

Laboration 1

Förnamn Efternamn

2024-11-23

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 $n = 40, 100, 30$, för normalfördelade, likformigt fördelade och exponentielfördelade stickprov, respektive. 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 utforskar 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 (1) är att 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 (2) är att 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 exponentielfördelade 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 vektorer enligt instruktionerna

```
set.seed(20010310)
normalstickprov1 <- rnorm(40, 24, 24) # antal simulerade 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)
```

För att visualisera stickprovet plottar vi datan med hjälp av R-funktionen `hist` och jämför den med grafen av 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)
x <- seq(from = -72, to = 96, length.out = 100) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
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) # Väljer övre och undre -gränser för att täcka hela sti
lines(x, dnorm(x, 24, 24))
mtext("Histogram över normalfördelade stickprov (n=40)", outer = TRUE, cex = 1.2, font = 2)

```

Histogram över normalfördelade stickprov (n=40)

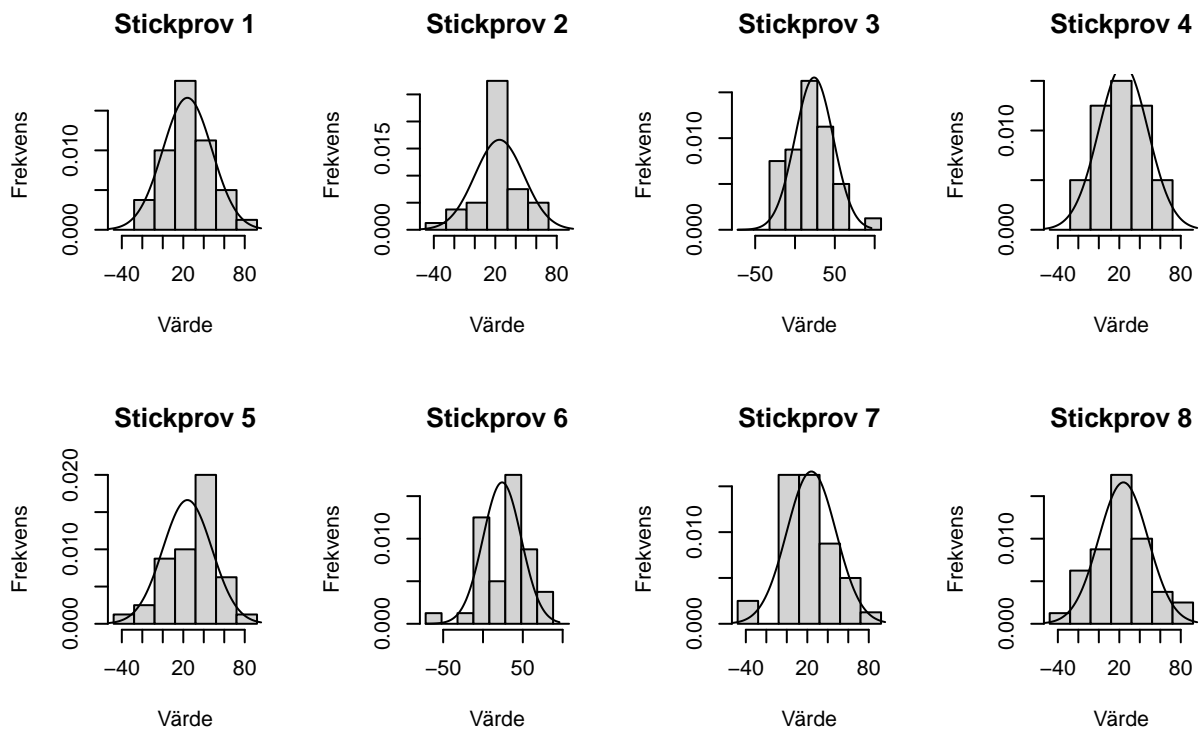


Diagram 1: Histogram

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.

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

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

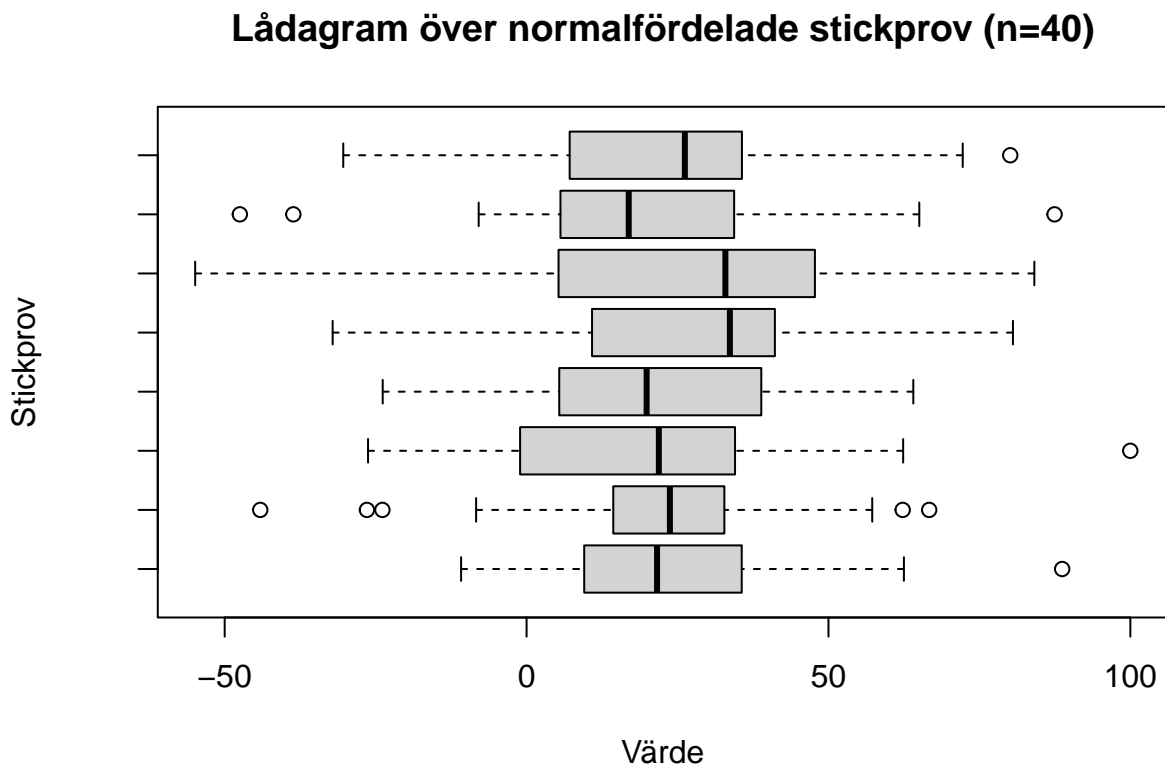


Diagram 2: Lådagram av Normalfördelad data

Om vi tittar på lådagrammen i figur 2, så ser asymmetrin större ut. Tydligast 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.

Till slut vänder vi oss till normalfördelningsplott, 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))
```

```
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

```

I figur 3 noterar vi att i de flesta stickproven, så ligger de allra flesta punkter på en rät linje. De 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.

