STK1110

Obligatorisk oppgave 2

Jonas Semprini Næss

25. januar 2022

NB! Kode på oppgave 1 er skrevet i samarbeid med "erlek@math.uio.no"

Oppgave 1:

a.) *Lag et 95 % konfidensintervall for forventet antocyaninnhold μ basert på målingene over.*

Løsning:

Vi antar fra opplysningene at vi har tilnærmet normalfordelte målinger

$$X_1, \ldots, X_{15} \sim N(\mu, \sigma)$$

for antocyaninnhold slik at $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ og $Z = \frac{\bar{X}-\mu}{\sigma \backslash \sqrt{n}} \sim N(\mu, \sigma)$. Dermed vil et generelt konfidensintervall se slik ut

$$P(-Z_{\alpha\backslash 2} \leq \frac{\bar{X}-\mu}{\sigma \backslash \sqrt{n}} \leq Z_{\alpha\backslash 2}) = 1-\alpha$$

$$P(-\bar{X} - Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\bar{X} + Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

hvor $\alpha = 0.05$ slik at det teoretiske intervallet er gitt ved

$$(\bar{X} - Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha\backslash 2}\frac{\sigma}{\sqrt{n}}) \qquad \blacksquare$$

Ønsker vi dessuten å lage et empirisk konfidensintervall for μ tar vi simpelthen å bruker det empiriske gjennomsnittet og variansen gitt ved

$$\begin{aligned}
\mu &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \mu)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i \right)^2 \\
&\Downarrow \\
S &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i \right)}
\end{aligned}$$

som gir følgende konfidensintervall

$$(\bar{X} - Z_{\alpha\backslash 2}\frac{S}{\sqrt{n}}, \bar{X} + Z_{\alpha\backslash 2}\frac{S}{\sqrt{n}}) \qquad \blacksquare$$

b.) *På Wikipedia kan vi lese at forventet antocyaninnhold i blåbær er* 558 *mg/100g. Nå skal du bruke simuleringer til å late som om du måler antocyan i 15 prøver med blåbær veldig mange ganger. Generer 10000 datasett, hvert av størrelse* $n = 15$*, bestående av observasjoner av de stokastiske variablene* $X_1 \dots X_{15} \sim N(558, 30)$*. Du kan bruke rnorm()-funksjonen i R til dette. Selv om du har simulert fra en fordeling med kjent forventning og varians, skal du late som om begge disse er ukjent i det følgende. Lag et 95 % konfidensintervall for* μ *som i punkt a), basert på hvert av de simulerte datasettene, slik at du får 10000 intervaller. Tell opp andelen av disse intervallene som inneholder verdien 558. Kommenter og forklar*

Løsning:

Som løsning av denne oppgaven vil det være fingert å enten lage en stor $N \times n$ matrise hvor $N = 10000$, $n = 15$ eller en, 1-dimensjonal array for å samle alle dataene. I det følgende R-scriptet er det benyttet en 1-d array og en enkel if-test for å sjekke hvor mange ganger $\mu = 558$ forekommer i de 10000 forskjellige konfidensintervallene.

```
#Oppgave 1b.
count = 0
mu_exact = 558
sd_exact = 30
iterations = 10000

for (j in 1:iterations){
  data_sett = array(1:15)
  for (i in 1:15){
    X_est = rnorm(15 ,mu_exact, sd_exact)
    data_sett[i] = X_est
  }
  mu_est = mean(data_sett)
  sd_est = sd(data_sett)

  con_int_low = mu_est - 1.96*sd_est/sqrt(15)
  con_int_high = mu_est + 1.96*sd_est/sqrt(15)
  if (mu_exact >= con_int_low && mu_exact <= con_int_high){
    count = count + 1
  }
}
print(count/iterations)
```

Hvilket gir at forekomsten av $\mu = 558$ er

```
> print(count/iterations)
[1] 0.9266.
```

Altså forekommer $\mu = 558$ i rundt 93% av konfidensintervallene hvilket virker som et rimelig svar siden vi opperer med et relativt stort datasett og gjør sekvensielle beregninger av gjennomsnittet og standardavviket underveis som kan gi små avvik fra eksakte verdier for μ og σ.

c.) *Bruk nå i stedet det tilnærmede intervallet for store utvalg, altså*

$$\left(\bar{X} - 1.96\frac{S}{\sqrt{15}}, \bar{X} + 1.96\frac{S}{\sqrt{15}}\right)$$

med

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{15-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

og beregn dette for hvert av 10000 *datasett, generert som i b). Hvor stor andel av intervallene inneholder* $\mu = 558$*? Kommenter og forklar resultatet.*

