# STK1000 – Obligatorisk oppgave 2 (bio)

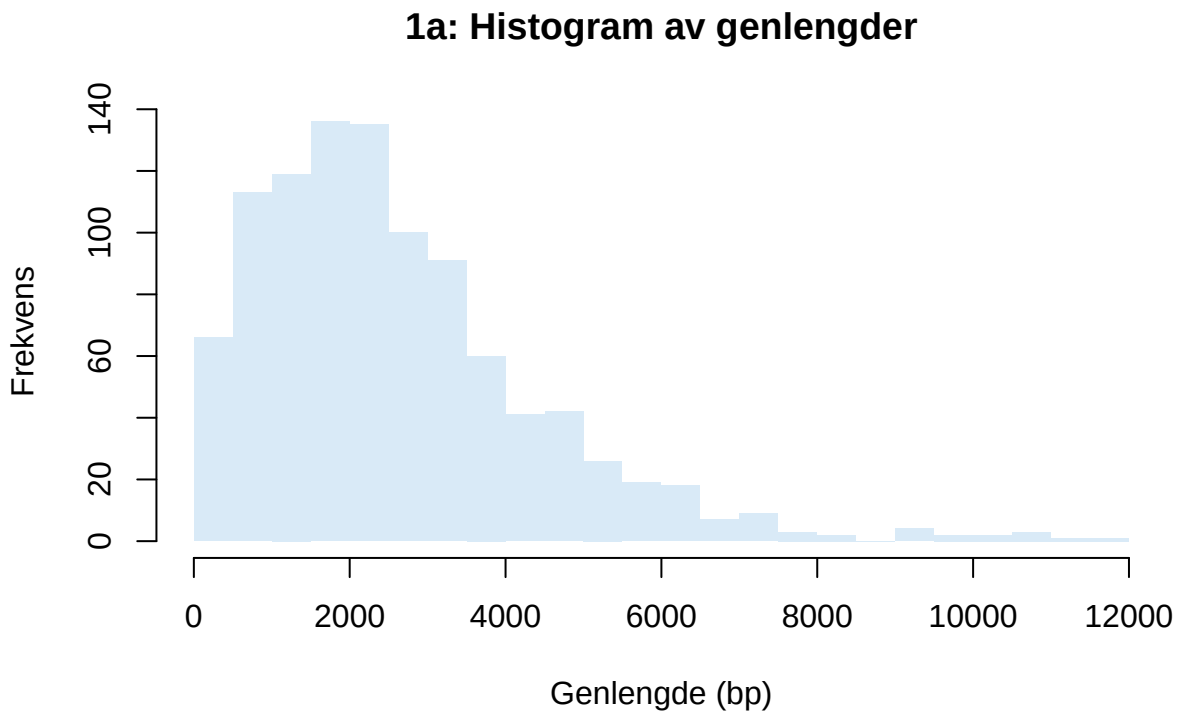
xxxxxxx-xxxxx

## Oppgave 1 – Utvalgsfordelinger (genlengder)

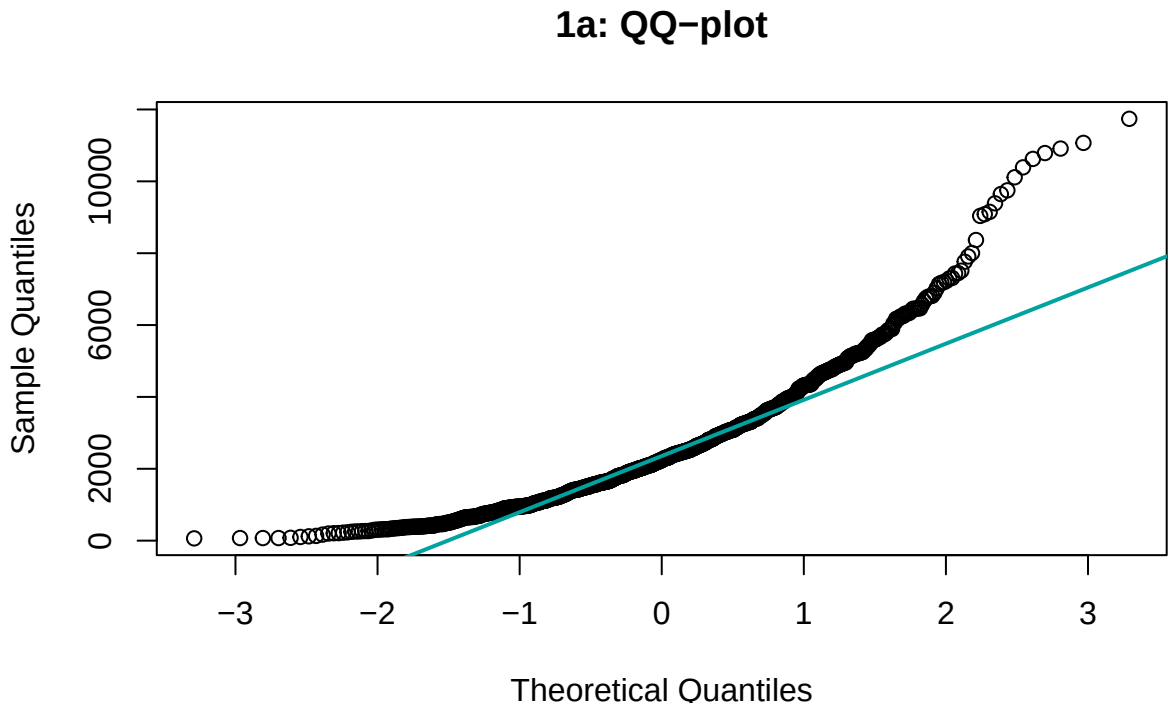
```
# Data
datapath <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/oblig2/gene.txt"
genes <- read.table(datapath, header=TRUE, sep=";")
Gene.Lengths <- genes$Gene.Lengths
```

