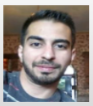
02323- Statistik



Muhammad Ali Khan Bangash

Brugernavn	Email	Studienummer	Uddannelse
s092512	s092512@student.dtu.dk	s092512	diploming. Softwaretek.

Indhold

a)

Datamateriale består af et datasæt af 145 forskellige observationer som skal bruges til at undersøge BMI (Body Mass Index) blandt Danmarks befolkning. Datamateriale kan kategoriseres i 5 følgende variabler :

- Højde
- Vægt
- Køn
- Urbanitet
- Fastfood

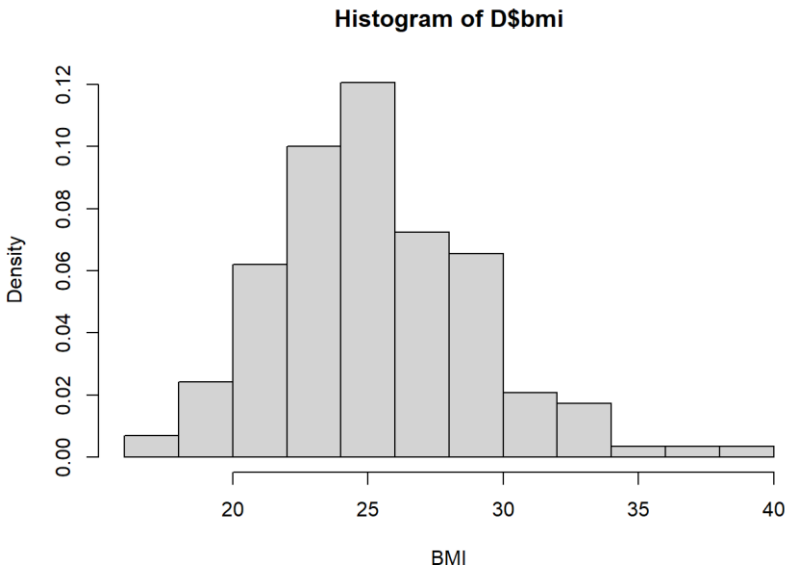
I projektet er respondenternes højde målt i cm, vægt i kg. Urbanitet er kategoriseret fra 1-5 og fastfood er respondentens indtægt af fastfood målt i dage pr år. Højde, vægt og fastfood er kvantitativ variabler. Derimod er urbanitet og køn kategoriseret variabler der løber fra 1-5 og 0-1 respektive. Der mangler ikke nogen værdier i datasættet. Man kan tage summary af datasættet for at undersøge kvartiler, median, minimum og maksimum.

```
     height          weight          gender          urbanity        fastfood
 Min.   :154.0   Min.   : 50.00   Min.   :0.0000   Min.   :1.000   Min.   :  0.00
 1st Qu.:166.0   1st Qu.: 65.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:  6.00
 Median :173.0   Median : 75.00   Median :1.0000   Median :4.000   Median :  6.00
 Mean   :173.9   Mean   : 76.74   Mean   :0.5034   Mean   :3.669   Mean   : 21.04
 3rd Qu.:182.0   3rd Qu.: 87.00   3rd Qu.:1.0000   3rd Qu.:5.000   3rd Qu.: 24.00
 Max.   :196.0   Max.   :130.00   Max.   :1.0000   Max.   :5.000   Max.   :365.00
```

b)

BMI kan beregnes ud fra følgende formel:

$$BMI = \frac{vægt}{højde^2}$$



For at beskrive fordelingen af BMI-værdierne i Histogram, skal man regne middelværdien og median af BMI. Hvis gennemsnit er mindre end median vil fordeling være venstreskæv og hvis gennemsnittet er større end median, er fordelingen højre skæv. Disse to værdier kan regnes ud i R studio og det kan ses ved udregning at middelværdien/gennemsnittet er 25.25 og median er 24.69, hvilket vil sige at middelværdien er større end median og derfor kan det konstateres at fordelingen er højre skæv.

BMI kan ikke være negativ da i formlen $BMI = \frac{vægt}{højde^2}$, er højden og vægt både en positiv tal og kan ikke være negativ.

