BMI Projekt Statistik


fig.algin = 'center'

Danmark Tekniske Universitet

02323 Introduktion til statistik

12.oktober.2020

Thomas Hohnen, s195455

Beskrivende Analyse

Opgave a

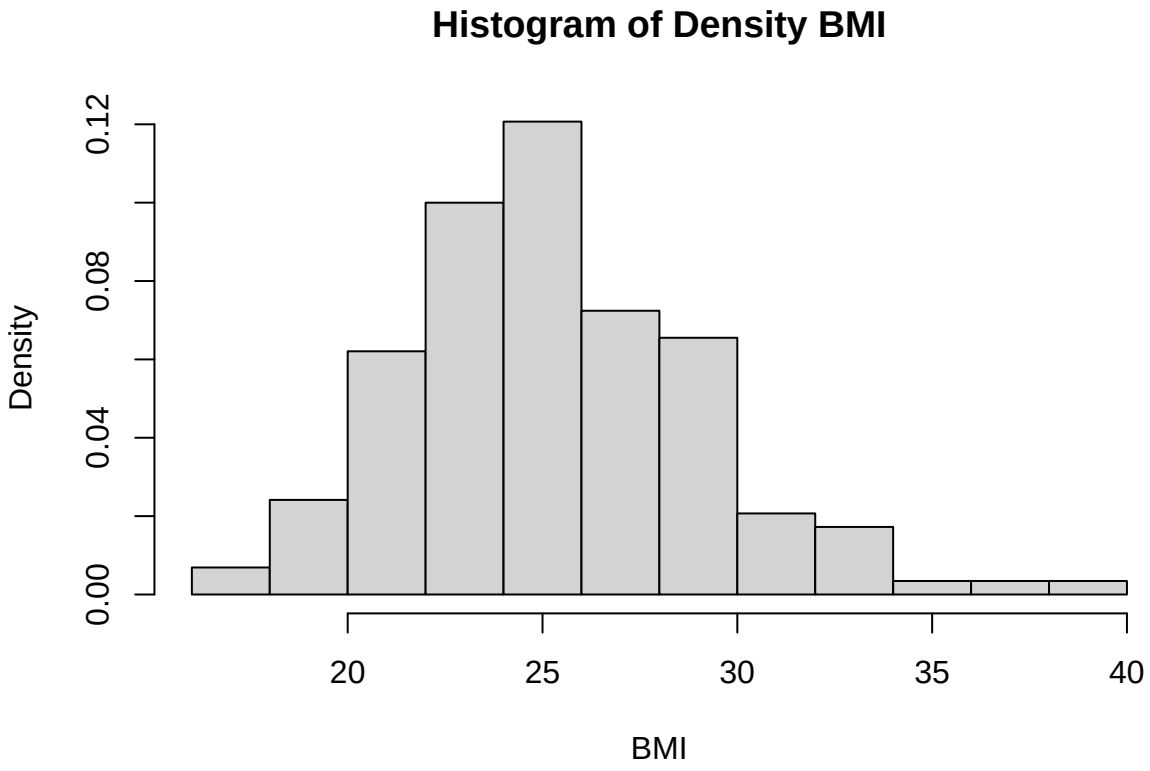
Lav en kort beskrivelse af datamaterialet. Hvilke variable indgår i datasættet? Er der tale om kvantitative eller kategoriseret variable. Er der nogle manglende værdier i variablene? Hvor mange observationer indgår der?

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

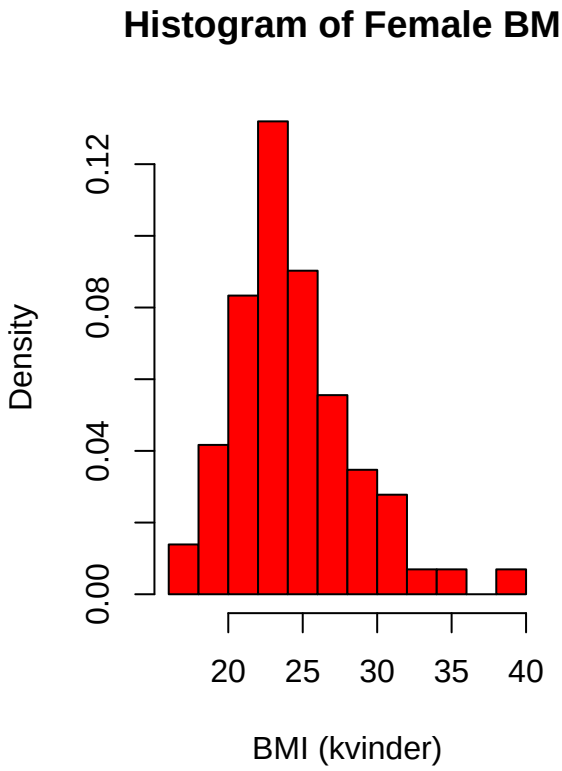
Lav et density histogram for BMI. Beskriv fordelingen af BMI-værdierne i stikprøven ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan BMI være negativ? Er der stor spredning i observationerne?

