Oblig 1

Av Furkan Kaya

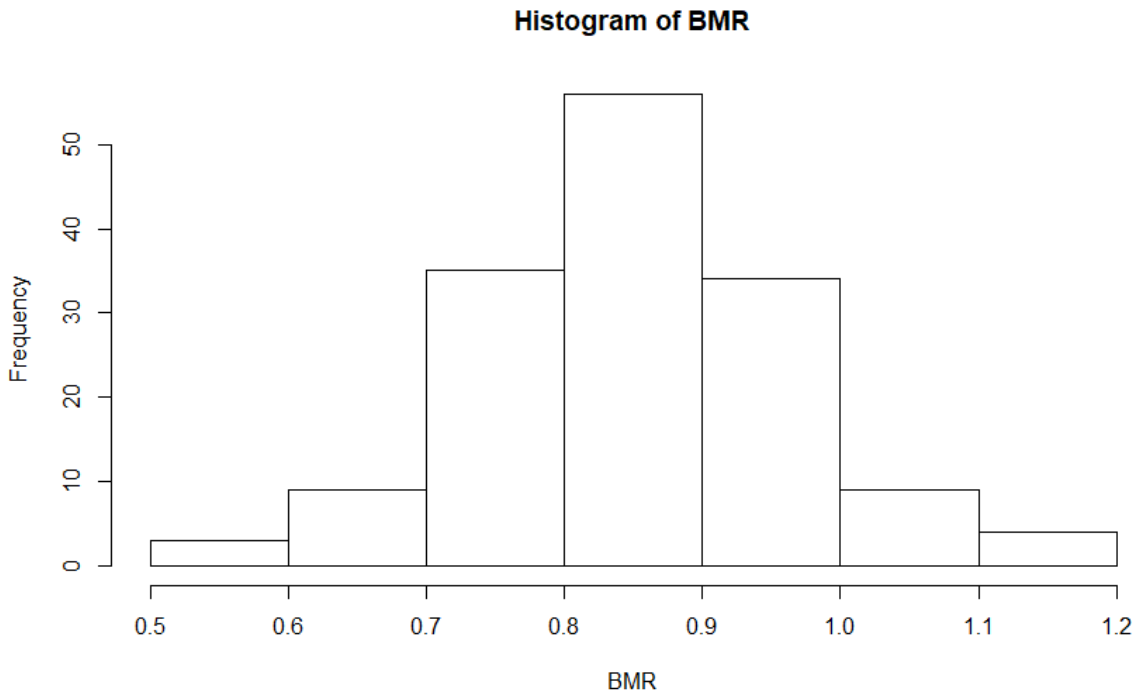
STK1000: Innføring i anvendt statistikk

Figur 1: viser en Sebrafink

Oppgave 1:

a)

I denne første oppgaven skulle vi lage et histogram av BMR for de 150 sebrafinkene. Vi begynte kodingen med å hente tabellen. Det blir vist nedenfor. Jeg legger her til at jeg legger ved koden for hver oppgave på slutten av oppgaven. Resultatet ble:



Figur 2: viser histogram over BMR

Av histogrammet ser vi at over 50 sebrafinker har en BMR på mellom 0.8 og 0.9 mL O2 per minutt. Histogrammet er såkalt unimodalt, noe som betyr at det har en topp. Og at det er symmetrisk.

```
> data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data/zebrafin
ch.txt"
> zebrafinch <- read.table(data,header=TRUE)
> attach(zebrafinch)
> hist(BMR)
```

b)

Vårt formål i oppgave b) er å finne gjennomsnitt og median av BMR. Svarene følger i koden nedenfor. For begge er intensjonen at man skal finne senteret av fordelingen. Men forskjellen mellom dem er at medianen er mer resistent mot outlier-verdier sammenlignet med gjennomsnittet.

```
> mean(BMR)
[1] 0.8485003
> median(BMR)
[1] 0.8397846
```

c)

Vi skal nå finne standardavvik og interkvartil til BMR-verdiene. Kvartil gir spredningen til en fordeling, mens interkvartil (IQR) er distansen mellom tredje og første kvartil. Det blir hovedsakelig brukt til å identifisere outliers og er resistent. Standardavvik (SD) er også å måle spredning, men er lite resistent fordi det er basert på gjennomsnitt fremfor median.

