

STK1110 - Statistiske metoder og dataanalyse høsten 2022

Løsningsforslag til oppgaver i læreboka for uke 37

Lærebok: *Modern Mathematical Statistics with Applications (3rd ed.)* - Devore, Berk & Carlton

Exercise 6.29

Vi vet at hvis $U_1, \dots, U_n$ er uavhengige og χ^2 -fordelte stokastiske variabler med $k_1, \dots, k_n$ frihetsgrader, så er summen av disse variablene χ^2 -fordelt med $k_1 + \dots + k_n$ frihetsgrader.

Så hvis $X \sim \chi_\nu^2$, kan vi skrive $X = \sum_{i=1}^{\nu} X_i$, der $X_i \stackrel{u.i.f.}{\sim} \chi_1^2$.

Her er $\mu = E(X_i) = 1$ og $\sigma^2 = V(X_i) = 2$ (jf. side 315 i læreboka).

Siden X_i -ene er uavhengige og identisk fordelte stokastiske variabler, kan vi bruke sentralgrenseteoremet (jf. side 298 i læreboka)

$$\lim_{\nu \rightarrow \infty} P\left(\frac{X - \nu}{\sqrt{2\nu}} \leq z\right) = \lim_{\nu \rightarrow \infty} P\left(\frac{\sum_{i=1}^{\nu} X_i - \nu\mu}{\sqrt{\nu\sigma}} \leq z\right) = \Phi(z),$$

der $\Phi(z)$ er den kumulative standardnormalfordelingen.

Dette betyr at $X \sim \chi_\nu^2$ er tilnærmet normalfordelt, når antall frihetsgraden ν er stort.

Exercise 6.30

På samme måte som i Exercise 6.48 kan vi uttrykke $X \sim \chi_\nu^2$ som $X = \sum_{i=1}^{\nu} X_i$, der $X_i \stackrel{u.i.f.}{\sim} \chi_1^2$.

Dermed har vi $\frac{X}{\nu} = \frac{1}{\nu} \sum_{i=1}^{\nu} X_i = \bar{X}$.

Videre, siden $X_i \sim \chi_1^2$, vet vi at $\mu = E[X_i] = 1$ og $\sigma^2 = V(X_i) = 2$.

Ifølge store talls lov (jf. side 324 i læreboka), har vi da

$$\frac{X}{\nu} = \bar{X} \rightarrow E[X_i] = 1, \quad \text{når } \nu \rightarrow \infty.$$

Exercise 8.24

a) Med følgende kode lager vi et Q-Q plot, som er vist i Figur 1.

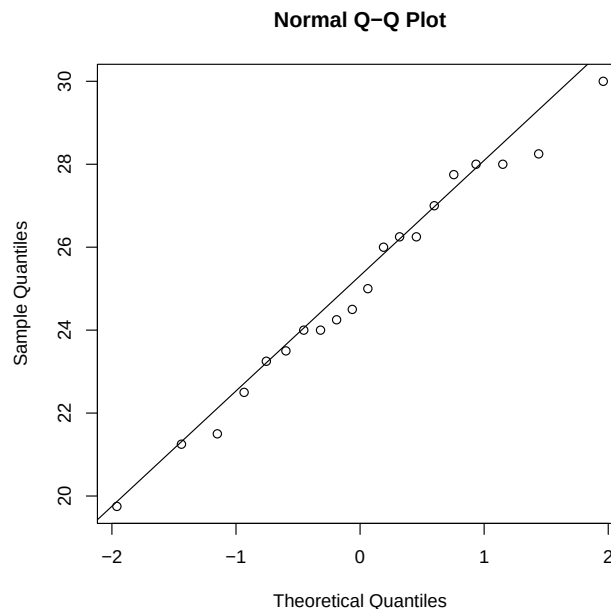
```
> # Leser data fra web.  
> ACT.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe8-34.txt", header = TRUE)  
> # Trekker ut vektoren med ACT-scorene.  
> ACT.data = ACT.data$ACT
```

```

>
> # QQ plot
> qqnorm(ACT.data)
> qqline(ACT.data)

```

Fra Q-Q plotet kan vi se at punktene grovt sett ligger på en rett linje. Dermed konkluderer vi at det er rimelig å anta normalfordeling.



