Laboration 1 - MT4001

Daniel Svedlund, Sebastijan Babic

2024-11-18

Sammanfattning

I denna laboration använder vi oss av grafisk analys i form av histogram, lådagram, normalfördelningsplottar och stolpdiagram, vi har några stickprov som är normal,likformigt och exponential-fördelade och letar då efter det minsta antalet observationer som krävs för att visa att stickproven är normalfördelade eller inte. Detta talet n visar sig ofta vara så stor som 100 för att vi ska säkert kunna avgöra att fördelningen inte är normalfördelningen. Vilken plottyp för analys man använder sig av beror på stickproven och effektiviteten av dem plottyper hos dem olika analyser skiljer sig åt.

Det undersöktes även data om den genomsnittliga konsumtionen av alkohol som mätts i 18 st. OECD-länder. Vi undersöker om konsumtionen av någon alkoholtyp följer en normalfördelning eller inte. Vi ser att ölkonsumtion verkar följa en normalfördelning medan vin och spritkonsumtionen inte gör det.

Det undersöktes även hur Sverige ligger till i öl, vin- och spritkonsumtion bland dem 18 OECD-länder och ser att vi ligger under det genomsnittliga konsumtionen hos alla alkoholtyper. Irland toppar ölkonsumtionen, Italien vinkonsumtion och Japan spritkonsumtion.

Uppgift 2 - Kommer data från en normalfördelning?

För att besvara om en okänd fördelning kan vara en normalfördelning använder vi oss av ofta approximationer som sedan undersöks grafiskt (åtminstone här). Vi ska också svara hur stort n behöver vara, dvs antalet observationer för att få en rimlig approximation.

Följande fördelningar har allihopa väntevärde och standardavvikelse a där a är dem två sista siffrorna i vårat personnummer, 13.

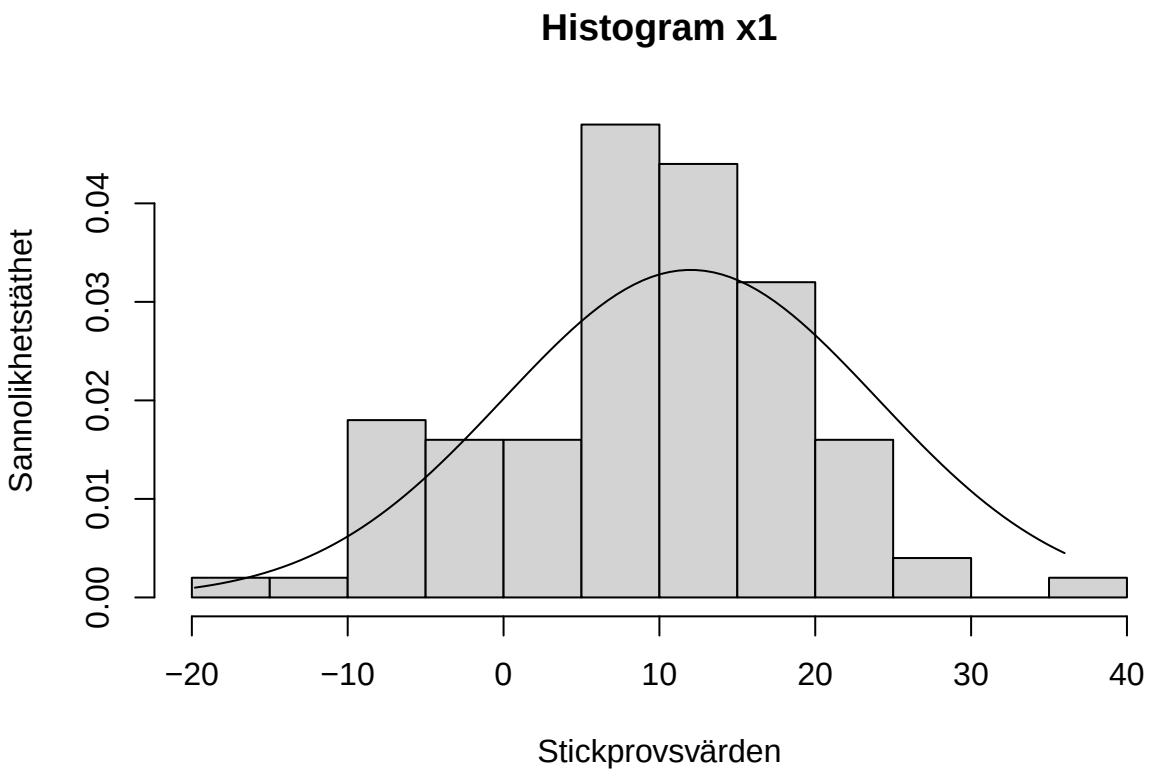
Uppgift 2.1 Normalfördelade data

```
set.seed(20040911)

n <- 100

x1 <- rnorm(n, 12, 12)

hist(x1,
     main = "Histogram x1",
     xlab = "Stickprovsvärden",
     ylab = "Sannolikhetstäthet",
     prob = TRUE)
x <- seq(from = min(x1), to = max(x1), length.out = 100) # tag minsta x:-10
# och största x:25 för linjens intervall
# kan göra manuellt och sätta från -10 till 25 men detta gör
#det enklare om man vill på något sätt förändra
# stickproven
lines(x, dnorm(x, 12, 12))
```

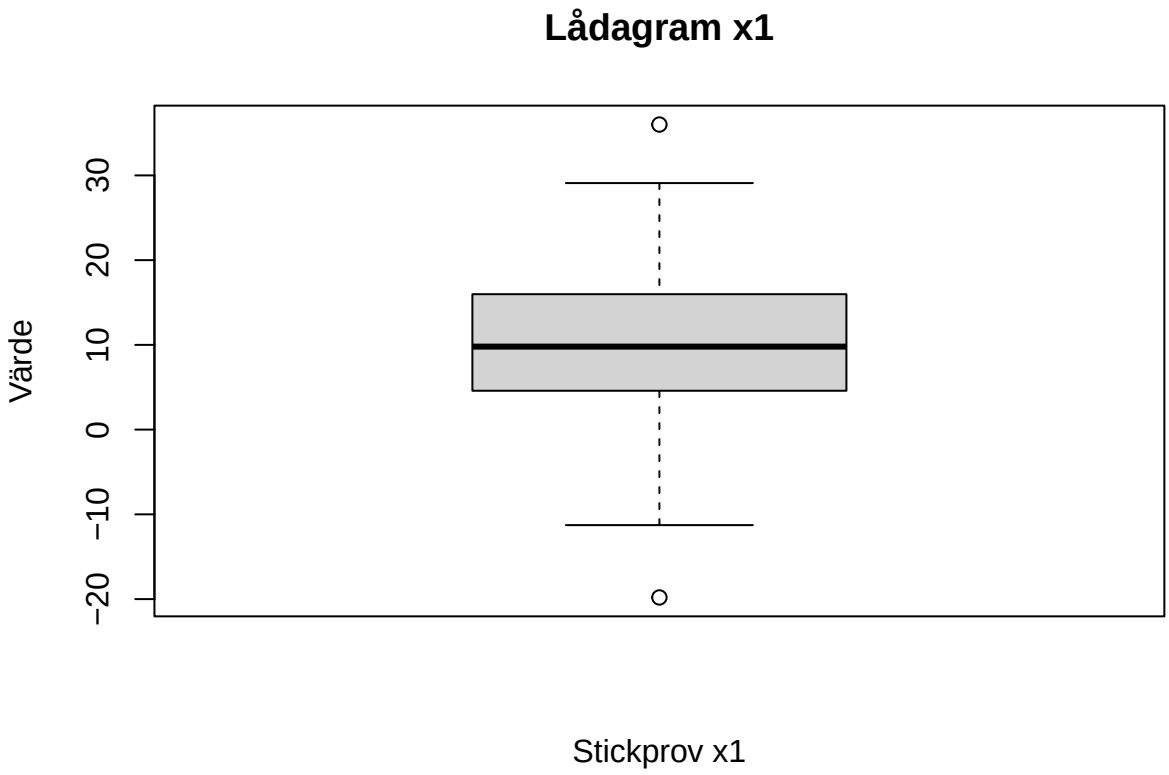
```
boxplot(x1,
        main = "Lådagram x1",
        xlab = "Stickprov x1",
        ylab = "Värde")
```

```
qqnorm(x1,
       main = "Normalfördelningsplot x1",
       xlab = "Teoretiska kvantiler",
       ylab = "Observerade kvantiler")
qqline(x1)
```

