BMI Projekt Statistik

Danmark Tekniske Universitet

02323 Introduktion til statistik

21.oktober.2020

Thomas Hohnen, s195455

Beskrivende Analyse

Opgave a

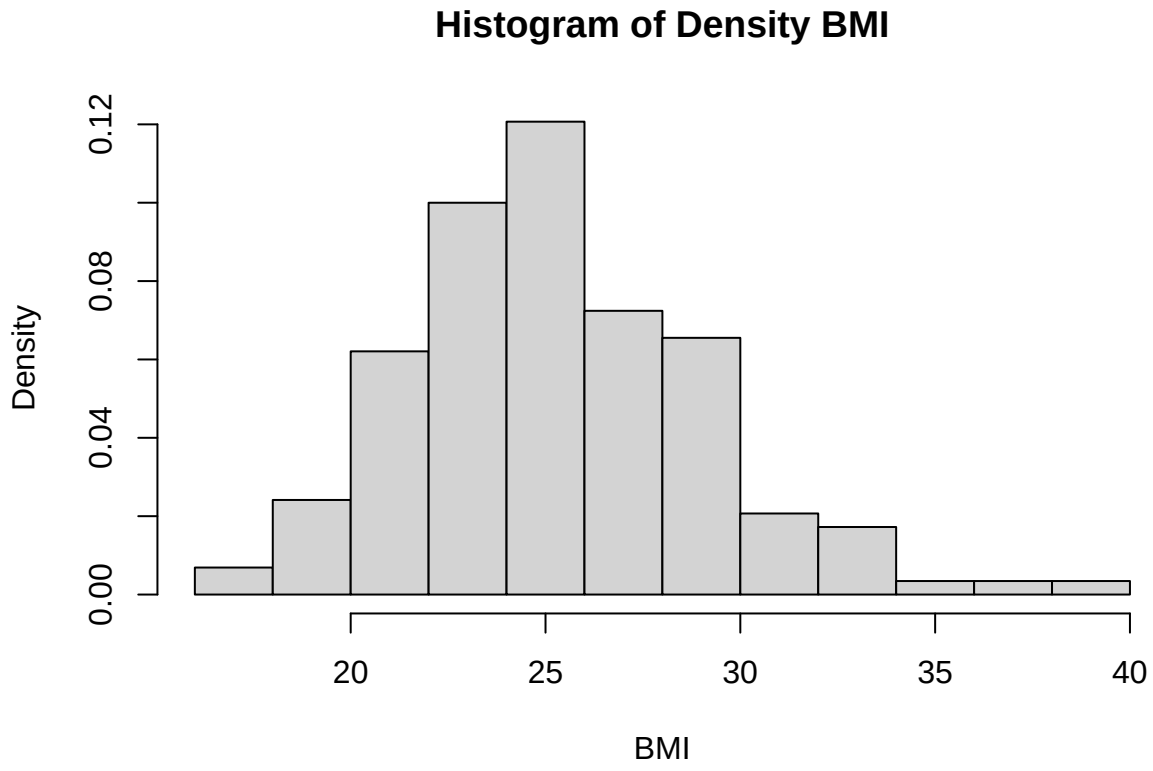
Lav en kort beskrivelse af datamaterialet. Hvilke variable indgår i datasættet? Er der tale om kvantitative eller kategoriseret variable. Er der nogle manglende værdier i variablene? Hvor mange observationer indgår der?

c(180, 185, 180, 168, 173, 161), c(80, 98, 80, 60, 83, 78), c(1, 1, 1, 0, 1, 0), c(5, 1, 5, 4, 5, 3), c(24, 6, 6, 24, 24, 6)

Dette datasæt er sat om som en form for tabel. Der er 5 forskellige variabler som også repræsenterer kolonerne. Disse variabler er "height, weight, gender, urbanity, fastfood". I alt er der blevet lavet 725 forskellige observationer som er blevet delt ud for de forskellige tabeller. Der er i alt 3 af variablerne som er kategoriseret disse er gender, urbanity og fastfood. Både højde og vægt er begge kvantitative.

Opgave b

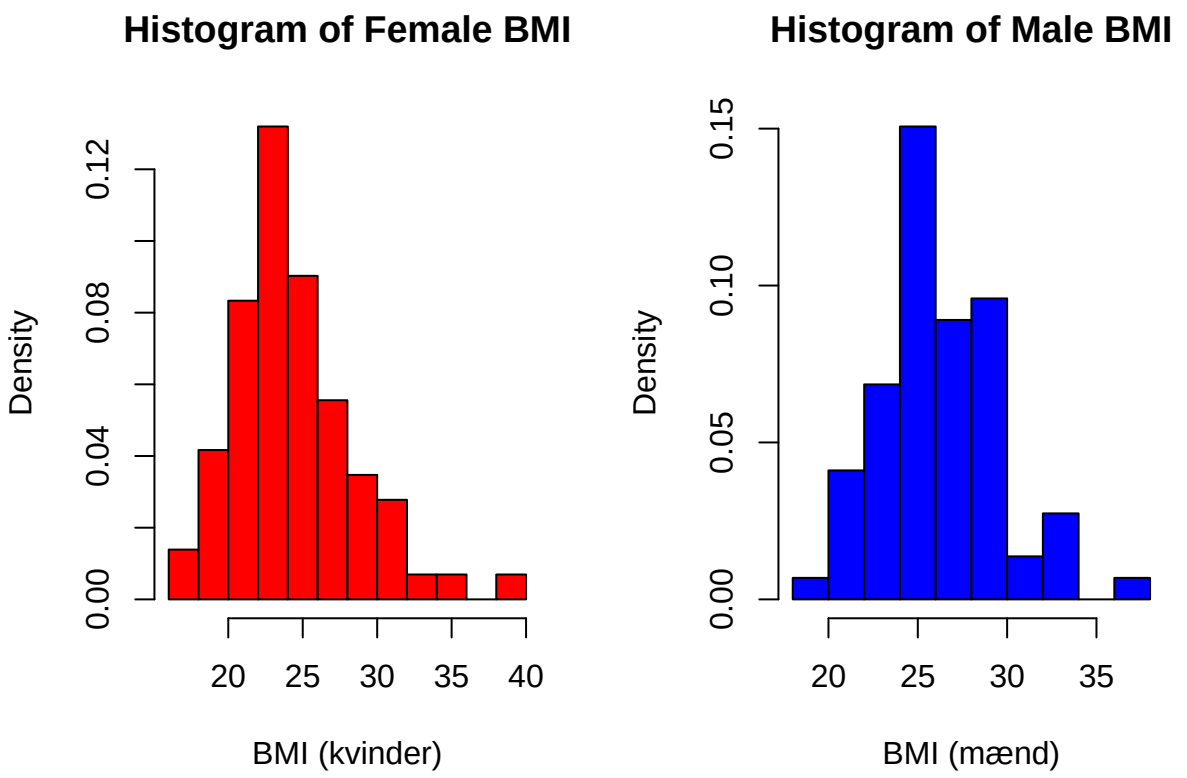
Lav et density histogram for BMI. Beskriv fordelingen af BMI-værdierne i stikprøven ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan BMI være negativ? Er der stor spredning i observationerne?



