STK1000 – Obligatorisk innlevering 2

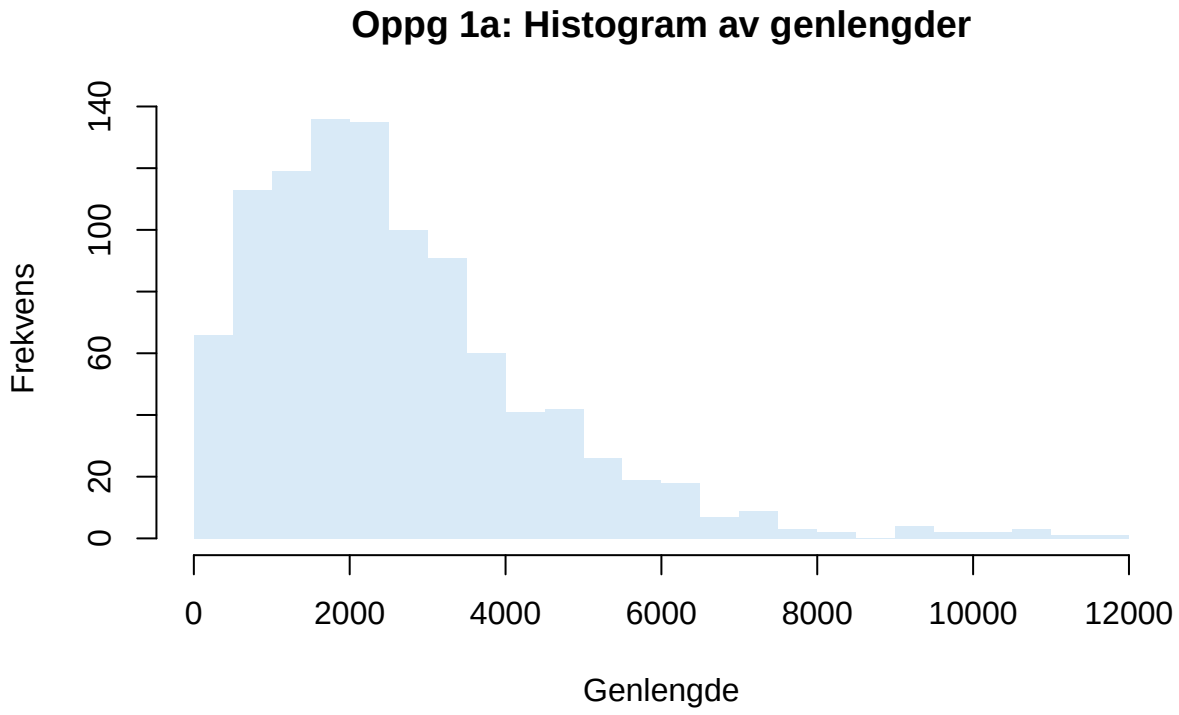
xxxxxxx-xxxxx

Oppgave 1 – Genlengder (populasjon og utvalgsfordeling)

