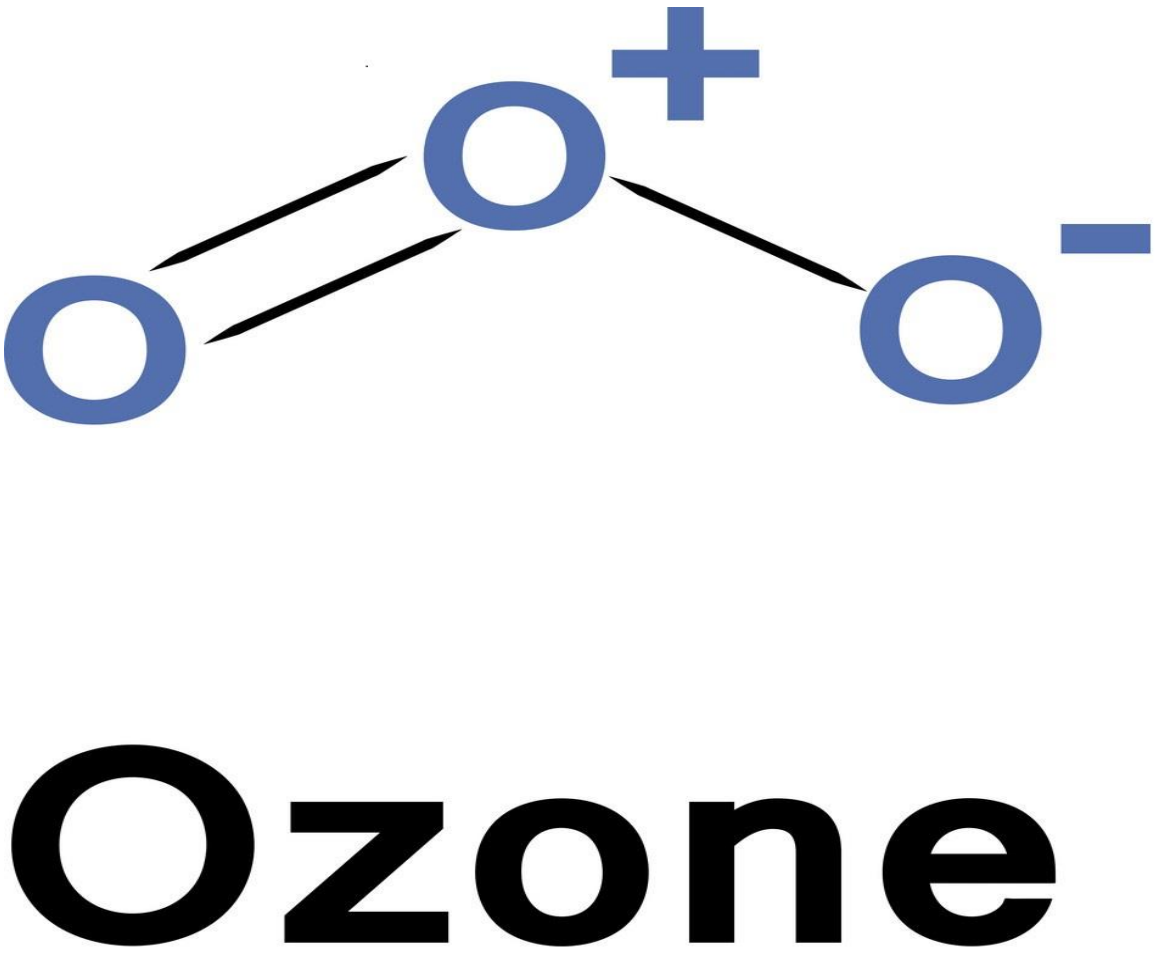
Oblig 2

Av Furkan Kaya

I STK1000: Innføring i anvendt statistikk

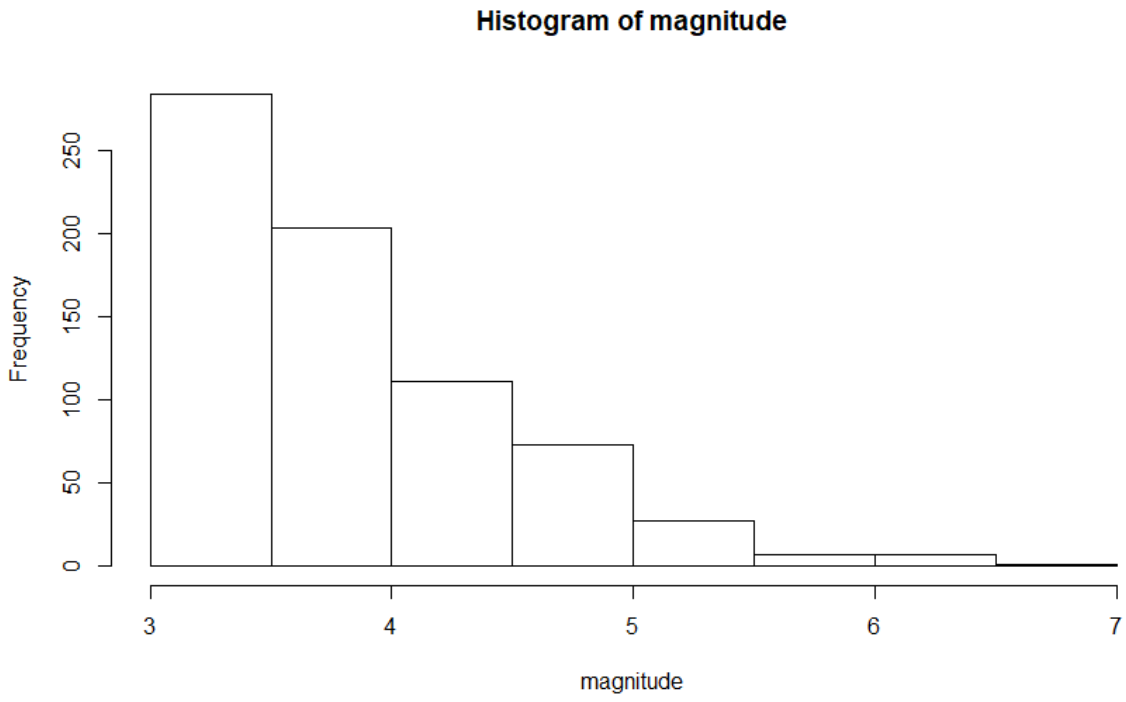


Figur 1: viser molekylet ozon som består av tre oksygenatomer med en dobbeltbinding

Oppgave 1:

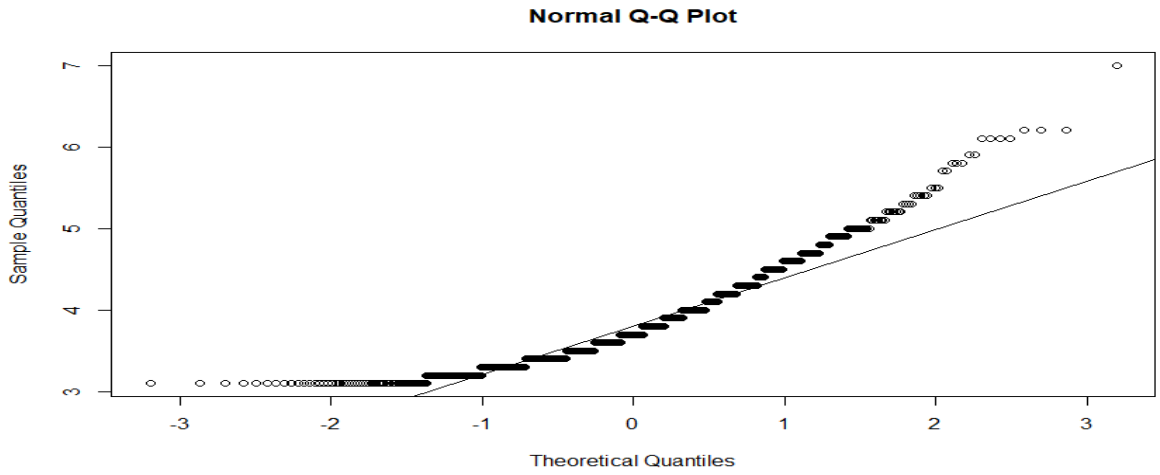
a)

Vi skal i denne oppgaven lage et histogram av styrken på jordskjelvene mellom 1980 og 1982.



Figur 2: et histogram av variabelen "magnitude" i datasettet fra USA

Denne virker på figur 2 å ikke være normalfordelt. For sikkerhets skyld bruker vi også qqline og qqnorm-funksjonen på Rstudio ettersom det er pålagt i oppgaven.



Figur 3: plott som viser qqnorm og qqline funksjonene i Rstudio

Et histogram er normalfordelt om det er symmetrisk rundt en unimodal topp. Vi ser klart at histogrammet på figur 2 er venstre-forskjøvet. Normal-fordelingsplotet indikerer det samme, at verdien er ganske forskjøvet.

```
> data <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/earthq
uakes.txt"
> earthquakes <- read.table(data, header=TRUE)
> magnitude <- earthquakes$magnitude
> hist(magnitude)
> qqnorm(magnitude)
> qqline(magnitude)
```

b)

Vi skal regne ut gjennomsnittet og standardavviket til størrelsen til alle jordskjelvene. Det følger som kode med svar inkludert nedenfor:

```
> mean(magnitude)
[1] 3.874334
> sd(magnitude)
[1] 0.6623718
```

c)

I oppgave c) skal vi trekke ut et utvalg på 50 fra variabelen magnitude og regne ut gjennomsnittet av disse.

```
> sample1 <- sample(magnitude, 50)
> mean(sample1)
[1] 3.834
```

Her velger jeg å legge ut disse 50 verdiene for sikkerhets skyld.