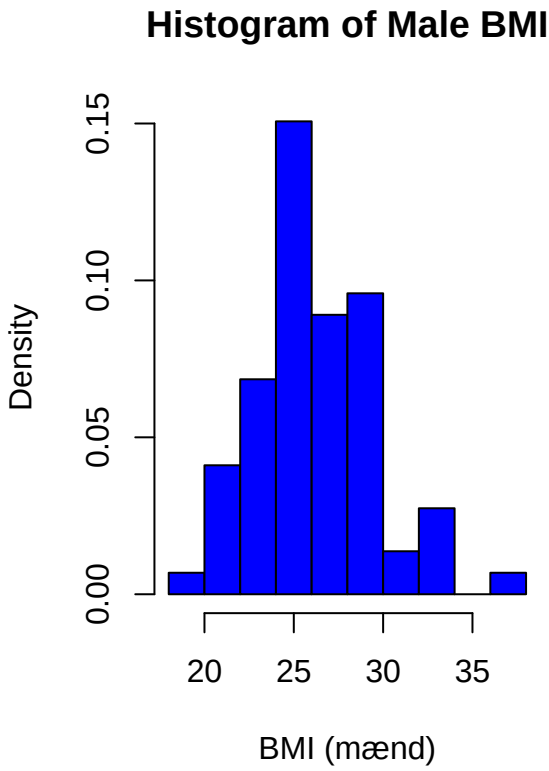


I det ovenstående histogram kan man se fordelingen af de forskellige BMI-værdier. Dette er et empirisk tæthedsdiagram som betyder at hvis alle arealerne af søjlerne bliver lagt sammen vil det give 1. Denne mængde af data er en højreskæv fordeling siden der går en lille hale ud mod højre side af histogrammet. Dette kan ses siden størstedelen af den grupperet data er omkring 25 og efter det kommer der en lille hale til højre.

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?