I det ovenstående histogram kan man se fordelingen af de forskellige BMI-værdier. Dette er et empirisk tæthedsdiagram som betyder at hvis alle arealerne af søjlerne bliver lagt sammen vil det give 1. Denne mængde af data er en højreskæv fordeling, siden der går en lille hale ud mod højre side af histogrammet ud fra midten af histogrammet hvor den største del af data'et er gupperet.For at den skulle være symetrisk skulle alt data være centreret omkring vores middelværdi. Når man kigge på dette histogram kender vi ikke den præcise middelværdi men det kan ses at det ligger omkring 20-25 da det er der vores data er grupperet.

Opgave c

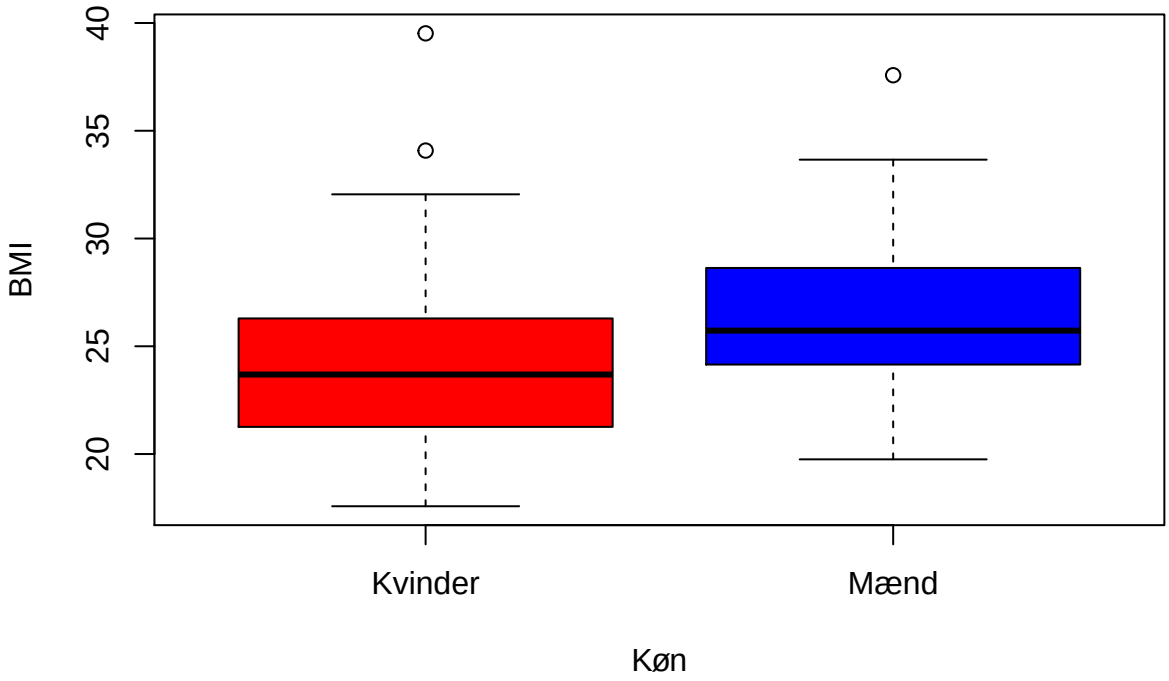
Lav et density histogram af BMI for hhv. kvinder og mænd. Beskriv de empiriske fordelinger af BMI for mænd og kvinder ud fra disse histogrammer, som i det forrige spørgsmål. Ser der ud til at være forskel i mænd og kvinders BMI?



Hvis man først kigger på fordelingen over BMI for kvinder, kan man se at den ligner utroligt meget den samlede fordeling. Den er en smule forskudt men ellers er der stadig en smule højreskev. Dette er fordi størstedelen af observationerne ligger i midten og fordeler sig ud på begge sider, men der kommer stadig en lille hale mod højre. Når man kigger på fordelingen over BMI for mænd, er det en smule anderledes. Her er der dog stadig tale om en højeskæv fordeling. Dette er fordi observationerne ikke fordeler sig lige på hver side af midten, men der igen kommer en lille hale mod højre side. Desuden kan det også ses at der er en forskel i mænd og kvinders BMI. Dette kan ses ved at BMI for kvinder har sit højdepunkt lige under 25 og begynder derfor at dykke. Hvorimod for mænd har den højdepunktet på 25 og begynder på at falde lidt men stiger igen med det samme. Dette betyder at kvinder generelt har en bedre BMI end mænd har.

