Code

02402 - Statistics Projekt 1: Bmi undersøgelse

Daniel F. Hauge - s2011687

Code

Dette projekt består af en beskrivende analyse og en statistisk analyse af et datasæt med mennesker. Projektet forsøger at kaste lys på BMI med statistik. Projektet henvender sig til læserer som er familiær med projekt beskrivelsen fra statistik kurset 02402 fra DTU og befærder sig komfortabelt i statistiske begreber og metoder.

Projektet er lavet som en R notebook, og indeholder derfor områder hvor R er brugt som redskab til udregning og plot optegninger. Bokse med skriften `Code` indikere at der er blevet kodet noget i R til at udregne, tegne eller gemme værdier til senere brug. R koden kan findes på følgende måder:

- Se den medfølgende ".R" fil, eller via på: https://github.com/DanielHauge/02402-statistics/blob/master/Project1/projekt_code_only.R (https://github.com/DanielHauge/02402-statistics/blob/master/Project1/projekt_code_only.R)
- Se den medfølgende ".rmd" fil, eller på: https://github.com/DanielHauge/02402-statistics/blob/master/Project1/project.rmd (https://github.com/DanielHauge/02402-statistics/blob/master/Project1/project.rmd)
- Projektet kan også findes i renderet form med kode afsnit på: https://htmlpreview.github.io/?https://github.com/DanielHauge/02402-statistics/blob/master/Project1/project.nb.html (https://htmlpreview.github.io/?https://github.com/DanielHauge/02402-statistics/blob/master/Project1/project.nb.html)

Beskrivende analyse

a) Data

Datamaterialet indeholder 145 observationer med 7 egenskaber på mennesker for projektets problemstilling. Variable på hver observation er som følger:

Variable Navn	Måle type	Måle enhed	Forklaring
gender	Kategoriserende	0 (Mand) / 1 (Kvinde)	Personens køn angivet som som et tal ved 0 eller 1.
height	Kvantitativt	cm	Personens højde i centimeter.
weight	Kvantitativt	kg	Personens vægt målt i kilogram
urbanity	Kategoriserende	tal i intervallet [1 - 5]	Befolkningstætheden i byen personen bor. 1: Udenfor bymæssig bebyggelse, 2: Under 10.000 indbyggere, 3: [10.00-49.999] Indbyggere, 4: [50.000-99.999] Indbyggere, 5: Over 100.000 indbyggere
fastfood	Kvantitativt	dage pr. år	Antal dage per år personen har spist fastfood.
bmi	Kvantitativt	Bmi	Dette er en enhed som prøver at beskrive kroppens størelses forhold, baseret på vægt og højde. Bmi står for “Body Mass Index”.
logbmi	Kvantitativt	Log(Bmi)	Dette er bmi’en der er blevet log transformeret.

Nedenfor ses første og sidste observation som et præsenterende eksempel:

Code

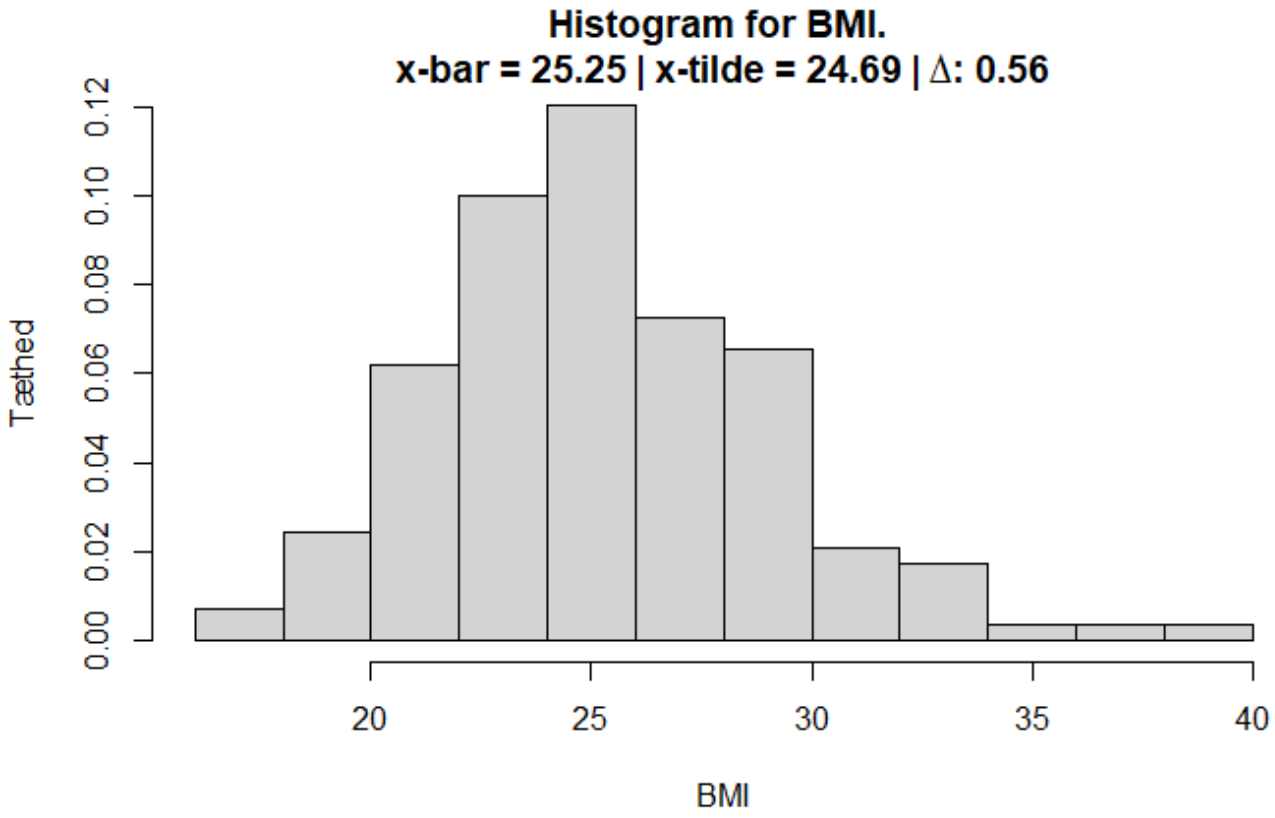
	height <int>	weight <int>	gender <int>	urbanity <int>	fastfood <dbl>	bmi <dbl>	logbmi <dbl>
1	180	80	1	5	24	24.69136	3.206453
145	163	105	0	3	0	39.51974	3.676800