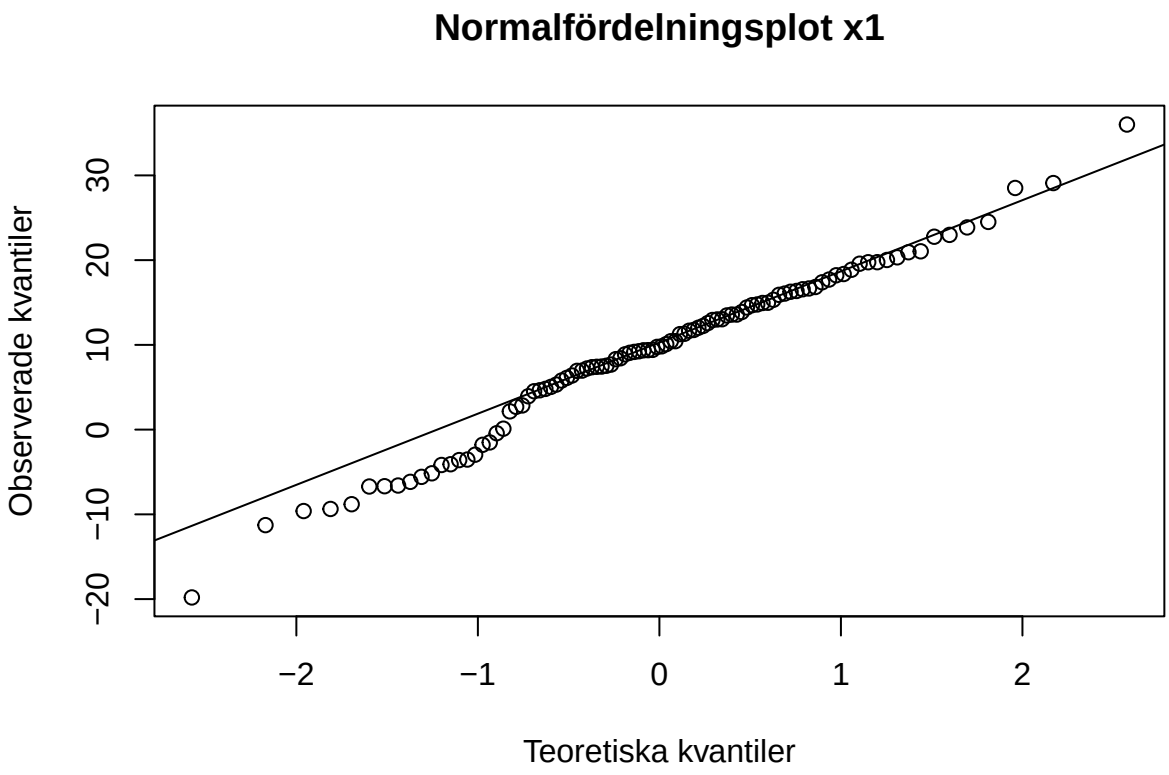
```
# definierar 8 stickprov som x1,x2,...,x8 med en normalfördelning

set.seed(20040911)
x1 <- rnorm(n, 12, 12) # kommer att bli samma stickprov som ovan då vi har samma seed
x2 <- rnorm(n, 12, 12) # detta blir ett stickprov annorlunda från x1, likaså de nedan
x3 <- rnorm(n, 12, 12)
x4 <- rnorm(n, 12, 12)
x5 <- rnorm(n, 12, 12)
x6 <- rnorm(n, 12, 12)
x7 <- rnorm(n, 12, 12)
x8 <- rnorm(n, 12, 12)
```



Figur 1: Histogram över en stickprov x1 av storlek n = 100.

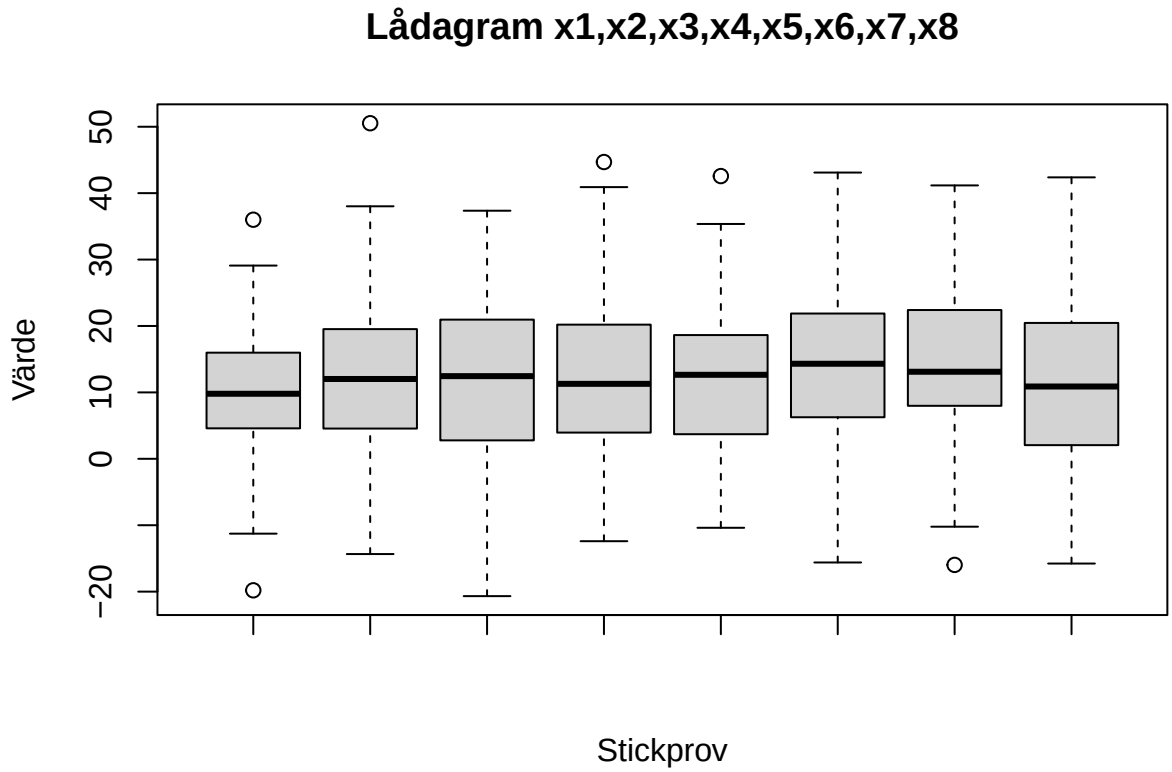


Figur 2: Lådagram över en stickprov x1 av storlek n = 100.



Figur 3: Normalfördelningsplot över en stickprov x1 av storlek n = 100.

```r
boxplot(x1, x2, x3, x4, x5, x6, x7, x8, xlab = "Stickprov", ylab = "Värde", main = "Lådagram x1,x2,x3,x
```



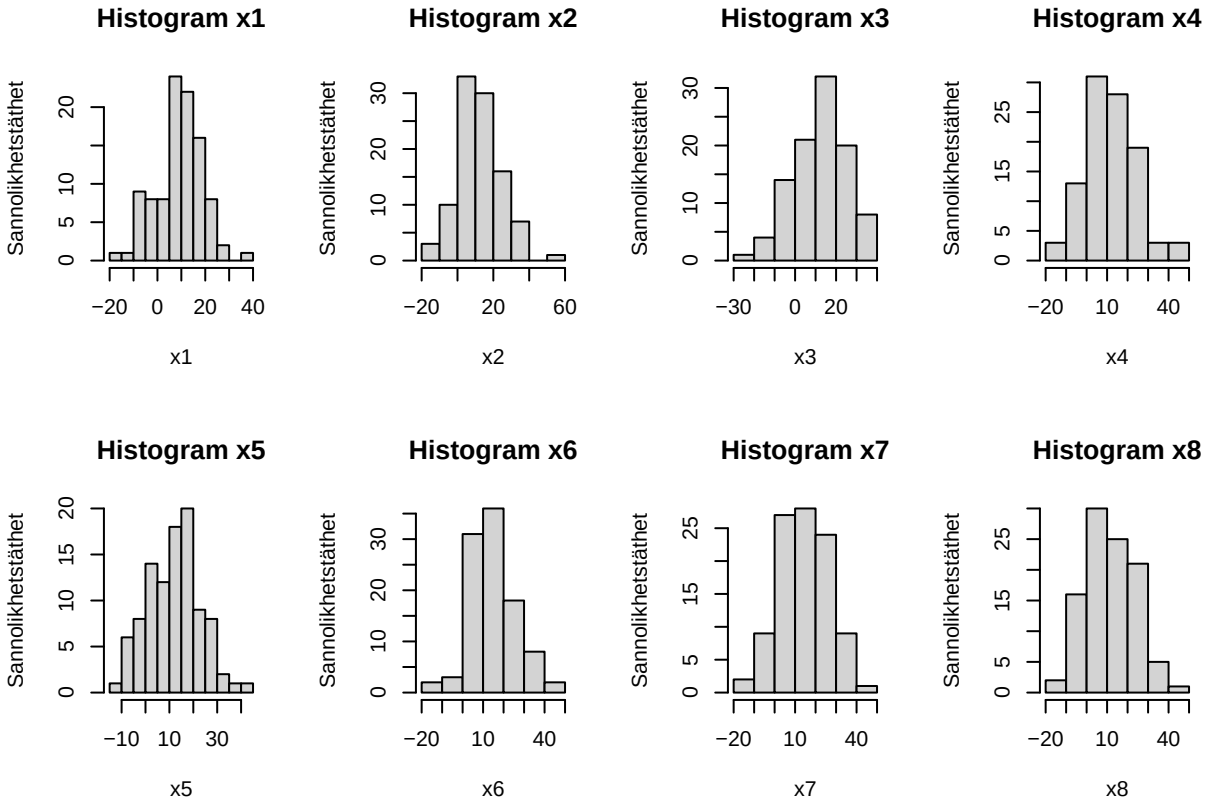
Figur 4: 8 lådagram i en samma plot som visar 8 olika stickprov av storlek 10.

```r
old_par <- par(mfrow = c(2, 4)) # 2 rader, 4 kolonner
hist(x1, main = "Histogram x1", ylab = "Sannolikhetstäthet")
hist(x2, main = "Histogram x2", ylab = "Sannolikhetstäthet")
hist(x3, main = "Histogram x3", ylab = "Sannolikhetstäthet")
hist(x4, main = "Histogram x4", ylab = "Sannolikhetstäthet")
hist(x5, main = "Histogram x5", ylab = "Sannolikhetstäthet")
hist(x6, main = "Histogram x6", ylab = "Sannolikhetstäthet")
hist(x7, main = "Histogram x7", ylab = "Sannolikhetstäthet")
hist(x8, main = "Histogram x8", ylab = "Sannolikhetstäthet")
```

```r
par(old_par)
```

Då n = 100 kan man hyfsat enkelt avgöra fördelningen på x1,x2,...,x8 följer en normalfördelning eller inte på lådagrammen (figur 2) och histogrammen (figur 1). Däremot är det fortfarande svårt att avgöra om det är en normalfördelning eller inte på normalfördelningsplotten (figur 3).

Lådagrammen (figur 2) och histogrammen (figur 1) är ungefär lika effektiva för att avgöra detta, med det menas att vi behöver ungefär lika stor n på både för att med säkerhet kunna avgöra fördelningen men däremot kan lådagramen vara lite vilseledande som vi kommer se i uppgift 2.2 så histogrammen (figur 1) verkar vara effektivast här då det krävs minst värde på n för att avgöra om normalfördelat eller inte jämfört med dem andra plottyper.



