.73" "3.32" "5.8" "47.48" ...
 $ Beställare          : chr [1:2044] "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ
 $ Analys              : chr [1:2044] "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-an
 $ Registreringstid    : POSIXct[1:2044], format: "2021-12-17 16:20:00" "2021-08-12 11:49:00" "2021-06-
 $ Tekniskt godkännande: POSIXct[1:2044], format: "2021-12-17 18:14:00" "2021-08-12 13:58:00" "2021-06-
 $ Ankomsttid          : chr [1:2044] "43085.727083333331" "42958.5" "42888.827777777777" "42819.75" ..
 $ Ledtid              : chr [1:2044] "3.2638888888888891E-2" "8.1944444444444445E-2" "8.5416666666666
 $ Mätvärde            : num [1:2044] 0.07 0.08 0.07 0.06 0.06 0.06 0.06 9.32 0.06 0.06 ...
>
```

Vad finns i en data frame

```
> acovid <- read.csv2("rjh-acovid.csv", fileEncoding = "UTF-8-BOM")
> str(acovid)
>'data.frame':  2044 obs. of  9 variables:
 $ Kön                 : chr  "K" "K" "K" "M" ...
 $ Ålder..År.          : chr  "6,73" "3,32" "5,8" "47,48" ...
 $ Beställare          : chr  "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN"
 $ Analys              : chr  "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-Co
 $ Registreringstid    : chr  "2021-12-17 16:20" "2021-08-12 11:49" "2021-06-03 19:28" "2021-03-26 17:54"
 $ Tekniskt.godkännande: chr  "2021-12-17 18:14" "2021-08-12 13:58" "2021-06-03 21:55" "2021-03-26 20:14"
 $ Ankomsttid          : chr  "2021-12-17 17:27" "2021-08-12 12:00" "2021-06-03 19:52" "2021-03-26 18:00"
 $ Ledtid              : chr  "00:47:00" "01:58:00" "02:03:00" "02:14:00" ...
 $ Mätvärde            : num  0.07 0.08 0.07 0.06 0.06 0.06 0.06 9.32 0.06 0.06 ...
> head(acovid)
  Kön Ålder..År.       Beställare           Analys Registreringstid Tekniskt.godkännande       Ankomstti
1   K       6,73 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-12-17 16:20     2021-12-17 18:14 2021-12-17 17:2
2   K       3,32 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-08-12 11:49     2021-08-12 13:58 2021-08-12 12:0
3   K        5,8 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-06-03 19:28     2021-06-03 21:55 2021-06-03 19:5
4   M      47,48 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-03-26 17:54     2021-03-26 20:14 2021-03-26 18:0
5   K      17,44 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-03-05 23:49     2021-03-06 01:18 2021-03-05 23:5
6   K      14,85 ÖSJ   AKUT  BARN S-anti-SARS-CoV-2 2021-02-24 06:48     2021-02-24 14:18 2021-02-24 08:3
```

Arbeta med en data frame

```
> median(acovid$Mätvärde)
[1] NA
> median(acovid$Mätvärde, na.rm=TRUE)
[1] 0.1
```

Utdrag ur en data frame

```
> acovid[5,3]
[1] "ÖSJ   AKUT  BARN"
>
> acovid[2:7,c(1,3,9)]
  Kön       Beställare Mätvärde
2   K ÖSJ   AKUT  BARN     0.08
3   K ÖSJ   AKUT  BARN     0.07
4   M ÖSJ   AKUT  BARN     0.06
5   K ÖSJ   AKUT  BARN     0.06
6   K ÖSJ   AKUT  BARN     0.06
7   M ÖSJ   AKUT  BARN     0.06
>
```

Ta ut rader md bara kvinnor

```
> acovid.kvinnor <- acovid[ acovid$Kön == "K", ]
```

Vad vi säger är

- Alla rader där kolumnen "Kön" har värdet "K"
- Alla kolumner

Tidyverse vs base R

```
> library(tidyverse)
> acovid.k.positiva <- filter(acovid, Kön == "K" & Mätvärde > 20 )
```

Vill vi hålla oss till "base R" går det också:

```
> acovid.k.positiva <- acovid[ (acovid$Kön == "K" & acovid$Mätvärde > 20), ]
```

Vi kan enkelt kolla att vi fick med det vi ville:

```
> nrow(acovid.k.positiva)
[1] 160
> unique(acovid.k.positiva$Kön)
[1] "K"
> summary(acovid.k.positiva$Mätvärde)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.43   41.27   74.00   88.70  129.25  317.00
```

Tidyverse vs base R

```
> acovid.arb$Datum.A <- as.Date(acovid.arb$Ankomsttid)
> str(acovid.arb)
'data.frame':   2044 obs. of  10 variables:
 $ Kön                 : chr  "K" "K" "K" "M" ...
 $ Ålder..År.          : chr  "6,73" "3,32" "5,8" "47,48" ...
 $ Beställare          : chr  "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" "ÖSJ   AKUT  BARN" ...
 $ Analys              : chr  "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" "S-anti-SARS-CoV-2" ...
 $ Registreringstid    : chr  "2021-12-17 16:20" "2021-08-12 11:49" "2021-06-03 19:28" "2021-03-26 17:54" ...
 $ Tekniskt.godkännande: chr  "2021-12-17 18:14" "2021-08-12 13:58" "2021-06-03 21:55" "2021-03-26 20:14" ...
 $ Ankomsttid          : chr  "2021-12-17 17:27" "2021-08-12 12:00" "2021-06-03 19:52" "2021-03-26 18:00" ...
 $ Ledtid              : chr  "00:47:00" "01:58:00" "02:03:00" "02:14:00" ...
 $ Mätvärde            : num  0.07 0.08 0.07 0.06 0.06 0.06 0.06 9.32 0.06 0.06 ...
 $ Datum.A             : Date, format: "2021-12-17" "2021-08-12" "2021-06-03" "2021-03-26" ...
> summary(acovid.arb$Datum.A)
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.         NA's
"2020-12-31" "2021-01-20" "2021-02-10" "2021-02-23" "2021-03-02" "2021-12-30"        "142"
```

Var jobbar du: getwd() och setwd()

```
> getwd()
```

Ibland kan det vara bra att kunna byta arbetskatalog, och givetvis kan R det med

```
> setwd("C:/Min/nya/arbetakatalog")
```

Om du vill använda dig av en relativ sökväg går det också bra:

```
> setwd("../arbetakatalog2")
```

Notera att R inte använder sig av bakåtslaskar även på en Windows-dator: vill du använda dig av dem måste d

```
> setwd("C:\\Min/nya\\arbetakatalog")
```

Avsaluta din R-session

```
> q() # eller
> quit()
```

Alla de kommandon du skrivit sparas i ".Rhistory"