Løsning:

I løsning av oppgaven endres det kun på estimeringsmetoden for hvert datasett slik at de numeriske tilnærmingene for μ og σ nå må legges til. Det kan se slik ut

```
#Oppgave 1c.)
count = 0
mu_exact = 558
sd_exact = 30
iterations = 10000

for (j in 1:iterations){
  data_sett = array(1:15)
  mu_est = 0
  for (i in 1:15){
    X_est = rnorm(15 ,mu_exact, sd_exact)
    data_sett[i] = X_est
  }
  mu_est = sum(data_sett)/15
  sd_est = 0
  for (i in 1:15){
    sd_est = sd_est + (data_sett[i] - mu_est)^2
  }
  sd_est = sqrt(sd_est/14)
  con_int_low = mu_est - 1.96*sd_est/sqrt(15)
  con_int_high = mu_est + 1.96*sd_est/sqrt(15)
  if (mu_exact >= con_int_low && mu_exact <= con_int_high){
    count = count + 1
  }
}
print(count/iterations)
```

Hvilket gir at forekomsten av $\mu = 558$ er

```
> print(count/iterations)
[1] 0.9287.
```

Som tilsier at den empiriske tilnærmingen for μ er rimelig siden den samsvarer med resultatet for den teoretiske μ i b.)

d.) *Trekk 10000 datasett som i b) og lag et* 95% *konfidensintervall for for hvert av dem. Hvor stor andel av intervallene inneholder* $\sigma = 30$*?*

Løsning:

For å konstruere et konfidensintervall for σ husker vi at

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} = \frac{n\hat{\sigma}^2}{\sigma^2}$$

hvor $\frac{n\hat{\sigma}}{\sigma^2} \sim \chi^2_n$ slik at $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Dermed vil et tilsvarende konfidensintervall være gitt ved

$$\chi^2_{1-\alpha\backslash 2} \leq X^2_n \leq \chi^2_{\alpha\backslash 2}$$

$$\Rightarrow \chi^2_{1-\alpha\backslash 2} \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{\alpha\backslash 2}$$

$$\Rightarrow \frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha\backslash 2}} \geq \sigma^2 \geq \frac{n\hat{\sigma}^2}{\chi^2_{\alpha\backslash 2}}$$

$$\Rightarrow \sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha\backslash 2}}} \geq \sigma \geq \sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{\alpha\backslash 2}}}$$

$$\Downarrow$$

$$\boxed{\left(\sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{\alpha\backslash 2}}}, \sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha\backslash 2}}}\right)}$$

Implementerer vi så dette resultatet kan det gi følgende løsning

```
#Oppgave 1d.)
count = 0
mu_exact = 558
sd_exact = 30
```

```
iterations = 10000

chi2.kvant.low = qchisq(0.025,df=14)
chi2.kvant.up = qchisq(0.975,df=14)

for (j in 1:iterations){
 data_sett = array(1:15)
 mu_est = 0
 for (i in 1:15){
  X_est = rnorm(15, mu_exact, sd_exact)
  data_sett[i] = X_est
 }
 mu_est = sum(data_sett)/15
 sd_est = 0
 for (i in 1:15){
  sd_est = sd_est + (data_sett[i] - mu_est)^2
 }

 sd_est = sqrt(sd_est/14)

 low.sd=(14)*sd_est^2/chi2.kvant.up
 up.sd=(14)*sd_est^2/chi2.kvant.low

 cont_int_low = sqrt(low.sd)
 cont_int_high= sqrt(up.sd)
 if (sd_exact >= cont_int_low && sd_exact <= cont_int_high){
  count = count + 1
 }
}

print(count/iterations)
```

Hvilket gir at forekomsten av $\sigma = 30$ er

```
> print(count/iterations)
[1] 0.9503.
```

Som igjen er veldig rimelig grunnet datasettet og de estimeringene vi gjør.

e.) *Under antakelsen om normalfordeling er* $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$, $i = 1, \ldots, n$, *med* $\mu = 558$ *og* $\sigma = 30$. *Anta nå at* $Z_1, \ldots, Z_{15}$ *i virkeligheten er t-fordelt med 7 frihetsgrader, altså* $Z_1, \ldots, Z_{15} \overset{uif}{\sim} t_7$. *Trekk nå* 10000 *datasett fra denne fordelingen ved å*

1. Trekke $z_1, \ldots, z_{15}$ fra t_7 med R-funksjonen rt().
2. La $x_i = \mu + \sigma z_i, \ i = 1, \ldots, n.$

Gjenta deretter oppgave b) med de nye datasettene. Hvor robust er metoden for å lage konfidensintervall for forventningsverdien for antakelsen om normalfordeling?

