

Statistik for odontologer - Eksamensopgave 2024

25. marts, 2024

Opgave 1

1.1

```
d <- read.csv("http://causal.sund.ku.dk/f24/85.csv", header=TRUE, stringsAsFactors=TRUE)
```

1.2

Datasættet opdeles i en gruppe med mundtørhed (*Xerostomi, Ja*) og en gruppe uden mundtørhed (*Xerostomi, Nej*). En anden mulighed kunne være at opdele datasættet i dem som tager Sertralin, og dem som ikke gør. Begge opdelinger har samme formål - at give et modtageren bedre overblik over datasættet. Der undersøges herefter sammenhængen mellem mundtørhed og andre faktorer som alder, køn, brugen af det antidepressive middel Sertralin, og cariestilfælde. For at lave tabel 1, opstilles de forskellige variable, som vist nedenfor, med en forklaring på, hvilke typer variable der er tale om, for overskuelighedens skyld.

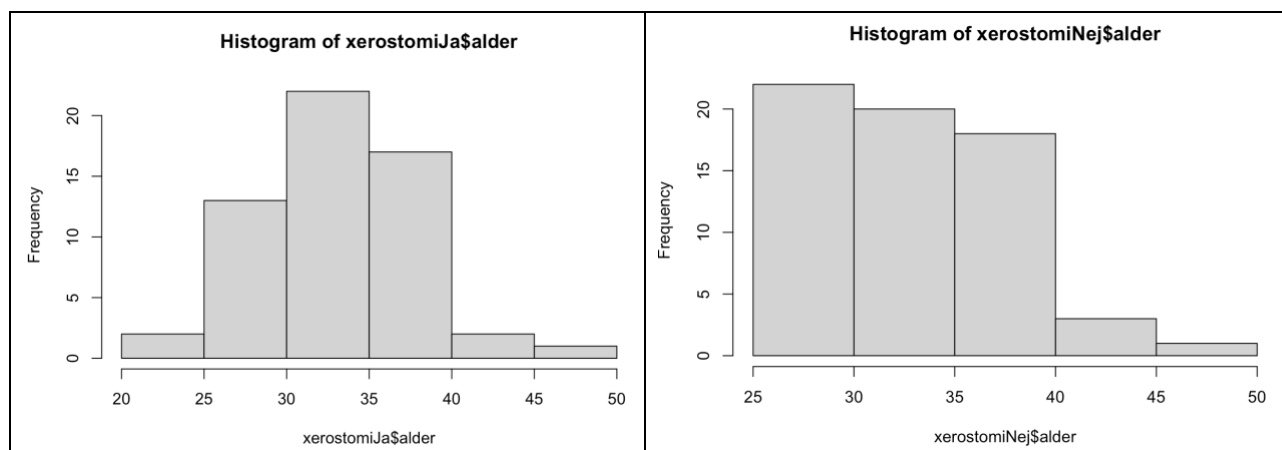
Sertralin: Om personen i datasættet tager Sertralin eller ej, er en binær kategorisk variabel, hvor 1 angiver, at personen tager Sertralin og 0 angiver, at de ikke tager stoffet. Denne variabel kan beskrives med hyppighed (frekvens), samt den relative frekvens.

Xerostomi: Hvorvidt personen lider af mundtørhed, er en ligeledes en binær kategorisk variabel, som beskrives med hyppighed og den relative frekvens.

Caries: Angiver om en person har et eller flere huller i tænderne, hvor 1: ja og 0: nej, hvorfor denne variabel også er en kategorisk binær variabel, der også kan beskrives med hyppigheden og den relative frekvens.

MDI: Major Depression Inventory, MDI, er en kategorisk variabel opdelt i grupperne normal depression, let depression, moderat depression og svær depression. Denne kategoriske variabel beskrives ligeledes ved hjælp af hyppigheden og den relative frekvens.

Alder: Personernes alder målt i år er en numerisk variabel. For at undersøge om personerne i datasættets alder er normalfordelt, plottes alder og xerostomi ind i et histogram (et for gruppen med xerostomi, og et for gruppen uden).



Ud fra ovenstående histogrammer ses det, at den gruppe med mundtørhed er nogenlunde normalfordelt med symmetri omkring midten og en karakteristisk "klokkeform". Den gruppe uden mundtørhed er derimod ikke normalfordelt, og den er venstreforskudt. Jeg vælger derfor at bruge median og IQR for at beskrive denne variabel, og det giver desuden modtageren et hurtigt og let overblik over datasættet.