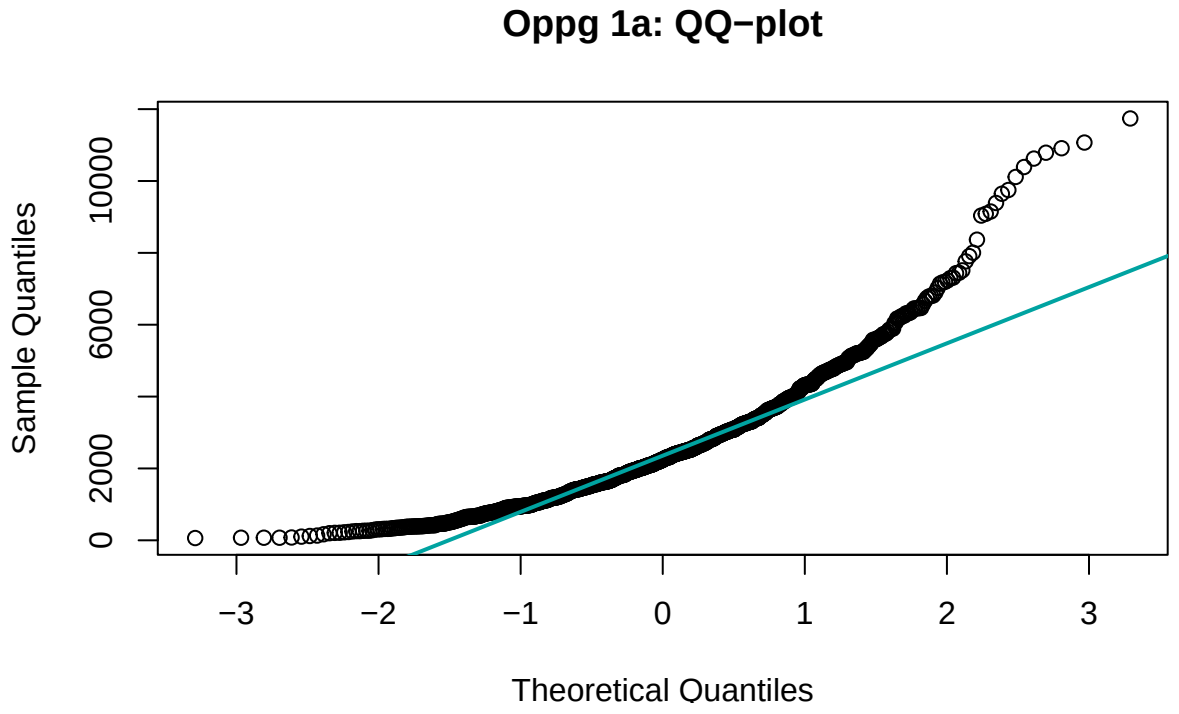
```
# Farger + plot-font
mck_blue      <- "#0033A0"
mck_teal      <- "#00A3A1"
mck_grey      <- "#6D6E71"
mck_lightblue <- "#D9EAF7"
op <- par(family = "Arial")

# 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) Populasjon: mu og sigma + plott

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

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



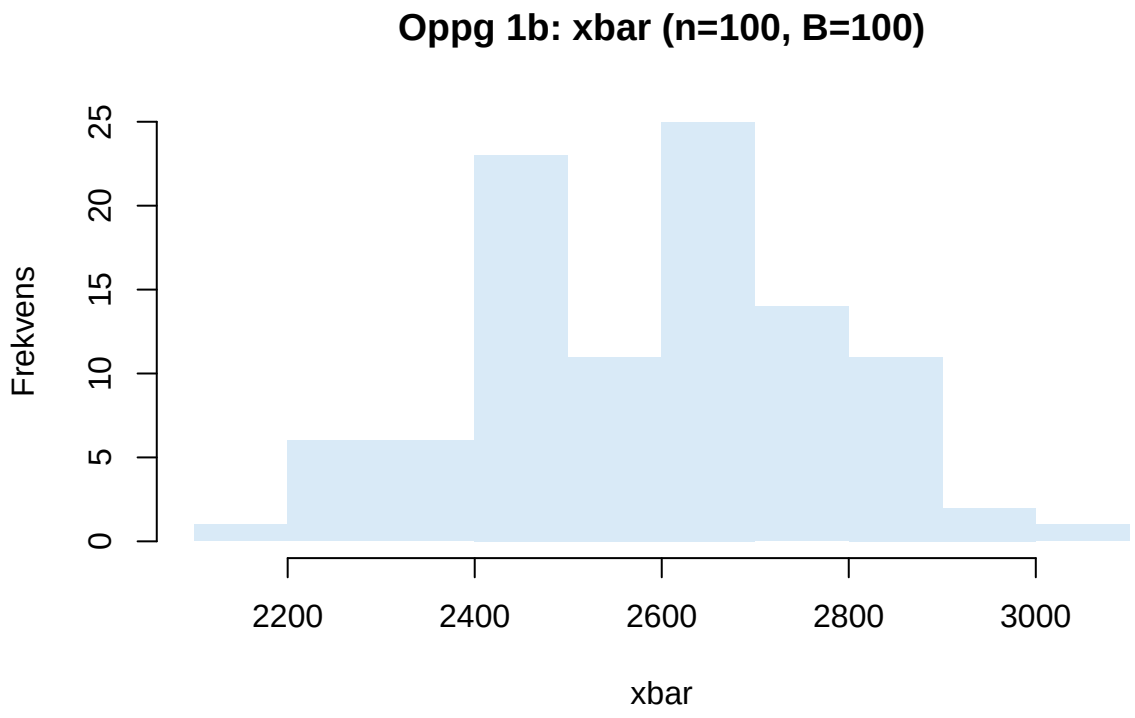

