STK1110-h21: Obligatorisk innlevering 2

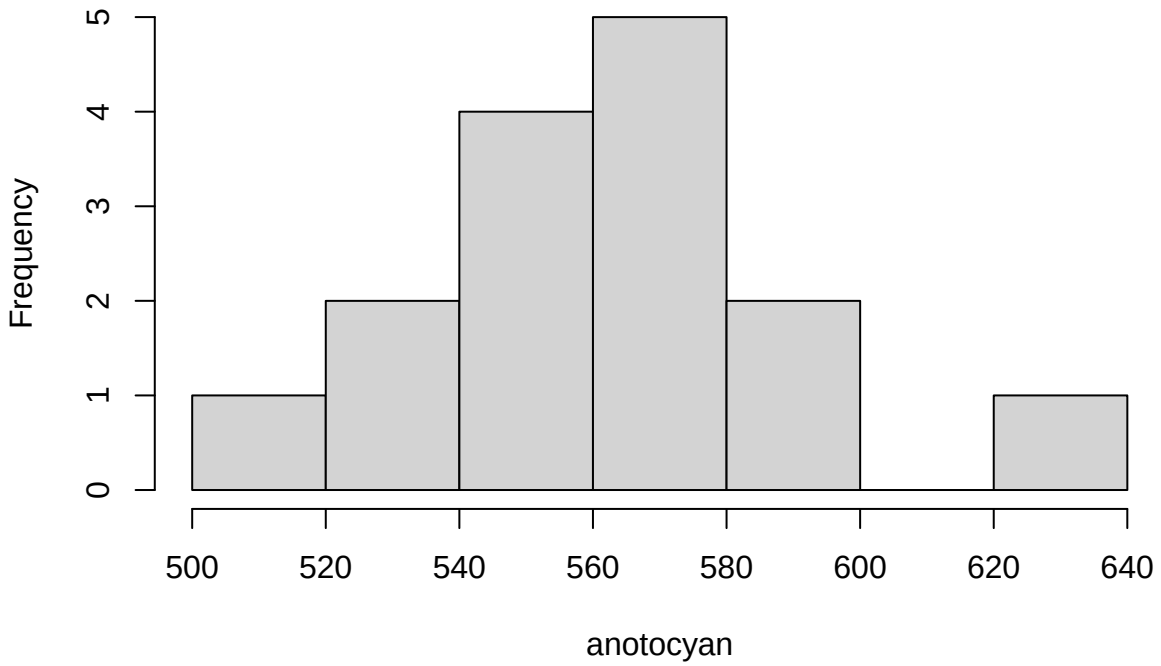
Martin Mihle Nygaard (martimn)

Oppgave 1 — Bare blåbær

Deloppgave (a) — 95% konfidensintervall

Jeg lager et histogram over målingene, $X_1, X_2, ..., X_n$, for å forsikre meg om at de er sånn passe normalfordelt.

```
anotocyan <- c(525, 587, 547, 558, 591, 531, 571, 551, 566, 622, 561, 502, 556, 565, 562)
hist(anotocyan, main = '')
```



Og det ser det ut som de er. $X_1, X_2, ..., X_n$ kan da sees på som tilfeldige uttrekninger fra en normalfordeling med forventningsverdi μ og standardavvik σ. Ettersom antall observasjoner n er mindre enn tommelfingerreglen 40, tar jeg utgangspunkt i t-fordelingen. Fra læreboka (Devore s. 403) bruker jeg proposisjon (8.14),

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

hvor $\bar{X}$ er snittet av observasjonene, s standardavviket og $t_{\alpha,n-1}$ ønsket andel α av tetthetsfunksjonen til t-fordelingen med $n-1$ frihetsgrader. Jeg setter inn for disse og gjør beregningene i R.

```
alpha <- 1 - 0.95
anotocyan.mean <- mean(anotocyan)
anotocyan.sd   <- sd(anotocyan)
anotocyan.n    <- length(anotocyan)
KI <- c(anotocyan.mean - qt(1-alpha/2, anotocyan.n-1) * anotocyan.sd / sqrt(anotocyan.n),
        anotocyan.mean + qt(1-alpha/2, anotocyan.n-1) * anotocyan.sd / sqrt(anotocyan.n))
```

Dette gir 95% konfidensintervallet (543.851, 575.483).

