

Eksamensopgave 2023

1) Datapræsentation

1.

Filnavn på mit personlige datasæt:

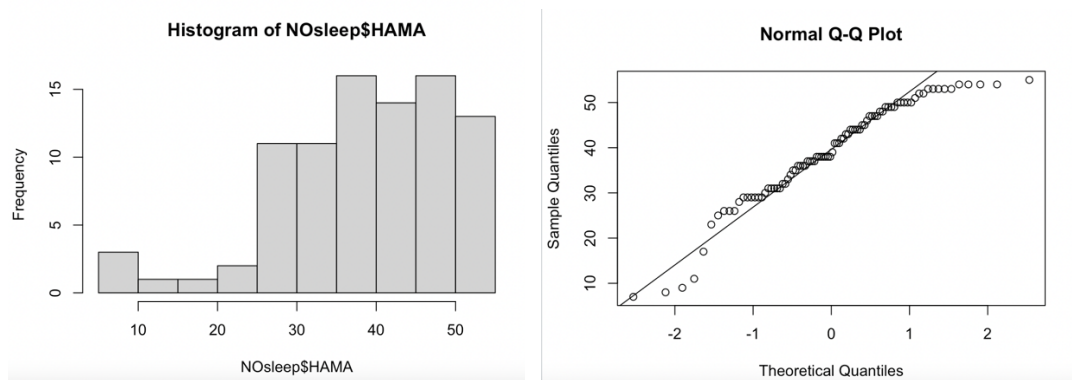
```
d <- read.csv("http://causal.sund.ku.dk/f23/15.csv", header=TRUE, stringsAsFactors=TRUE)
```

2.

I denne opgave tages der udgangspunkt i den videnskabelige hypotesen om at personer med angst har en øget risiko for tandskæren og denne parameter kan være påvirket af andre faktorer. Jeg ønsker derfor at udføre en tabel 1, der sammenligner personer med angst der sover uroligt om natten, med personer der har angst og *ikke* sover uroligt om natten. Jeg har derfor valgt, at dele datasættet op på baggrund af gruppen sleep. Datasættet opdeles i "NOsleep" svarende til personer uden søvnforstyrrelse som angives med værdien 0, og gruppen "YESsleep" svarende til personer med søvnforstyrrelser med værdien 1. Dermed opnår man et bedre overblik over sit data.

Datasættet indeholder 5 variable henholdsvis "HAMA", "alkohol", "sex", "bruxism" og "sleep". I dette datasæt vil variableerne "HAMA" og "bruxism" være de numeriske variable (diskret), hvor de resterende er kategoriske binære variable. En numerisk variabel beskrives via dens typiske værdi (gennemsnit eller median) og dens variation (spredning eller IQR). Jeg tegner et histogram for de numeriske variable for at vurdere om datasættet er symmetrisk fordelt eller ej. Derudover laves et QQ-plott for at understøtte vores histogram da vi kun tager udgangspunkt i en lille stikprøve.

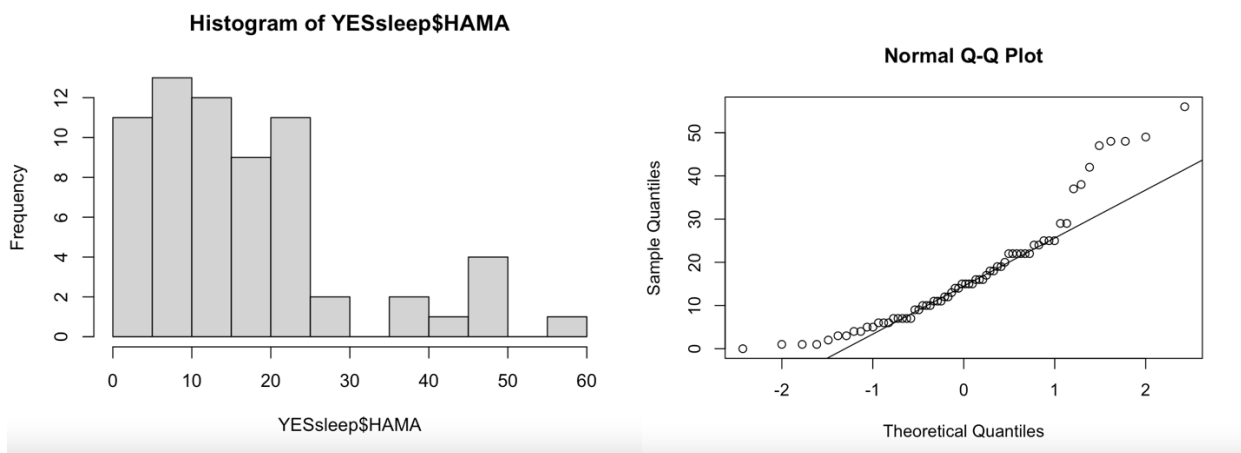
Histogram og QQ-plot af 'HAMA' for gruppen med *NO* sleep (ingen søvnforstyrrelser):



Jeg vurderer, at dataet i histogrammet af HAMA for gruppen er skævt fordelt mod højre side, men at dataet i QQ-plottet af HAMA for gruppen er symmetrisk fordelt med få afvigelser. På baggrund af QQ-plottet bestemmer jeg dataet til at være symmetrisk. Derfor bestemmer jeg middelværdien og spredningen for den typiske værdi og variation:

Middelværdien for gruppen med personer uden søvnforstyrrelser er 39.01 og spredning er 11.16.

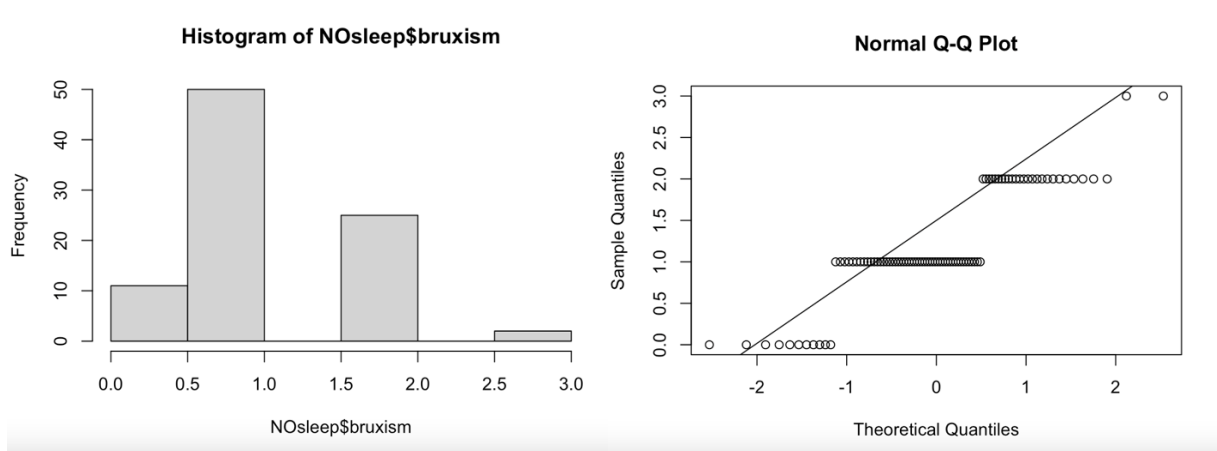
Histogram og QQ-plot af 'HAMA' for gruppen med *YES* sleep (søvnforstyrrelser):