Normalfördelningsplottar över normalfördelade stickprov (n = 40)

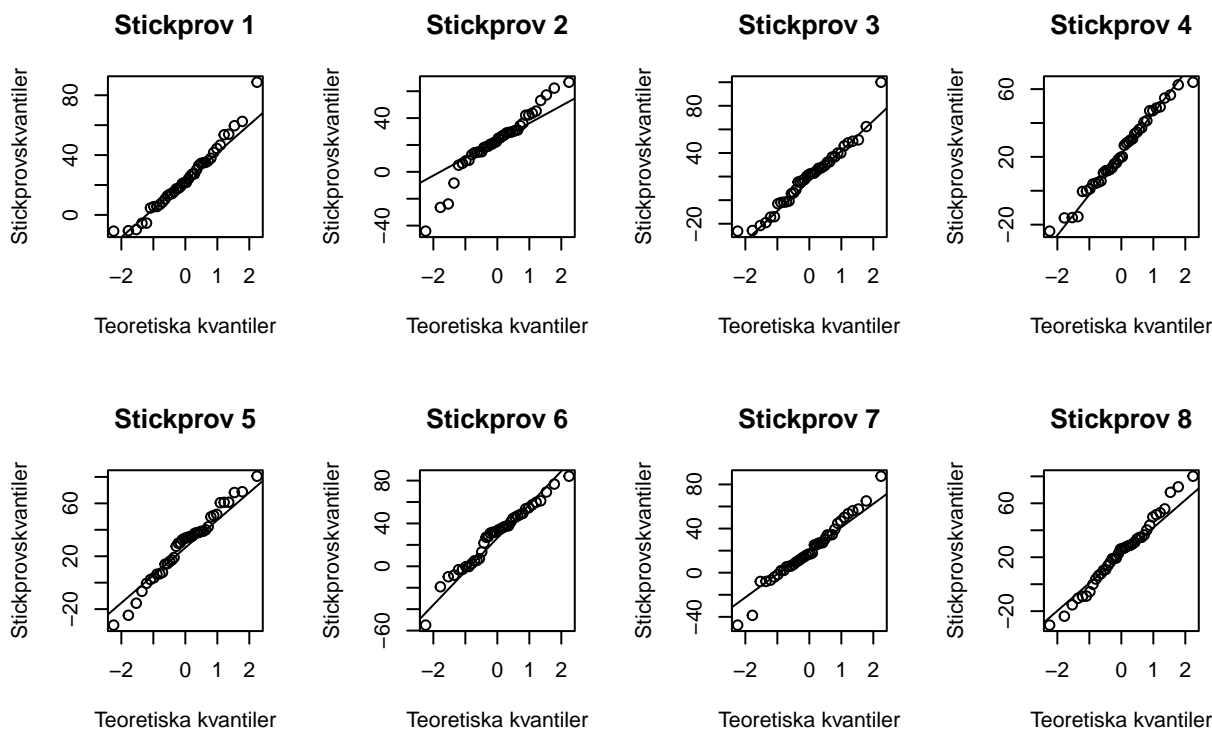


Diagram 3: Normalfördelningsplott av normalfördelad data

Likformigt fördelade data

För likformigt fördelade data gör vi samma sak. Med hjälp av vår teoretiska uppgift så sätter vi gränspunkterna för våra stickprov enligt instruktionerna så att väntevärdet och standardavvikelse blir 24. Annars gör vi allt precis som tidigare.

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

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 = 20),
     xlab = "Värde",
     ylab = "Frekvens",
     main = "Stickprov 1",
     prob = TRUE)
x <- seq(from = -72, to = 96, length.out = 100) # Väljer övre och undre -gränser för att täcka hela sti
lines(x, dnorm(x, 24, 24)) # Grafen för Täthetsfunktionen av en normalfördelad slumpvariabel med väntev
```

```

hist(likformstickprov2,
     breaks = seq(from = -48, to = 96, by = 20),
     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 = 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(likformstickprov4,
     breaks = seq(from = -48, to = 96, 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(likformstickprov5,
     breaks = seq(from = -48, to = 96, by = 20),
     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", outer = TRUE, cex = 1.2, font = 2)

```

Histogram över likformigt fördelade stickprov

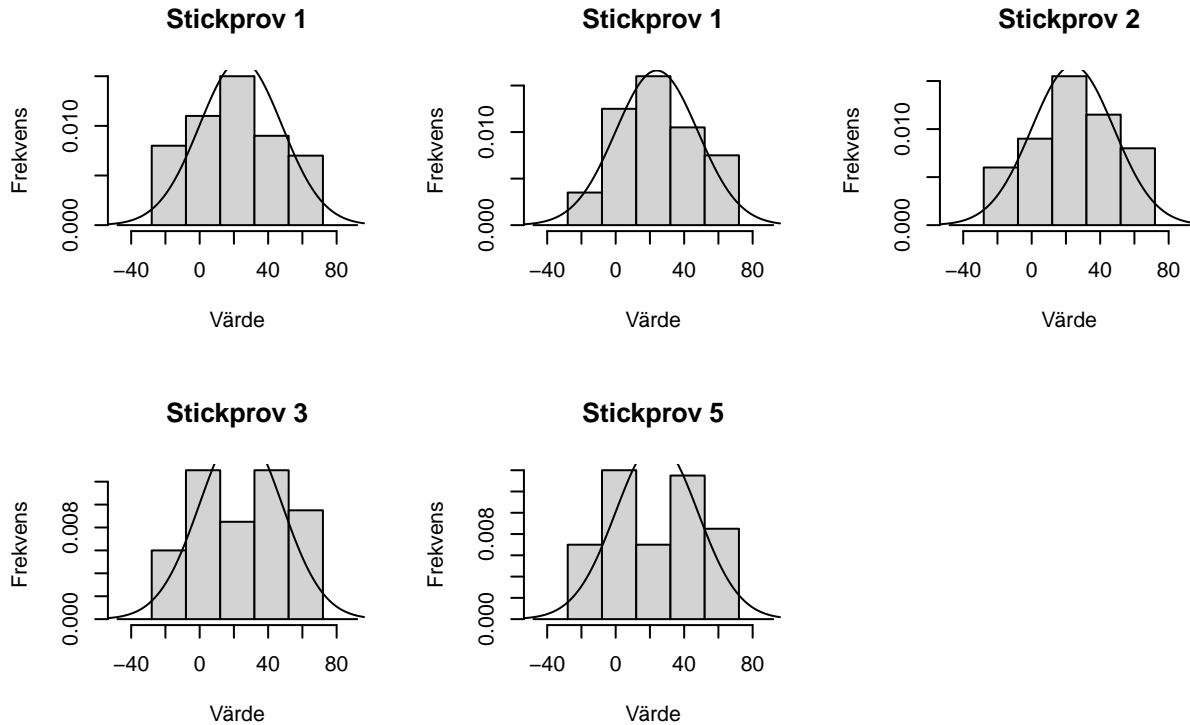


Diagram 4: Histogram av likformigt fördelad data

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 karaktäristiskt av en normalfördelning. Vidare är svansarna inte långa, 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.

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",
        xlab = "Värde",
        ylab = "Stickprov"
)
```