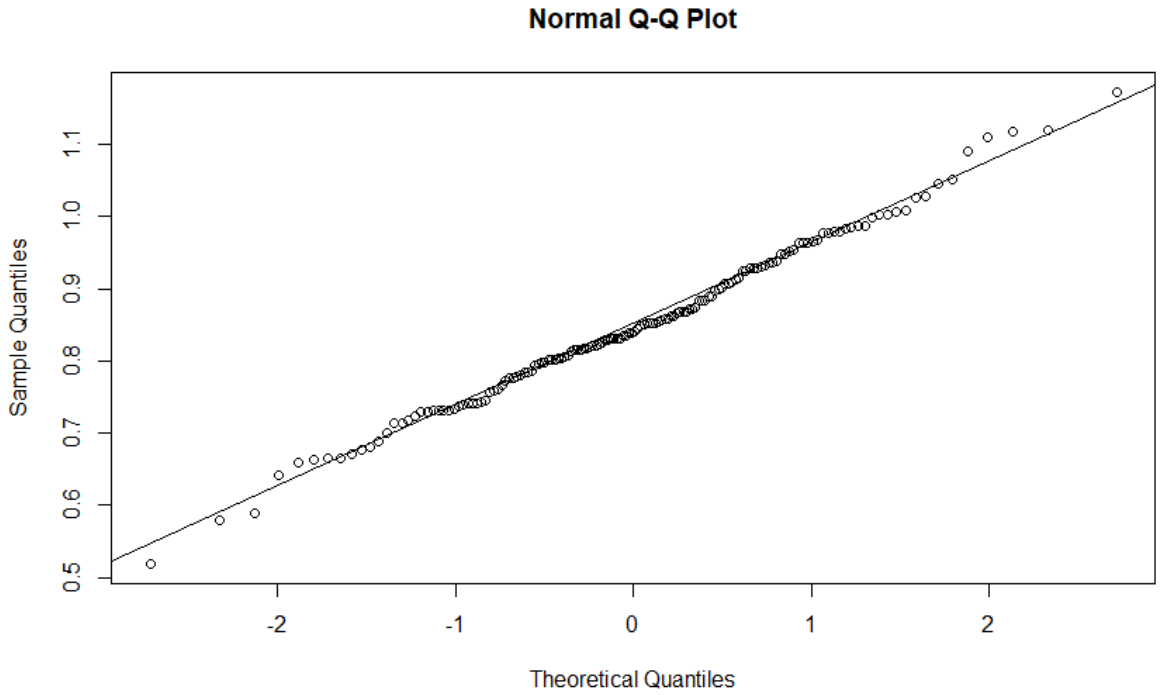
```
> IQR(BMR)
[1] 0.1517731
> sd(BMR)
[1] 0.1134001
```

IQR beskriver på en måte de midtre 50 prosent ved at 75 prosent – 25 prosent = 50 prosent. SD gir hele fordelingen. Av dette ser vi at IQR av BMR er større enn SD. Vi vet da fra pensum at når SD = 0 eller IQR = 0, så har alle observerte verdier samme verdi. Når spredningen fra gjennomsnittet eller medianen blir større, så øker også henholdsvis SD eller IQR. Av dette tolker vi da det slik at det er større spredningen rundt medianen og i 50 prosent-delen av BMR, mens hele fordelingen er jevnere fordelt.

d)

Vi skal i denne oppgaven vurdere om BMR er normalfordelt ved å bruke qqplot og qqline funksjonene i Rstudio. Ved å ha begge funksjoner, så kan vi sammenligne dem. Hvis verdiene i qqplot er omtrent på linjen fra qqline, så indikerer det at det er normalfordeling. Da ser vi på

figuren vi fikk i R.



Figur 3: skjematikk som viser qqplot og qqline av BMR

Figur 3 viser at punktene ligger omtrent på linjen. Det gir da at det er en normalfordeling.

```
> qqnorm(BMR)
> qqline(BMR)
```

e)

Vi skal finne den standardiserte verdien av BMR = 0.8. Dette gjøres ved hjelp av likningen:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Hvor Z er altså den standardiserte verdien. Og de to andre variablene er funnet i b) og c). Innsatt gir dette da:

$$Z = \frac{0.8 - 0.8485003}{0.1134001} = -0.4276918$$

Først og fremst er det verdt å merke seg at den standardiserte verdien er negativ. Det er altså 0.43 standardavvik mindre enn gjennomsnittlig verdi.

f)

Oppgaven ønsker at vi skal finne sannsynligheten for å finne en sebrafink med BMR under 0.6. Da bruker vi (1) igjen og får at:

$$Z = \frac{0.6 - 0.8485003}{0.1134001} = -2.191359$$

Ved bruk av tabell A i pensumboken gir dette oss en sannsynlighet på p = 0.0143. Vi kan sammenligne dette med verdiene fra zebrafinch hvor 2 ved manuell opptelling man kan se at det er 2 verdier som er under 0.6. Når vi bruker 2/150 får vi et svar som sier oss at 0.0133. Altså i nærheten av det vi skal finne.

g)

Samme oppgave som den forrige, men vi skal her finne sannsynligheten for å finne BMR over 1. Ligning (1) gir en standardisert Z på 1.33598. Vi skal altså finne X > 1, som gir at siden Z for 1.33598 er 0.9082. Noe som gir p = 1 – 0.9082 = 0.0918 sannsynlighet.

