Laboration 1

Ottilia Andersson, David Sermoneta

2024-11-24

Sammanfattning

Den här laborationen går ut på att utforska olika metoder för att avgöra huruvida data i ett stickprov kan anses komma från en normalfördelning eller ej.

I del 1 besvarar vi två frågor: (a) Hur stort stickprov behöver man ha för att kunna avgöra om data är normalfördelade? och (b) Vilken grafisk metod är bäst för att avgöra om data är normalfördelade? Vi kommer fram till att stickprovsstorleken bör vara åtminstone $n = 40$ för normalfördelade, $n = 100$ för likformigt fördelade, respektive $n = 30$ för exponentialfördelade stickprov. Om man inte vet vilken fördelning ett stickprov kommer ifrån är det således säkrast att välja $n = 100$ som minsta storlek på sitt stickprov. Vidare tycker vi att bland de tre metoder vi undersöker (boxplot, normalfördelningsplot och histogram), så är normalfördelningsplotten det mest effektiva sättet att enkelt med ögat avgöra om data är normalfördelade eller ej.

I del 2 använder vi dessa kunskaper och utforkar ett datamaterial över alkoholkonsumtionen i ett urval av OECD-länder. Vi kan inte konstatera något samband mellan konsumtionen av de olika alkoholtyperna. Vi diskuterar de olika ländernas placering i förhållande till varandra.

Uppgift 1: Kommer data från en normalfördelning?

I den här uppgiften vill vi undersöka hur vi kan avgöra om ett stickprov kommer från en normalfördelning eller inte. Vi har två huvudsakliga syften. Det första är att (a) ta reda på hur mycket data som behövs för att kunna besvara den frågan, med andra ord: hur stort behöver ett stickprov vara? (Hur stort värde på n behövs?) Det andra är att (b) komma fram till vilken metod vi anser vara mest effektiv för ändamålet, utifrån att bedömningen görs helt grafiskt.

För att besvara dessa två frågor har vi simulerat olika stickprov från för oss redan kända fördelningar: normalfördelade, likformigt fördelade respektive exponentialfördelade data. Vi har tittat på tre olika typer av grafiska metoder: boxplot, histogram och normalfördelningsplot. Vi kommer här nedan att redovisa våra resultat utifrån fördelning, följt av en avslutande diskussion.

Normalfördelade data

För att undersöka vår redan normalfördelade data definierar vi följande åtta vektororer enligt instruktionerna.

```
set.seed(20010310)
normalstickprov1 <- rnorm(40, 24, 24) # antal genererade värden, väntevärde, standardavvikelse
normalstickprov2 <- rnorm(40, 24, 24)
normalstickprov3 <- rnorm(40, 24, 24)
normalstickprov4 <- rnorm(40, 24, 24)
normalstickprov5 <- rnorm(40, 24, 24)
normalstickprov6 <- rnorm(40, 24, 24)
normalstickprov7 <- rnorm(40, 24, 24)
normalstickprov8 <- rnorm(40, 24, 24)
```

Normalfördelade data: Histogram

För att visualisera stickprovet plottar vi data med hjälp av R-funktionen `hist` och jämför den med grafen för täthetsfunktionen till en normalfördelad slumpvariabel med väntevärde och standardavvikelse 24.

```
old_par <- par(mfrow = c(2,4), oma = c(0,0,3,0))

hist(normalstickprov1,
     breaks = seq(from = -48, to = 96, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 1",
     prob = TRUE)
# Väljer övre och undre gränser för att täcka hela stickprovet
x <- seq(from = -72, to = 96, length.out = 100)
# Grafen för täthetsfunktionen av en normalfördelad slumpvariabel
# med väntevärde och standardavvikelse lika med 24
lines(x, dnorm(x, 24, 24))

hist(normalstickprov2,
     breaks = seq(from = -48, to = 96, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 2",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(normalstickprov3,
     breaks = seq(from = -72, to = 120, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 3",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(normalstickprov4,
     breaks = seq(from = -48, to = 96, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 4",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(normalstickprov5,
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 5",
     breaks = seq(from = -48, to = 96, by = 20),
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
```

```r
lines(x, dnorm(x, 24, 24))

hist(normalstickprov6,
     breaks = seq(from = -72, to = 120, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 6",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(normalstickprov7,
     breaks = seq(from = -48, to = 96, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 7",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(normalstickprov8,
     breaks = seq(from = -48, to = 96, by = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 8",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))
mtext("Histogram över normalfördelade stickprov (n=40)", outer = TRUE, cex = 1.2, font = 2)
```

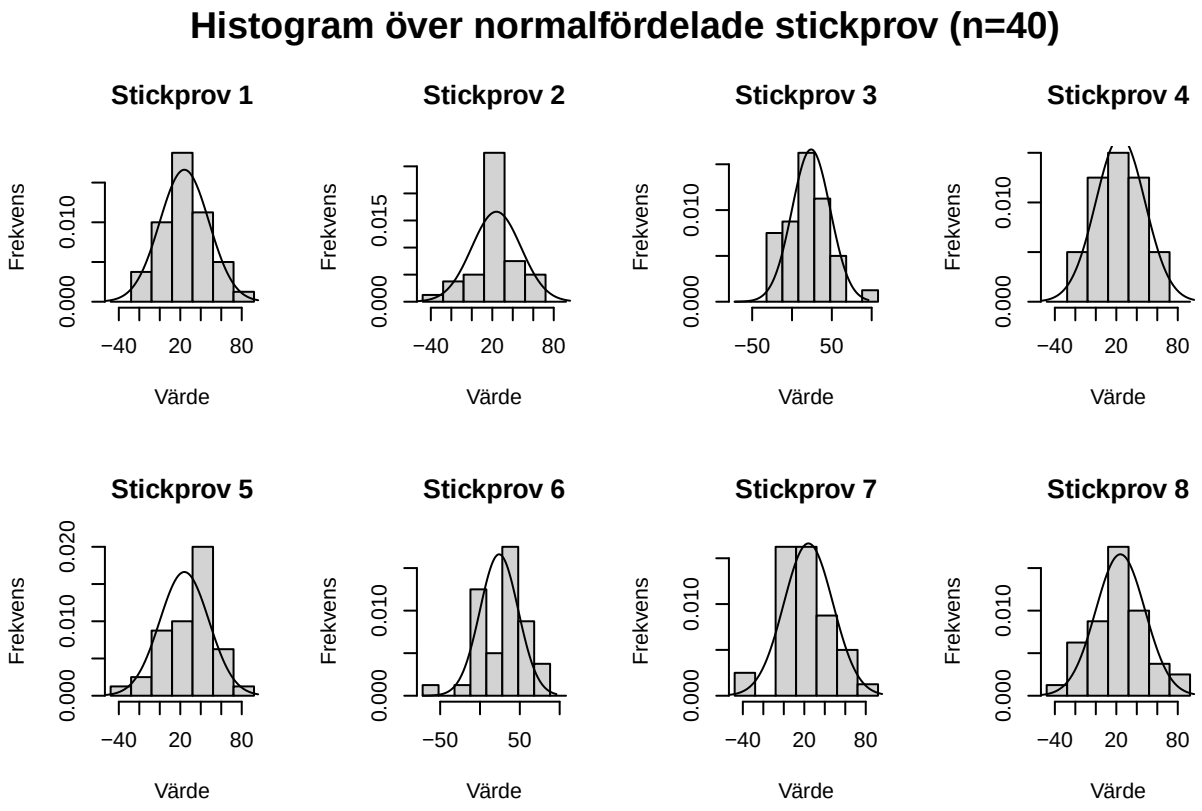

Diagram 1: Histogram över åtta normalfördelade stickprov, alla med n=40. Linjen är täthetsfunktionen till en normalfördelning med väntevärde och standardavvikelse 24.

Vi ser i figur 1 att våra stickprov är rätt så symmetriska, och endast i vissa fall skev åt antingen höger eller vänster. Med vissa undantag följer frekvensen av stickproven samma kurva som bestäms av täthetsfunktionen för en normalfördelning.

Normalfördelade data: Lådagram

För att visualisera data med hjälp av lådagram använder vi oss av den inbyggda R-funktionen `boxplot()` och matar in samma data som tidigare:

```
boxplot(normalstickprov1, normalstickprov2, normalstickprov3, normalstickprov4,
        normalstickprov5, normalstickprov6, normalstickprov7, normalstickprov8,
        horizontal = TRUE,
        main = "Lådagram över normalfördelade stickprov (n=40)",
        xlab = "Värde",
        ylab = "Stickprov")
```

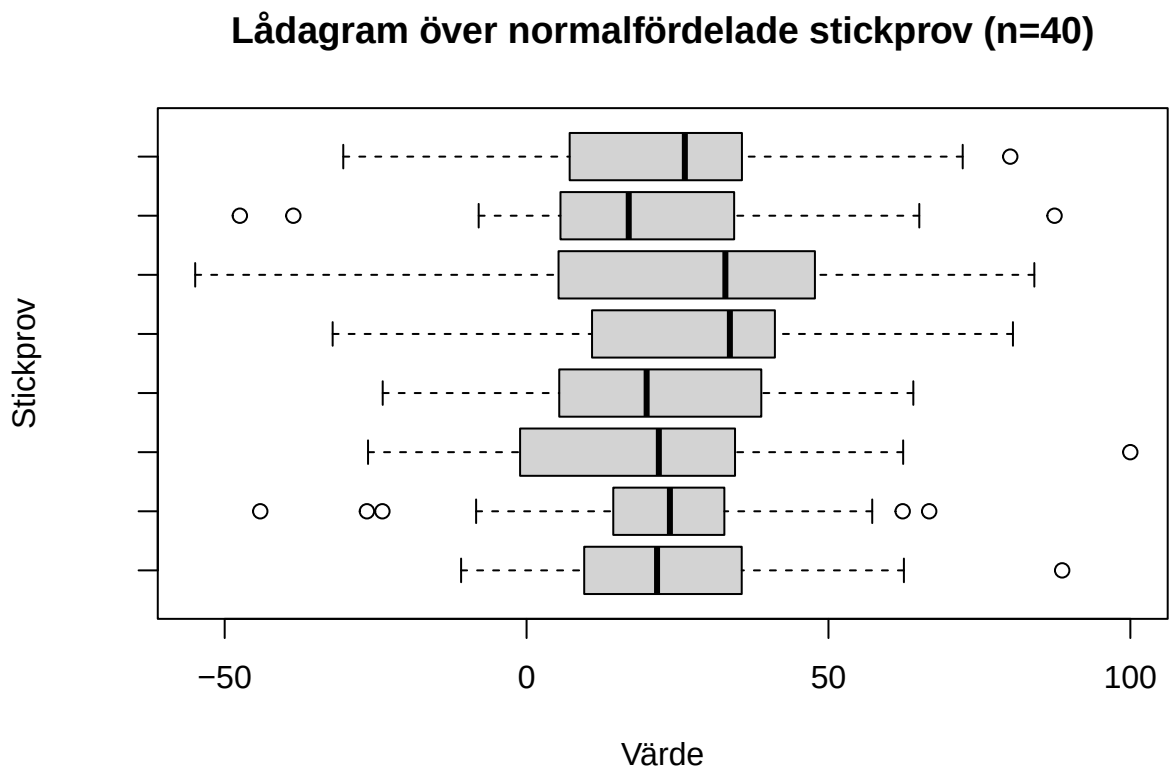

Diagram 2: Lådagram över åtta normalfördelade stickprov, alla med n=40.