Jeg vurderer, at dataet i histogrammet af HAMA for gruppen *YES* sleep er skævt fordelt mod venstre side, men at dataet i QQ-plottet af HAMA for gruppen *YES* sleep er nogenlunde symmetrisk fordelt dog med få afvigelser. På baggrund af QQ-plottet bestemmer jeg dataet til at være symmetrisk. Derfor bestemmer jeg middelværdien og spredningen for den typiske værdi og variation.

Middelværdien for gruppen med personer med urolig nattesøvn er 17.05 og spredning er 13.14.

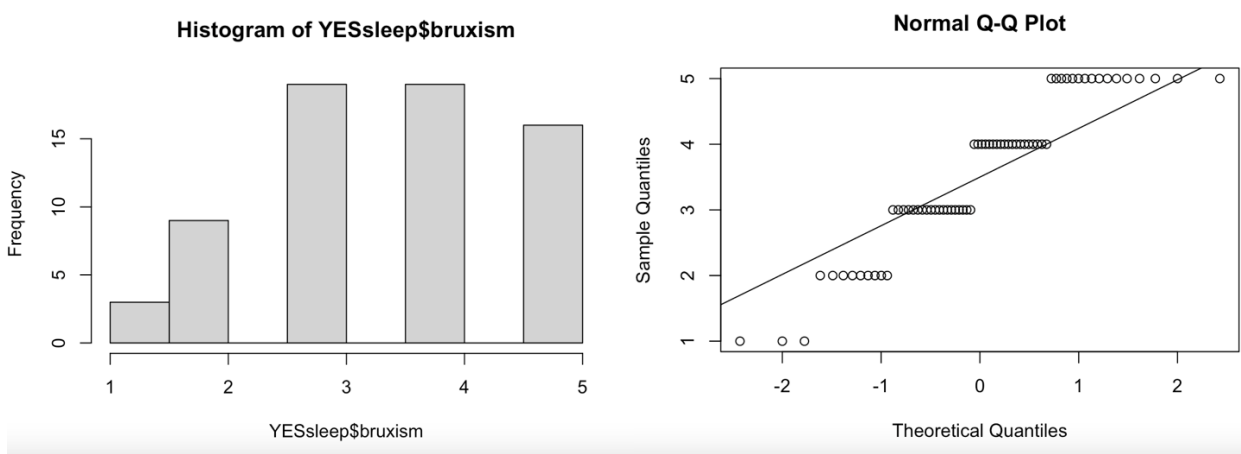
Histogram og QQ-plot af ”bruxism” for gruppen med *NO* sleep (ingen søvnforstyrrelser):



Jeg vurderer, at histogrammet og QQ-plottet for bruxism ved gruppen med *NO* sleep er skævt fordelt, og derfor bestemmer jeg medianen og IQR som den typiske værdi og variation.

For gruppen af personer med *NO* sleep er medianen 1 og IQR er 1.

Histogram og QQ-plot af ”bruxism” for gruppen med *YES*sleep (søvnforstyrrelser):



Jeg vurderer, at histogrammet og QQ-plottet for bruxism ved gruppen med *YES* sleep er skævt fordelt, og derfor bestemmer jeg medianen og IQR som den typiske værdi og variation.

For gruppen af personer med *NO* sleep er medianen 4 og IQR 1.

I datasættet er der tre kategoriske binære variable hhv. "alkohol", "sex" og "sleep". En kategorisk variabel beskrives via dens frekvens og relative frekvens. Da jeg har opdelt mit datasæt efter "sleep" bestemmer jeg frekvensen og den relative frekvens for "alkohol" og "sex".

Jeg bestemme frekvensen af "alkohol" i en opsummeringstabel for gruppen i *NO* sleep idet tabellen angiver frekvensen. Jeg bestemmer den relative frekvens ved at dividere frekvensen med det samlede antal deltagere i gruppen *NO* sleep og ganger med 100%. Det samlede antal forsøgsparticipanter i gruppen *NO* sleep er 88, som er aflæst i RStudio:

	Alkohol (ja=1)	Alkohol (nej=0)
Frekvens	16	72
Relativ frekvens	18.18%	81.82%

Jeg bestemme frekvensen på samme måde af gruppen "alkohol" i en opsummeringstabel i gruppen for *YES* sleep. Det samlede antal forsøgsparticipanter i gruppen *YES* sleep er 66, som er aflæst i RStudio:

	Alkohol (ja=1)	Alkohol (nej=0)
Frekvens	15	51
Relativ frekvens	22.73%	77.27%

Til at bestemme frekvensen af "sex" laver jeg ligeledes en opsummeringstabel for hver gruppe, og bestemmer den relative frekvens på samme måde som overstående. For gruppen med *NO* sleep:

	Mand	Kvinde
Frekvens	44	44
Relativ frekvens	50%	50%

For gruppen med *YES* sleep:

	Mand	Kvinde
Frekvens	29	37
Relativ frekvens	43.94%	56.06%

Opsummering i en **tabel 1**:

TABEL 1		
N=154	Personer UNDEN søvnforstyrrelser (NOsleep)	Personer MED søvnforstyrrelser (YESsleep)
Antal forsøgspersoner	88	66
<u>HAMA</u>	39.01 (11.16)	17.05 (13.14)
Mean (SD)		
<u>Bruxism</u>	1 (1)	4 (1)
Median (IQR)		
<u>Sex</u>		
(Relativ frekvens)		
- Mand	44 (50%)	29 (41.94%)
- Kvinde	44 (50%)	37 (56.04%)
<u>Alkohol</u>		
(Relativ frekvens)		
- 1 (Ja)	16 (18.18%)	15 (22.73%)
- 0 (Nej)	72 (81.82%)	51 (77.27%)

2) Sammenhæng mellem angst og tandskæren

1.