Opgave d

Lav et boksplot af BMI opdelt efter køn. Benyt derefter plottet for at beskrive den observerede fordeling af BMI for mænd og kvinder. Er fordelingerne symmetriske eller skæve? Ser det umiddelbart ud til at der er forskel mellem fordelingerne? Er der nogle ekstreme observationer?



I disse to plot kan man bedre se at BMI for mænd er generelt højere end at den er for kvinder som der også blev beskrevet i tidligere opgave. Hvis man nu kigger på plottet for kvinder kan man argumentere for at det denne gang er symmetirsk hvis man forkaster de ekstreme observationer. Dette kan man se på at middelværdien ligger tæt på midten og der er ca. lige meget data rundt om. Samtidigt kan man også sige at data'en for mænd ikke har ændret sig siden den stadig er højreskæv. På disse 2 plots kan det også ses at der er nogle ekstreme observationer som ligger imellem 35-40 for både mænd og kvinder hvilket kunne have en effekt på det endelige Box plot.

Opgave e

Udfyld tabellen med de opsummerende størrelser for BMI for hele stikprøven og derefter separat for mænd og kvinder. Beskriv hvilken ekstra information der kan udledes fra tabellen sammenlignet med boxplottet?

Variabel	Antal obs.	Stikprøve-gennemsnit	Stikprøve-varians	stikprøve-standard-afvigelse	Nedre kvartil	Median	Øvre kvartil
	n	(x)	(s^2)	(s)	(Q1)	(Q2)	(Q3)
Alle	145	25,25	14,69	3,83	22,59	24,69	39,52
Kvinder	72	24,22	16,42	4,05	21,26	23,69	26,29
Mænd	73	26,27	11,07	3,33	24,15	25,73	28,63

Der er nogle forskellige informationer som denne tabel giver i stedet for at kigge på vores boxplot. Som udgangspunkt gælder vores boxsplots informationer kun for de følgende kvartiller: 0, 25, 50, 75 og 100. Boxplottet er rigtig god til at se hvordan vores data er spredt ud over hele datasættet omkring vores middelævrdi. Dog siger vores boxsplot ikke noget omkring vores varians, spredning, gennemsnit og den

samlede antal af obsevertioner. Det eneste som vi kan se når det kommer til antal observationer er at der ca. er taget lige mange observationer for hhv mænd og kvinder.

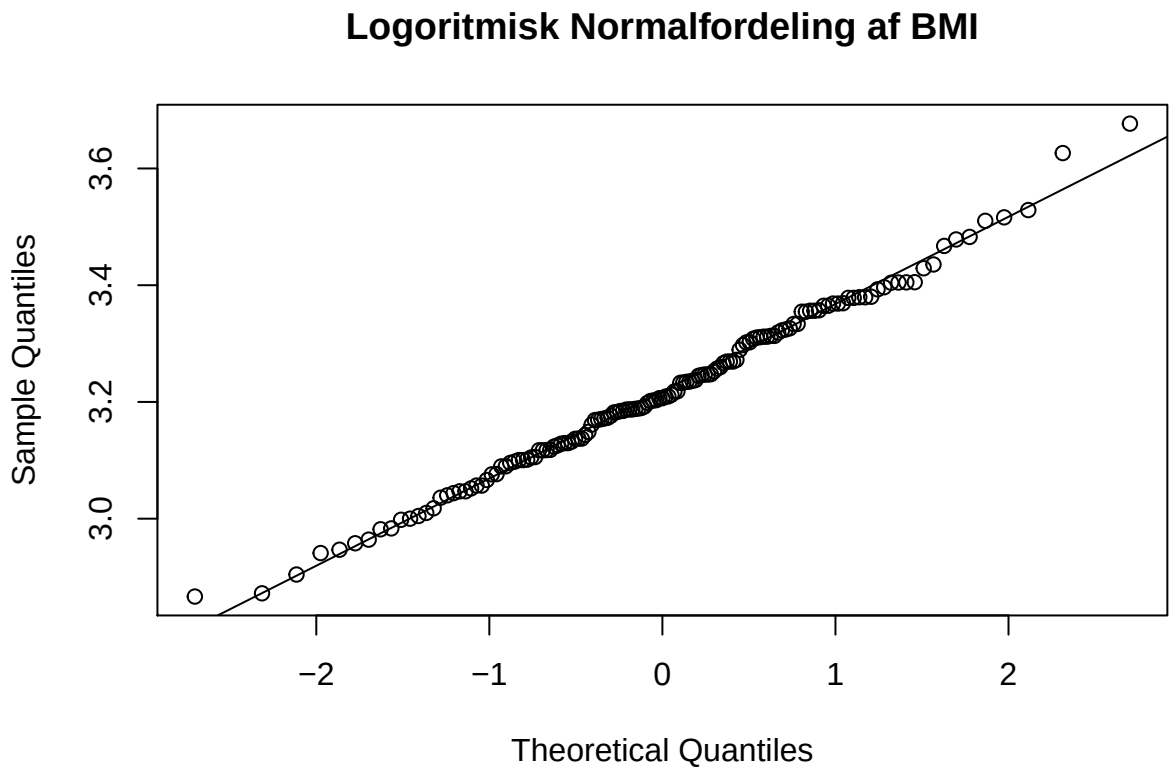
Statistisk Analyse

Opgave f

Opskriv en statistisk model for logoritmen af til BMI for hele befolkningen, hvor der ikke skelner mellem kvinder og mænd. Estimer modellens parameter (middelværdi og standardafvigelse). Fortag modelkontrol af de antagede forudsætninger. Idet, konfidensintervaller og hypotesetest her involverer fordelingen ef gennemsnit, kan det være nyttigt også at inddrage den centrale grænseværdisætning i argumentationen.

