

Laboration 1 - MT4001

Daniel Svedlund, Sebastijan Babic

2024-11-18

Sammanfattning

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 vårt personnummer, 13.

1. Vilken är det minsta stickprovsstorleken som behövs för att den fördelning ni simulerar från skall avslöja sig som normal eller icke-normal?
2. Vilken grafisk metod anser ni är mest effektiv för att avgöra om ett stickprov är normalfördelat eller inte? Motivera med de olika grafiska metoderna (boxplot, histogram, normalfördelningsplott)

Uppgift 2.1 Normalfördelade data

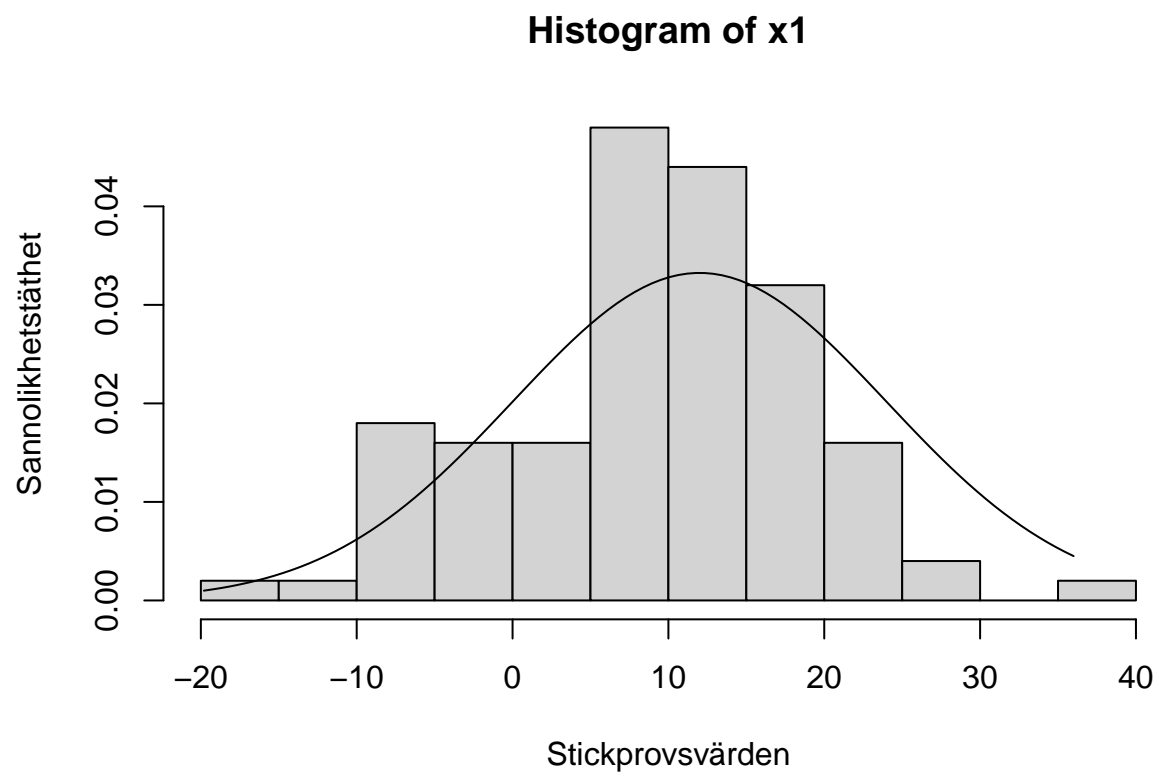
```
set.seed(20040911)

n <- 100

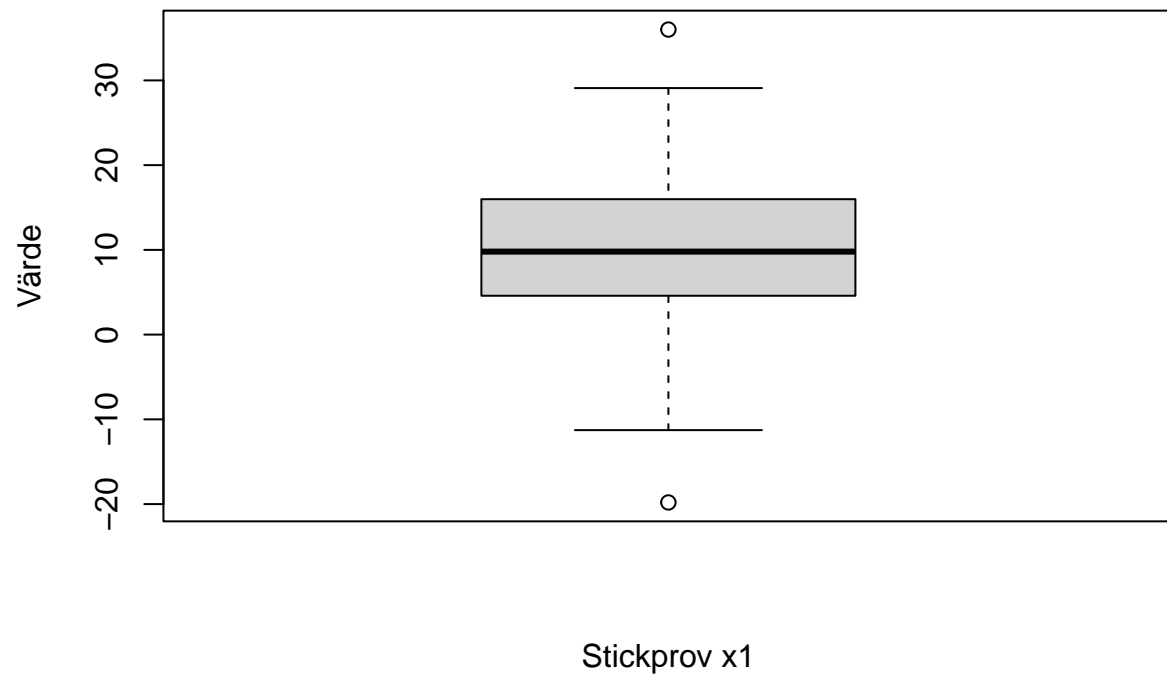
x1 <- rnorm(n, 12, 12)

hist(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 linjen.
# kan göra manuellt och sätta från -10 till 25 men detta gör det enklare om man vill på något sätt förä
# stickproven
lines(x, dnorm(x, 12, 12))