Figur 1: Q-Q plot for ACT scorene.

b) Basert på resultatet fra punkt a), antar vi at dataene er normalfordelte.

Ifølge side 403 i læreboka er et $100(1 - \alpha)\%$ konfidensintervall for μ i dette tilfellet gitt ved

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right],$$

hvor $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ og $S = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Ved å bruke dataene gitt i oppgaven og $\alpha = 0.05$, får vi konfidensintervallet $[23.79, 26.31]$ ved R-koden nedenfor. (Merk at vi i R får den kritiske verdien $t_{\frac{\alpha}{2}, n-1}$ ved kommandoen `qt(1-alpha/2, df = n-1)`.)

```

1 > # Estimerer for mu og sigma.
2 > mu.hat = mean(ACT.data)
3 > sigma.hat = sd(ACT.data)
4 >
5 > n = length(ACT.data)
6 > alpha=0.05
7 >
8 > # Konfidensintervall for mu
9 > mu.CI = c(
10 +   mu.hat - qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n),
11 +   mu.hat + qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n)
12 + )
13 > print(mu.CI)
14 [1] 23.79121 26.30879

```

c) 21 ligger lavere enn den nedre konfidensgrensen av konfidensintervallet vi fant i punkt b). Så vi konkluderer at kalkulusstudentene i gjennomsnitt har en bedre ACT score enn en gjennomsnittlig student.

Exercise 8.28

a) Ifølge side 403 i læreboka er et $100(1 - \alpha)\%$ konfidensintervall for μ i dette tilfellet gitt ved

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right],$$

hvor $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ og $S = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Ved å bruke dataene gitt i oppgaven og $\alpha = 0.05$, får vi konfidensintervallet $[0.89, 0.96]$ ved R-koden nedenfor. (Merk at vi i R får den kritiske verdien $t_{\frac{\alpha}{2}, n-1}$ ved kommandoen `qt(1-alpha/2, df = n-1)`.)

```
> # Leser data fra web.
> cadence.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe8-38.txt",
>
> # Trekker ut vektoren med verdiene for ‘‘cadence’’ (antall skritt per sekund)
> cadence.data = cadence.data$cadence
>
> # Estimerer for mu og sigma.
> mu.hat = mean(cadence.data)
> sigma.hat = sd(cadence.data)
>
> n = length(cadence.data)
> alpha=0.05
>
> # Konfidensintervall for mu
> mu.CI = c(
+   mu.hat - qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n),
+   mu.hat + qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n)
+ )
> print(mu.CI)
[1] 0.8876158 0.9633842
```

At et 95% konfidensintervall for μ = “forventet antall skritt per sekund (cadence)” er $[0.89, 0.96]$, betyr at vi er 95% sikre på at μ ligger mellom 0.89 skritt pr sekund og 0.96 skritt per sekund. (Her betyr formuleringen “95% sikre” (engelsk: “95% confident”) at hvis vi gjentatte ganger hadde beregnet konfidensintervall ved hjelp av formelen ovenfor, ville 95% av intervallene inneholdt den sanne verdien av μ .)

b) Ifølge side 405 i læreboka er et $100(1 - \alpha)\%$ prediksjonsintervall for en ny observasjon er gitt ved

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} S \sqrt{1 + \frac{1}{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} S \sqrt{1 + \frac{1}{n}} \right]$$

Ved å bruke dataene gitt i oppgaven og $\alpha = 0.05$, får vi prediksjonsintervallet $[0.75, 1.10]$ ved følgende R-kode.

```
> # Prediksjonsintervall for en ny observasjon
> mu.PI = c(
+   mu.hat - qt(1-alpha/2, df = n-1)*sigma.hat*sqrt(1+1/n),
+   mu.hat + qt(1-alpha/2, df = n-1)*sigma.hat*sqrt(1+1/n)
+ )
> print(mu.PI)
[1] 0.7518927 1.0991073
```

At et 95% prediksjonsintervall for en ny observasjon er $[0.75, 1.10]$, betyr at vi er 95% sikre på at en ny observasjon vil ligge mellom 0.75 skritt pr sekund og 1.10 skritt per sekund. (Her betyr formuleringen “95% sikre” (engelsk: “95% confident”) at hvis vi gjentatte ganger hadde beregnet prediksjonsintervall ved hjelp av formelen ovenfor, ville 95% av intervallene inneholdt den nye observasjonen.)