Løsning:

Likeledes som i oppgavene over lager vi en 1-dimensjonal array på samme størrelse som tidligere. Forskjellen herfra ligger i at vi henter verdier for Z_i fra t-fordelingen istedenfor normalfordelingen. Dermed vil en mulig løsning se slik ut

```
#Oppgave 1e.)

count = 0
mu_exact = 558
sd_exact = 30
iterations = 10000

for (j in 1:iterations){
  data_sett = array(1:15)
  mu_est = 0
  for (i in 1:15){
    Z_i = rt(15, 7)
    x_i = mu_exact + sd_exact*(Z_i)
    data_sett[i] = x_i
  }
  mu_est = mean(data_sett)
  sd_est = sd(data_sett)
  con_int_low = mu_est - 2.145*sd_est/sqrt(15)
  con_int_high = mu_est + 2.145*sd_est/sqrt(15)
  if (mu_exact >= con_int_low && mu_exact <= con_int_high){
      count = count + 1
    }
   }
print(count/iterations)
```

Hvilket gir at forekomsten av $\mu = 558$ er

```
> print(count/iterations)
[1] 0.948.
```

Dette sier oss at konfidensintervallmetoden er meget robust i henhold til antagelsen om normalfordeling fordi forekomsten $\mu = 558$ under begge analysene ligger mellom 93% og 95%.

f.) *Trekk datasett som i e) og lag deretter 95% konfidenintervall for standardavviket til X_i slik som i d). Merk imidlertid at $Var(Z_i) = \frac{7}{7-2}$ slik at variansen til X_i nå er $\tilde{\sigma^2} = Var(X_i) = Var(\mu + \sigma Z_i) = \sigma^2 Var(Z_i) = 1.4\sigma^2$. Det er altså $\tilde{\sigma}$ du skal lage konfidensintervall for, og sjekke andelen intervaller som inneholder $\tilde{\sigma}$. Sammenlign med resultatene fra d) og kommenter.*

Løsning:

Prosedyren i denne oppgaven bruker både fremgangsmåten fra e.) samt konfidensintervallet fra d.). Vi husker at et 95% konfidensintervall for σ er gitt ved

$$\left(\sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{\alpha,n}}}, \sqrt{\frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha,n}}} \right)$$

men siden vi ønsker å finne for $\tilde{\sigma} = 1.4\sigma$ transformeres intervallet til

$$\left(\sqrt{\frac{n\hat{\sigma}^2}{1.4\chi^2_{\alpha,n}}}, \sqrt{\frac{n\hat{\sigma}^2}{1.4\chi^2_{1-\alpha,n}}} \right)$$

hvor, som sist, $\hat{\sigma} = S$. Dette kan implementers følgelig

```
 #Oppgave 1f.)
 count = 0
mu_exact = 558
sd_exact = 30
iterations = 10000
chi_low = qchisq(0.025, (14), lower.tail=FALSE)
chi_high = qchisq(0.025, (14), lower.tail=TRUE)

for (j in 1:iterations){
  data_sett = array(1:15)
  mu_est = 0
  for (i in 1:15){
    Z_i = rt(15, 7)
    x_i = mu_exact + sd_exact*(Z_i)
    data_sett[i] = x_i
```

```
 }
  mu_est = mean(data_sett)
  sd_est = 0

  for(i in 1:15){
    sd_est = sd_est + (data_sett[i] - mu_est)^2
  }

  sd_tilde = 1.4*(sd_est/14)

  low.sd= (14)*sd_tilde/chi_low
  up.sd= (14)*sd_tilde/chi_high

  cont_int_low = sqrt((1/1.4)*low.sd)
  cont_int_high = sqrt((1/1.4)*up.sd)

  if (sd_exact >= cont_int_low && sd_exact <= cont_int_high){
      count = count + 1
    }
 }
print(count/iterations)
```

Hvilket gir at forekomsten av $\sigma = 30$ er

```
> print(count/iterations)
[1] 0.7927.
```

Her kan vi observere at modellen for konfidensintervall for σ ved hjelp av t-fordelingsmetoden ikke er like god ei robust som for normalfordelingen ettersom forekomsten av σ er markant lavere.