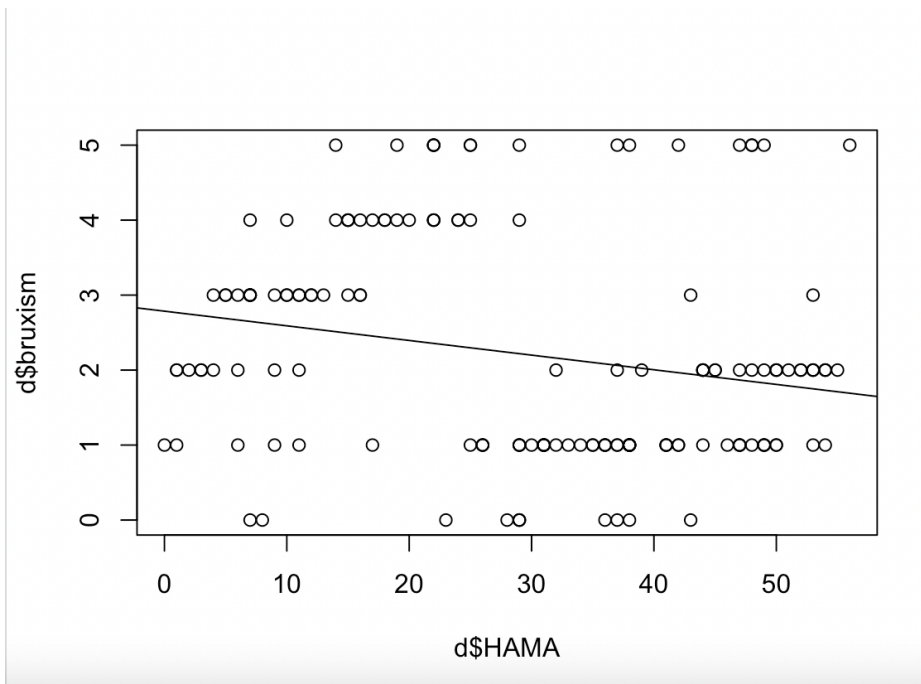
I denne opgave er den afhængige variabel tandskæren (bruxism), som er en numerisk variabel, og den uafhængige variabel er angst (HAMA), som er en numerisk variabel. Derfor vælger jeg en lineær regression som den statistiske model.

2.

Modellen estimeres via RStudio kommando. Herfra får jeg følgende værdier til at skrive den estimerede model op som:

$$Bruxism = 2.786460 - 0.019551 \cdot HAMA$$

Til at visualisere data tegner vi et scatter plot, hvortil den afhængige variabel ”bruxism” er indsat på y-aksen, og den uafhængige variabel ”HAMA” er indsat på x-aksen:



Skæringspunktet med y-aksen for den rette linje svarer til parameteren $a = 2.79$ og hældningen for den rette linje er $b = -0.0196$. Jeg fortolker resultatet som den gennemsnitlige sværhedsgrad af bruxisme er omkring 3 (2.79), men sværhedsgraden aftager med -0.0196 hver gang man bevæger sig en gang højere op på angst skalaen.

3.

Jeg bestemmer et 95% konfidensinterval og på baggrund af resultatet er jeg 95% sikker på, at intervallet $[-0.005, -0.34]$ indeholder den sande værdi af hældningskoefficienten mellem ”HAMA” og ”bruxism” i populationen. Jeg er også 95% sikker på, at intervallet $[2.3, 3.27]$ indeholder den sande værdi for skæringspunktet i populationen.

4.

For at undersøge om effekten af graden angst på graden af tandskæren er statistisk signifikant vælger jeg at se på p-værdien. RStudio tester altid en nulhypotese om, at parametrene er lig med 0. Nulhypotesen for denne parameter siger hældningskoefficienten er lig med 0. Det vil betyde, at graden af bruxisme holdes konstant selvom graden af angst ændre sig. Begge parametre har en p-værdi der er mindre end 5%, hvor $p < 0.001$, og dermed kan jeg forkaste nulhypoteserne. Det betyder, at der er statistisk evidens for en signifikant sammenhæng mellem graden af angst og bruxisme.

5.

Til at prædiktere den gennemsnitlige grad af tandskæren ved en angstgrad på 20, skal man indsætte 20 ind i den estimerede model lavet i opgave 2.2. Hertil får jeg, at ved en angstgrad på 20 ligger man på 2 (2.39544) i bruxisme skalaen. Jeg valgt af runde ned, da skalaen er angivet i hele tal. Vi kan herefter bestemme et 95% referenceinterval for angstgrad ved at benytte formlen:

$$95\% \text{ referenceinterval} = \text{middelværdi} \pm 2 \cdot \text{spredning}$$

Middelværdien i denne opgave er bestemt til 2.395 og spredningen bestemmer vi til 1.442 som er aflæst i RStudio under "residual standard error". Dermed er et 95% referenceinterval for angstgrad på 20 givet ved $[-0.489; 5.279]$. Det vil sige, at 95% af personer med en angstgrad på 20 vil kunne befinde sig på bruxisme skalaen fra 0-5.

6.

Vi undersøgte, om der er statistisk evidens for om der er sammenhængen mellem graden af tandskæren og graden af angst, hvor vi testede en nulhypotese om at graden af bruxisme holdes konstant selvom graden af angst ændre sig. Denne nulhypotese blev forkastet, da begge parametre har en $p < 0,001$. På baggrund af dette kan vi konkludere, at der er signifikant sammenhæng mellem graden af angst og bruxisme. Denne sammenhæng blev undersøgt med en lineær regression. Ud fra resultaterne kan jeg konkludere, at graden af bruxisme falder i gennemsnit med -0.0196 for hver gang med bevæger sig én gang højere op på angst skalaen. Det betyder, at graden af bruxisme falder, jo højere man scorer på angst-skalaen.

3) Sammenhæng mellem tandskæren og søvnforstyrrelser

1.

I denne opgave er den afhængige variabel søvnforstyrrelser "sleep", som er en kategorisk binær variabel, og den uafhængige variabel er graden af tandskæren "bruxisme", som er en numerisk variabel. Jeg vælger derfor at lave en logistisk regression som den statistiske model.

2.

En logistisk regressionsmodel med en uafhængig variabel er defineret som en lineær sammenhæng på log Odds-skala og er givet ved følgende ligning:

$$\log Odds(y) = a + b \cdot x$$

Ud fra outputtet på RStudio kan jeg skrive den estimerede model op på log Odds-skalaen som:

$$\log Odds(\text{sleep}) = -5.9836 + 2.6303 \cdot \text{bruxisme}$$

Den estimerede model viser, at 2.6303 svarer til betydning af bruxisme for modellen svarende til hældningskoefficienten. Yderligere angiver -5.9836 skæring med y-aksen.