2 rows

b) Bmi fordeling

Vi kan ved hjælp fra et histogram over tæthed for BMI nemmere beskrive hvordan stikprøvens BMI oberservationer fordeler sig. Nedenfor ses et histogram over Tætheden for BMI, hver søjle i histogrammet har et interval på 2. Y aksen eller højderne på søjlerne beskriver hvor stor en andel der ligger inden i det tilsvarende søjle interval. X aksen er BMI.

Code

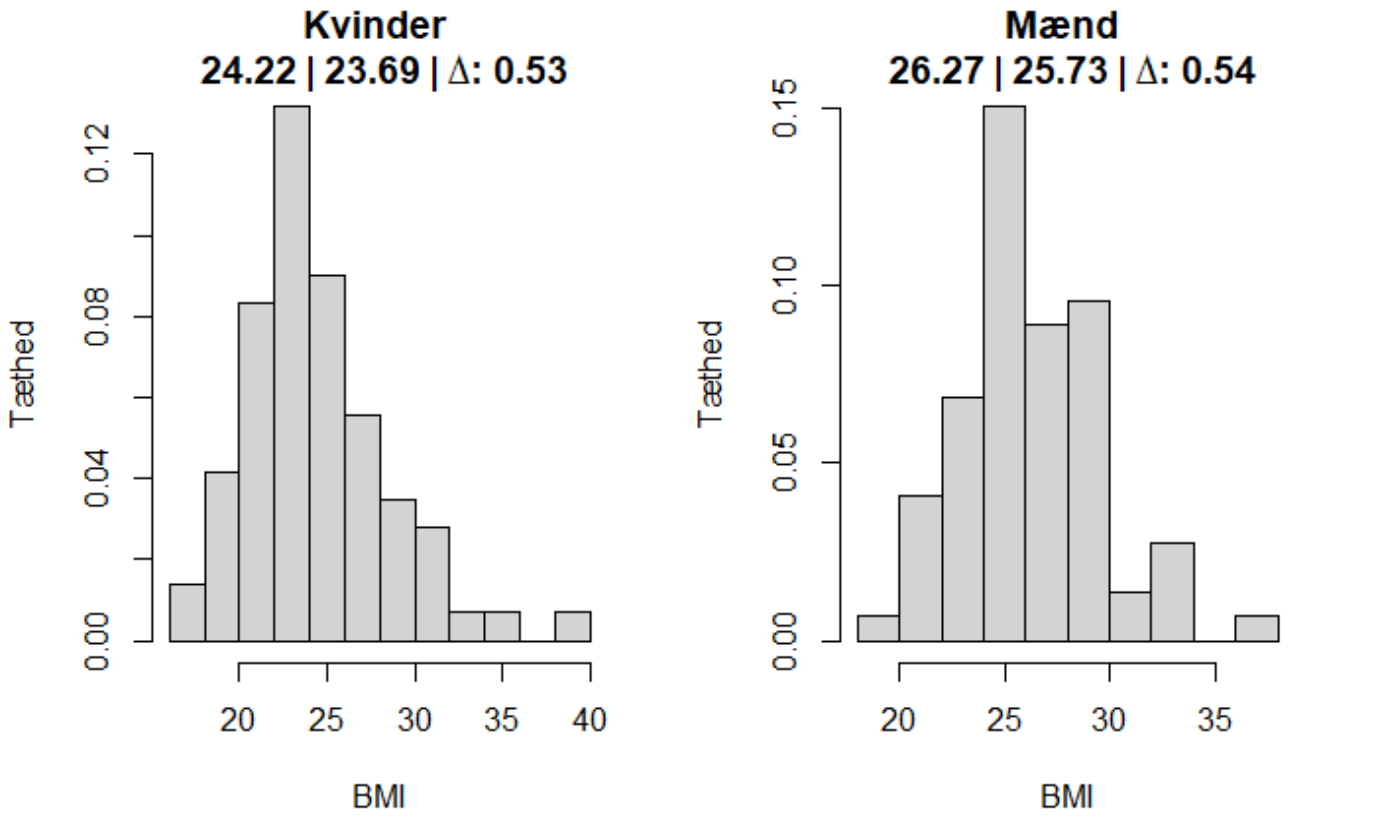


Histogrammet følger noget der måske godt kunne ligne en normal fordeling. Men histogrammet er ikke helt symmetrisk omkring den centrale data tendens, vi har altså en højreskæv fordeling af BMI. Det ses at stikprøven indeholder en længere hale på den overvægtige ende, derfor kan det også ses at BMI middelværdien i stikprøven er over medianen værdien.

c) Segmentering af data