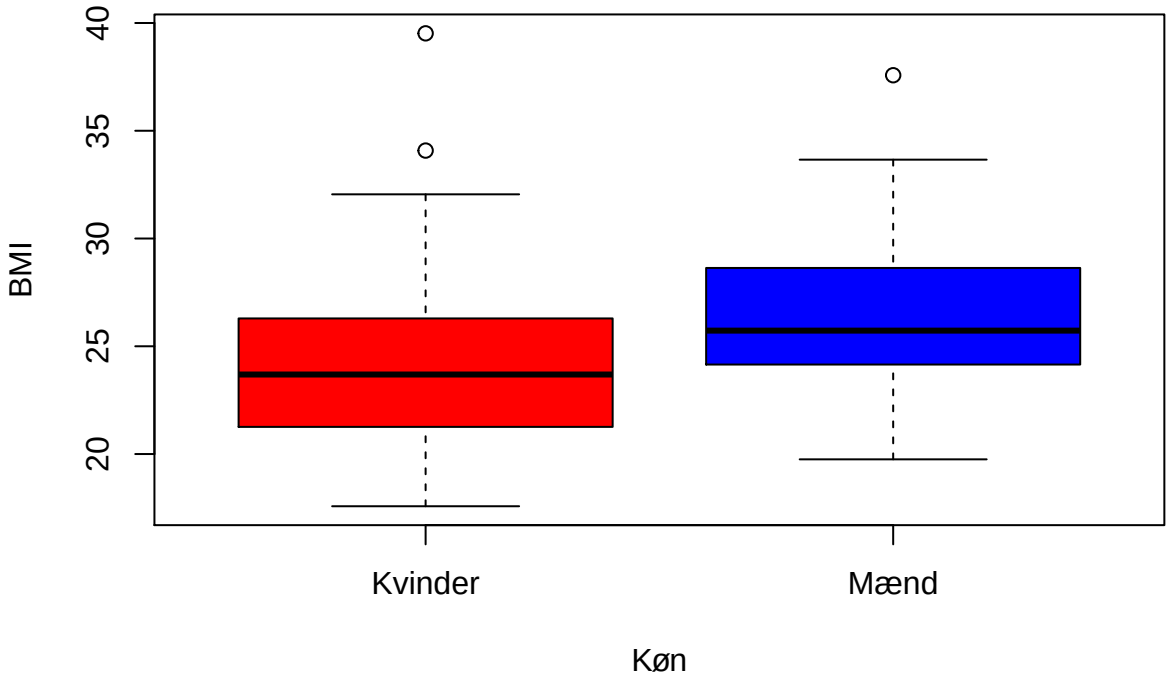




Hvis man først kigger på fordelingen over BMI for kvinder, kan man se at den ligner utroligt meget den samlede fordeling. Den er en smule forskudt men ellers er der stadig en symmetrisk fordeling. Dette er fordi størstedelen af observationerne ligger i midten og fordeler sig ud på begge sider. Når man kigger på fordelingen over BMI for mænd, er det en smule anderledes. Her er der ikke længere en symmetrisk fordeling men en skæv fordeling. Dette er fordi observationerne ikke fordeler sig lige på hver side af midten. 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 hvilket giver den symmetriske fordeling. 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 forkvinder som der også blev beskrevet i tidligere opgave. Man kan se på plottet for kvinder at det er en symmetrisk fordeling da medianen ligger ca. i midten af observationerne derudover kan man også se at fordelingen for mænd er skæv da medianen ikke ligger i midten. 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