Det tilhørende 95% konfidensinterval aflæses til [1.876, 3.659]. Jeg er 95% sikker på, at dette interval indeholder den sande værdi for effekten af bruxisme på log Odds-skalaen. RStudio tester nulhypotesen, at parameteren er lig med 0, og der derfor ikke er sammenhæng mellem bruxisme og søvn. Her fik vi et resultat, hvor begge parametre har en $p < 0,001$, og dermed kan vi forkaste nulhypoteserne. På baggrund af dette kan vi konkludere, at der er signifikant sammenhæng mellem bruxisme og søvn.

3.

Sandsynlighedsskalaen i en logistisk regressionsmodel er opgivet med resultater fra overstående opg:

$$p(\text{sleep} = 1) = \frac{e^{a+bx}}{1+e^{a+bx}} \rightarrow (\text{sleep} = 1) = \frac{e^{-5.9836 + 2.6303 \cdot \text{bruxisme}}}{1+e^{-5.9836 + 2.6303 \cdot \text{bruxisme}}}$$

4.

Sandsynligheden for søvnforstyrrelser ved en sværhedsgrad af tandskæren på 2 udregnes på RSTUDIO til 0.32673 svarende til 32.67%:

$$(\text{sleep} = 1) = \frac{e^{-5.9836 + 2.6303 \cdot 2}}{1 + e^{-5.9836 + 2.6303 \cdot 2}} = 0.32673$$

5.

For at estimere risikorationen for søvnforstyrrelser med en bruxismegrad på 1 og 4 bestemmer man først den estimerede risiko for dette.

Estimeret risiko for søvnforstyrrelser med en sværhedsgrad af tandskæren på 4:

$$(\text{sleep} = 1) = \frac{e^{-5.9836 + 2.6303 \cdot 4}}{1 + e^{-5.9836 + 2.6303 \cdot 4}} = 0.9894142$$

Estimeret risiko for søvnforstyrrelser med en sværhedsgrad af tandskæren på 1:

$$(sleep = 1) = \frac{e^{-5.9836 + 2.6303 \cdot 1}}{1 + e^{-5.9836 + 2.6303 \cdot 1}} = 0.03378727$$

Risikoratioen bestemmes som forholdet mellem de estimerede risikoer:

$$\frac{0.9894142}{0.03378727} = 29.28364$$

Derfor er risikoen for søvnforstyrrelser med en bruxismegrad på 4 altså 29.28 gange risikoen for søvnforstyrrelser med en bruxismegrad på 1.

6.

e^b er en odds-ratio der beskriver effekten af den uafhængige variabel. Derfor bruges hældningskoefficienten i den logistisk regression til udregning af odds-ratio, som blev estimeret til $b=2.6303$. Jeg udregner Odds-ratioen til 13.88, som betyder, at hver gang tandskæringsgraden stiger med én, bliver odds for søvnforstyrrelser 13.88 gange større. Når sværhedsgraden af tandskæren stiger med 2 enheder vil ændring i odds for søvnforstyrrelser derfor være 192,6 gange større.

7.

For at udregne sværhedsgraden af tandskæring, hvor risikoen for søvnforstyrrelser først krydser 75%, løser jeg ligningen vha. Maple, som isolerer værdien af bruxisme:

$$0.75 = \frac{e^{-5.9836 + 2.6303 \cdot bruxisme}}{1 + e^{-5.9836 + 2.6303 \cdot bruxisme}}$$

$$bruxisme = 2.6925$$

Vi skal have en bruxisme på 2.69 (over 2) før at risikoen for søvnforstyrrelser krydser 75%.

8.

Intercept svarer til vores parameter $a=-5.983$, som angiver skæringspunktet for linjen på y-aksen på logOdds skalaen. Jeg finder intercept på Odds-skalaen ved at tage den eksponentielle funktion af parameteren a , hvor RStudio udregner det til 0.0025. Denne værdi vil angive skæringspunktet på Odds-skalaen. RStudio udregner også en nulhypotesen, som svarer til at være parametrene er lig

med 0, og der derfor ikke er sammenhæng mellem søvnforstyrrelser og bruxisme. P-værdierne er $p < 0,001$, og dermed kan vi forkaste nulhypoteserne og konkludere at gennemsnitlige odds for søvnforstyrrelser er forskellig fra 0, når bruxisme er lig med 0.

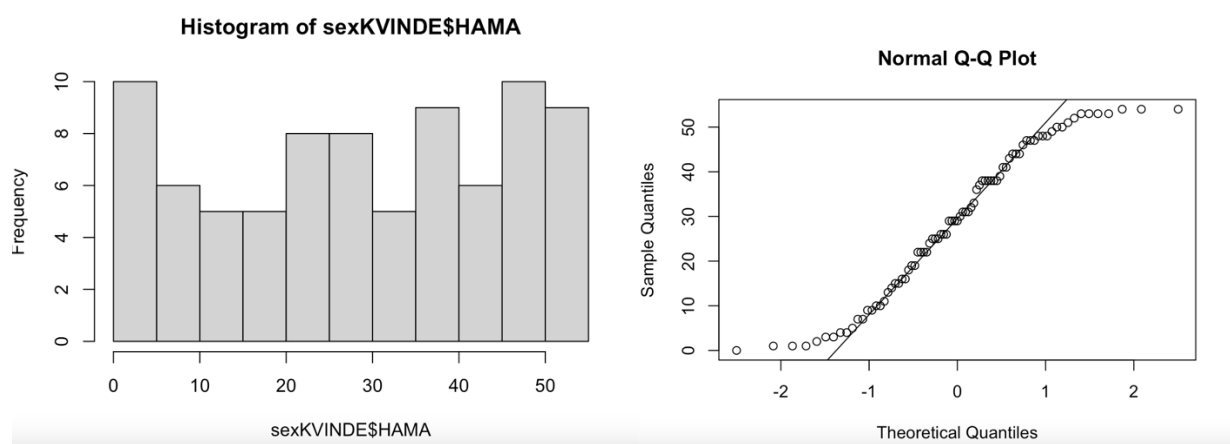
9.