```r
qqnorm(Gene.Lengths, main="Oppg 1a: QQ-plot")
qqline(Gene.Lengths, col=mck_teal, lwd=2)
```



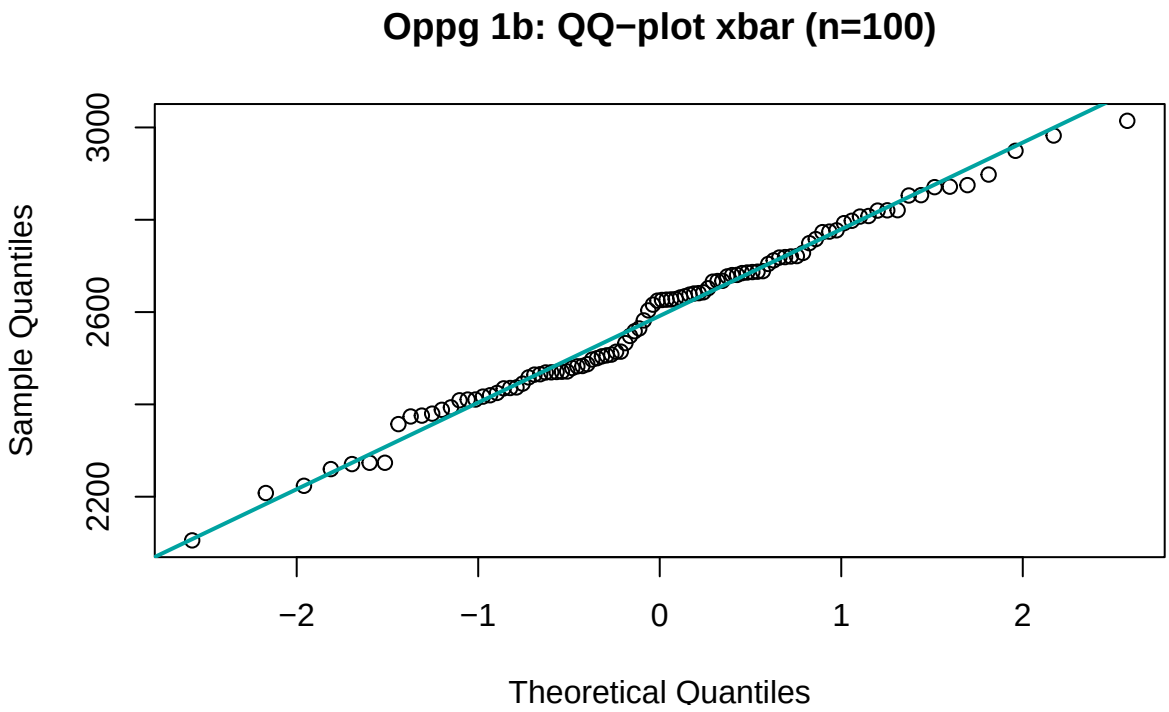
Tolkning 1a: Fordelingen er moderat høyreskjev (hale mot høyre). QQ-plottet viser avvik i høyre hale, så normalantakelsen er bare delvis rimelig. Nivå og spredning er ca. 2610.39 og 1817.44.

1b) Utvalg n=100: ett $\bar{x}$ + fordeling av $\bar{x}$

```
set.seed(100)
samp100_once <- sample(Gene.Lengths, 100, replace=TRUE)
xbar100_once <- mean(samp100_once)
cat("1b) Ett utvalg n=100, xbar =", round(xbar100_once,2), "\n")
1b) Ett utvalg n=100, xbar = 2767.98

B <- 100
means100 <- replicate(B, mean(sample(Gene.Lengths, 100, replace=TRUE)))

cat("1b) mean(xbar) =", round(mean(means100),2),
    " sd(xbar) =", round(sd(means100),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(means100, breaks="FD", col=mck_lightblue, border=NA,
     main="Oppg 1b: xbar (n=100, B=100)", xlab="xbar", ylab="Frekvens")
```



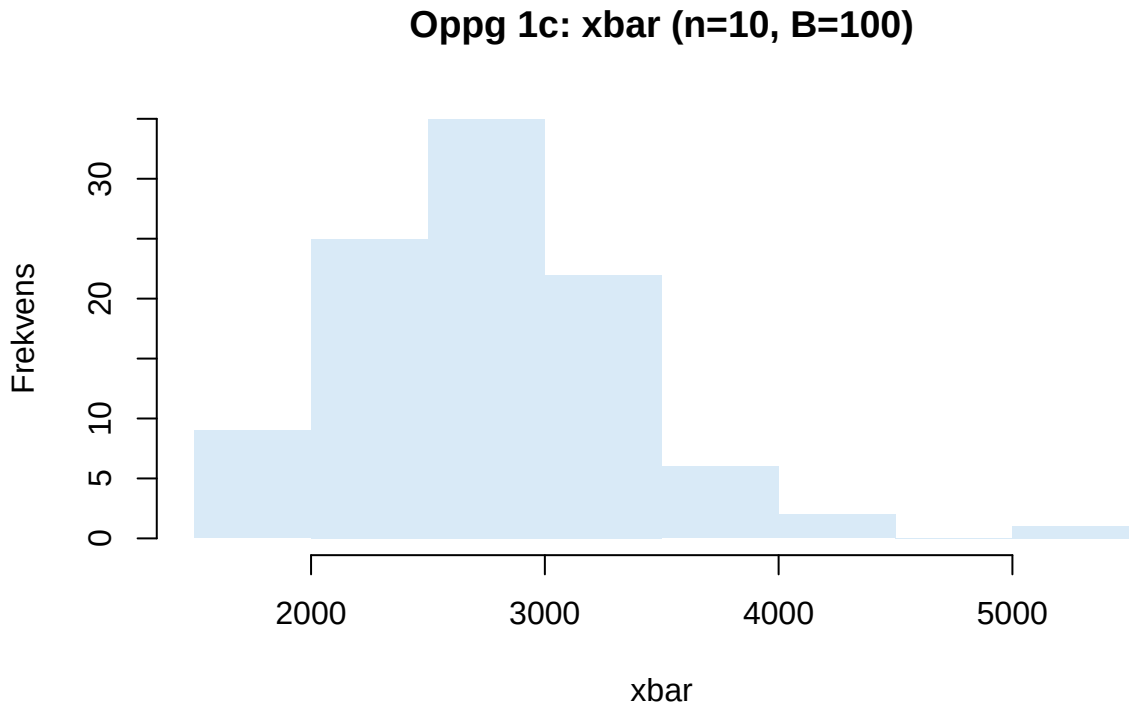