Om vi tittar på lådagrammen i figur 2, så tycker vi att asymmetrin framgår tydligare. I synnerhet syns detta i hur medianen varierar. Men något som övertygar oss om att det ändå är en normalfördelning (förutom att vi redan vet om det), är att morrhåren har rätt proportion till själva lådan. Det vill säga, första och tredje kvartilerna av stickproven befinner sig i en rätt snäv låda, som ju är karaktäristiskt för en normalfördelning, och minsta och största värdena ligger långt ut på morrhåren, där de är få.

Normalfördelade data: Normalfördelningsplot

Till slut vänder vi oss till normalfördelningsplot, och här igen, använder vi oss av `R`:s inbyggda funktioner `qqnorm()` och `qqline()` för att se hur våra stickprov förhåller sig till en normalfördelning.

```
old_par <- par(mfrow = c(2,4), oma = c(0,0,3,0))
```

```r
qqnorm(normalstickprov1,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 1")
qqline(normalstickprov1)

qqnorm(normalstickprov2,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 2")
qqline(normalstickprov2)

qqnorm(normalstickprov3,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 3")
qqline(normalstickprov3)

qqnorm(normalstickprov4,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 4")
qqline(normalstickprov4)

qqnorm(normalstickprov5,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 5")
qqline(normalstickprov5)

qqnorm(normalstickprov6,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 6")
qqline(normalstickprov6)

qqnorm(normalstickprov7,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 7")
qqline(normalstickprov7)

qqnorm(normalstickprov8,
      xlab = "Teoretiska kvantiler",
      ylab = "Stickprovskvantiler",
      main = "Stickprov 8")
qqline(normalstickprov8)

mtext("Normalfördelningsplottar över normalfördelade stickprov (n = 40)",
      outer = TRUE, cex = 1.2, font = 2)
```

I figur 3 noterar vi att i de flesta stickproven, så ligger de allra flesta punkter på en rät linje. Det förekommer avvikelser, men givet att vi endast har $n = 40$ datapunkter (och att avvikelserna inte är systematiska) tycker vi att det här ger en bra idé om att stickproven faktiskt är normalfördelade.

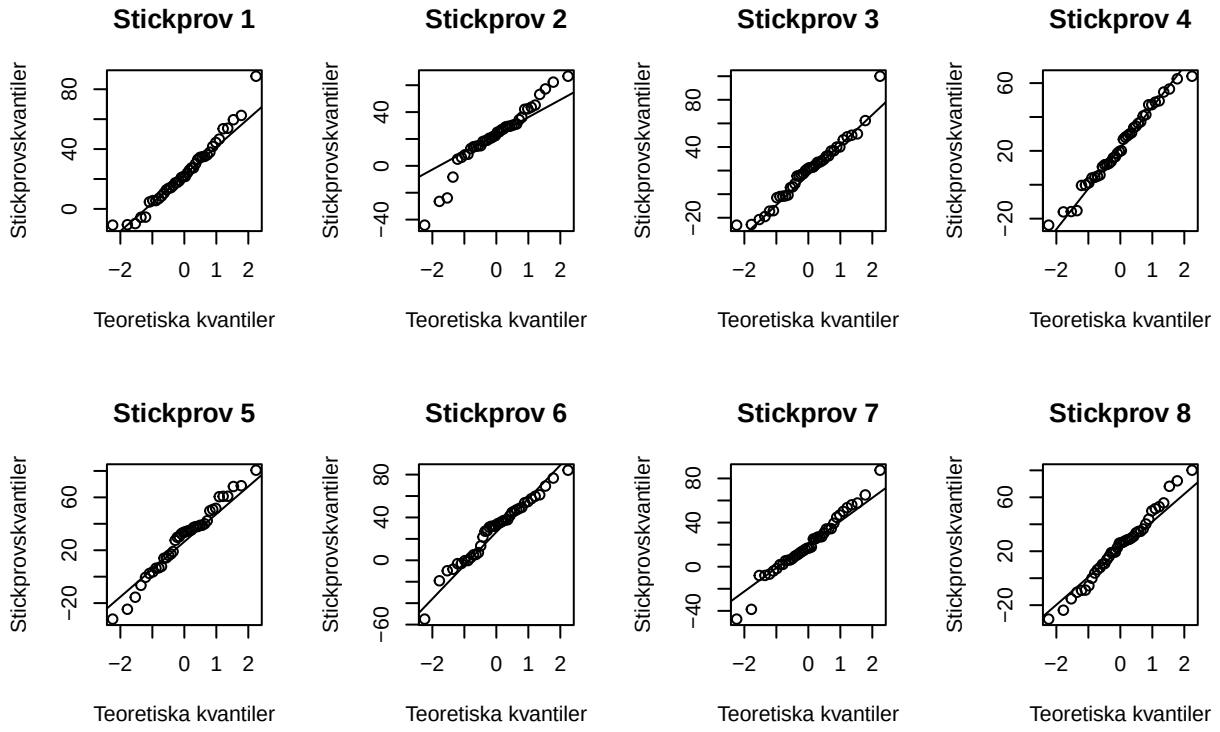

Diagram 3: Normalfördelningsplott över åtta normalfördelade stickprov, alla med n=40.

Likformigt fördelade data

För likformigt fördelade data gör vi samma sak. Med hjälp av resultatet från den teoretiska uppgiften bestämmer vi gränspunkterna för våra stickprov, enligt instruktionerna, så att väntevärde och standardavvikelse blir 24. Annars gör vi allt precis som tidigare.

```
set.seed(20010310)
 #     antal genererade värden,            min,       max
likformstickprov1 <- runif(100, 24*(1-sqrt(3)), 24*(1+sqrt(3)))
likformstickprov2 <- runif(100, 24*(1-sqrt(3)), 24*(1+sqrt(3)))
likformstickprov3 <- runif(100, 24*(1-sqrt(3)), 24*(1+sqrt(3)))
likformstickprov4 <- runif(100, 24*(1-sqrt(3)), 24*(1+sqrt(3)))
likformstickprov5 <- runif(100, 24*(1-sqrt(3)), 24*(1+sqrt(3)))
```

Likformigt fördelade data: Histogram

Vi börjar med att visualisera stickproven med hjälp av histogram, och sätter dem mot täthetsfunktionen av en normalfördelad slumpvariabel med väntevärde och standard avvikelse 24.

```
old_par <- par(mfrow = c(2,3), oma = c(0,0,3,0))

hist(likformstickprov1,
     breaks = seq(from = -48, to = 96, by = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 1",
     prob = TRUE)
```

```r
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(likformstickprov2,
     breaks = seq(from = -48, to = 96, by = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 1",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(likformstickprov3,
     breaks = seq(from = -48, to = 96, by = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 2",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(likformstickprov4,
     breaks = seq(from = -48, to = 96, by = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 3",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))

hist(likformstickprov5,
     breaks = seq(from = -48, to = 96, by = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 5",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100)
lines(x, dnorm(x, 24, 24))
mtext("Histogram över likformigt fördelade stickprov (n=100)",
      outer = TRUE, cex = 1.2, font = 2)
```

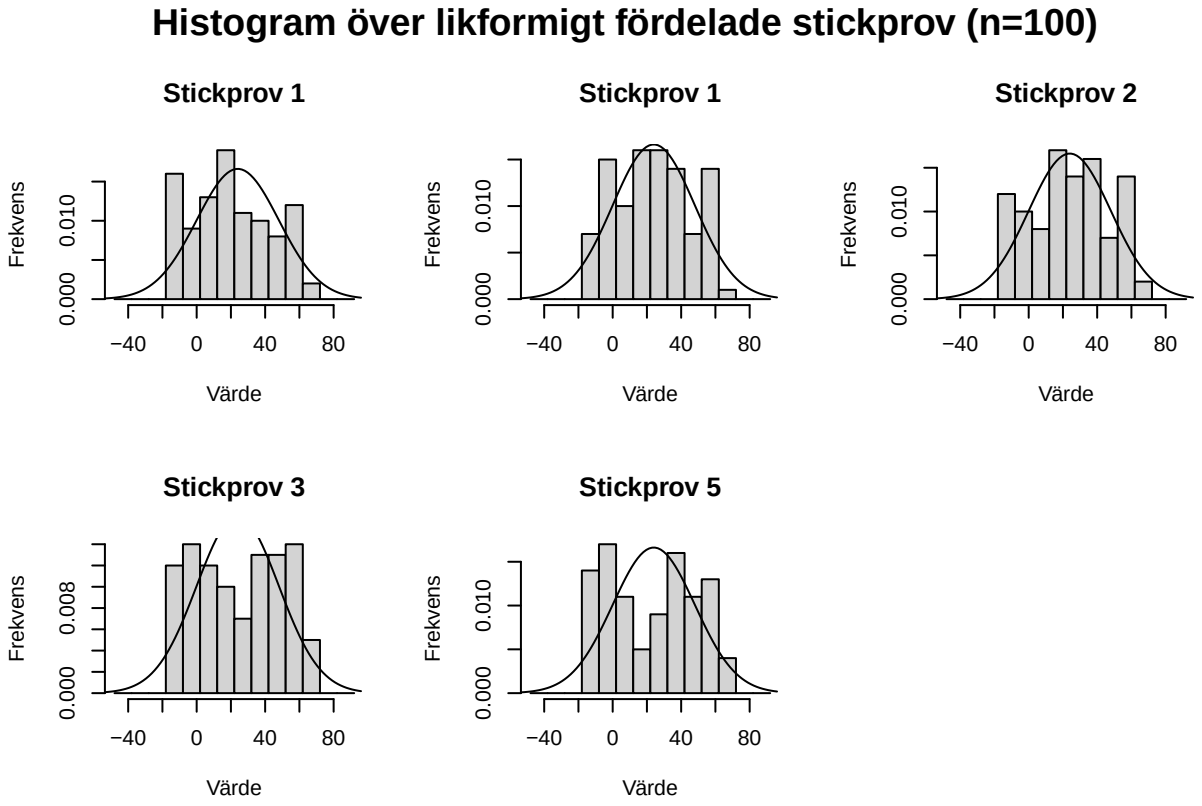

Diagram 4: Histogram över fem likformigt fördelade stickprov.