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

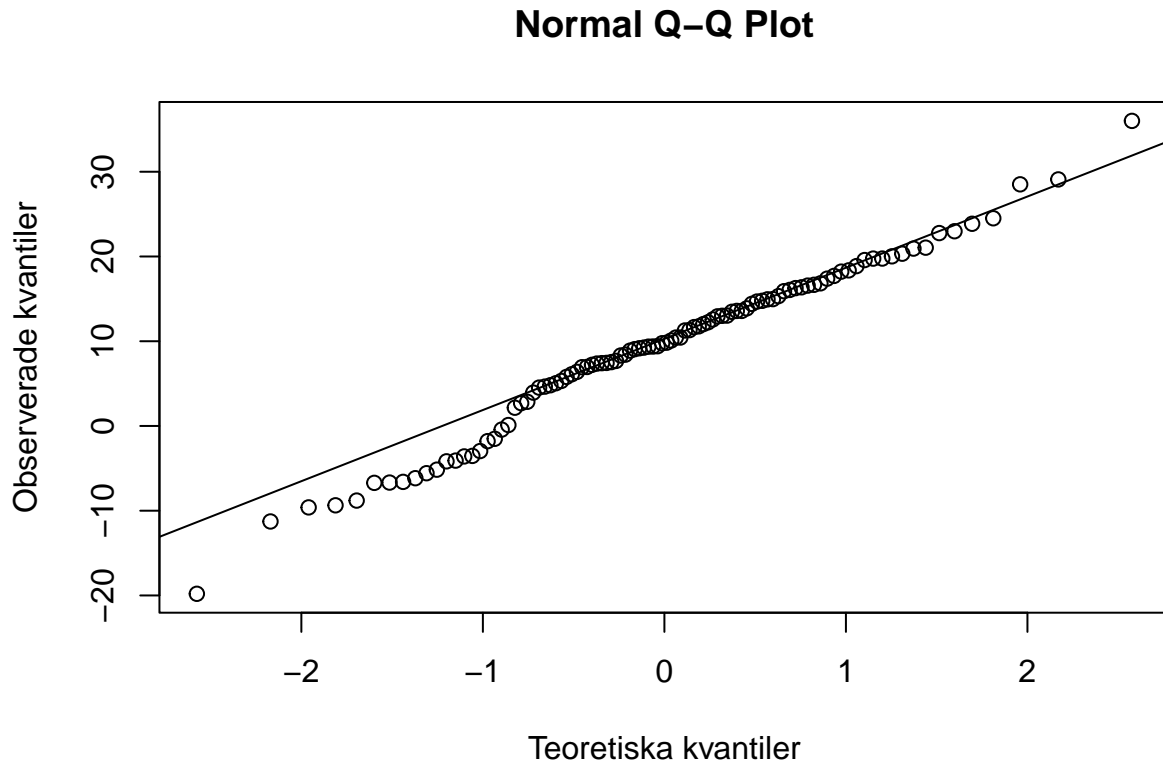


Figur 1: Histogram över en stickprov x_1 av storlek $n = 100$.



Figur 2: Lådagram över en stickprov x_1 av storlek $n = 100$.

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

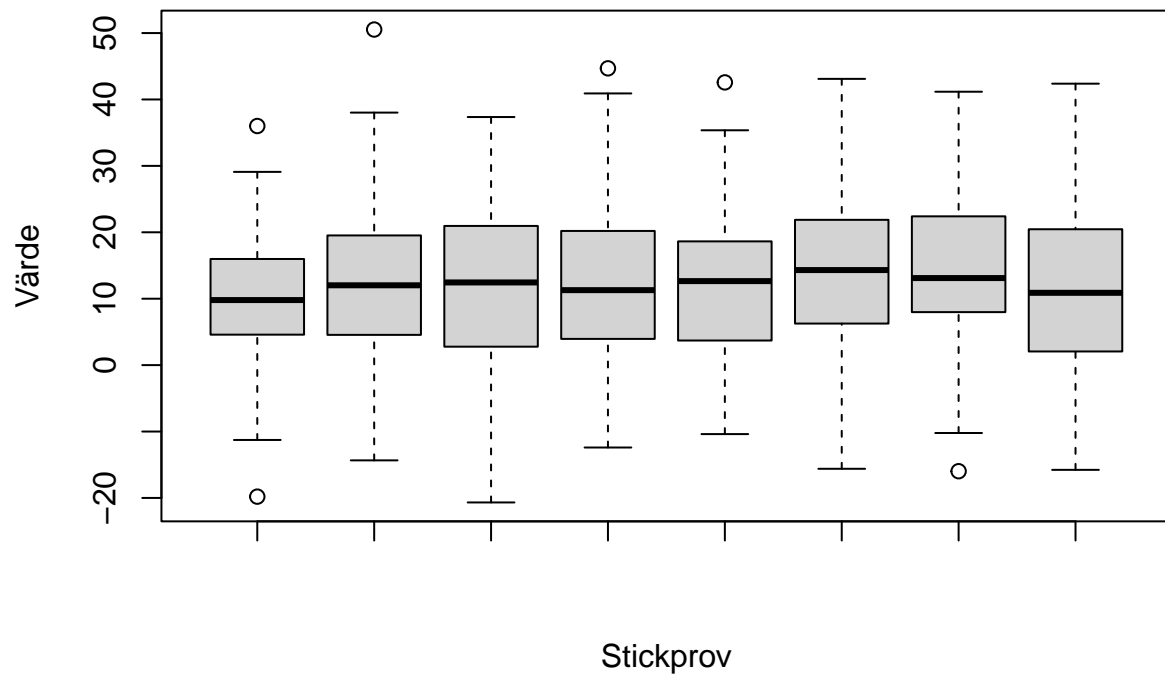


Figur 3: Normalfördelningsplot över en stickprov x_1 av storlek $n = 100$.

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

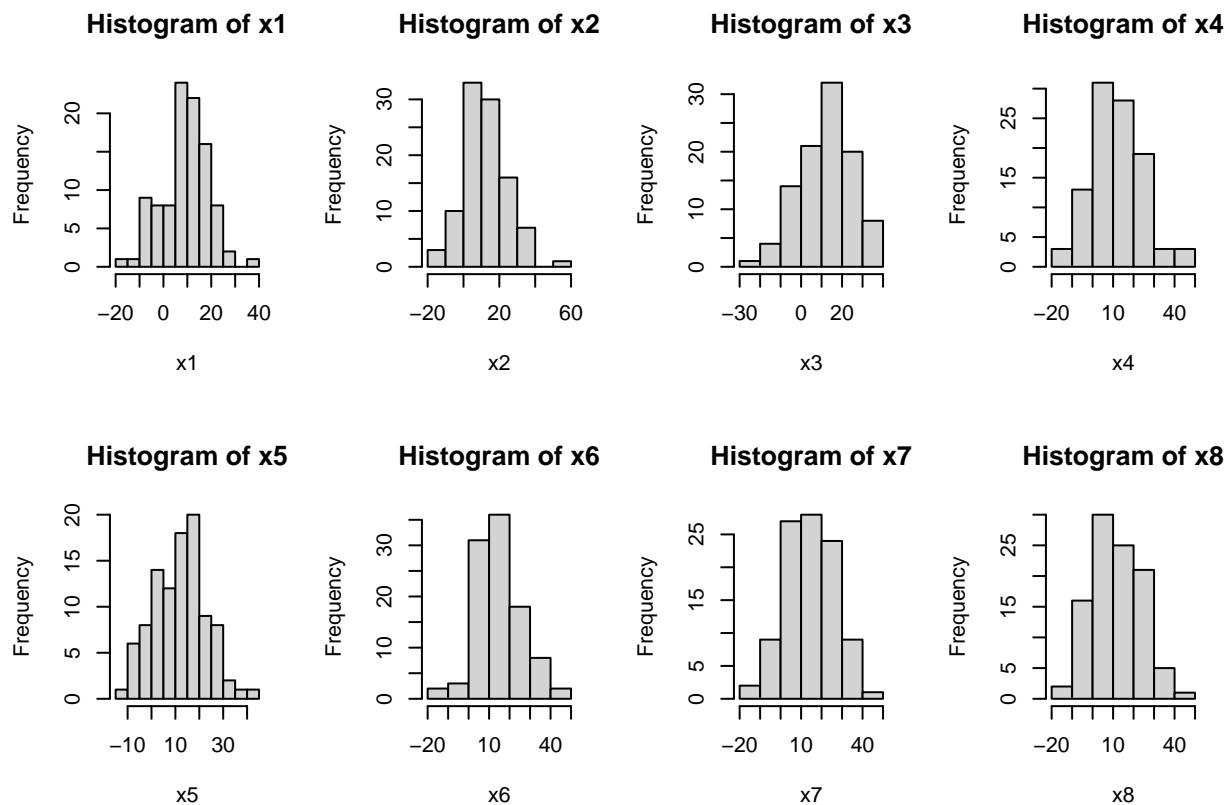
set.seed(20040911)
x1 <- rnorm(n, 12, 12) # kommer att bli samma stickprov som ovan då vi har samma seed
x2 <- rnorm(n, 12, 12) # detta blir ett stickprov annorlunda från x1, likaså de nedan
x3 <- rnorm(n, 12, 12)
x4 <- rnorm(n, 12, 12)
x5 <- rnorm(n, 12, 12)
x6 <- rnorm(n, 12, 12)
x7 <- rnorm(n, 12,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

boxplot(x1, x2, x3, x4, x5, x6, x7, x8, xlab = "Stickprov", ylab = "Värde")
```