```
qqnorm(means100, main="Oppg 1b: QQ-plot xbar (n=100)")
qqline(means100, col=mck_teal, lwd=2)
```



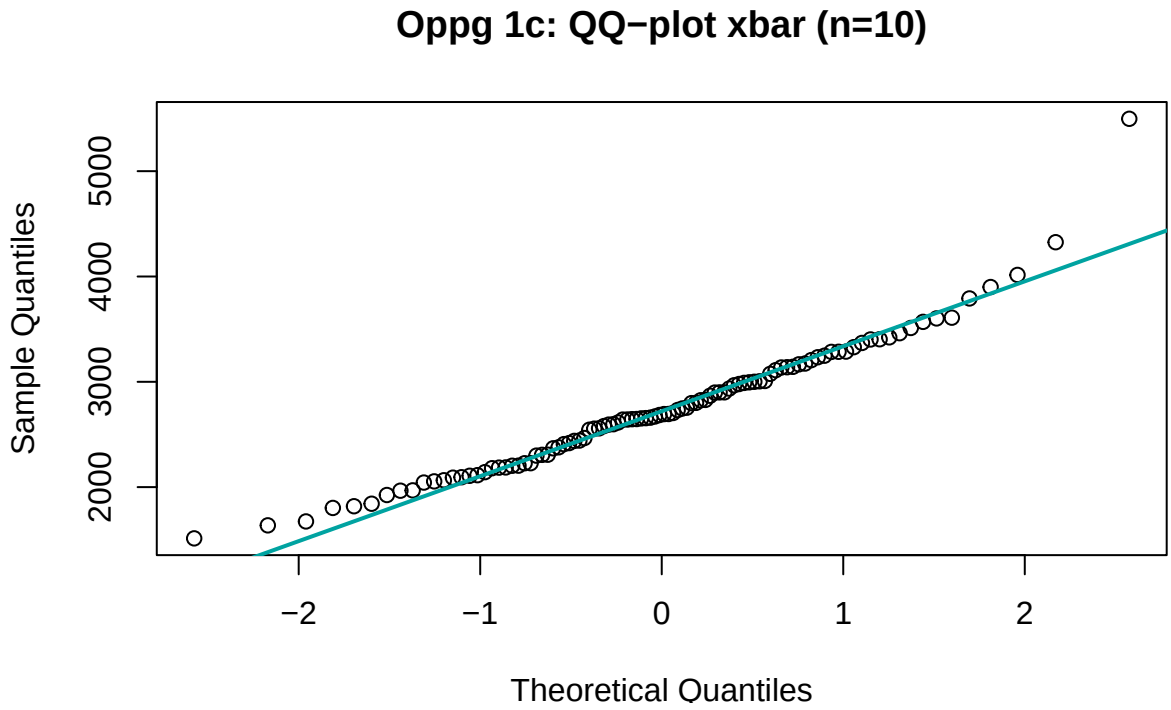
Tolkning 1b: $\bar{x}$ (n=100) er smal og tilnærmet normal. `sd(x̄)=` 184.03 ligger nær teori 181.74, altså stabile gjennomsnitt.

1c) Utvalg n=10: fordeling av $\bar{x}$

```
set.seed(101)
B <- 100
means10 <- replicate(B, mean(sample(Gene.Lengths, 10, replace=TRUE)))

cat("1c) mean(xbar) =", round(mean(means10),2),
    " sd(xbar) =", round(sd(means10),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(means10, breaks="FD", col=mck_lightblue, border=NA,
     main="Oppg 1c: xbar (n=10, B=100)", xlab="xbar", ylab="Frekvens")
```




```
qqnorm(means10, main="Oppg 1c: QQ-plot xbar (n=10)")
qqline(means10, col=mck_teal, lwd=2)
```



Tolkning 1c: $\bar{x}$ (n=10) er bredere og mer variabel enn for n=100; `sd(x̄)=` 620.58 > 181.74. Mindre utvalg gir mer variasjon.

1d) P($\bar{x} > 3000$) via CLT

```
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, fordi $\bar{x}$ samler seg tettere rundt .