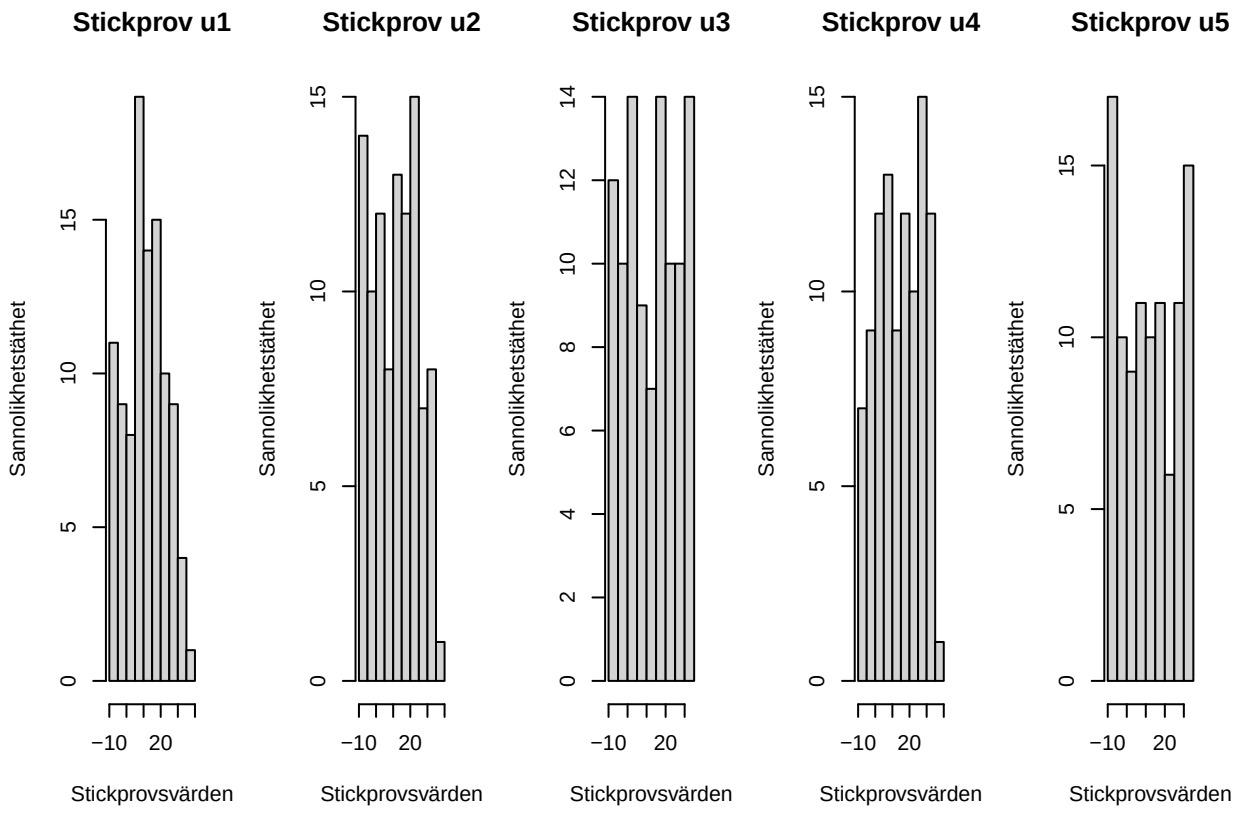
Figur 5: 8 histogram i en och samma plot som visar 8 olika stickprov av storlek 10 från innan.

Uppgift 2.2 - Likformigt fördelade data

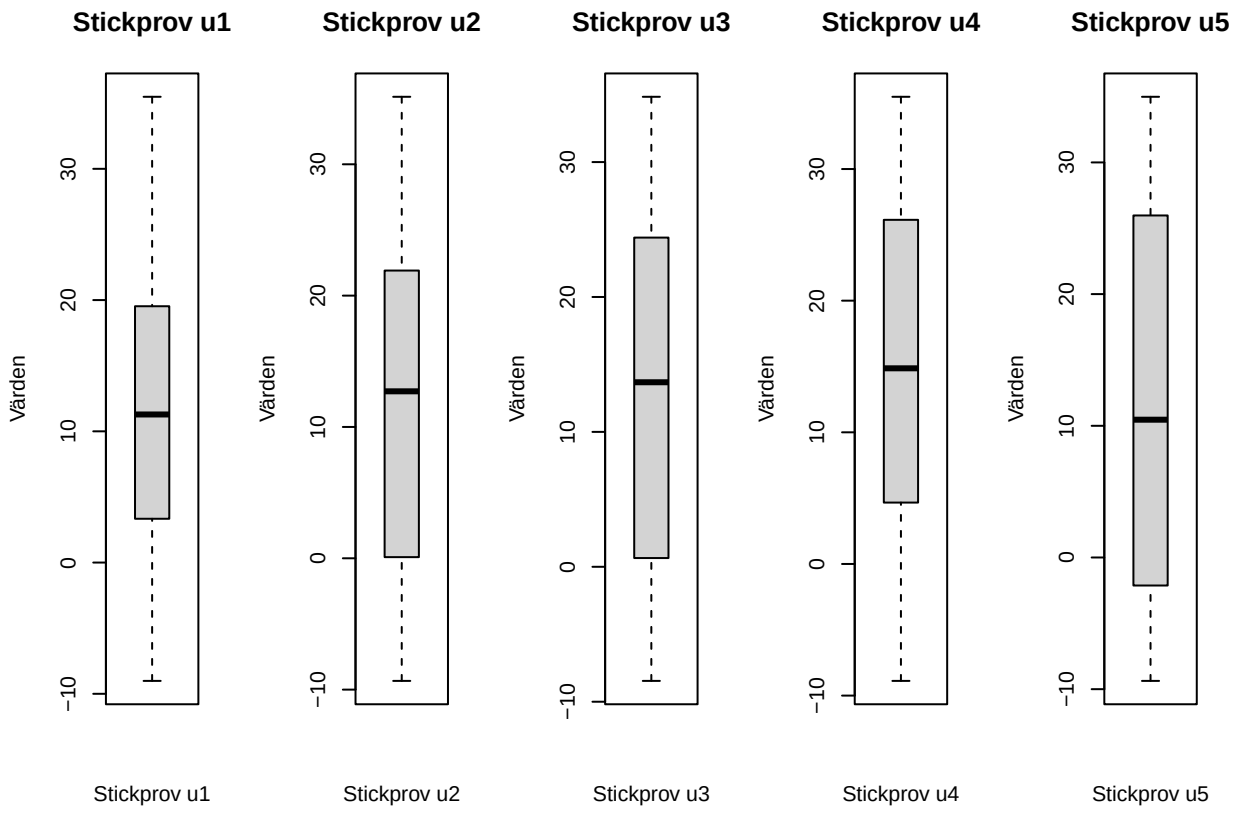
```r
set.seed(20040911)

# välj a från uppgift 1
a <- 13
n <- 100

# definiera gränser
alpha <- a * (1 - sqrt(3))
beta <- a * (1 + sqrt(3))

# generera slumpmässiga data
u1 <- runif(n, min = alpha, max = beta)
u2 <- runif(n, min = alpha, max = beta)
u3 <- runif(n, min = alpha, max = beta)
u4 <- runif(n, min = alpha, max = beta)
u5 <- runif(n, min = alpha, max = beta)

# plotta histogram
old_par <- par(mfrow = c(1, 5))  # skapa layout för flera plots och repetera för alla 3 typer av plotta
hist(u1, main = "Stickprov u1", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(u2, main = "Stickprov u2", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(u3, main = "Stickprov u3", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(u4, main = "Stickprov u4", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(u5, main = "Stickprov u5", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
```

```r
par(old_par)

# plotta lådagram
old_par <- par(mfrow = c(1, 5))
boxplot(u1, main = "Stickprov u1", xlab = "Stickprov u1", ylab = "Värden")
boxplot(u2, main = "Stickprov u2", xlab = "Stickprov u2", ylab = "Värden")
boxplot(u3, main = "Stickprov u3", xlab = "Stickprov u3", ylab = "Värden")
boxplot(u4, main = "Stickprov u4", xlab = "Stickprov u4", ylab = "Värden")
boxplot(u5, main = "Stickprov u5", xlab = "Stickprov u5", ylab = "Värden")
```