Oppgave 2:

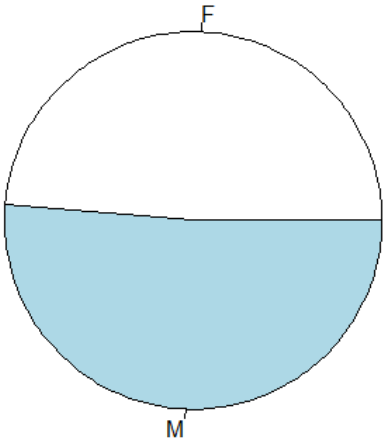
a)

Forskjellen er at en kategorisk variabel plasserer et tilfelle i en av flere grupper eller kategorier, mens en kvantitativ variabel gir en numerisk verdi som kan bruke aritmetikk på. I eksempelet er kjønn en kategorisk variabel. Noe «tungt jaktet» og «lett jaktet» i kolonne to også kan karakteriseres som. Variablene som gir konsentrasjonen av enten kortisol eller testeron er kvantitative variabler.

b)

Først lager vi tabell og pie-chart for kjønn.

```
> table(sex)
sex
 F  M
72 76
> pie(table(sex))
```

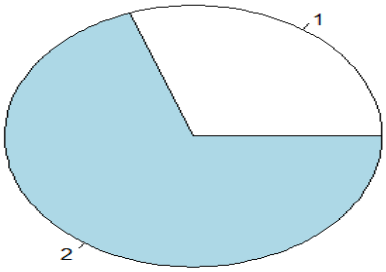


Figur 4: viser pie-chart over kjønnene på ulvene

Vi ser at det er noen flere mannlige ulver enn kvinnelige (om man kan bruke begrepene som de er for mennesker selvsagt).

Neste punkt er da å gjøre det samme for hvorvidt populasjonen av ulver enten er tung jaktet eller lett jaktet.

```
> table(population)
population
  1   2
 45 103
> pie(table(population))
```



Figur 5: viser pie-chart over om ulven tilhører en populasjon som er tungt jaktet (2) eller lett jaktet (1)

Vi ser at den store majoriteten av befolkningen er tungt jaktet. Det betyr da for eksempel at den er i proksimitet til menneskeheten.

c)

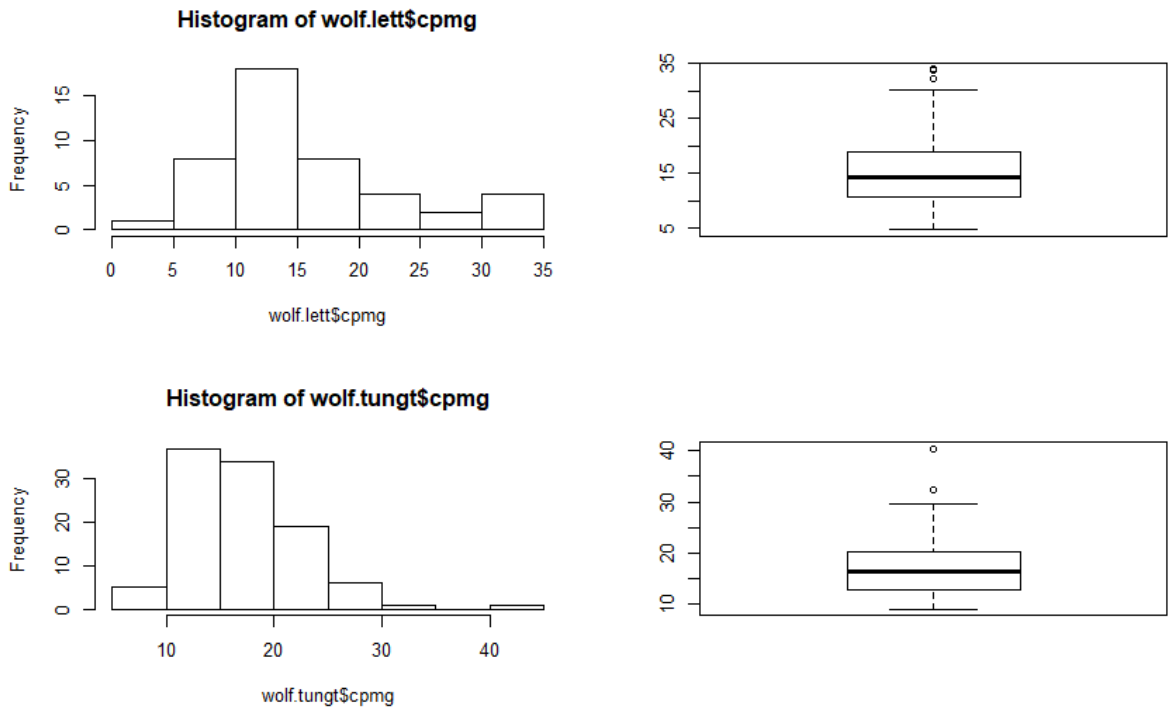
Ifølge oppgaveteksten skal vi dele datasettet inn i to. Et for tung jaktet og et for lett jaktet. Kommandoene kommer nedenfor.

```
> wolf.lett <- wolf[wolf[,"population"]==1,]
> wolf.tungt <- wolf[wolf[,"population"]==2,]
```