### 1a) Fordeling i populasjonen

```
mu    <- mean(Gene.Lengths, na.rm=TRUE)    # gjennomsnitt
sigma <- sd(Gene.Lengths,   na.rm=TRUE)    # standardavvik
cat("1a) mu =", round(mu,2), "  sigma =", round(sigma,2), "\n")
1a) mu = 2610.39   sigma = 1817.44

hist(Gene.Lengths, breaks="FD", col=mck_lightblue, border=NA,
     main="1a: Histogram av genlengder", xlab="Genlengde (bp)", ylab="Frekvens")
```



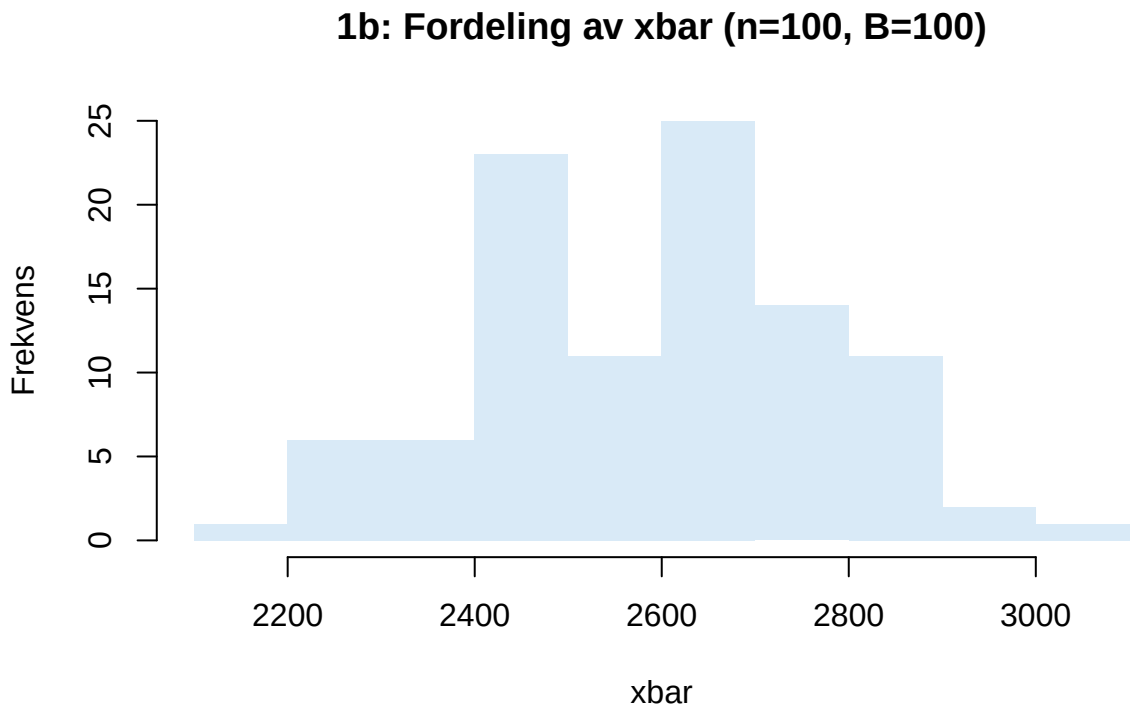

```
qqnorm(Gene.Lengths, main="1a: QQ-plot"); qqline(Gene.Lengths, col=mck_teal, lwd=2)
```



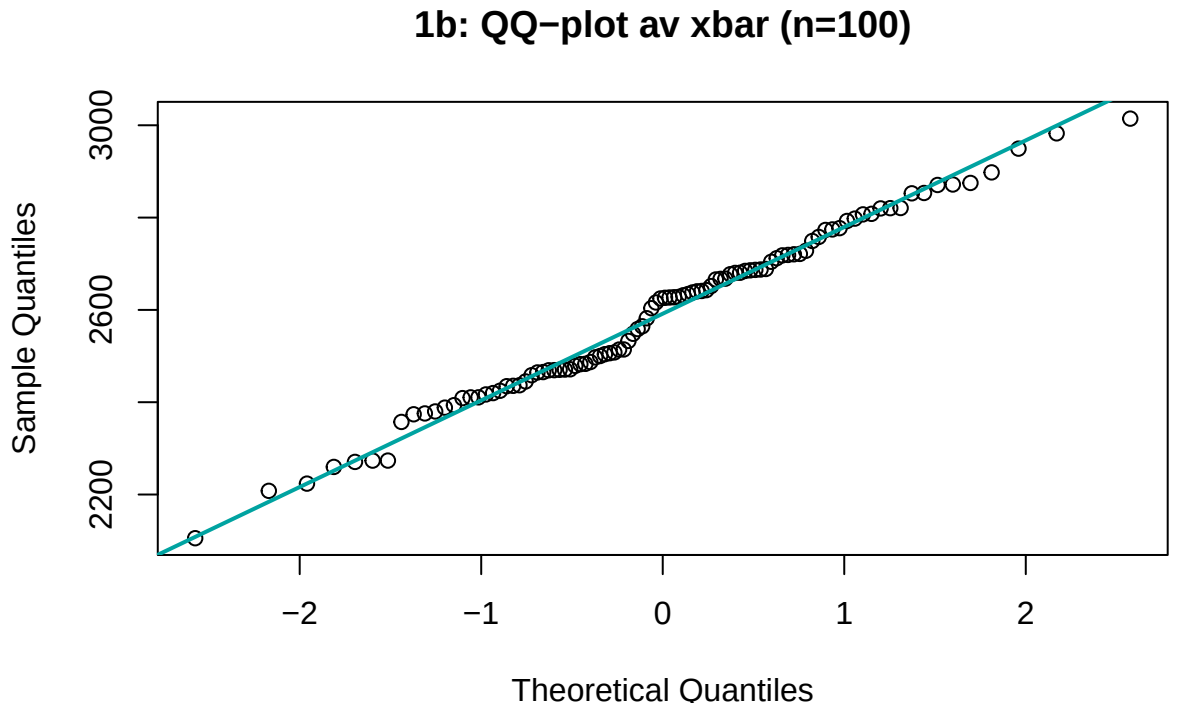

**Tolkning 1a:** Moderat høyreskjev fordeling; QQ-plot avviker i høyre hale. Normalantakelse er bare delvis rimelig. Nivå/spredning 2610.39 / 1817.44.