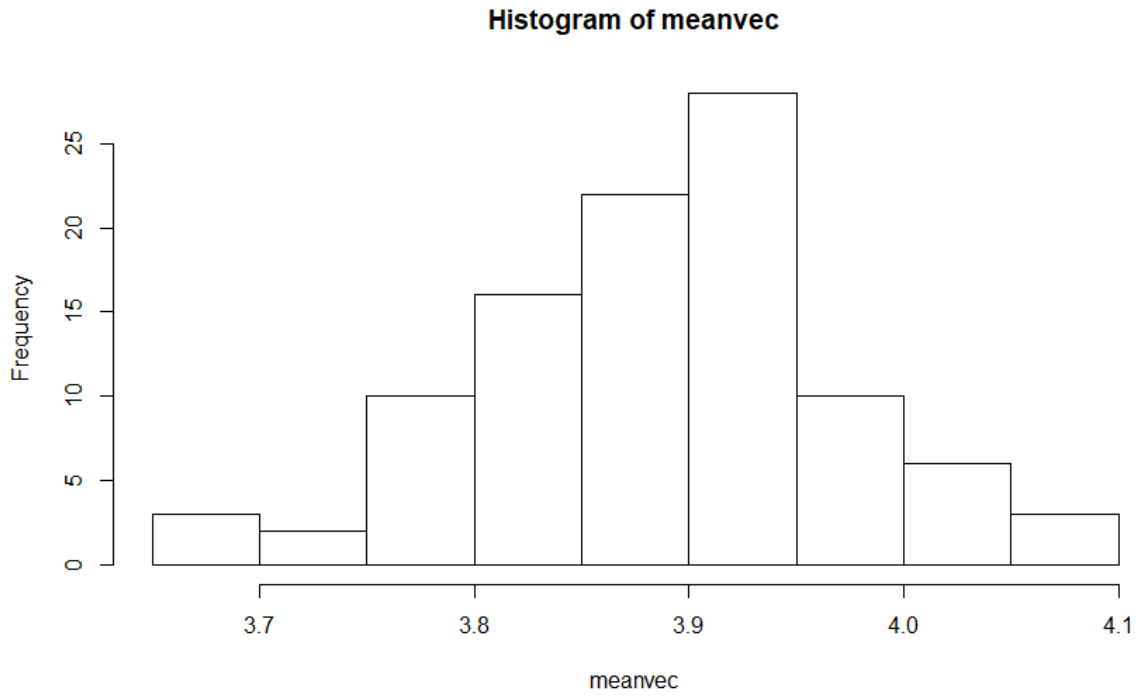
```
> sample1
 [1] 3.4 3.7 4.9 3.7 3.9 4.9 3.1 4.1 3.4 3.9 5.8 3.1 4.5 3.2 3.4 3.4 3.7 4.
8 3.5 3.2 4.2 3.3 3.6 3.7 3.9 3.2 3.2 3.9 3.4
[30] 3.4 6.1 3.3 4.0 4.4 3.6 4.4 4.1 3.5 3.8 3.4 4.1 4.0 3.7 3.6 3.3 3.2 3.
5 4.5 4.3 3.5
```

d)

Vi bruker en oppgitt R-kode til å finne hundre utvalg og så finne et gjennom snitt av dem. Deretter skal vi lage et histogram av dem.

```
> meanvec <- rep(0, 100)
> for(i in 1:100) {
```

```
+     sample.now <- sample(magnitude, 50)
+     meanvec[i] <- mean(sample.now)
+ }
> hist(meanvec)
```



Figur 4: viser et histogram som viser de hundre utvalg og gjennomsnittet av dem

Fordelingen er mer normalfordelt enn den i oppgave a, men er fortsatt forskjøvet. De viser at et sannsynlig gjennomsnitt er mellom 3.8 og 3.95. Og bekrefter svaret vi fikk i oppgave 1 b).

e)

Vi skal her finne μ_x og σ_x for både teoretisk og empirisk verdi. Jeg velger å begynne med teoretisk først, før jeg går videre inn på empirisk verdi.

Teoretisk:

$$\mu_x = \mu = 3.874334$$

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.6623718}{\sqrt{713}} = 0.0248$$

Empirisk:

$$\mu_x = \mu = 3.88812$$

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.08721543}{\sqrt{100}} = 8.7215 * 10^{-3}$$

Ved å sammenligne disse teoretiske verdiene, så ser vi at gjennomsnittene er omtrent de samme. Men det er derimot en forskjell i standardavvik og denne forskjellen er forholdsvis stor til standardavvik å være.

f)

Både oppgave d) og e) skal gjentas men da med henholdsvis 10 og 100 jordskjelv i hvert utvalg fremfor 50 jordskjelv.