#Ved å studere tabellene over dataene etter at de har blitt delt inn i to, kan vi bekrefte at det skjer en kategorisk inndeling som det påpekes ovenfor.

d)

Nedenfor følger en oversikt over kortisol for både tung jaktet og lett jaktet ulvebestand.



Figur 6: er en oversikt over kortisol for de to ulvebestandene. Lett jaktet følger øverst, mens tungt jaktet følger nederst

Dette kommer i form av histogram og boksplott slik oppgaveteksten tilsier. I histogrammet for tung jaktet ulvebestand ser vi at det blir produsert mer kortisol sammenlignet med den lett jaktede ulvebestanden. Vi ser også en enkelt topp for sistnevnte i intervallet 10-15, mens det er flere topper av nesten lik størrelse for den tung jaktede bestanden. Boksplottetene gir at det for den lett jaktede er et toppunkt som er som er mer enn outlier enn det for den tung jaktede. Interkvartilene virker forholdsvis like. Det samme kan man si om forholdet til medianen.

```
> par(mfrow=c(2,2))
> hist(wolf.lett$cpmg)
> boxplot(wolf.lett$cpmg)
> hist(wolf.tungt$cpmg)
> boxplot(wolf.tungt$cpmg)
```

Standard kode som før, men med et unntak. Her indikerer dollartegnet at man ønsker variabelen cpmg i henholdsvis wolf.lett og wolf.tung.

e)

Resultatene nedenfor er for gjennomsnittet og medianen for de respektive inndelingene for ulvebestandene fra tidligere oppgaver.

```
> mean(wolf.lett$cpmg)
[1] 15.56222
> median(wolf.lett$cpmg)
[1] 14.24
> mean(wolf.tungt$cpmg)
[1] 17.07495
> median(wolf.tungt$cpmg)
[1] 16.32
```

Vi ser at gjennomsnittlig kortisol i lett jaktet ulv er lavere enn i gjennomsnittlig kortisol tung jaktet ulv. Denne tendensen går igjen i median også. Ellers er medianen mindre enn gjennomsnittet. Det betyr at det finnes outlier-verdier som fører til dette fenomenet. Disse verdiene ser vi også på boksplottet i figur 6.

f)

For å svare på denne oppgaven må vi først finne standardavviket og de andre variablene i femtallsoppsummeringen.

Gjennomsnitt med standardavvik for hver av populasjonene blir:

$Wolf.lett = 15.56222 \pm 7.298785$

$Wolf.tungt = 17.07495 \pm 5.543389$

For femtallsoppsummeringen trenger vi kvartil 1,3, median, minste verdi og maksverdi. Disse er da (i tabellform):

Populasjon	Lett jaktet	Tungt jaktet
Minste verdi	4.75	8.91
Maks verdi	34	40.43
Kvartil 1	10.79	12.58
Kvartil 3	18.92	20.185
Median	14.24	16.32

Med standardavviket og gjennomsnittet får vi en beskrivelse av populasjonene som er litt rudimentær føler jeg. Det har et gjennomsnitt som enkelt kan være offer for outlier-verdier og et standardavvik som tilsier at det er et stort spenn. Spesielt gjelder dette for lett jaktet ulvebestand, mens verdiene er bedre for tungt jaktede. Sånn rent avlesningsmessig sett. Spennet er heller ikke så stort for tungt jaktet. Personlig foretrekker jeg femtallsoppsummeringen fordi man får med mye mer. Gjennomsnitt med standardavvik fungerer fint om man skal lese av fort og få en viss forståelse. Men for en dypere forståelse av kortisolkonsentrasjonen ville jeg ha foretrukket femtallsoppsummering ettersom det gir deg minste og maksverdi, samt median (som kan sammenlignes med gjennomsnitt). Kvartilene kan også på en måte sammenlignes med standardavviket i det å forstå spennet.