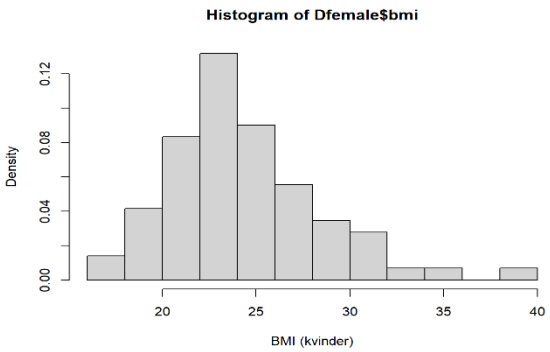
Spredning kan findes ved at beregne standardafvigelse ved hjælp af følgende formel:

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

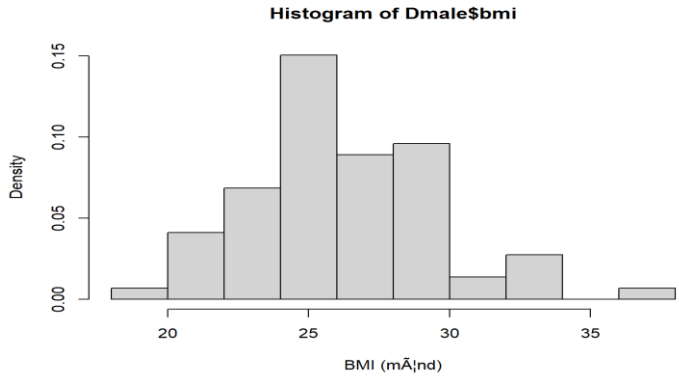
-

Resultater = 3.83, hvilket er ikke meget spredning i observationer.

c)



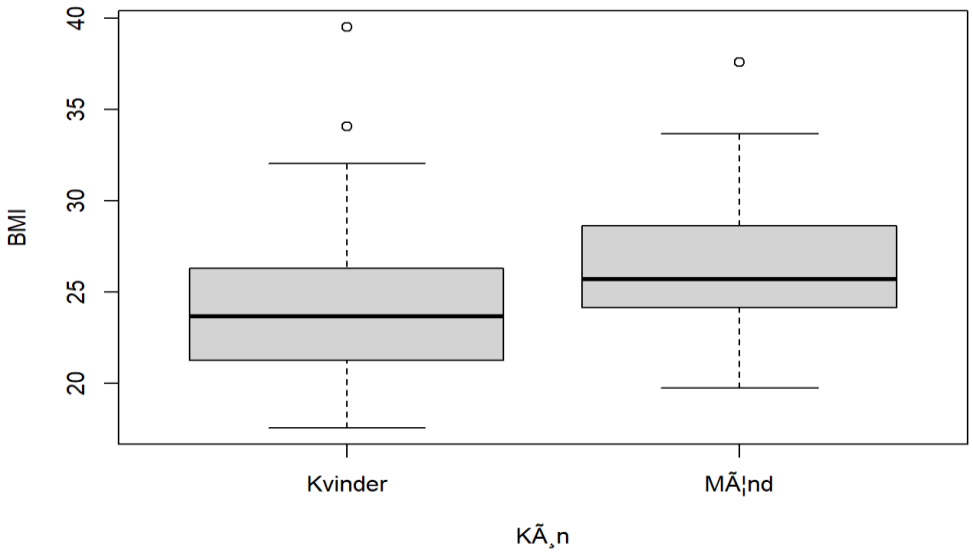
Figur 3 Histogram for kvinder



Figur 4 :Histogram for Mænd

Ved udregninger I R studio kan des ses at kvinder i gennemsnit har BMI på 24.21og median på 23.68 og ligeledes har mænd en gennemsnit i BMI på 26.26 og median på 25.72. Der kan konstateres at begge køn har en gennemsnit større end median, hvilket vil sige at fordelingen er højre skæv.

d)