For 10 jordskjelv

```
> meanvec1 <- rep(0, 100)
> for(i in 1:100) {
+     sample.now <- sample(magnitude, 10)
+     meanvec1[i] <- mean(sample.now)
+ }
> mean(meanvec1)
[1] 3.8944
> sd(meanvec1)
[1] 0.1980068
```

$$\mu_x = \mu = 3.8944$$

$$\sigma_x = 0.01980068$$

For 100 jordskjelv

```
> meanvec2 <- rep(0, 100)
> for(i in 1:100) {
+     sample.now <- sample(magnitude, 100)
+     meanvec2[i] <- mean(sample.now)
+ }
>
> mean(meanvec2)
[1] 3.87446
> sd(meanvec2)
[1] 0.05698758
```

$$\mu_x = 3.87446$$

$$\sigma_x = 5.698758 * 10^{-3}$$

g)

Sannsynligheten for at man trekker et utvalg med gjennomsnitt større enn 4.5 for n = 10, 50 og 100. Det gir da P(Z>4.5).

For n = 10:

$$Z = \frac{(4.5 - 3.8944)}{0.1980068} = 3.1862$$

Som gir med tabell A i pensumboken,

$$P\ (Z > 4.5) = 1 - 0.9993 = 0.0007$$

For n = 50:

h)

Forklaring av begrepene bias og varians til en observator.

Med observator bias mener vi tendensen til å se hva vi forventer å se, eller hva vi ønsker å se. Det betyr for eksempel at man har noe man tror kommer til å se i en studie og at man ser på eksperimentet med fordommer.

Varians kan forklares som at flere observatører ser forskjellige data ut i fra samme populasjon. Her kan de ta, som i eksemplene ovenfor, prøver som gir utfall som varierer. Altså da variasjon mellom resultat oppnådd av flere observatører av samme material.

Oppgave 2:

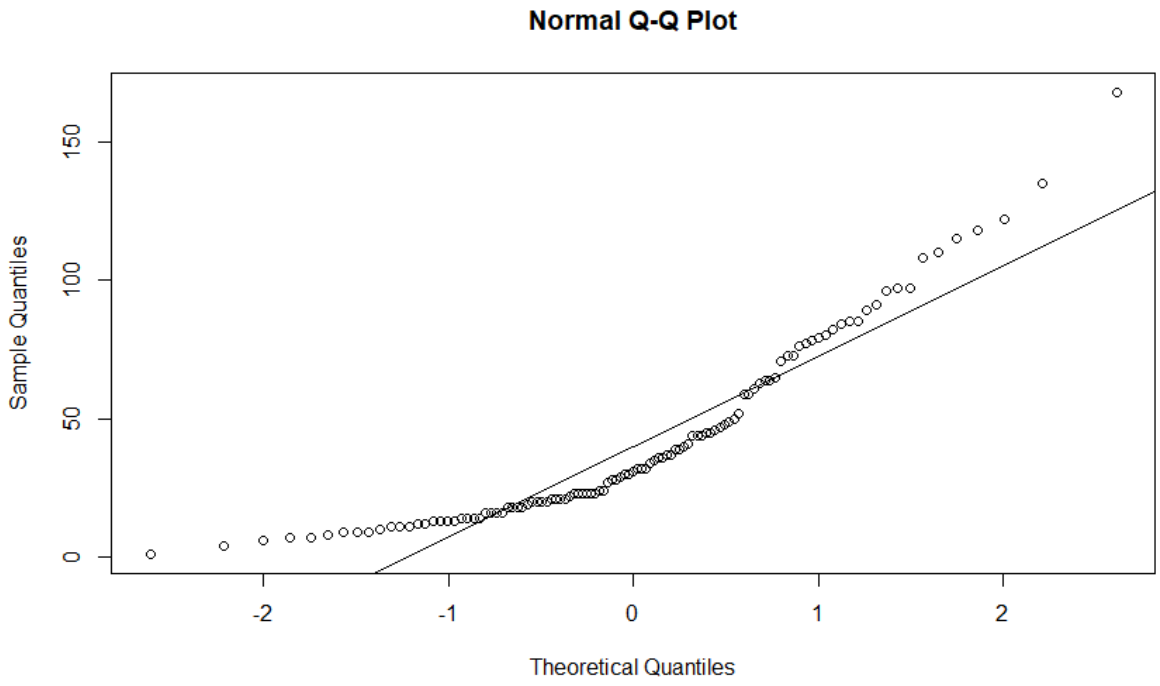
a)

I den første oppgaven skal vi se på hvordan ozonlaget forandrer seg i løpet av de 111 målinger. Målingen er gitt ppb (parts per billion). Det virker som om det er en måling for hver dag.

```
> newyork$ozone
  [1]  41  36  12  18  23  19   8  16  11  14  18  14  34   6  30  11   1
11   4  32  23  45 115  37  29  71  39  23  21
 [30]  37  20  12  13 135  49  32  64  40  77  97  97  85  10  27   7  48
35  61  79  63  16  80 108  20  52  82  50  64
 [59]  59  39   9  16 122  89 110  44  28  65  22  59  23  31  44  21   9
45 168  73  76 118  84  85  96  78  73  91  47
 [88]  32  20  23  21  24  44  21  28   9  13  46  18  13  24  16  13  23
36   7  14  30  14  18  20
```