Først skrives den statistiske model op uden de rigtige tal. Dette gør at vi kan se hvordan modellen endeligt vil komme til at se ud.

inline equation: $X \sim LN(\alpha, \beta^2)$ *and i.i.d where* $i = 1, ..., n$

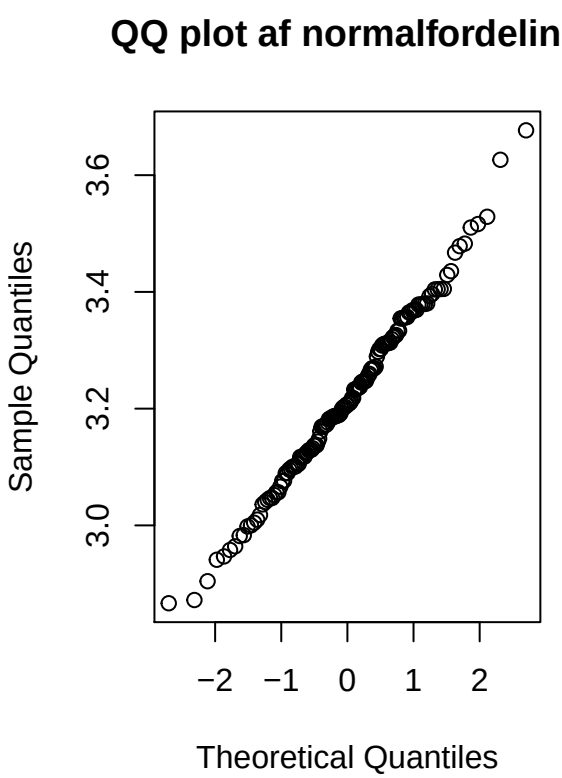


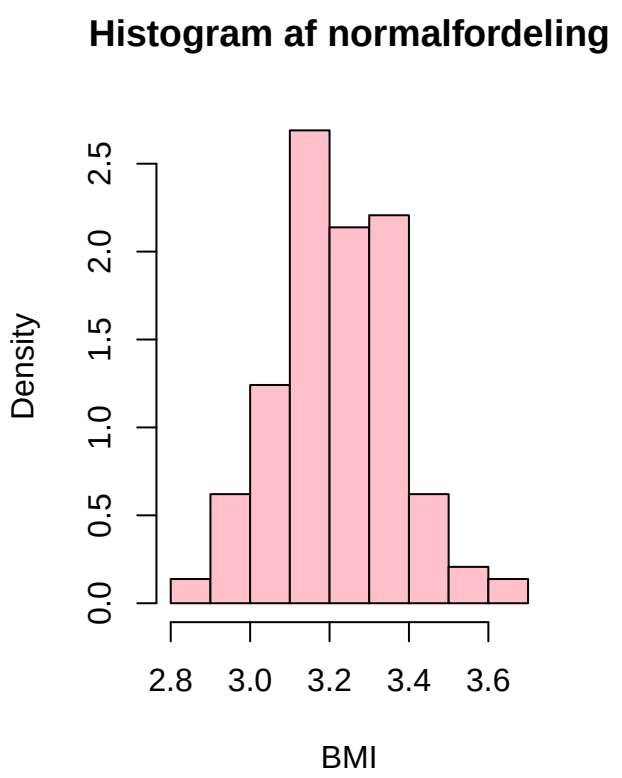
Den ovenstående graf hviser den logaritmiske normalfordeling af BMI værdierne. Ud fra dette kan vi se at den data der er blevet taget passer meget bedre nu til en normalfordeling. Dette betyder også er fordelingen til disse observationer vil nu være symetrisk og ikke længere højreskæv.

For at regne varians og middelværdi ud er der blevet tage brug af kommandoerne "var" og "mean" i r. Herefter er disse tal blevet sat ind i den statistiske model.

$$\mu log(bmi) = 3.218, \sigma log(bmi) = 0.022$$

$X_i \sim LN($ 3.2176405,0.0221646) *and i.i.d where* $i = 1, ..., n$





Her kan det ses at QQ plottet følger en normalfordeling hvilket betydet at ud fra dette plot er den data normalfordelt. Derudover kan det også ses ved histogrammet et dette ikke længere er højreskæv men derimod nu er symmetrisk og har den normalle klokkeform. Ud fra dette kan der konkluderes at den mængde data der arbejdes med er normalfordelt.

Opgave g

Angiv formlen for et 95% konfidensinterval for middelværiden af logoritmen til BMI for hele befolkningen. Indsæt tal og bereng intervallet. Angiv derefter et 95% konfidensinterval for medianen af BMI for hele befolkningen.

Det første der skal gøres er at skrive den ønskede formål op og derefter få den udfyldt med de forskellige tal ved at bruge R. Den nedenstående formel bruges til at finde konfidensintervallet på 95%.

$$\bar{x} \pm t_{0.975} \cdot \frac{\sigma}{\sqrt{n}}$$

Dette gør at vi kan beregne hvad disse 2 tal er ved at bruge R hvilket giver det følgende resultat:

Middelværiden blev tidligere regnet ud til: 3.2176405

Dette gør at vores t-fordelings 97,5% kvartil til: 1.9764596

$$3.28 \pm 1.977 \cdot \frac{0.149}{\sqrt{145}}$$

Her er det vigtigt at udpege at vi regner med de logoritmiske værdier. Dette er ikke hvad der skal arbejdes med. Grundet dette bliver de endelige værdier oplytet i e som følgende.

$e^{3.218} = 24.97$

Resultatet uden at oplyte det i e:

$$[3.24, 3.19]$$

Resultatet efter at den er blevet oplytet i e:

$[e^{3.242}, e^{3.242}] = [25.59, 24.37]$

Opgave h

Udfør en hypotese test med henblik på at undersøge, om middelværdien af logoritmen til BMI er forsekllige fra log(25). Dette kan gøres ved at teste følgende hypotese:

$$H_0 : \mu logBMI = Log(25)$$
$$H_1 : \mu logBMI \neq Log(25)$$

For at løse denne opgave skal der testes om vores nul hypotese passer. Dette bliver gjort ved først at finde vores t-test størelse og derefter kan p-værdien blive regnet ud. Hvis vores p-værdi ender med at være over 5% acceptere vi vores null hypotese, altå er vores signifikansniveau på 0,05. Hvis ikke det er sagen bliver den forkaster frem for den anden hypotese.