Oppgave 2:

a.) *Lag et boksplott som viser fordelingen av observasjonene. Kommenter hva du finner.*

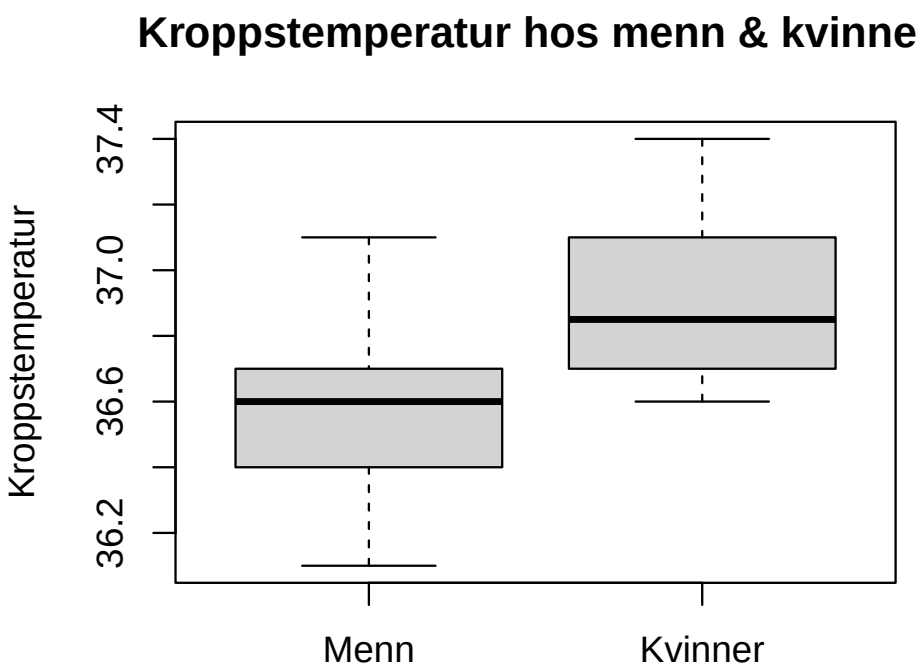
Løsning:

De gitte dataene gir følgende boksplott

```
#Oppgave 2a.)
```

```
data <-
    read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/temp.txt",
header=T)

boxplot(data, ylab= "Kroppstemperatur",
      main="Kroppstemperatur hos menn & kvinner")
```



Det kan observeres fra plotet at mediantemperaturen til kvinner er noe høyere enn menns (jmf. den svarte streken på plotet), samt at de lavere verdiene for kroppstemperaturen til kvinner ligger rundt medianen til mennene.

b.) *Lag normalfordelingsplott for de to observasjonssettene, altså ett for menn og ett for kvinner. Kommenter hva du ser.*

Løsning:

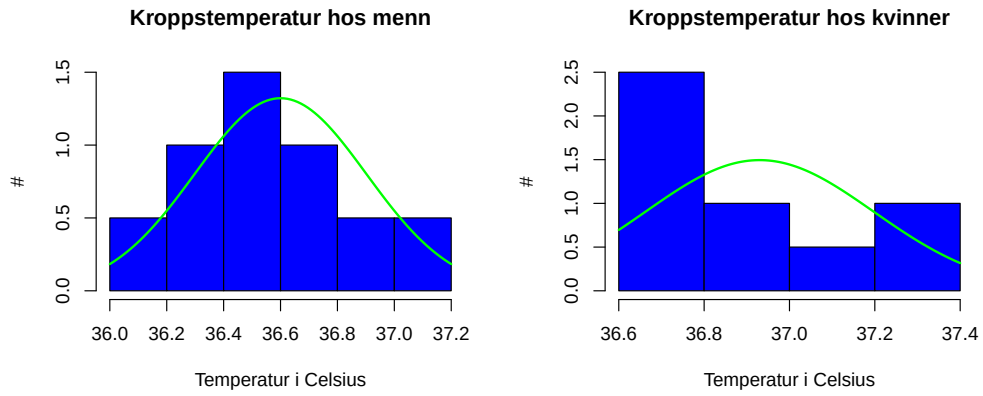
For å sjekke om dataene er realiseringer av normalfordelingen kan det være beleilig å enten lage et normalsannsynlighets (Q-Q)plot eller simpelthen plotte et histogram over fordelingen med den respektive normalfordelingskurven. I løsning av oppgaven velger vi det sist nevnte. Det gir

```
 #Oppgave 2b.)
 hist(data$Menn,
     freq = FALSE,
     col="blue",
     xlab="Temperatur i Celsius",
     ylab="#",main="Kroppstemperatur hos menn")

curve(dnorm(x,
     mean = mean(data$Menn),
     sd = sd(data$Menn)),
     col = "green",
     lwd = 2,
     add = TRUE)

hist(data$Kvinner,
   freq = FALSE,
   col="blue",
   xlab="Temperatur i Celsius",
   ylab="#",main="Kroppstemperatur hos kvinner")

curve(dnorm(x,
     mean = mean(data$Kvinner),
     sd = sd(data$Kvinner)),
     col = "green",
     lwd = 2,
     add = TRUE)
```