Da der er naturlige forskellige mellem kvinder og mænd, bør vi også kigge på dataen opdelt for kvinder og mænd. Nedenfor ses histogrammer i samme stil som tidligere, dog opdelt i mænd og kvinder.

Code

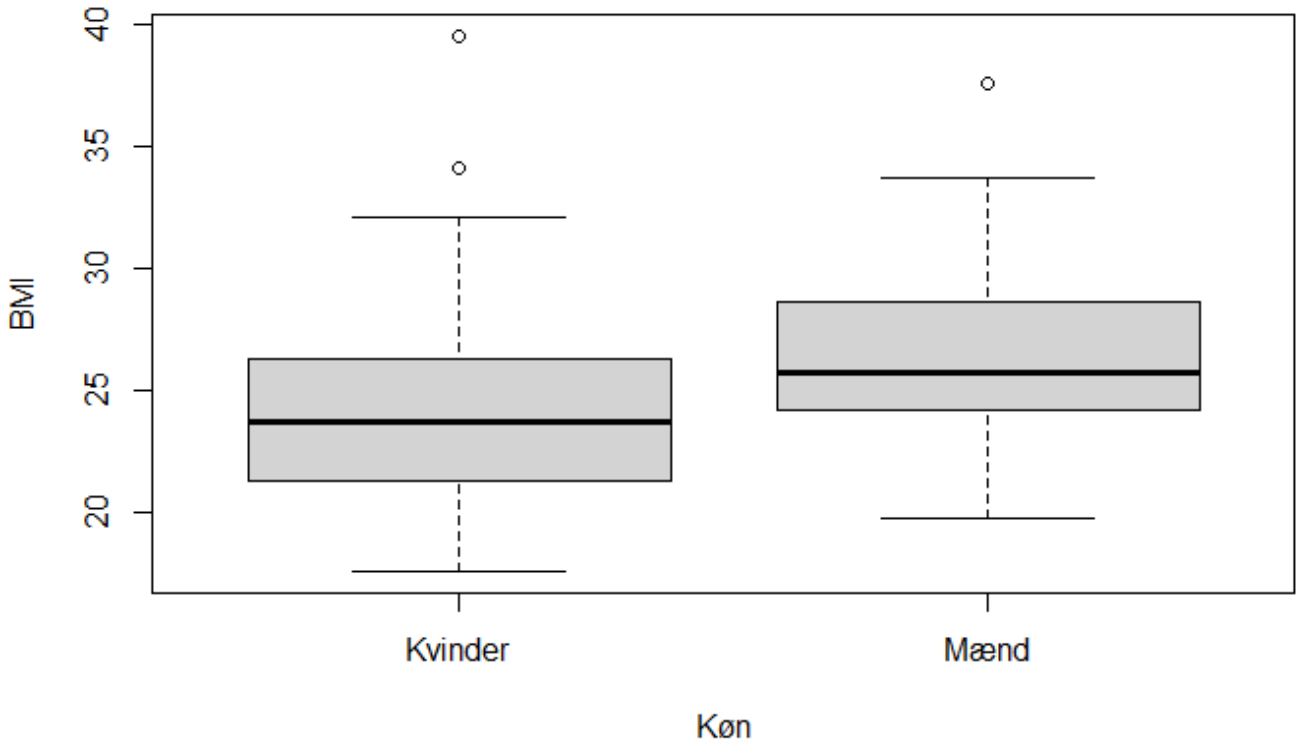


Det samme mønster opstår, med en højre skæv fordeling. Dog lader det til at stikprøven for mænd er fordelt en lille smule mere skæv set ved en højere difference mellem middelværdien og medianen. Generelt ser det ud til at mænd ligger lidt højere på BMI end kvinder.

d) Box-plot og Sammenligninger

Histogrammer kan illustrere tætheden for BMI'en for en given stikprøve. Dog kan det være vanskeligt at sammenligne 2 stikprøvers fordelinger med histogrammer. Derfor kan samme præsentation laves med et boxplot istedet for histogrammer. Nedenfor ses et box-plot af fordelingen af kvinder og mænd.