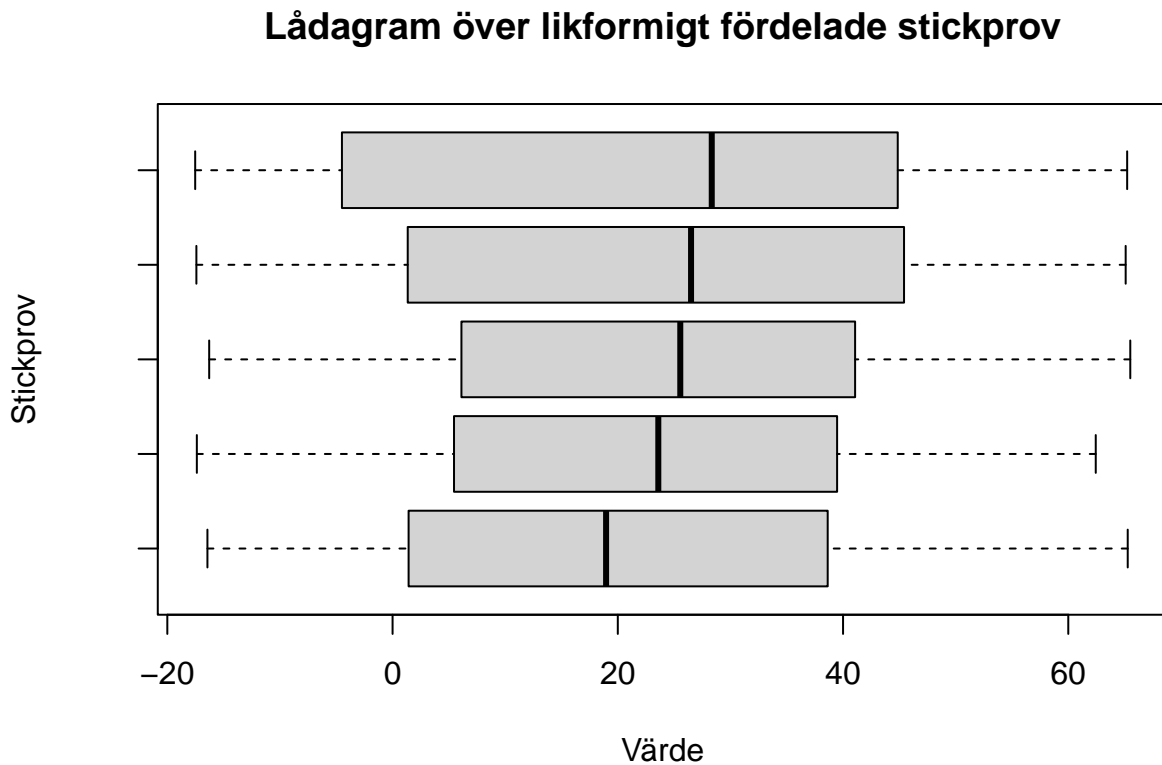


Diagram 5: Lådagram för likformigt fördelad data

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 datan 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.

```
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)

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", outer = TRUE, cex = 1.2, font = 2)

```

Normalfördelningsplottar över likformigt fördelade stickprov

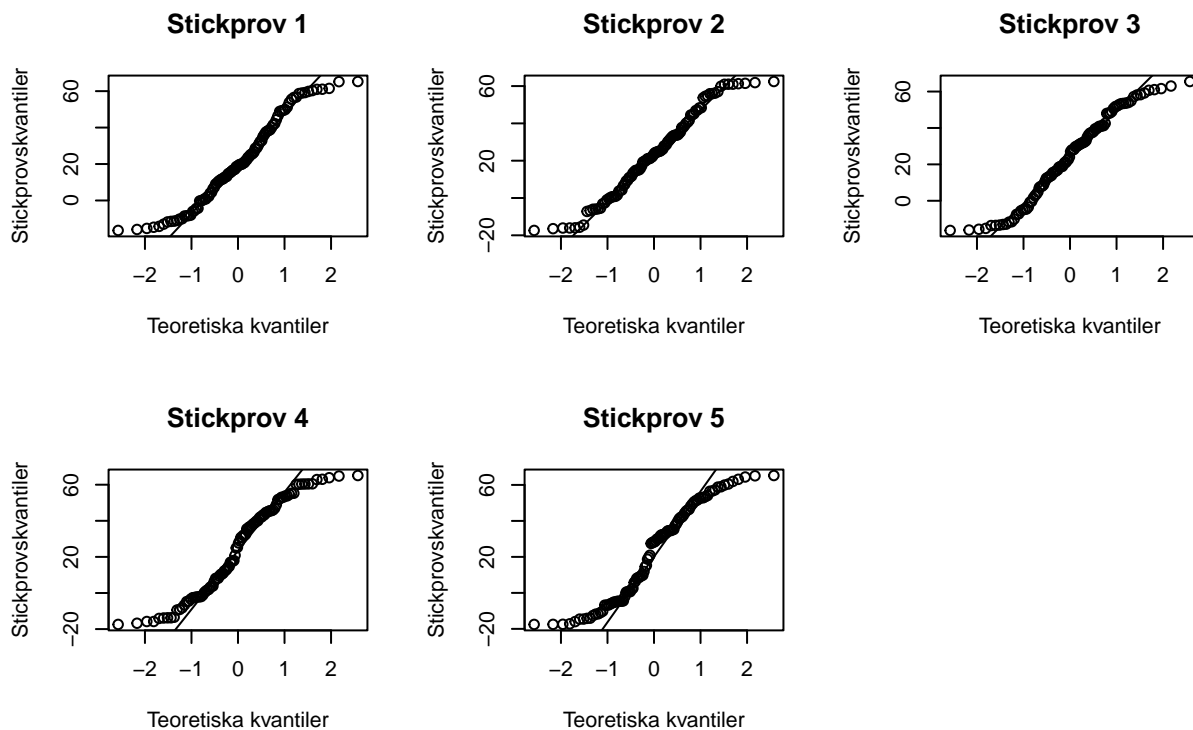


Diagram 6: Normalfördelningsplot av likformigt fördelad data

I figur 6 ser vi tydligast att detta inte kan vara en normalfördelning. Alla stickprov följer en slags "S" form, och är lika täta genom alla kvantiler, vilket inte ska ske för normalfördelade data.

