STK1000 - oblig 1

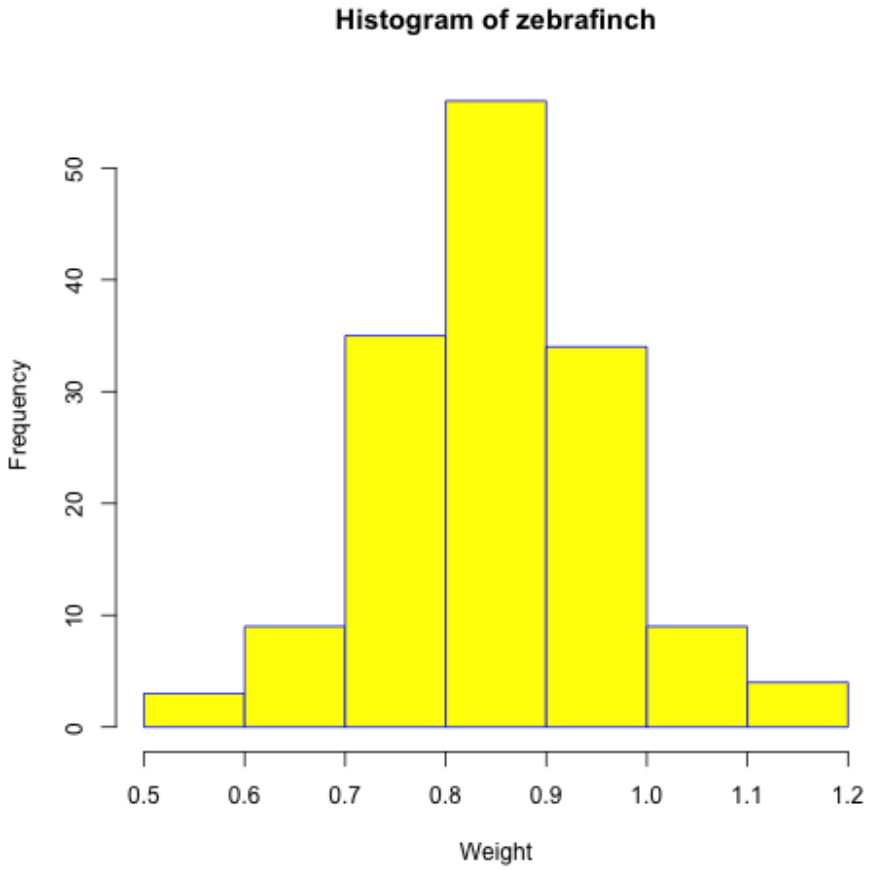
Mikael B. Kiste

20. september 2018

1

a

Her er et histogram av BMR verdiene til sebrafinkene. Histogrammet ser normalfordelt ut, med en god del sebrafinker som har en BMR verdi nær gjennomsnittet og et kraftig avtagende antall finker jo lenger man går vekk fra gjennomsnittet.



Figur 1: Histogram av BMR-verdiene til 150 zebrafisker

b

Ved å bruke kommandoen `mean` i R får jeg gjennomsnitt og `median` gir meg median.

$$\bar{x} = 0.8485003\text{mL}\frac{o_2}{\text{m}}$$
$$\text{mean} = 0.8397846\text{mL}\frac{o_2}{\text{m}}$$

Gjennomsnittet tar summen av måleverdien til alle datapunktene før det deler på antall datapunkter mens median er verdien på det midterste elementet i en ordnet (numerisk økende) liste av datapunktene.