Figur5: Boksplot for mænd og kvinders BMI

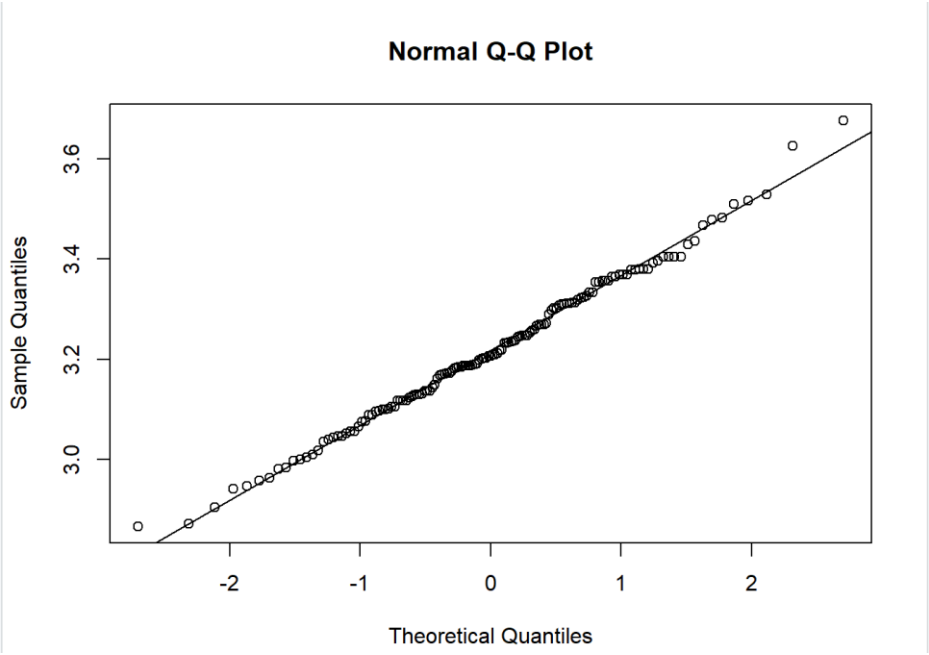
Det kan dog ses tydeligt på de boksblot for mænd og kvinders BMI at der er symmetrisk fordeling ved kvinderne omkring median, derimod er fordeling skæv hos mænd da median ikke er placeret mellem første og sidste kvartil. Kvinders øvre og nedre kvartiler er lavere end mænd hvilket er ensbetydende med at spredning er større hos kvinder end mænd.

e)

Variable: BMI	Antal obs.	Stikprøve-gennemsnit	Stikprøve-varians	Stikprøve-standard-afvigelse	Nedre Kvartil	Median	Øvre kvartil
	n	($\bar{x}$)	(s^2)	(s)	(Q1)	(Q2)	(Q3)
Alle	145	25,24795	14,6808	3,832243	22,58955	24,69136	27,6305
Kvinder	72	24,2164	16,41787	4,051897	21,25850	23,68911	26,29172
mænd	73	26,26536	11,06872	3,326969	24,15167	25,72552	28,63404

I den øvre tabel kan det ses at man har en an præcis tal på median og middelværdi. Det kan dog også bekræftes at kvinder har lavere BMI i alle kvartiler end mænd.

f)



Statistic model for logaritmen til BMI kan skrives på formlen :

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, \ldots, n.$$

Udregninger fra R studio viser at formlen kan skrives med følgende tal:

$$X_i(3.22, 0.149^2) where\ i = 1,..145$$

Hvor 3.22 er middelværdien for log BMI af vores datasæt og 0.149 er standardafvigelse af log BMI af datasættet.

g)

Formlen for konfidensinterval for middelværdien kan beregned ved hjælp af følgende formel

$$\bar{x} \pm t_{1-\alpha/2} * \frac{s}{\sqrt{n}}$$

$Middelværdien\ \bar{x} = 3.22$

$t - kvantiler\ findes\ ved\ at\ at\ taste\ qt(0.975, n-1), hvor\ n = antal\ af\ obs, t = 1.977$

$s = standardafvigelse\ af\ logBMI = 0.149$

$n = 145$

Indsætter ovenstående værdier ind i formlem for konfidensinterval får vi :

$$3.22 \pm 1.977 * \frac{0.149}{\sqrt{145}} = 3.244{,}3.196$$

$$3.22 + \frac{1.977 \cdot 0.149}{\sqrt{145}} \xrightarrow{\text{at 5 digits}} 3.2445$$

$$3.22 - \frac{1.977 \cdot 0.149}{\sqrt{145}} \xrightarrow{\text{at 5 digits}} 3.1955$$

For at finde 95% af konfidensinterval kan man tage eksponentiel værdi af det middelværdi som vi har fundet ovenover dvs.:

exp(3.244) og exp (3.1994) eller i R-studio ved følgende formel :

```
KI <- t.test(D$logbmi, conf.level=0.95)$conf.int
KI
exp(KI)
```

Resultat = [24.36635 , 25.58684]

h)