Sex: Personens køn er en kategorisk binær variabel. Her beskrives det med hyppighed af det pågældende køn, samt den andel antallet udgør i procent ud af det samlede antal deltagere.

Tabel 1 for datasæt 85:

| | Xerostomi, Ja | Xerostomi, Nej |
|----------------------------|---------------|----------------|
| Antal deltagere (%) | 57 (47,11) | 64 (52,89) |
| Køn (%) | | |
| Mand | 25 (20,66) | 43 (35,54) |
| Kvinde | 32 (26,45) | 21 (17,36) |
| Alder, median (IQR) | 33,65 (7) | 33,29 (7,25) |
| MDI (%) | | |
| < 20: normal | 6 (10,53) | 21 (32,81) |
| 20-24: let | 7 (12,28) | 10 (15,63) |
| 25-29: moderat | 8 (14,03) | 6 (9,37) |
| > 29: svær | 36 (63,16) | 27 (42,19) |
| Sertralin | | |
| Ja | 51 (89,47) | 27 (42,19) |
| Nej | 6 (10,53) | 37 (57,81) |
| Caries | | |
| Ja | 42 (73,68) | 15 (23,44) |
| Nej | 15 (26,32) | 49 (76,56) |

Opgave 2

2.1

For at undersøge, hvordan graden af depression (MDI) afhænger af både alder og køn anvendes multipel lineær regression. Denne statistiske model er passende til at beskrive denne sammenhæng, idet den tillader analyse af flere uafhængige variable, *køn* (binær) og *alder* (kontinuerlig), på en kontinuerlig afhængig variabel, *MDI*. I modellen tages både hovedeffekt og interaktionsleddet i betragtning, for at undersøge om effekten er signifikant.

Ud fra outputtet i R-studio kan den estimerede sammenhæng opskrives som følger:

$$MDI = \begin{cases} 35,67 - 0,30 \cdot age & , \text{ hvis } k\ddot{o}n = \text{mand} \\ 28,84 + 0,09 \cdot age & , \text{ hvis } k\ddot{o}n = \text{kvinde} \end{cases}$$

For mænd er begyndelsesværdien, dvs. start MDI-scoren, i gennemsnit 35,67 til alderen 0 år. Denne værdi aftager med 0,30 for hvert år der går. For kvinder er begyndelsesværdien 28,84, med i modsætning til mændene falder denne MDI score/værdi ikke. Den stiger med 0,9 for hver enhedsændring i alderen for kvinderne.

De to modeller muliggør undersøgelse af, hvordan alder påvirker graden af depression (MDI) hos de to køn, således at man også kan undersøge eventuelle forskelle mellem de to køn.

2.2

Konfidensintervallet er en metode, der bruges til at begrænse en parameter. De parametre der findes, er *alder*, som er hældningen, og *graden af depression (DMI)*, som er skæring med y-aksen x.

95%-konfidensintervallet for a-værdien: Jeg er 95% sikker på, at intervallet [6,48 ; 51,20] indeholder den sande skæring med y-aksen i stikprøven. Skæring er i dette tilfælde et mål for graden af depression, MDI.

95%-konfidensintervallet for b-værdien:

Jeg er 95% sikker på, at den sande værdi af hældningskoefficienten, dvs. parameteren *alder*, ligger i intervallet [-0,55 ; 0,75].

2.3

Nulhypotesen er, at alder ikke har en signifikant betydning for graden af depression. Ud fra ovenstående test, som blev undersøgt i R-studio fra opgave 2.2, hvor 95%-konfidensintervallet blev udregnet, fik jeg en p-værdi på 0,015. Begge parametre har dermed en p-værdi, som er under 0,05 (5%), og hermed kan nulhypoteserne *forkastes*. Der er med andre ord signifikant sammengæng mellem de uafhængige variable (alder og køn) og den afhængige variabel (MDI).