Deloppgave (b) — Simulert t-fordeling konfidensintervall

```
teller <- 0
for (i in 1:10000) {
  uttrekk      <- rnorm(15, mean = 558, sd = 30)     # trekk ut 15 observasjoner
  uttrekk.mean <- mean(uttrekk)                      # finn snittet av disse
  uttrekk.sd   <- sd(uttrekk)                        # finn standardavvik
  intervall    <- c(uttrekk.mean - qt(0.975, 15-1) * uttrekk.sd / sqrt(15),
                    uttrekk.mean + qt(0.975, 15-1) * uttrekk.sd / sqrt(15))
  if (intervall[1] <= 558 & 558 <= intervall[2]) {   # hvis 558 er innenfor KI
    teller <- teller + 1                             # ==> øk teller med én
  }
}
```

Dette gir 9516 antall simulerte konfidensintervaller som inneholder 558. Dette utgjør 95.16% av totalt antall simuleringer. Hvis koden min er lusløs, er dette akkurat som intervallet forutså; 95%-konfidensintervall betyr at verdien vi undersøker skal være innafor 95% av tiden, som er omtrent tilfellet i disse simuleringene.

Deloppgave (c) — Simulert konfidensintervall for store utvalg

Gjør akkurat samme steg som forrige gang, bare med alternativt intervall, som spesifisert i oppgaven.

```
teller <- 0
for (i in 1:10000) {
  uttrekk      <- rnorm(15, mean = 558, sd = 30)     # trekk ut 15 observasjoner
  uttrekk.mean <- mean(uttrekk)                      # finn snittet av disse
  uttrekk.sd   <- sd(uttrekk)                        # finn standardavvik
  intervall    <- c(uttrekk.mean - 1.96 * uttrekk.sd / sqrt(15),
                    uttrekk.mean + 1.96 * uttrekk.sd / sqrt(15))
  if (intervall[1] <= 558 & 558 <= intervall[2]) {   # hvis 558 er innenfor KI
    teller <- teller + 1                             # => øk teller med én
  }
}
```

Dette gir (i en eksempelkjøring) 9311 antall simulerte konfidensintervaller som inneholder 558. Dette utgjør 93.11% av totalt antall simuleringer. Altså, dette blir slett ikke et 95% konfidensintervall; andelen innenfor intervallet blir for lavt. Dette er som forventet, siden vi har så få observasjoner (15), som strider med betingelsen «stort utvalg».

Deloppgave (d) — Konfidensintervall for σ

Jeg bruker formuleringen av konfidensintervall for varians og standardavvik fra læreboka (Devore, s. 410).

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right) \tag{1}$$

Hvor $\chi^2_{\alpha,v}$ er kvantilfunksjonen chi-kvadrat fordelingen. Jeg bruker så `R` som i tidligere deloppgaver.

```
teller <- 0
for (i in 1:10000) {
  uttrekk      <- rnorm(15, mean = 558, sd = 30)   # trekk ut 15 observasjoner
```

```
  uttrekk.var  <- var(uttrekk)                      # finn standardavvik
  intervall    <- sqrt(c((15-1) * uttrekk.var / qchisq(0.975, 15-1, lower.tail = TRUE),
                         (15-1) * uttrekk.var / qchisq(0.975, 15-1, lower.tail = FALSE)))
  if (intervall[1] <= 30 & 30 <= intervall[2]) {   # hvis 30 er innenfor KI
    teller <- teller + 1                           # ==> øk teller med én
  }
}
```

I en eksempelkjøring, får vi 9513 intervaller som inneholder 30. Dette utgjør en andel på 95.13%. Som er svært nær forventningen på 95%.

Deloppgave (e) — Konfidensintervall for μ med t-fordelt populasjon

```
teller <- 0
for (i in 1:10000) {
  uttrekk   <- rt(15, 7)
  x         <- 558 + uttrekk * 30
  intervall <- c(mean(x) - qt(0.975, 15-1) * sd(x) / sqrt(15),
                 mean(x) + qt(0.975, 15-1) * sd(x) / sqrt(15))
  if (intervall[1] <= 558 & 558 <= intervall[2]) {   # hvis 30 er innenfor KI
    teller <- teller + 1                             # ==> øk teller med én
  }
}
```

I en eksempelkjøring, får vi 9513 intervaller som inneholder 30. Dette utgjør en andel på 95.13%.

Øh, hvis oppgaven er tiltenkt å ta stilling til robusthet som beskrevet i denne Wikipedia artikkelen[1], er jeg usikker. Jeg får jo samme resultat som i deloppgave (b), men vet ikke om dette betyr at metoden er «robust» under en eller annen streng definisjon.