1e) Bias og varians for $\bar{x}$

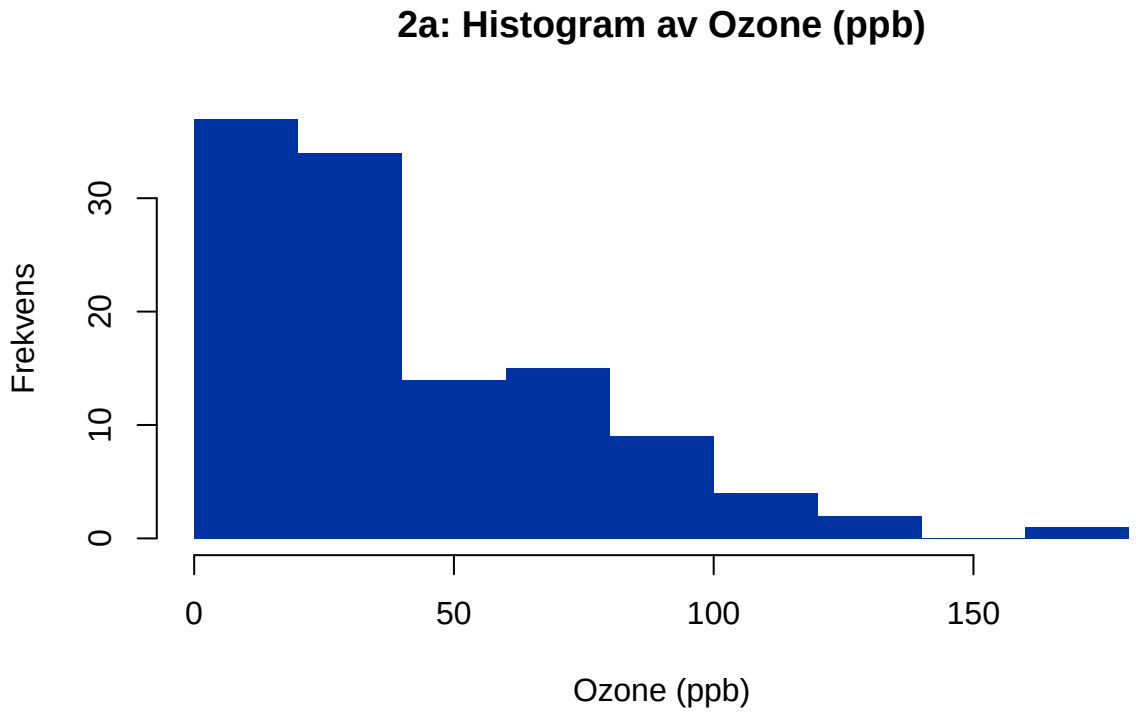
```
cat("1e) xbar er ubiasert (bias=0). Varians(xbar)=sigma^2/n: større n gir mer presise anslag.\n")
1e) xbar er ubiasert (bias=0). Varians(xbar)=sigma^2/n: større n gir mer presise anslag.
par(op)
```

Oppgave 2 – Ozon i NYC (mai–sept 1973)

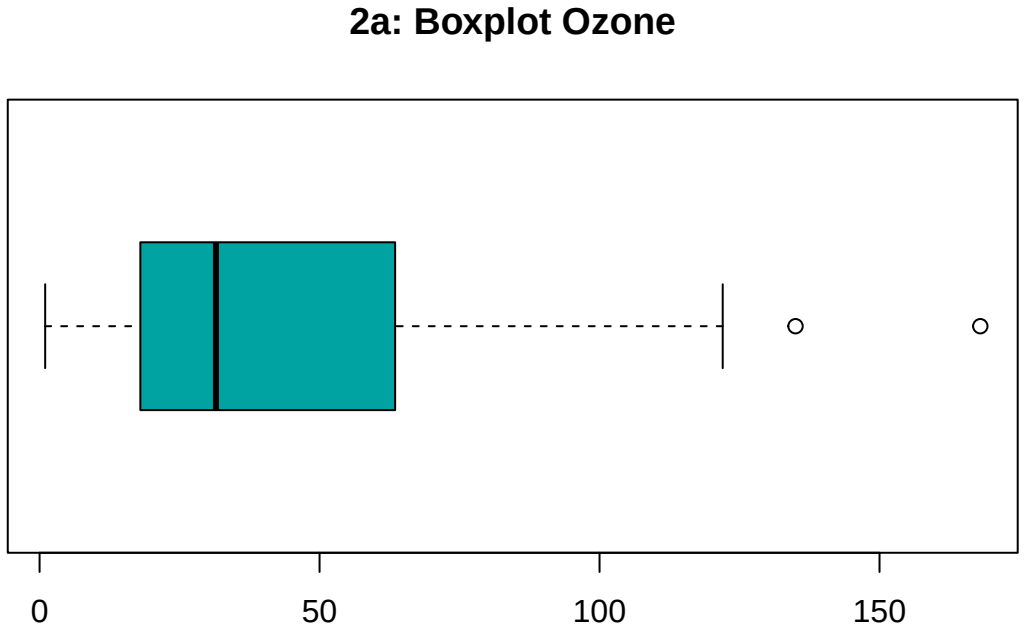
```
mck_blue      <- "#0033A0"
mck_teal      <- "#00A3A1"
mck_grey      <- "#6D6E71"
mck_lightblue <- "#D9EAF7"
op <- par(family = "Arial")

aq    <- airquality
O3    <- aq$Ozone
Temp  <- aq$Temp
Month <- aq$Month
Day   <- aq$Day
```