Figur 4: 8 lådagran i en samma plot som visar 8 olika stickprov av storlek 10 och deras utseende.

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



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

```
par(old_par)
```

Då $n = 100$ kan man hyfsat enkelt avgöra fördelningen på x_1, x_2, \dots, x_8 följer en normalfördelning eller inte. Däremot är det fortfarande svårt att avgöra fördelningen på normalfördelningsplotten. Lådagrammen och histogrammen 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 lådagrammen vara lite vilseledande som vi kan se i uppgift 2.2 så histogrammen verkar vara effektivast här.

Uppgift 2.2 - Likformigt fördelade data

```
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)

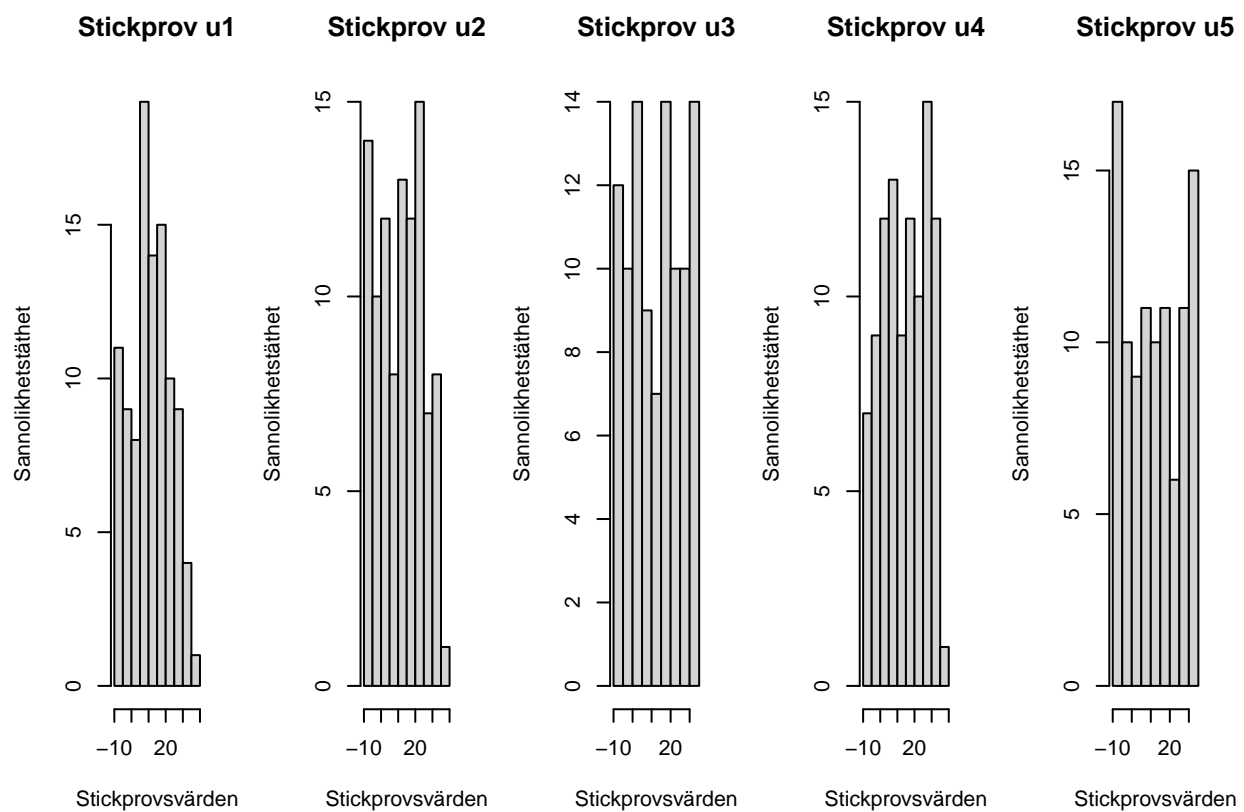
# plotta histogram
old_par <- par(mfrow = c(1, 5)) # skapa layout för flera plots och repetera för alla 3 typer av plotta
hist(u1, main = "Stickprov u1", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(u2, main = "Stickprov u2", xlab = "Stickprovsvärden", ylab = "Sannolikhetstäthet",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

par(old_par)

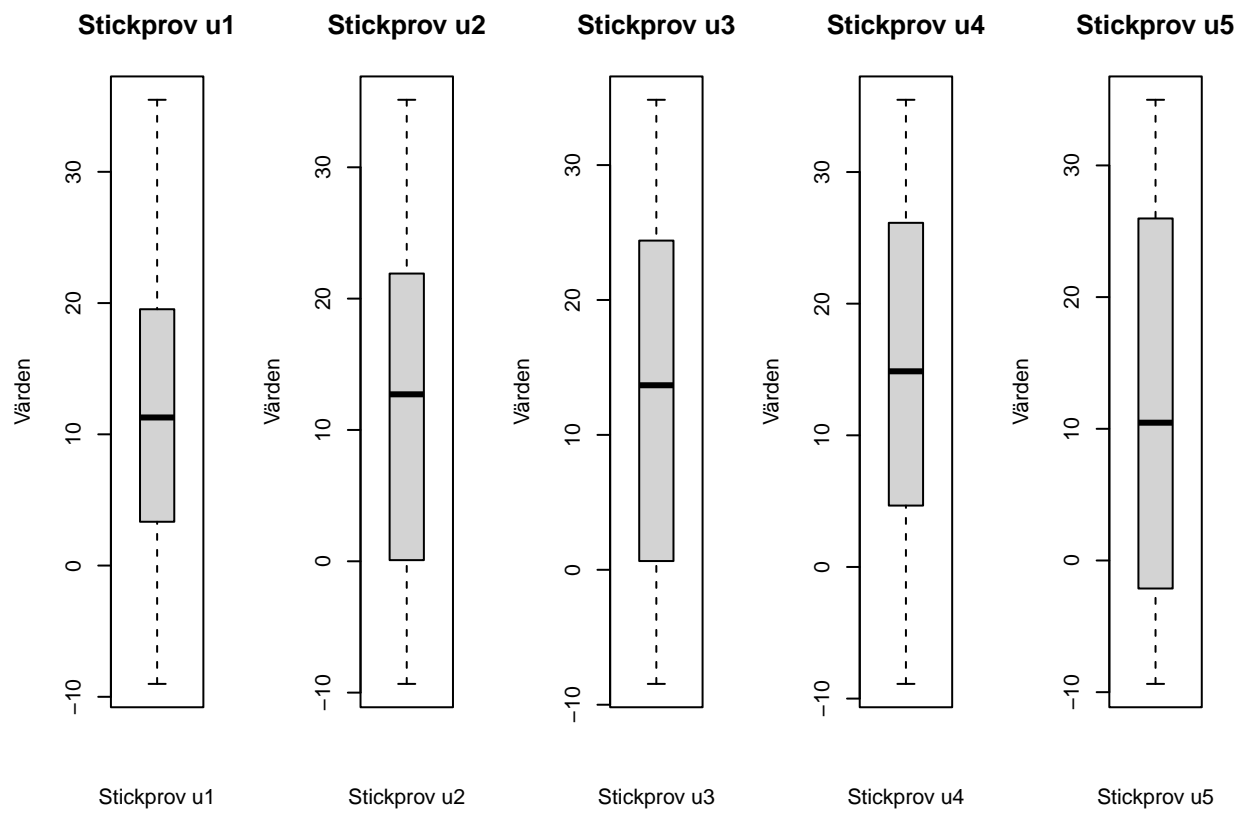
old_par <- par(mfrow = c(1, 5))
boxplot(u1, main = "Stickprov u1", xlab = "Stickprov u1", ylab = "Värden")
boxplot(u2, main = "Stickprov u2", xlab = "Stickprov u2", ylab = "Värden")
boxplot(u3, main = "Stickprov u3", xlab = "Stickprov u3", ylab = "Värden")
boxplot(u4, main = "Stickprov u4", xlab = "Stickprov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

par(old_par)

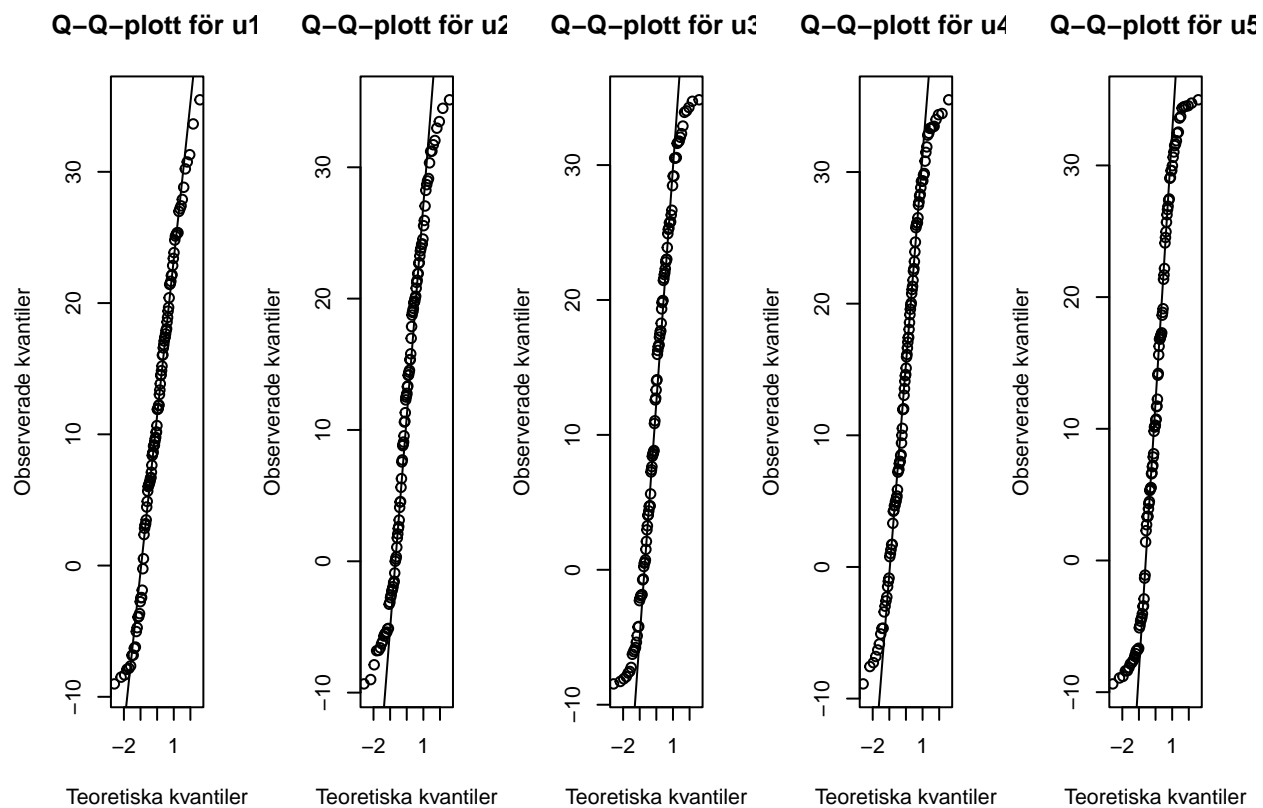
old_par <- par(mfrow = c(1, 5))
qqnorm(u1, main = "Q-Q-plott för u1", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u1)
qqnorm(u2, main = "Q-Q-plott för u2", xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(u2)
qqnorm(u3, main = "Q-Q-plott för u3", xlab =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



Figur 6: Histogram, lådagran och normalfördelningsplottar för 5 stickprov u_1, u_2, u_3, u_4, u_5 av likformig fördelning av storlek 10 med gränserna $13 \pm 13\sqrt{3}$.



Figur 7: Histogram, lådagran och normalfördelningsplottar för 5 stickprov u_1, u_2, u_3, u_4, u_5 av likformig fördelning av storlek 10 med gränserna $13 \pm 13\sqrt{3}$.



Figur 8: Histogram, lådagran och normalfördelningsplottar för 5 stickprov u_1, u_2, u_3, u_4, u_5 av likformig fördelning av storlek 10 med gränserna $13 \pm 13\sqrt{3}$.