Exponentialfördelade data

Nu undersöker vi våra exponentialfördelade stickprov!

```
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)
```

Vi visualiserar igen med histogram:

```
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)
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", outer = TRUE, cex = 1.2, font = 2)

```

Histogram över exponentialfördelade stickprov

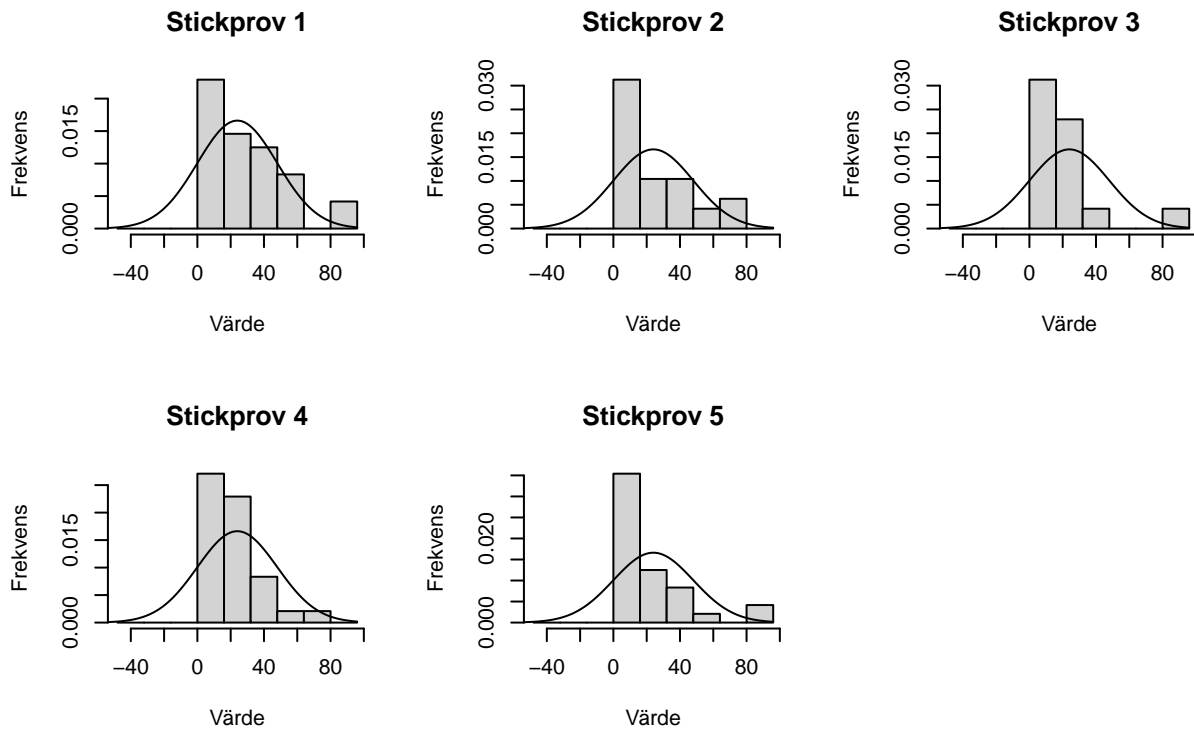


Diagram 7: Histogram

I figur 7 är det väldigt enkelt att se, trots storleken på stickprovet, att datan ej är normalfördelad. Det som ger bort det är (1.), den största stapeln ligger aldrig på värdet 24, dvs det teoretiska väntevärdet. Och (2.) faktumet att inga staplar ligger till vänster om värdet 0.

Nu undersöker vi lådagrammen för våra stickprov.

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

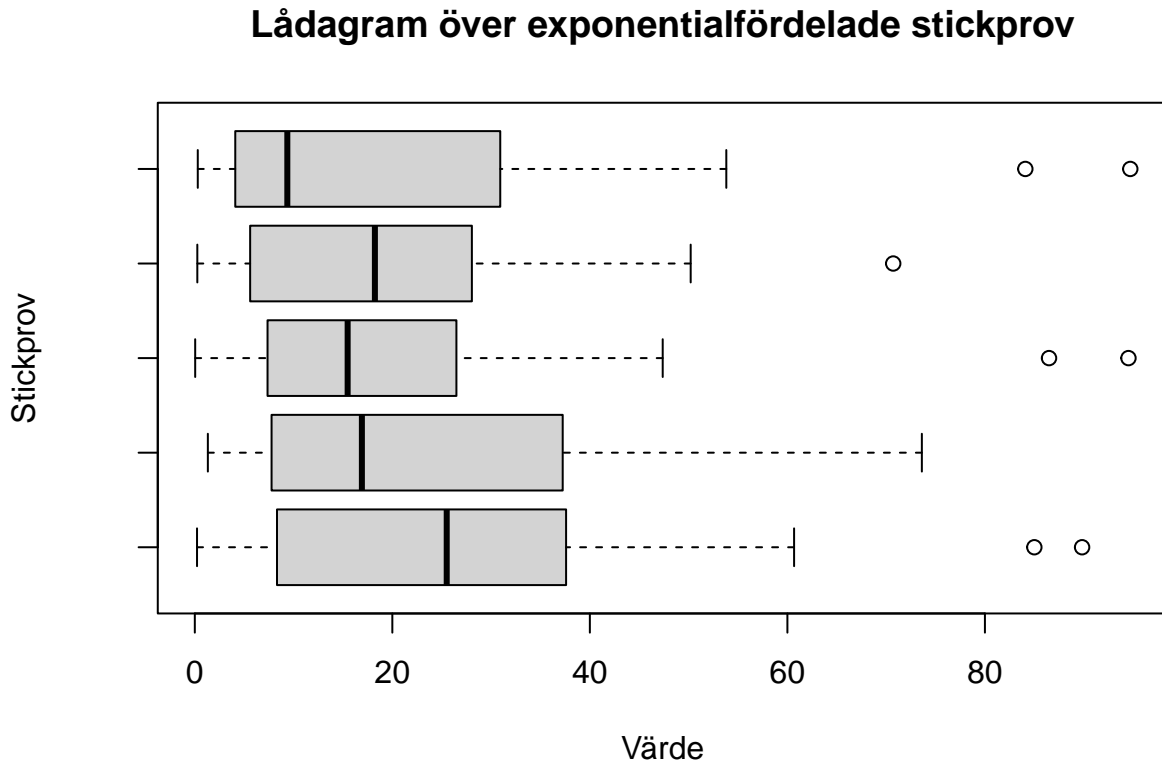


Diagram 8: Lådagram av exponentialfördelade stickprov

I figur 8 är det också extremt tydligt att stickproven inte är normalfördelade. Morrhåren är inte alls lika långa, och vi har många avvikande datapunkter långt ut till höger, medans vi har inga alls till vänster. Medianen är också långt ifrån det teoretiska värdet 24 i många av stickproven.