Exercise 8.60

Basert på resultatet fra punkt a) i Exercise 8.34, antar vi at dataene er normalfordelte.

Ifølge side 410 i læreboka er et $100(1 - \alpha)\%$ konfidensintervallet for σ i dette tilfellet gitt ved

$$\left[S \sqrt{\frac{n-1}{\chi_{\frac{\alpha}{2}, n-1}^2}}, S \sqrt{\frac{n-1}{\chi_{1-\frac{\alpha}{2}, n-1}^2}} \right]$$

Ved å bruke dataene gitt i Exercise 8.34 og $\alpha = 0.01$, får vi konfidensintervallet $[1.89, 4.48]$ ved R-koden nedenfor. (Merk at vi i R får den kritiske verdien $\chi_{\frac{\alpha}{2}, n-1}^2$ ved kommandoen `qchisq(p=1-alpha/2, df=n-1)`, og vi får den kritiske verdien $\chi_{1-\frac{\alpha}{2}, n-1}^2$ ved kommandoen `qchisq(p=alpha/2, df=n-1)`.)

```
> # Leser data fra web.
> ACT.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe8-34.txt", header=TRUE)
> # Trekker ut vektoren med ACT-scorene.
> ACT.data = ACT.data$ACT
>
> # Estimat for sigma.
> sigma.hat = sd(ACT.data)
>
> n = length(ACT.data)
> alpha=0.01
>
> # Konfidensintervall for sigma
> sigma.CI = c(
+   sigma.hat*sqrt((n-1)/qchisq(p=1-alpha/2, df=n-1)),
+   sigma.hat*sqrt((n-1)/qchisq(p=alpha/2, df=n-1))
+ )
> print(sigma.CI)
[1] 1.887463 4.481447
```

Ifølge kommentaren nederst på side 410 i læreboka, krever konfidensintervallet at dataene er uavhengige og (tilnærmet) normalfordelte. (Merk at konfidensintervallet for μ gjelder også for data som ikke er normalfordelte hvis n er tilstrekkelig stor; jf. nederst side 406 i læreboka.)

Exercise 8.61

a) Med følgende kode lager vi et Q-Q plot, som er vist i Figur 2.

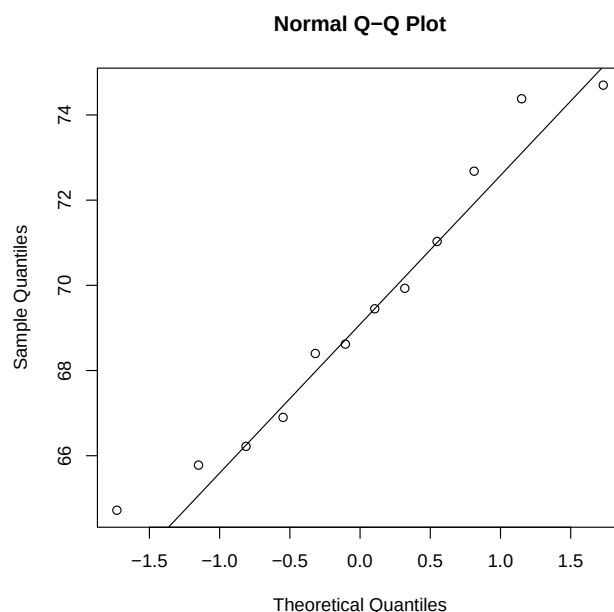
```
> # Leser data fra web.
> conductor.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe8-47.txt", header=TRUE)
> # Sjekker dataene.
> head(conductor.data)
  conductor time
1  Bernstein 71.03
2    Bohm    72.68
3 Furtwangler 74.38
4   Karajan  66.90
5 Leinsdorf  65.78
6    Mazur   69.45
```

```

> # Trekker ut vektoren med tidene.
> conductor.data = conductor.data$time
>
> # QQ plot
> qqnorm(conductor.data)
> qqline(conductor.data)

```

Fra Q-Q plotet kan vi se at punktene grovt sett ligger på en rett linje. Dermed konkluderer vi at det er rimelig å anta normalfordeling.