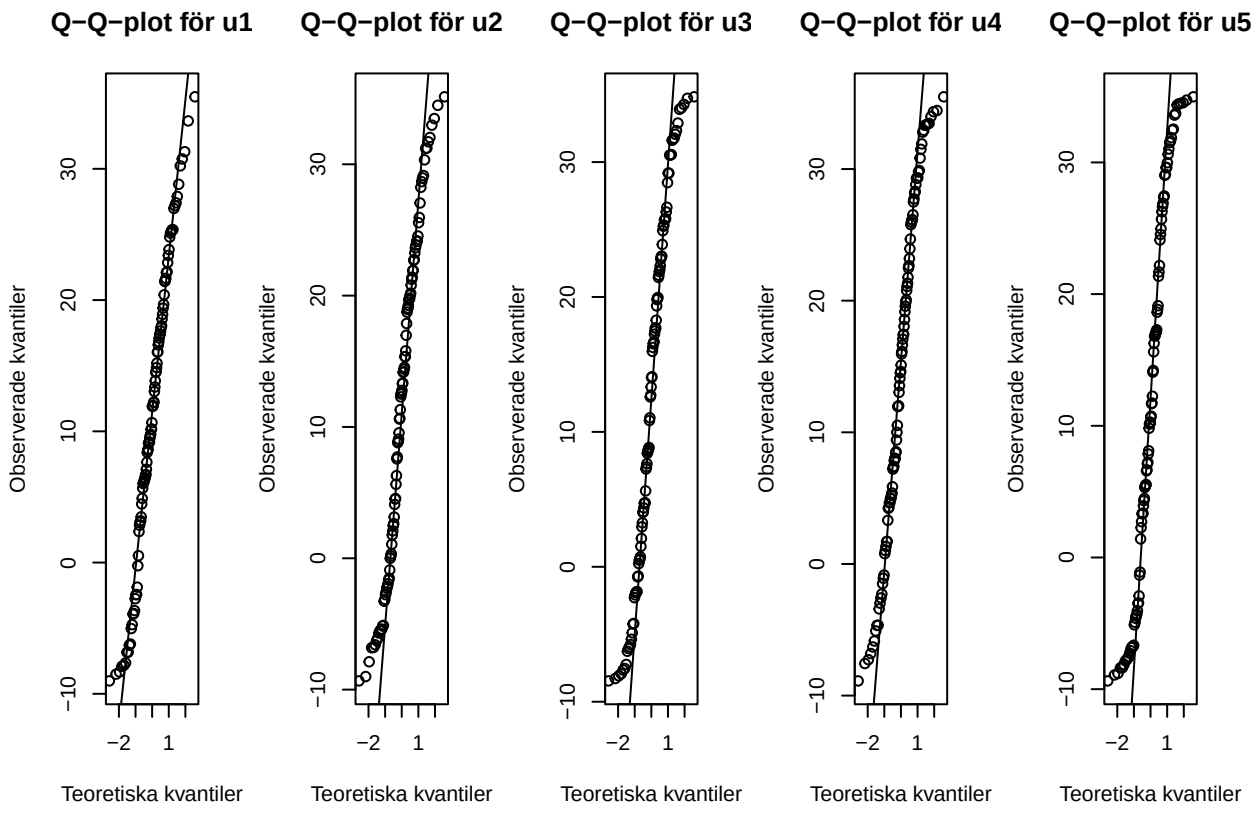
```r
par(old_par)

# plotta normalfördelningsplot
old_par <- par(mfrow = c(1, 5))
qqnorm(u1, main = "Q-Q-plot för u1", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u1)
qqnorm(u2, main = "Q-Q-plot för u2", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u2)
qqnorm(u3, main = "Q-Q-plot för u3", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u3)
qqnorm(u4, main = "Q-Q-plot för u4", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u4)
qqnorm(u5, main = "Q-Q-plot för u5", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u5)
```



Figur 6: Histogram, lådagram och normalfördelningsplottar för 5 oberoende stickprov u1,u2,u3,u4,u5 av storlek (antalet observationer) 10 med gränserna 13 +/- 13sqrt(3) enligt uppgift 1.



Figur 7: Histogram, lådagram och normalfördelningsplottar för 5 oberoende stickprov u1,u2,u3,u4,u5 av storlek (antalet observationer) 10 med gränserna 13 +/- 13sqrt(3) enligt uppgift 1.



Figur 8: Histogram, lådagram och normalfördelningsplottar för 5 oberoende stickprov u1,u2,u3,u4,u5 av storlek (antalet observationer) 10 med gränserna 13 +/- 13sqrt(3) enligt uppgift 1.

```r
par(old_par)  # återställ layout
```

Vi kan med hyfsat säkerhet säga att då n = 100 ser vi att stickproven absolut inte följer en normalfördelning. Däremot är det fortfarande svårt att avgöra om deras egentliga fördelning, histogrammen (figur 6) har ofta stora avvikelser från vad förväntas av en likformig fördelning (en konstant värde på y-axeln). Lådagrammen (figur 7) kan vara lite misledande då lådorna ser någorlunda ut som en normalfördelning men stickprov av likformig fördelning har också medianen i sitt väntevärde.

Normalfördelningsplotten (figur 8) verkar vara mest effektiv för att avgöra om stickproven följer en normalfördelning eller inte, den visar dessutom motsatsen till vad dem andra kanske gör. Lådagrammen (figur 7) visar att vi kanske har en normalfördelning ty medianen är ungefär lika med väntevärdet som är något vi kan förvänta oss hos en normalfördelning. Histogrammen (figur 6) visar en någorlunda normalfördelning men inte riktigt helt pga. dem avvikelser som nämndes tidigare, det tyder på att vi inte har en normalfördelning.

Uppgift 2.3 - Exponentialfördelade data

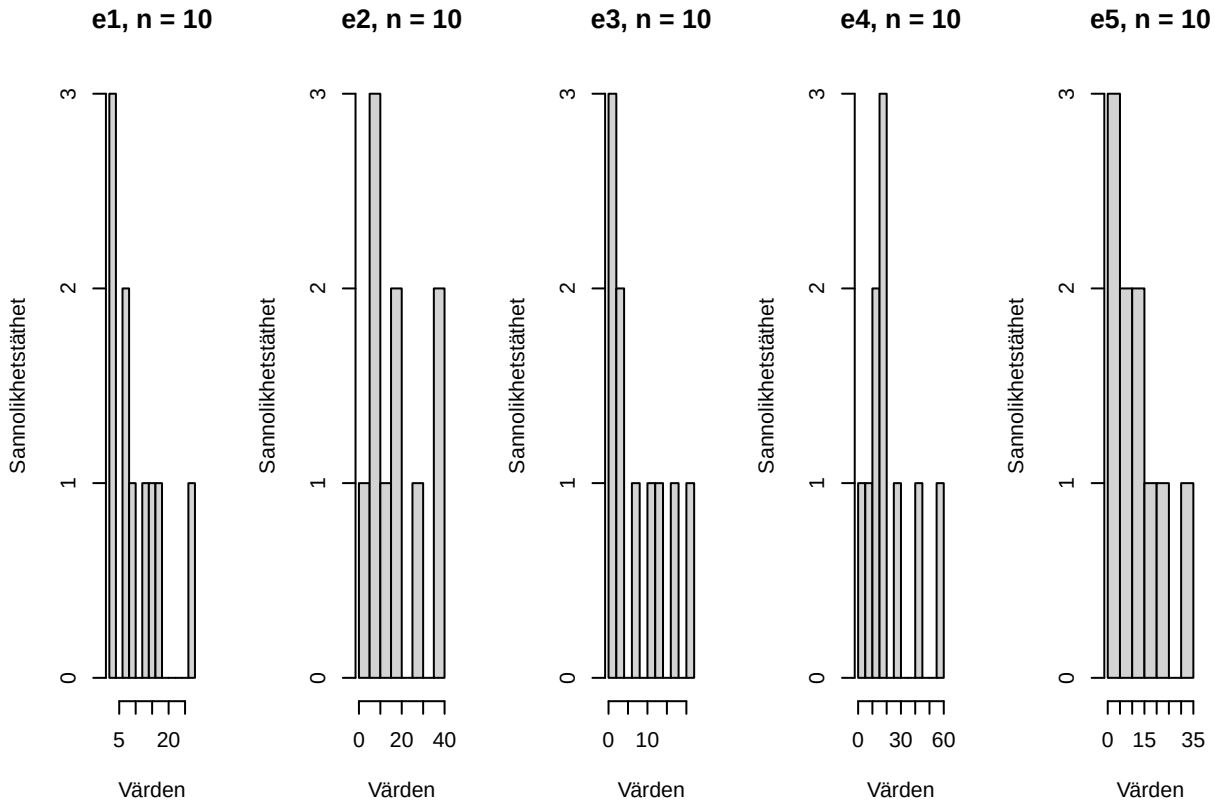
```
set.seed(20040911)

# välj a såsom uppgift 1 kräver
a <- 13
n <- 10

# definiera parametern beta
beta <- 1/a


# generera 5 oberoende stickprov
e1 <- rexp(n, r = beta)
e2 <- rexp(n, r = beta)
e3 <- rexp(n, r = beta)
e4 <- rexp(n, r = beta)
e5 <- rexp(n, r = beta)

# plotta histogram
old_par <- par(mfrow = c(1, 5))
hist(e1, main = paste("e1, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e2, main = paste("e2, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e3, main = paste("e3, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e4, main = paste("e4, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e5, main = paste("e5, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
```

```
par(old_par)
```

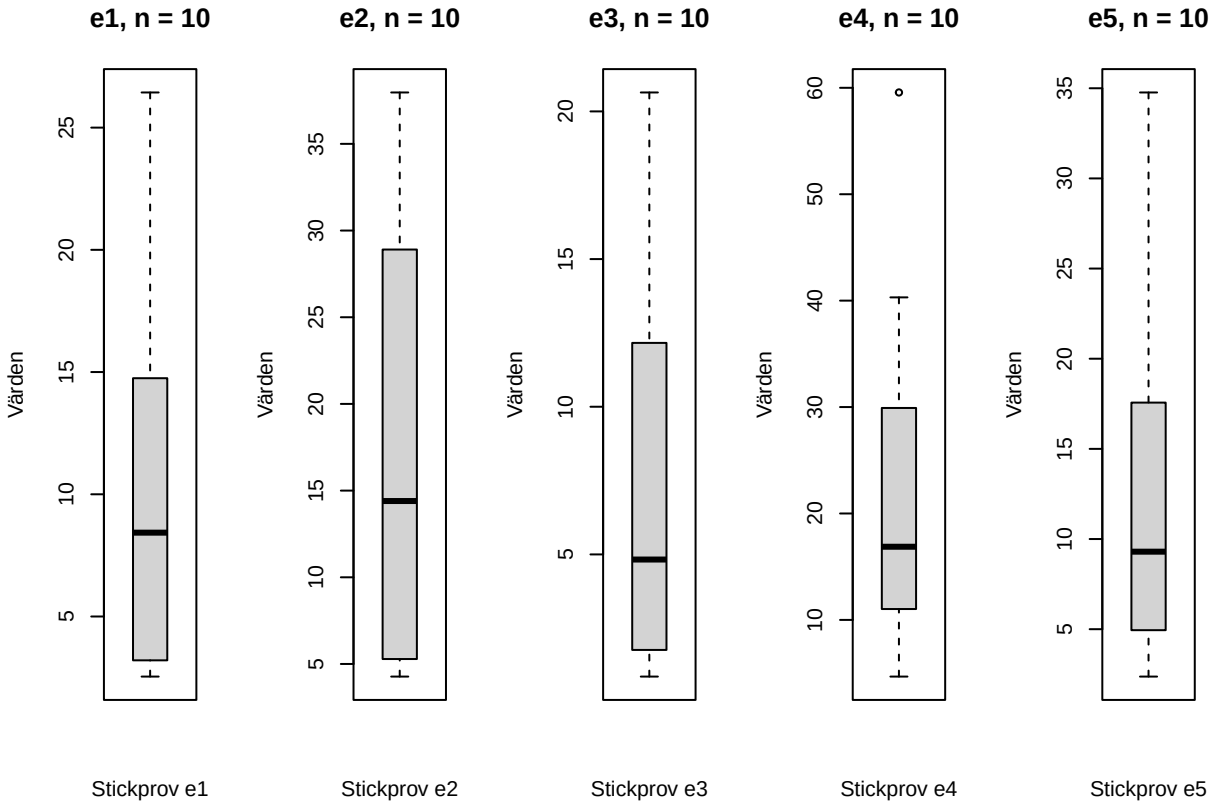
```
# plotta lådagram
old_par <- par(mfrow = c(1, 5))
boxplot(e1, main = paste("e1, n =", n), xlab = "Stickprov e1", ylab = "Värden", breaks = 10)
boxplot(e2, main = paste("e2, n =", n), xlab = "Stickprov e2", ylab = "Värden", breaks = 10)
boxplot(e3, main = paste("e3, n =", n), xlab = "Stickprov e3", ylab = "Värden", breaks = 10)
boxplot(e4, main = paste("e4, n =", n), xlab = "Stickprov e4", ylab = "Värden", breaks = 10)
boxplot(e5, main = paste("e5, n =", n), xlab = "Stickprov e5", ylab = "Värden", breaks = 10)
```

```
par(old_par)
```