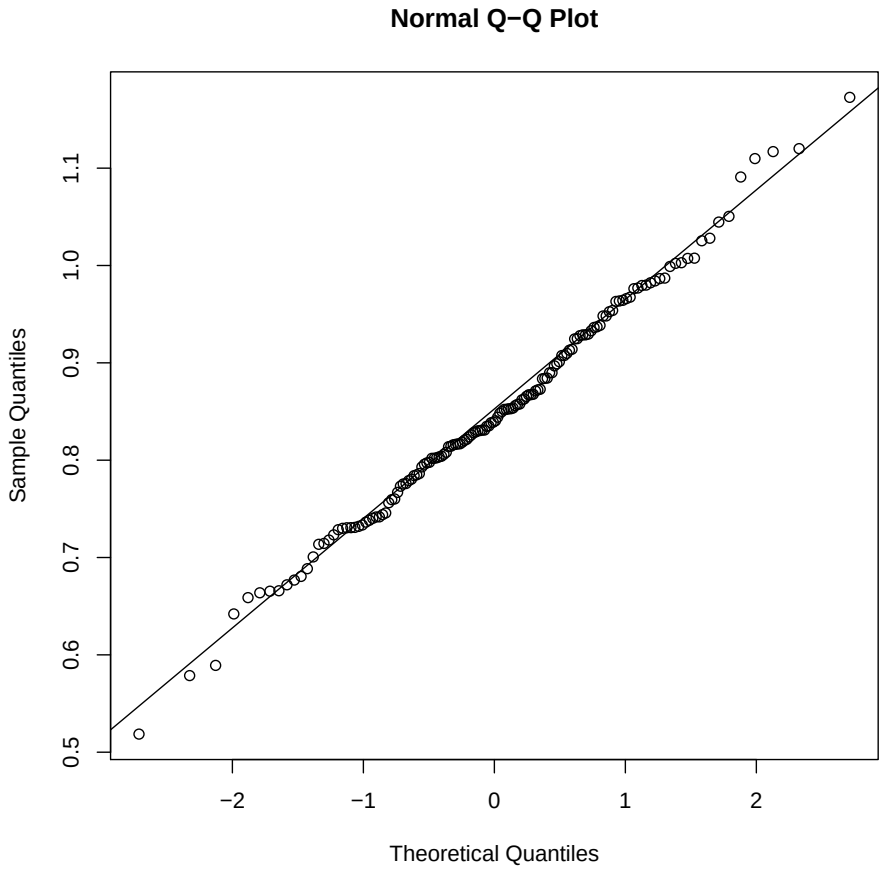
c

Ved å bruke kommandoen `IQR` i R får jeg inter-kvartil og `sd` gir meg standardavvik.

$$\text{IQR} = 0.1517731\text{mL}\frac{o_2}{\text{m}}$$
$$\sigma = 0.1134001\text{mL}\frac{o_2}{\text{m}}$$

d

Ved å bruke kommandoene `qqnorm()` i R får jeg inter-kvartil og `sd()` gir meg standardavvik.



Figur 2: Figuren viser målte verdier mot standardavviket vi fikk tidligere i oppgaven

Plottet indikerer hvordan de målte verdiene har fordelt seg i forhold til standardavviket. Igjen ser man at de fleste datapunktene legger seg rundt gjennomsnittet, men i tillegg får man et inntrykk av hvordan målverdiene avviker fra den teoretiske perfekte normalfordelingen (over eller under den lineære funksjonen). Man kan se at dataen passer ganske bra til en normalfordeling.

e

Standardiserte verdien til BMR: Den standardiserte verdien, eller 'z-score', antyder hvor mange standardavvik en målverdi er fra gjennomsnittet. Altså

$$z = \frac{x - \mu}{\sigma}$$

For BMR = 0.8 mL O_2/min gir det:

$$z = \frac{x - \mu}{\sigma} = \frac{0.8 - 0.8485003}{0.1134001} \approx -0.43$$

Målverdien på $0.8\text{mL}\frac{\text{o}_2}{\text{m}}$ er -0.43 standardavvik unna gjennomsnittet.

f

Ved å bruke kommandoen `pnorm()` i R får jeg prosentilen til en BMR på 0.6

$$p(x < 0.6) = 0.01421294$$

g

På samme måte som i forrige oppgave bruker jeg `pnorm()`, men denne gangen tar jeg én og trekker i fra sannsynligheten for å finne ut sannsynligheten for at en fink har en BMR som er OVER dette.)

$$p(x > 1.0) = 0.09077876$$

2

a

Kvantitative variabler er variabler som har en konkret tallverdi. Dette tillater at flere nyttige numeriske operasjoner kan utføres på et datasett av kvantitative variabler. Konsentrasjon av kortisol og testosteron er også kvantitative
Kvalitative variabler derimot kan være mer abstrakte. Som for eksempel øyenfarge.
Kategoriske variabler er alltid medlemmer av et gitt sett av mulige verdier. Innen vitenskap er det for eksempel fortsatt slik at nesten utelukkende alle dyr kun kan ha ett av to kjønn: han (maskulin) eller hun (feminin). Kjønn er altså et eksempel på en kategorisk variabel. I dette tilfellet er også populasjon kategorisk når den kun deler populasjonen inn i ulv som er hardt jaktet eller ikke. Legg merke til at denne også er en kvalitativ variabel. Kategoriske variabler kan være enten kvalitative eller kvantitative.

b

Ved å bruke kommandoene `pie(table(variabel))` i R får jeg et kakediagram over de kategoriske variablene.