Deloppgave (f) — Konfidensintervall for σ med t-fordelt populasjon

```
tilde_sigma <- sqrt(1.4) * 30
teller      <- 0
for (i in 1:10000) {
  uttrekk   <- rt(15, 7)
  z         <- 558 + uttrekk * tilde_sigma
  intervall <- sqrt(c((15-1) * var(z) / qchisq(0.975, 15-1, lower.tail = TRUE),
                      (15-1) * var(z) / qchisq(0.975, 15-1, lower.tail = FALSE)))
  if (intervall[1] <= tilde_sigma & tilde_sigma <= intervall[2]) {
    teller <- teller + 1
  }
}
```

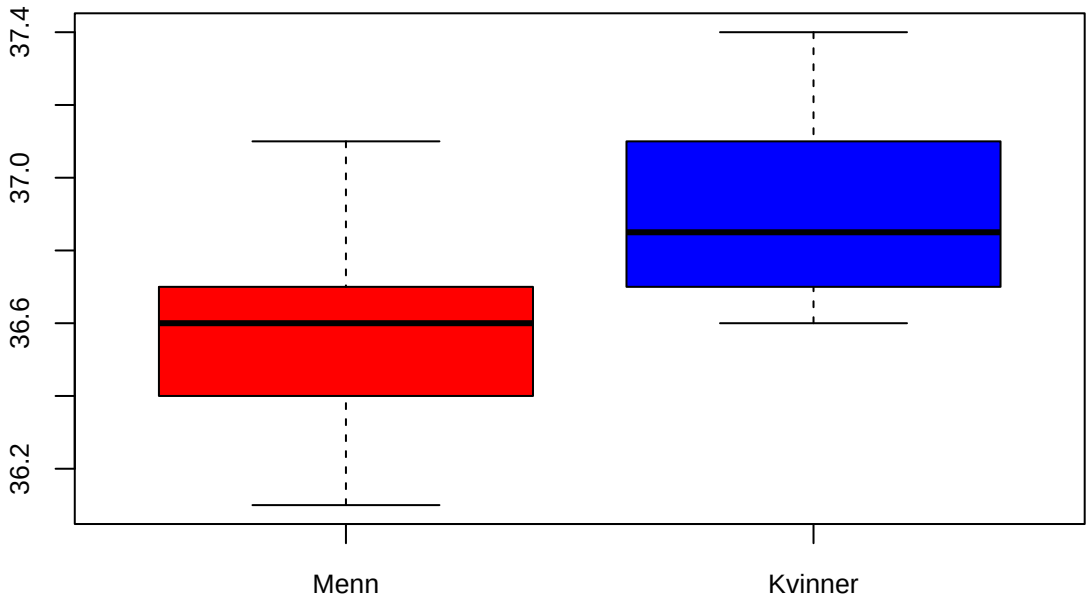
I en eksempelkjøring, får vi 7874 intervaller som inneholder $\tilde{\sigma} = \sqrt{1.4} \cdot \sigma = \sqrt{1.4} \cdot 30$. Dette utgjør en andel på 78.74%. Dette er mindre enn 95%, som var det jeg siktet på. Tror dette kommer av at betingelsen om en normalfordelt populasjon for bruk av (1) er brutt; populasjonen er nå t-fordelt i stedet, og tydeligvis ikke nærme nok normalfordeling.

Oppgave 2 — Hete kropper

[1] https://en.wikipedia.org/wiki/Robust_confidence_intervals

Deloppgave (a) — Boksplott

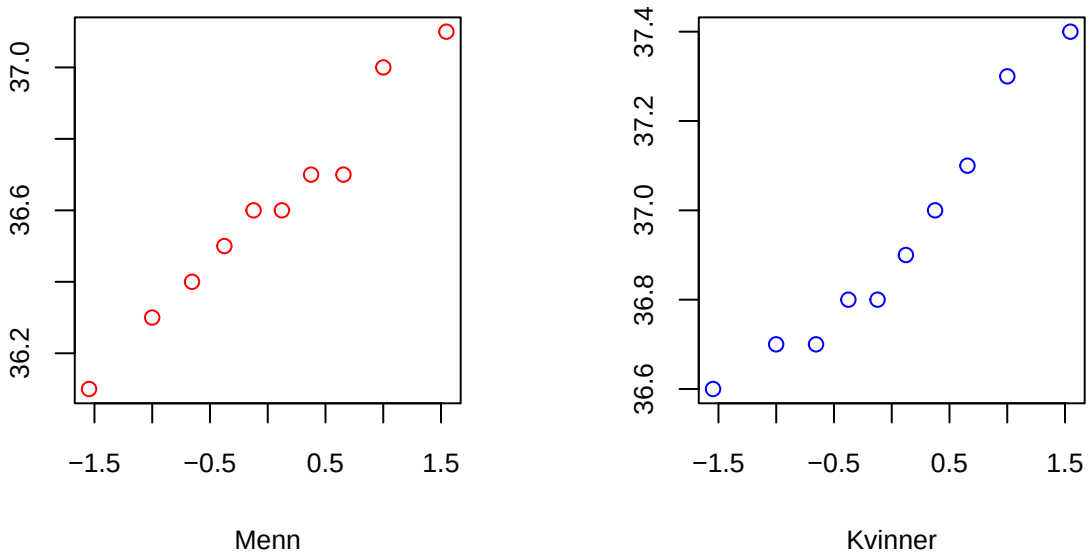
```
par(ps = 10)
temp <- read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt",
                   header = TRUE)
boxplot(temp, col = c("red", "blue"))
```



Det ser ut som menn har en lavere kroppstemperatur enn kvinner i dette datasettet. Men medianen er ikke ekstremt forskjellig.