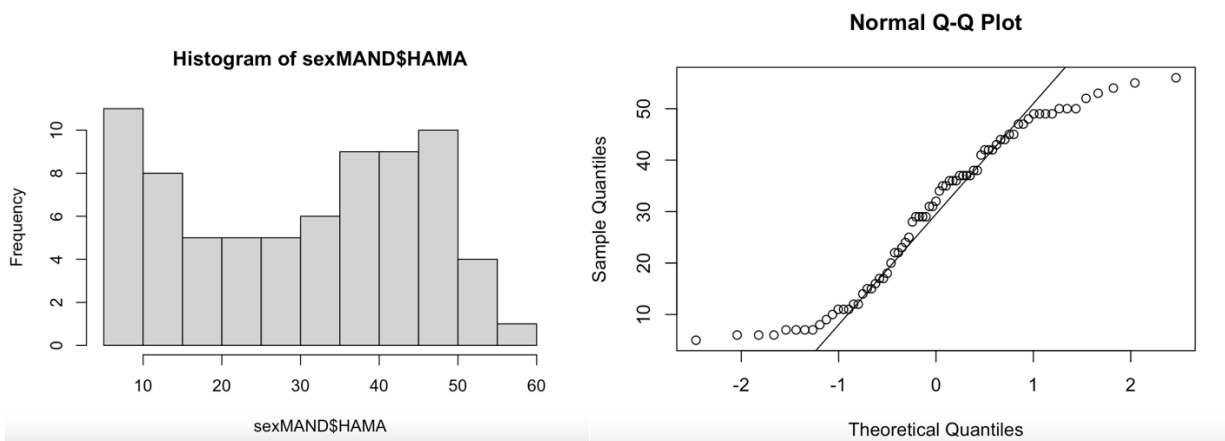
Jeg kan også indsætte 0 ind i den estimerede modellen for at få odds for søvnforstyrrelser ved en tandskæringsgrad på 0 og får resultatet 0.0025. Det stemmer overens med overstående opg, hvor jeg udregnede skæringspunktet på y-aksen på Odds-skalaen, der ligeledes svarer til en tandskæringsgrad på 0, som RStudio udregnede til 0.0025. Dette stemmer overens med den estimerede model.

4) Sammenhæng mellem angst og køn

1.

I denne opgave er den afhængige variabel angstniveau, som er en numerisk variabel, og den uafhængige variabel er køn, som er en kategorisk binær variabel. Jeg vælger derfor, at lave en t-test for to populationer som den statistiske model. Da samme person ikke er tilstede i begge grupper er dataet uparret, og jeg udfører derfor en uparret t-test. Det er et krav, at observationerne er tilnærmelsesvis normalfordelte for denne anvendte statistiske model. Der udføres derfor et histogram og QQ-plot for hver gruppe for at afgøre dette:





Jeg konkludere at histogrammerne er ikke symmetriske, hvilket skyldes, at vi tager udgangspunkt i en lille stikprøve. Derfor understøttes det med QQ-plots, hvor data er symmetriske. Det konkludere på baggrund af dette, at dataet følger tilnærmelsesvis en normalfordeling, og jeg kan derfor stole på konklusionerne fra t-testet.

Efter at have udført en uparret t-test kan vi aflæse, at den gennemsnitlige angstniveau for kvinder er 29.07407 (29), hvor den for mænd ligger på 30.17808 (30). Jeg udregner den gennemsnitlige forskel i angstniveau for mænd og kvinder til at være 1.10401. Det tilhørende 95% konfidensinterval aflæses til $[-6.27; 4.065]$. Jeg er 95% sikker på, at intervallet fra -6.27; 4.065 indeholder den sande forskel i den gennemsnitlige angstniveau blandt mænd og kvinder.

2.

Rstudio tester altid en nulhypotese, hvor parametrene er lig med 0. Nulhypotesen for den uparret t-test er, at de to middelværdier er i populationerne (mand og kvinde) er ens. Ved at aflæse p-værdien til 0.6736, som er en p-værdi på over 5% og vi kan derfor *ikke* forkaste nulhypotesen om at der er en signifikant forskel i angstniveau mellem køn, hvilket betyder, at der er ingen evidens for om køn har indvirkning på angstniveau.

5) Sammenhæng mellem søvnforstyrrelser og alkoholindtag

1.

I denne opgave er den afhængige variabel søvnforstyrrelser, som er en kategorisk binær variabel, og den uafhængige variabel er alkohol, som er en kategorisk binær variabel. Derfor vælger jeg at lave χ^2 -test som den statistiske model

2.

Jeg aflæser fra min udførte tabel 1, hvor jeg efterfølgende kan opstille 2x2-optællingstabel, der angiver frekvenserne for de observerede data som jeg tager udgangspunkt i:

	Personer UNDEN søvnforstyrrelser (NOsleep)	Personer MED søvnforstyrrelser (YESsleep)
Alkohol = 0 (nej)	72	51
Alkohol = 1 (ja)	16	15

Nulhypotesen for χ^2 -testen er, at fordelingen af søvnforstyrrelser er ens uafhængigt om man drikker alkohol eller ej og derfor har alkohol ikke en signifikant indvirkning på søvn. Efter udførslen af testen, aflæser jeg p-værdien for nulhypotesen til at være $p = 0.6219$ og derfor over 5%. Derfor kan vi ikke forkaste nulhypotesen, hvilket betyder, at der er ingen evidens for om alkohol har indvirkning på søvnforstyrrelser. Der er altså samme risiko for søvnforstyrrelser hvis man indtager alkohol og modsat hvis man ikke indtager alkohol for at kunne sove.

3.

Der gælder, at en χ^2 -test altid skal suppleres med en relevant beskrivelse af effektstørrelsen til at beskrive forskellen i fordelingen af søvnforstyrrelser blandt dem, der drikker alkohol med dem, der ikke drikker alkohol.

Jeg udregner først den estimerede risiko for de to grupper, før jeg kan udregne den relative risiko.

Den estimerede risiko for søvnforstyrrelser blandt dem, der drikker alkohol:

$$risiko = \frac{15}{15 + 16} = 0.483871$$

Derfor er risikoen 48.39% for søvnforstyrrelser, når man drikker alkohol for at kunne sove.

Den estimerede risiko for søvnforstyrrelser blandt dem, der ikke drikker alkohol:

$$risiko = \frac{51}{51 + 72} = 0.4146341$$

Derfor er risikoen 41.46% for søvnforstyrrelser, når man ikke drikker alkohol for at kunne sove.

Dernæst kan jeg bestemme den relative risiko, når jeg har de estimerede risikoer, ved at udregne forholdet mellem risikoen for søvnforstyrrelser blandt dem, der drikker alkohol med dem, der ikke drikker alkohol.

$$\frac{0.4839}{0.4146} = 1.167149$$

Risikoen for søvnforstyrrelser for personer der drikker alkohol er 1.17 gange større risiko for søvnforstyrrelser blandt dem der *ikke* drikker alkohol.

4.

Jeg bestemmer først odds for hver gruppe blandt dem, der drikker alkohol med dem, der ikke drikker alkohol.