Figur 2: Q-Q plot for data fra Exercise 8.61.

b) Basert på resultatet i punkt fra a), antar vi at dataene er normalfordelte.

Ifølge side 410 i læreboka er et $100(1 - \alpha)\%$ konfidensintervallet for σ i dette tilfellet gitt ved

$$\left[S \sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}, n-1}}}, S \sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}, n-1}}} \right]$$

Ved å bruke dataene gitt i oppgaven og $\alpha = 0.05$, får vi konfidensintervallet $[2.34, 5.60]$ ved R-koden nedenfor. (Merk at vi i R får den kritiske verdien $\chi^2_{\frac{\alpha}{2}, n-1}$ ved kommandoen `qchisq(p=1-alpha/2, df=n-1)`, og vi får den kritiske verdien $\chi^2_{1-\frac{\alpha}{2}, n-1}$ ved kommandoen `qchisq(p=alpha/2, df=n-1)`.)

```

> # Estimat for sigma.
> sigma.hat = sd(conductor.data)
>
> n = length(conductor.data)
> alpha=0.05
>
> # Konfidensintervall for sigma
> sigma.CI = c(
+   sigma.hat*sqrt((n-1)/qchisq(p=1-alpha/2, df=n-1)),
+   sigma.hat*sqrt((n-1)/qchisq(p=alpha/2, df=n-1))
+ )
> print(sigma.CI)

```

```
[1] 2.335524 5.597773
```

At et 95% konfidensintervall for σ er $[2.34, 5.60]$, betyr at vi er 95% sikre på at σ ligger mellom 2.34 og 5.60 minutter. (Her betyr formuleringen “95% sikre” (engelsk: “95% confident”) at hvis vi gjentatte ganger hadde beregnet konfidensintervall ved hjelp av formelen ovenfor, ville 95% av intervallene inneholdt den sanne verdien av σ .)

c) Vi ser av konfidensintervallet i punkt b) at vi er 95% sikre på at σ ligger 2.34 og 5.60 minutter. Altså er det variasjon i hvor lang tid det tar å spille Beethvens niende symfoni, så det er ikke tilfellet at tiden det tar er helt bestemt av komponistens nedtegnning av verket.

Exercise 8.65

a) Ifølge side 403 i læreboka er et $100(1 - \alpha)\%$ konfidensintervall for μ i dette tilfellet gitt ved

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right],$$

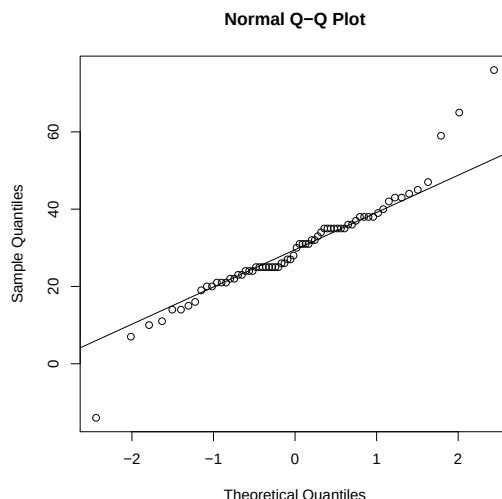
$$\text{hvor } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ og } S = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Ved å bruke dataene gitt i oppgaven og $\alpha = 0.05$, får konfidensintervallet $[26.61, 32.94]$ ved R-koden nedenfor. (Merk at vi i R får den kritiske verdien $t_{\frac{\alpha}{2}, n-1}$ ved kommandoen `qt(1-alpha/2, df = n-1)`.)

```
> # Leser data fra web.
> wtgain.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe8-51.txt",
> # Trekker ut vektoren med vektene.
> wtgain.data = wtgain.data$wtgain
>
> # Estimerer for mu og sigma.
> mu.hat = mean(wtgain.data)
> sigma.hat = sd(wtgain.data)
>
> n = length(wtgain.data)
> alpha=0.05
>
> # Konfidensintervall for mu, basert på t-fordeling
> mu.CI = c(
+   mu.hat - qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n),
+   mu.hat + qt(1-alpha/2, df = n-1)*sigma.hat/sqrt(n)
+ )
> print(mu.CI)
[1] 26.61491 32.94391
```

b) Med følgende kode lager vi et Q-Q plot, som er vist i Figur 3.

```
> # QQ plot
> qqnorm(wtgain.data)
> qqline(wtgain.data)
```



Figur 3: Q-Q plot for data fra Exercise 8.65.