Deloppgave (b) — Normalfordelingsplott

```
par(mfrow = c(1,2), pty = "s", ps = 10)
qqnorm(temp$Menn, col = "red", xlab = "Menn", ylab = "", main = NULL)
qqnorm(temp$Kvinner, col = "blue", xlab = "Kvinner", ylab = "", main = NULL)
```



Det ser ut til at dataene sammenfaller ganske greit med normalfordeling, ettersom datapunktene ligger

omtrent på en rett linje. Det virker som temperaturene målt på kvinner er noe mer kurvet, som kan tyde på en viss skjevhet (dette ser man litt i boksplottet også).

Deloppgave (c) — Normalfordelte, små, utvalg med ukjent, lik, varians

La den stokastiske variabelen X være kroppstemperaturen en tilfeldig kvinne og Y til en tilfeldig mann. Jeg antar at observasjonene $X_1, ..., X_m$ og $Y_1, ..., Y_n$ er uavhengig identisk fordelt med henholdsvis forventning μ_K og μ_M.

Jeg antar at variansene er like, altså $\sigma_K^2 = \sigma_M^2 = \sigma^2$, men ukjente. Ettersom både X og Y er antatt normalfordelte, vil også transformasjonen $\bar{X} - \bar{Y}$ også være normalfordelt med forventning $\mu_K - \mu_M$. En estimator for σ^2 som tar høyde for forskjellig utvalgstørrelse (altså at $m \neq n$) er

$$S_p^2 = \frac{m-1}{m+n-2} S_K^2 + \frac{n-1}{n+m-2} S_M^2 \tag{2}$$

hvor $S_{\{K,M\}}^2$ er den observerte variansen.

Jeg kan nå konstruere en ny stokastisk, standardisert, variabel T. Denne er t-fordelt ettersom vi har for få observasjoner til å anta «stort» utvalg. Altså,[2]

$$T = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_K - \mu_M)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2} \tag{3}$$

Hypotesetest

Jeg ønsker å undersøke om $\mu_K \neq \mu_M$, eller ekvivalent $\mu_K - \mu_M \neq 0$. Jeg formulerer derfor nullhypotesen H_0 og alternativhypotesen H_a som følger:

- H_0: $\mu_K - \mu_M = \Delta_0 = 0$
- H_a: $\mu_K - \mu_M \neq \Delta_0 = 0$

Jeg bruker (3) som testvariabel, hvor $\mu_K - \mu_M$ erstattes med Δ_0. Dette fungerer ettersom denne t-fordelingen er sentrert rundt 0 (standardisert), og jeg ønsker å undersøke om differansen Δ_0 er langt nok unna 0, gitt et visst standardavvik. Jeg ønsker å være 95% sikker, eller «konfident», om du vil. Derfor formuleres forkastingsbettingelsen slik:

$$H_0 \text{ forkastes dersom } \left\{ \begin{matrix} t \leq t_{\alpha/2, m+n-2} \\ t \geq t_{1-\alpha/2, m+n-2} \end{matrix} \right\} \text{ hvor } t = \frac{\left(\bar{X} - \bar{Y}\right) - \Delta_0}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ og } \alpha = 1 - 95\%.$$

Konfidensintervall

Jeg ønsker et intervall som det er $1 - \alpha$ sannsynlighet for at T befinner seg i. Øvre og nedre skranke for dette intervallet blir $\pm t_{\alpha/2, m+n-2}$. Jeg manipulerer dette intervallet til heller å beskrive $(\mu_K - \mu_M)$:

[2] `TODO`: kjapt forklar hvor $\sqrt{\frac{1}{m} + \frac{1}{n}}$ kommer fra.

$$
\begin{aligned}
1-\alpha &= \mathrm{P}\left(t_{\alpha/2,m+n-2} \leq \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \leq t_{1-\alpha/2,m+n-2} \right) \\
&= \mathrm{P}\left(t_{\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \leq \left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M) \leq t_{1-\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \right) \\
&= \mathrm{P}\left(-\left(\bar{X}-\bar{Y}\right)+t_{\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \leq -\mu_K-\mu_M \leq -\left(\bar{X}-\bar{Y}\right)+t_{1-\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \right) \\
&= \mathrm{P}\left(\left(\bar{X}-\bar{Y}\right)-t_{\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \geq \mu_K-\mu_M \geq \left(\bar{X}-\bar{Y}\right)-t_{1-\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \right) \\
&= \mathrm{P}\left(\left(\bar{X}-\bar{Y}\right)-t_{1-\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \leq \mu_K-\mu_M \leq \left(\bar{X}-\bar{Y}\right)-t_{\alpha/2,m+n-2} S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \right)
\end{aligned}
$$