For at undersøge om middelværdien for logBMI er forskellige fra log(25), testes der følgende hypotese:

$H_0: \mu_{logBMI} = \log(25)$
$H_1: \mu_{logBMI} \neq \log(25)$

Test størrelsen kan beregnes ved følgende formel:

$$T_{obs} = \frac{\bar{x} - u_0}{s/\sqrt{n}}$$

Vi indsætter værdierne :

$$\frac{3.22 - \log(25)}{\frac{0.149}{\sqrt{145}}} \xrightarrow{\text{at 5 digits}} 0.096980$$

$T_{obs} = \frac{3.22 - \log(25)}{0.149/\sqrt{145}} = 0.097$

For at beregne p-værdien bruger vi følgende formel i R-studio:

$p - værdien = 2 * pt(-abs(T_{obs}), df = n - 1)$
$hvor\ T_{obs} = 0.097, n = 145$
$p - værdien = 0.922$

Ifølge hypotesen, signifikansniveauet alpha $\alpha = 5\ \% = 0.05\ \%$. Hvis p-værdien er mindre end signifikansniveauet, forkastes hypoteseteorien. Da vores p-værdi = 0.922 > 0.5 kan det konkluderes at der nul hypotesen H0 holder vand. Da median ikke er forskellig fra 25, kan det også konkluderes at mere end halvdelen af befolkning må være overvægtig.

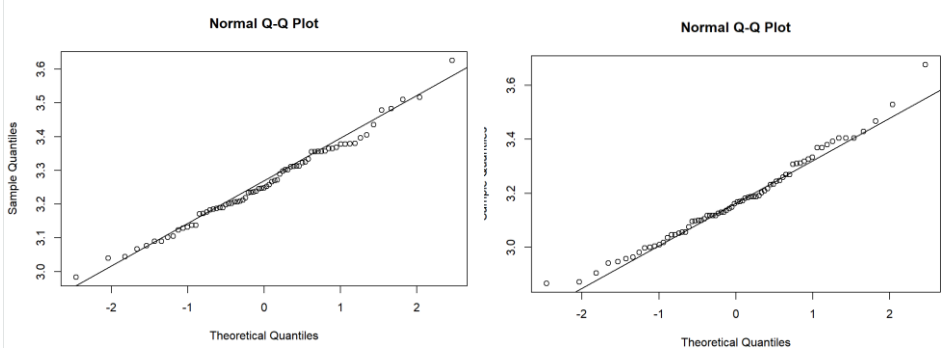
I)

Statistik model for logaritmen for BMI kan skrives ved hjælpe af følgende formel :

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, \ldots, n.$$

$$hvor\ \mu = middelværdien, \sigma = standard\ afvigelse$$
$$for\ kvinder:\ X_i(3.17{,}0.16^2), hvor\ i = 1..72$$
$$for\ mænd: X_i(3.26{,}0.12^2), hvor\ i = 1..73$$

Når mam sammenligner QQ-plots for mænd og kvinder, ser normalfordeling ud til at følge den rette linje.



Qq plot male *Qq plot- female*

j)

Formlen for konfidensinterval for middelværdien kan beregnes ved hjælp af følgende formel :

$$\bar{x} \pm t_{1-\alpha/2} * \frac{s}{\sqrt{n}}$$

For kvinder:

$Middelværdien\ \bar{x} = 3.17$

$t - kvantiler\ findes\ ved\ at\ at\ taste\ qt(0.975, n-1), hvor\ n = antal\ af\ obs, t = 1.944$

$s = standardafvigelse\ af\ logBMI = 0.16$

$n = 72$

Indsætter ovenstående værdier ind i formlen for konfidensinterval får vi :

$$3.17 \pm 1.944 * \frac{0.16}{\sqrt{72}} = 3.136, 3.21$$

exp(3.136), exp (3.21) = 23.02372 , 24.82047

For mænd:

$Middelværdien\ \bar{x} = 3.260$

$t - kvantiler\ findes\ ved\ at\ at\ taste\ qt(0.975, n-1), hvor\ n = antal\ af\ obs, t = 1.993$

$s = standardafvigelse\ af\ logBMI = 0.124$

$n = 73$

Indsætter ovenstående værdier ind i formlen for konfidensinterval får vi :

$$3.260 \pm 1.993 * \frac{0.124}{\sqrt{73}} = 3.231677\ , 3.289498$$

exp(3.232), exp (3.299) = 25.32209 , 26.82940

	Nedre grænse af KI	**Øvre grænse af KI**
Kvinder	23.02	24.82
Mænd	25.32	26.82

k)