Sammenhængen mellem graden af depression og alder, samt køn blev analyseret ved multipel lineær regression. Da p-værdien var under 0,05, var sammenhængen mellem disse parametre signifikant. For mændene var det sådan, at for hver enhedsændring i alder på 1 år, faldt den gennemsnitlige grad af depression med 0,30. For kvinderne steg den gennemsnitlige grad af depression med 0,09 for hvert ændring i år (95%-konfidensintervallet $[-0,55 ; 0,75]$). Ved alderen 0 år var den gennemsnitlige grad af depression 35,67 for mænd, og 28,84 for kvinder (95%-konfidensintervallet $[6,48 ; 51,20]$)

2.4

Hvis ikke man var interesseret i at undersøge, om køn også havde en effekt på graden af depression, men kun alder, kunne man have anvendt den simple lineære regressionsmodel til at beskrive sammenhængen mellem MDI og alder. I opgave 2.1 fastslog jeg, at modellen både inkluderede hovedeffekten og interaktionseffekten, for at jeg kunne undersøge om effekten var signifikant. Her var der altså både forskel på hældningskoefficienten, samt skæringspunkt med y-aksen mellem, hos de to modeller. En mere simpel model til at beskrive sammenhængen mellem depressionsscore, alder og køn, kunne være at man kun inddrogede hovedeffekten, dvs. alderen. Efter denne analyse i R-studio fås følgende model:

$$MDI = \begin{cases} 30,49 - 0,13 \cdot age & , \text{hvis køn} = \text{mand} \\ 36,66 - 0,13 \cdot age & , \text{hvis køn} = \text{kvinde} \end{cases}$$

Som forventet ses to ligninger med samme hældningskoefficient, dog med forskellige skæringspunkter med y-akser, de er med andre ord parallelle, men forskudte. For at opsummere, viser ovenstående regression, at den gennemsnitlige depressionsscore aftager med 0,13 for hvert år der går, og at kvinder i gennemsnit har en højere depressionsscore end mænd til alderen 0 år.

Derudover fås en p-værdi til 0,0079, som ligeledes er under 0,05, hvilket betyder at der igen er signifikant sammenhæng mellem MDI, alder og køn. P-værdier angiver signifikansniveauet for en given statistisk test, hvor en lavere p-værdi indikerer at resultatet er mere signifikant. P-værdien for denne model kun med hovedeffekt på 0,0079 er endnu lavere end en p-værdi på 0,015 fra resultatet i opgave 2.3, hvilket betyder, at resultatet er mere signifikant med kun en hovedeffekt frem for med hovedeffekt og interaktionsleddet.

Det er essentielt i statistisk, at man ønsker *den mest simple model, der forklarer data bedst muligt*, hvorfor den endelige model, der bedst beskriver sammenhængen mellem MDI, alder og køn er den multiple lineære regression kun med hovedeffekt.

2.5

Jeg anvender den multiple lineære regression med en hovedeffekt til at forudsæ den gennemsnitlige depressionsscore for en mand på 30 år.

$$MDI_{\text{hovedeffekt}} = 30,49 - 0,13 \cdot 30 = 26,59$$

For en 30-årig mand ses det, at den gennemsnitlige depressionsscore ifølge modellen er 26,59. Denne depressionsscore ligger inden for intervallet for en let depression (let depression: 25 – 29).

Sidebemærkning:

I tillæg testede jeg den model med både hovedeffekt og interaktionsleddet, for at se om forskellen var betydeligt.

$$MDI_{\text{hovedeffekt+interaktionseffekt}} = 35,67 - 0,30 \cdot 30 = 26,67$$

Det ses her, at det er en meget lille forskel på de to modeller. Dette stemmer også overens med p-værdierne, som begge lå en del under 0,05.

2.6

Referenceintervallet referer til det interval, der rummer en specifik procentdel af observationerne for en bestemt variabel i datasættet. Fra kommandoen udregnet i opgave 2.4, kan en *residual standard error* aflæses til 10,56 svarende til spredningen, og fra opgave 2.5 fandt jeg depressionsscoren blandt 30-årige mænd, som jeg nu bruger i denne opgave. Da referenceintervallet er defineret som $\pm 2 \cdot \text{spredningen}$, kan det udregnes og man får dermed et 95%-referenceinterval på [5,47 ; 47,71].

Det ovenstående 95%-referenceinterval [5,47 ; 47,71] viser det område, hvor man med 95% sikkerhed forventer, at den gennemsnitlige depressionsscore vil være for 30-årige mænd i dette datasæt. Dette interval strækker sig over alle fire grupper på MDI-skalaen, fra normal depression til svær depression.

2.7

Alder er her en afhængig kontinuert variabel, og køn er en binær variabel, hvorfor man kan anvende en uparret t-test, til at undersøge om den gennemsnitlige alder er den samme for mænd og for kvinder. Grunden til, at det er den uparrede t-test jeg vælger, er fordi den sammenligner gennemsnittet af to uafhængige stikprøver, og analyserer om der er signifikant forskel mellem dem. Her kan den samme person ikke optræde i begge grupper. Altså ønsker jeg at sammenligne den gennemsnitlige alder for mænd og for kvinder, og idet observationerne i de to grupper ikke er forbundede, er den uparrede t-test den mest passende model at anvende.