Code



Her er det meget mere tydeligt, at stikprøven for mænd har en tendens til at være mere overvægtige end kvinder. Box-plottet viser også meget tydeligt enkeltstående ekstreme tilfælde, her ses 2 ekstreme tilfælde i stikprøven for kvinder, og en enkelt i stikprøven for mænd. Det ligner også at den interkvartile rækkevidde (IQR) for stikprøven over mænd er mindre end for kvinder, hvilket kunne tyde på at BMI afviger mindre, altså spredningen er mindre for mænd end kvinder.

Med dette box-plot er det altså nemmere at sammenligne og få en fornemmelse af fordelingerne mellem stikprøverne for mænd og kvinder.

e) Nøgletal

Plots giver en god fornemmelse af hvordan data'en er fordelt og ser ud. Dog giver plots ikke nogle eksakte tal som kan bruges i beskrivelserne. Nedenfor er relevante nøgletal fundet og indskrevet i tabellen.

Code

Variabel	Antal Observationer	Middelværdi	Varians	Standard afvigelse	Nedre Kvartil	Median	Øvre kvartil
Alle	145	25,24	14,68	3,83	22,5	24,69	27,63
Kvinder	72	24,21	16,41	4,05	21,25	23,68	26,29
Mænd	73	26,26	11,06	3,32	24,15	25,72	28,63

Med en tabel udfyldt med nøgletal, kan vi mere præcist beskrive fordelingerne. Med box-plottet kunne man med gode øjne spotte sig til at stikprøven for mænd har en mindre spredning, men disse nøgletal afslører det direkte. Med histogrammerne og box-plottet kan præcise værdier ikke altid aflæses som eksempelvis nedre og øvre kvatiler. Disse plots afslører ofte heller ikke en præcis tæthed eller data centralitet, men giver blot en illustrerende ide om fordelingen. Nøgletal giver et indblik i præcis hvor data'ens centrale tendens er, og hvilke størelser stikprøvernes fordeling kan karakteriseres ved.

Statistisk analyse

f) Logaritmisk perspektiv

Den statiske model antages at være log normalt fordelt. Hermed:

$X_i \sim LN(\alpha, \beta^2)(\beta > 0), i = 1, \dots n$

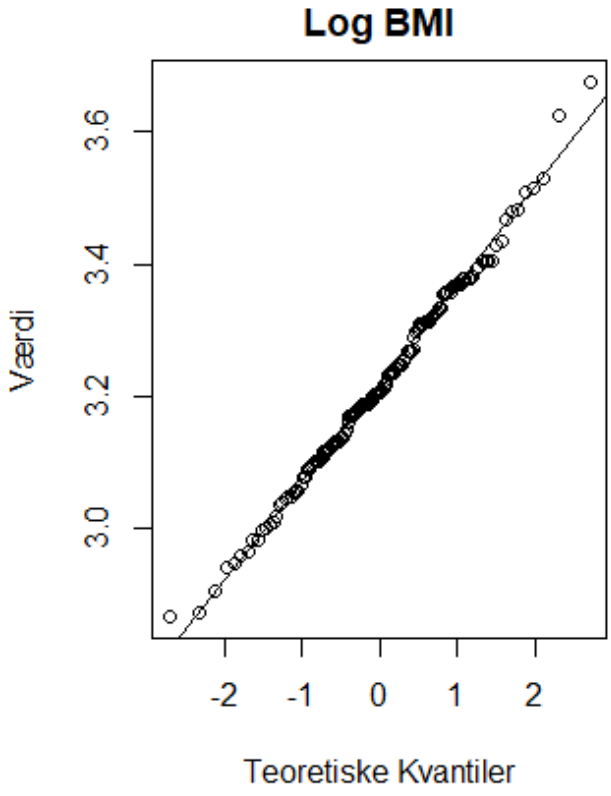
Code

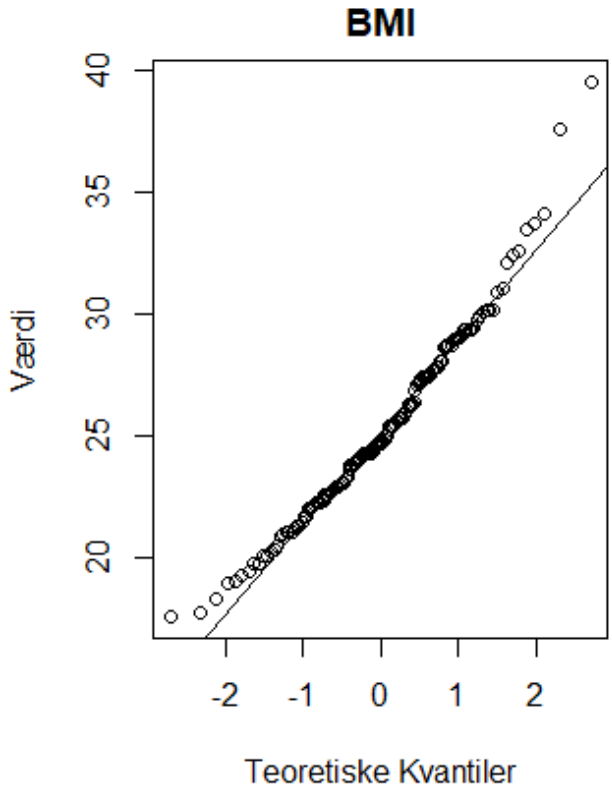
Estimationer baseret på stikprøven: - $\alpha = 3.217641$ - $\beta = 0.1488778$