Figur 1: Histogram og normalplott av dataene

Hvor vi observerer fra plottene at hverken dataene for menn eller kvinner er eksakt normalfordelte, men at antagelsen om normalfordeling hos menn er betraktelig bedre enn hos kvinner (jmf. jevnere haler og hovedvekten av dataene faller inn under gjennomsnittet).

c.) *Anta at variansen er den samme for de to utvalgene, og test med signifikansnivå* 5% *om det er noen forskjell i forventet kroppstemperatur. Beregn P-verdien, og lag et* 95% *konfidensintervall for denne forskjellen.*

Løsning:

For å løse oppgaven er det fornuftig å benytte seg av hypotesetesting for å sjekke om det er rimelig å anta at det forskjellig forventet kroppstemperatur mellom kvinner og menn. Dermed har vi hypotesene

$$H_0 : \mu_k - \mu_m = 0$$

$$H_a : \mu_k - \mu_m \neq 0$$

.

Siden vi antar normalfordeling på dataene benytter vi oss av en two sample T-test, gitt ved

$$T = \frac{\bar{X} - \bar{Y} - (\mu_k - \mu_m)}{\sqrt{S_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}}.$$

Hvor, i vårt tilfelle $m = n$ siden utvalgsstørrelsen er like store. Videre er S_p^2 gitt ved

$$\begin{aligned} S_p^2 &= \frac{m-1}{m+n-2}S_1^2 + \frac{m-1}{m+n-2}S_2^2 \\ &= \frac{m-1}{2m-2}\left(S_1^2 + S_2^2\right) \\ &= \frac{1}{2}S_1^2 + \frac{1}{2}S_2^2 \end{aligned}$$

hvor vi vet at S_1^2, S_2^2 er utvalgsvariansen til de respektive dataene. Herfra bruker vi R til å teste for tilfelle hvor H_0 er antatt riktig.

```
#Oppgave 2c.)

m = 10

SP = ((m-1)/(2*m-2))*(var(data$Kvinner)+var(data$Menn))

Tvalue = (mean(data$Kvinner)-mean(data$Menn))/(sqrt(SP*(2/m)))

print(round(c(SP, Tvalue), 4))
```

som gir verdiene

```
> print(round(c(SP, Tvalue), 4))
[1] 0.0812 2.5901.
```

Siden variansen er antatt lik for de utvalgene kan vi nå finne den kritiske t-verdien med $\alpha = 0.05$ og frihetsgrader

$$\begin{aligned} \nu &= \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2 \backslash m\right)^2}{m-1} + \frac{\left(S_2^2 \backslash n\right)^2}{n-1}} \\ &= \frac{\left(2\frac{S^2}{m}\right)^2}{\frac{2\left(S^2 \backslash m\right)^2}{m-1}}, \qquad \boxed{S^2 = S_1^2 + S_2^2} \\ &= \frac{4}{\frac{2}{m-1}} = 2(m-1) = 18. \end{aligned}$$

Bruker vi så t-tabellen gir det

$$t_{\text{crit}} = t_{\alpha \backslash 2, 18} = 2.1009.$$

For å kunne avvise nullhypotesen med 95% sikkerhet ($\alpha = 0.05$) må følgende likning være oppfylt

$$|t_{\text{crit}}| \leq T$$

$$2.1009 \leq 2.5901.$$

Dette er grunnet at differansen mellom T og 0 (vår nullhypotese) er større enn t_{crit}. Sluttvis ønsker vi å finne P-verdien og konstruere et 95% konfidensintervall for forskjellen i forventet kroppstemperatur. P-verdien til utvalgene korresponderer til sanmsynligheten for at nullhypotesen er sann gitt en signifikansverdi. Altså vil nullhypotesen påstås sann dersom P-verdien er større enn α.