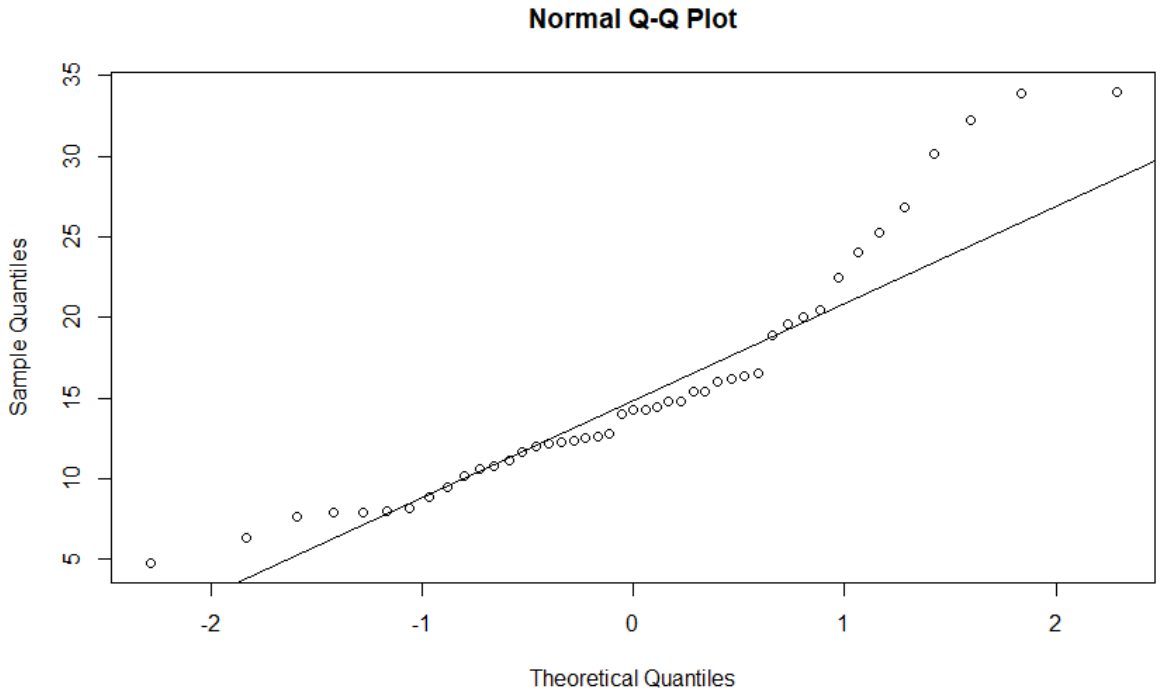
```
sd(wolf.lett$cpmg)
[1] 7.298785
> sd(wolf.tungt$cpmg)
[1] 5.543389
> quantile(wolf.lett$cpmg)
   0%   25%   50%   75%  100%
```

```
 4.75 10.79 14.24 18.92 34.00
> quantile(wolf.tungt$cpmg)
    0%    25%    50%    75%   100%
 8.910 12.580 16.320 20.185 40.430
> max(wolf.lett$cpmg)
[1] 34
> min(wolf.lett$cpmg)
[1] 4.75
> max(wolf.tungt$cpmg)
[1] 40.43
> min(wolf.tungt$cpmg)
[1] 8.91
> median(wolf.lett$cpmg)
[1] 14.24
> median(wolf.tungt$cpmg)
[1] 16.32
```

#viser forskjellige koder for å få forskjellige verdier som brukes i femtallsoppsummering og det manglende standardavviket for kortisol.

