

Statistik 1 2009

Övning 2, v38

Hämta följande datafil: <http://www.abo.fi/fak/mnf/mate/jc/statistik1/LWPatienter.sav>

Filen innehåller data för 140 astmapatienter med avseende på följande variabler:

Variabelnamn:	Beskrivning:
Patno	Patient number
Treatm	Treatment
Sex	Gender
Age	Age (yr)
Weight	Weight (kg)
Height	Height (cm)
Smoke	Smoking habits

Starta R och sedan R Commander med kommando:

library(Rcmdr)

Importera den ovannämnda datafilen till R Commander (*Data-Import...-From SPSS...*) och namnge den **LWFPatienter**. I många situationer vill man utföra statistiska analyser separat för vissa delmängder av data, t ex skilt för män och kvinnor i detta exempel. Ävenom många analysfunktioner i R Commander tillåter valet av delmängder, kan det vara smidigt att skapa ett permanent dataset som innehåller den önskade delmängden ifall man behöver utföra flera olika typer av analyser. För att exempelvis skapa en datamatrix som innehåller variabelvärden endast för kvinnor, gör på följande sätt. Välj *Data-Active...-Subset...* och fyll det önskade villkoret i rutan *Subset expression*. Programmet förväntar sig ett uttryck med logiska operatorer, t ex följande uttryck

SEX == "Female"

väljer alla rader i datamatriken där variabeln SEX har värdet Female. Det finns två likhetstecken efter varandra och värdet placeras här mellan citationstecken pga att det är en textsträng. **Notera att programmet skiljer mellan gemener och versaler både i variabelnamn och i variabelvärden!** Man kan lätt granska dessa egenskaper antingen i rutan *Edit data set* eller i *View data set* innan man börjar skapa delmängder. För numeriska variabler behövs ej citationstecknen, t ex med uttrycket **WEIGHT < 70** skulle man välja samtliga individer vars vikt understiger 70kg. I rutan *Name for new data set* kan man ange ett nytt namn för datasetet. Det lönar sig att använda informativa namn, t ex **LWFPatienterKvinnor** i exemplet ovan. Om man har under en arbetsession importerat eller skapat fler än en datamatrix, kan man växla mellan dem och välja den aktiva datamatriken genom att gå till menyn *Data-Active...-Select active data set*. Observera att analyser görs utifrån den datamatriken som senast valts.

1. Skapa två nya datamatrixer utifrån LWFPatienter, en för kvinnor och en för män (den senare erhålls med uttrycket **SEX == "Male"**). Bestäm skilt för männen och kvinnorna korrelationsmatrisen (använd Spearmans koefficient) för variablerna AGE, WEIGHT och HEIGHT (menyn *Statistics-Summaries-Correlation Matrix*). Obs! Håller man CTRL-knappen i botten, kan flera variabler väljas med musens

- vänsterklick. Testa om korrelationen mellan WEIGHT och HEIGHT skiljer sig signifikant från 0 på 5%-nivån (skilt för män och kvinnor). Detta kan göras i menyn *Statistics-Summaries-Correlation test*. Välj nu datamatrixen **LWFPatienter** som den aktiva (*Data-Active...-Select active data set*) och konstruera en matris med spridningsdiagram (scatterplot) för AGE, WEIGHT och HEIGHT separat för männen och kvinnorna. Matrisen med spridningsdiagram erhålls från menyn *Graphs-Scatterplot matrix*, välj de lämpliga variablerna och sätt *Plot by groups* lika med **SEX**. I grafikfönstret kan man exportera grafiken till en rad olika format via menyn *File-Save as*. Då kan bilderna lätt importeras till ett ordbehandlingsprogram såsom Word.
2. Fortsätt med datamatrixen **LWFPatienter**. Gör en regressionsanalys med WEIGHT som beroende variabel och HEIGHT som förklarande variabel separat för männen och kvinnorna. Detta görs via *Statistics-Fit models-Linear regression*. Välj de rätta variablerna och fyll i det lämpliga logiska uttrycket i rutan *Subset expression* (t ex **SEX == "Female"** för att utföra analysen för kvinnor). Notera att varje gång man anpassar en statistisk modell, anges den automatiskt ett namn i rutan *Enter name for model* (man kan även ange ett eget namn). Programmet lagrar alltså uppgifterna om anpassningen i ett objekt som bär detta namn. Det möjliggör att man kan återskapa resultaten för en valbar modell då man utför flera analyser. Den aktiva modellen väljes från *Models-Select active model*. Statistiska sammanfattningen av modellen anpassningen kan då erhållas på nytt i utskriftsfönstret med *Models-Summarize model*. Konfidensintervallen för modellparametrar kan erhållas från *Models-Confidence intervals*.
 3. Rita spridningsdiagram för WEIGHT och HEIGHT där regressionslinjerna är synliga. Välj *Graphs-Scatterplot* och lägg variablerna i x- och y-rutan, samt välj **SEX** i *Plot by groups* (då används skilda symboler för män och kvinnor).
 4. Titta på fördelningen över rökvanorna (variabel SMOKE) med hjälp av ett stapeldiagram och ett cirkeldiagram (*Graphs-Bar graph* och *Graphs-Pie chart*).