Odds for søvnforstyrrelser blandt dem drikker alkohol og har søvnforstyrrelser og dem der drikker alkohol og ikke har søvnforstyrrelser:

$$\frac{15}{16} = 0.938$$

Odds for søvnforstyrrelser blandt dem der ikke indtager alkohol bestemmes ved at udregne forholdet mellem dem der ikke drikker alkohol og har søvnforstyrrelser og dem der ikke drikker alkohol og ikke har søvnforstyrrelser:

$$\frac{51}{72} = 0.7083$$

Odds-ratioen bestemmer jeg som forholdet mellem de to odds:

$$\frac{0.938}{0.7083} = 1.32$$

Derfor er odds for søvnforstyrrelser for dem der indtager alkohol 1.32 gange odds for søvnforstyrrelser blandt de der ikke drikker alkohol.

6) Sammenhæng mellem tandskæren, angst og søvnforstyrrelser

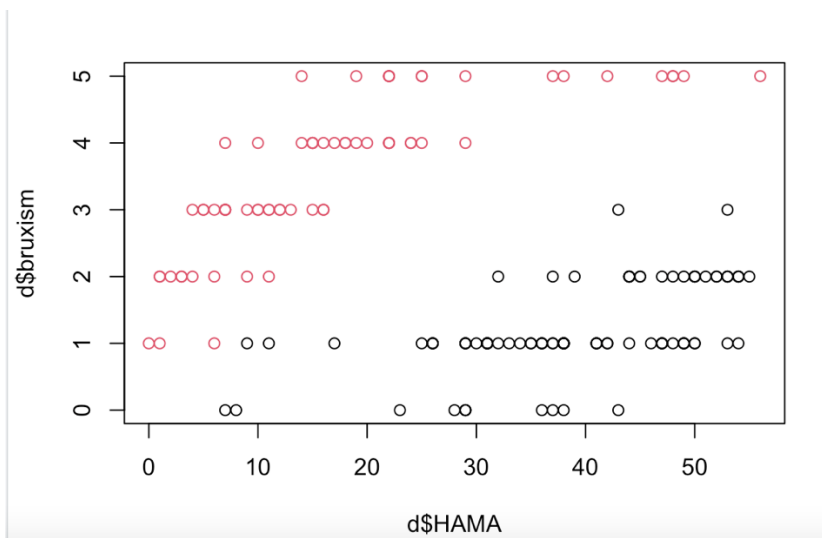
1.

I denne opgave er den afhænge variabel tandskæren, som er en numerisk variabel, og at denne variabel afhænger af både angst og søvnforstyrrelser, som er en numerisk og kategorisk binær variabel som begge er en uafhængig variabel. Derfor vælger jeg en multipel lineær regression som den statistiske model. Man kan inddrage en kategorisk uafhængig variabel i en regressionsmodel; som en hovedeffekt eller både som en hovedeffekt og en interaktionseffekt.

Den kategoriske variabel "sleep" skal omdannes til talværdier enten værdien 0 eller 1, som i vores opgave allerede er tildelt:

$$1(\text{sleep} = \text{søvnforstyrrelser}) = \begin{cases} 1, & \text{hvis sleep} = \text{søvnforstyrrelser} \\ 0, & \text{hvis sleep} = \text{ingen søvnforstyrrelser} \end{cases}$$

For at visualisere data udfører jeg et scatter plot af tandskæren med angst og farvelægger punkterne afhængig af, om personerne har søvnforstyrrelser eller ej. Her er de røde punkter dem der har søvnforstyrrelser modsat er de sorte punkter, som er personer uden søvnforstyrrelser:



Vi ønsker at udføre den mest simple statistisk model, som kan forklare vores data bedst muligt. Jeg starter derfor med den mest komplekse model og ser om denne skal simplificeres. Derfor starter jeg med at lave den mest komplicerede model, som er en multipel lineær regressionsmodel med både hovedeffekt og interaktionseffekt. En multipel lineær regressionsmodel for de observerede data med både hovedeffekt og interaktionseffekt er givet ved følgende ligning, hvor x_1 er numerisk og x_2 er kategorisk binær:

$$y_i = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + b_3 \cdot x_{2i} \cdot x_{1i} + e_i$$

Jeg indsætter vores variabler ind i modellen der beskriver sammenhængen af tandskæren mellem angst "HAMA" og søvnforstyrrelser "sleep":

$$tandskæren = a + b_1 \cdot HAMA_i + b_2 \cdot 1(\text{sleep}_i = \text{søvnforstyrrelser}) + b_3 \cdot 1(\text{sleep}_i = \text{søvnforstyrrelser}) + e_i$$

Jeg vil nu skrive modellen op i to separate tilfælde:

$$Tandskæren = \begin{cases} a + b_2 + (b_1 + b_3) \cdot HAMA_i + e_i, & \text{hvis sleep} = \text{søvnforstyrrelser} \\ a + b_1 \cdot HAMA_i + e_i, & \text{hvis sleep} = \text{ikke har søvnforstyrrelser} \end{cases}$$

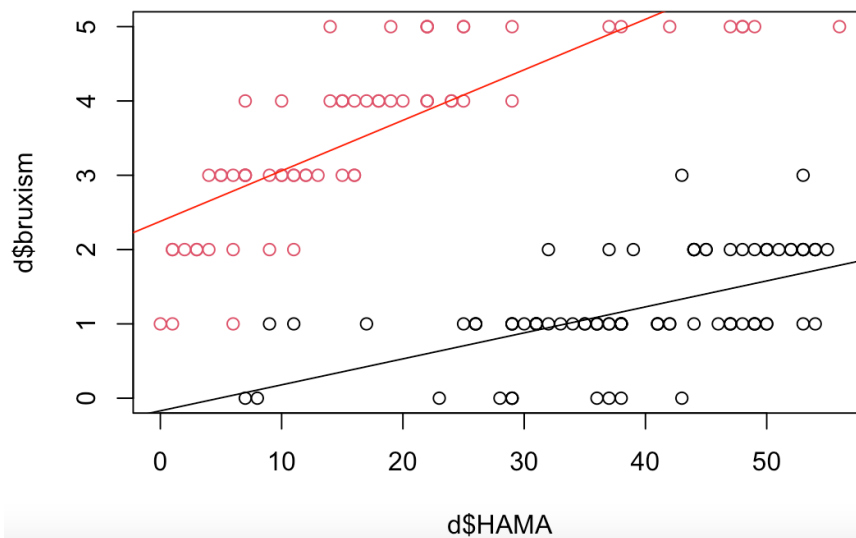
Jeg kan nu indsætte de estimerede parametre a , b_1 , b_2 og b_3 i den multiple regressionsmodel med både hovedeffekt og interaktionseffekt og beskrives ved:

$$Tandskæren = \begin{cases} -0.17 + 2.55 + (0.035 + 0.033) \cdot HAMA_i, & \text{hvis sleep} = \text{søvnforstyrrelser} \\ -0.17 + 0.035 \cdot HAMA_i, & \text{hvis sleep} = \text{ikke har søvnforstyrrelser} \end{cases}$$

Overstående forkortes til:

$$Tandskæren = \begin{cases} 2.38 + 0.068 \cdot HAMA_i, & \text{hvis sleep} = \text{søvnforstyrrelser} \\ -0.17 + 0.035 \cdot HAMA_i, & \text{hvis sleep} = \text{ikke har søvnforstyrrelser} \end{cases}$$

Hovedeffekten er her beskrevet ved $b_2 = 2.55$, som beskriver forskellen i skæringspunktet på y-aksen. For personer med søvnforstyrrelser er den 2.55 højere end skæringspunktet for linjen for personer uden søvnforstyrrelser. Interaktionseffekten er beskrevet ved $b_3=0.033$, hvor personer med søvnforstyrrelser har en hældningskoefficient der er 0.033 gange større end dem uden søvnproblemer. Den røde linje er et udtryk for personer med søvnforstyrrelser, og den sorte linje er et udtryk for ingen søvnforstyrrelser.