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

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

qqnorm(expstickprov1,
        xlab = "Teoretiska kvantiler",
        ylab = "Stickprovskvantiler",
        main = "Stickprov 1")
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", outer = TRUE, cex = 1.2, font = 2)

```

Normalfördelningsplottar över exponentialfördelade stickprov

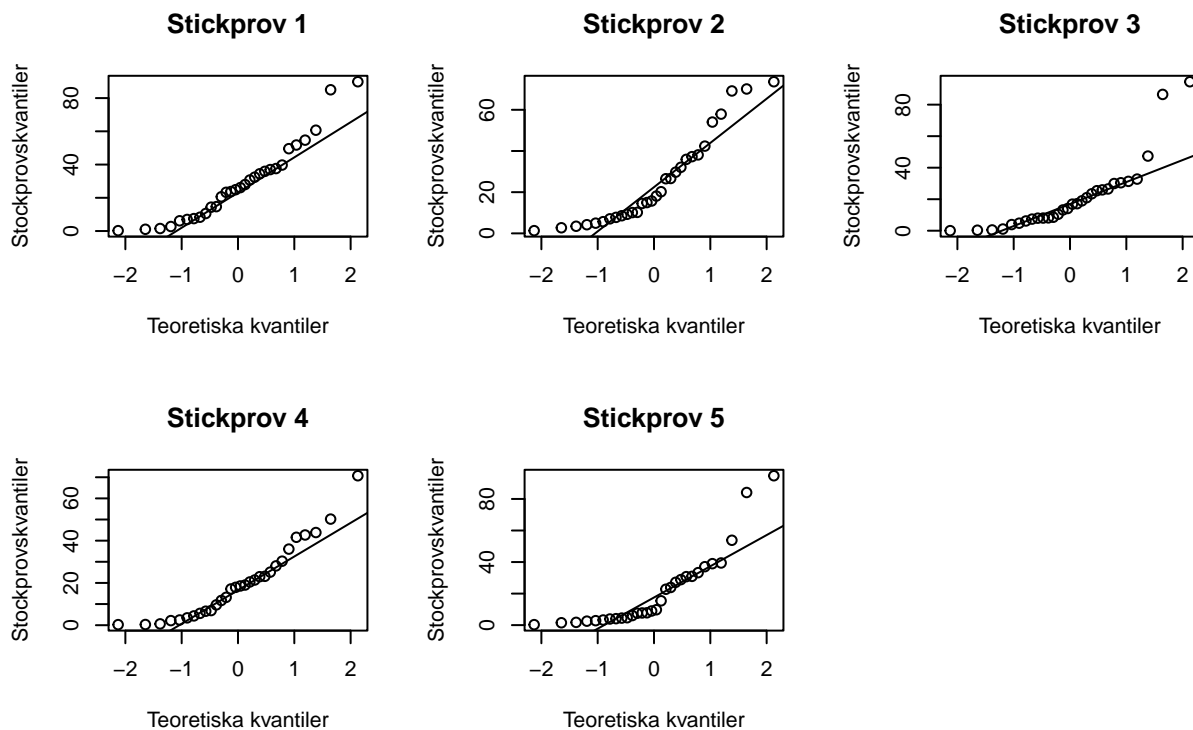


Diagram 9: Normalfördelningsplot av exponentialfördelade stickprov.

I figur 9 ser vi att punkterna inte ens sitter på en rät linje! Och igen, avvikande punkterna är mycket mer extrema åt ett håll än det andra. Det känns tydligt att våra stickprov inte är normalfördelade.

Disukssion och slutsats

Vi tyckte att för relativt låga värden på n för normalfördelade och exponentialfördelade 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 låga värden på n . Detta tror vi har att göra med faktumet att likformig data ändå delar några egenskaper med normalfördelad sådan. Som att t.ex. datan är symmetriskt fördelad längst medelvärde. Men ökar man antalet observationer i stickprovet får man en klarare bild, särskilt av normalfördelningsplotten (figur 9), som visar tydligt hur stickprovskvantilerna inte alls svarar mot de hos en normalfördelning.

För normalfördelade stickproven kändes det som att histogrammen gjorde ett bra jobb av att visa hur normalfördelad datan 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 dem 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 tycker vi nog att normalfördelningsplotten är enklast att avläsa, och ger en tydlig bild på om ett givet stickprov är normalfördelat eller inte.

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)