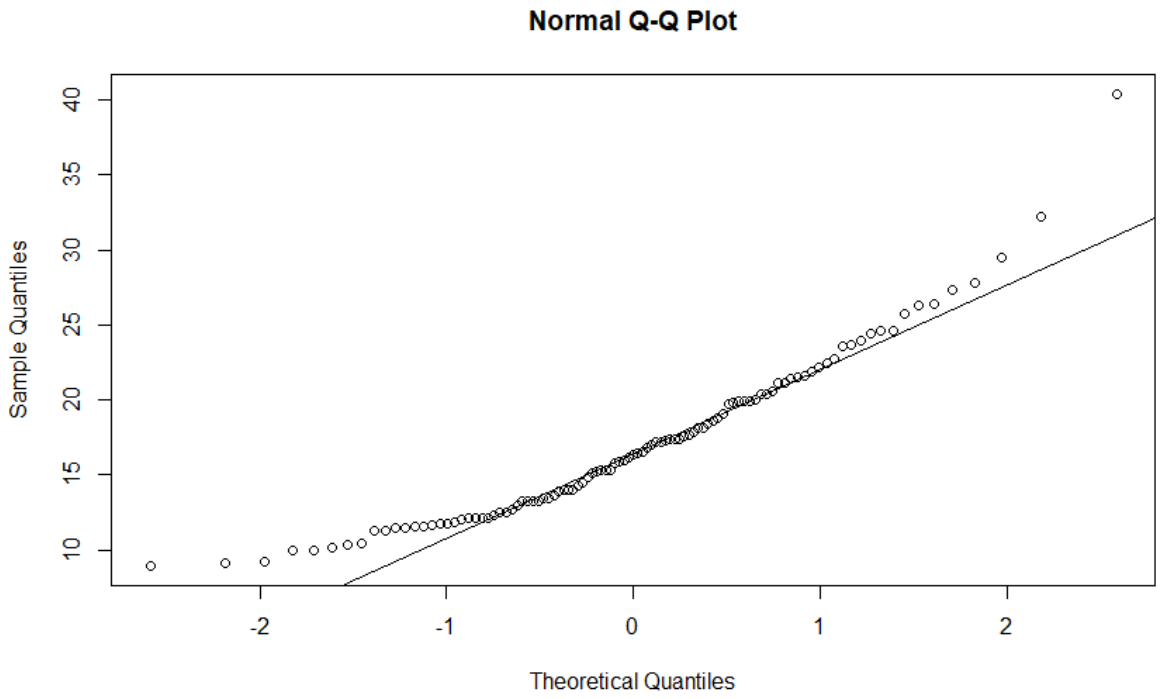
g)

Som tidligere benytter jeg meg av qqnorm og qqline funksjonene for å se om kortisolkonsentrasjonen er normalfordelt.



Figur 7: viser qqnorm og qqline for lett jaktet ulvebestand

Figur 7 viser hvordan qqnorm og qqline er for lett jaktet ulvebestand. Vi ser at den vanskelig kan betegnes som normalfordelt ettersom det er mange verdier på slutten og begynnelsen som ikke følger normal-linjen. Men i midten er det en del som samsvarer med et slags toppunkt.



Figur 8: viser qqnom og qqline for tungt jaktede ulvbestand

Figur 8 viser hvordan forholdet mellom qqnorm og qqline-funksjonen i Rstudio er. Her ser vi en klarere sammenheng og kan erklære at tungt jaktet ulvebestand statistikken er normalfordelt.

```
> qqnorm(wolf.lett$cpmg)
> qqline(wolf.lett$cpmg)
> qqnorm(wolf.tungt$cpmg)
> qqline(wolf.tungt$cpmg)
```

#viser hvordan qqnorm og qqline funksjonene blir brukt for de to populasjonsfordelingene.

Oppgave 3:

a)

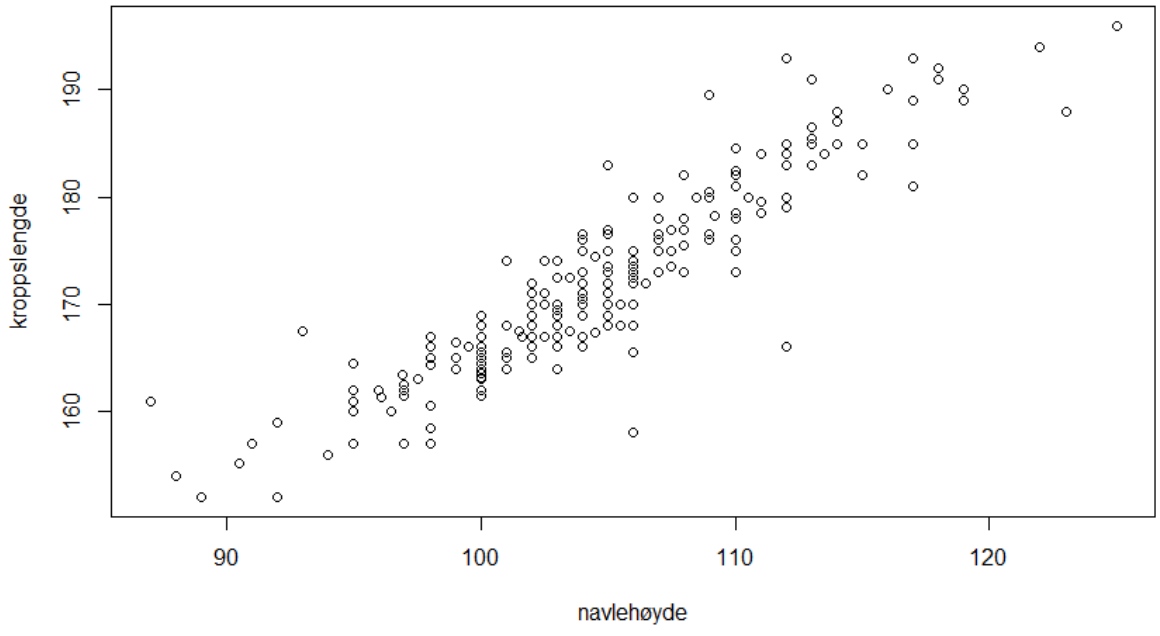
Vi bruker summary-funksjonen i Rstudio til å lage et sammendrag av statistikken fra databasen.

```
> summary(vitruvisk$kjonn)
  K   M
150  73
> summary(vitruvisk$kroppslengde)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  152.0   166.0   172.0   172.3   178.0   196.0
> summary(vitruvisk$fot.navle)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   87.0   101.0   104.0   104.8   109.0   125.0
> summary(vitruvisk$navle.isse)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.00   65.00   67.00   67.34   70.00   81.00
> summary(vitruvisk$favn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  146.0   165.0   171.0   172.4   180.0   202.0
```