Siste linje gir meg formuleringen av konfidensintervallet for differansen mellom forventningene til de to variablene X og Y:

$$
\left((\bar{x}-\bar{y}) - t_{1-\alpha/2,m+n-2} \cdot s_p\sqrt{\frac{1}{m}+\frac{1}{n}}, (\bar{x}-\bar{y}) - t_{\alpha/2,m+n-2} \cdot s_p\sqrt{\frac{1}{m}+\frac{1}{n}} \right)
$$

P-verdi

$$
\begin{aligned}
P\text{-verdi} &= P\left(-t \geq T \geq t \mid \mu_K-\mu_M = 0\right) \\
&\approx \text{ område under } t\text{-kurven før } -t \text{ og etter } t \\
&= 1-\int_{-t}^{t} t_{m+n-2} \\
&= 2\cdot\left(1-F_{m+n-2}(t)\right)
\end{aligned}
$$

Hvor F_v er den kumulative fordelingsfunksjonen for t-fordelingen med v frihetsgrader. Her utnytter jeg at t-fordelingen er symmetrisk, som er hvorfor jeg ganger med to (for nedre og øvre hale).

Sjekk i `R`

```
# Hypotesetest
alpha    <- 1-0.95
Δ_0      <- 0
x_strek  <- mean(temp$Kvinner)
y_strek  <- mean(temp$Menn)
m        <- length(temp$Kvinner)
n        <- length(temp$Menn)
S_p      <- sqrt(((m-1)/(m+n-2)) * var(temp$Kvinner) + ((n-1)/(n+m-2)) * var(temp$Menn))
t        <- ((x_strek-y_strek) - Δ_0) / (S_p * sqrt(1/m + 1/n))
skranker <- c(qt(alpha/2, m+n-2), qt(1-alpha/2, m+n-2))

# Konfidensintervall
KI <- c((x_strek-y_strek) - qt(1-alpha/2, m+n-2) * S_p * sqrt(1/m + 1/n),
        (x_strek-y_strek) - qt(alpha/2, m+n-2) * S_p * sqrt(1/m + 1/n))

# P-verdi
P_verdi <- 2*(1-pt(t, m+n-2))
```

Dette gir $t = 2.5900616$, som ligger utenfor intervallet $(-2.100922, 2.100922)$. H_0 bør derfor forkastes. P-verdien er i samsvar med denne konklusjonen, siden P-verdi $= 0.0184813 < \alpha = 0.05$. Konfidensintervallet blir $(0.0623213, 0.5976787)$.

Jeg får god overensstemmelse med Rs innebygde t-test.

```
t.test(temp$Kvinner, temp$Menn)
```

```
##
##  Welch Two Sample t-test
##
## data:  temp$Kvinner and temp$Menn
## t = 2.5901, df = 17.734, p-value = 0.01863
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06203301 0.59796699
## sample estimates:
## mean of x mean of y
##     36.93     36.60
```

Deloppgave (d) — Ulik varians

I tilfellet med forskjellig varians må jeg justere litt på (3).[3]

$$
\begin{aligned}
T &= \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{\frac{S_K^2}{m}+\frac{S_M^2}{n}}} \\
&= \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{\frac{S_K^2}{m}+\frac{S_M^2}{n}}} \cdot \frac{\left(\sqrt{\sigma_K^2/m+\sigma_M^2/n}\right)^{-1}}{\left(\sqrt{\sigma_K^2/m+\sigma_M^2/n}\right)^{-1}} \\
&= \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{\sigma_K^2/m+\sigma_M^2/n}} \cdot \frac{\left(\sqrt{\frac{S_K^2}{m}+\frac{S_M^2}{n}}\right)^{-1}}{\left(\sqrt{\sigma_K^2/m+\sigma_M^2/n}\right)^{-1}} \\
&= \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{\sigma_K^2/m+\sigma_M^2/n}} \cdot \left(\sqrt{\frac{\frac{S_K^2}{m}+\frac{S_M^2}{n}}{\sigma_K^2/m+\sigma_M^2/n}}\right)^{-1} \\
&= Z \cdot \left[\sqrt{\frac{U}{\nu}}\right]^{-1}, \text{ hvor } Z = \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{\sigma_K^2/m+\sigma_M^2/n}} \sim N(0,1) \text{ og } \frac{U}{\nu} = \frac{\frac{S_K^2}{m}+\frac{S_M^2}{n}}{\sigma_K^2/m+\sigma_M^2/n}
\end{aligned}
$$