For at beregne test størrelsen for sammenligning af to stikprøver:

$$t_{obs} = \frac{(x_1 - x_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_{obs} = \frac{(3.17 - 3.26) - 0}{\sqrt{\frac{0.16^2}{72} + \frac{0.124^2}{73}}} = -3.78$$

Frihedsgrader kan bestemmes ved hjælp af følgende formel:

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

$$\frac{\left(\frac{0.16^2}{72}+\frac{0.124^2}{73}\right)^2}{\frac{\left(\frac{0.16^2}{72}\right)^2}{71}+\frac{\left(\frac{0.124^2}{73}\right)^2}{72}}$$

$$v = 133.7508765 \approx 133.75$$

P værdien kan regnes ved formlen:

2*pt(-abs(3.78),df=133.75-1)

p-værdi = 0. 0002360425

Som hovedregel bruger vi signifikansniveauet på 5% = 0.05 og derfor kan det konstateres at p-værdien < signifikansværdi dvs. 0.00023 < 0.05 og derfor kan nul hypotesen forkastes.

L)

Hypotesetest er unødvendigt da allerede ved nærmere undersøgelse af data fra spørgsmål delen J kunne vi konstatere at konfidensintervaller for mænd og kvinder er forskellige fra hinanden og derfor kunne det konkluderes at de er signifikant forskellige.

m)

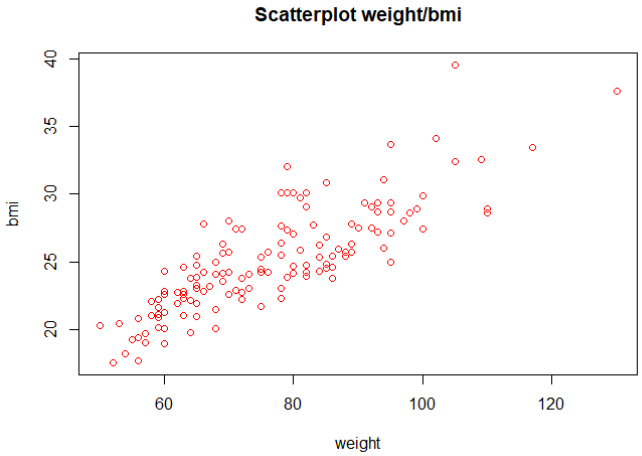
Korrelation mellem BMI og vægt kan findes ved hjælp af følgende formel :

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

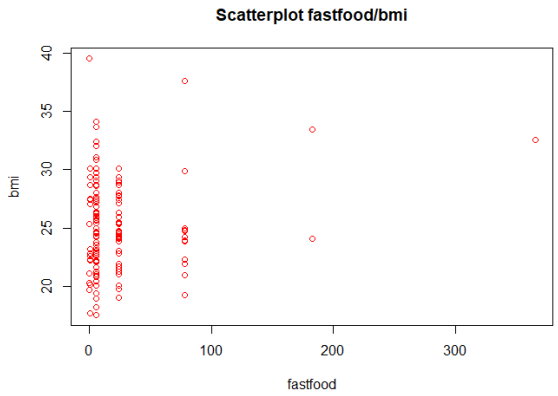
Korrelation	**R kode**	**Resultat**
BMI:vægt	cor(D[,c("weight","bmi")], use="pairwise.complete.obs")	0.828261
BMI:fastfood	cor(D[,c("fastfood","bmi")], se="pairwise.complete.obs")	0.1531578

Vægt:fastfood	cor(D[,c("weight","fastfood")], use="pairwise.complete.obs")	0.279322

På nedenstående scatterplot mellem vægt og BMI kan det ses at BMI er direkte proportionale med vægt, dvs. at når den ene stiger så stiger den anden variabel også.



På nedenstående scatterplot mellem fastfood og bmi tyder det heller ikke på at de to variabler forudsiger hinanden. Det vil sige at der ikke er sammenhæng mellem de to variabler.



På nedenstående scatterplot mellem vægt og fastfood, tyder der heller ikke noget på at der er sammenhæng mellem de to variabler. Dvs. at den ene forudsager ikke den anden.