### 1b) Utvalg n = 100

```
set.seed(100)
x100_once <- sample(Gene.Lengths, 100, replace=TRUE)
xbar100_once <- mean(x100_once)
cat("1b) xbar (ett utvalg, n=100) =", round(xbar100_once,2), "\n")
1b) xbar (ett utvalg, n=100) = 2767.98

# 100 utvalg (n=100) i tråd med hint
meanvec <- rep(NA, 100)
for(i in 1:100){
  sample.now <- sample(Gene.Lengths, 100, replace=TRUE)
  meanvec[i] <- mean(sample.now)
}
cat("1b) mean(xbar) =", round(mean(meanvec),2),
    "  sd(xbar) =", round(sd(meanvec),2),
    "  teori sigma/sqrt(100) =", round(sigma/sqrt(100),2), "\n")
1b) mean(xbar) = 2591.2   sd(xbar) = 184.03   teori sigma/sqrt(100) = 181.74

hist(meanvec, breaks="FD", col=mck_lightblue, border=NA,
     main="1b: Fordeling av xbar (n=100, B=100)", xlab="xbar", ylab="Frekvens")
```




```
qqnorm(meanvec, main="1b: QQ-plot av xbar (n=100)"); qqline(meanvec, col=mck_teal, lwd=2)
```