Figur 3: kjønnsfordelingen blandt ulvepopulasjonen er svært jevnt fordelt, hvilket nesten alltid er tilfellet pga naturlig seleksjon. Omtrent to tredeler er tungt jaktet

c

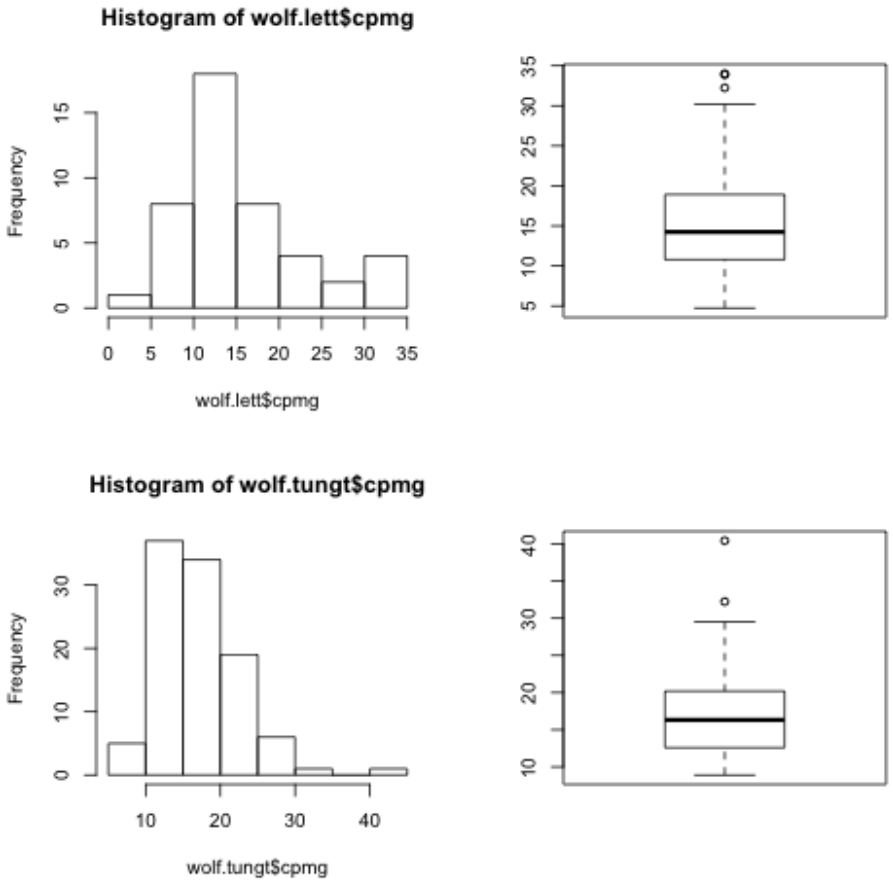
Denne oppgaven løser jeg ved å bruke kommandoene `wolf.lett <- wolf[wolf[,population"]==1,]` og `wolf.tungt <- wolf[wolf[,population"]==2,]`, slik jeg ble fortalt i oppgaven.

d

Her bruker jeg de følgende kommandoene:

- `par(mfrow=c(2,2))`
- `hist(wolf.lett)`
- `boxplot(wolf.lett)`
- `hist(wolf.tungt)`
- `boxplot(wolf.tungt)`

Hvordan disse fungerer er kommentert i koden



Figur 4: Her er histogram og boxplott av kortisolkonsentrasjonen til ulv som har blitt lett jaktet (øverst) og tungt jaktet (nederst)

Man kan se at ulv som har blitt tungt jaktet har høyere konsentrasjon av kortisol.

e

I begge tilfeller er medianen mindre enn gjennomsnittet. Dette antydes også gjennom histogrammene