Det kan med et QQ-plot ses at følge en log normal fordelings model. Det ses på plotsne nedenfor at de observerede BMI'er følger en teoretisk normal fordeling langt bedre ved logaritmisk transformation. Den logaritmisk transformeret data følger altså en normal fordeling bedre, da punkterne (data'en) følger linjen relativt godt. Det ses at rå BMI buer lidt op ved høje og lave værdier.

Code

Code





g) Konfidensinterval

Konfidensinterval kan findes med følgene formel:

- $\tilde{x} \pm t_{0.975} \times \frac{s}{\sqrt{(n)}}$

Hvor $\tilde{x}$ er middelværdien, s er standard afvigelsen og $t_{0.975}$ er t fordelingens 97.5 kvantile i $n-1$ frihedsgrader, hvor n er antal observationer af data'en og derfor 144 frihedsgrader i dette interval for logbmi.

Udregner i R med følgende:

- $\tilde{x} \approx 3.217$
- $s \approx 0.148$
- $n = 145$
- $t_{0.975} \approx 1.97$

Code

$KI \approx [3.11; 3.33]$

h) Hypotese

Givet hypotesen at middelværdien for den logaritmiske transformerede BMI er log(25) kan testes ved at finde P-værdien.

P-værdien for hypotesen findes ved formel:

- $p - value = 2 \times P(T > |t_{obs}|)$

Hvor T følger t fordelingen med (n-1) frihedsgrader.

Teststørelsen t_{obs} findes ved formel:

- $t_{obs} = \frac{\tilde{x} - \mu_0}{s/\sqrt{n}}$

Hvor $\tilde{x}$ er stikprøve middelværdien, s er stikprøve standard afvigelsen, n er antal observationer og μ_0 er null hypotesen.

Indsætter værdier og udregner med følgende i R:

- $\mu_0 = log(25)$
- $\tilde{x} = 3.217641$
- $s = 0.1488778$
- $n = 145$

Code

- $t_{obs} \approx -0.0999$
- $p \approx 0.92$

Med den givne hypotese fås en meget høj P værdi, som svare til at at der stortset ingen evidens er imod hypotesen. Givet denne data, er det ifølge denne hypotese test slet ikke umuligt at populations middelværdien for LogBMI er Log(25). Her afvises hypotesen ikke, da det formodes at signifikansniveauet er typisk ved 0.05. Med andre ord er der altså ikke statistisk signifikans til at afvise hypotesen, da p værdien ikke er mindre signifikans nivueat. Vi kan med denne hypotese test ikke konkludere at mere end halvdelen af populationen er overvægtig, men vi kan dog konkludere at det godt kunne lade sig gøre i følge den givne data.

En hypotese test kan også laves med R med funktionen `t.test()`, den giver samme resultater.

Code

i) Opdelte modeler

Samme procedure som i del f)

Kvinder:

- $K_i \sim LN(\alpha_k, \beta_k^2)(\beta_k > 0), i = 1, \dots n_k$

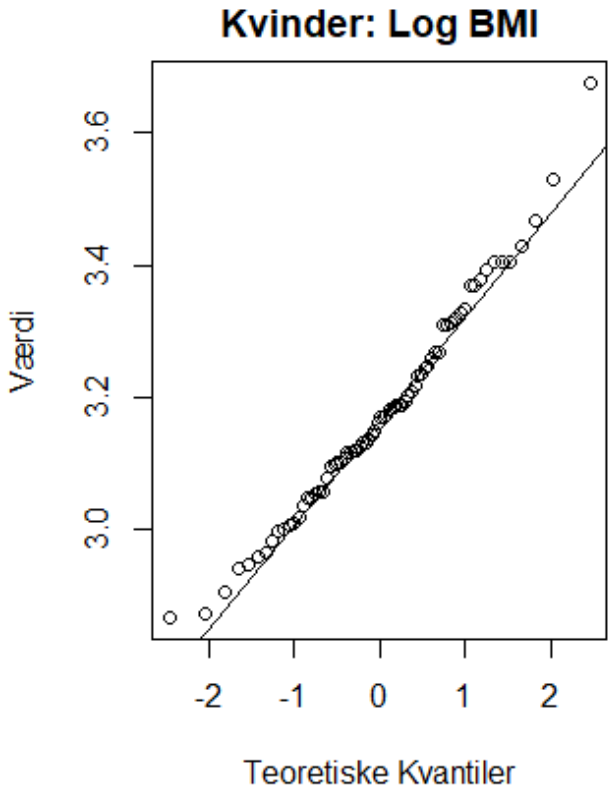
Code

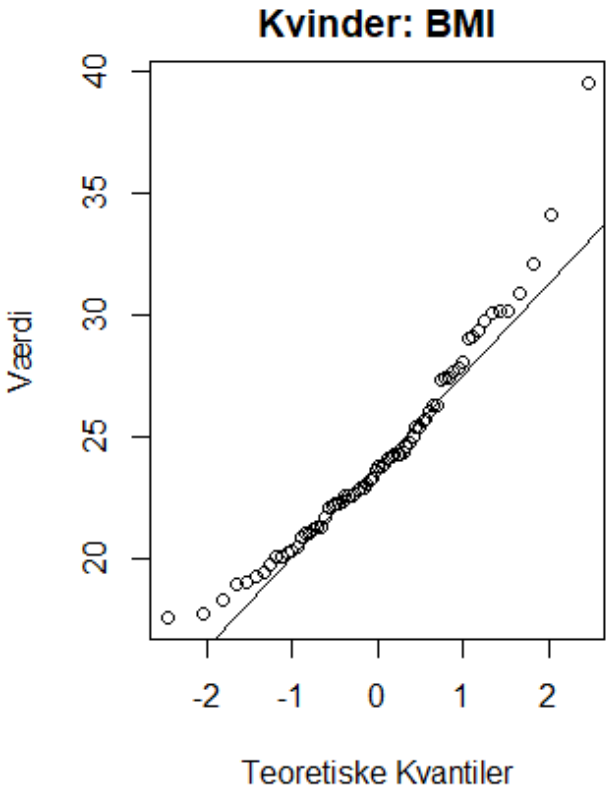
Estimationer baseret på stikprøven: - $\alpha_k \approx 3.17409681$ - $\beta_k \approx 0.15988773$

QQ-plots nedenfor afslører at log transformerede bmi'er passer bedre i en normal fordelings model end rå BMI. Herved en log normal fordeling.

Code

Code





Mænd:

- $M_i \sim LN(\alpha_m, \beta_m^2)(\beta_m > 0), i = 1, \dots n_m$

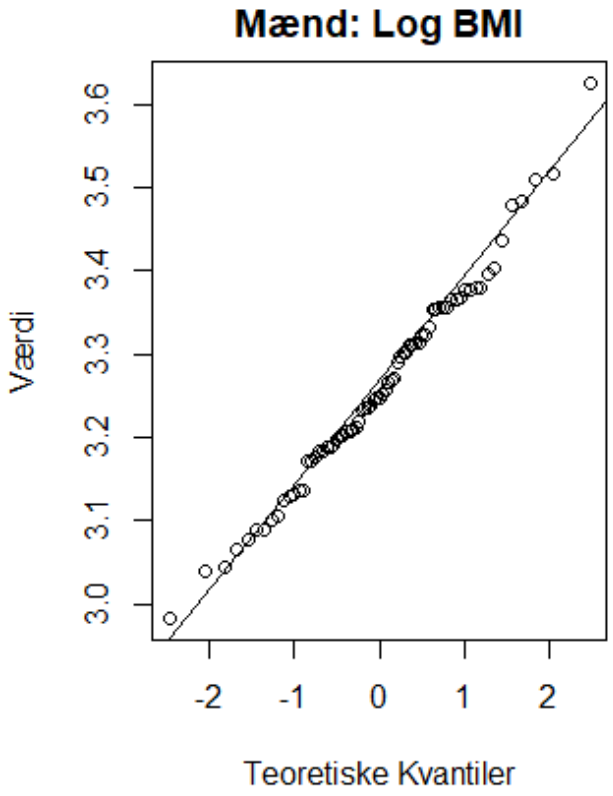
Code

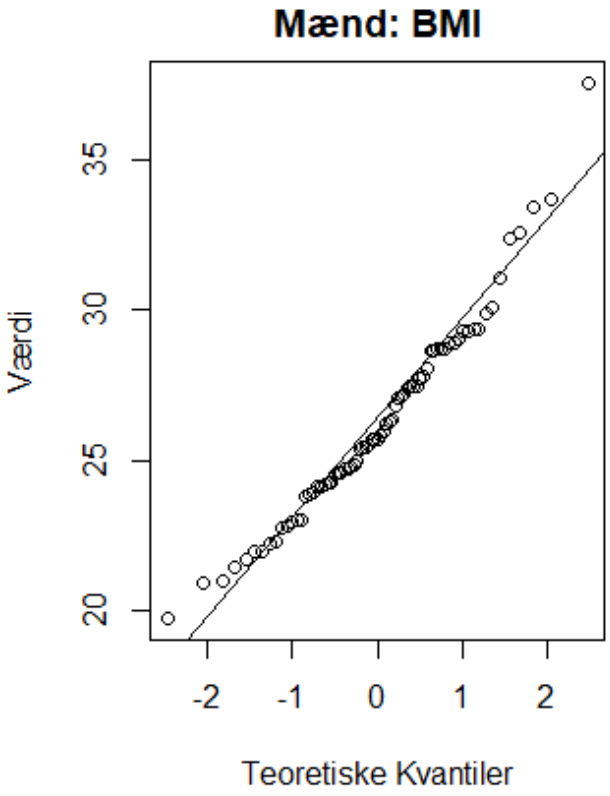
Estimationer baseret på stikprøven: - $\alpha_m \approx 3.26058769$ - $\beta_m \approx 0.12391144$