Først bliver de formler skrevet op som der skal bruges til at teste vores hypotese:

$$t_{obs} : \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$\text{p-værdi} : 2 \cdot P(T > |t_{obs}|)$$

Herefter bliver de vores data sat ind i formlerne for at kunne regne begynde at påvise vores null hypotese. Først bliver teststørelsen fundet. Først bliver der fundet teststørelsen:

$$t\text{obs} = \frac{3.22 - log(25)}{0.15/\sqrt{145}} = -0.09$$

Herefter kan der nu blive regnet på p-værdien:

$$\text{p-værdi} : 2 \cdot P(T > |t_{obs}|) = 2 \cdot (1 - F(|t_{obs}|)) = 0.92$$

Her skal der efterfølgende blive sat de rigtige tal ind. Vores t værdi er som altid $n - 1$ hvilket gør at vi herefter kan udregne vores tal gennem R.

Dette betyder at vi ikke forkaster vores nulhupotese siden vores p værdi er over vores signifikansniveau. Dog skal det også siges at p-værdien er meget høj i forhold til hvad man normalt ender med og der er meget lidt bevis mod H0. Dette betyder at der heller ikke kan konkluderes at overhalvdelen er overvægtig siden man først er overvægtig hvis man har en BMI på over 25. Det eneste som der er kokluderet her er at middelværdien for logBMI er ikke er forskellige fra log(25) og ikke om det var en overvægt af folk med BMI på 25.

Opgave i

Angiv statistiske modeller for logoritmen til BMI for hendholdsvis kvinder og mænd. Fortag modelkontrol af de antagede forudsætninger i de to modeller. Estimer modellens parametre.

For at løse denne opgave skal der først blive skrevet de 2 statisktiske modeller op for hhv mænd og kvinder. Herefter vil der blive lavet 2 forsekellige qqplots for både logBMI af kvinder og mænd men også den normale BMI af kvinder og mænd. For at få udfyldt disse statiskemodeller med de relevente tal bliver der gjort brug af R for at vinde hhv middelværdig og varians for kvinder og mænd. Disse tal bliver sat ind på α og β pladserne.

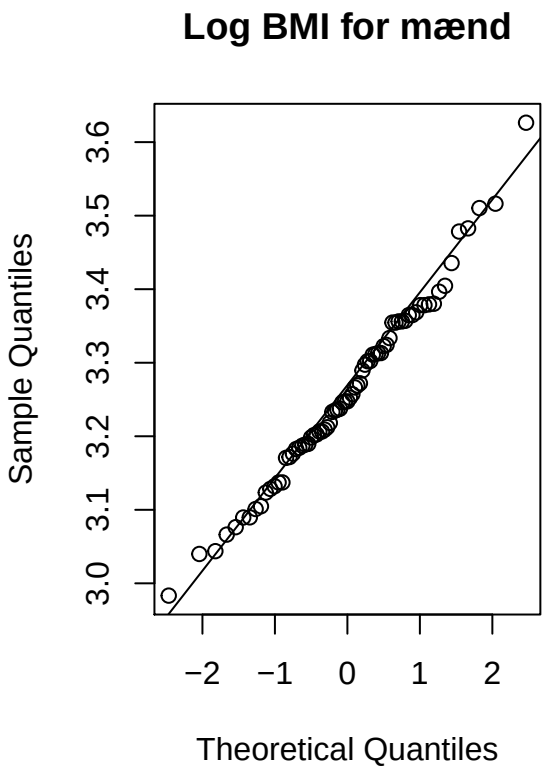
Først for den statistiske model for kvinder:

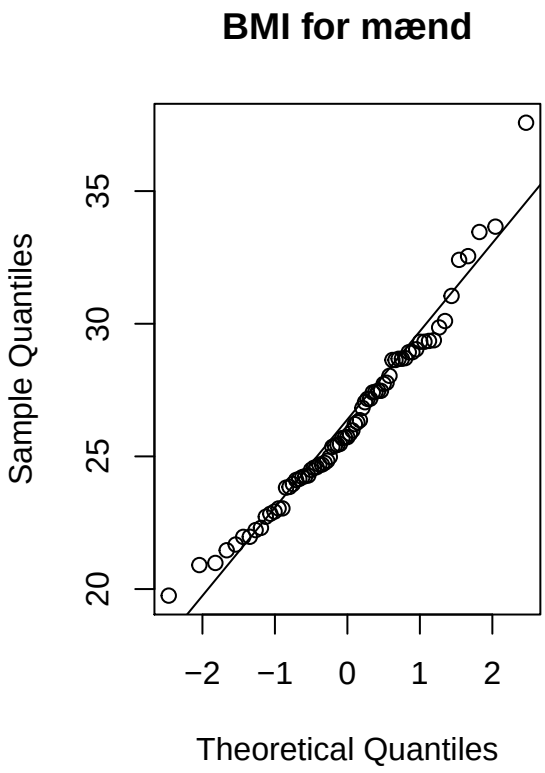
$$X_i \sim LN(3.174, 0.0255) \text{ and i.i.d where } i = 1, ..., n$$