I følge verdiene ovenfor virker det som om det er lite ozon i mai over New York. I juni stiger dette og holder seg høyt i løpet av hele sommeren. I september synker det igjen. Dette kan ha sammenheng med temperaturen ettersom sommermånedene stor sett er varmere enn vår og høst-månedene. Dette kan igjen ha sammenheng med at varmen kommer av mer radiasjon som slipper igjennom til New York.

```
> qqnorm(newyork$ozone)
> qqline(newyork$ozone)
```



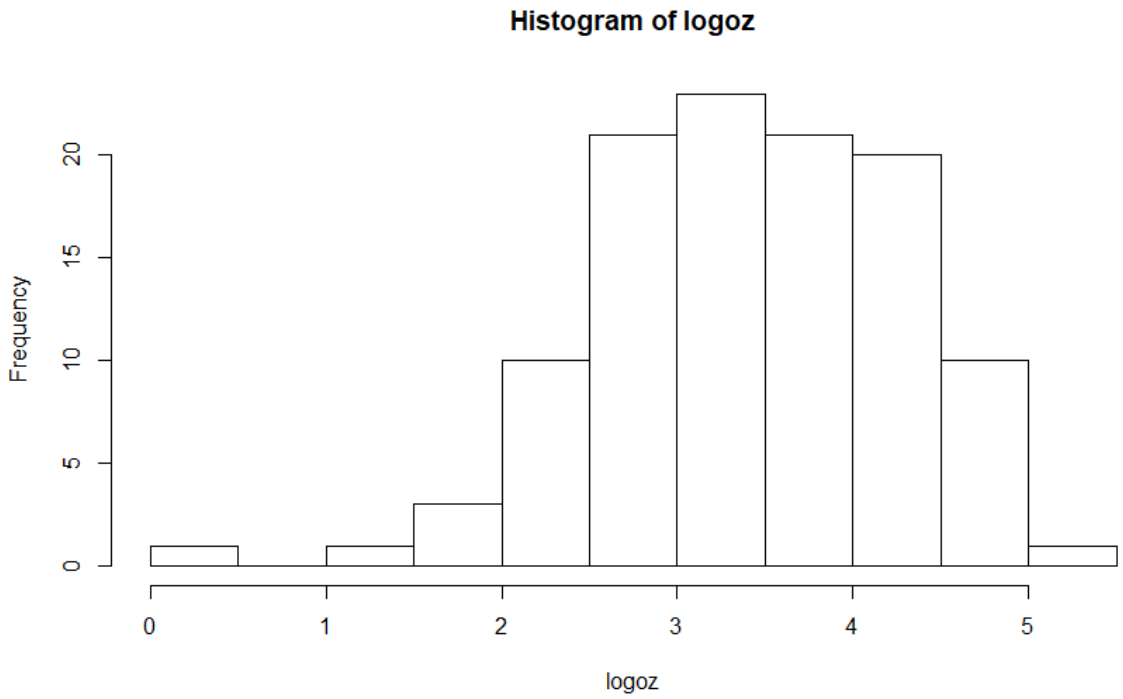
Figur 5: viser fordelingen av ozon-målingene

Vi ser på figur 5 at det ikke er normalfordelte verdier.

b)

Skal se på et histogram over log-verdiene.

```
> logoz <- log(newyork$Ozone)
> hist(logoz)
```



Figur 5: viser histogram over log-verdiene

Ut i fra figure 5 vil jeg si at det ikke er normalfordelt. Dette skyldes at det ikke er normalfordelt rundt det unimodale toppunkt.

c)

Man bruker z når det er snakk om n ≥ 30 fordi da tilsier sentral grense teoremet at fordelingen blir normal. Når $n \leq 30$, så bruker man t-fordelingen.

Antagelser ved t-fordelingen er at den er normalfordelt. Standardavviket er ukjent. Man tar da og tilnærmer standardavviket med standarderror av prøve gjennomsnitt. Log-verdiene er nærmere en normalfordeling. Derfor bør de brukes.

d)

Intensjonen er at man skal finne et 95 prosent konfidensintervall med Rstudio. Dette gjøres så nedenfor.

```
> t.test(logoz)

        One Sample t-test

data:  logoz
t = 41.564, df = 110, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.253057 3.578798
sample estimates:
mean of x
 3.415927
```

Så skal vi bruke ex() funksjonen på øvre og nedre grense oppgitt i konfidensintervallet ovenfor.

```
> exp(3.253057)
[1] 25.8693
> exp(3.578798)
[1] 35.83045
```

e)

I denne oppgaven ser vi på 90 og 99 prosent konfidensintervaller.

```
> t.test(logoz, conf.level = 0.9)

        One Sample t-test

data:  logoz
t = 41.564, df = 110, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 3.279598 3.552257
sample estimates:
mean of x
 3.415927
```

Det over var for 90 prosent. Så ser vi på 99 prosent.

```
> t.test(logoz, conf.level = 0.99)

        One Sample t-test

data:  logoz
t = 41.564, df = 110, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 3.200500 3.631355
sample estimates:
mean of x
 3.415927
```

Så setter vi dem i tabell for å sammenligne dem bedre.