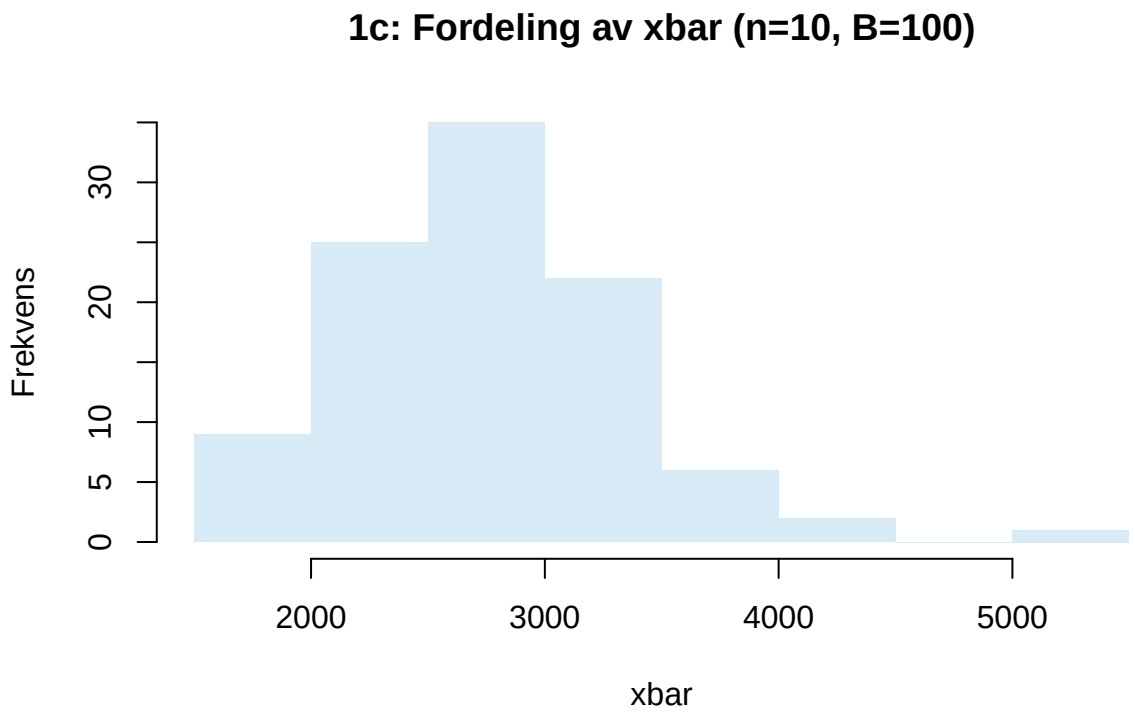



**Tolkning 1b:** $\bar{x}$ (n=100) er smal og tilnærmet normal. sd($\bar{x}$) `sigma/sqrt(100)` → stabile gjennomsnitt.

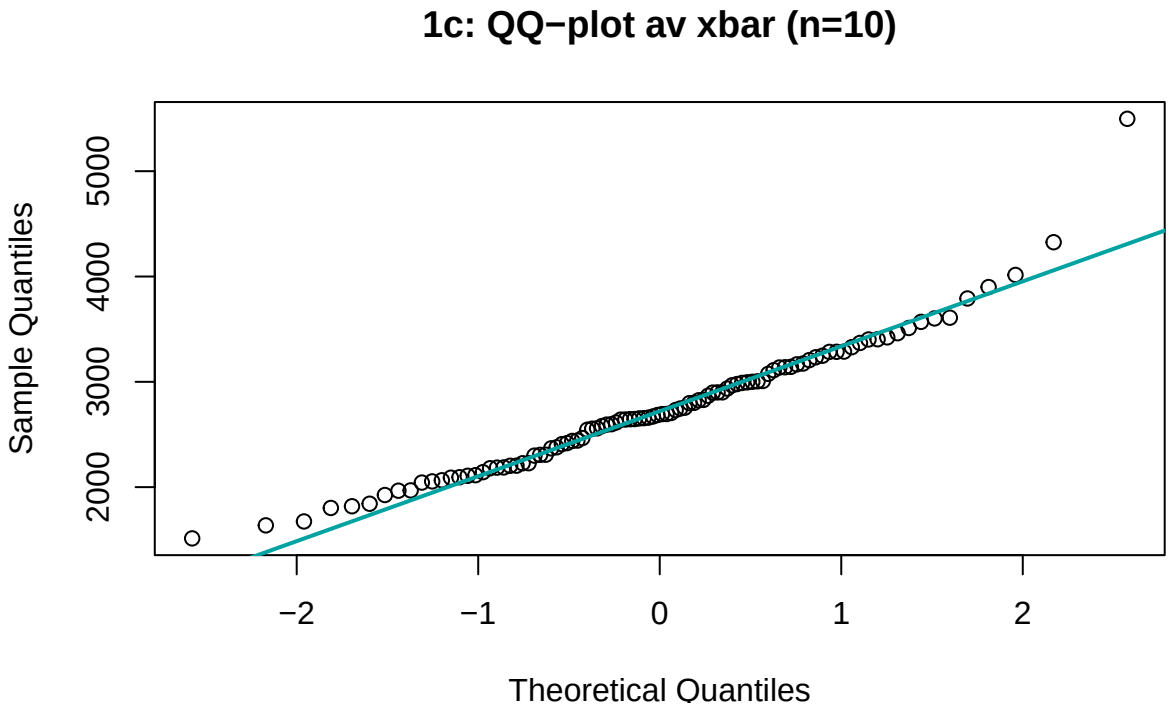
### 1c) Utvalg n = 10

```
set.seed(101)
meanvec10 <- rep(NA, 100)
for(i in 1:100){
```

```
  sample.now <- sample(Gene.Lengths, 10, replace=TRUE)
  meanvec10[i] <- mean(sample.now)
}
cat("1c) mean(xbar) =", round(mean(meanvec10),2),
    "  sd(xbar) =", round(sd(meanvec10),2),
    "  teori sigma/sqrt(10) =", round(sigma/sqrt(10),2), "\n")
1c) mean(xbar) = 2747.58   sd(xbar) = 620.58   teori sigma/sqrt(10) = 574.72

hist(meanvec10, breaks="FD", col=mck_lightblue, border=NA,
     main="1c: Fordeling av xbar (n=10, B=100)", xlab="xbar", ylab="Frekvens")
```




```
qqnorm(meanvec10, main="1c: QQ-plot av xbar (n=10)"); qqline(meanvec10, col=mck_teal, lwd=2)
```




**Tolkning 1c:** $\bar{x}$ (n=10) er bredere og mer variabel enn for n=100; sd($\bar{x}$) større som forventet.

### 1d) Sannsynlighet P($\bar{x} > 3000$)

```
p10  <- 1 - pnorm(3000, mean=mu, sd=sigma/sqrt(10))
p100 <- 1 - pnorm(3000, mean=mu, sd=sigma/sqrt(100))
cat("1d) P(xbar > 3000): n=10  =", signif(p10,3), "\n")
1d) P(xbar > 3000): n=10  = 0.249
cat("1d) P(xbar > 3000): n=100 =", signif(p100,3), "\n")
1d) P(xbar > 3000): n=100 = 0.016
```