I figur 4 noterar vi att histogrammen passar kurvan av normalfördelningens täthetsfunktion sämre, där staplarna till höger/vänster om medianen är lika stora eller större som de i medianen. Detta är långt ifrån karaktäritiskt av en normalfördelning. Vidare är "svansarna" inte långa (vilket vi vill se hos normalfördelade data), och staplarna minskar inte tillräckligt gradvist ju längre ifrån vi befinner oss från medelvärdet. Givet storleken på stickproven känns det tydligt att vår data ej kommer från en normalfördelning.

Likformigt fördelade data: Lådagram

Nu undersöker vi lådagrammen för våra stickprov och ser vad vi kan utvinna för information.

```
boxplot(likformstickprov1,
        likformstickprov2,
        likformstickprov3,
        likformstickprov4,
        likformstickprov5,
        horizontal = TRUE,
        main = "Lådagram över likformigt fördelade stickprov (n=100)",
        xlab = "Värde",
        ylab = "Stickprov"
        )
```

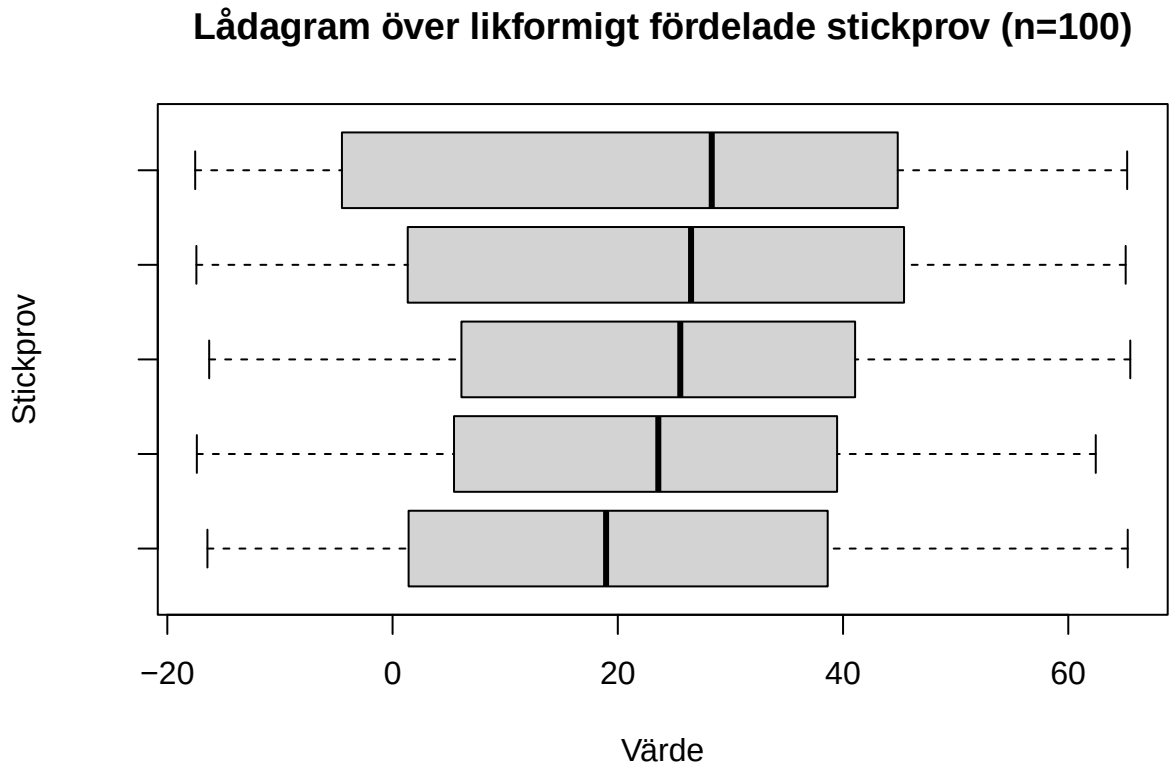

Diagram 5: Lådagram över fem likformigt fördelade stickprov.

I figur 5 ser vi ungefär samma tendenser som i histogrammen. Vidare är morrhåren inte alls särskilt långa i proportion till varsin lådas storlek, vilket tyder på att data inte följer en normalfördelning. Annars ser symmetrin bra ut, och längden på morrhåren är också ungefär lika långa på båda sidorna, så vi undersöker vidare genom normalfördelningsplottar.

Likformigt fördelade data: Normalfördelningsplot

```
old_par <-par(mfrow = c(2,3), oma = c(0,0,3,0))
qqnorm(likformstickprov1,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Stickprov 1")
qqline(likformstickprov1)
```

```r
qqnorm(likformstickprov2,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Stickprov 2")
qqline(likformstickprov2)

qqnorm(likformstickprov3,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Stickprov 3")
qqline(likformstickprov3)

qqnorm(likformstickprov4,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Stickprov 4")
qqline(likformstickprov4)

qqnorm(likformstickprov5,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Stickprov 5")
qqline(likformstickprov5)

mtext("Normalfördelningsplottar över likformigt fördelade stickprov (n=100)",
      outer = TRUE, cex = 1.2, font = 2)
```

I figur 6 ser vi att alla stickprov följer ett slags "S-form", och är lika täta genom alla kvantiler, vilket inte ska ske för normalfördelade data. Av alla tre metoder vi undersökt tycker vi att vi här, i normalfördelningsplotten, tydligast ser att dessa stickprov *inte* kan komma från en normalfördelning.

Exponentialfördelade data

Nu undersöker vi våra expnentialfördelade stickprov!

```r
set.seed(20010310)
expstickprov1 <- rexp(30, 1/24) # antal simulerade observationer, 1/väntevärdet
expstickprov2 <- rexp(30, 1/24)
expstickprov3 <- rexp(30, 1/24)
expstickprov4 <- rexp(30, 1/24)
expstickprov5 <- rexp(30, 1/24)
```

Exponentialfördelade data: Histogram

Vi visualiserar igen med histogram:

```r
old_par <- par(mfrow = c(2,3), oma = c(0,0,3,0))

hist(expstickprov1,
      breaks = seq(from = -48 , to = 96, length.out = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 1",
     prob = TRUE)
```

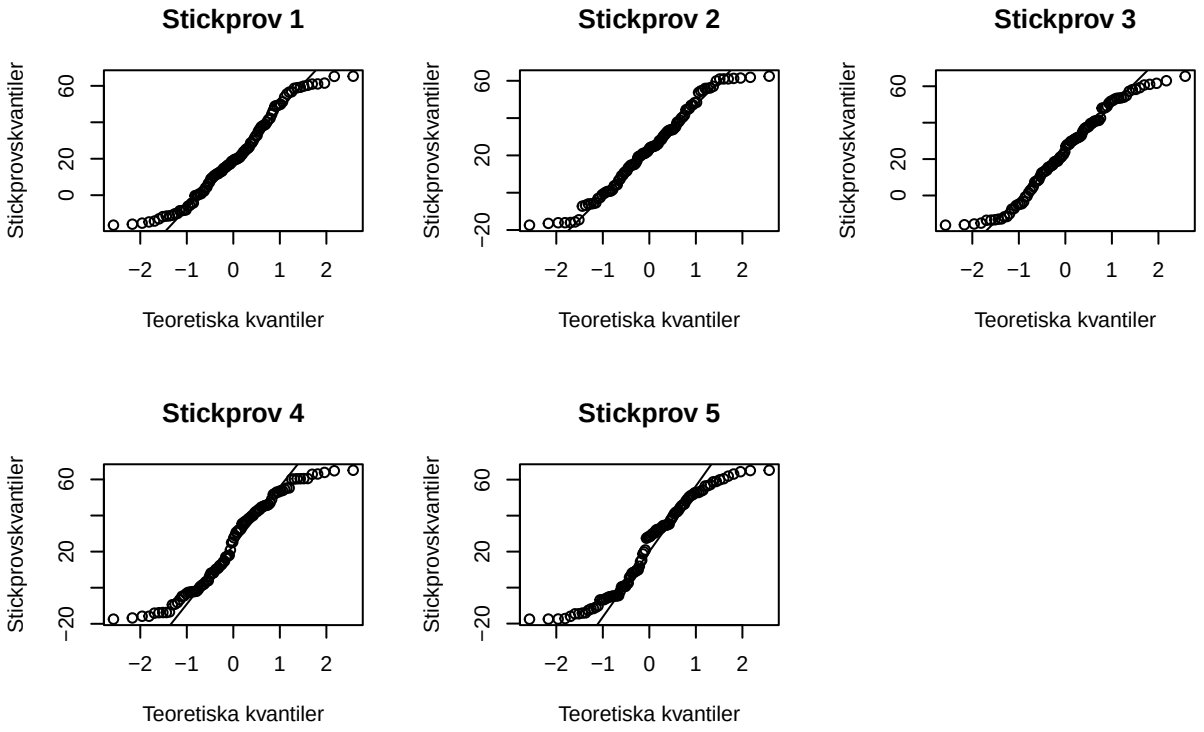

Diagram 6: Normalfördelningsplot över fem likformigt fördelade stickprov.

```
x <-  seq(from = -72 , to = 96, length.out = 100)
lines(x, dnorm(x,24,24))

hist(expstickprov2,
      breaks = seq(from = -48 , to = 96, length.out = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 2",
     prob = TRUE)
x <-  seq(from = -72 , to = 96, length.out = 100)
lines(x, dnorm(x,24,24))

hist(expstickprov3,
      breaks = seq(from = -48 , to = 96, length.out = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 3",
     prob = TRUE)
x <-  seq(from = -72 , to = 96, length.out = 100)
lines(x, dnorm(x,24,24))

hist(expstickprov4,
      breaks = seq(from = -48 , to = 96, length.out = 10),
     xlab = "Värde",
     ylab = "Frekvens",
```

```
     main = "Stickprov 4",
     prob = TRUE)
x <-  seq(from = -72 , to = 96, length.out = 100)
lines(x, dnorm(x,24,24))

hist(expstickprov5,
      breaks = seq(from = -48 , to = 96, length.out = 10),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 5",
     prob = TRUE)
x <-  seq(from = -72 , to = 96, length.out = 100)
lines(x, dnorm(x,24,24))

mtext("Histogram över exponentialfördelade stickprov (n=30)",
      outer = TRUE, cex = 1.2, font = 2)
```