U kan tilnærmes med χ^2_ν fordelingen. Da brukes en justert v, som tilnærmes med de observerte variansene S_K^2 og S_M^2 slik:

$$
\nu = \frac{\sigma_K^2/m+\sigma_M^2/n}{\left(\frac{\sigma_K^2}{m}\right)^2/(m-1)+\left(\frac{\sigma_M^2}{n}\right)^2/(n-1)} \approx \frac{S_K^2/m+S_M^2/n}{\left(\frac{S_K^2}{m}\right)^2/(m-1)+\left(\frac{S_M^2}{n}\right)^2/(n-1)}
$$

Denne ν-en brukes tilnærme riktig t-fordeling:

$$
T = \frac{\left(\bar{X}-\bar{Y}\right)-(\mu_K-\mu_M)}{\sqrt{S_K^2/m+S_M^2/n}} \sim t_\nu
$$

[3]Her er jeg lite stødig, og innrømmer ærlig at jeg regurgiterer forelesningsfoilene.

Hypotesetest

Som i forrige deloppgave, men forkastingsbettingelsen blir

$$H_0 \text{ forkastes dersom } \left\{ \begin{matrix} t \leq t_{\alpha/2,\nu} \\ t \geq t_{1-\alpha/2,\nu} \end{matrix} \right\} \text{ hvor } t = \frac{\left(\bar{X} - \bar{Y}\right) - \Delta_0}{\sqrt{S_K^2/m + S_M^2/n}} \text{ og } \alpha = 1 - 95\%.$$

Konfidensintervall

Utledningen er analog til forrige oppgave. Konfidensintervallet blir

$$\left((\bar{x} - \bar{y}) - t_{1-\alpha/2,\nu} \cdot \sqrt{s_k^2/m + s_m^2/n}, (\bar{x} - \bar{y}) - t_{\alpha/2,\nu} \cdot \sqrt{s_k^2/m + s_m^2/n} \right)$$

***P*-verdi**

Igjen, som i forrige oppgave, men med ν innsatt for $m + n - 2$.

$$\begin{aligned} P\text{-verdi} &= P\left(-t \geq T \geq t \mid \mu_K - \mu_M = 0\right) \\ &\approx \text{ område under } t\text{-kurven før } -t \text{ og etter } t \\ &= 1 - \int_{-t}^{t} t_\nu \\ &= 2 \cdot (1 - F_\nu(t)) \end{aligned}$$

Utregning i `R`

```
# Hypotesetest
v        <- ( (var(temp$Kvinner)/m + var(temp$Menn)/n)
           / ((var(temp$Kvinner)/m)^2/(m-1) + (var(temp$Menn)/n)^2/(n-1)))
t        <- ( ((x_strek-y_strek) - Δ_0)
           / sqrt(var(temp$Kvinner)/m + var(temp$Menn)/n))
skranker <- c(qt(alpha/2, v), qt(1-alpha/2, v))

# Konfidensintervall
KI <- c((x_strek-y_strek)-qt(1-alpha/2, v) * sqrt(var(temp$Kvinner)/m+var(temp$Menn)/n),
        (x_strek-y_strek)-qt(alpha/2, v)   * sqrt(var(temp$Kvinner)/m+var(temp$Menn)/n))

# P-verdi
P_verdi <- 2*(1-pt(t, v))
```

Dette gir $t = 2.5900616$, som ligger utenfor intervallet $(-1.9621379, 1.9621379)$. H_0 bør derfor fortsatt forkastes (om jeg ikke har regna feil). P-verdien er fortsatt i samsvar: P-verdi $= 0.0097236 < \alpha = 0.05$. Konfidensintervallet blir $(0.0800038, 0.5799962)$.

Det virker som antagelsen om ulik varians drastisk styrker H_a. Litt spekulasjon, men fra plottet i deloppgave (a) kan vi se at observasjonene av kvinnenes temperatur har noe mindre varians; hvis denne variansen hadde vært mer jevnstor med mennenes (altså antagelse om lik varians), ville de to fordelingene overlappet mer, og desto mer observasjonene overlapper, jo vanskeligere er det å fastslå med sikkerhet at de er forskjellige? Derfor vil sikkerheten øke (og P-verdien minske).