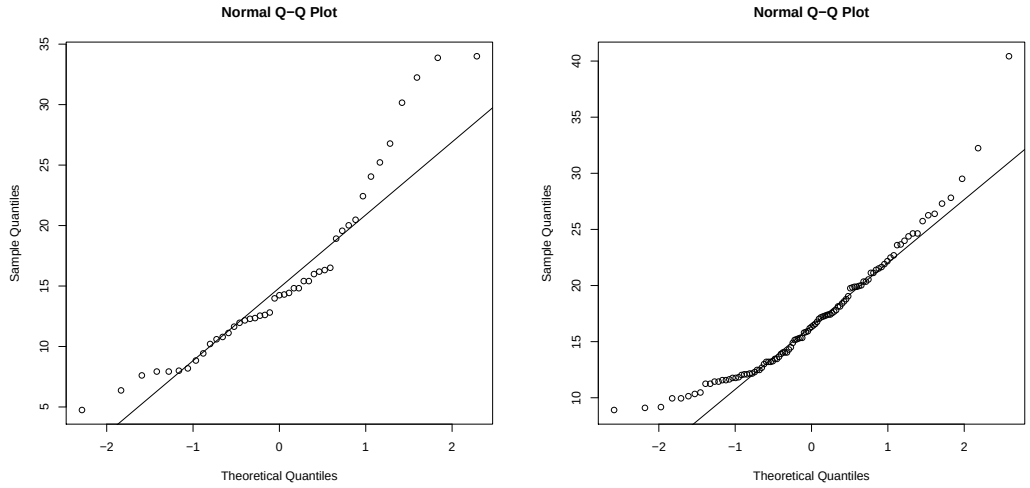
Tabell 1: Tabell over gjennomsnitt, median og standardavvik

	Mean	Median	Standard deviation
Lett	15.56222	14.24	7.298785
Tungt	17.07495	16.32	5.543389

f

For normalfordelte datasett holder det å oppgi gjennomsnitt og standardavvik for å tilstrekkelig beskrive fordelingen. I andre tilfeller trenger man mer informasjon og da kan det være nyttig å bruke en femtallsoppsummering som i tillegg gir minimumverdi, maximumverdi og kvartiler. For å svare på spørsmålet må vi altså se på populasjon og avgjøre om de er normalfordelt eller ikke

g



(a) Lett jaktet ulv (b) Tungt jaktet ulv

Figur 5: QQ plot over kortisolnivå for populasjonene

Fra figur 5 ser det ut til at tungt jaktet ulv er normalfordelt mens lett jaktet ulv ikke er det. Altså kan en bruke gjennomsnitt og standardavvik for å tilstrekkelig beskrive den tungt jaktede populasjonen mens for den lett jaktede populasjonen bør man bruke femtallsoppsummering.

3

a

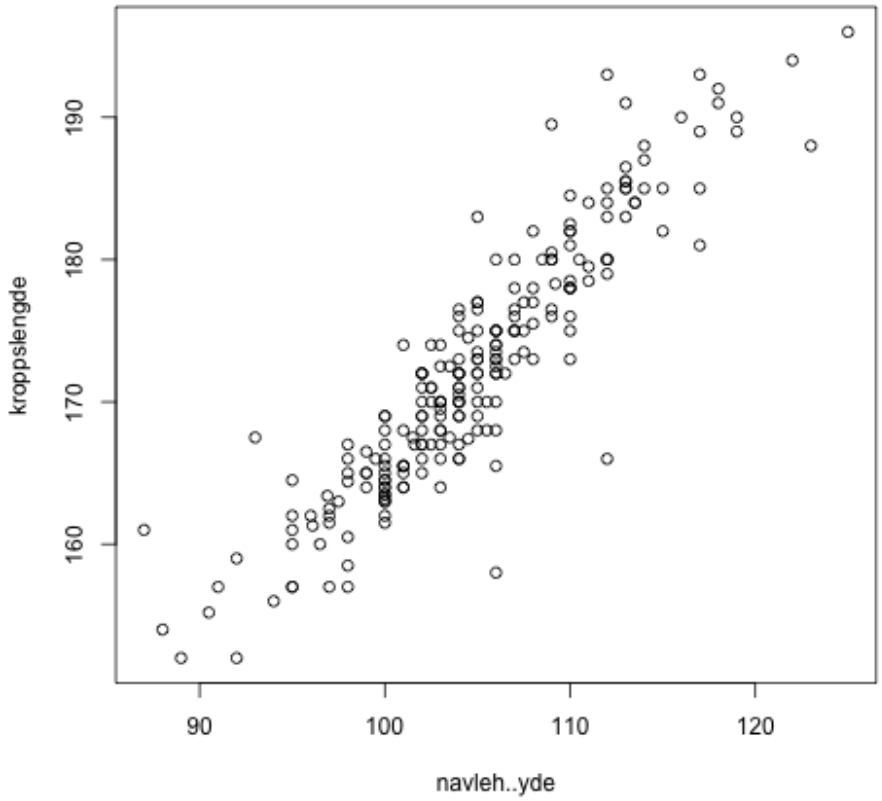
I denne oppgaven bruker jeg kommandoen `summary()`
150 kvinner og 73 menn deltok i forsøket