**Tolkning 1d:** Sannsynligheten blir mye mindre når n øker ($\bar{x}$ tettere rundt ).

### 1e) Bias og varians

**Kort:** $\bar{x}$ er **ubiasert** for  (bias=0). **Var($\bar{x}$)= $^2$/n** → større n gir mer presise anslag.

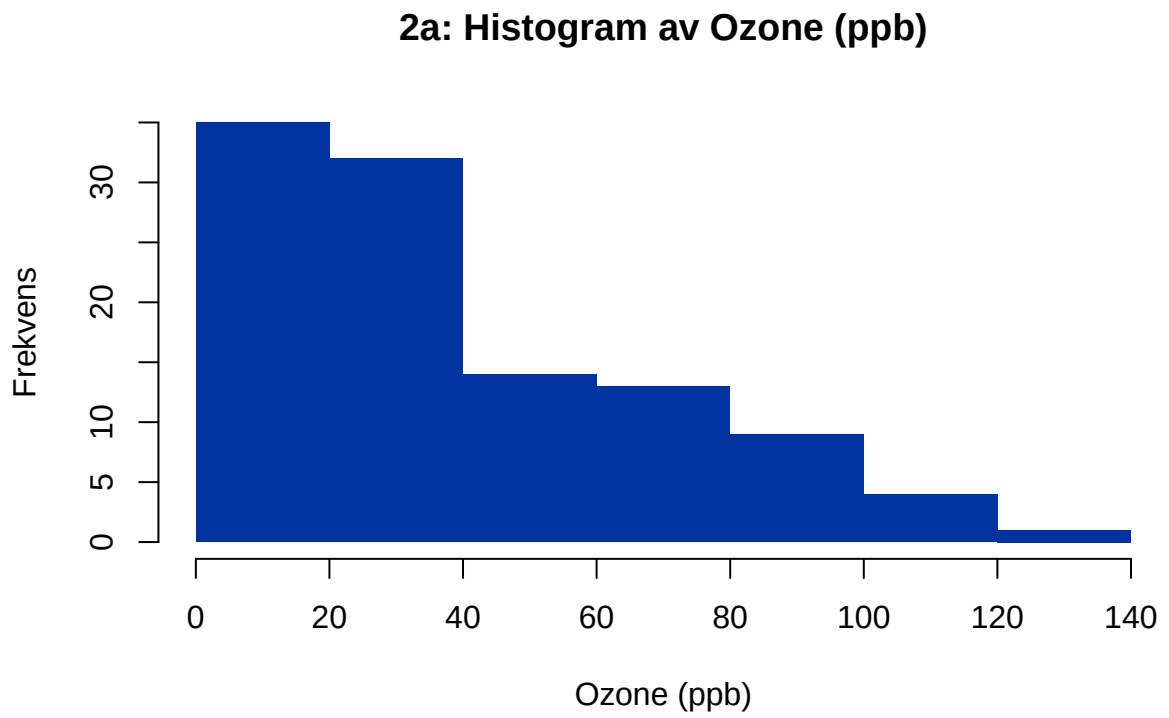
---

## Oppgave 2 – Signifikanstester og konfidensintervaller (ozon NYC 1973)

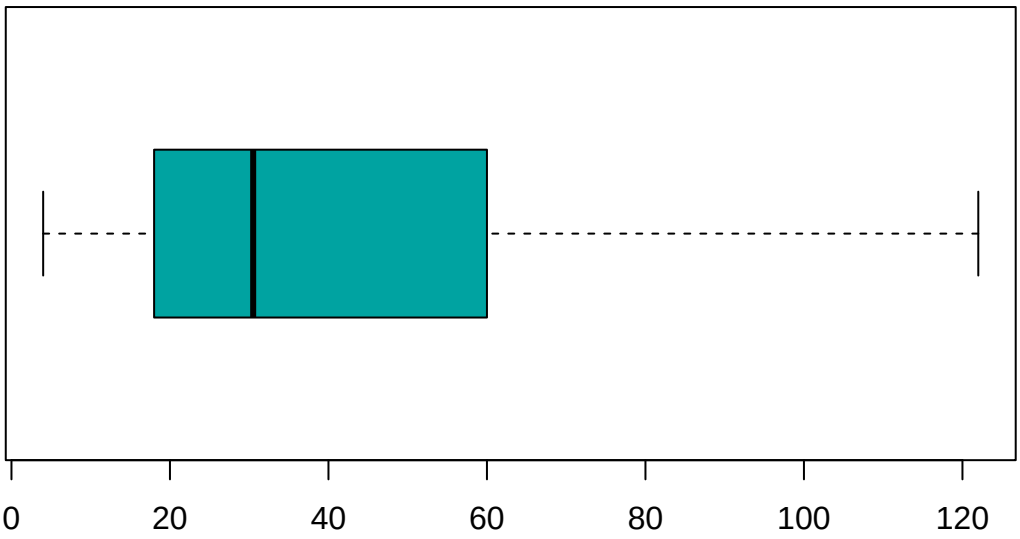
```
# Data
datapath2 <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/oblig2/ozone.txt"
newyork <- read.table(datapath2, header=TRUE)
Ozone <- newyork$Ozone
Temp  <- newyork$Temp
Month <- newyork$Month
Day   <- newyork$Day
```

### 2a) Oppsummering + antakelser

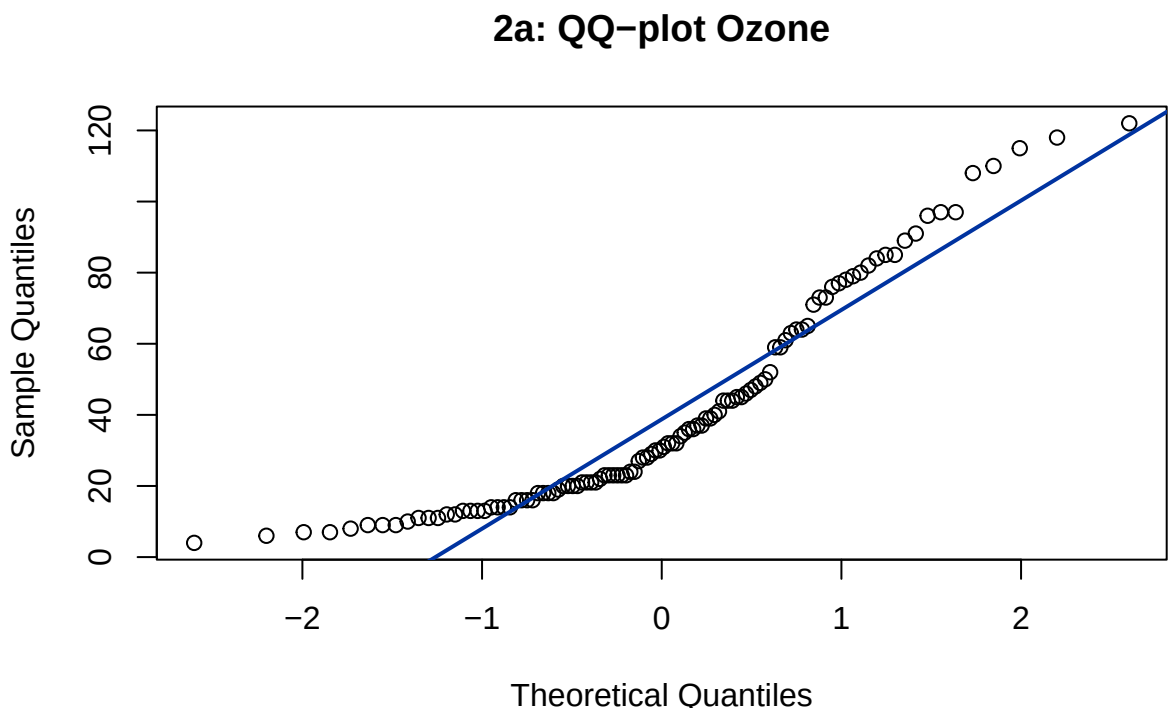
```
mean_O3   <- mean(Ozone, na.rm=TRUE)
median_O3 <- median(Ozone, na.rm=TRUE)
sd_O3     <- sd(Ozone, na.rm=TRUE)
iqr_O3    <- IQR(Ozone, na.rm=TRUE)
summary(Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   18.00   30.50   40.45   59.50  122.00

hist(Ozone, breaks="FD", col=mck_blue, border=NA,
     main="2a: Histogram av Ozone (ppb)", xlab="Ozone (ppb)", ylab="Frekvens")
```




```
boxplot(Ozone, horizontal=TRUE, col=mck_teal, main="2a: Boxplot Ozone")
```




```
qqnorm(Ozone, main="2a: QQ-plot Ozone"); qqline(Ozone, col=mck_blue, lwd=2)
```




```
cat("2a) mean=", round(mean_O3,1), " median=", round(median_O3,1),
    " sd=", round(sd_O3,1), " IQR=", round(iqr_O3,1), "\n")
2a) mean= 40.5  median= 30.5  sd= 29.8  IQR= 41.5
```