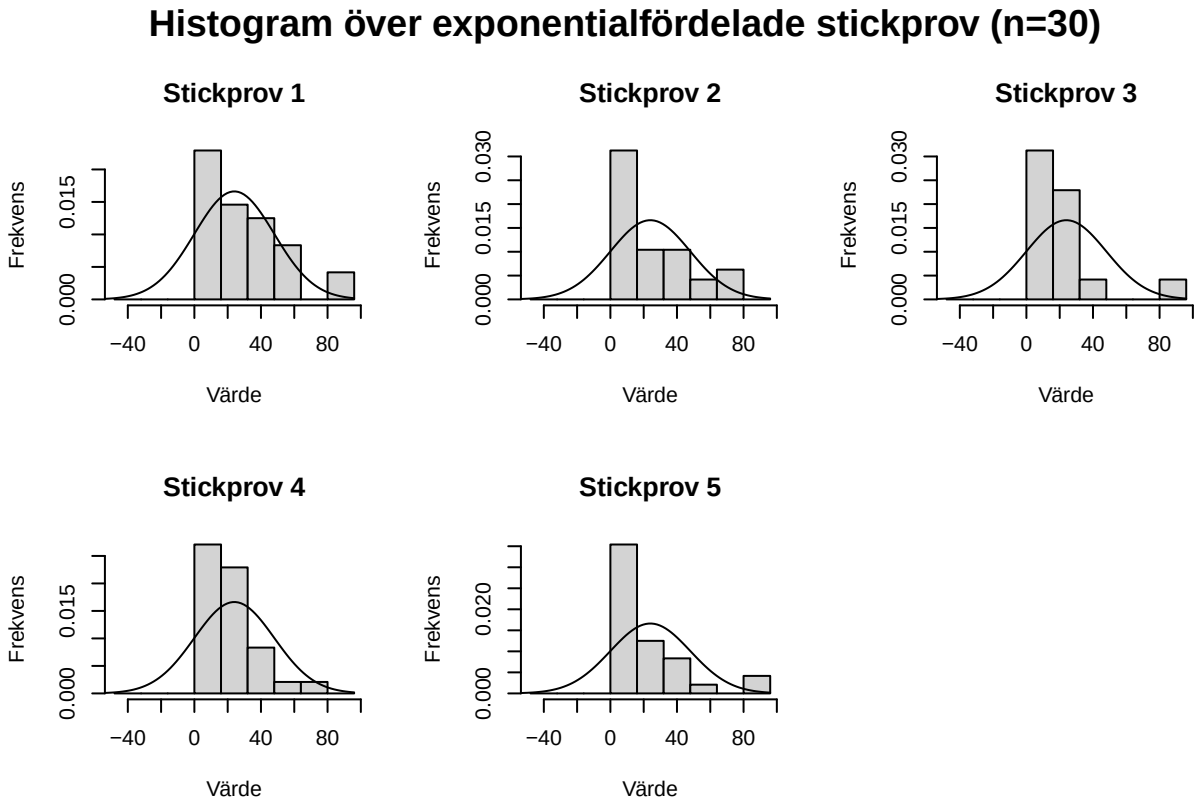

Diagram 7: Histogram över fem exponentialfördelade stickprov.

I figur 7 är det enkelt att se, trots storleken på stickprovet ($n = 30$), att data ej är normalfördelad. Det som övertygar oss om det är dels att den största stapeln aldrig ligger aldrig på värdet 24, dvs det teoretiska väntevärdet, utan konstant *under*, dels faktumet att inga staplar ligger till vänster om värdet 0. Alla stickprov är skeva åt höger och uppvisar *inte* den symmetri som vi vill se hos normalfördelade data.

Exponentialfördelade data: Lådagram

Nu undersöker vi lådagrammen för våra exponentialfördeladestickprov.

```
boxplot(expstickprov1,
        expstickprov2,
        expstickprov3,
        expstickprov4,
        expstickprov5,
        horizontal = TRUE,
        main = "Lådagram över exponentialfördelade stickprov (n=30)",
        xlab = "Värde",
        ylab = "Stickprov")
```

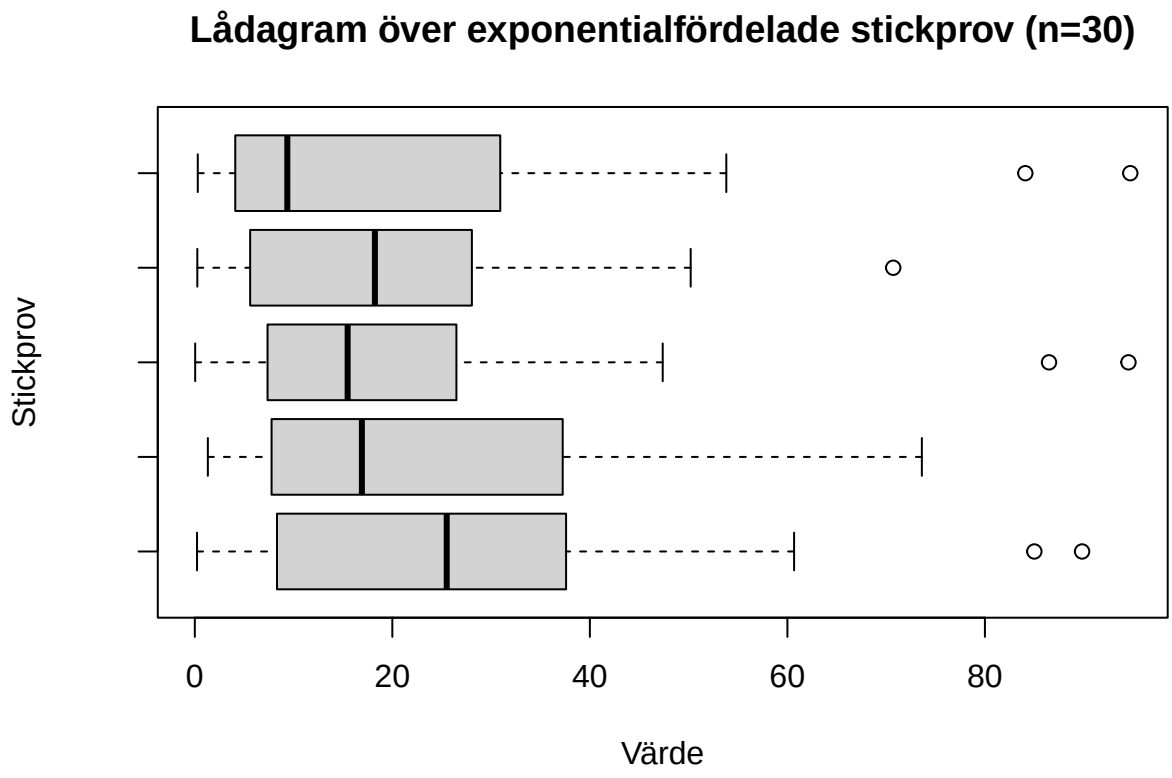

Diagram 8: Lådagram över fem exponentialfördelade stickprov.

Även med lådagrammen, som vi ser i figur 8, är det tydligt att stickproven inte är normalfördelade. Morrhåren är inte lika långa på bägge sidor om lådan, och alla avvikande datapunkter ligger långt ut till höger, medan inga alls ligger till vänster. Medianen är också långt ifrån det teoretiska värdet 24 i många av stickproven.

Exponentialfördelade data: Normalfördelningsplot

Även om vi är rätt så säkra på våra observationer, så undersöker vi även normalfördelningsplottar över våra stickprov.

```
old_par <- par(mfrow = c(2,3), oma = c(0,0,3,0))

qqnorm(expstickprov1,
       xlab = "Teoretiska kvantiler",
       ylab = "Stockprovskvantiler",
       main = "Stickprov 1")
```

```r
qqline(expstickprov1)

qqnorm(expstickprov2,
       xlab = "Teoretiska kvantiler",
       ylab = "Stockprovskvantiler",
       main = "Stickprov 2")
qqline(expstickprov2)

qqnorm(expstickprov3,
       xlab = "Teoretiska kvantiler",
       ylab = "Stockprovskvantiler",
       main = "Stickprov 3")
qqline(expstickprov3)

qqnorm(expstickprov4,
       xlab = "Teoretiska kvantiler",
       ylab = "Stockprovskvantiler",
       main = "Stickprov 4")
qqline(expstickprov4)

qqnorm(expstickprov5,
       xlab = "Teoretiska kvantiler",
       ylab = "Stockprovskvantiler",
       main = "Stickprov 5")
qqline(expstickprov5)

mtext("Normalfördelningsplottar över exponentialfördelade stickprov (n=30)",
      outer = TRUE, cex = 1.2, font = 2)
```

I figur 9 ser vi att punkterna inte sitter på en rät linje, och de avvikande punkterna är mycket mer extrema åt ett håll än det andra. Vi anser att det utifrån detta är tydligt att våra stickprov inte är normalfördelade.