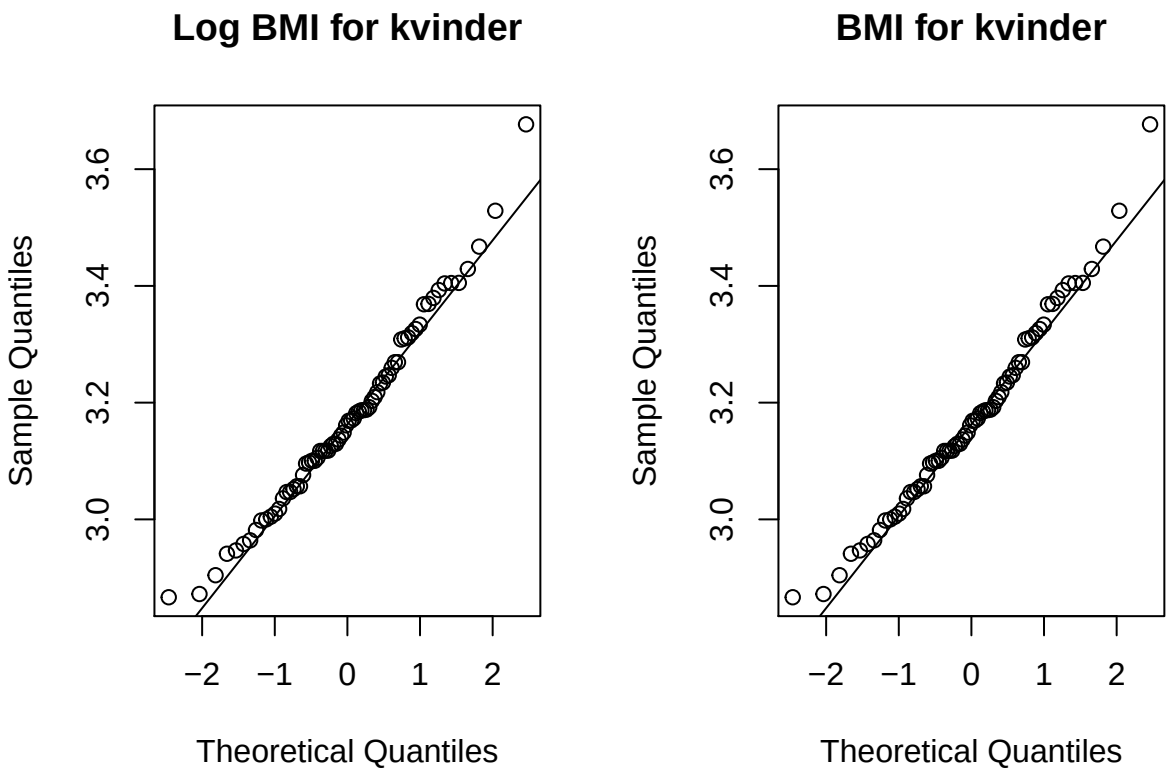
Derefter for mænd:

$$X_i \sim LN(3.260, 0.0153) \text{ and i.i.d where } i = 1, ..., n$$

Herefter er der blevet lavet 4 forskellige qq plot. Dette vil hjælpe os med at se om vores statistiske model passer. Der er først blevet lavet for mænd og derefter kvinder.







Her kan vi se at normalfordelingen får både mænd og kvinder er blevet mere symmetirks end en den logoritmiske fordeling for hhv. mænd og kvinder. Dette kan man også se hvad at kigge på de ekstreme variabler som ikke længere er helt så ekstreme som de var før. Dette kan man tydeligt se hvad at kigge på den linje der følger hvert plots som repræsentere den "perfekte" normalfordeling. Der ligger også nogle histogrammer i Bilag hvor man kan se at den at den er blevet mere klokke formet end det var før.

Opgave j

Beregn 95% konfidensintervaller for middelværdien af logoritmen til BMI for hhv mænd og kvinder. Benyt disse til at bestemme 95% konfidensintervaller for median af BMI for hhv mænd og kvinder. Udfyld tabellen.

Middelværdien blev regnet ud i tidligere opgave for både mænd og kvinder. Først bliver der kigget på mænds konfidensinterval:

$$3.26 \pm 1.99 \cdot \frac{0.12}{\sqrt{73}}$$

Herefter oplyfter vi hele resultatet i e hvilket giver:

$$[25.32; 26.82]$$

Derefter for kvinder:

$$3.17 \pm 1.99 \cdot \frac{0.16}{\sqrt{72}}$$

Hvor man igen oplyfter hele resultatet i e hvilket giver:

$$[23.02; 24.82]$$

Herefter kan de forskellige konfidensintervalls værdier blive sat ind i en tabel.

	Nedre grænse af KI	Øvre grønse af KI
Kvinder	23.02	24.82
Mænd	25.32	26.83

Opgave k

Undersøg ved en hypotesetest, om der kan påvises en forskel på mænd og kvinders BMI. Opskriv hypotesen og angiv signifikansniveauet, formlen for teststørelsen, samt teststørelsens fordeling. Indsæt tal, og bereng teststørelsen og p-værdien

Det første der skal gøres er at opstille vores hypotese test. Hvor vi også beslutter at vores signifikansniveau er på 5%.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu 1 - \mu 2 \neq 0$$

Disse 2 hypoteser er opsat sådan pga. de 2 grupper ikke har noget af gøre med hindanen. Altså hvis de 2 middelværdier trukket fra hindanden er lige med nul er der en sammenhæng hvis ikke er der ingen. Det første der skal gøres er at sætte de forskellige formler op der skal bruges.