Tabell 2: Fem-punkts oppsummering av de fire variablene

	min	Q_1	μ	Q_3	max
kroppslengde	152.0	166.0	172.3	178.0	196.0
fot.navle	87.0	101.0	104.8	109.0	125.0
navle.isse	52.00	65.00	67.34	70.00	81.00
favn	146.0	165.0	172.4	180.0	202.0

b

I denne oppgaven bruker jeg kommandoen `plot()` i for å lage et spredningsplot.



Figur 6: Her er et plot av kroppslengde som funksjon av navlehøyde

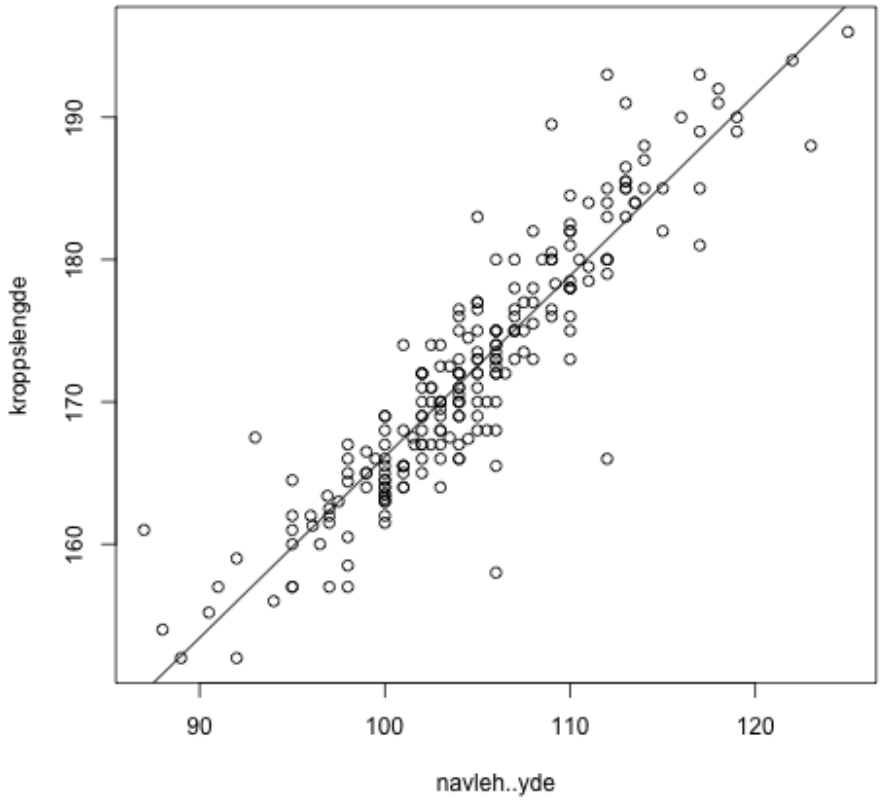
Man kan se at det er høy korrelasjon mellom kroppslengde og navlehøyde

c

I denne oppgaven bruker jeg kommandoen `cor()` for å få korrelasjonen. En korrelasjon på 0.9140397 antyder igjen at det er en god sammenheng mellom kroppslengde og navlehøyde.