Statistisk Analyse

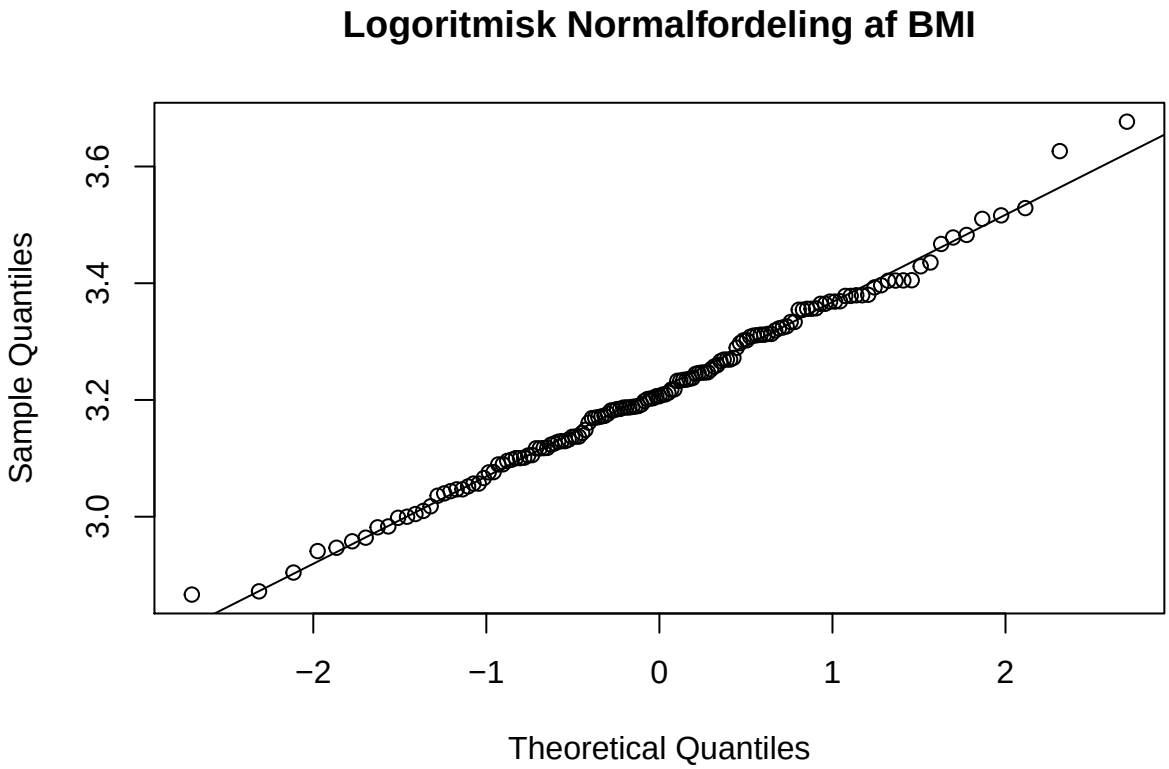
Opgave f

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.

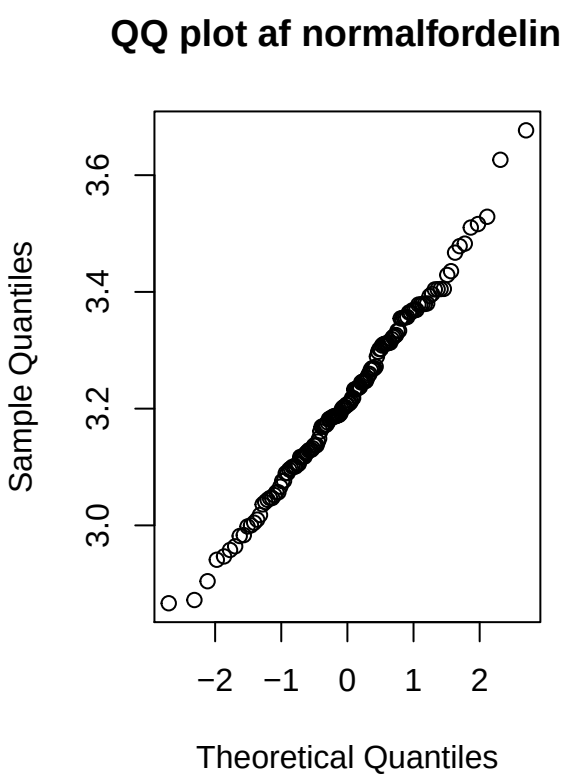
inline equation: $X \sim LN(\alpha, \beta^2)$ *and i.i.d where* $i = 1, ..., n$

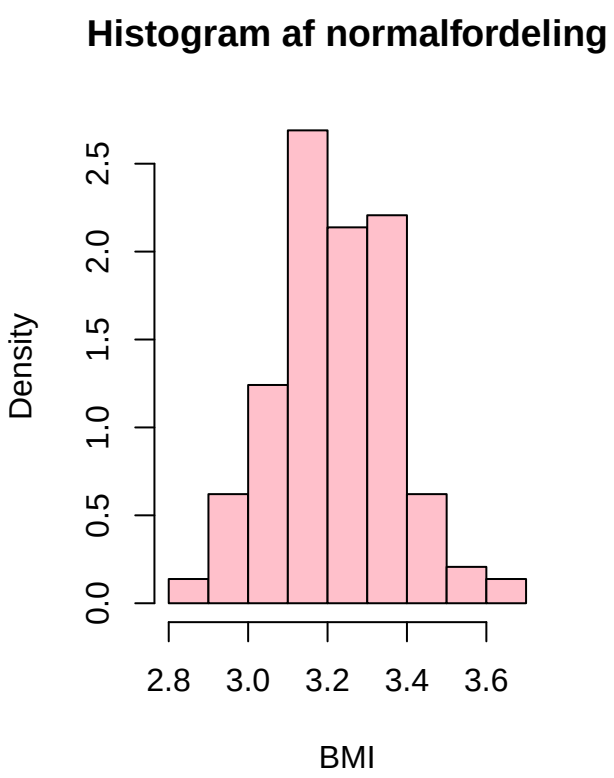


Den ovenstående graf hviser den logaritmiske normalfordeling af BMI værdierne. Ud fra dette kan vi se at den data der er blevet taget passer meget bedre nu til en normalfordeling. Dette betyder også er fordelingen til deisse observationer vil nu være symetrisk og ikke længere højreskæv.

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.

$X_i \sim LN($ 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.

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

Middelværiden blev tidligere regnet ud til: 3.2176405

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

3.2176405 ± 1.9764596