#Summary-funksjonen blir brukt på alle kolonner.

Når det gjelder andre spørsmål som må besvares er de alle oppgitt i blå kodetekst ovenfor.

b)



Figur 9: viser kroppslengde på y-aksen og navlehøyde på x-aksen

Ut i fra figur 9 ser vi at høyere navlehøyde medfører lengre kroppslengde. Som en ekstrabonus kan vi legge til at figuren også viser en korrelasjon som kan være nært opptil 1 basert på øyemål.

```
> plot(vitruvisk$fot.navle,vitruvisk$kroppslengde,xlab="navlehøyde",
+      ylab="kroppslengde")
```

c)

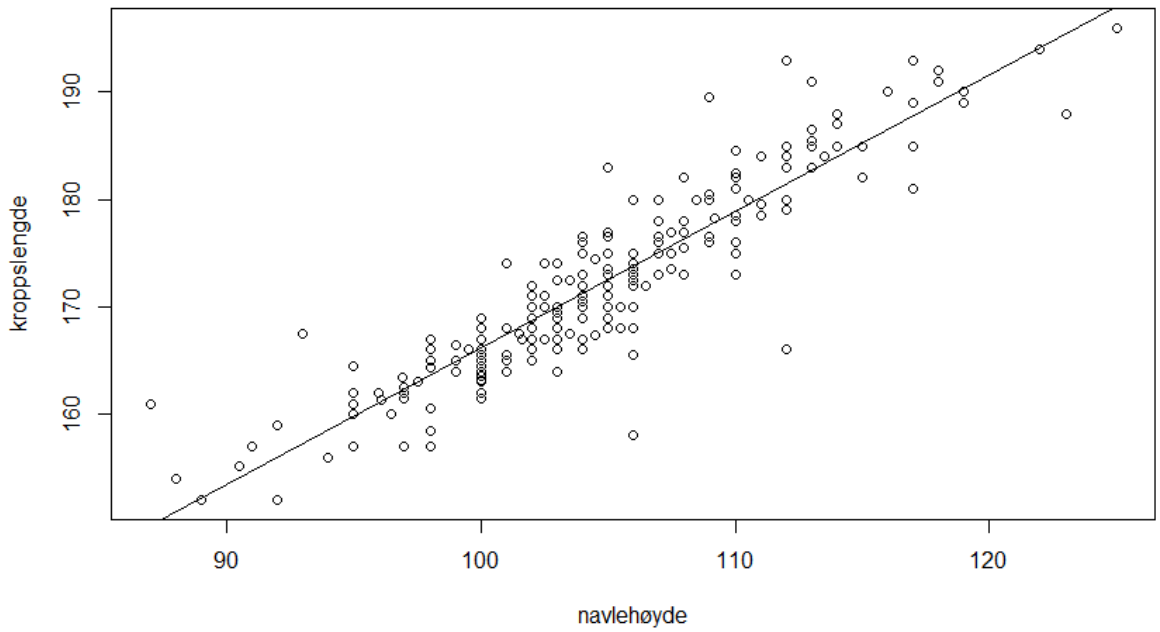
I forrige oppgave ble det gitt en korrelasjonsverdi basert på øyemål. I denne gaven skal vi se om dette stemmer med faktisk verdi.

```
> cor(vitruvisk$fot.navle,vitruvisk$kroppslengde)
[1] 0.9140397
```

Svaret tolkes slik at perfekt korrelasjon er -1 eller 1, avhengig av retning på plottet. Positiv «stigningtall» gir da 1 og vi ser at en verdi på 0.914 tilsier at det er veldig god korrelasjon og at vi får bekreftet at høy navlehøy gir lang kroppslengde.

d)

Man ønsker en regresjonslinje med basis i koden nedenfor for oppgave 3b).