Disukssion och slutsats

Vi tyckte att för relativt låga värden på n för normalfördelade ($n = 40$) och exponentialfördelade ($n = 30$) stickproven, gick det rätt enkelt att dra en slutsats om huruvida stickproven var normalfördelade eller inte. För de likformiga stickproven var det inte lika uppenbart för dessa låga värden på n, utan vi ansåg att vi behövde ett större värde på n för att säkert kunna säga att de likformiga stickproven *inte* var normalfördelade. Detta tror vi har att göra med faktumet att likformig data ändå delar några egenskaper med normalfördelad data. Som att t.ex. data är symmetriskt fördelad i förhållande till medelvärdet. Men ökar man antalet observationer i stickprovet får man en klarare bild, särskilt av normalfördelningsplotten (figur 9), som visar tydligt hur stickprovskvantilerna inte svarar mot de hos en normalfördelning.

För de normalfördelade stickproven tyckte vi att histogrammen gjorde ett bra jobb med att visa hur normalfördelad data var. Även om det fanns undantag i vissa staplar, så var trenden tydlig. Allt var tydligt symmetrisk, och svansarna tedde sig så som täthetsfunktionens svansar.

För de exponentialfördelade stickproven kändes det lite godtyckligt, men lådagrammen (figur 5) var nog enklast att avläsa för att avgöra huruvida stickproven inte var normalfördelade. Hur mycket de skevade åt vänster, hur olika långa morrhåren var, samt hur avvikande värdena av stickproven alla fann sig på höger sida.

Skulle man behöva välja en metod utav dessa tre - i praktiken kan vi ju hamna i en situation där vi inte vet vilken fördelning vårt stickprov kommer ifrån! - tycker vi nog att normalfördelningsplotten är enklast att avläsa, och ger en tydlig bild på om ett givet stickprov är normalfördelat eller inte. Vi menar att ögat är bättre

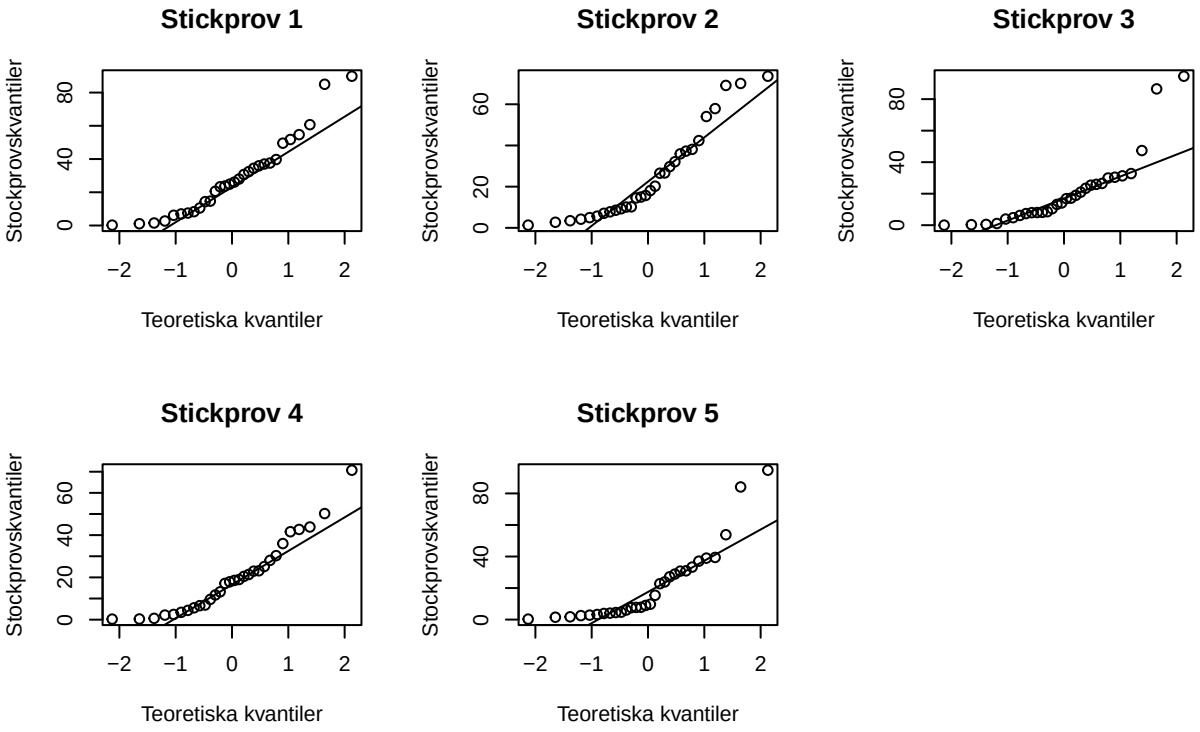

Diagram 9: Normalfördelningsplot över fem exponentialfördelade stickprov.

på att avgöra om punkter ligger på en rät linje, än om staplar följer den klockformade täthetsfunktionen, liksom att avgöra avstånd mellan kvartiler (i lådagrammet), och därför anser vi att normalfördelningsplotten är det bästa valet.

Uppgift 2: Explorativ dataanalys

I den här delen ska vi undersöka ett verkligt datamaterial med hjälp av de metoder som vi bekantat oss med ovan. Våra data är statistik över alkoholkonsumtionen i ett urval av OECD-länder, uppdelat på öl, vin och sprit.

Vi börjar med att läsa in data i en dataframe, och extraherar vektorer utifrån kolonnerna, så att vi kan plotta variablerna var för sig och mot varandra.

```
alkohol_OECD <- read.csv("olvinsprit.csv", header = TRUE) # läser in data

land <- alkohol_OECD$Land # extraherar och läser in kolonner som vektorer
öl <- alkohol_OECD$beer
vin <- alkohol_OECD$vin
sprit <- alkohol_OECD$sprit

alkohol_OECD # visar data
```

```
##              Land beer vin sprit
## 1         Sverige   56  16   2.9
## 2         Danmark   98  32   2.7
## 3         Finland   79  10   5.7
## 4           Norge   56  11   2.4
## 5         Belgien   98  30   2.6
## 6       Frankrike   41  47   7.2
## 7          Irland  155  13   5.3
## 8         Italien   29  54   2.7
## 9         Holland   80  20   4.7
## 10        Schweiz   57  42   2.4
## 11 Storbritannien   97  20   3.9
## 12       Tyskland  119  26   5.3
## 13      Osterrike  106  36   3.2
## 14            USA   85   7   4.8
## 15         Kanada   70  10   4.3
## 16     Australien   89  21   2.6
## 17    Nya_Zeeland   78  19   2.3
## 18          Japan   55  10   8.2
```

Vi noterar att vi för varje variabel har 18 datapunkter. Nu är vi redo att undersöka våra data!

Är data normalfördelad?

Vår första fundering är huruvida data kan anses komma från en normalfördelning. Vi använder de metoder vi nyss har lärt oss och ser efter.

Öl

```
old_par <- par(mfrow = c(1, 3), oma=c(0,0,3,0))

qqnorm(öl,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Normalfördelningsplot")
qqline(öl)

boxplot(öl,
```

```
        ylab = "Konsumtion",
        main = "Boxplot")

hist(öl,
     xlab = "Konsumtion",
     ylab = "Frekvens",
     main = "Histogram")

mtext("Konsumtion av öl i ett antal OECD-länder", outer = TRUE, cex = 1.2, font = 2)
```

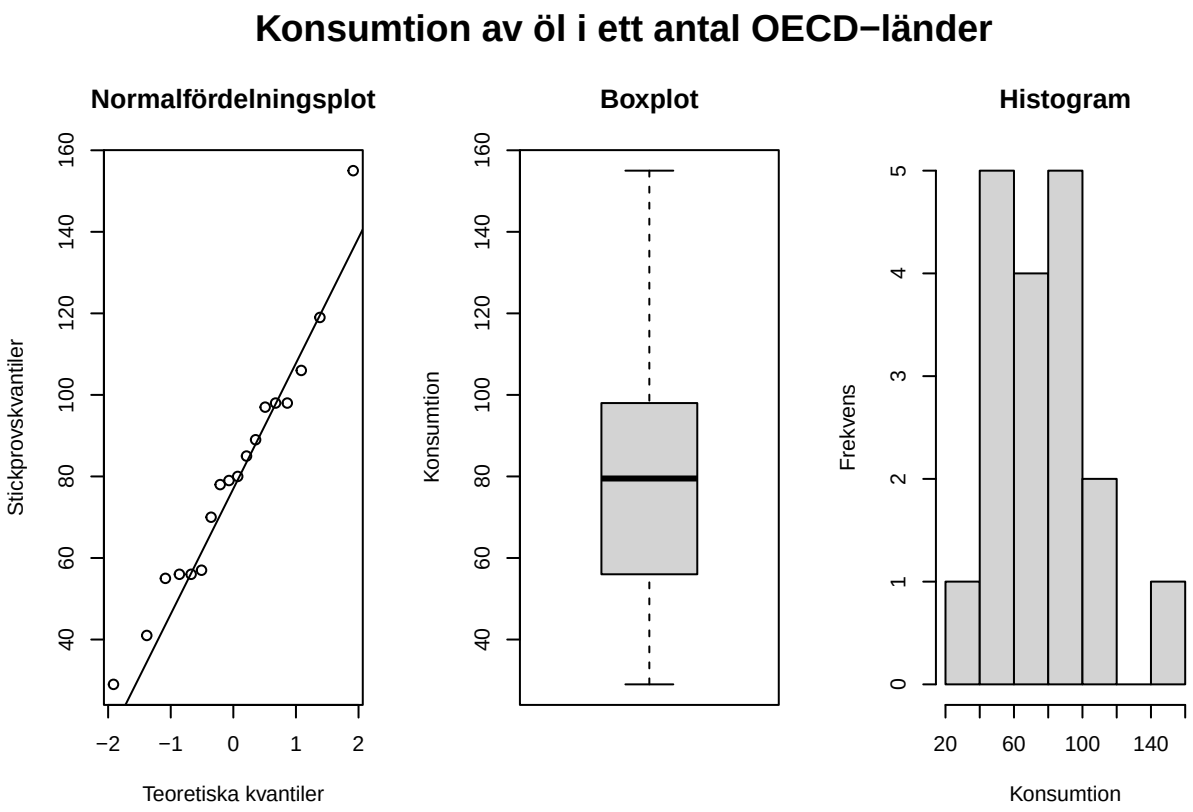

Diagram 10: Ölkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

I figur 10 ser vi tre olika plottar över konsumtionen av öl. Alla tre plottar indikerar att data kan komma från en normalfördelning. I normalfördelningsplotten ser vi att punkterna ligger längs en hyfsat rät linje. I boxplotten ser vi symmetri i lådan (medianen och kvartilerna), och längden på "svansarna" motsvarar ungefär vad vill se i en boxplot om data vore normalfördelad (även om dessa inte är helt symmetriska). I histogrammet ser vi något som påminner om normalfördelningens täthet: väntevärdet ligger någonstans kring 70-80, de allra flesta datapunkterna hamnar mellan 40 och 100, och en mindre mängd datapunkter sprider sig symmetriskt på bägge sidor om väntevärdet. Detta är vad vi väntar oss av normalfördelad data. Vi menar att öl kan anses komma från en normalfördelning.