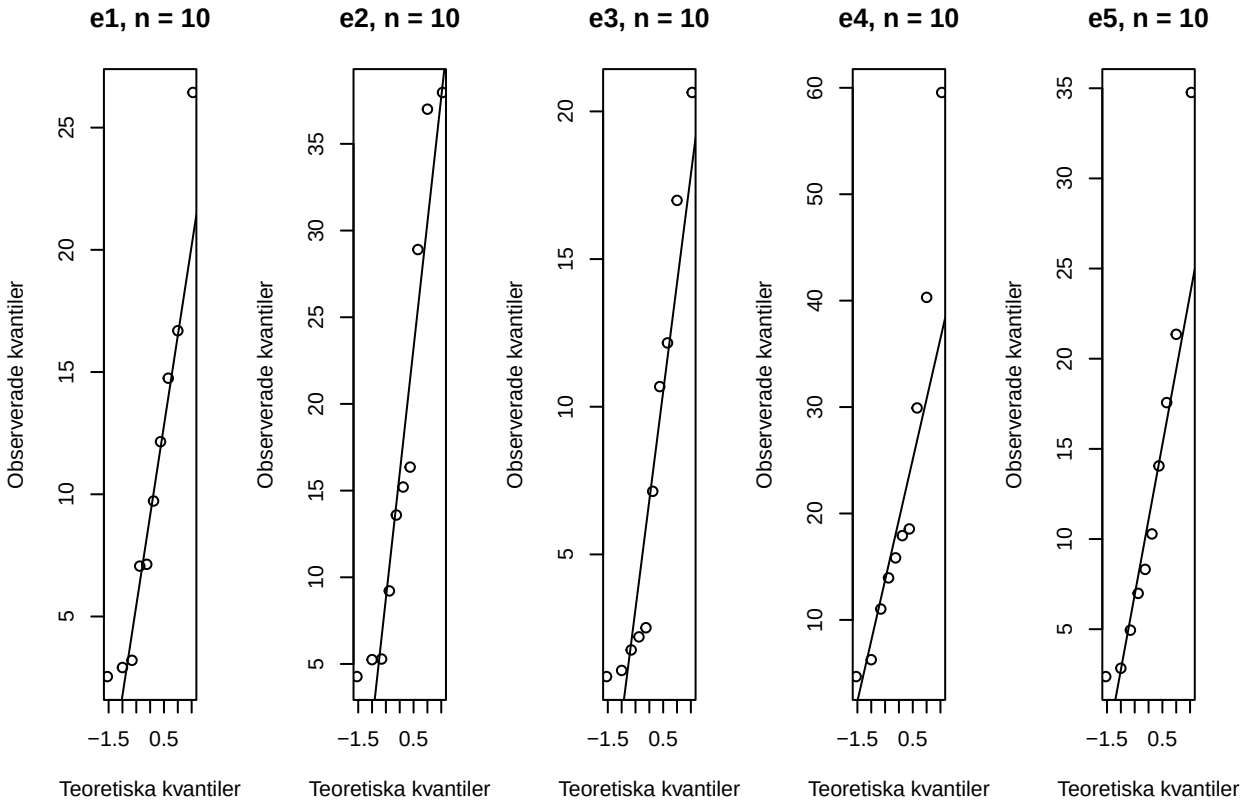
```
# plotta normalfördelningsplot
old_par <- par(mfrow = c(1, 5))
qqnorm(e1, main = paste("e1, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e1)
qqnorm(e2, main = paste("e2, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e2)
qqnorm(e3, main = paste("e3, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e3)
qqnorm(e4, main = paste("e4, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e4)
qqnorm(e5, main = paste("e5, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e5)
```



Figur 9: Histogram för 5 oberoende stickprov där antalet observationer är 10 med parametern 1/13



Figur 10: Lådagram för 5 oberoende stickprov av storlek 10.



Figur 11: Normalfördelningsplottar för 5 oberoende stickprov av storlek 10.

```r
par(old_par)
```

Redan när n = 10 kan vi se hos lådagrammen (figur 10) ganska tydligt se att det inte är en normalfördelning pga. dem låga medianer. Därmed kan vi nog säga att lådagrammen (figur 10) verkar mest effektiv här eftersom normalfördelningsplotten (figur 11) och histogrammen (figur 9) kräver ungefär n = 100 för att avgöra att stickproven inte följer en normalfördelning, dvs ett större värde på n för att med säkerhet se att i det här fallet inte är ett normalfördelning.

Uppgift 3: Explorativ dataanalys

Vi ska utforska csv-filen "olvinsprit.csv" som har 4 variabler "Land", "beer", "vin", och "sprit", där den genomsnittliga konsumtionen har mätts i 18 st OECD-länder.

```
data <- read.csv("olvinsprit.csv", header = TRUE)

land <- data$Land
beer <- data$beer
vin <- data$vin
sprit <- data$sprit
```

Vi kan börja utforska om någon alkoholtyp kan anses vara normalfördelad och gör detta först genom att skapa histogram för vardera alkoholtyp där vi även ritar in kurvan för en normalfördelning med väntevärde = medelvärdet av alkoholtypen samt standardavvikelse = standardavvikelsen för alkoholtypen.

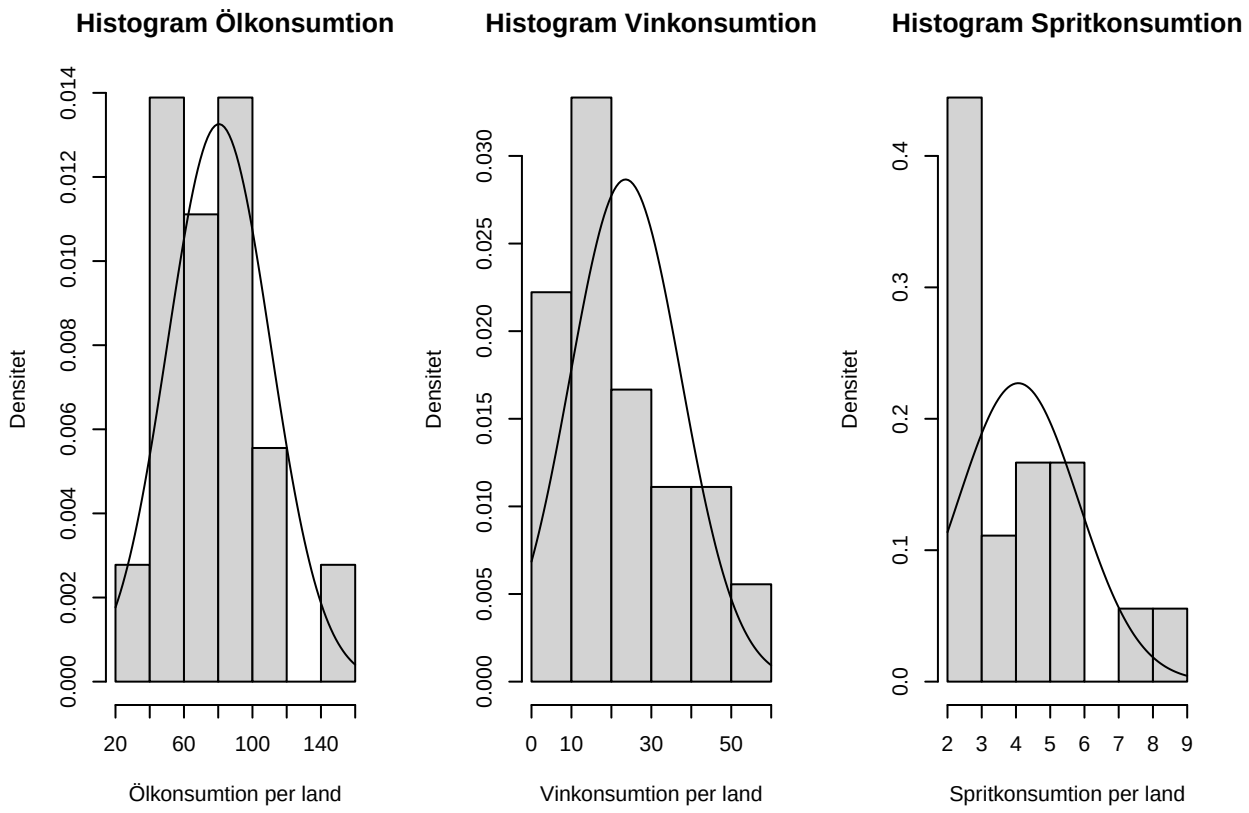
Från figur 12 ser ölkonsumtionen ut att kunna vara normalfördelad medan vin- och spritkonsumtionen mer liknar en exponentialfördelning.

```
old_par <- par(mfrow = c(1, 3)) # 1 rad, 3 kolonner

hist(beer, prob = TRUE, main = "Histogram Ölkonsumtion",
     xlab = "Ölkonsumtion per land", ylab ="Densitet")
x <- seq(from = 20, to = 160, length.out = 100)
lines(x, dnorm(x, mean(beer), sd(beer)))

hist(vin, prob = TRUE, main = "Histogram Vinkonsumtion",
     xlab = "Vinkonsumtion per land", ylab ="Densitet")
x <- seq(from = 0, to = 60, length.out = 100)
lines(x, dnorm(x, mean(vin), sd(vin)))

hist(sprit, prob = TRUE, main = "Histogram Spritkonsumtion",
     xlab = "Spritkonsumtion per land", ylab ="Densitet")
x <- seq(from = 2, to = 9, length.out = 100)
lines(x, dnorm(x, mean(sprit), sd(sprit)))
```