**Tolkning 2a:** Høyreskjev fordeling med uteliggere; normalantakelsen er svak.
**Antakelser t-prosedyre:** uavhengige observasjoner, omtrent normalitet på gjennomsnitt (CLT hjelper), ingen ekstreme avvik som dominerer. **t (ikke z)** fordi er ukjent.

### 2b) Test mot 30 ppb + 90 % og 99 % KI

```
# H0: mu = 30, H1: mu > 30 (ensidig)
t_greater_90 <- t.test(Ozone, mu=30, alternative="greater", conf.level=0.90)
t_greater_99 <- t.test(Ozone, mu=30, alternative="greater", conf.level=0.99)
t_greater_90

    One Sample t-test

data:  Ozone
t = 3.6395, df = 107, p-value = 0.0002112
alternative hypothesis: true mean is greater than 30
90 percent confidence interval:
 36.74982      Inf
sample estimates:
mean of x
  40.4537
t_greater_99

    One Sample t-test

data:  Ozone
t = 3.6395, df = 107, p-value = 0.0002112
alternative hypothesis: true mean is greater than 30
99 percent confidence interval:
 33.67016      Inf
sample estimates:
mean of x
  40.4537

cat("2b) ensidig p-verdi =", signif(t_greater_90$p.value,3), "\n")
2b) ensidig p-verdi = 0.000211
cat("2b) 90% KI =", paste(round(t_greater_90$conf.int,1), collapse=" ; "), "\n")
2b) 90% KI = 36.7 ; Inf
cat("2b) 99% KI =", paste(round(t_greater_99$conf.int,1), collapse=" ; "), "\n")
2b) 99% KI = 33.7 ; Inf
```

**Tolkning 2b:** Konkluder om >30 ut fra p-verdi. 99 % KI er bredere enn 90 %; sjekk om 30 er utenfor.

### 2c) Juli/august vs. mai/juni/september

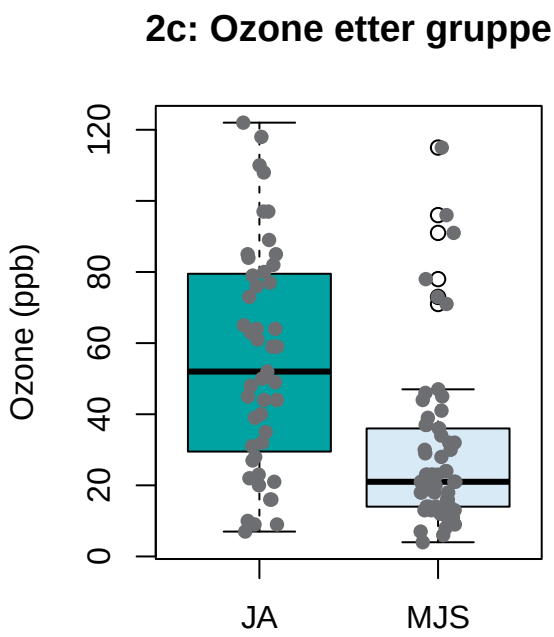
```
# Del etter måned (i tråd med hint)
oz.juli.august   <- newyork[newyork$Month %in% c(7,8), "Ozone"]
oz.mai.juni.sept <- newyork[newyork$Month %in% c(5,6,9), "Ozone"]

# Welch to-utvalgs t-test
tt2 <- t.test(oz.juli.august, oz.mai.juni.sept)
tt2

    Welch Two Sample t-test

data:  oz.juli.august and oz.mai.juni.sept
t = 4.9526, df = 80.631, p-value = 3.957e-06
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 16.06035 37.63270
sample estimates:
mean of x mean of y
 55.61702  28.77049
```

```r
m_JA  <- mean(oz.juli.august,   na.rm=TRUE)
m_MJS <- mean(oz.mai.juni.sept, na.rm=TRUE)
cat("2c) mean(JA) =", round(m_JA,1),
    "  mean(MJS) =", round(m_MJS,1),
    "  p-verdi =", signif(tt2$p.value,3), "\n")
```

```
2c) mean(JA) = 55.6   mean(MJS) = 28.8   p-verdi = 3.96e-06
```

```r
# --- 2c: figur (boksplot + punkter) ---
par(mfrow = c(1, 2))

boxplot(
  oz.juli.august, oz.mai.juni.sept,
  names = c("JA", "MJS"),
  col = c(mck_teal, mck_lightblue),
  main = "2c: Ozone etter gruppe",
  ylab = "Ozone (ppb)"
)

stripchart(
  list(JA = oz.juli.august, MJS = oz.mai.juni.sept),
  vertical = TRUE, method = "jitter",
  pch = 16, col = mck_grey, add = TRUE
)

par(mfrow = c(1, 1))
```