```
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 har ofta stora avvikelser från vad förväntas av en likformig fördelning (en konstant värde på y-axeln). Lådagrammen kan vara lite misledande då lådorna ser någorlunda ut som en normalfördelning men i verklighet inte är det.

Normalfördelningsplotten verkar vara mest effektiv för att avgöra om stickproven följer en normalfördelning eller inte, den visar dessutom motsatsen till vad dem andra kanske gör. Boxplot 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 visar en någorlunda normalfördelning men inte riktigt helt pga. dem avvikelser som nämndes tidigare, det tyder på att vi inte har en normalfördelning eftersom vi skulle följa en mycket mer steg-för-steg ökning på denna axel och därmed få något mer likt 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)

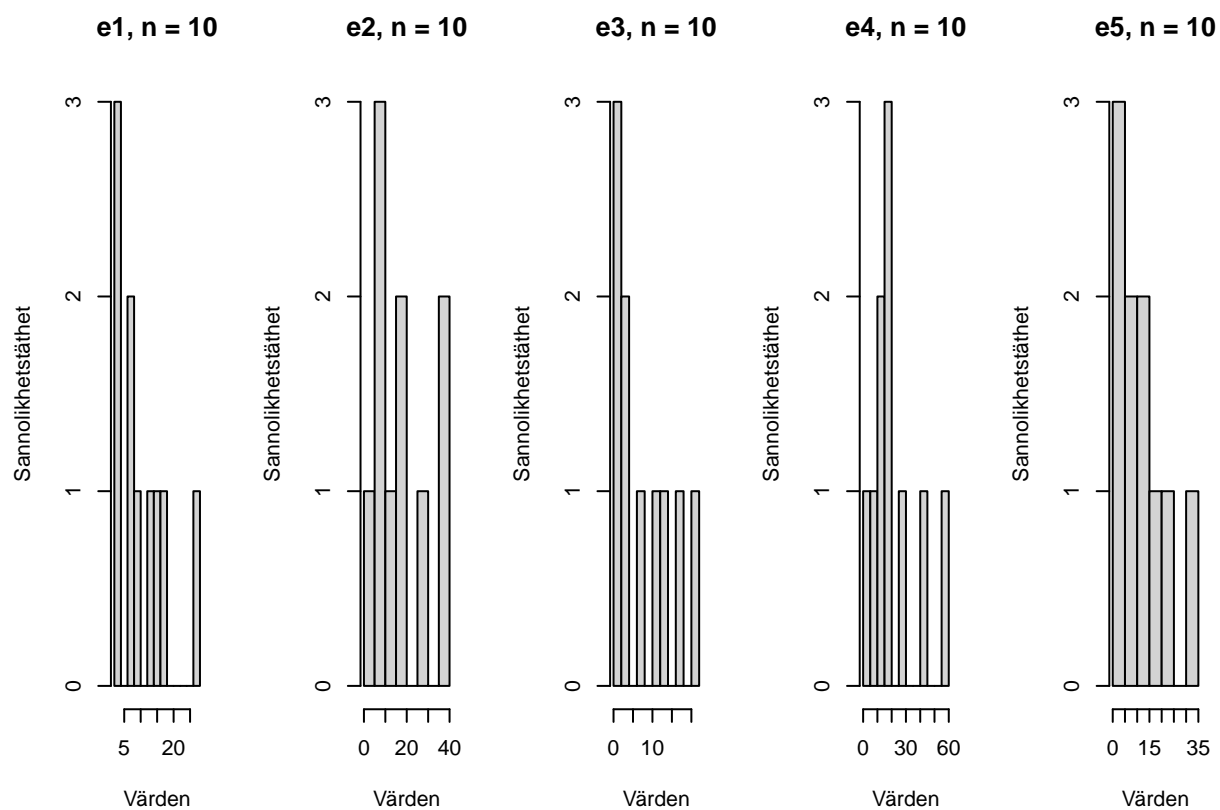
# plotta histogram
old_par <- par(mfrow = c(1, 5))
hist(e1, main = paste("e1, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e2, main = paste("e2, n =", n), xlab = "Värden", ylab = "Sannolikhetstäthet", breaks = 10)
hist(e3, main = paste("e3, n =", n), xlab = "Värden", ylab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