Deloppgave (e) — F-test

Følger metoden beskrevet i seksjon 10.5 (Devore s. 527). Jeg formulerer følgende hypoteser om variansen: - H_0: $\sigma_K^2 = \sigma_M^2$ - H_a: $\sigma_K^2 \neq \sigma_M^2$ Som teststatistikk bruker jeg forholdet $f = s_K^2/s_M^2$, som har en F-fordeling

med parametre $v_1 = m - 1$ og $v_2 = n - 1$. Forkastingsbetingelsene blir

$$H_0 \text{ forkastes dersom } \left\{ \begin{matrix} f \leq F_{\alpha/2,m-1,n-1} \\ f \geq F_{1-\alpha/2,m-1,n-1} \end{matrix} \right\} \text{ hvor } f = \frac{s_K^2}{s_M^2} \text{ og } \alpha = 1 - 95\%.$$

```
m <- length(temp$Kvinner)
n <- length(temp$Menn)
f <- var(temp$Kvinner)/var(temp$Menn)
f.forkast <- c(qf(  alpha/2, m-1, n-1),
               qf(1-alpha/2, m-1, n-1))
f.konfint <- c(f*qf(alpha/2, m-1, n-1),
               f/qf(alpha/2, m-1, n-1))
```

Basert på `R` koden over, ser det ut som $f \approx 0.782$ *ikke* ligger i forkastingsområdet til H_0, altså innenfor intervallet $(0.248, 4.026)$, som impliserer at den alternative hypotesen bør forkastes. Et 95% konfidensintervall for f blir $(0.194, 3.147)$. Jeg får god overensstemmelse med den innebygde testen:

```
var.test(temp$Kvinner, temp$Menn)
```

```
##
##  F test to compare two variances
##
## data:  temp$Kvinner and temp$Menn
## F = 0.78171, num df = 9, denom df = 9, p-value = 0.7197
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.194165 3.147149
## sample estimates:
## ratio of variances
##          0.7817073
```

Deloppgave (f) — Prediksjonsintervall

Med utgangspunkt i normalfordeling på både X og Y, vet jeg at forventningsrette estimatorer for disse er snittet, henholdsvis $\bar{X}$ og $\bar{Y}$. Videre, vet jeg også at observasjonene er uavhengig identisk fordelt, derfor er det rimelig å anta at neste observasjon har samme forventning som de foregående. Med disse antagelsene blir forventningen da

$$\mathrm{E}(X_{11} - Y_{11}) = \mathrm{E}(X_{11}) - \mathrm{E}(Y_{11}) = \mathrm{E}(X) - \mathrm{E}(Y) = \bar{X} - \bar{Y}.$$

Jeg prøver å finne forventningen og variansen til uttrykket $X_{11} - Y_{11} - (\bar{X} - \bar{Y})$:

$$\begin{aligned} \mathrm{E}\left[X_{11} - Y_{11} - (\bar{X} - \bar{Y})\right] &= \mathrm{E}\left[X_{11}\right] - \mathrm{E}\left[Y_{11}\right] - \mathrm{E}\left[\bar{X}\right] + \mathrm{E}\left[\bar{Y}\right] \\ &= \bar{X} - \bar{Y} - \bar{X} + \bar{Y} = 0 \\ \mathrm{Var}\left[X_{11} - Y_{11} - (\bar{X} - \bar{Y})\right] &= \mathrm{Var}\left[X_{11}\right] + \mathrm{Var}\left[Y_{11}\right] + \mathrm{Var}\left[\bar{X}\right] + \mathrm{Var}\left[\bar{Y}\right] \\ &= \sigma_K^2 + \sigma_M^2 + \frac{\sigma_K^2}{m} + \frac{\sigma_M^2}{n} \end{aligned}$$

Videre, hvis jeg antar at variansene til X og Y er like — som foreslått i forrige deloppgave — kan jeg bruke det vektede snittet til de observerte variansene S_p^2 fra (2).

$$\sigma_K^2 + \sigma_M^2 + \frac{\sigma_K^2}{m} + \frac{\sigma_M^2}{n} = S_p^2 + S_p^2 + \frac{S_p^2}{m} + \frac{S_p^2}{n} = S_p^2 \left(2 + \frac{1}{m} + \frac{1}{n}\right)$$

Hvis variansen er kjent, blir fordelingen til uttrykket $X_{11} - Y_{11} - (\bar{X} - \bar{Y})$ normalfordelt ettersom dette er en lineærkombinasjon av antatt normalfordelte stokastiske variable. Hvis variansen er ukjent er det snakk om T-fordeling[4]. Så — jeg tror — vi kan konstruere en stokastisk variabel

$$T_{11} = \frac{X_{11} - Y_{11} - (\bar{X} - \bar{Y}) - 0}{S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}.$$

Denne T_{11} kan jeg manipulere for å finne et prediksjonsintervall, akkurat som tidligere med konfidensintervall.