2a) Oppsummer tall og figurer

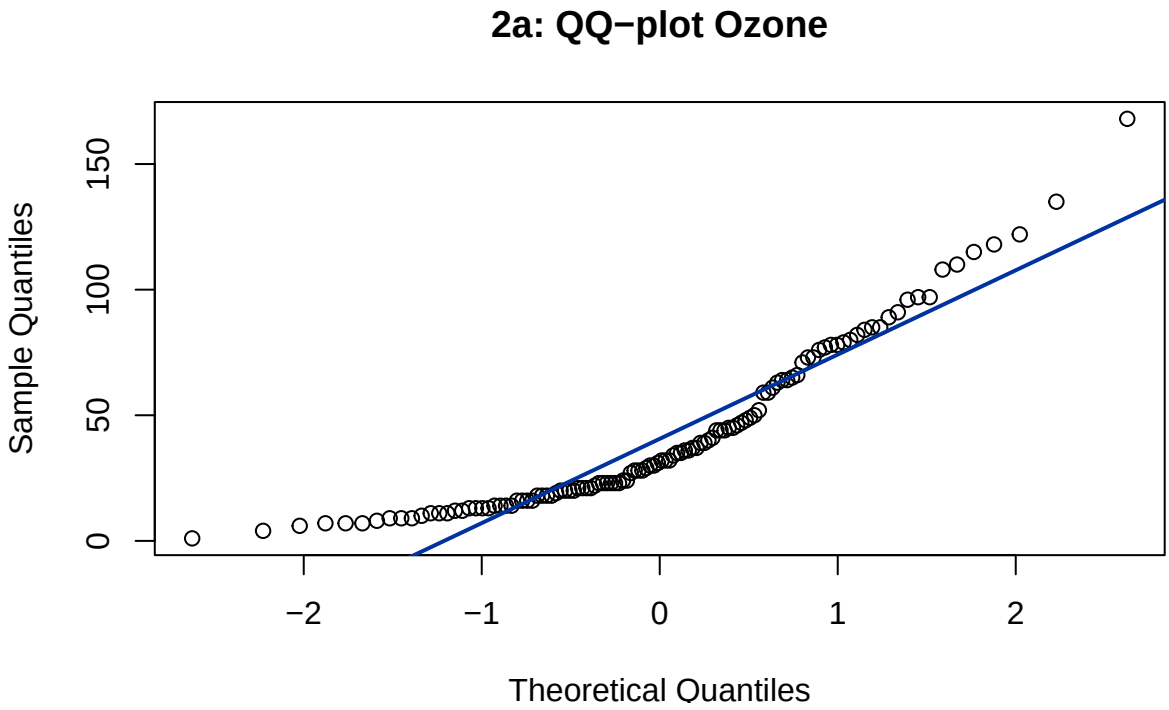
```
mean_O3   <- mean(O3, na.rm=TRUE)
median_O3 <- median(O3, na.rm=TRUE)
sd_O3     <- sd(O3, na.rm=TRUE)
iqr_O3    <- IQR(O3, na.rm=TRUE)

mean_O3; median_O3; sd_O3; iqr_O3
[1] 42.12931
[1] 31.5
[1] 32.98788
[1] 45.25
summary(O3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00   18.00   31.50   42.13   63.25  168.00      37

hist(O3, breaks="FD", col=mck_blue, border=NA,
     main="2a: Histogram av Ozone (ppb)",
     xlab="Ozone (ppb)", ylab="Frekvens")
```




```
boxplot(O3, horizontal=TRUE, col=mck_teal, main="2a: Boxplot Ozone")
```




```
qqnorm(O3, main="2a: QQ-plot Ozone"); qqline(O3, col=mck_blue, lwd=2)
```



Tolkning 2a: Ozone er høyreskjev med noen uteliggere. Gjennomsnitt 42.1, median 31.5, sd 33, IQR 45.2. Normalantakelsen er svak, men utvalgsstørrelsen gjør t-metoder greie med litt forsiktighet.