Vin

```
old_par <- par(mfrow = c(1, 3), oma=c(0,0,3,0))

qqnorm(vin,
       xlab = "Teoretiska kvantiler",
```

```
       ylab = "Stickprovskvantiler",
       main = "Normalfördelningsplot")
qqline(vin)

boxplot(vin,
        ylab = "Konsumtion",
        main = "Boxplot")

hist(vin,
     breaks = seq(from = 0, to = 70, by = 10),
     xlab = "Konsumtion",
     ylab = "Frekvens",
     main = "Histogram")

mtext("Konsumtion av vin i ett antal OECD-länder", outer = TRUE, cex = 1.2, font = 2)
```

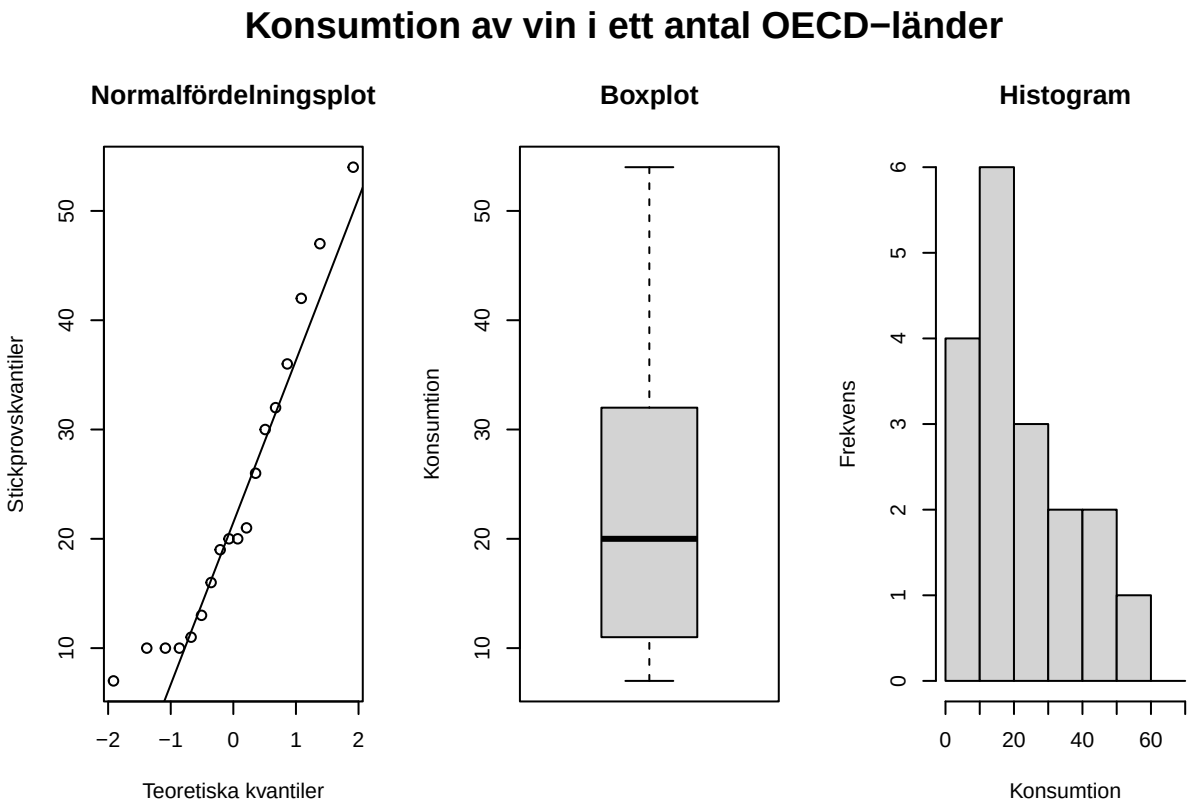

Diagram 11: Vinkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

Konsumtionen av vin fördelat mellan de olika länderna ser vi i figur 11. Jämfört med öl ser vi här större skevhet. I normalfördelningsplotten är det fler outliers utanför den önskade räta linjen och i boxplotten ser vi en avsaknad av symmetri - tyngdpunkten ligger nedåt och vi har längre "svans" uppåt. Detsamma ser vi i histogrammet. Om konsumtionen av vin vore normalfördelad hade vi velat se ungefär lika mycket data till vänster om tyngdpunkten som till höger. Det gör vi inte helt nu. Alltså anser vi *inte* att man kan säga att vin kommer från en normalfördelning, utifrån det vi ser i våra data.

Sprit

```
old_par <- par(mfrow = c(1, 3), oma=c(0,0,3,0))

qqnorm(sprit,
       xlab = "Teoretiska kvantiler",
       ylab = "Stickprovskvantiler",
       main = "Normalfördelningsplot")
qqline(sprit)

boxplot(sprit,
        ylab = "Konsumtion",
        main = "Boxplot")

hist(sprit,
     xlab = "Konsumtion",
     ylab = "Frekvens",
     main = "Histogram")
mtext("Konsumtion av sprit i ett antal OECD-länder", outer = TRUE, cex = 1.2, font = 2)
```

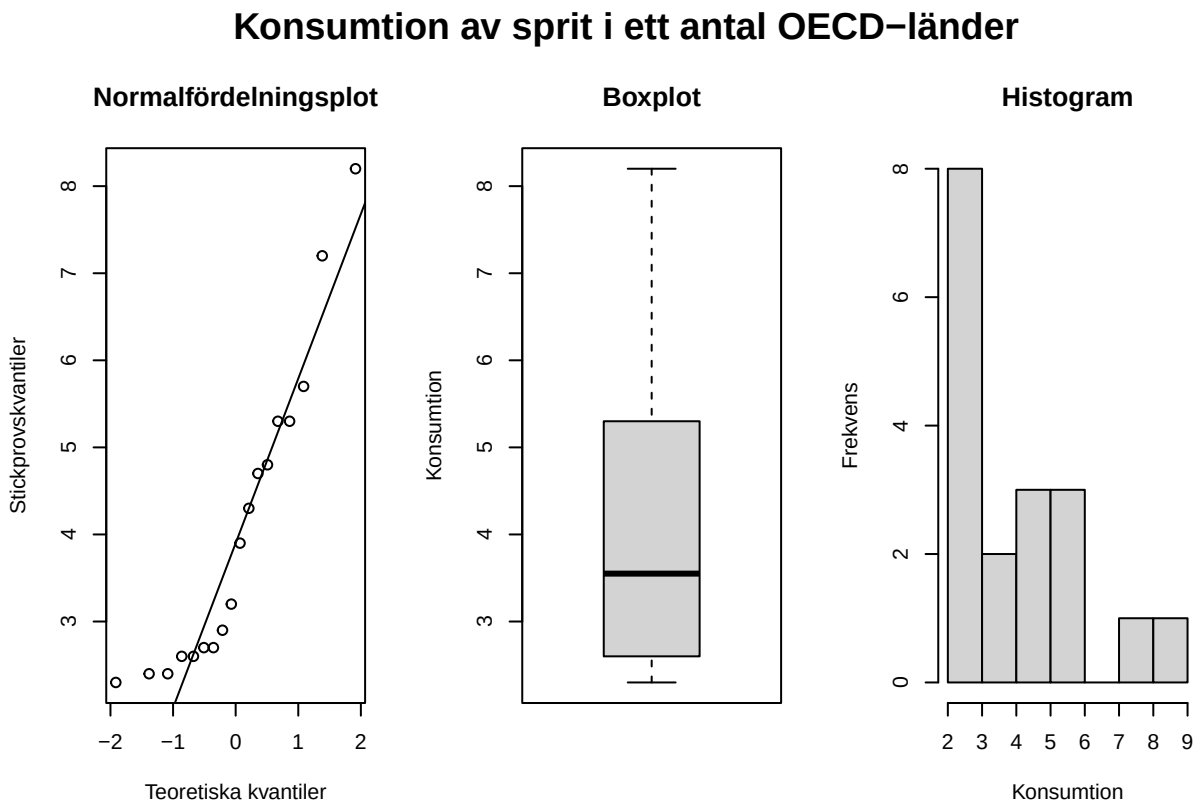

Diagram 12: Spritkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

Slutligen betraktar vi spritkonsumtionen i figur 12, genom de tre olika diagramtyperna. Här ser vi en något ytterligare större skevhet än för vinkonsumtionen. Särskilt tydligt är detta i histogrammet, där vi ser att typvärdet ligger i den första klassen. I boxplotten är "svansarna" ojämna och i normalfördelningsplotten har vi flera datapunkter utanför den räta linjen. Vi anser, utifrån våra data, att spritkonsumtionen *inte* kommer från en normalfördelning.