$$\begin{aligned}
1 - \alpha &= \mathrm{P}\left[t_{\alpha/2,m+n-2} < T_{11} < t_{1-\alpha/2,m+n-2}\right] \\
&= \mathrm{P}\left[t_{\alpha/2,m+n-2} < \frac{X_{11} - Y_{11} - (\bar{X} - \bar{Y}) - 0}{S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}} < t_{1-\alpha/2,m+n-2}\right] \\
&= \mathrm{P}\left[\bar{X} - \bar{Y} + t_{\alpha/2,m+n-2} \cdot S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}} < X_{11} - Y_{11} < \bar{X} - \bar{Y} + t_{1-\alpha/2,m+n-2} \cdot S_p\sqrt{2 + \frac{1}{m} + \frac{1}{n}}\right]
\end{aligned}$$

Setter jeg inn de observerte snittene, m, n, og α, får jeg følgende 95% prediksjonsintervall for $X_{11} - Y_{11}$:

$$\left(\bar{x} - \bar{y} + t_{.025,18} \cdot s_p\sqrt{2 + \tfrac{2}{10}},\ \bar{x} - \bar{y} + t_{.975,18} \cdot s_p\sqrt{2 + \tfrac{2}{10}}\right)$$

Dette er altså et intervall vi skal være 95% sikre på at differansen mellom kroppstemperaturen til neste observerte kvinne og mann faller innenfor. Dette i motsetning til konfidensintervallet fra deloppgave (c), hvor vi ønsket å finne et intervall for den forventede gjennomsnittforskjellen.

Dette intervallet er betydelig bredere enn konfidensintervallet for $\mu_1 - \mu_2$. Dette kommer av at usikkerheten rundt $\mu_1 - \mu_2$ minsker med mengden data; mens $X_{11} - Y_{11}$ vil *alltid* ha en usikkerhet proporsjonal med den faktiske variansen i dataene (reflektert i det konstante leddet under rottegnet).

Oppgave 3 — Fedre i tidsklemma

Deloppgave (a) — Hypotesetest, p-verdi

Dette er et «stort» utvalg. La p_M og p_F være henholdsvis andelen mødre og fedre i «tidsklemma», og $\hat{p}_{\{M,F\}}$ er de observerte andelene. La $\hat{p}$ være et vektet gjennomsnitt av de to populasjonsandelene, altså

$$\hat{p} = \frac{m}{m+n}\hat{p}_M + \frac{n}{n+m}\hat{p}_F.$$

Hvor m og n er antall mødre og fedre spurt, henholdsvis (begge er 3000 i denne undersøkelsen). Jeg formulerer følgende hypoteser:

- H_0: $p_M - p_F = 0$
- H_a: $p_M - p_F \neq 0$

Og setter forkastingsbettingelsen

$$H_0 \text{ forkastes dersom } \begin{Bmatrix} z \leq z_{\alpha/2} \\ z \geq z_{1-\alpha/2} \end{Bmatrix} \text{ hvor } z = \frac{\hat{p}_M - \hat{p}_F}{\sqrt{\hat{p}(1-\hat{p})\,(1/m + 1/n)}} \text{ og } \alpha = 1 - 95\%.$$

P-verdien er gitt ved $2 \cdot (1 - \Phi(|z|))$.

Jeg bruker `R` til å regne ut de faktiske verdiene.

[4] Jeg antar med $m + n - 2$ frihetsgrader, men det er jeg oppriktig talt usikker på.

```r
pM <- 441
pF <- 486
m <- 3000
n <- 3000
pM_hatt <- pM/m
pF_hatt <- pF/n
p_hatt  <- (m*pM_hatt)/(m+n) + (n*pF_hatt)/(m+n)
z <- (pM_hatt - pF_hatt) / sqrt(p_hatt*(1/m+1/n))
forkast <- c(qnorm(alpha/2), qnorm(1-alpha/2))
p_verdi <- 2*(1-pnorm(abs(z)))
```

Test verdien $z = -1.4779939$ ligger innenfor skrankene -1.959964 og 1.959964. H_0 beholdes, og H_a forkastes. P-verdien er 0.1394094, som ikke er mindre enn signifikansnivået på 5%.

Altså, forskjellen mellom mødre og fedre er ikke signifikant.

Deloppgave (b) — Sjekk i R

```r
prop.test(c(441, 486), c(3000, 3000), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(441, 486) out of c(3000, 3000)
## X-squared = 2.5836, df = 1, p-value = 0.108
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.033286474  0.003286474
## sample estimates:
## prop 1 prop 2
##  0.147  0.162
```

Jeg vil si at forskjellen fortsatt *ikke* er signifikant. Men jeg får forskjellig p-verdi. Fra p-verdien her ser man at et slikt resultat vil forekomme 10.8% av tiden, hvis det ikke er noen forskjell mellom mødre og fedre (altså om H_0 er sann).