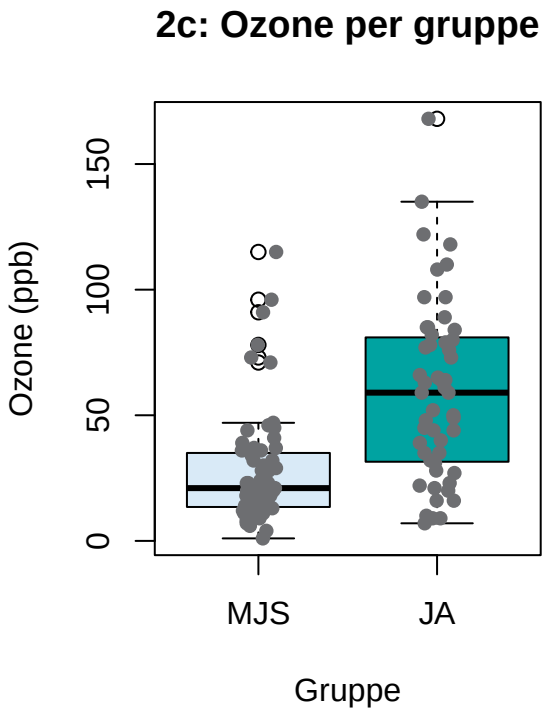
2b) Test H0: =30 (H1: >30) + 90% og 99% KI

```
t_greater_90 <- t.test(O3, mu=30, alternative="greater", conf.level=0.90)
t_greater_99 <- t.test(O3, mu=30, alternative="greater", conf.level=0.99)
t_greater_90

    One Sample t-test

data:  O3
t = 3.9601, df = 115, p-value = 6.511e-05
alternative hypothesis: true mean is greater than 30
90 percent confidence interval:
 38.18143      Inf
sample estimates:
mean of x
 42.12931
t_greater_99

    One Sample t-test

data:  O3
t = 3.9601, df = 115, p-value = 6.511e-05
alternative hypothesis: true mean is greater than 30
99 percent confidence interval:
 34.9034     Inf
sample estimates:
mean of x
 42.12931
```

Tolkning 2b: Ensidig p-verdi = 6.51×10^{-5}. Ved 5 %-nivå: avviser H0 og konkluderer >30.
90 % KI: [38.2, ∞]; 99 % KI: [34.9, ∞]. 99 % er bredere enn 90 %.

2c) Juli+august vs. mai+juni+september

```
keep <- !is.na(O3) & !is.na(Month)
O3_keep <- O3[keep]; Month_keep <- Month[keep]

grp <- ifelse(Month_keep %in% c(7,8), "JA", "MJS")
grp <- factor(grp, levels=c("MJS","JA"))

tt2 <- t.test(O3_keep ~ grp)
tt2

    Welch Two Sample t-test

data:  O3_keep by grp
t = -5.5552, df = 82.798, p-value = 3.282e-07
alternative hypothesis: true difference in means between group MJS and group JA is not equal to 0
95 percent confidence interval:
 -42.85199 -20.25619
sample estimates:
mean in group MJS  mean in group JA
         27.98438          59.53846

grp_mean <- tapply(O3_keep, grp, mean, na.rm=TRUE)
grp_mean
     MJS       JA
27.98438 59.53846

par(mfrow=c(1,2))
boxplot(O3_keep ~ grp, col=c(mck_lightblue, mck_teal),
        main="2c: Ozone per gruppe", xlab="Gruppe", ylab="Ozone (ppb)")
stripchart(O3_keep ~ grp, vertical=TRUE, method="jitter",
           pch=16, col=mck_grey, add=TRUE)
par(mfrow=c(1,1))
```




```
par(op)
```

Tolkning 2c: Middel JA=59.5, MJS=28. P-verdi = 3.28×10^{-7} → signifikant forskjell (retning: JA > MJS).