Fra Q-Q plotet kan vi se at punktene i sentrumsregionen ligger på linjen, men én lav og noen høye verdier avviker fra linjen. Så man kan sette et spørsmåltegn ved om dataene er normalfordelte. Men siden $n = 68$ er ganske mange observasjoner, vil sentralgrenseteoremet tre i kraft slik at konfidensintervallet er tilnærmet gyldig (jf. side 392 i læreboka).

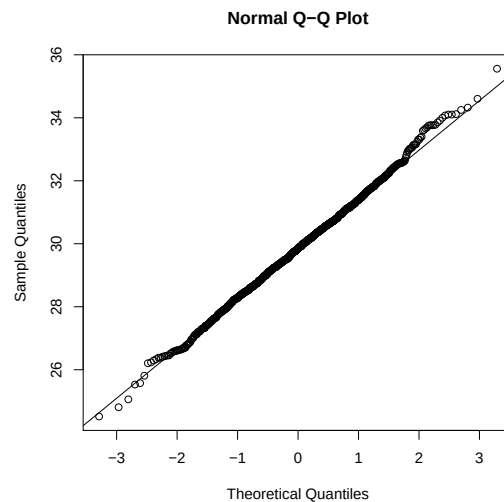
c) og d) Vi bruker R-pakken `boot` til å generere 999 bootstrap utvalg og beregne gjennomsnittet for hvert av dem. Vi bestemmer så bootstrap estimatet for standarfeilen og bruker det til å bestemme et tilnærmet 95% konfidensintervall; jf. intervallet gitt øverst på side 414 i læreboka. Se R-koden.

```
> # c)
> # Vi vil bruke R-pakken "boot".
> library(boot)
>
> # Vi definerer en funksjon som skal brukes til å beregne gjennomsnittene for bootstrap-utvalgene
> my.mean = function(x, indices) {mean(x[indices])}
>
> # Vi trekker 999 bootstrap utvalg.
> set.seed(1)
> mean.boot = boot(data=wtgain.data, statistic=my.mean, R=999)
>
> # d)
> # Vi bestemmer bootstrap estimatet for standardfeilen og bruker det til å bestemme et tilnærmet 95% konfidensintervall
> se.boot=sd(mean.boot$t)
> mu.CI.boot = c(
+   mu.hat - 1.96*se.boot,
+   mu.hat + 1.96*se.boot,
+ )
> print(mu.CI.boot)
[1] 26.77541 32.78341
```

e) Med følgende kode lager vi et Q-Q plot, som er vist i Figur 4.

```
> # e)
> # QQ plot
> qqnorm(mean.boot$t)
> qqline(mean.boot$t)
```

Fra Q-Q plotet ser vi at punktene grovt sett ligger på en rett linje. Dermed konkluderer vi at det er rimelig å anta at bootstrap utvalget fra punkt c) er normalfordelt.



Figur 4: Q-Q plot for 999 bootstrap utvalg fra Exercise 8.65 c).

f) Vi bruker bootstrap utvalgene fra punkt c) til å bestemme persentil intervallet; jf. side 414 i læreboka. Se R-koden.

```
> # f)
> # Regner ut persentil intervallet
> boot.ci(mean.boot, conf=0.95, type=c("perc"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = mean.boot, conf = 0.95, type = c("perc"))

Intervals :
Level      Percentile
95%      (26.63, 32.85 )
Calculations and Intervals on Original Scale
>
```

g) Vi oppsummerer de konfidensintervallene vi har funnet:

- Fra punkt a) har vi at et 95% konfidensintervall basert på t-fordelingen er: $[26.61, 32.94]$
- Fra punkt d) har vi at et tilnærmet 95% konfidensintervall basert på normalfordeling og bootstrap estimat for standardfeilen er: $[26.77, 32.78]$
- Fra punkt f) har vi at bootstrap persentil intervallet er: $[26.63, 32.85]$

Vi ser at de tre konfidensintervallene er ganske like.