d

Her er koden justert litt i forhold til det som ble oppgitt i oppgaveteksten



Figur 7: Lineær regresjon utført på kroppslengde som funksjon av navlehøyde

e

Tabell 3: Koeffisienter for modellen

Koeffisienter	a	b
	38.897	1.273

I denne oppgaven brukte jeg `print(fit)` for å få ut koeffisientene til den lineære tilpasningen av funksjonen. Det ser ut til at en økning på navlehøyde med én cm vil, i gjennomsnitt, føre til en høydeøkning på 1.273 cm. Stigningstallet er mer enn én, men fortsatt en konstant verdi (pga. den lineære tilpasningen).

f

Her er det rett og slett å bare ta produktet med det lineære forholdet vi fant i forrige oppgave mens man husker på å inkludere verdien hvor den lineære funksjonen skjærer y aksen. Altså funksjonsverdien til den lineære funksjonen med riktige koeffisienter a og b

$$f(121) = a + bx = 192.93$$

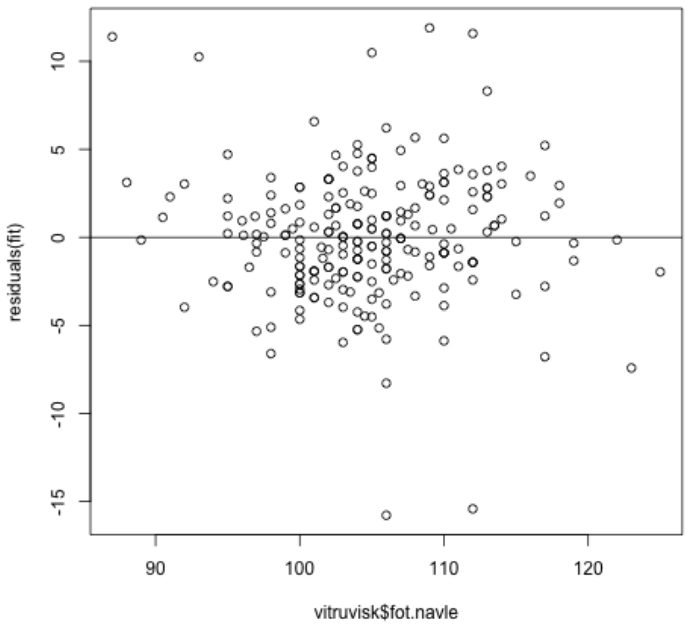
g

Tabell 4: ...

	min	Q_1	μ	Q_3	max
kroppslengde	152.0	166.0	172.3	178.0	196.0
fot.navle	87.0	101.0	104.8	109.0	125.0
navle.isse	52.00	65.00	67.34	70.00	81.00
favn	146.0	165.0	172.4	180.0	202.0

Fra tabellen kan man se at det er en viss andel av variasjonen i kroppslengde som forklares av navlehøyde. Dette samsvarer godt med det vi fant ut i forrige oppgave.

h



Figur 8: Plott over residualene av navlehøyde i forhold til den lineære modellen

Det vil alltid være noen uteligger for enhver rimelig regresjonstilpasning. Her ser det ut til at residualene er nogelunde jevnt fordelt, hvilket kan være et godt tegn. Mange datapunkter som ligger enten over eller under modellen vil aldri være tilfellet ettersom regresjonstilpasningen ville justert for dette. Det man kan prøve å se etter, derimot, er om det ser ut til at residualene følger en systematisk trend som ikke er jevnt fordelt. For eksempel, hvis man prøver en lineær tilpasning på et andregradspolynom, ville man kanskje sett at datapunktene i midten ligger under det modellen skulle tilsi mens datapunktene lenger ut ligger over. Men, som sagt, dette ser ikke ut til å være tilfellet her.

Appendix

3.1 Oppgave 1

```
# hard kodet input data
zebrafinch <- c(0.7562718,0.7841234,0.8535867,0.82394,0.7804905,0.9

# instansiering av essensielle statistiske storrelser til fordeling
mean(zebrafinch)          #the mean (0.8485003)
sd(zebrafinch)            #Standard deviation
median(zebrafinch)        #median
IQR(zebrafinch)           #Interquartile

# Denne linjen printer den standardiserte verdien til 0.8
print((0.8-mean(zebrafinch))/sd(zebrafinch))

# gjor klar et plot
png(file = "zebrafish.pdf")
# ploter et histogram av fordelingen
hist(zebrafinch,xlab = "Weight",col = "yellow",border = "blue")
# lagrer filen
dev.off()

qqnorm(zebrafinch) #lager qq plot
qqline(zebrafinch) #lager qq plot

# faar prosentil til 0.6
pnorm(0.6, mean = mean(zebrafinch), sd = sd(zebrafinch))
# faar prosentil for 'ikke 0.6'. ie
# ie. sannsynligheten for at en fink har mer enn 0.6 BMR
1 - pnorm(1, mean = mean(zebrafinch), sd = sd(zebrafinch))
```

