Brev till Block 2

Hej!
Här kommer ert andra brev. Ofta, när det krävs kortare texter, använder vi mig av anslagstavlan istället för formen brev.

Idag ska ni få ett litet läromål i statistik: **variabler och skalor**.

Läromål i det här brevet
Lägg märke till att det finns olika typer av variabler, och att varje typ av variabel har en tillhörande typ av skala. Det får konsekvenser för vilka statistiska metoder du kan använda för att beskriva eller utföra beräkningar med en viss variabel.

Läs om variabler och skalor i brevet,
se filmen i Explore Statistics with R:
https://play.ki.se/media/StatRx_W1_F1+Welcome+to+Explore+Statistics+with+R/0_y56pkarw

eller använd någon annan text. I brevet finns några litteraturtips. Det finns även R-kod att testa: Gör ett enkelt stapeldiagram.

Installera R
Kolla Learning Sequence: Download and install R.
Vad betyder Learning Sequence? Det är som ett kapitel i materialet Explore Statistics with R.
Om du har problem, skriv en rad i kursens diskussionsforum ”Ladda ner och installera R”.

Paketet ISwR
När du arbetar med kursboken, Introductory Statistics with R **(ISwR)**, är det viktigt att du har datorn till hands och testar och kanske ändrar i bokens kod-exempel. För att kunna följa med är det lämpligt att installera ett paket som hör till boken och som har bokens förkortning som namn, ISwR.

Vad är ett paket?
Många statistik- och matematikprogram levereras som en grundinstallation med allmän funktionalitet. Till det kan köpa eller få tillägg med särskilda funktioner. Tilläggen kan heta modul, paket, toolbox, add-on eller liknande. I R används termen paket och det finns för närvarande mer än 3000 paket med diverse funktionalitet att ladda ner från CRAN (The Comprehensive R Archive Network, r-project.org). Ett paket kan innehålla funktioner och även dataset. Appendix B (ISwR sid 293-323) listar de dataset som finns i paketet ISwR.

####### här kommer åtta rader som du kan kopiera in i R console
#Snabb repetition, kommando i R för att installera bokens paket:
install.packages("ISwR")
#glöm inte att också ladda paketet innan du använder det:
library(ISwR)
#testa att paketet har laddats ordentligt genom att be att få titta i datasetet stroke:
stroke
######
du bör få upp en lång datatabell som börjar så här:
> stroke

	sex	died	dstr	age	dgn	coma	diab	minf	han	dead	obsmonths
1	Male	1991-01-07	1991-01-02	76	INF	No	No	Yes	No	TRUE	0.16339869
2	Male	<NA>	1991-01-03	58	INF	No	No	No	No	FALSE	59.60784314

Nu över till grundläggande statistisk terminologi, variabler och skalor.

Variabler
En variabel är ett namn som används för att representera ett okänt värde. Det är användbart när du vill beskriva en beräkning utan att använda konkreta värden. På sidan 5 (ISwR) ser du tre variabler i en känd formel för att beräkna body mass index:
bmi <- weight / height^2
När man väljer bland statistiska metoder att visualisera eller räkna på variabler måste man ha klart för sig vilken typ av variabel man har att göra med. Man delar in variabler i kvalitativa och kvantitativa.

En kvalitativ variabel är icke-numerisk, tex kön, yrke, civilstånd, bilmärke. Det finns inget naturligt sätt att sätta variabelns värden i någon ordning. Ändå är det brukligt att använder sig av siffror när man kodar och lagrar kvantitativa data, tex kvinna=0, man=1.
Kvantitativa, numeriska, variabler delas in i diskreta och kontinuerliga. Antal barn i familjen är en kvantitativ variabel, men den kan bara anta vissa värden, heltal, därför är det en diskrekt variabel. Likaså är antalet bakteriekolonier som växer på en agarplatta en diskret variabel. En kvantitativ variabel kan också vara kontinuerlig, som till exempel temperatur. Man känner igen den kontinuerliga variabeln genom att den kan anta vilket värde som helst inom ett intervall. Mellan två angivna temperaturer kan man alltid ange en temperatur.

Skalor
Lägg märke till hur olika typer av variabler passar ihop med olika typer av skalor. Kvalititativa variabler kan placeras in i nominalskala. Nominal betyder namn och för tanken till att ordningsföljden inte har någon betydelse. Du kan rita ett stapeldiagram som visar könsfördelningen på arbetsplatsen. I så fall är x-axeln en nominalskala. Här följer en enkel strategi i R för att rita ett stapeldiagram. Klistra in detta i R och se hur det blir.

```
#först skapar vi en vektor med antalet kvinnor och antalet män
gender <- c(33,24)
#i barplotkommandot lägger vi in namn för staplarna, samt en titel för diagrammet
barplot(gender,names.arg=c("Male","Female"), main="Gender")
```

Vad hander om du lägger till argumentet ylim=c(0,50) i kommandot?
Pröva gärna att läsa dokumentationen för barplot så här:

?barplot()

I vilken ordning löparna går i mål i en tävling är exempel på kvantitativ diskret data som passar på en ordinalskala. Här har det betydelse i vilken ordning observationerna läggs, men man kan inte säga något om avståndet mellan två observationer. Kanske gick en löpare i mål långt före klungan, kanske gick tre personer i mål nästan samtidigt. Den informationen har gått förlorad när man använder ordinalskala. Att ange temperatur med Celsius skala är ett exempel på intervallskala. Tio graders skillnad betyder samma sak oavsett var på skalan du börjar. Däremot kan man inte säga att 20 grader celcius är dubbelt så varmt som 10 grader celcius eftersom skalan inte startar i en meningsfull nollpunkt. För att multiplikation och division ska vara meningsfullt så krävs data på en kvotskala, så som när man anger temperatur i Kelvin som utgår ifrån absoluta nollpunkten.

Kommentar om statistiklitteratur
Det kan vara bra att komplettera ISwR med en text om statistik. I kursplanen listas två böcker. Detta brev berör variabler och skalor. Om det kan man läsa i
kap 1, Lind/Marchal/Wathen, Statistical techniques in business & economics

Om du inte har någon bok vill vi rekommendera resurser på internet

Tänk på att du har gratis tillgång till eböcker på kib.ki.se
Ta chansen att se vad som finns där.

På internet finns gratis bland annat detta:
https://docs.tibco.com/data-science/GUID-4306E8B6-8C8F-4AC0-911F-B6B22FF03C3C.html

Även wikipedia kan vara riktigt bra på grundläggande statistik
http://en.wikipedia.org/wiki/Statistics