```

Konsumtion av öl i ett antal OECD-länder

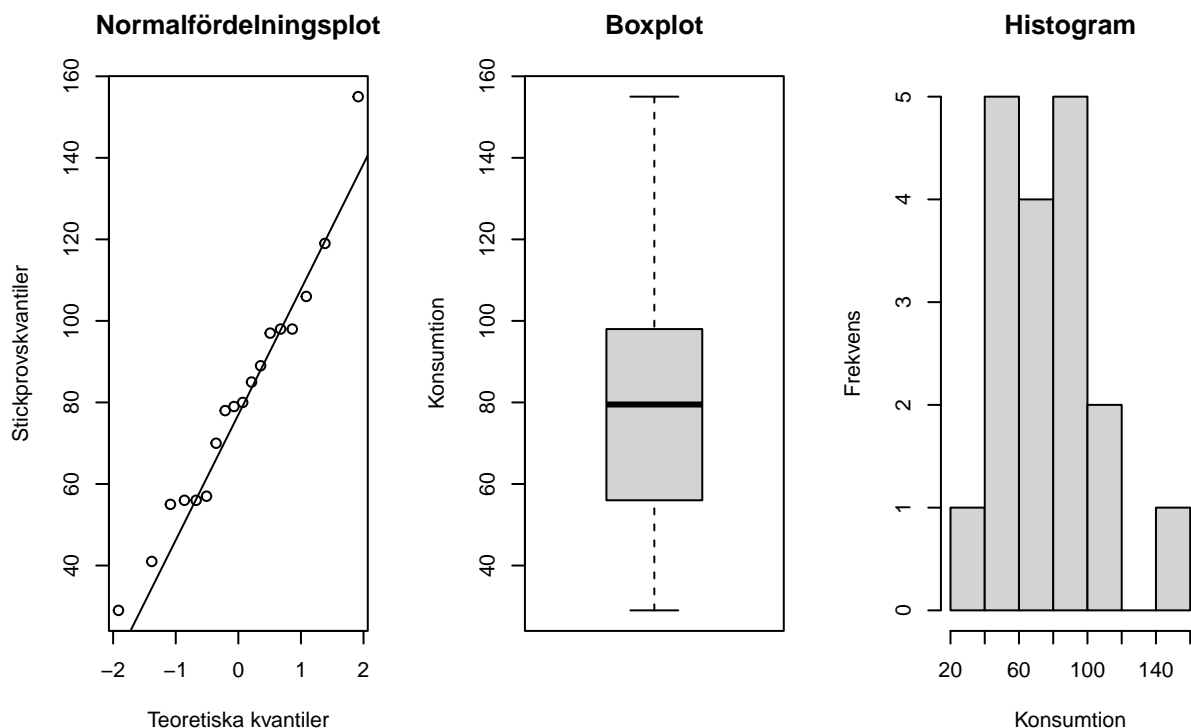


Diagram 10: Ölkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

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 vi ser 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)

```

Konsumtion av vin i ett antal OECD-länder

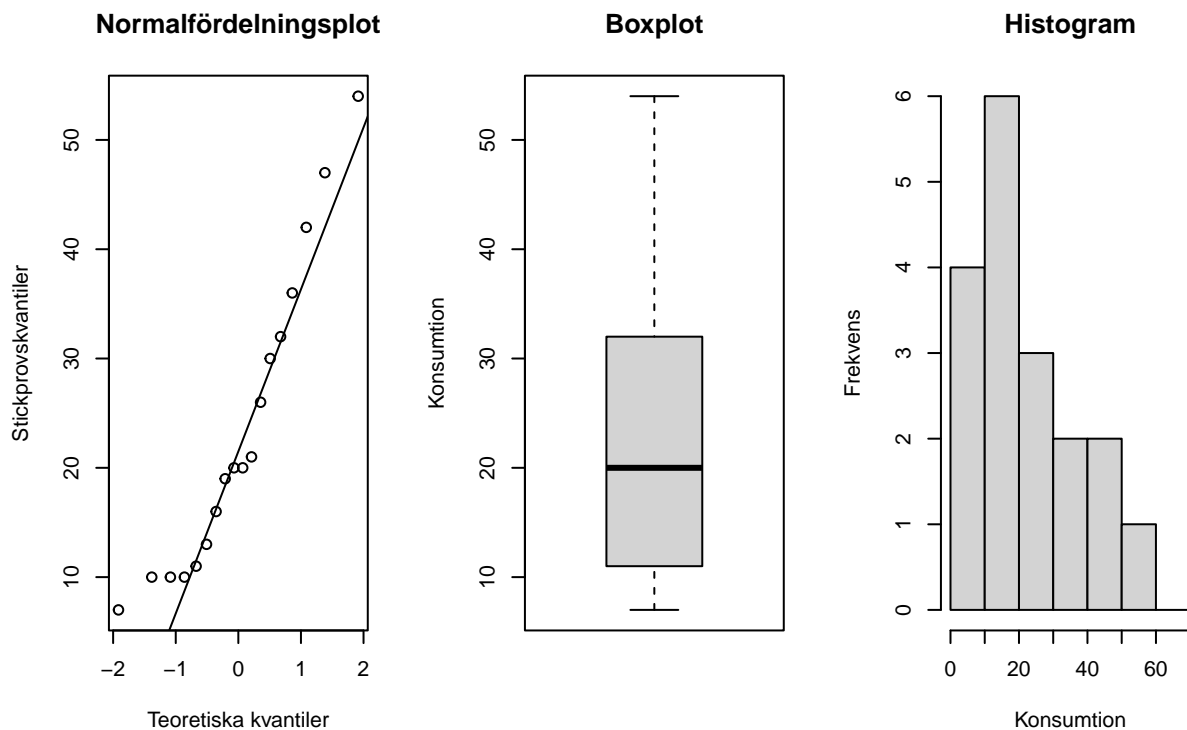


Diagram 11: Vinkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

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)
```

Konsumtion av sprit i ett antal OECD-länder

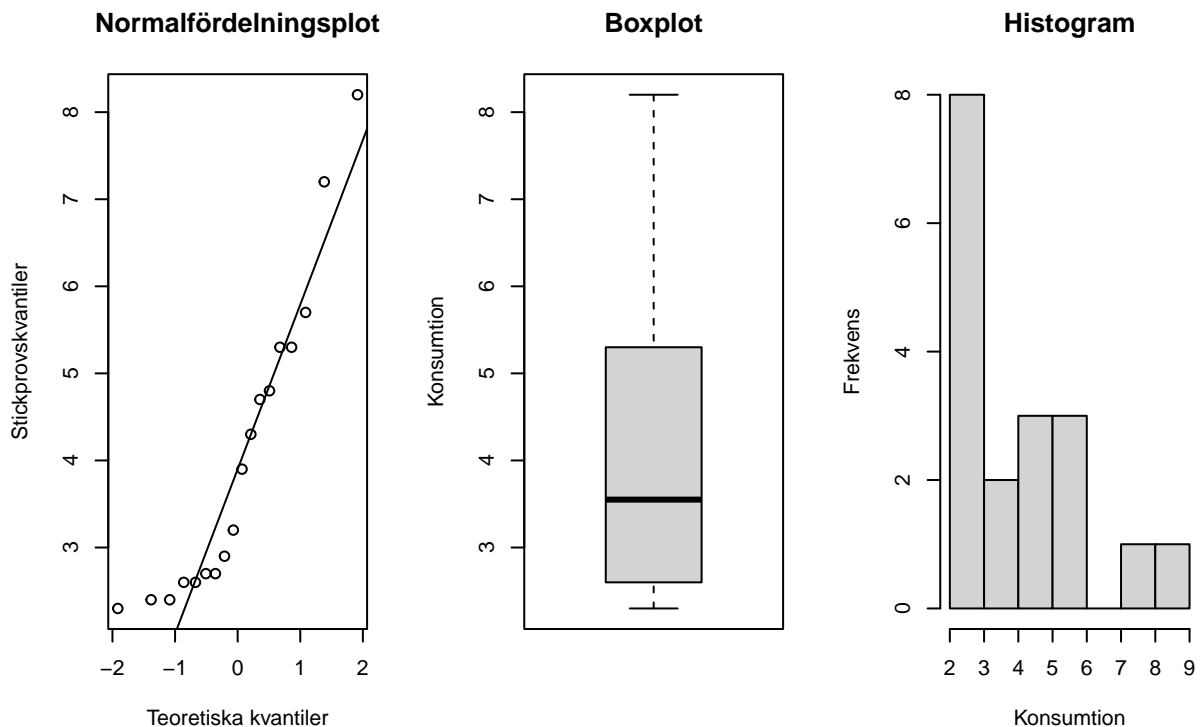


Diagram 12: Spritkonsumtionen i 18 olika OECD-länder, presenterat i tre olika plottar: normalfördelningsplot, boxplot och histogram.

