Oppgave 2

a) Snittet av målingene ($n=15$) er $\bar{x} = 559.7$ og standard-avviket er $s = 28.6$. Med disse verdiene og med antagelsen om at populasjonen er normalfordelt med forventning μ og varians σ^2, så lager vi et 95% konfidens-intervall for forventning μ basert på målingene våre med:

$$\left(\bar{x} - t_{0.025, n-1} \frac{s}{\sqrt{n}}, \ \bar{x} + t_{0.025, n-1} \frac{s}{\sqrt{n}} \right)$$

hvor den kritiske t-verdien ved $\frac{\alpha}{2} = 0.025$ og frihetsgrader $df = n-1 = 15-1 = 14$ er $t_{0.025, 14} = 2.145$.

Setter inn alle verdiene og får intervallet:

$$\underline{\underline{(543.9, \ 575.5)}}$$

Vi kan da si med 95% sikkerhet at den virkelig verdien for μ vil ligge mellom 543.9 mg/100g og 575.5 mg/100g.

b) Begynner først med å liste opp målingene og finne lengen som vi vet er $n = 15$:

```r
data = c(525,587,547,558,591,531,571,551,566,622,561,502,556,565,562)
n = length(data)
```

Lager så $m = 10000$ datasett, hvor hvert datasett inneholder de stokastiske variablene $X_1, X_2, \ldots, X_{15} \overset{uif}{\sim} N(588, 30^2)$ ved hjelp av funksjonen rnorm(15, 588, 30).

For hvert av disse datasettene lager vi et 95% konfidensintervall for μ som i a) og teller hvor mange av disse intervallene som inneholder den virkelig verdien $\mu = 558$:

```r
set.seed(1)
m <- 10000
count <- 0
for (i in 1:m) {
    sim_data <- rnorm(n, 558, 30)
    sim_mean <- mean(sim_data)
    sim_sd   <- sd(sim_data)
    t_0.025  <- 2.145
    conf_low <- sim_mean - t_0.025 * sim_sd / sqrt(n)
    conf_hig <- sim_mean + t_0.025 * sim_sd / sqrt(n)
    if (conf_low < 558 && conf_hig > 558) {
        count <- count + 1
    }
}
print(count)
```

Dette ga count = 9527.

Vi fikk altså at 95.27 % av konfidensintervallene inneholdte $\mu = 558$. Dette stemmer godt med teorien bak konfidensintervall, hvor "95% sikkerhet" tilsvarer at 95 % av intervallene vi får ved mange målinger vil inneholde den virkelige verdien for μ.

c) Det tilnærmede intervallet for store utvalg er:

$$\bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

Dette er det samme som 95% intervallet, utenom at 1.96 er satt inn for t-verdien. Dette skyldes av grenseverdien for t-verdien ved $\frac{\alpha}{2} = 0.025$ som går mot 1.96 når $n \to \infty$ ($t_{0.025, n-1} \overset{n \to \infty}{=} 1.96$). Lager det tilnærmede intervallet for de 10000 datasettene som i b) og finner antallet av dem som inneholder $\mu = 558$:

```r
m <- 10000
count <- 0
for (i in 1:m) {
    sim_data <- rnorm(n, 558, 30)
    sim_mean <- mean(sim_data)
    sim_sd   <- sd(sim_data)
    conf_low <- sim_mean - 1.96 * sim_sd / sqrt(n)
    conf_hig <- sim_mean + 1.96 * sim_sd / sqrt(n)
    if (conf_low < 558 && conf_hig > 558) {
        count <- count + 1
    }
}
print(count)
```

Dette gir count = 9313. Altså 93.13 % av de tilnærmede intervallene for store utvalg inneholdte $\mu = 558$. Det er 2.14 % lavere enn det vi fikk i b). Dette gir mening ettersom vi effektivt sett har gjort intervallene smalere med 1.96 istedenfor 2.145, som betyr at færre intervaller vil inneholde $\mu = 558$.

d) Siden vi antar at populasjonen er normalfordelt så finner vi et 95 % konfidensintervall for variansen σ^2 ved:

$$\left(\frac{n-1}{\chi^2_{0.025,\, n-1}} \cdot s^2 ,\ \frac{n-1}{\chi^2_{0.975,\, n-1}} \cdot s^2 \right)$$

hvor de kritiske chi-verdiene er $\chi^2_{0.025,14} = 26.119$ og $\chi^2_{0.975,14} = 5.629$. Kvadratet av dette intervallet gir 95 % konfidens intervall for σ. Regner dette for $m = 10000$ datasett som i b) og teller hvor mange av disse intervallene som inneholder den virkelige verdien $\sigma = 30$:

```r
m <- 10000
count <- 0
for (i in 1:m) {
    sim_data <- rnorm(n, 558, 30)
    sim_sd   <- sd(sim_data)
    chi_low  <- 26.119
    chi_high <- 5.692
    conf_low <- sqrt((n-1) / chi_low * sim_sd^2)
    conf_hig <- sqrt((n-1) / chi_high * sim_sd^2)
    if (conf_low < 30 && conf_hig > 30) {
        count <- count + 1
    }
}
print(count)
```

Dette gir count = 9482. Altså 94.82 % av konfidensintervallene inneholder $\sigma = 30$. Dette stemmer godt med det vi ønsket å få.

e)