Konf.intervall	Min	Maks
90	3.279598	3.552257
95	3.253057	3.578798
99	3.200500	3.631355

Vi ser at minimumsverdiene blir mindre etter hvert som vi skal være sikrere (altså går opp i prosent på intervallene) og maksverdiene blir høyere.

f)

Først tester vi som oppgaveteksten sier at vi skal gjøre.

```
> logoz.juni <- logoz[newyork$Month == 6]
> logoz.july <- logoz[newyork$Month == 7]
> t.test(logoz.juni, logoz.july)

        Welch Two Sample t-test

data:  logoz.juni and logoz.july
t = -2.755, df = 18.041, p-value = 0.01301
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.140604 -0.153718
sample estimates:
mean of x mean of y
 3.236673  3.883834
```

Det er noen forskjeller mellom juni og juli måneder i ozon-laget, men alt i alt er de ganske like ettersom de er sommermåneder.

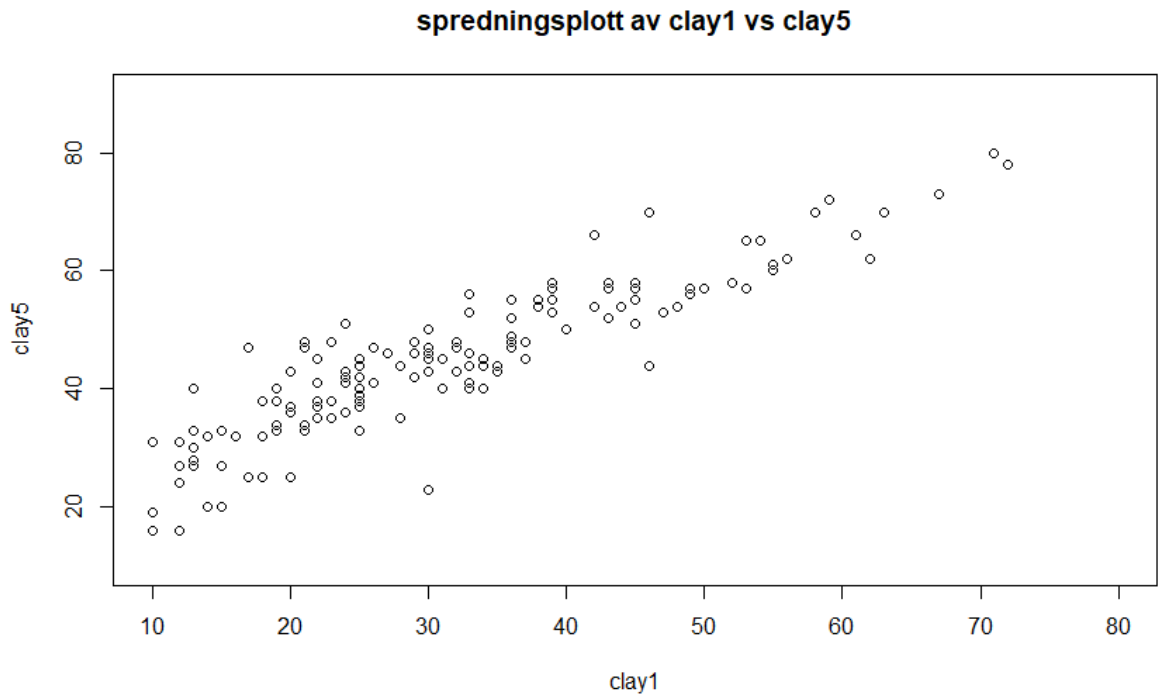
Nullhypotese: det skal ikke være noe relasjon mellom juni og juli måned

Alternativ hypotese: mengde ozon i juni og juli er relaterte til hverandre som en følge av at de er begge sommermåneder

Oppgave 3:

a)

Etter å ha brukt oppgitt i oppgaven og gjort de nødvendige forandringer fikk vi følgende figur vist på figur 6. Basert på figuren ser det ut som om det er et lineært forhold mellom de to variabler. Dette skyldes at r virker å være r = 1 (i nærheten av det i hvert fall).