Her er den første formel der skal bruges:

$$t\text{obs } = \frac{(\bar{x}1 - \bar{x}2) - \delta 0}{\sqrt{s1^2/n1 + s2^2/n2}}$$

Dette giver resultatet:

$$t\text{obs } = 3.637$$

Herefter den anden formel som skal bruges til at beregne vores antal af frihedsgrader:

$$v = \frac{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^2}{\frac{(s1^2/n1)^2}{n1-1} + \frac{(s2^2/n2)^2}{n2-1}}$$

Dette giver resultatet:

$$v = 133.75$$

Når disse tal så er blevet regnet ud. Bliver det sat ind i den følgende formel:

$$p \text{ -vardi } = 2 \cdot P\left(T > |t_{\text{obs }}|\right) = 2 \cdot \left(1 - F\left(|t_{\text{obs }}|\right)\right) = 3.92 \times 10^{-4}$$

$$\text{p-værdi} : 2 \cdot P(T > |3.637|) = 0.00039$$

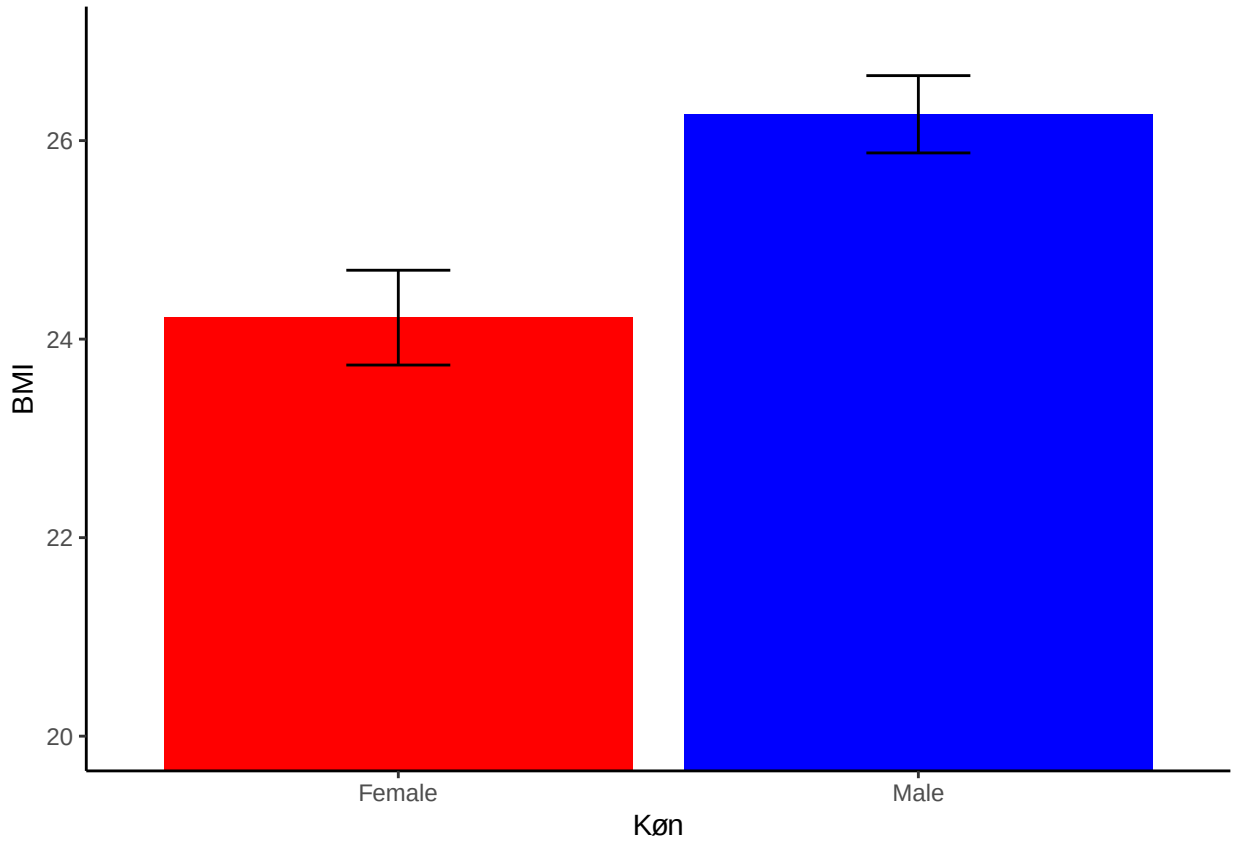
Dette betyder at vores nul hypotese ikke kan bekraftes og derfor bliver den forkastet siden vores p-værdi er under vores signifikansniveau. Desuden definere v antallet af vores frihedsgrader og bliver derfor brugt til at regne p-værdien ud i R.

Opgave l

Kommenter om det er nødvendigt at udfører en hypotesetest i det forrige spørgsmål, eller den samme konklusion kunne opnås ud fra konfidensintervallerne alene?

Ja det ville man godt kunne sige ud fra konfidensintervallerne alene. Dette er fordi hvis man kigger på det laveste inteveste interval for mænd er det stadig højere end det højeste interval for kvinder. Dette betyder at man kan konkludere at der ikke er nogen overlapning mellem mænd og kvinders BMI.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Korrelation