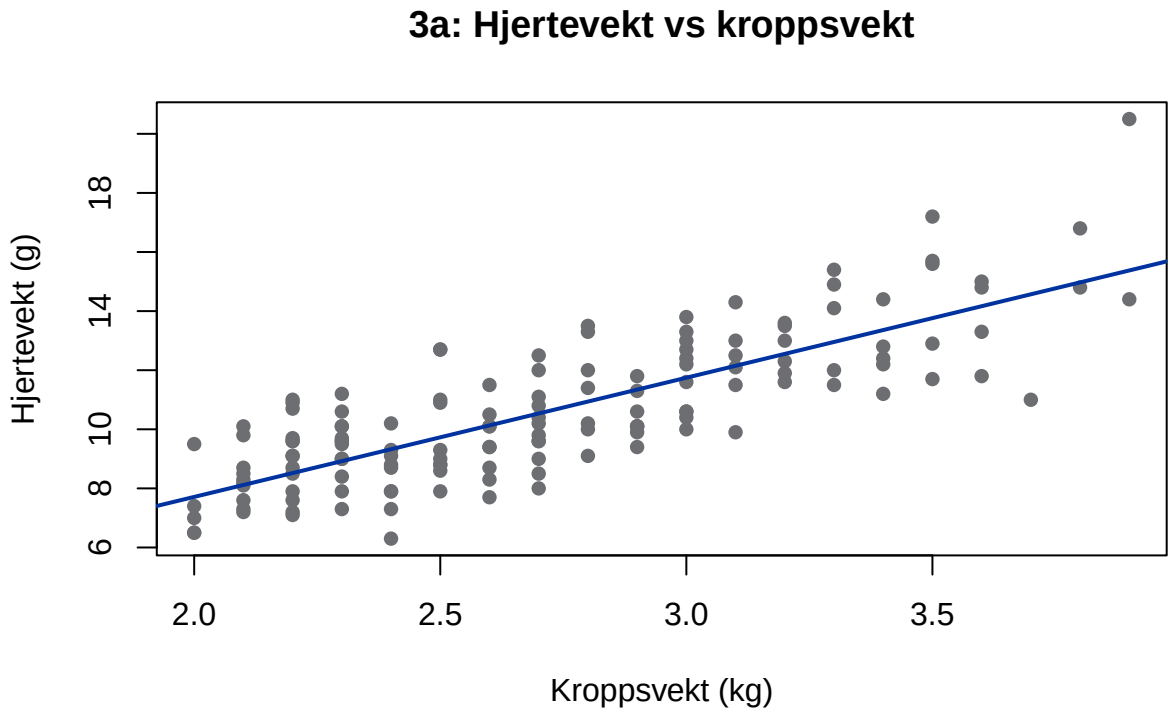

**Tolkning 2c:** Si om forskjellen er **signifikant** ($p<0.05$) og **hvilken gruppe** som er høyest (JA vs. MJS).

## Oppgave 3 – Lineær regresjon (katter: Bwt → Hwt)

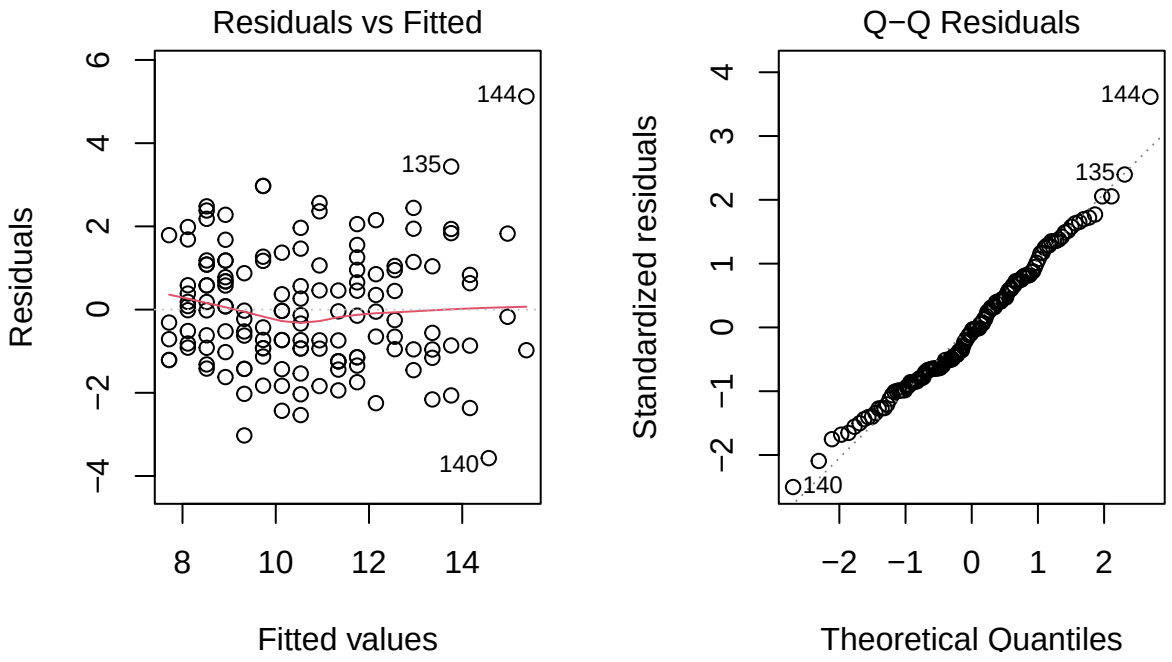
```
if (!requireNamespace("MASS", quietly=TRUE)) install.packages("MASS")
library(MASS)
dat <- MASS::cats
bwt <- dat$Bwt
hwt <- dat$Hwt
```

### 3a) Spredningsplott + modell + antagelser

```
plot(bwt, hwt, pch=16, col=mck_grey,
     xlab="Kroppsvekt (kg)", ylab="Hjertevekt (g)",
     main="3a: Hjertevekt vs kroppsvekt")
fit <- lm(hwt ~ bwt)
abline(fit, col=mck_blue, lwd=2)
```




```
# Enkel diagnostikk (antagelser)
par(mfrow=c(1,2))
plot(fit, which=1)
plot(fit, which=2)
```




```
par(mfrow=c(1,1))
```

**Tolkning 3a:** Klar positiv trend; rett linje virker rimelig. Residual-plot uten sterk kurve/heteroskedastisitet og omtrent rett QQ-plot støtter antagelsene greit.