Vi har at generelle P-verdier er lik arealet til venstre og høyre for T-kurven. Altså

$$P_{\text{Verdi}} = P(-2.59 \leq T \leq 2.59)$$

og siden T-kurven i dette tilfellet er symmetrisk om midtpunktet vil arealet på venstre og høyresiden være like. Dermed kan vi skrive P-verdien som

$$P_{\text{Verdi}} = 2P(T \leq 2.59)$$

Herfra kan vi simulere løsning ved hjelp av R

```
pvalue = 2*pt(Tvalue, df=18, lower.tail = F)
```

```
> print(pvalue)
[1] 0.01848131
```

hvor vi observerer at $P_{\text{Verdi}} \leq \alpha$ som betyr at vi kan avvise nullhypotesen med 95% sikkerhet. Herfra skal vi konstruere et 95% konfidensintervall for forksjellen i forventet kroppstemperatur, med $\alpha = 0.05$. For en vilkårlig pooled t-test gir det

$$\left(\bar{X} - \bar{Y}\right) \pm t_{\alpha\backslash 2,\nu}\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$$

hvor $m = n = 10$ og $\nu = 18$. Løser vi med R får vi

```
CF = qt(0.025, 18, lower.tail = FALSE)
error = CF*sqrt((var(data$Kvinner) + var(data$Menn))/10)
diff = mean(data$Kvinner)-mean(data$Menn)
leftside <- diff - error
rightside <- diff + error
print(c(leftside, rightside))
```

```
> print(c(leftside, rightside))
[1] 0.06232131 0.59767869
```

og dermed kan vi være 95% sikkre på at

$$0.06232131 \leq \bar{X} - \bar{Y} \leq 0.59767869.$$

Kan tilleggsvis sjekke med t.test() i R at alt vi har beregnet er korrekt.

```
t.test(x = data$Kvinner,
       y=data$Menn,
       mu = 0,
       paired = FALSE,
       var.equal = TRUE,
       conf.level = 0.95)
```

```
Two Sample t-test

data:  data$Kvinner and data$Menn
t = 2.5901, df = 18, p-value = 0.01848

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:

 0.06232131 0.59767869

sample estimates:
mean of x mean of y
    36.93     36.60
```

d.) *Gjennomfør testen og beregn P-verdien også i det tilfellet der en ikke antar felles varians. Diskuter og forklar resultatene.*

Løsning:

Når man ikke antar felles varians må vi beregne de nye frihetsgradene gitt

ved

$$\nu = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{\left(S_1^2 \backslash m\right)^2}{m-1} + \frac{\left(S_2^2 \backslash n\right)^2}{n-1}}$$

$$= \frac{\left(\frac{S_1^2+S_2^2}{m}\right)^2}{\frac{\frac{S_1^4+S_2^4}{m^2}}{m-1}}.$$

T verdien er den samme som i oppgave c.), og følgende R script kan lages

```
v = ((var(data$Menn)
   +var(data$Kvinner))/m)^2/(((var(data$Menn)^2 +
   var(data$Kvinner)^2)/(10^2))/(9))

pvalue2 = 2*pt(Tvalue, df=v, lower.tail = F)

CF2 = qt(0.025, v, lower.tail = FALSE)

error2 = CF2*sqrt((var(data$Kvinner) + var(data$Menn))/m)

diff = mean(data$Kvinner)-mean(data$Menn)

l2<- diff - error2
r2 <- diff + error2
```

```
> print(pvalue2)
[1] 0.01863038
> print(c(l2, r2))
[1] 0.06203301 0.59796699
```

Sjekker så om dette stemmer overens med t.test()

```
t.test(x = data$Kvinner, y=data$Menn, mu = 0, paired = FALSE,
   var.equal = FALSE, conf.level = 0.95)
```

```
  Welch Two Sample t-test

  data:  data$Kvinner and data$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
```

hvor vi observerer at verdiene stemmer godt overens selv om testen runder av frihetsgraden til tre desimaler og man således kan få små avrundingsfeil. En kan også observere at nullhypotesen kan forkastes i dette tilfellet også.

e.) *Utled og gjennomfør en F-test for å sjekke om det er noen grunn til å påstå at variansene er forskjellige. Sjekk mot var.test() i R .*

Løsning:

I denne oppgaven skal vi sjekke om vi med pålitelig grunnlag kan påstå om variansene er forskjellige. Dette gjør vi ved hjelp av en F-test. Observatoren til F-testen er da gitt ved

$$F = \frac{S_1^2 \backslash \sigma_1^2}{S_2^2 \backslash \sigma_2^2}$$

der

$$\nu_1 = m - 1, \ \nu_2 = n - 1.$$

Konstruerer så hypotesen

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

hvor vi observerer at nullhypotesen vil anta at forholdet mellom S_1^2 og S_2^2 er 1. Altså kan vi skrive

$$f = \frac{S_1^2}{S_2^2} \sim f_{m-1,n-1}.$$

Setter vi så inn for verdien til variansen for begge kjønn gir det

$$f = 1.279.$$

Dermed kan vi ved en two-tailed test avvise nullhypotesen dersom

$$f \geq F_{\alpha\backslash 2, m-1, n-1}$$

som ved hjelp av R gir oss

```
Fratio = var(data$Menn)/var(data$Kvinner)
alpha = 0.05
fcrit = qf(alpha/2,9,9, lower.tail = F)
```

hvor, ettersom vi benytter oss av den største variansen bruker en right-tailed test, som gir verdiene

```
> print(c(Fratio, fcrit))
[1] 1.279251 4.025994.
```

Her er det åpenbart at $|f_{\text{crit}}| > F$ som betyr at vi ikke kan avvise nullhypotesen, og den korresponderende P-verdien er

$$P_{\text{Verdi}} = 2P(F \geq f_{\text{crit}})$$

som ved hjelp av R gir

```
Fratio = var(data$Menn)/var(data$Kvinner)

PvalueF = 2*pf(Fratio, df1=9, df2=9, lower.tail = F)
```

```
> print(PvalueF)
[1] 0.7196901.
```

Sjekker sluttvis resultatene opp mot var.test() funksjonen

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

```
  F test to compare two variances

  data:  data$Menn and data$Kvinner

  F = 1.2793, 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:
   0.3177479 5.1502577
```

```
sample estimates:
ratio of variances
          1.279251
```

som viser at verdiene våre stemmer overens med de fra testen.

f.) *Se nå på situasjonen der en vurderer å innhente to nye målinger. La* X_{11} *være verdien for kvinnen og* Y_{11} *verdien for mannen, slik at forskjellen er* $X_{11}Y_{11}$*. Vi antar nå at alle observasjonene er normalfordelte med samme varians. Begrunn at et rimelig anslag for* $X_1 1 Y_{11}$ *er differansen mellom gjennomsnittet av de 10 eksisterende målingene for kvinner og menn, altså* $\bar{X} - \bar{Y}$*. Hva er fordelingen til* $X_{11}Y_{11} - (\bar{X} - \bar{Y})$*? Bruk dette til å lage et* 95% *prediksjonsintervall for* $X_{11}Y_{11}$*, altså et intervall som med sannsynlighet* 0.95 *inneholder* $X_{11}Y_{11}$*. Dette er gjennomgått for ett-utvalgs-situasjonen på forelesning. Forklar hva som er forskjellen mellom et slikt intervall og et konfidensintervall for* $\mu_1\mu_2$ *. Hvordan skal et prediksjonsintervall tolkes? [Hint: Siden alle variablene er normalfordelte, er* $X_{11}Y_{11} - (\bar{X} - \bar{Y})$ *også det. Det er derfor nok å beregne forventning og varians for å finne fordelingen til denne størrelsen.]*

Løsning:

Vi antar at variansen mellom menn og kvinner lik og vi ønsker å lage et prediksjonsintervall for forskjellen $X_{11} - Y_{11}$. Det gir at $E[X_{11} - Y_{11}] = \mu_X - \mu_Y$ med respektive estimator $\tilde{X}_n - \tilde{Y}_n$ for utvalgsmengden $m = n = 10$. Siden vi har antatt at datamengden er realiseringer av normalfordelingen vil likeledes en linjærkombinasjon av stokatiske variable $X, Y \sim N(\mu, \sigma)$ og gjennomsnittene deres $\bar{X}, \bar{Y} \sim N(\mu, \sigma)$ være normalfordelte. Altså er $X_{11}Y_{11} - (\bar{X} - \bar{Y}) \sim N(\mu, \sigma)$.