Opgave m

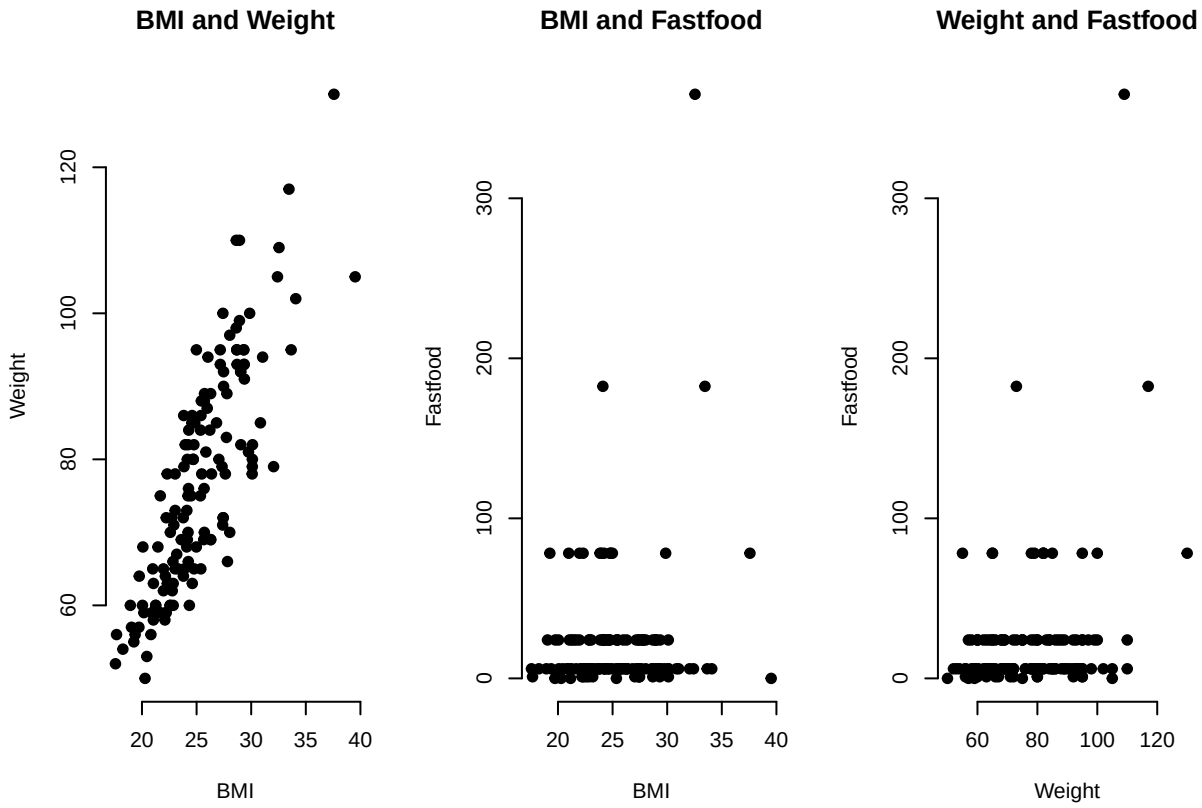
Angiv formlen til beregning af korrelation mellem BMI og vægt. indsæt tal og bereng korrelation. Beregen desuden de resterende parvise korrelationer, der indvoldvere BMI, vægt og fastfood. Lav scatterplots, der illustrer de parvise sammenhænge mellem disse variable. Vurder om sammenhængden mellem plots og korrelation er forsvundet.

Der skal bruges nogle forskellige formler til at renge det ud. Den første er kovariansen:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

Den anden er korrelationskeoficient:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$$



Som man kan se på de 3 plot der er blevet lavet er der kun en korrelation mellem Vægt og BMI dette kan man også se hvis man regner tallene ud med de 2 ovenstående formler. Hvis man gør det får man følgende:

Korrelationskeoficient mellem BMI og Vægt:

$$0.828261$$

Korrelationskeoficient mellem BMI og Fastfood:

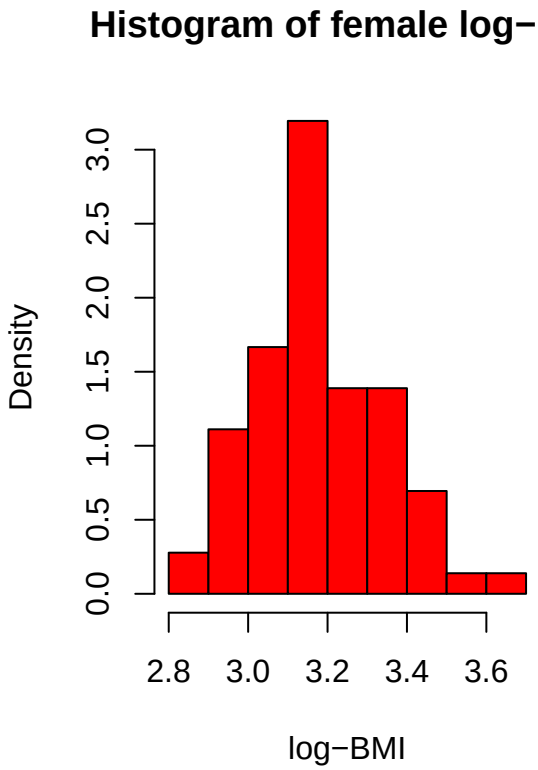
$$0.1531578$$

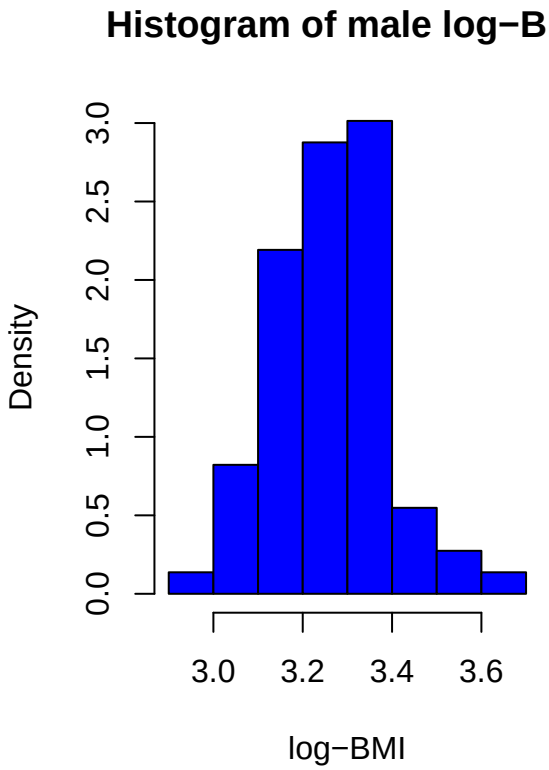
Korrelationskeoficient mellem Vægt og Fastfood:

$$0.2793223$$

Dette gør så at vi kan se at det er en sammenhæng mellem BMI og vægt fordi den er tæt på 1 hvorimod de andre er tættere på 0.

Bilag


Histogram of female log−BMI
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2.8
3.0
3.2
3.4
3.6
log−BMI


Histogram of male log−BMI
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.0
3.2
3.4
3.6
log−BMI