QQ-plots nedenfor afslørere at log transformerede bmi'er passer bedre i en normal fordelings model end rå BMI. Herved en log normal fordeling.

Code

Code





j) Konfidensinterval over median

Først kan konfidensintervalerne for middelværdien i logbmi udregnes.

Code

logbmi (middel)	Nedre	Øvre
Kvinder	3.136525	3.211669
Mænd	3.231677	3.289498

Hernæst kan vi transformere tilbage til bmi med ekspoential funktionen. Så fås følgende konfidens intervaller for mænd og kvinders medianer.

Code

bmi (median)	Nedre	Øvre
Kvinder	23	24.8
Mænd	25.3	26.8

k) Køns forskelle

Vi kan undersøge om der er statistisk signifikans i dataen til at påvise en forskel på kvinders og mænds BMI, det kan vi med hypotesen at deres middelværdier er ens.

- $H_0 : \mu_k = \mu_m \implies \mu_k - \mu_m = 0$
- $H_1 : \mu_k \neq \mu_m$
- $\alpha = 0.05$

Her bruges den log transformerede bmi, da den er lidt nemmere at arbejde med, da den følger en normal fordeling.

Nu da vi prøver at beskrive noget om 2 uafhængige forskellige populationer, altså mænd og kvinder, er det nødvendigt at bruge andre/udvidet formler end tidligere.

Nu bruges følgende formler til udregning af teststørelse, frihedsgrader og p værdi: (NB. $\tilde{x}$, S, n er altså middelværdi, standard afvigelse og antal observationer for stikprøven og ikke populationen, og δ_0 er middelværdi differensen i hypotesen, hvilket i denne hypotese er 0):

- $t_{obs} = \frac{(\tilde{x_1} - \tilde{x_2}) - \delta_0}{/\sqrt{S_1^2/n + S_2^2/n}}$
- $V = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$

Indsætter følgende og udregner med R.

Code

- Logbmi middelværdi i stikprøven for mænd: $\tilde{x_1} = 3.26058$
- Logbmi middelværdi i stikprøven for kvinder: $\tilde{x_2} = 3.17409$
- Standardafvigelse for mænd : $s_1 \approx 0.123$
- Standardafvigelse kvinder : $s_2 \approx 0.159$
- Antal observationer for mænd: $n_1 = 73$
- Antal observationer for kvinder: $n_2 = 72$

Hermed fås:

Code

- $t_{obs} \approx 3.637$
- $V \approx 133.75$

Nu er det blot at "slå op" i t fordelingen med test størelsen og frihedsgraderne i formlen:

- $p - value = 2 \times P(T > |t_{obs}|)$

Her fås p værdien:

Code

- $p \approx 0.00039$

Det er en lille p værdi, hvilket betyder at der er stærk evidens for at hypotesen ikke er sand. Givet signifikans niveauet på 0.05 er hypotesen afvist, da p værdien er mindre end signifikans niveauet. Med andre ord viser undersøgelsen at der er en forskel på kvinder og mænds logbmi i populationen med minimum 95% sikkerhed.

Der fås samme resultater med `t.test()` funktionen i R.

l) Inference med konfidensintervaler

Hvis vi kigger på 95% konfidensintervallerne igen, ser vi at den højeste værdi for kvinder er mindre end den laveste værdi for mænd.

Middel LogBmi	Nedre	Øvre
Kvinder	3.136525	3.211669
Mænd	3.231677	3.289498

Det betyder at konfidensintervallerne ikke overlapper og derfor ikke har nogen fællesværdier.

Hvis vi er 95% sikre på at middelværdierne i 2 forskellige uafhængige populationer ligger i 2 intervaller som ingen fællesværdier har, er vi også 95% sikre på at de 2 populationers middelværdi er forskellig.

Derfor kunne vi blot ved at kigge på konfidensintervalerne, afvise hypotesen om at kvinder og mænd har samme logbmi middelværdi givet signifikans niveaut 0.05.

m) Korrelation

Korrelation af stikprøvens variable kan findes ved at finde korrelations koofficienten med formler:

- $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \tilde{x})(y_i - \tilde{y})$
- $r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \tilde{x}}{s_x})(\frac{y_i - \tilde{y}}{s_y}) = \frac{s_{xy}}{s_x \times s_y}$

Hvor $\tilde{x}, x, s_x$ er den tilsvarende middelværdi, observations værdi og standard afvigelse for den givne observations variabel.