### 3b) Skjæringspunkt og stigningstall + 95 % KI

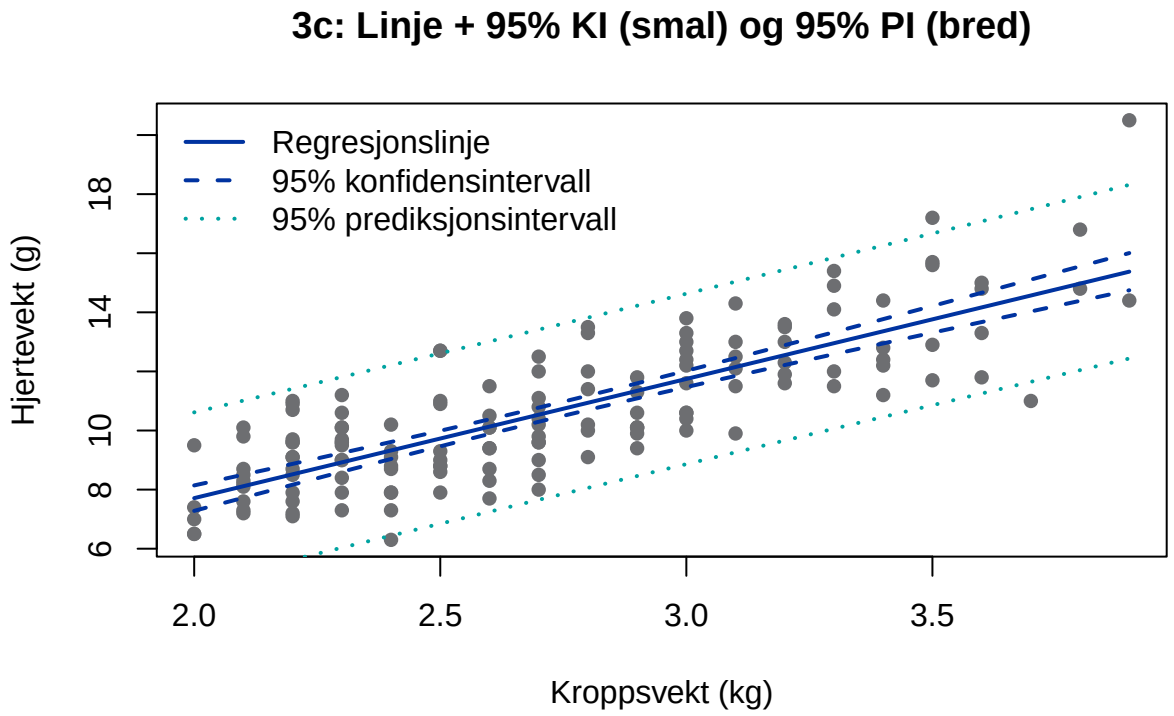
```
sfit <- summary(fit)
b0 <- sfit$coefficients[1,1]                # intercept
b1 <- sfit$coefficients[2,1]                # slope
se.b1 <- sfit$coefficients[2,2]
df <- fit$df.residual

lower <- b1 + qt(0.025, df) * se.b1
upper <- b1 + qt(0.975, df) * se.b1

cat("3b) Intercept =", round(b0,2), "  Slope =", round(b1,2), "\n")
3b) Intercept = -0.36   Slope = 4.03
cat("3b) p-verdi for slope =", signif(sfit$coefficients[2,4],3), "\n")
3b) p-verdi for slope = 6.97e-34
cat("3b) 95% KI for slope = [", round(lower,2), ", ", round(upper,2), "]\n", sep="")
3b) 95% KI for slope = [3.54, 4.53]
```

**Tolkning 3b:** Stigningstallet er økningen i forventet hjertevekt (g) ved +1 kg kroppsvekt. Veldig lav p-verdi → forkast H0: slope=0  klar lineær sammenheng.

### 3c) Prediksjons- og konfidensintervall + prediksjon ved 3.0 kg

```
xval <- seq(min(bwt), max(bwt), length.out=200)
newd <- data.frame(bwt = xval)
pi <- predict(fit, newdata=newd, interval="prediction")  # prediksjon (bred)
ci <- predict(fit, newdata=newd, interval="confidence")  # konfidens (smal)
```

```r
plot(bwt, hwt, pch=16, col=mck_grey,
     xlab="Kroppsvekt (kg)", ylab="Hjertevekt (g)",
     main="3c: Linje + 95% KI (smal) og 95% PI (bred)")
lines(xval, ci[,"fit"], col=mck_blue, lwd=2)
lines(xval, ci[,"lwr"], col=mck_blue, lty=2, lwd=2)
lines(xval, ci[,"upr"], col=mck_blue, lty=2, lwd=2)
lines(xval, pi[,"lwr"], col=mck_teal, lty=3, lwd=2)
lines(xval, pi[,"upr"], col=mck_teal, lty=3, lwd=2)
legend("topleft",
       legend=c("Regresjonslinje","95% konfidensintervall","95% prediksjonsintervall"),
       col=c(mck_blue, mck_blue, mck_teal), lty=c(1,2,3), lwd=2, bty="n")
```




```r
pred_3 <- predict(fit, newdata=data.frame(bwt=3.0), interval="prediction")
cat("3c) Predikert hjertevekt ved 3.0 kg = ",
    round(pred_3[1],2), " g,  95% PI = [",
    round(pred_3[2],2), ", ", round(pred_3[3],2), "]\n", sep="")
```

```
3c) Predikert hjertevekt ved 3.0 kg = 11.75 g,  95% PI = [8.86, 14.63]
```

**Tolkning 3c:** KI gjelder forventet respons (linja). PI gjelder én katt (bredere pga. individvariasjon). Se tall for 3.0 kg over.