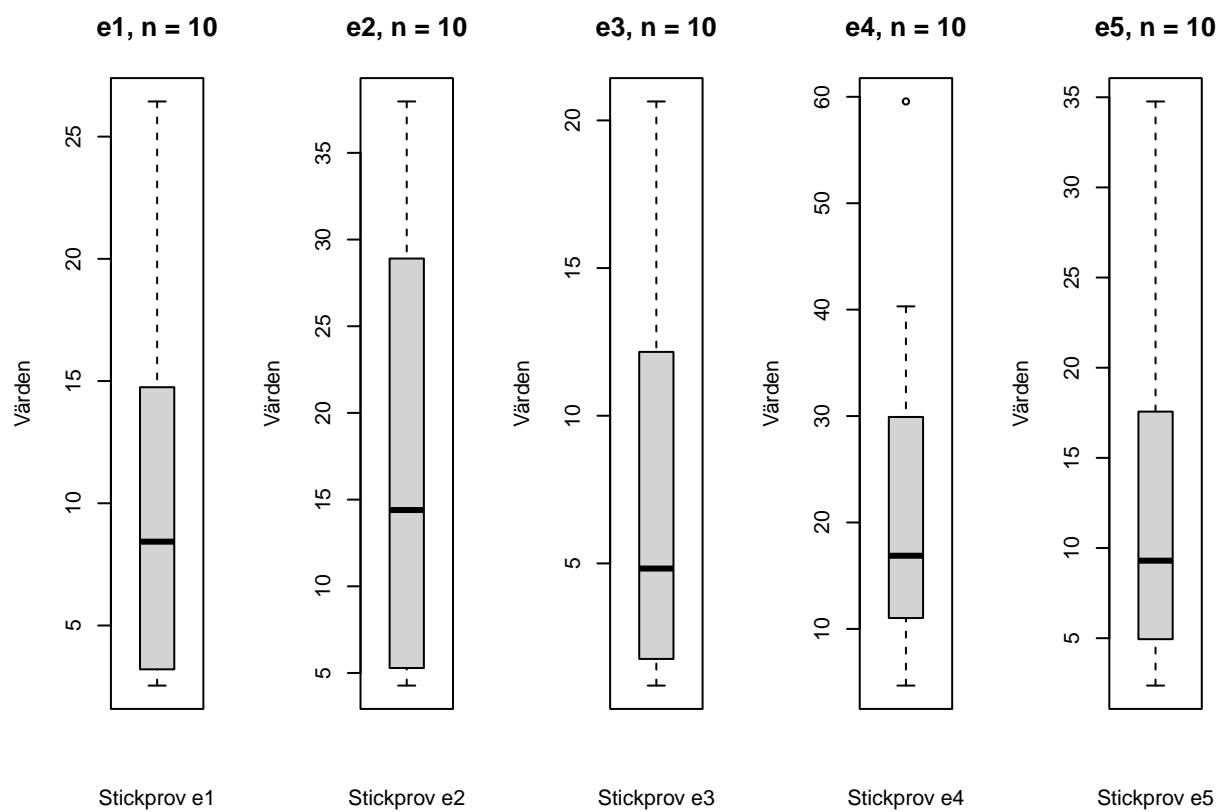
```
# plotta histogram
old_par <- par(mfrow = c(1, 5))
boxplot(e1, main = paste("e1, n =", n), xlab = "Stickprov e1", ylab = "Värden", breaks = 10)
boxplot(e2, main = paste("e2, n =", n), xlab = "Stickprov e2", ylab = "Värden", breaks = 10)
boxplot(e3, main = paste("e3, n =", n), xlab = "Stickprov e3",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

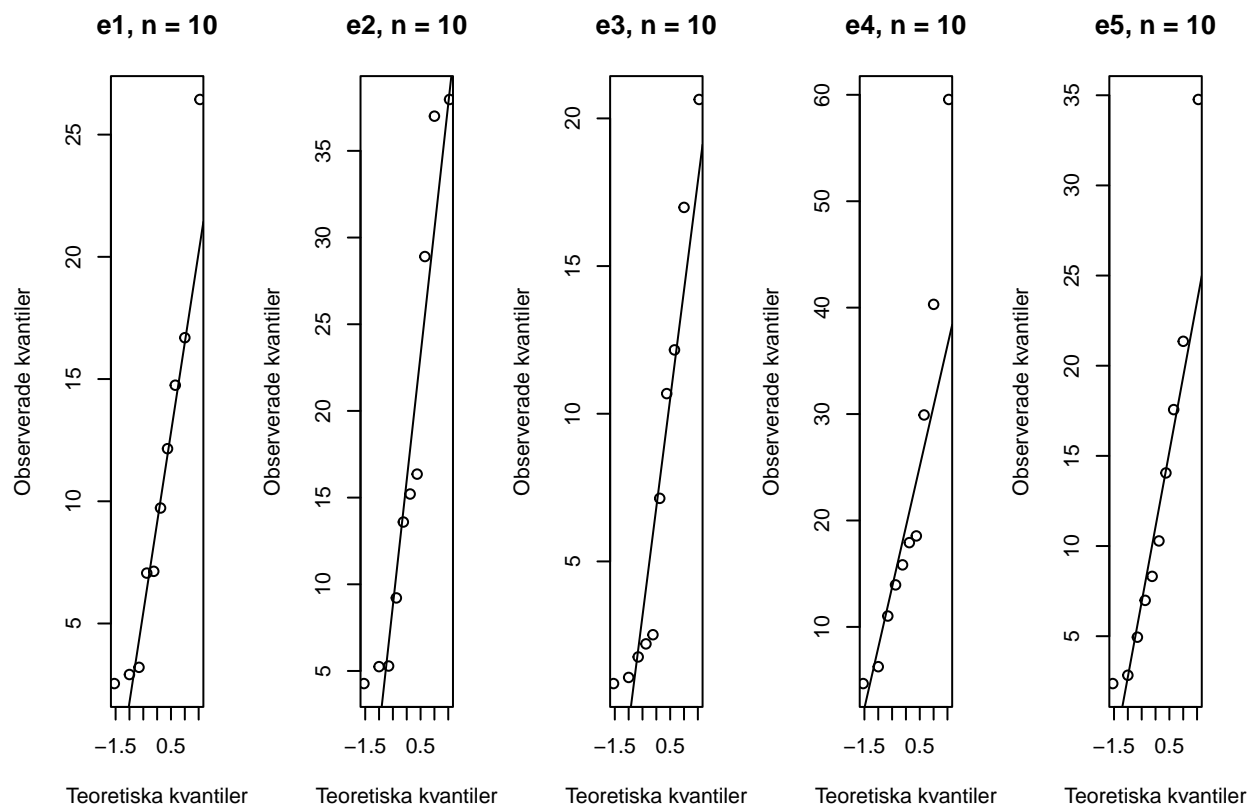
```
# plotta histogram
old_par <- par(mfrow = c(1, 5))
qqnorm(e1, main = paste("e1, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e1)
qqnorm(e2, main = paste("e2, n =", n), xlab = "Teoretiska kvantiler", ylab = "Observerade kvantiler")
qqline(e2)
qqnorm(e3, main = paste("e3, 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 exponentialfördelade stickprov där antalet observationer är 10 med parametern $1/13$



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



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

```
par(old_par)
```

Redan när $n = 10$ kan vi se hos lådagrammen ganska tydligt se att det inte är en normalfördelning pga. de låga medianer. Därmed kan vi nog säga att lådagrammen verkar mest effektiv här eftersom Q-Q plotten och histogrammen kräver ungefär $n = 100$ för att avgöra att stickproven inte följer en normalfördelning.

Sammanfattning

För att pålitligt bedöma om data är normalfördelade krävs ofta större stickprovsstorlekar ($n \geq 100$) men det beror på vad för fördelning det är vi jobbar med.

- Normalfördelade stickprov kräver $n = 100$ där histogrammen är mest effektiv.
- Likformigt fördelade stickprov kräver $n = 100$ där normalfördelningsplotten är mest effektiv.
- Exponentialfördelade stickprov kräver $n = 10$ där lådagrammen verkar vara mest effektiv.

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 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.

```
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)
```



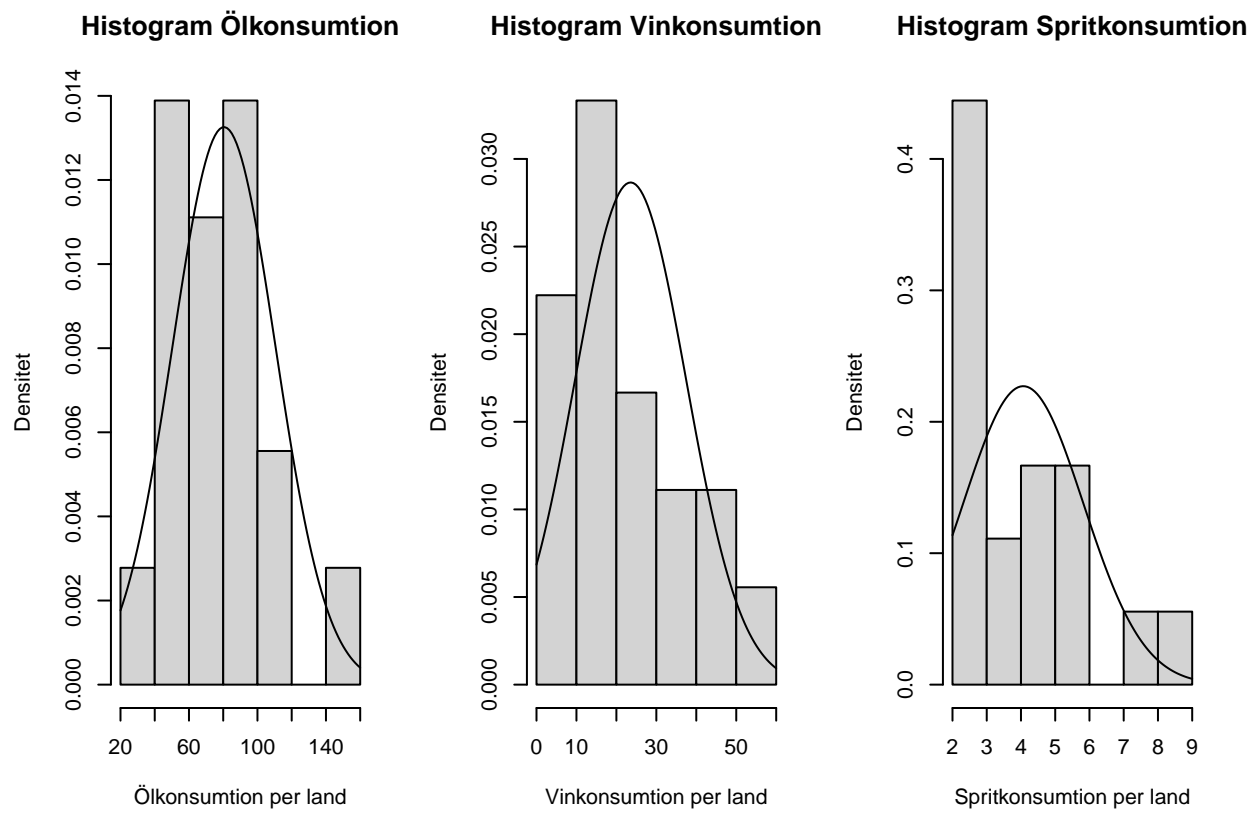
Från figuren ser ölkonsumtionen ut att kunna vara normalfördelad medan vin och sprit mer liknar en exponentialfördelning.