Figur 10: gir en regresjonslinje for statistiske verdier i oppgave 3b)

```
> fit <- lm(vitruvisk$kroppslengde ~ vitruvisk$fot.navle)
> abline(fit)
```

#fit-koden i oppgaveteksten mangler ~, dette måtte legges til på egen hånd.

e)

Vi skal bruke summary(fit) til å finne koeffisientene.

```
> summary(fit)

Call:
lm(formula = vitruvisk$kroppslengde ~ vitruvisk$fot.navle)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7843  -1.9667  -0.0568   2.1695  11.8981

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         38.89675    3.98888   9.751   <2e-16 ***
vitruvisk$fot.navle  1.27252    0.03799  33.499   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.605 on 221 degrees of freedom
Multiple R-squared:  0.8355,   Adjusted R-squared:  0.8347
F-statistic:  1122 on 1 and 221 DF,  p-value: < 2.2e-16
```

#Funksjonen kaller inn likningen fra forrige oppgave. Og den finner residuals, krysningspunkt og fempunktsoppsummeringen.

For sikkerhets skyld, krysningspunktet er 38.89675 og stigningstallet er 1.27525.

f)

Vi skal bruke modellen i oppgave d) til å finne kroppslengden til en person med en navlehøyde på 121 cm.

$$y\,(x) = 1.27525x + 38.89675$$

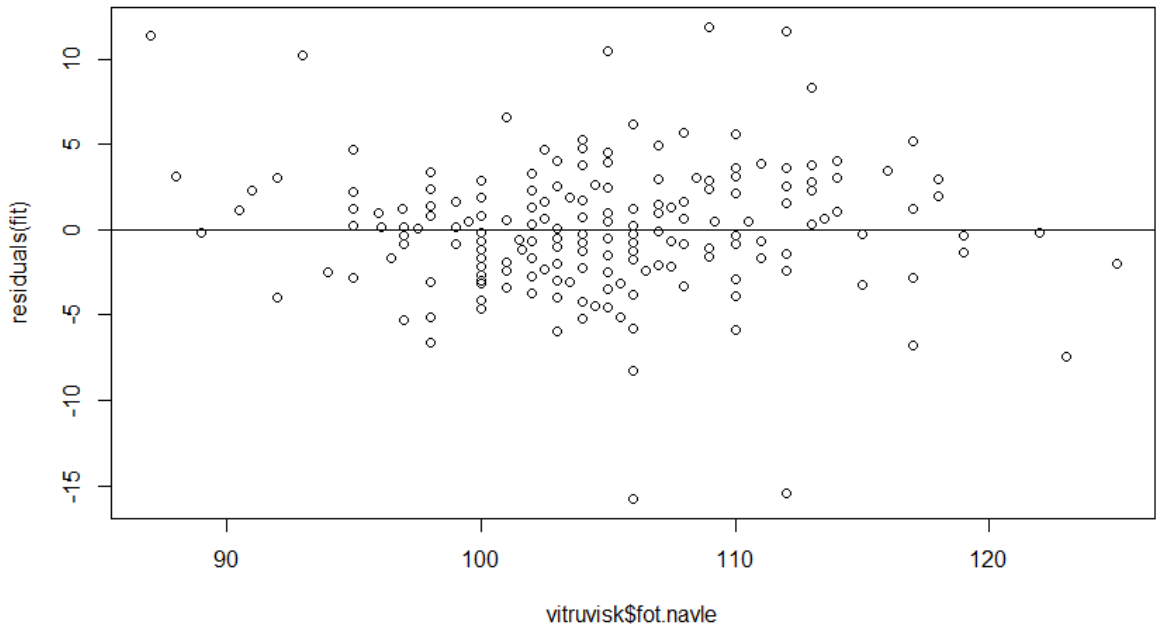
$$y(121) = 193.202\ cm$$

g)

Vil først påpeke at jeg er usikker om jeg har forstått oppgaven riktig. Burde kanskje ha spurt om veiledning eller gjort oppgaven med andre studenter, men føler at man forstår det først om man gjør et oppgavesett av seg selv. Gjør derfor mitt beste og avventer respons.

Ved å se på medianen vil jeg si at verdiene er 1- 0.0568 = 0.9432 bestemt av navlehøyden. Sammenlignet med korrelasjon og dets fundament i gjennomsnitt har vi at sistnevnte er mer påvirket av outlier-verdier sammenlignet med summary-funksjonen.

h)



Figur 11: viser plottet av residualer

På figur 11 ser man flere outlier-verdier. Disse ligger på 10 for positiv del av aksen og -15 for negativ del av aksen. Da tenker jeg altså at de ligger på en horisontal linje på den verdien.

En residual er altså forskjellen mellom en observert verdi til respons variabelen og verdien forutsett av regresjonslinjen. Ut i fra figur 11 vil jeg si at den passer ganske bra. Det er få outlier-verdier og de fleste observerte verdier ligger innenfor standardavvik.