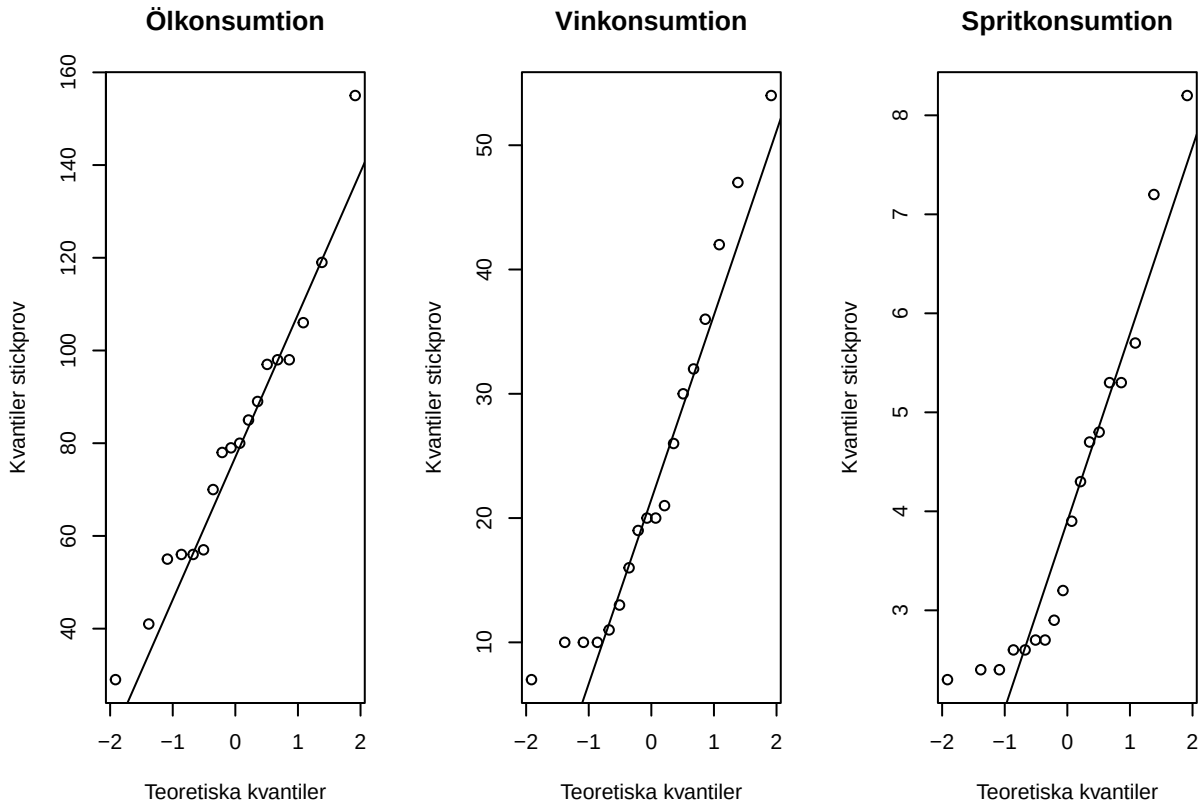
```
par(old_par)
```

```
old_par <- par(mfrow = c(1, 3)) # 1 rad, 3 kolonner

qqnorm(beer, main = "Ölkonsumtion", xlab = "Teoretiska kvantiler", ylab = "Kvantiler stickprov")
qqline(beer)

qqnorm(vin, main = "Vinkonsumtion", xlab = "Teoretiska kvantiler", ylab = "Kvantiler stickprov")
qqline(vin)

qqnorm(sprit, main = "Spritkonsumtion", xlab = "Teoretiska kvantiler", ylab = "Kvantiler stickprov")
qqline(sprit)
```

```
par(old_par)
```



Figur 12: Data från öl, vin- och spritkonsumtion plottat med histogram och jämförd med den teoretiska normalfördelningen, representerad av kurvan.

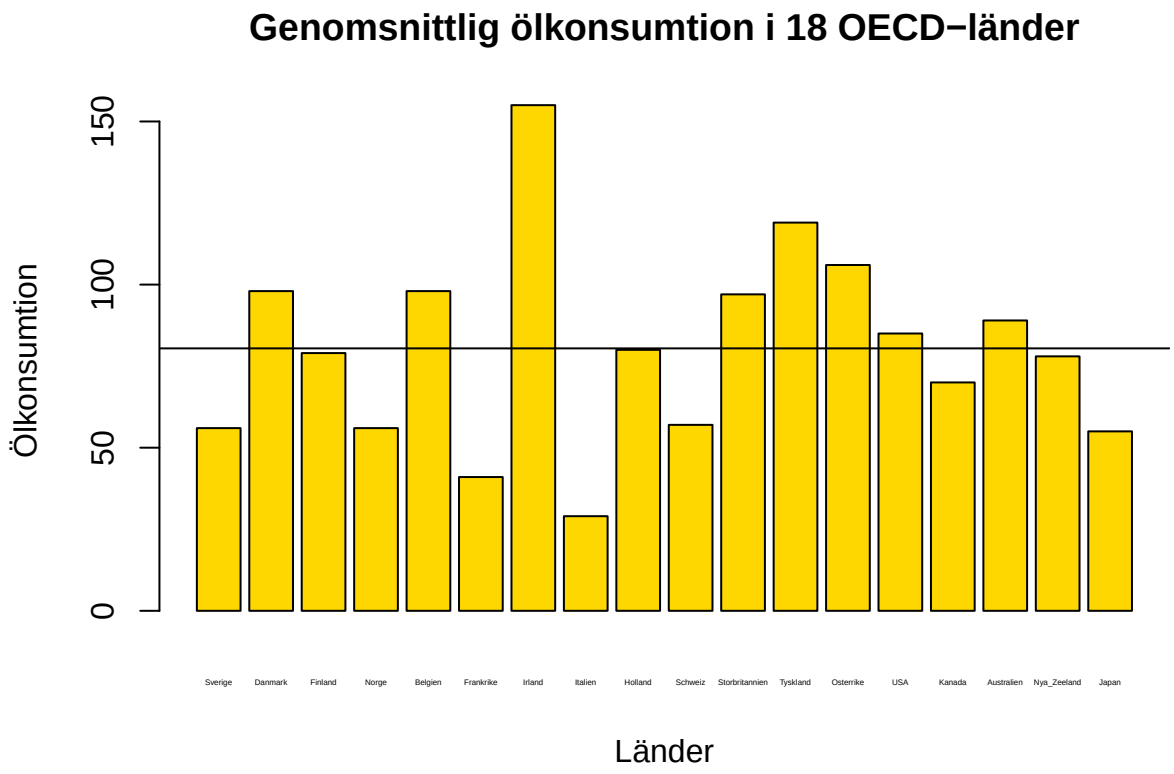


Figur 13: Normalfördelningsplottar som visar hur nära alkoholkonsumtionen för alla länder av öl, vin och sprit ligger den teoretiska normalfördelningen.

Vi kan fortsätta utforska om någon av vektorerna öl, vin eller sprit kommer från en normalfördelning genom att använda normalfördelningsplottar. Vi ser i figur 13 att det är svårare att dra någon slutsats här men spritkonsumtionen avviker något mer än de andra från den räta linjen vilket även här talar för att den ej är normalfördelad.