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="Konsumtion av öl")

plot(gg_alkohol)
```

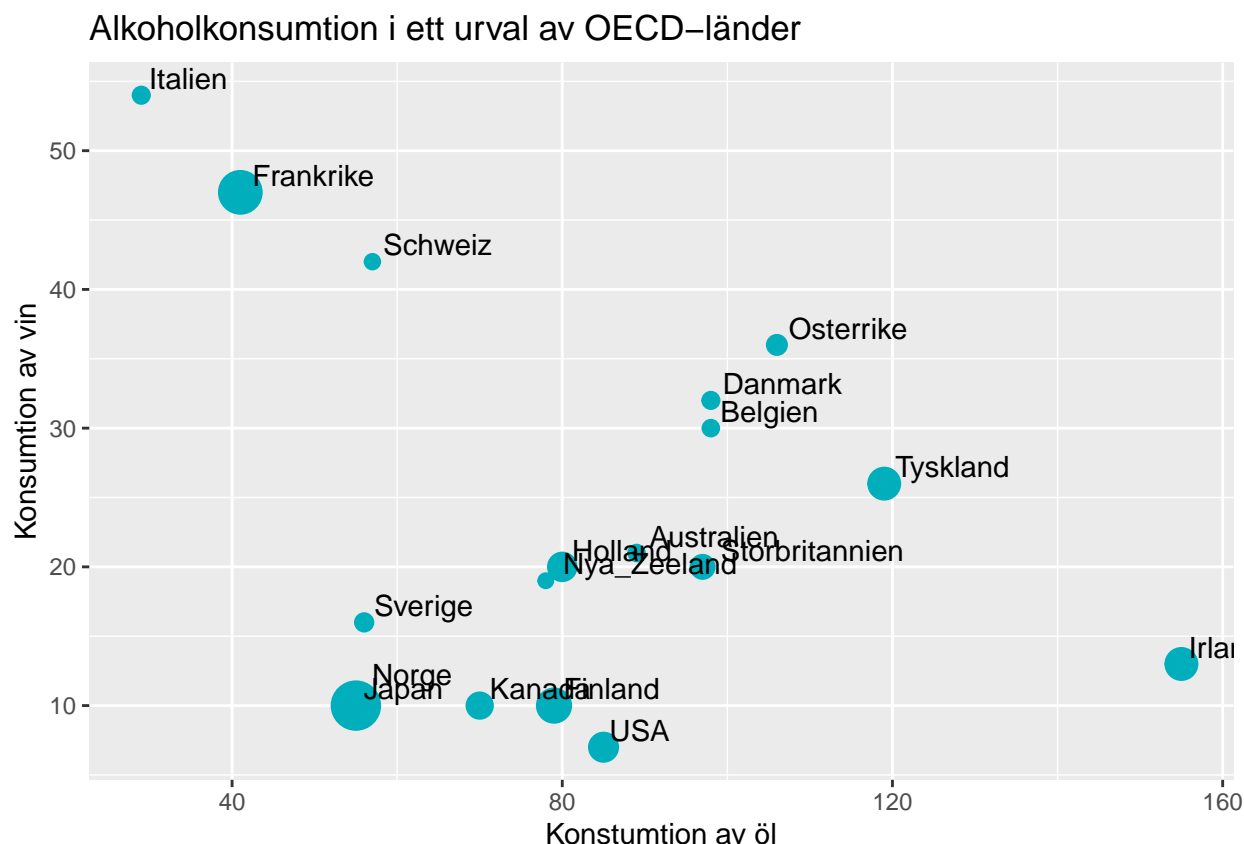


Diagram 13: Alkoholkonsumtionen i 18 OECD-länder, öl på x-axeln och vin på y-axeln. Storleken på datapunkterna indikerar konsumtion av starksprit: Ju större punkt, desto högre konsumtion av starksprit.

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'
```

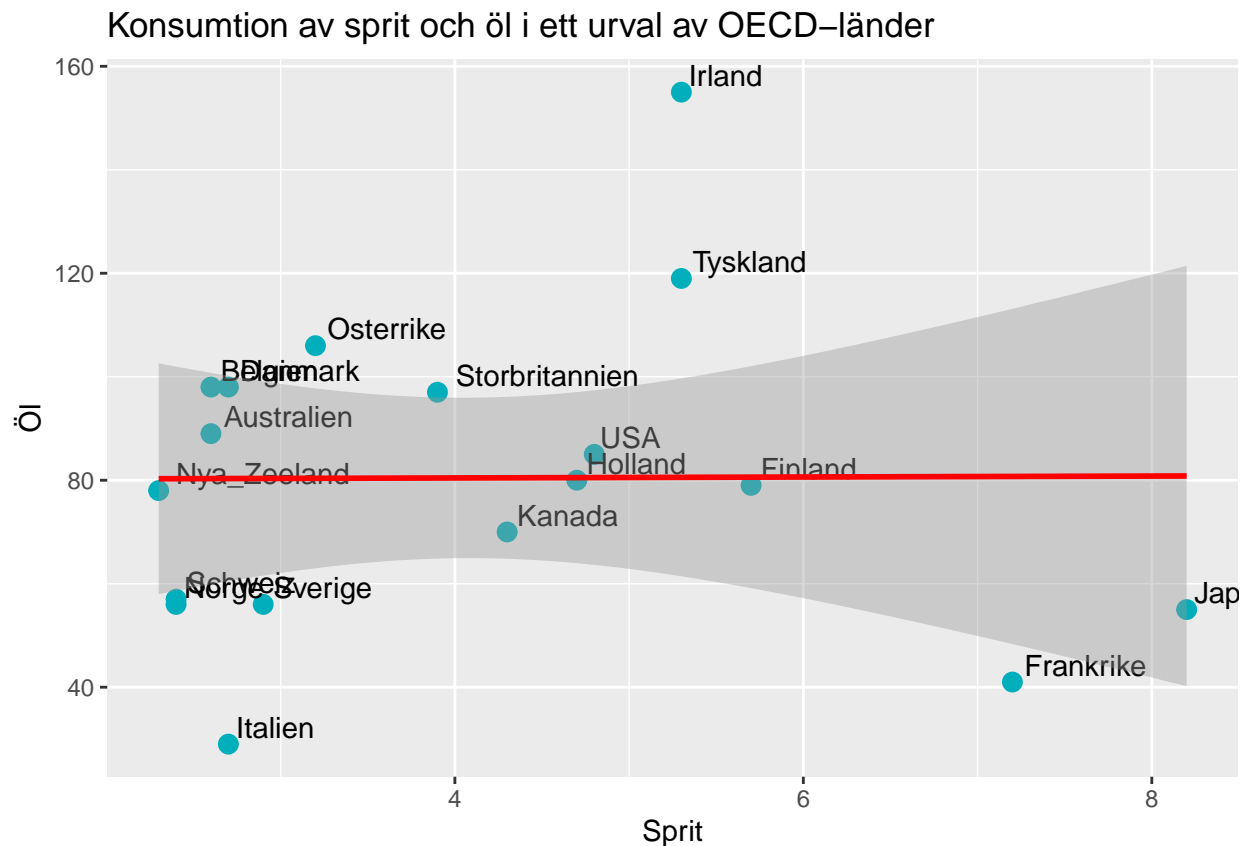


Diagram 14: Sambandet mellan konsumtionen av sprit och vin, i 19 OECD-länder. Sprit på x-axeln och öl på y-axeln. Det skuggade området indikerar konfidensintervall.