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

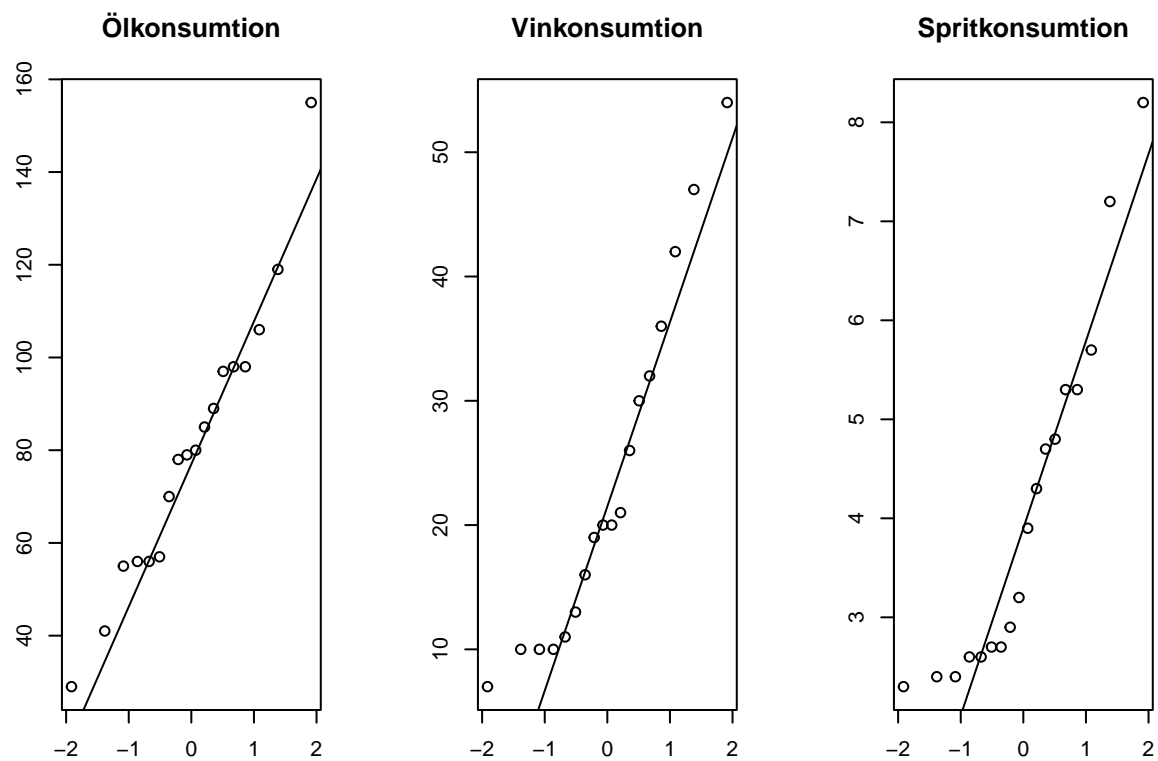
qqnorm(beer, main = "Ölkonsumtion", xlab = " ", ylab = " ")
qqline(beer)


qqnorm(vin, main = "Vinkonsumtion", xlab = " ", ylab = " ")
qqline(vin)

qqnorm(sprit, main = "Spritkonsumtion", xlab = " ", ylab = " ")
qqline(sprit)
```



Figur 12: Data från öl, vin och sprit plottat som histogram och jämförd med den teoretiska normalfördelningen, representerad av linjen.



 Figur 13: Normalfördelningsplottar som visar hur nära den totala konsumtionen för alla länder av antingen öl, vin eller sprit ligger en normalfördelning.

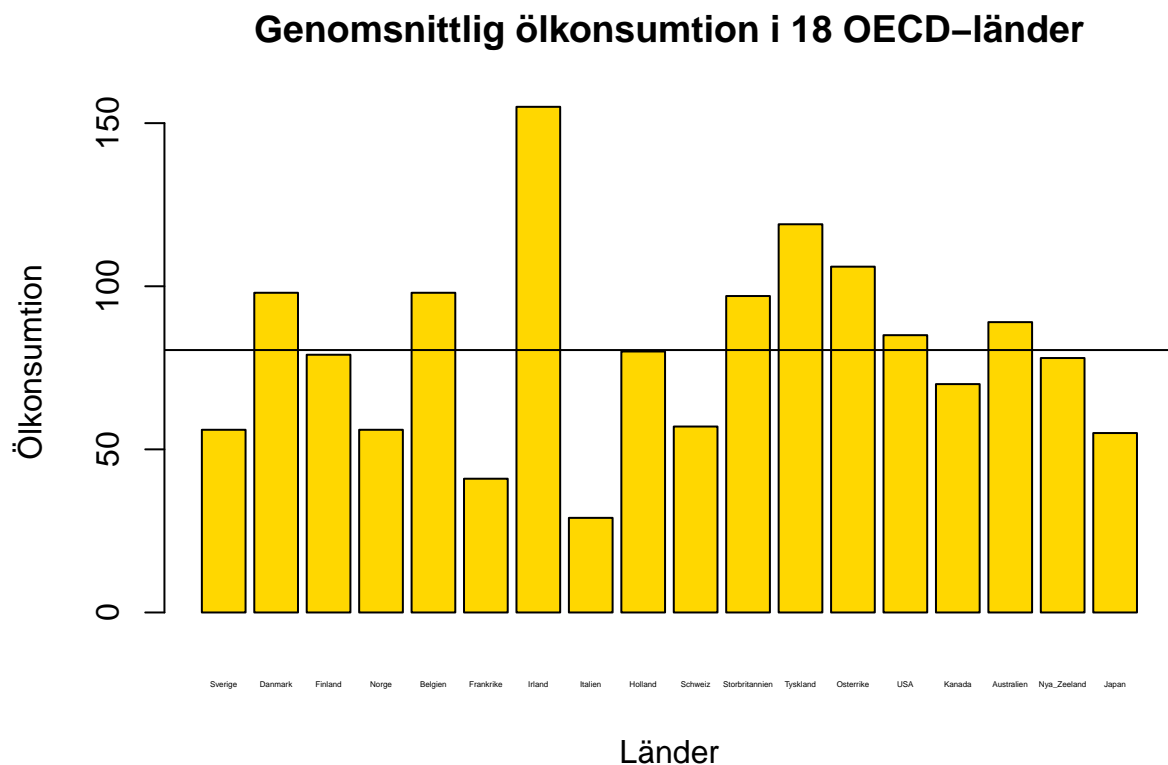
```
par(old_par)
```

Från normalfördelningsplottarna i figur är det svårare att dra någon slutsats men spritkonsumtionen avviker något mer än de andra från räta linjen vilket även detta talar för att den ej är normalfördelad.

För att jämföra alkoholkonsumtionen av öl, vin och sprit mellan länderna kan vi jämföra detta i 3 stolpdia-gram, 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)
```