Oppgave 3 – Månedstemperatur (kun 3a–3c)

```
mck_blue      <- "#0033A0"
mck_teal      <- "#00A3A1"
mck_grey      <- "#6D6E71"
mck_lightblue <- "#D9EAF7"
op <- par(family = "Arial")

mpath <- "https://www.uio.no/studier/emner/matnat/math/STK1000/data/obligdata/oblig1/blindernmonthly.tx
m     <- read.table(mpath, header=TRUE, sep=";")

data8 <- m[m$month==8, ]   # august
data9 <- m[m$month==9, ]   # september
```

3a) Oppsummer

```
summary(data8$temp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.10   15.32   16.10   16.14   16.70   20.50
summary(data9$temp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.80   10.85   12.00   11.78   12.57   15.00
```

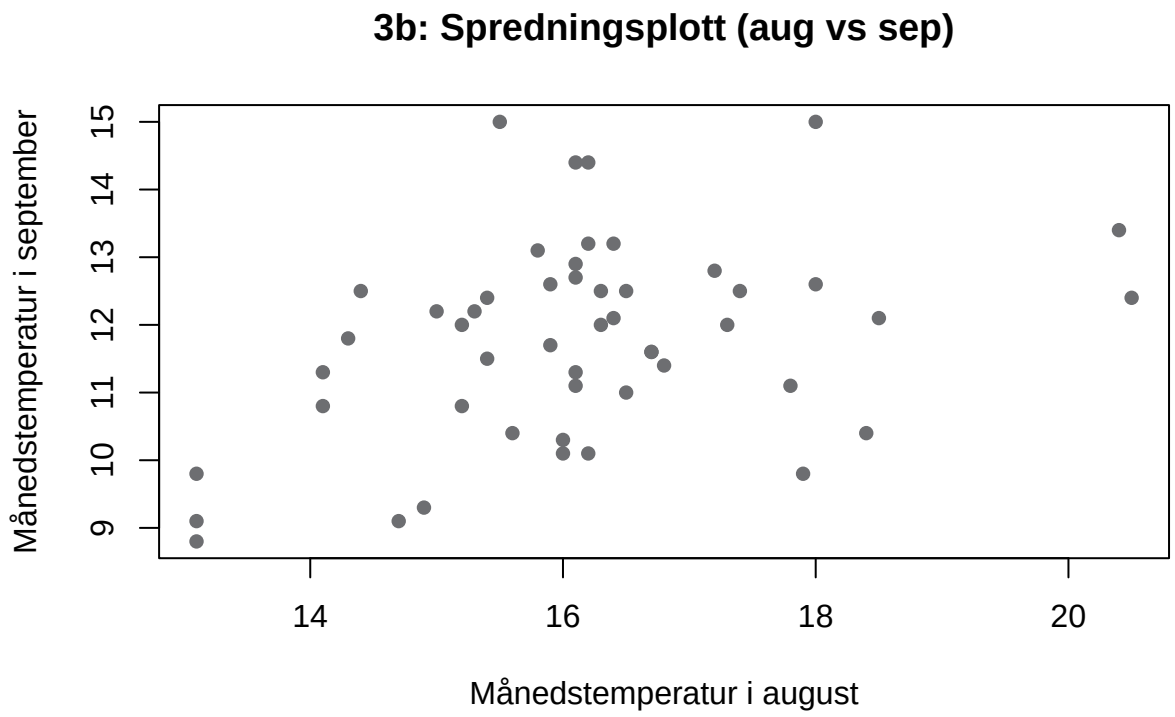
```r
meanAug   <- mean(data8$temp,   na.rm=TRUE)
medianAug <- median(data8$temp, na.rm=TRUE)
iqrAug    <- IQR(data8$temp,    na.rm=TRUE)

meanSep   <- mean(data9$temp,   na.rm=TRUE)
medianSep <- median(data9$temp, na.rm=TRUE)
iqrSep    <- IQR(data9$temp,    na.rm=TRUE)

cat("3a) Median august =", round(medianAug,1),
    " | Median september =", round(medianSep,1),
    " | IQR august =", round(iqrAug,1),
    " | IQR september =", round(iqrSep,1), "\n")
```

```
3a) Median august = 16.1  | Median september = 12  | IQR august = 1.4  | IQR september = 1.7
```

Tolkning 3a: August er varmere enn september: median august 16.1 mot september 12. IQR-ene (1.4 og 1.7) viser tilsvarende spredning.

3b) Spredningsplott (august på x, september på y)

```r
augTemp  <- data8$temp
septTemp <- data9$temp

plot(augTemp, septTemp,
     pch=16, col=mck_grey,
     xlab="Månedstemperatur i august",
     ylab="Månedstemperatur i september",
     main="3b: Spredningsplott (aug vs sep)")
```



Tolkning 3b: Positiv sammenheng: varmere august henger sammen med varmere september.

3c) Korrelasjon

```
r <- cor(augTemp, septTemp, use="complete.obs")
r
[1] 0.383473

styrke  <- if (abs(r)>=0.7) "sterk" else if (abs(r)>=0.4) "moderat" else "svak"
retning <- if (r>=0) "positiv" else "negativ"
cat("3c) r =", round(r,2), "→", styrke, " ", retning, " sammenheng.\n", sep="")
3c) r =0.38→svak positiv sammenheng.
par(op)
```

Tolkning 3c: r=0.38 gir svak positiv sammenheng; tydelig, men ikke veldig sterk.