For at afgøre om vi skal bruge den mest komplicerede model med både hovedeffekt og interaktionseffekt ift vores data, tager jeg udgangspunkt i p-værdien for interaktionseffekten. Her er vil nulhypotesen være, at interaktionseffekten $B_3 = 0$, som siger at de to linjer er parallelle. Da vores p-værdi er $p < 0,001$ dvs langt under 5% kan vi forkaste nulhypotesen om at linjerne er parallelle. Det betyder, at der er forskel i hældningskoefficienten og der er derfor statistisk evidens for en sammenhæng mellem graden af tandskæren og graden af angst, når man sammenligner dem, der har søvnproblemer, med dem der ikke har søvnproblemer. Derfor kan modellen ikke simplificeres, når vi har empirisk evidens for at benytte den mest komplicerede model.

2.

Den estimerede sammenhæng blev defineret i overstående opgave, som tager udgangspunkt i den mest kompliceret model:

$$Tandskæren = \begin{cases} 2.38 + 0.068 \cdot HAMA_i, & \text{hvis sleep} = \text{søvnforstyrrelser} \\ -0.17 + 0.035 \cdot HAMA_i, & \text{hvis sleep} = \text{ikke har søvnforstyrrelser} \end{cases}$$

Modellens parameter b svarer til hældningskoefficienten, og kan fortolkes som graden af tandskæren som ændrer sig med 0.068 for dem med søvnforstyrrelser for hver gang man stiger med én på angstskaalen. Hvorimod dem der ikke har søvnforstyrrelser, ændres graden af tandskæren med 0.035 for hver gang man stiger med én på angstskaalen. Det betyder, at når man har søvnforstyrrelser, så har man generelt en større tendens til tandskæren. Parameteren a svarer til skæringspunktet på y-aksen og fortæller, at den gennemsnitlige bruxisme grad for personer med søvnforstyrrelser ligger på 2 (2.38) på bruxismeskalaen, for personer uden angst (svarer til 0 på x-aksen). For personer uden søvnforstyrrelser har man gennemsnitligt en bruxisme grad på 0 (-0.17), når man ikke har angst (svare til 0 på x-aksen).

3.

Jeg finder 95% konfidensintervallet via RStudio. Her får jeg 4 intervaller, hvor jeg tager udgangspunkt i konfidensintervallet angstgrad, fordi interaktionsleddet ikke inddrages her. I dette interval tages der udgangspunkt i personer der ikke har søvnforstyrrelser. Det tilhørende 95% konfidensinterval aflæses til [0.0234; 0.047]. Jeg er 95% sikker på, at dette interval indeholder den sande værdi for sammenhængen mellem graden af tandskæren og graden af angst blandt de personer, der ikke har søvnforstyrrelser.

4.

I denne opgave anvender jeg 95% konfidensintervallet for angstgrad med søvnforstyrrelser, da der tages hensyn til et interaktionsled. I dette interval tages der udgangspunkt i personer med søvnforstyrrelser. Det tilhørende 95% konfidensinterval aflæses til [0.017; 0.05]. Jeg er 95% sikker på, at dette interval indeholder den sande værdi for sammenhængen mellem graden af tandskæren og graden af angst blandt de personer, der har søvnforstyrrelser med dem der ikke har søvnforstyrrelse.

5.

Hovedeffekten er den effekt der tillægges referencegruppen, som er gruppen uden søvnproblemer. Derfor svarer den til forskellen mellem de to værdier, som er angivet til $b_2=2.55$. Her vil nulhypotesen teste om hovedeffekten er lig med 0, hvilket vil betyde at begge linjer har samme skæringspunkt på y-aksen, og derfor kan simplificeres til en simpel lineær regression. Jeg aflæser p-værdien til $p<0,001$, og derfor forkaster jeg nulhypotesen. Det betyder, at der er signifikant forskel i hovedeffekt, og linjerne derfor har forskellig skæring på y-aksen. Jeg konkludere derfor at graden af bruxisme har betydning for, om man har søvnforstyrrelser eller ej.

6.

I denne opgave blev der undersøgt, hvordan graden af tandskæren afhænger af både graden af angst og søvnforstyrrelser. Der blev i denne forbindelse udført en multipel lineær regression. I dette datasæt er det den mest komplicerede model med både hovedeffekt og interaktionseffekt, som er ideel til at repræsenterer vores data. Dette blev der fundet frem til, fordi vi forkastede vores nulhypotese om at interaktionseffekten $B_3 = 0$, som siger at de to linjer er parallelle. Her aflæste jeg p-værdien til $p<0.001$, hvilket betyder at der er statistisk evidens for en sammenhæng mellem graden af tandskæren og graden af angst, når man sammenligner dem, der har søvnproblemer, med dem der ikke har søvnproblemer. Personer med søvnforstyrrelser er placeret højere på bruxisme skalaen, hvilket også understøttes af hovedeffekten, hvor vi fik en værdi på 2.55. Denne værdi

beskrev, at skæringen på y-aksen 2.55 højere for personer med søvnforstyrrelser, hvorfor de har større tendens til tandskæren. Yderligere udregnede jeg, at graden af tandskæren ændrer sig med 0.068 for dem med søvnforstyrrelser for hver gang man stiger med én på angstskaalen. Hvorimod dem der ikke har søvnforstyrrelser, ændres graden af tandskæren med 0.035 for hver gang man stiger med én på angstskaalen. Ud fra dette kan jeg konkludere, at personer med søvnforstyrrelser har højere tendens til stige kraftigere på bruxisme-skalaen. Det ses dog, at personer uden søvnforstyrrelser generelt har en øget tendens til at placerer sig højt på angstskaalen, og samtidigt placerer de personer sig lavt på bruxisme-skalaen.