Code

Indsætter følgende og beregner med R:

- $s_x \approx 3.83$
- $s_y \approx 15.20$
- $\tilde{x} \approx 25.24$
- $\tilde{y} \approx 76.73$
- $r = \frac{1}{144} \times ((\frac{24.7 - 25.24}{3.83})(\frac{80 - 76.73}{15.20}) + \cdots + (\frac{39.5 - 25.24}{3.83})(\frac{105 - 76.73}{15.20})) \approx 0.82826$

Code

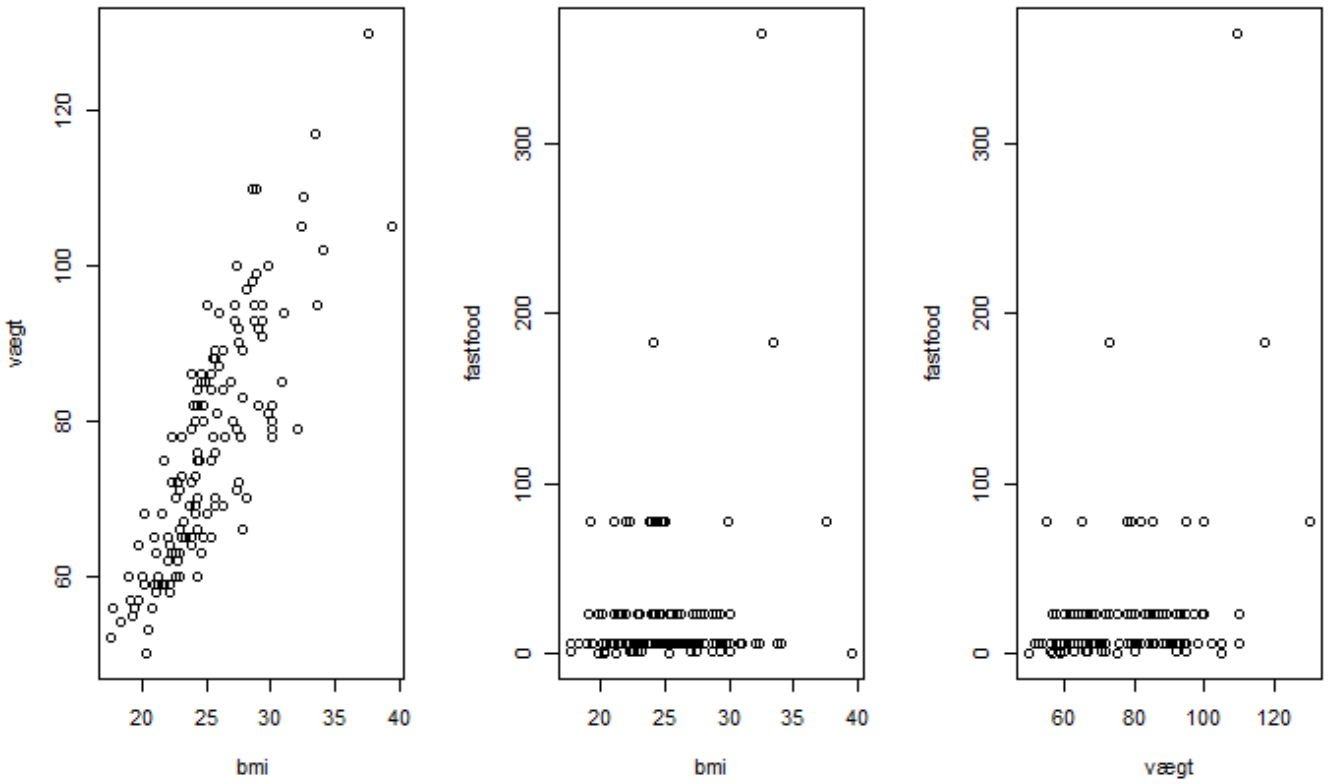
Vi har også korrelationerne mellem fastfood og bmi og vægt og fastfood.

- Korrelation af Vægt og Fastfood ≈ 0.279
- Korrelation af Bmi og Fastfood ≈ 0.153

Nedenfor ses scatter plots af data'en med de 3 parvise konfigurationer af BMI, Fastfood og Vægt.

Code

Code



Korrelations kooficienterne er som forventet. Korrelationen mellem vægt og bmi er omkring 0.8 hvilket i min optik viser korrelation, dermed ser man også på plottet at de data punkterne følger en lineær tendens, når bmi bliver høj bliver vægt også højere. Derudover ser vi også at korrelationerne for de konfigurationer der indeholder fastfood ikke er særlig høje, hvilket også kan ses på de meget flade "linjer" data'en danner. Dog kan dette skyldes data representationen. Måske ville det se anderledes ud hvis der var et præcist antal fastfood besøg.