3.2 Oppgave 2

```
# importerer data
data ="http://www.uio.no/studier/emner/matnat/math/STK1000/data/wol
# setter dataen inn i 'wolf'
wolf <- read.table(data,header=TRUE)
# summary gir femtallsoppsummering
summary(wolf)

#BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
# gjor klar et plot
png(file = "wolf_sex.pdf")
# her lages det to kolonner
par(mfrow=c(1,2))
# lager et kakediagram
pie(table(wolf$sex))
# lager et histogram
pie(table(wolf$population))
# lagrer filen
dev.off()

#CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
# setter den lett jaktede populasjonen inn i wolf.lett
wolf.lett <- wolf[wolf[,"population"]==1,]
# setter den tungt jaktede populasjonen inn i wolf.tungt
wolf.tungt <- wolf[wolf[,"population"]==2,]

#DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
# gjor klar et plot
png(file = "wolf_sammenlign.pdf")
# kommando som gjor at dette gaar fint
par(mfrow=c(2,2))
# plotter histogram av lett jaktet
hist(wolf.lett$cpmg)
# plotter boxplot av lett jaktet
boxplot(wolf.lett$cpmg)
# plotter histogram av tungt jaktet
hist(wolf.tungt$cpmg)
# plotter boxplot av tungt jaktet
boxplot(wolf.tungt$cpmg)
# lagrer filen
```

```
dev.off()

#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
# faar gjennomsnitt for lett
mean(wolf.lett$cpmg)
# faar medianen for lett
median(wolf.lett$cpmg)
# faar standardavvik for lett
sd(wolf.lett$cpmg)
# faar gjennomsnitt for tungt
mean(wolf.tungt$cpmg)
# faar gjennomsnitt for tungt
median(wolf.tungt$cpmg)
# faar standaravvik for tungt
sd(wolf.tungt$cpmg)

#GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
png(file = "TUNGT.pdf")
# faar qq plot av tungt jaktet
qqnorm(wolf.tungt$cpmg)
# teoretisk normalfordelt
qqline(wolf.tungt$cpmg)
# lagrer
dev.off()

png(file = "LETT.pdf")
# faar qq plot av lett jaktet
qqnorm(wolf.lett$cpmg)
# teoretisk normalfordelt
qqline(wolf.lett$cpmg)
# lagrer
dev.off()
```

3.3 Oppgave 3

```
#importerer
data = "http://www.uio.no/studier/emner/matnat/math/STK1000/data/vi
vitruvisk <- read.table(data,header=TRUE)

#AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
summary(vitruvisk)


#BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
# gjor klart for plotting
png(file = "3B.pdf")
#utforer plottet
plot(vitruvisk$fot.navle,vitruvisk$kroppslengde,xlab="navlehoyde",
ylab="kroppslengde")
# Lagrer filen
dev.off()


#CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
cor(vitruvisk$fot.navle,vitruvisk$kroppslengde)


#DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD

# gjor klart for plotting
png(file = "3D.pdf")
# utforer plottet
plot(vitruvisk$fot.navle,vitruvisk$kroppslengde,xlab="navlehoyde",
ylab="kroppslengde")
# gjor en lineaer tilpasning
fit <- lm(kroppslengde ~ fot.navle,data =  vitruvisk)
abline(fit)

# Save the file.
dev.off()



#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

```
# faar en oppsummering
summary(fit)
# printer tilpasningen
print(fit)


#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
# regner ut forventet verdi ut fra modellen
print(38.897 + 121*1.273)


#GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
# faar gjennomsnittet
summary(vitruvisk)


#HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
# gjor klart plottet
png(file = "3H.pdf")
# utforer plottet
plot(vitruvisk$fot.navle,residuals(fit))
abline(h=0)
# lagrer filen
dev.off()
```