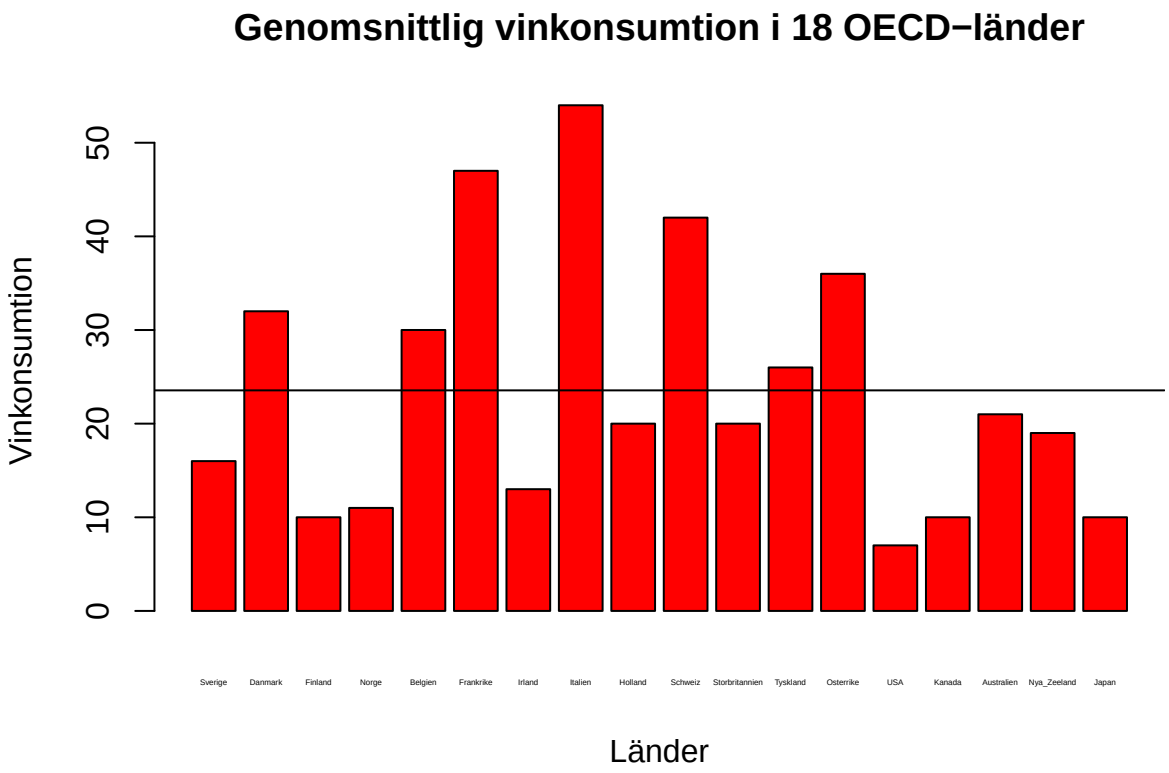
För att jämföra alkoholkonsumtionen av öl, vin och sprit mellan länderna kan vi jämföra detta i 3 stolpdiagram, ett för respektive alkoholtyp.

```
barplot(beer, names.arg = land, xlab ="Länder",
        ylab ="Ölkonsumtion", col ="gold",
        main ="Genomsnittlig ölkonsumtion i 18 OECD-länder",
        cex.names=.29)

abline(mean(beer), 0)
```



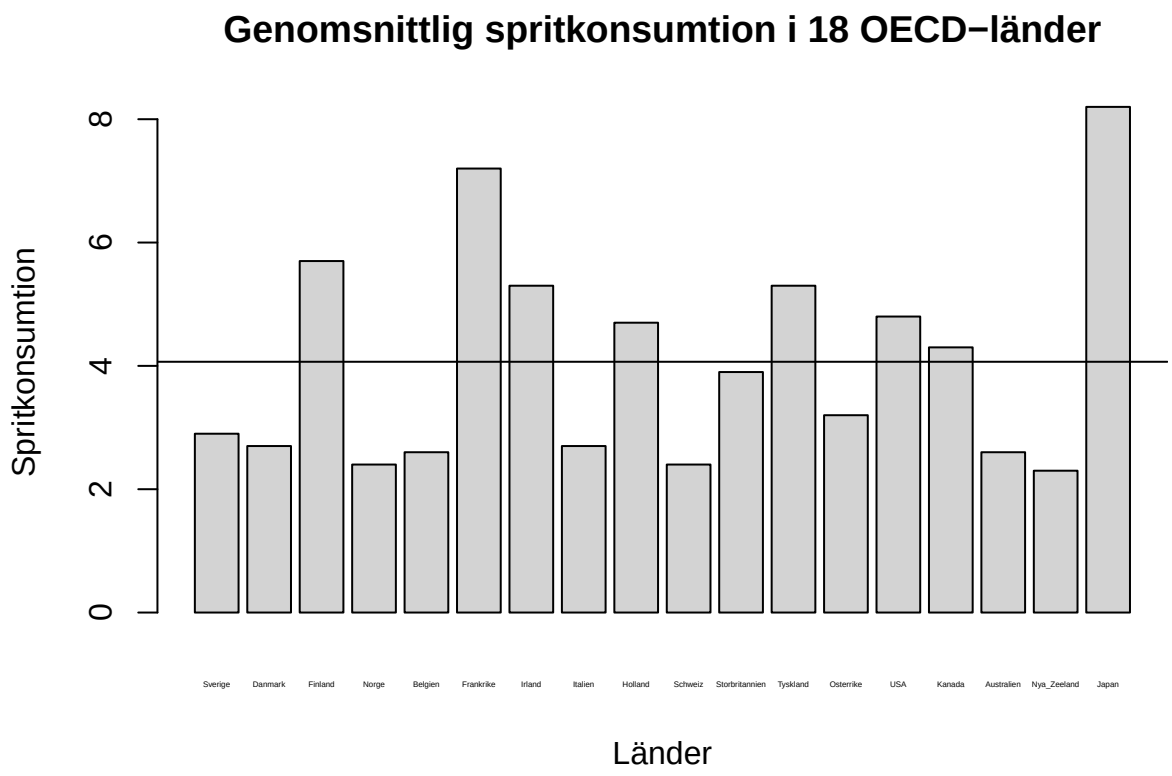
Figur 14: Stopldiagram med genomsnittlig ölkonsumtion samt en horisontell linje som visar medelkonsumtionen för alla länder.

```
barplot(vin, names.arg = land, xlab ="Länder",
        ylab ="Vinkonsumtion", col ="red",
        main ="Genomsnittlig vinkonsumtion i 18 OECD-länder",
        cex.names=.29)

abline(mean(vin), 0)
```



Figur 15: Stopldiagram med genomsnittlig vinkonsumtion samt en horisontell linje som visar medelkonsumtionen för alla länder.

```r
barplot(sprit, names.arg = land, xlab ="Länder",
        ylab ="Spritkonsumtion", col ="lightgrey",
        main ="Genomsnittlig spritkonsumtion i 18 OECD-länder",
        cex.names=.29)

abline(mean(sprit), 0)
```



Figur 16: Stopldiagram med genomsnittlig spritkonsumtion samt en horisontell linje som visar medelkonsumtionen för alla länder.

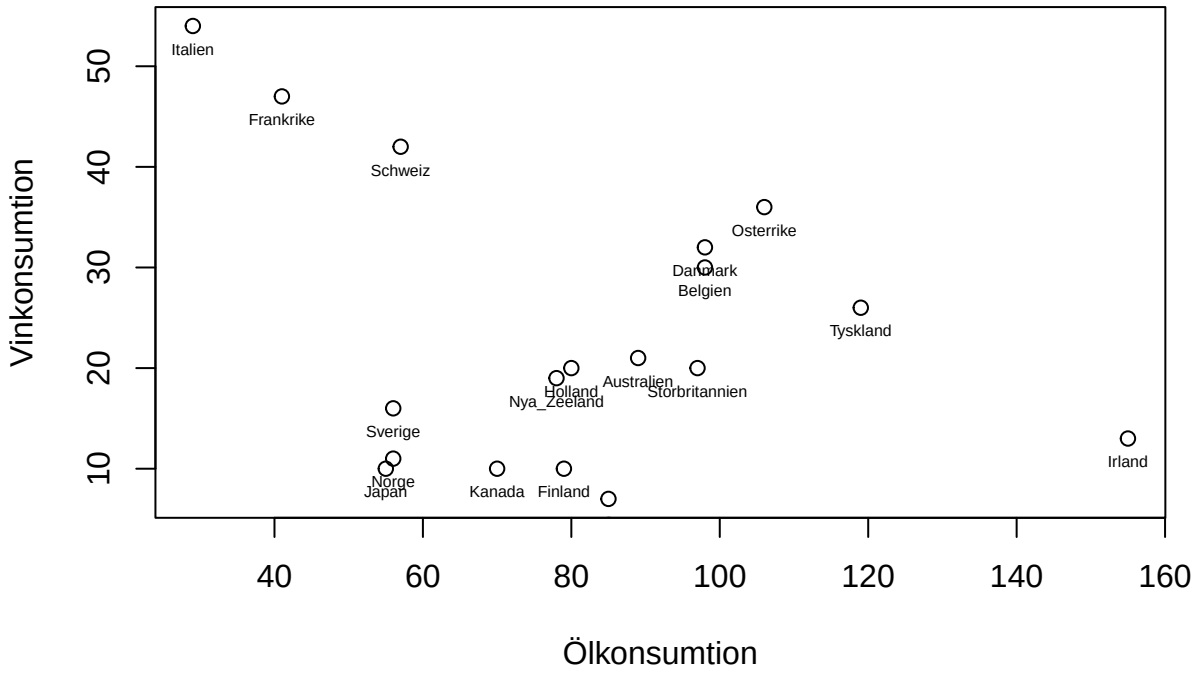
Vi ser från figurerna 14, 15 och 16 att Sverige verkligen inte avviker i någon av öl, vin- och spritkonsumtion.

Från figur 14 ser vi att Irland har överlägset högst ölkonsumtion medan figur 15 visar att Italien har högst vinkonsumtion följt av Frankrike. I figur 16 ser vi att högsta spritkonsumtionen sker i Japan, även där följt av Frankrike. Från stolpdiagrammen går det även att se att Tyskland är det enda landet som ligger över medel i samtliga alkoholtyper.

Ett någorlunda gemensamt drag hos de extrema länderna i öl- och vinkonsumtion är att de tenderar att endast ligga högt i den ena av dessa. Detta kan klart ses i figur 17 där de som ligger allra högst i antingen öl- eller vinkonsumtion ligger bland de lägsta i den andra.

Bland övriga länder kan man istället se en ganska tydlig positiv korrelation där ökad ölkonsumtion medför ökad vinkonsumtion. Samband mellan öl-sprit samt vin-sprit är svårare att säga något om.