På baggrund af 95% konfidensintervallerne er jeg 95% sikker på, at intervallet $[0.0234; 0.047]$ indeholder den sande værdi for sammenhængen mellem graden af tandskæren og graden af angst blandt de personer, der ikke har søvnforstyrrelser. Her kan jeg se, at min hældningskoefficient på 0.035 indgår i dette interval. Ligeledes er jeg 95% sikker på, at intervallet $[0.017; 0.05]$ indeholder den sande værdi for sammenhængen mellem graden af tandskæren og graden af angst blandt de personer, der har søvnforstyrrelser med dem der ikke har søvnproblemer. Her kan jeg se, at interaktionseffekten på 0.033 indgår i dette interval, og dermed stemmer 95% konfidensintervallet overens med mit resultater.

7.

Når jeg kigger tilbage på opgave 2.6, undersøgte jeg om der er sammenhæng mellem graden af tandskæren og graden af angst. I opgaven testede RStudio en nulhypotese om at graden af bruxisme holdes konstant selvom graden af angst ændre sig det vil sige ingen sammenhæng mellem de pågældende parametre. Denne nulhypotese blev forkastet, da begge parametre har en $p < 0,001$. På baggrund af dette konkluderede jeg, at der er signifikant sammenhæng mellem graden af angst og bruxisme. Dog viste resultaterne at graden af bruxisme falder, når man stiger på angstskaalen. Dette skyldes, at vi ikke tog udgangspunkt i søvnforstyrrelser i opgave 2.6. I overstående opgave tager vi højde for søvnforstyrrelser, som er en ekstra faktor. Vores hypotese til dennes statistiske undersøgelse tager udgangspunkt i at undersøge, om personer med angst generelt har en øget risiko for tandskæren, men også at denne sammenhæng kan være påvirket af andre faktorer. Vi kan på baggrund af vores statistisk undersøgelse konkludere at, når der ikke tages højde for søvnforstyrrelser vil angst ikke medføre en øget tendens til tandskæren, med derimod ses en lavere tendens til tandskæren, når man stiger på angstskaalen. Når der inddrages en ekstra faktor, som i vores undersøgelse er søvnforstyrrelser, ser man derimod, at graden af tandskæren øges, når vi bevæger os højere op på angstskaalen. Dermed kan jeg konkludere på baggrund af dette, at der er sammenhæng mellem graden af tandskæren og graden af angst, når der tages højde for andre faktorer såsom søvnforstyrrelse.

Appendix

```
d <- read.csv("http://causal.sund.ku.dk/f23/15.csv", header=TRUE, stringsAsFactors=TRUE)
names(d)
```

```
NOSleep<-subset(d,sleep=="0")#ingen søvnforstyrrelser
YESsleep<-subset(d,sleep=="1")#med søvnforstyrrelser
hist(NOSleep$HAMA)#Histogram for angst ved NOSleep
qqnorm(NOSleep$HAMA)
qqline(NOSleep$HAMA)
```

```
mean(NOSleep$HAMA)#Middelværdien for HAMA ved NO
sd(NOSleep$HAMA)#Spredningen for HAMA ved NO
```

```
hist(YESsleep$HAMA)#Histogram for angst ved YESsleep
qqnorm(YESsleep$HAMA)
qqline(YESsleep$HAMA)
mean(YESsleep$HAMA)#Middelværdien for HAMA ved YES
sd(YESsleep$HAMA)#Spredningen for HAMA ved YES
```

```
hist(NOSleep$bruxism)#Histogram for bruxisme ved NOSleep
qqnorm(NOSleep$bruxism)
qqline(NOSleep$bruxism)
median(NOSleep$bruxism)#Medianen for Nosleep
IQR(NOSleep$bruxism)#IQR for NOSleep
```

```
hist(YESsleep$bruxism)#Histogram for bruxisme ved YESsleep
qqnorm(YESsleep$bruxism)
qqline(YESsleep$bruxism)
median(YESsleep$bruxism)#Medianen for YESsleep
IQR(YESsleep$bruxism)#IQR for YESsleep
```

```
table(NOSleep$alkohol)
(16/88)*100
(72/88)*100
```

```
table(YESsleep$alkohol)
(15/66)*100
(51/66)*100
```

```
table(NOSleep$sex)
(44/88)*100
```

```
table(YESsleep$sex)
(29/66)*100
(37/66)*100
```

```
#OPGAVE 2
m<-lm(d$bruxism~d$HAMA)
summary(m)
plot(d$HAMA,d$bruxism) #scatter plot
abline(2.786460, -0.019551)
confint(m) #95% konfidensinterval
2.786460 - 0.019551*20
```

```
2.39544+2*1.442
2.39544-2*1.442
```

```
#OPGAVE 3
model<-glm(d$sleep~d$bruxism,family=binomial)#Model på log Odds-skala
summary(model)
confint(model)
(exp(-5.9836 + 2.6303*2))/(1+exp(-5.9836 + 2.6303*2))
(exp(-5.9836 + 2.6303*4))/(1+exp(-5.9836 + 2.6303*4))
(exp(-5.9836 + 2.6303*1))/(1+exp(-5.9836 + 2.6303*1))
0.9894142/0.03378727
exp(2.6303)
13.87793^2
exp(-5.9836)
(exp(-5.9836 + 2.6303*0))/(1+exp(-5.9836 + 2.6303*0))
```

```
#OPGAVE 4
sexKVINDE <- subset(d,sex=="kvinde")
sexMAND <- subset(d,sex=="mand")
hist(sexKVINDE$HAMA)
qqnorm(sexKVINDE$HAMA)
qqline(sexKVINDE$HAMA)
hist(sexMAND$HAMA)
qqnorm(sexMAND$HAMA)
qqline(sexMAND$HAMA)
t.test(d$HAMA~d$sex)
30.17808-29.07407
```

```
#OPGAVE 5
table(d$salkohol,d$sleep)
```

```
ma<-matrix(c(72,16,51,15), nrow=2, ncol=2)
```

```
ma
```

```
chisq.test(ma)
```

```
15/(15+16)
```

```
51/(51+72)
```

```
0.4839/0.4146
```

```
15/16
```

```
51/72
```

```
0.938/0.7083
```

```
#OPGAVE 6
```

```
plot(d$HAMA, d$bruxism, col = factor(d$sleep))
```

```
m1<-lm(d$bruxism~d$HAMA*d$sleep)#Multipel lineær regressionsmodel med BÅDE  
hovedeffekt og interaktionseffekt
```

```
summary(m1)
```

```
-0.17+2.55
```

```
0.035+0.033
```

```
abline(2.38,0.068, col="red")
```

```
abline(-0.17,0.035)
```

```
confint(m1)
```