Jämförande av variabler

Men vilka länder konsumerar vad? Och hur ligger Sverige till i jämförelse med andra länder? Finns det något samband mellan konsumtion av en viss spritsort och en annan? Låt oss plotta variablerna tillsammans, så ska vi försöka besvara frågor som dessa.

```
gg_alkohol <- ggplot(alkohol_OECD, aes(x=öl, y=vin, label=Land)) + # väljer data
  geom_point(col = "#00AFBB", size = sprit) + # lägger till punkter
  geom_text(hjust=-0.1, vjust=-0.3) + #  justerar labels position
  labs(title="Alkoholkonsumtion i ett urval av OECD-länder",
       y="Konsumtion av vin",
       x="Konstumtion av öl")

plot(gg_alkohol)
```

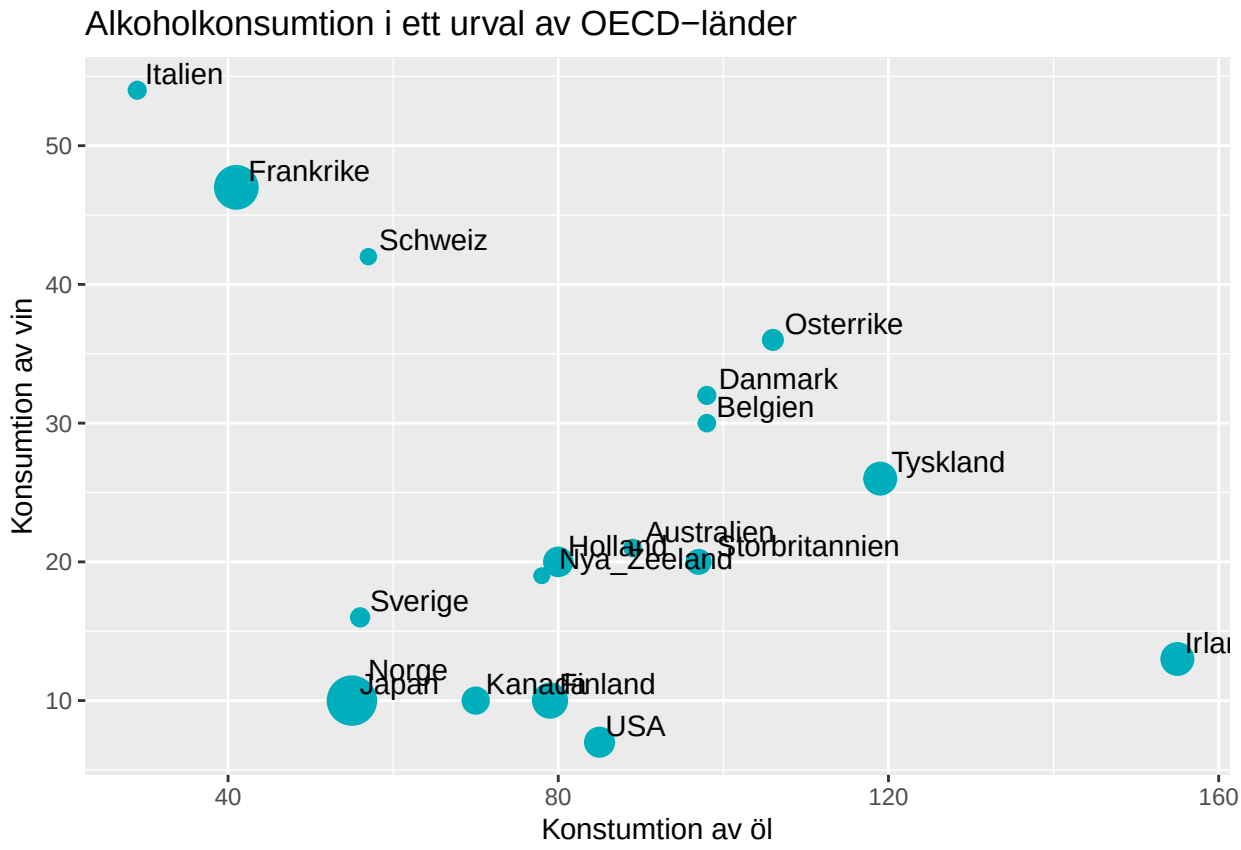

Diagram 13: Alkoholkonsumtionen i 18 OECD-länder, öl på x-axeln och vin på y-axeln. Storleken på datapunkterna indikerar konsumtion av starksprit: Ju större punkt, desto högre konsumtion av starksprit.

I figur 13 ser vi konsumtionen av öl plottat mot konsumtionen av vin. Även konsumtionen av starksprit syns i plotten, indikerat genom storleken på datapunkterna.

Vi kan konstatera att Italien är det land som konsumerar mest vin, följt av Frankrike och därpå Schweiz. Italien är samtidigt det land med minst konsumtion av öl, och Frankrike det land med näst minst konsumtion av öl. I vin vs öl-avseendet kan vi konstatera att Italien och Frankrike - och även Schweiz - sticker ut jämfört med flera av de andra länderna. Sverige konsumerar förhållandevis lite vin: det finns fler länder som konsumerar *mer* vin än Sverige, än vad det finns länder som konsumerar *mindre*.

Tittar vi på ölkonsumtionen är Irland det land som ligger klart i topp, med den högsta konsumtionen. Tvåan på listan över ölkonsumtion, Tyskland, kan anses ligga i överkant av ett bredare kluster av ölkonsumenter,

snarare. Tätt efter Tyskland följer Österrike, Danmark och Belgien. Sverige konsumerar förhållandevis lite öl: det finns fler länder som konsumerar *mer* öl än Sverige, än vad det finns länder som konsumerar *mindre*.

Japan är det land som konsumerar mest sprit, tätt följt av Frankrike. Det finns inget land som på egen hand sticker ut åt andra hållet, alltså med en jämförelsevis väldigt låg konsumtion av starksprit (det här följer av den typ av skevhet som dessa data har, vilket vi såg i avsnittet ovan där vi undersökte fördelningen för respektive dryckestyp). Sverige kan inte anses vara extremt i starkspritkonsumtionsavseende.

Låt oss nu betrakta hur sprit- och ölkonsumtionen förhåller sig till varandra.

```
sprit_öl <- ggplot(alkohol_OECD, aes(x=sprit, y=öl)) + # väljer data
  geom_point(col = "#00AFBB", size = 3) + # lägger till punkter
  geom_text(aes(label = land), hjust=-0.1, vjust=-0.3) + # justerar labels position
  geom_smooth(method="lm", col="red") + # lägger till trendlinje, inkl. konfidensintervall
  labs(title="Konsumtion av sprit och öl i ett urval av OECD-länder",
       x="Sprit",
       y="Öl")

plot(sprit_öl)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

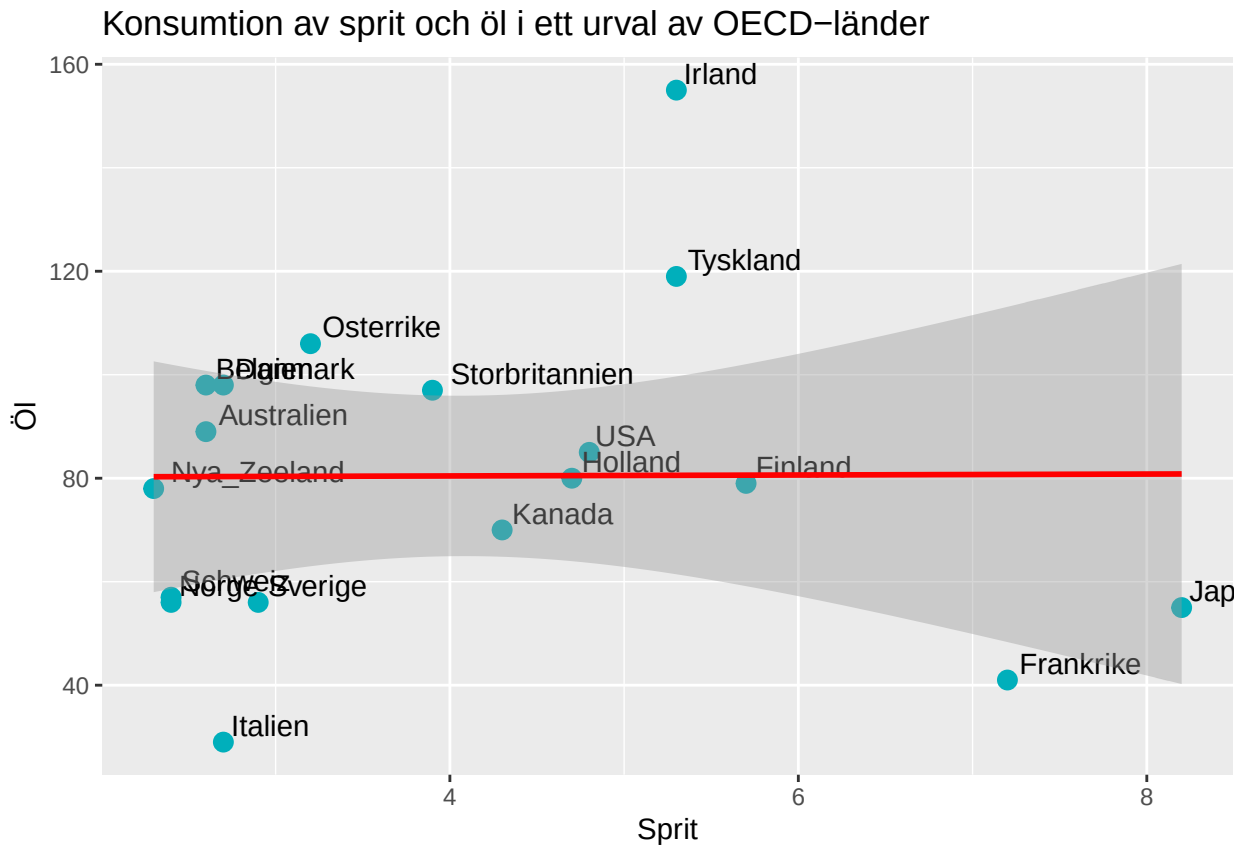

Diagram 14: Sambandet mellan konsumtionen av sprit och vin, i 19 OECD-länder. Sprit på x-axeln och öl på y-axeln. Det skuggade området indikerar konfidensintervall.

I figur 14 har vi plottat en trendlinje (röd) för hur konsumtion av sprit förhåller sig till konsumtion av öl. Det är tydligt att det inte finns något samband mellan dessa. Det breda konfidensintervallet (skuggat område) ser vi att *om* det finns ett samband, skulle det kunna vara både positivt eller negativt. Utifrån våra data kan vi inte avgöra vilket. Vi gör samma sak för öl och vin, respektive vin och sprit:

```
öl_vin <- ggplot(alkohol_OECD, aes(x=öl, y=vin)) + # väljer data
  geom_point(col = "#00AFBB", size = 3) + # lägger till punkter
  geom_text(aes(label = land), hjust=-0.1, vjust=-0.3) + # justerar labels position
  geom_smooth(method="lm", col="red") + # lägger till trendlinje, inkl. konfidensintervall
  labs(title="Konsumtion av öl och vin i ett urval av OECD-länder",
       x="Öl",
       y="Vin")

plot(öl_vin)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

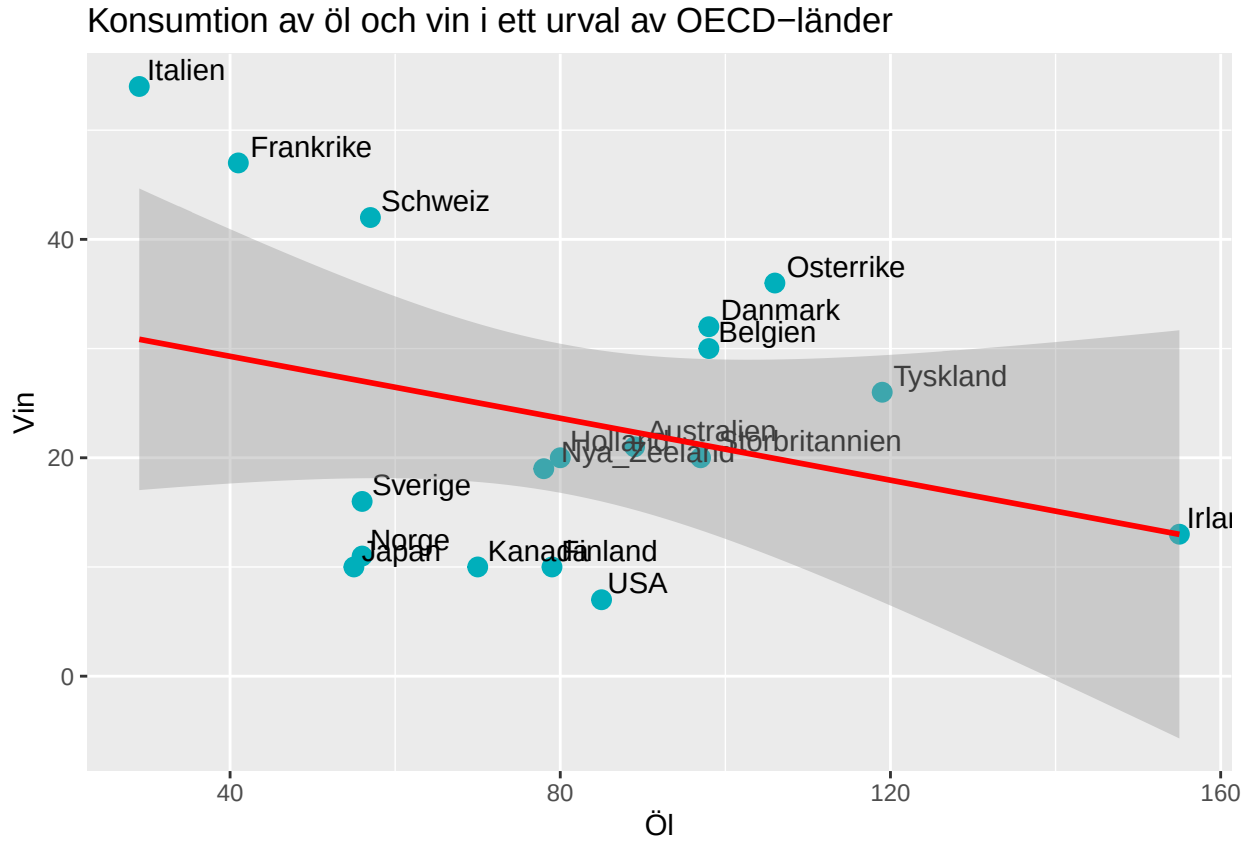

Diagram 15: Sambandet mellan konsumtionen av öl och vin, i 19 OECD-länder. Öl på x-axeln och öl på y-axeln. Det skuggade området indikerar konfidensintervall.

```
sprit_vin <- ggplot(alkohol_OECD, aes(x=sprit, y=vin)) + # väljer data
  geom_point(col = "#00AFBB", size = 3) + # lägger till punkter
  geom_text(aes(label = land), hjust=-0.1, vjust=-0.3) + # justerar labels position
  geom_smooth(method="lm", col="red") + # lägger till trendlinje, inkl. konfidensintervall
  labs(title="Konsumtion av sprit och vin i ett urval av OECD-länder",
       x="Sprit",
       y="Vin")

plot(sprit_vin)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

I figur 15 ser vi att det *skulle* kunna finnas ett *svagt negativt* samband mellan konsumtion av vin och öl

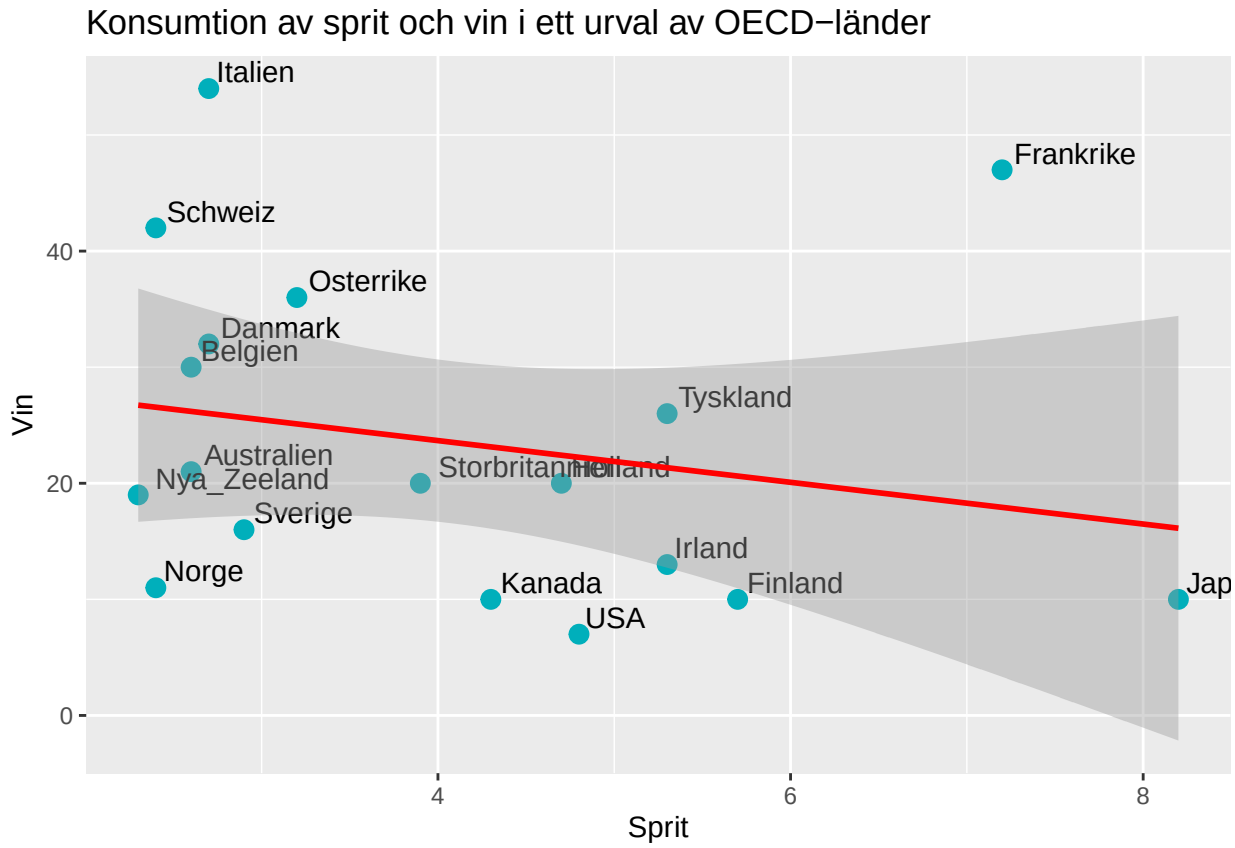

Diagram 16: Sambandet mellan konsumtionen av öl och vin, i 19 OECD-länder. Öl på x-axeln och öl på y-axeln. Det skuggade området indikerar konfidensintervall.

(alltså att ju högre ölkonsumtion, desto mindre vinkonsumtion). Men som vi ser på konfidensintervallen skulle också sambandet kunna vara det motsatta. Det som orsakar den svaga lutningen på den röda grafen är troligtvis våra extremer Italien, Frankrike och Irland. Italien och Frankrike konsumerar mycket vin och lite öl. Irland tvärtom. Sambandet mellan konsumtion av sprit och vin, som vi ser i figur 16, är med samma resonemang lika obefintligt.

Diskussion

Vi har undersökt alkoholkonsumtionen i ett urval av OECD-länder. Vår explorativa dataanalys har visat att konsumtionen av öl skulle kunna anses komma från en normalfördelning. Mer tveksamt är det med konsumtionen av vin, och än mer tveksamt är det för konsumtionen av sprit.

I topp bland vinkonsumenterna ser vi länder med stor vinexport, som Italien och Frankrike. Detsamma gäller konsumtionen av öl som domineras av Irland - ett land som är vida känt för sitt öl. Vi kan således konstatera att ett gemensamt drag hos de länder som sticker ut i konsumtion av en viss alkoholtyp är att det landet har stor produktion och export av just den varan. Frankrike och Italien: vin, Irland: öl. Dessa resultat var väntade. Möjligen hade vi väntat oss att Tyskland och Österrike skulle vara mer extrema i ölkonsumtion än vad våra data visar.

Vad gäller Sverige, visar vår analys att Sverige inte ligger i topp för någon alkoholtyp. Tvärtom befinner sig Sverige i den under halvan av länderna i vårt urval.

Genom vår analys kan vi vidare konstatera att det inte verkar finnas något samband mellan konsumtion av den ena vs den andra typen av alkohol. Med data från fler länder är det möjligt att vi skulle kunna skönja ett sådant samband för exempelvis vin och öl. Om vi utgår från vad vi kommer fram till i uppgift 1 är ett stickprov på n=18 faktiskt ett för litet stickprov för att kunna besvara frågan om normalfördelning.