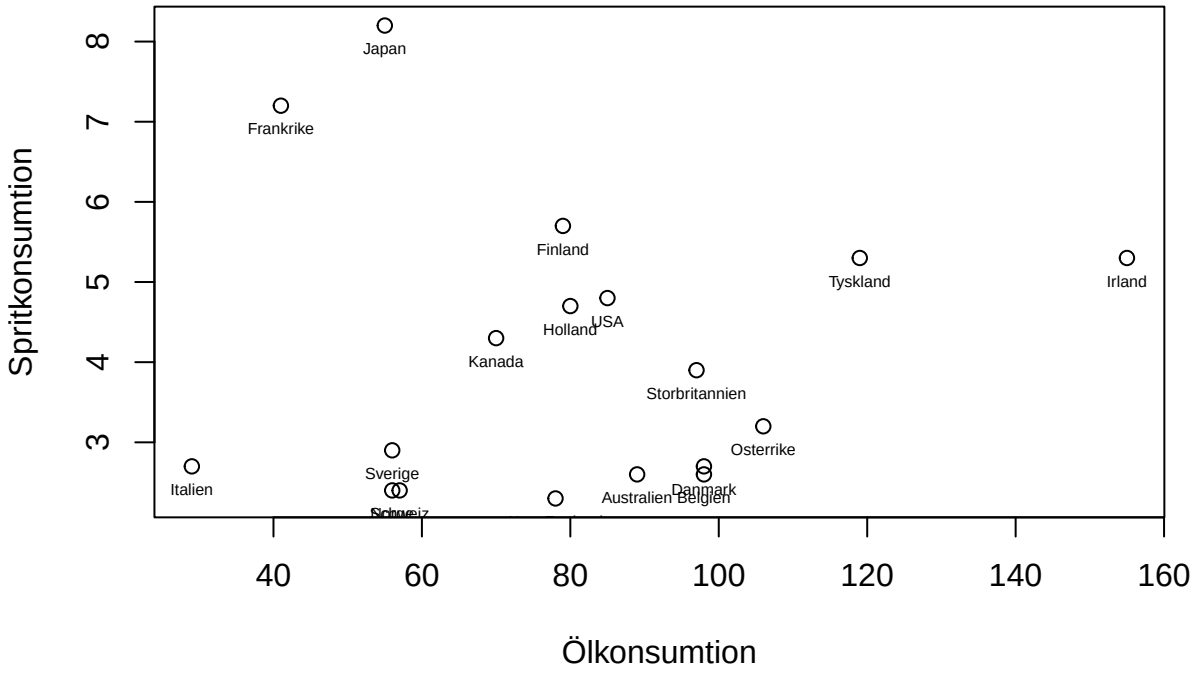
```
plot(beer, vin, xlab = "Ölkonsumtion", ylab = "Vinkonsumtion")
text(beer, vin, land, cex = 0.5, pos = 1)
```



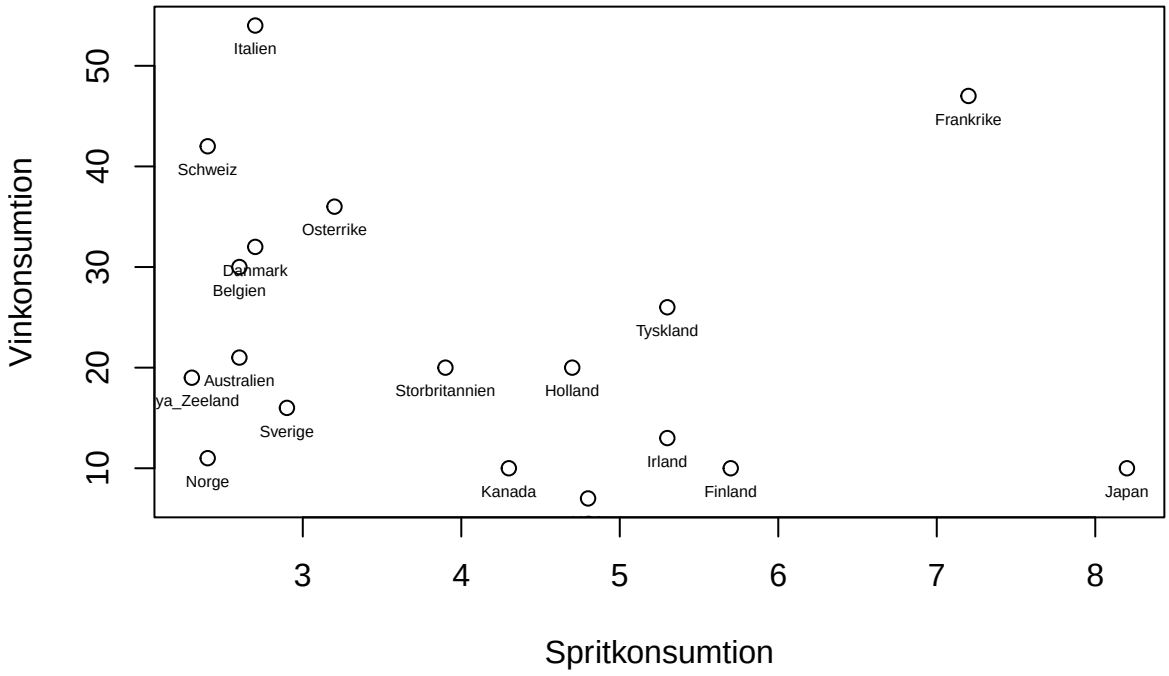
Figur 17: Scatterplot som jämför öl- och vinkonsumtionen i 18 OECD-länder.

```
plot(beer, sprit, xlab = "Ölkonsumtion", ylab = "Spritkonsumtion")
text(beer, sprit, land, cex = 0.5, pos = 1)
```

```
plot(sprit, vin, xlab = "Spritkonsumtion", ylab = "Vinkonsumtion")
text(sprit, vin, land, cex = 0.5, pos = 1)
```



Figur 18: Scatterplot som jämför öl- och spritkonsumtionen i 18 OECD-länder.

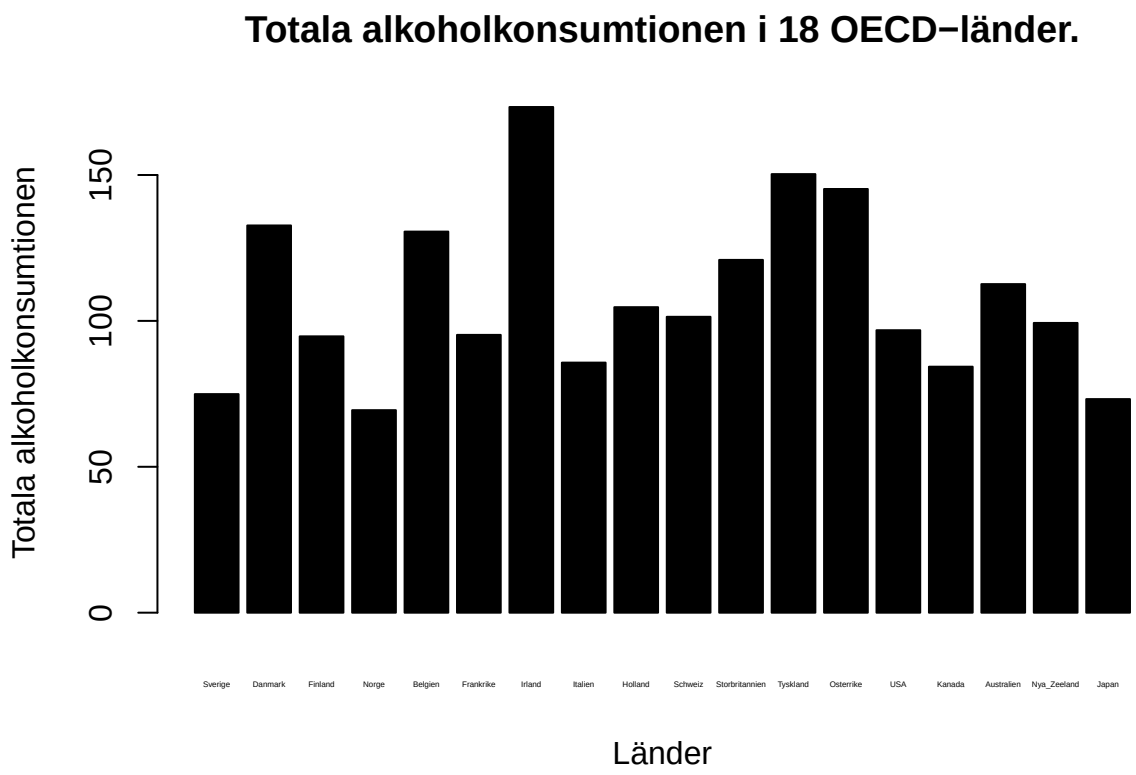


Figur 19: Scatterplot som jämför sprit- och vinkonsumtionen i 18 OECD-länder.

Vi kan också kolla på den totala alkoholkonsumtionen i länderna. Detta kan illustreras på ett smidigt sätt med ett stolpdiagram, figur 20. Här ser vi att Norge har lägsta totalkonsumtionen, tätt följt av Sverige och Japan.

Om man dock tänker sig att de negativa effekterna av alkohol är starkt korrelerad med procenten, vilket nog ändå får sägas rimligt blir detta dock en missvisande bild. T.ex. Japan hade ju den högsta spritkonsumtionen av samtliga länder, där procenten kan antas vara högre än för öl och vin men eftersom att sprit konsumeras i mindre mängder än öl och vin så ger det inte lika stor effekt på totala konsumtionen.

```
total_alk_kons = beer + vin + sprit # Skapa en vektor med totala alkoholkonsumtionen.

barplot(total_alk_kons, names.arg = land, xlab ="Länder",
        ylab ="Totala alkoholkonsumtionen", col ="black",
        main ="Totala alkoholkonsumtionen i 18 OECD-länder.",
        cex.names=.29)
```



Figur 20: Totala alkoholkonsumtionen i 18 OECD-länder.

Några sammanfattande slutsatser som kan dras är att ölkonsumtionen är den som mest troligt är normalfördelad. Inget land är extremt för samtliga alkoholtyper åt något håll. Enbart Tyskland ligger över medel i samtliga, men är istället inte högst i någon. Figur 20 visar ändå att Irland sticker ut lite när det gäller den totala konsumtionen och det finns en ganska stark negativ korrelation mellan öl- och vinkonsumtion när det gäller extrema länder. Att ett land ligger högt i öl- eller vinkonsumtion tyder alltså inte lika mycket på att de dricker stora mängder alkohol totalt som preferens för en viss alkoholtyp.