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'
```

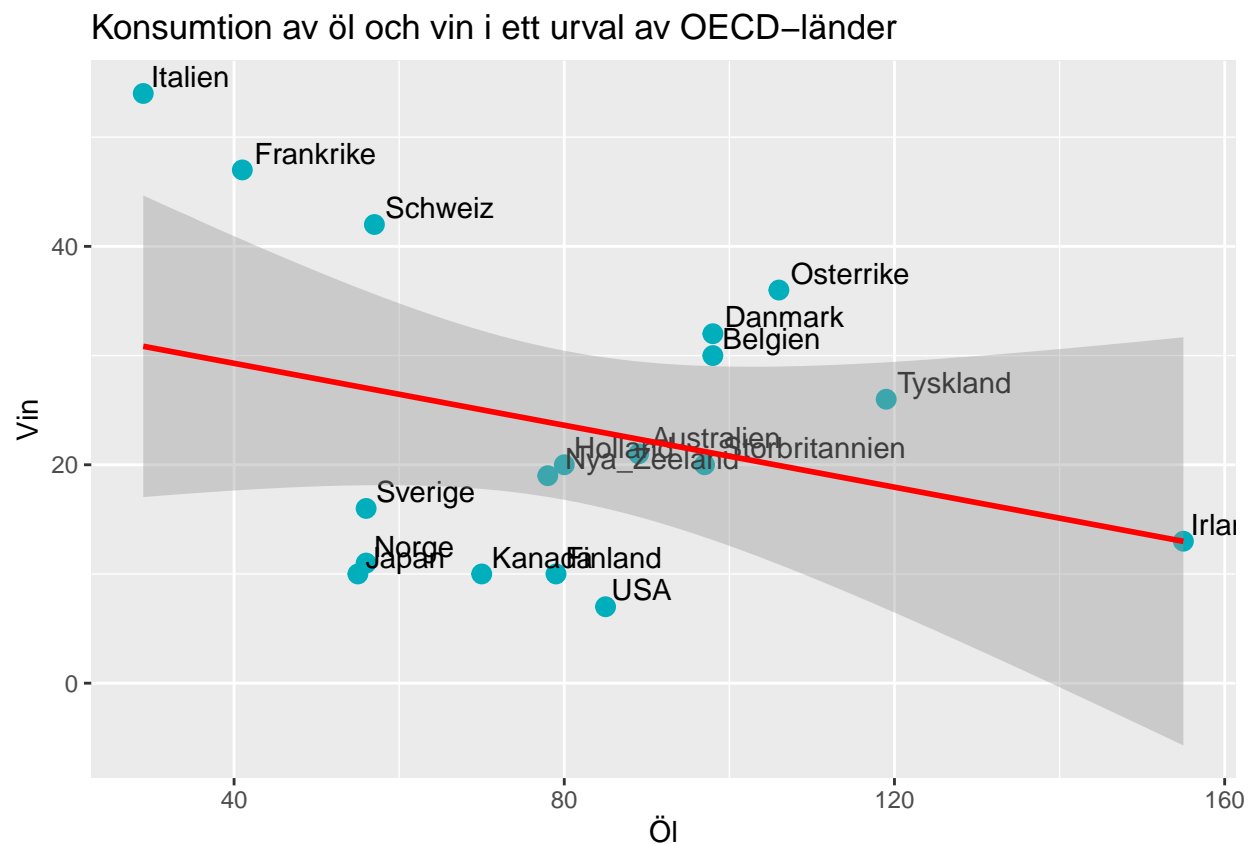


Diagram 15: Sambandet mellan konsumtionen av öl och vin, i 19 OECD-länder. Öl på x-axeln och vin på y-axeln. Det skuggade området indikerar konfidensintervall.

```

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

Konsumtion av sprit och vin i ett urval av OECD-länder

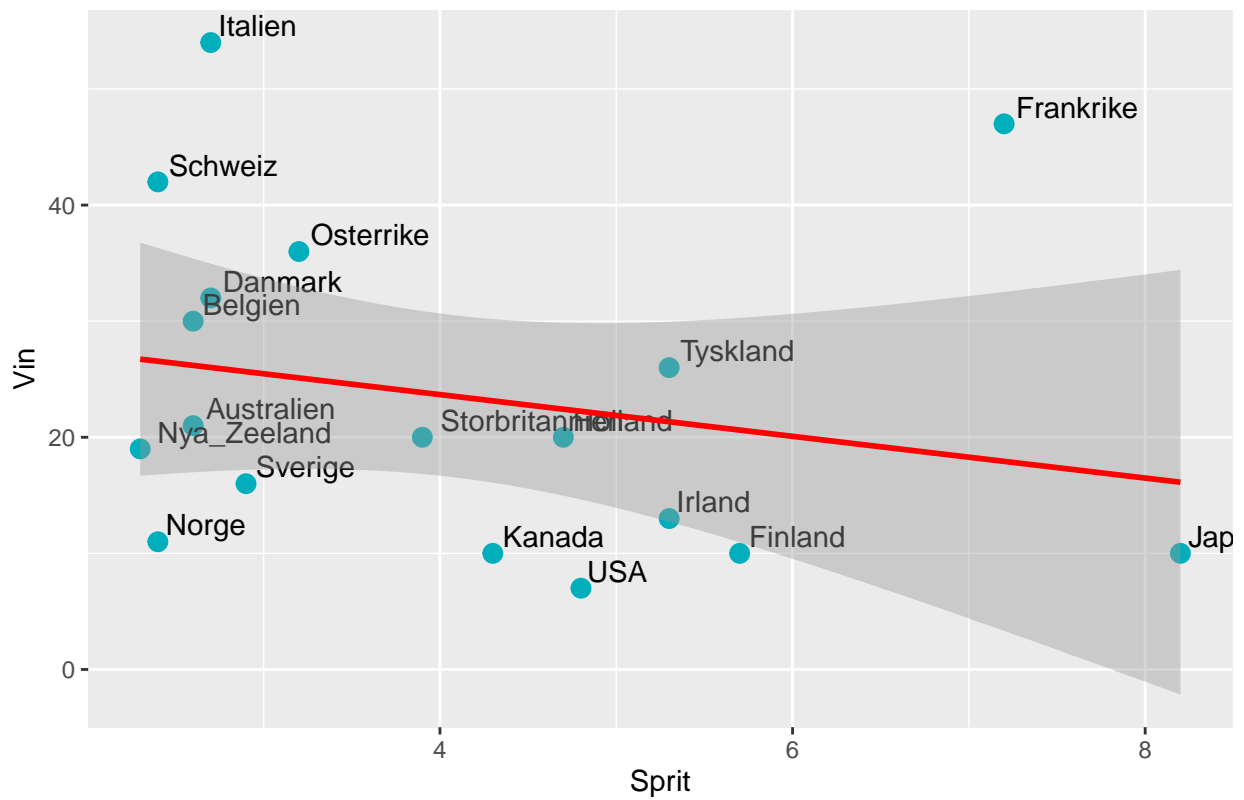


Diagram 16: Sambandet mellan konsumtionen av öl och vin, i 19 OECD-länder. Öl på x-axeln och vin på y-axeln. Det skuggade området indikerar konfidensintervall.

(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 mde 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.