Oppgave 3:

a.) *Er forskjellen mellom mødre og fedre signifikant? Formuler hypoteser, beregn en p-verdi, og konkluder. Kommenter kort.*

Løsning:

Det er rimelig å anta at vi arbeider med en binomisk problemstilling ettersom oppgaveteksten er formulert i en "ja, nei" form eller $P, \bar{P}$. Dermed

kan vi si at fedre som svarer ja er representert ved

$$X \sim \text{Bin}(3000, 0.162)$$

hvor $n = 3000$ og $p = 0.162$.På samme måte for mødre er

$$Y \sim \text{Bin}(3000, 0.147).$$

For å sjekke om det er en signifikant forskjell mellom mødre og fedre kan vi konstruere følgende hypotese

$$H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 \neq 0.$$

Hvor p_1, p_2 svarer til de respektive sannsynlighetene for at en person fra hvert kjønn svarer ja. Videre har vi også estimatorene $\hat{p_1}, \hat{p_2}$

$$\hat{p_1} = \frac{X}{n},\ \hat{p_2} = \frac{Y}{m}$$

for p_1, p_2 som er forventingsrette

$$\begin{aligned} E\left(\hat{p_1} - \hat{p_2}\right) &= E\left(\frac{X}{n} - \frac{Y}{m}\right) \\ &= \frac{1}{n} n p_1 - \frac{1}{m} m p_2 \\ &= p_1 - p_2 \end{aligned}$$

og man kan dermed skrive Z-funksjonen (scoren) som

$$Z = \frac{\hat{p_1} - \hat{p_2} - E\left(\hat{p_1} - \hat{p_2}\right)}{\text{SD}(\hat{p_1} - \hat{p_2})}.$$

Variansen til $\hat{p_1} - \hat{p_2}$ er gitt ved

$$\begin{aligned} \text{Var}(\hat{p_1} - \hat{p_2}) &= \text{Var}\left(\frac{X}{n}\right) + \text{Var}\left(\frac{Y}{m}\right), \qquad \boxed{\text{Var(-Y)} = \text{Var(Y)}} \\ &= \frac{1}{n^2} n p_1 (1 - p_1) + \frac{1}{m^2} m p_2 (1 - p_2) \\ &= \frac{p_1(1 - p_1)}{n} + \frac{p_2(1 - p_2)}{m} \end{aligned}$$

og siden nullhypotesen antar $p_1 - p_2 = 0$ gir det sluttvis

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - E\left(\hat{p}_1 - \hat{p}_2\right)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}}.$$

Bruker vi forutsetningen for nullhypotesen ser vi at $p_1 = p_2 = p$, og siden $n = m$ omskrives nevneren til

$$\sqrt{\frac{2p(1-p)}{n}}.$$

Men siden p er ukjent må den approksimeres ved

$$\hat{p} = \frac{1}{2}\left(\hat{p}_1 + \hat{p}_2\right).$$

Finurlig nok er som nevnt utavlgsmengdene like store som betyr at de "veier" like mye og da holder det å kun ta gjennomsnittet av de. Det gir

$$\hat{p} = \frac{1}{2}\left(0.162 + 0.147\right) = 0.1545$$

Bruker vi R for å løse siste del har vi

```
#Oppgave 3a.)
N = 3000 X = 486
Y = 441
p1 = X/N
p2 = Y/N
phat = (1/2)*(p1 + p2)
qhat = (1 - phat)
Z = (p1-p2)/(sqrt(phat*qhat*(2/N)))
```

```
>print(Z)
[1] 1.60737
```

og tilslutt sjekker vi P-verdien for å bestemme om vi skal forkaste nullhypotesen eller ei

$$P_{\text{Verdi}} = 2P(Z \leq 1.60737)$$

som igjen ved hjelp av R gir

```r
#Oppgave 3a.)
pvalue = 2*pt(Z, df=5998, lower.tail = F)
```

```
> pvalueopg3
[1] 0.1080259
```

og vi kan konkludere med at vi kan forkaste nullhypotesen dersom signifikansverdien er høyere enn 0.108 eller mer presist at $\alpha = 0.1 \vee \alpha < 0.1$.

b.) *Kontroller svaret ditt ved å bruke prop.test() i R.*

Løsning:

Bruker prop.test() i R og får

```r
#Oppgave 3b.)
prop.test( x = c(486,441) ,
        n = c(3000, 3000),
        alternative = "two.sided",
        correct = FALSE)
```

```
2-sample test for equality of proportions without continuity correction
              data:
c(486, 441) out of c(3000, 3000)

X-squared = 2.5836, df = 1, p-value = 0.108

alternative hypothesis: two.sided
95 percent confidence interval:
 -0.003286474  0.033286474
sample estimates:
prop 1 prop 2
0.162 0.147
```

Hvor vi observerer at P-verdien stemmer overens med det vi fikk i a.)

$$\int_a^b f(x)\, dx = G(\gamma) - F(\alpha)$$