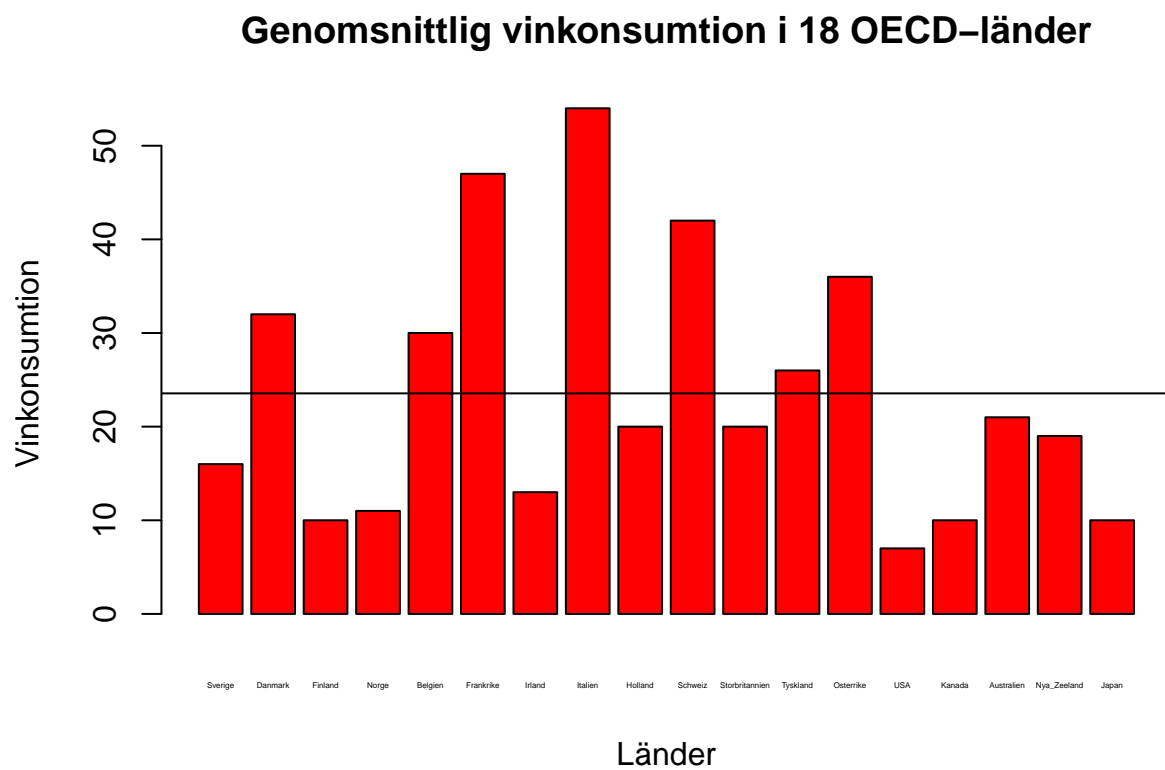
```
abline(mean(beer), 0)
```



Figur 14: Stolpdia-gram med genomsnittlig ölkonsumtion samt en horisontell linje som visar medelkonsum- tionen för alla länder.