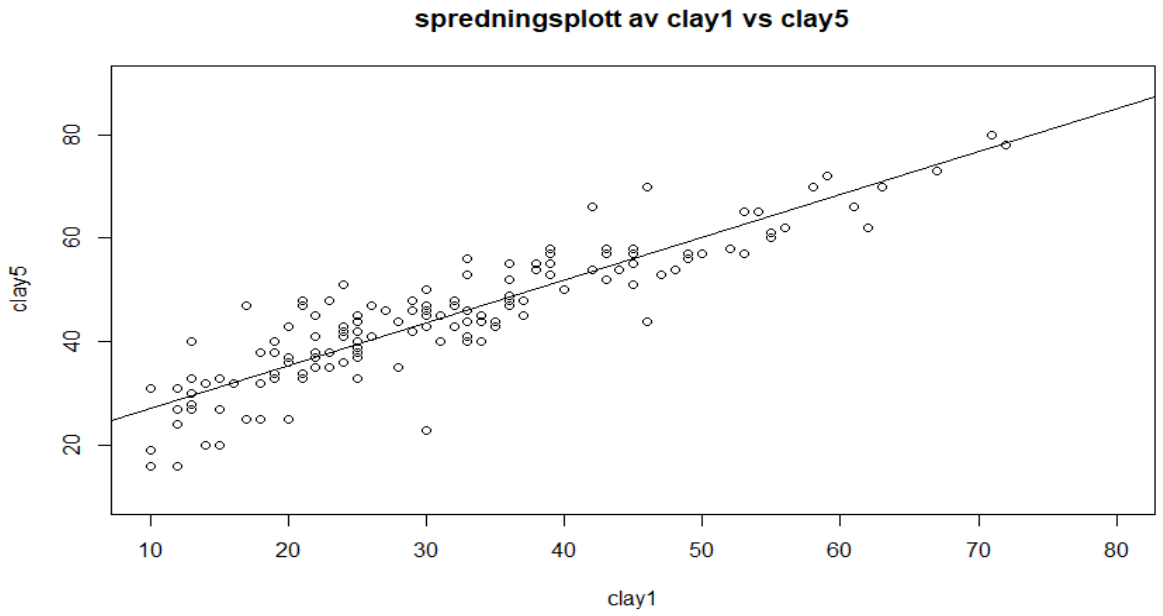


Figur 6: viser spredningsplott av clay1 vs clay5

b)

Ved å bruke koden nedenfor, så skal vi finne en regresjonslinje.

```
> fit <- lm(clay5 ~ clay1)
> abline(fit)
```



Figur 7: viser plottet fra figur 6 med en regresjonslinje

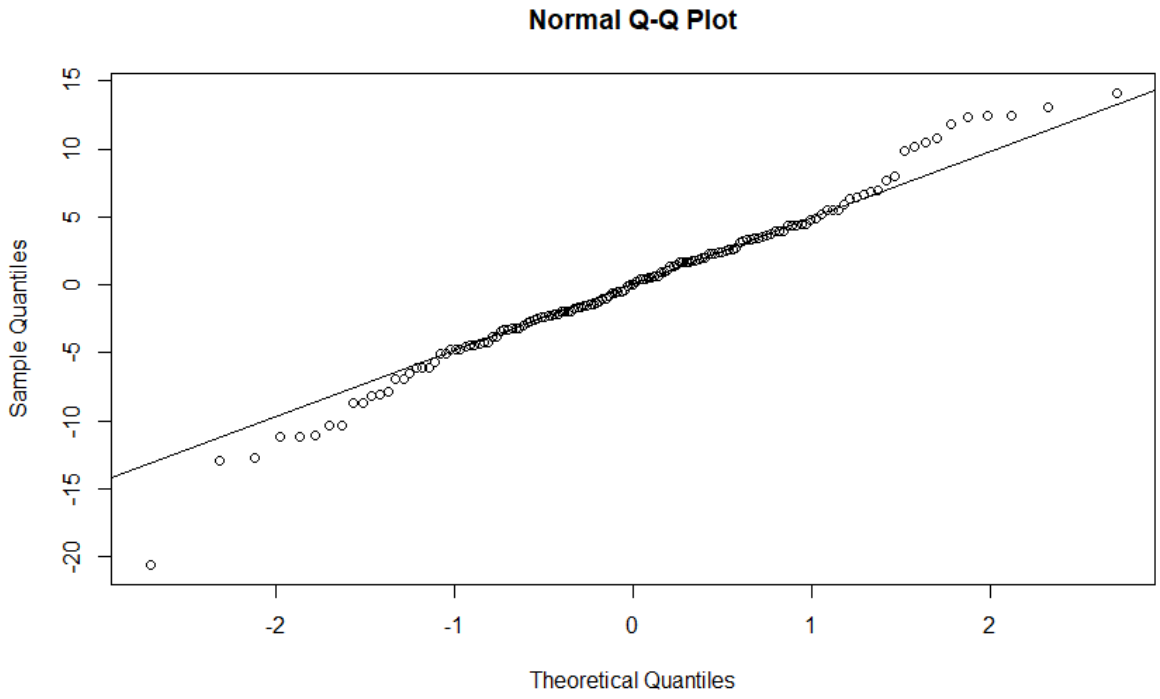
c)

Antagelsene som ligger til grunn for en lineær regresjonslinje er som følgende:

- Den skal forklare hvordan responsvariabelen y endrer seg når forklaringsvariabelen x skifter verdier
- Regresjonslinjen predikerer verdien av y for en gitt verdi av x
- Det skal være gitt på formen

$$y = b_0 + b_1 x$$

- Så har vi også residualer som er da: Residualer = observert y – predikert y. Eller skrevet på en annen måte, $y - (b_0 + b_1 x)$. Dette skal summere seg til 0.



Figur 8: viser residualer for oppgaven

Jeg føler at figur 8 viser at det omtrent summerer seg til 0 når man ser på positiv og negative deler av 0.

d)

Ved bruk av funksjonen summary(fit).

```
> summary(fit)

Call:
lm(formula = clay5 ~ clay1)

Residuals:
     Min       1Q   Median       3Q      Max
-20.6258  -3.1907   0.0055   3.3875  14.1500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.75856    1.15561   16.23   <2e-16 ***
clay1        0.82891    0.03377   24.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.687 on 145 degrees of freedom
Multiple R-squared:  0.806,    Adjusted R-squared:  0.8047
F-statistic: 602.4 on 1 and 145 DF,  p-value: < 2.2e-16
```

Modellens stigningstall og skjæringstall er:

$$y = 0.82891\, x + 18.75856$$

Vi skal så finne hvor mye lag 5 øker etter 5 år når lag 1 øker med 1 prosent for hvert år. Da gjør jeg det slik at jeg setter x = 1 og multipliserer dette tallet så med 5 for å få en sum.

$$(0.82891 * 1 + 18.75856) * 5 = 97.9383$$

Lag 5 blir da nesten fordoblet.

e)

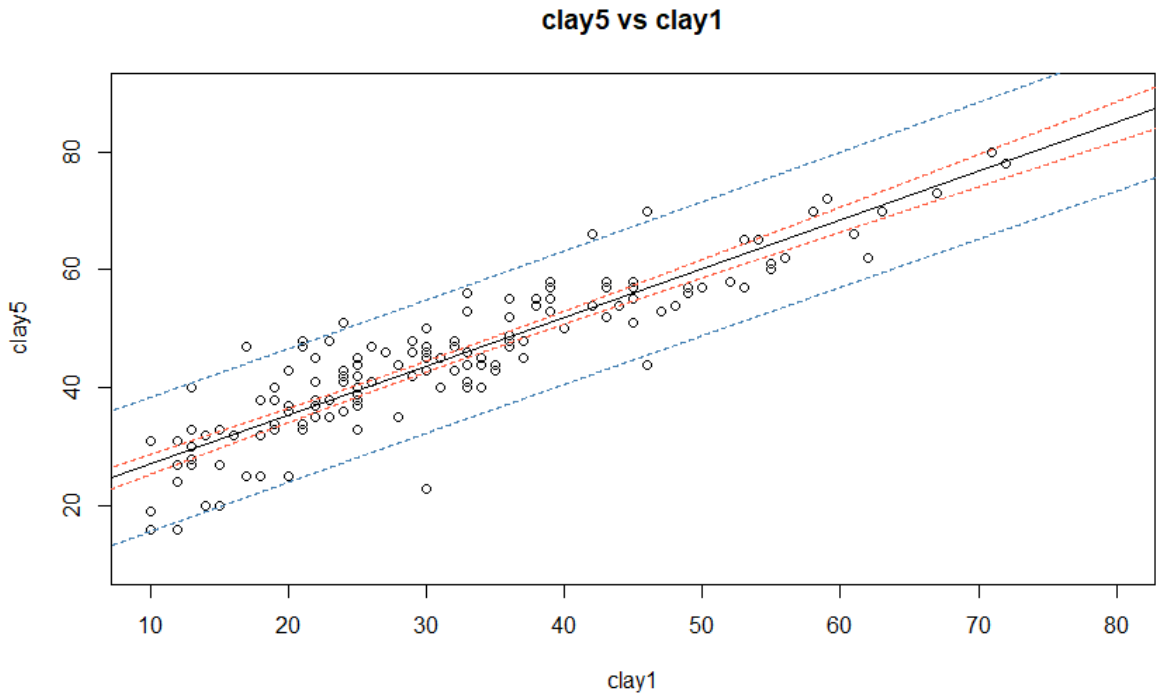
p- verdien er $< 2.2 * 10^{-16}$. En liten p-verdi indikerer bevis mot null-hypotesen. Det medfører da at null-hypotesen blir avvist. Null-hypotesen tilsier at det skal være ingen forhold mellom de to populasjoner. Vi ser da at det er et forhold mellom ettersom p-verdien er så liten som den er.

f)

Konfidensintervallet og koden som man skal bruke gir:

```
> b1
[1] 0.8289082
> se.b1
[1] 0.03377248
> lower
[1] 0.7730003
> upper
[1] 0.8848162
```

g)



Figur 9: viser prediksjonsintervall og konfidensintervall for clay5 vs clay1

Så skal jeg forklare hvordan disse intervallene skal tolkes og hvorfor det ene intervallet er bredere enn det andre. Prediksjonsintervallet er blått, mens konfidensintervallet er rødt. Vi forventer da at prediksjonsintervallet er bredere ettersom det er verdier man tror man kan se, mens konfidensintervaller er det man forventer å se/eller ønsker å se.

h)

I henhold til prediksjonsintervallet og plott på figur 9, skal det være mellom 55 og 80. Mens konfidensintervallet tilsier at det skal være mellom 60 og 65. Jeg antar at det skal være rundt 65.