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

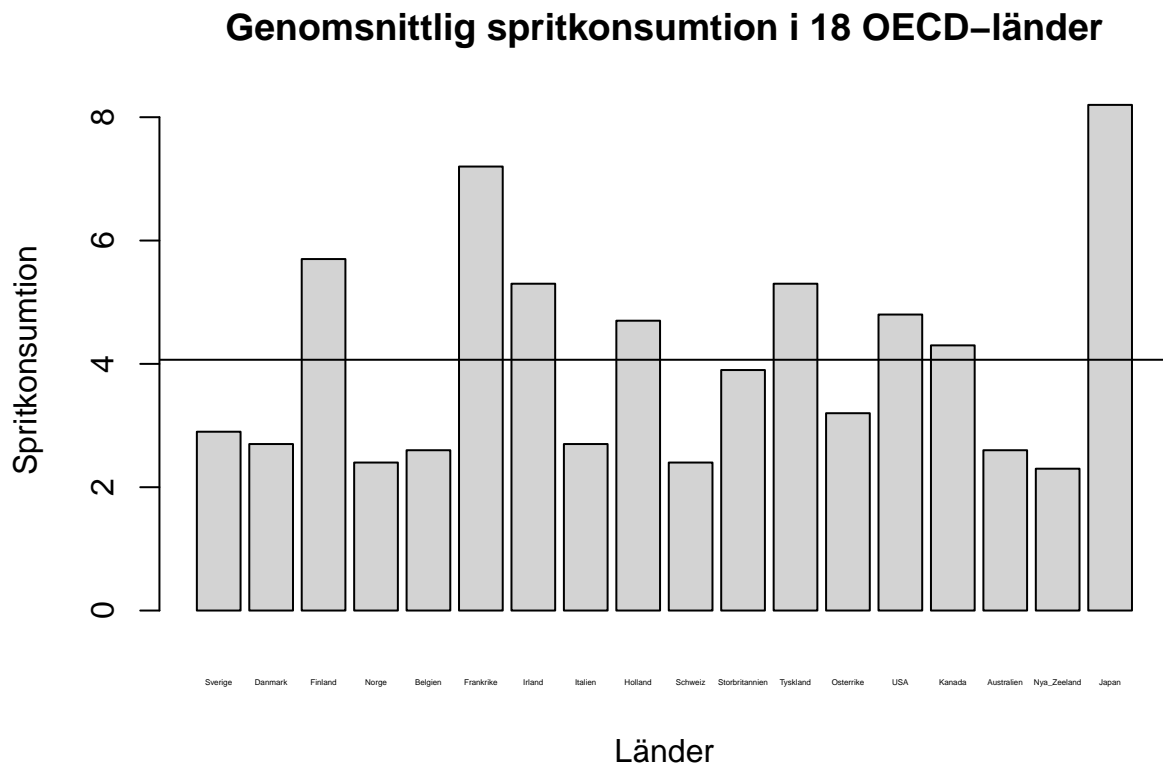
```
abline(mean(vin), 0)
```



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

```
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: Stolpldiagram med genomsnittlig spritkonsumtion samt en horisontell linje som visar medelkonsumtionen för alla länder.

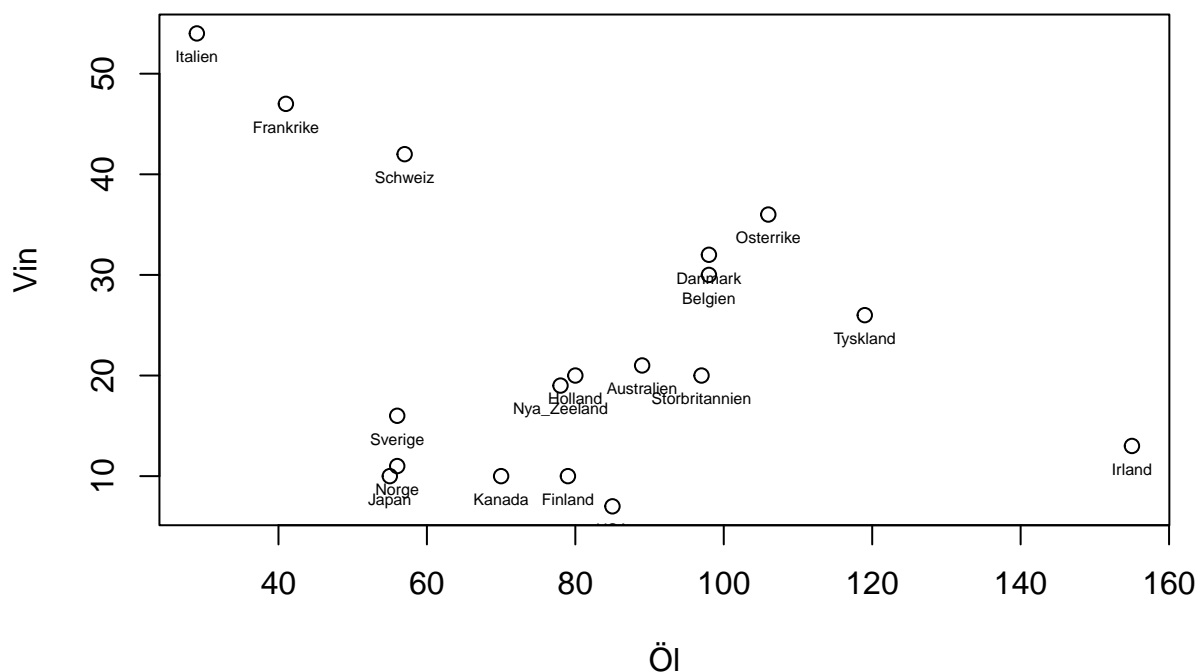
Vi ser att Sverige verkligen inte avviker i någon av öl, vin- och spritkonsumtion. Vi är landet lagom.

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

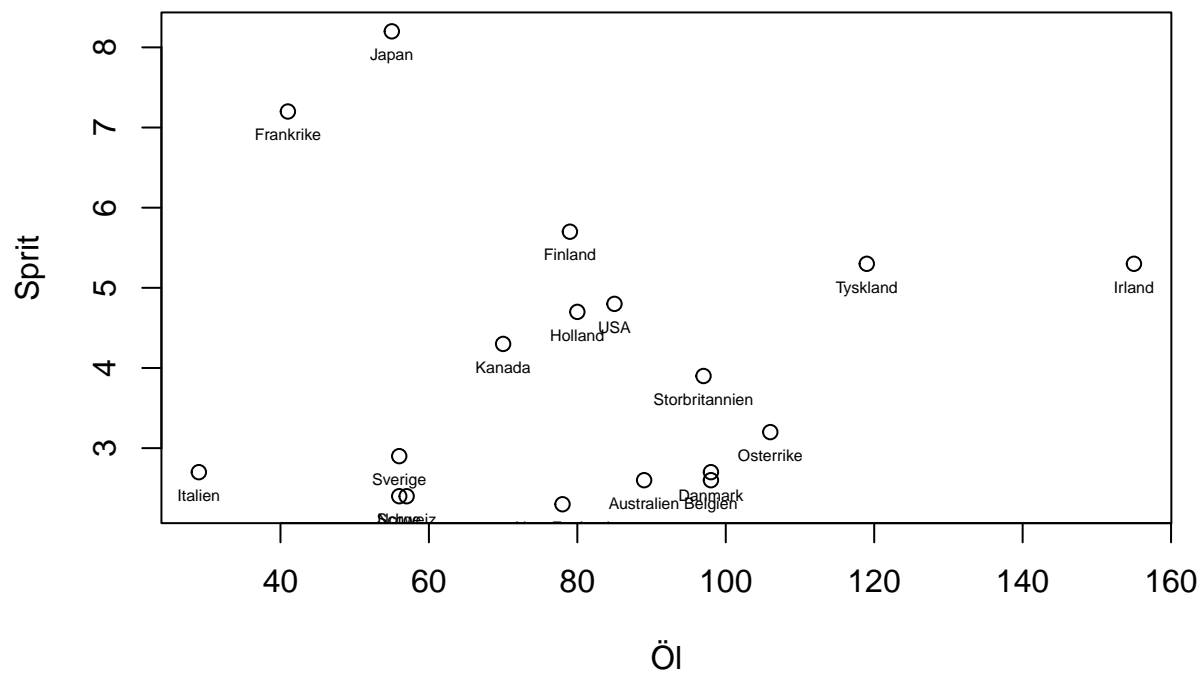
Ett någorlunda gemensamt drag hos de extrema länderna i öl- och vinkonsumtion är att de tenderar att endast ligga högt i den ena av dessa. Detta kan klart ses i figur (första textplot) där de som ligger allra högst i antingen öl- eller vinkonsumtion ligger bland de lägsta i den andra.

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.

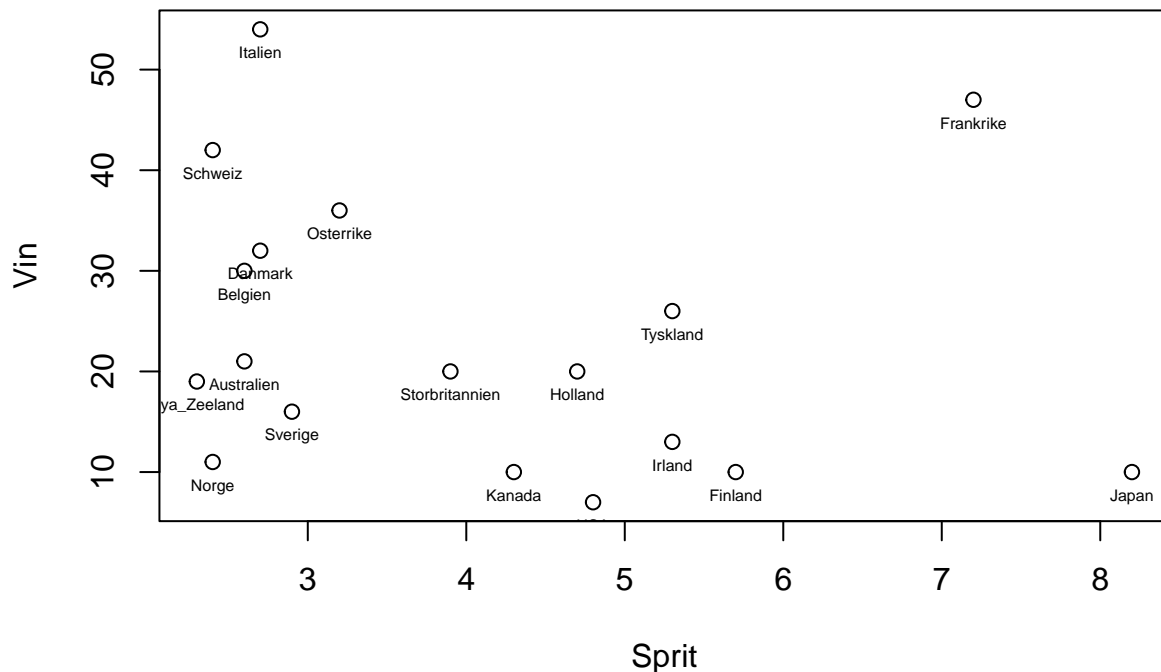
```
plot(beer, vin, xlab = "Öl", ylab = "Vin")
text(beer, vin, land, cex = 0.5, pos = 1)
```



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



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

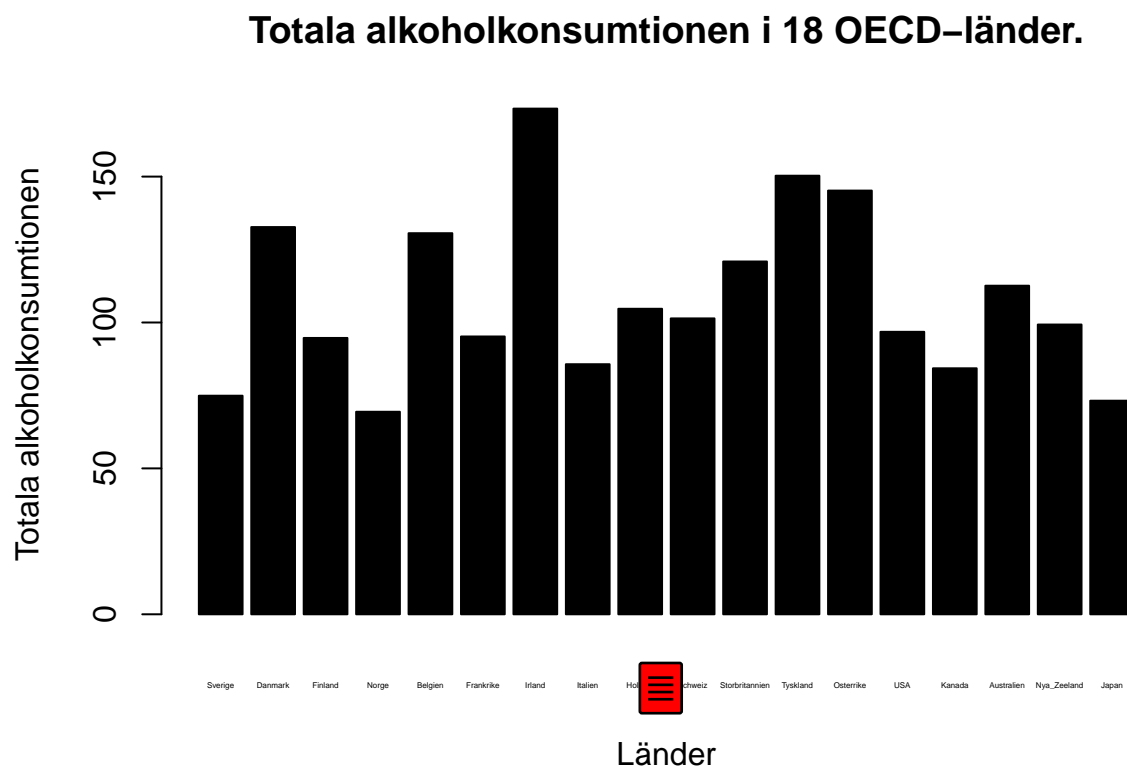
```
total_alk_kons = beer + vin + sprit # Skapa en vektor med totala alkoholkonsumtionen.
```

Vi kan avsluta med att kolla på den totala alkoholkonsumtionen i länderna. Detta kan illustreras på ett smidigt sätt med ett stolpdigram, figur 2. Här ser vi att Norge kniper förstaplatsen om man får uttrycka det så, 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.

Tycker ändå överlag att konsumtionen ligger ganska jämt fördelad över länderna, som tidigare nämndes är enbart Tyskland över medel för samtliga alkoholtyper men inte högst i någon av dem. Irland uppfyller stereotypen som öldrickarnation likaså Frankrike och Italien som vinkonsumerare. Spritkonsumtionen i Japan var något oväntad, men all in all ganska väntade resultat.

```
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 17: Totala alkoholkonsumtionen i 18 OECD-länder.