T-test analysen i R-studio giver os en gennemsnitsalder på 34,11 for gruppen med kvinderne og 32,96 for gruppen med mændene, men en gennemsnitlig forskel på 1,16 år.

Nulhypotesen er, at der ikke er forskel på alderen hos de to grupper, mænd og kvinder. P-værdien i t-testen er 0,1717, og da denne er over det typiske signifikansniveau på 0,5, kan nulhypotesen *ikke* forkastes. Med andre ord, er der ikke tilstrækkelig statistisk evidens for at afvise nulhypotesen, hvorfor der ikke er signifikant forskel på alderen hos mænd og kvinder baseret på denne statistiske analyse.

2.8

Baseret på analysen, hvor jeg undersøgte sammenhængen mellem graden af depression, alder og køn kan jeg konkludere følgende.

En multipel lineær regression blev udført for at undersøge, hvordan depressionsscoren afhænger af de uafhængige variable, køn og alder. Her var det både hovedeffekten og interaktionseffekten der blev medregnet. Resultaterne viste en signifikant effekt af alder og køn på depressionsscoren, med en p-værdi på 0,015 ($p < 0,05$). For hver stigning i alderen var der en estimeret ændring i depressionsscoren på $-0,30$ for mænd og $+0,09$ for kvinder, og et tilhørende 95%-konfidensinterval på $[-0,55 ; 0,75]$ for b-værdien, og et 95%-konfidensinterval på $[6,48 ; 51,20]$ for a-værdien.

For den multiple lineære regression kun med hovedeffekten, ses to ligninger med samme hældningskoefficient, og en p-værdi på 0,0079, som ligeledes er under 0,05, hvilket betyder at der igen er signifikant sammenhæng mellem MDI, alder og køn. Jo mindre p-værdien er, desto større evidens er der for, at nulhypotesen ikke stemmer overens med stikprøven, hvorfor den model, der bedst beskriver sammenhængen mellem MDI, alder og køn er den multiple lineære regression *kun* med hovedeffekt.

Yderligere foretog jeg en analyse med en uparret t-test for to grupper, hvor analysen i R-studio giver os en gennemsnitsalder på 34,11 kvinder og 32,96 for mænd, men en gennemsnitlig forskel på 1,16 år.

Det kan slutteligt konkluderes, at analyserne tyder på, at alderen har en signifikant påvirkning på depressionsscoren, mens der ikke er lige så stærk evidens til at påvise en signifikant forskel i depressionsscoren mellem mænd og kvinder ($p = 0,1717$). Det skal dog understreges, at der kan være behov for yderligere analyse for at kunne få en mere præcis forståelse for variablerne og deres indvirkning på hinanden.

Opgave 3

3.1

For at undersøge, hvordan sandsynligheden for at tage Sertralin afhænger af depressionsscoren anvendes logistisk regression. Depressionsscoren er en kategorisk uafhængig variabel, og sandsynligheden for at tage Sertralin er en binær afhængig variabel, hvorfor det er den logistiske regression der tages i brug.

3.2

Den logistiske model indsættes i R-studio ved brug af *glm-funktionen* og ud fra det outputtet indsættes værdierne, og herved fås følgende udtryk for den estimerede sammenhæng mellem indtagelse af Sertralin og graden af depression sandsynlighedsskalaen:

$$P_{\text{sertralin}} = \frac{e^{-6,67337+0,27717 \cdot MDI}}{1 + e^{-6,67337+0,27717 \cdot MDI}}$$

Dette er et udtryk for sandsynligheden for at tage Sertralin, som en funktion af depressionsscoren, hvor jeg har anvendt den naturlige logaritme til at omregne logaritmiske odds til sandsynligheder.

3.3

For at udregne odds-ratioen for effekten af depressionsscoren, anvendes oddsskalaen frem for sandsynlighedsskalaen, som vist i opgave 3.2.

$$\text{LogOdds}_{\text{sertralin}} = e^{-6,67337+0,27717 \cdot MDI}$$

Fra ovenstående modellen kan odds-ratioen for effekten af depressionsscoren beregnes, eftersom effekten af den uafhængige variabel (Sertralin) kan udtrykkes som en relativ forskel i odds, altså odds ratio, vha. e^b . Hvor b er koefficienten i den ovenstående model.

$$e^b = e^{0,27717} = 1,319391$$

Dets tilhørende 95%-konfidensinterval er [1,12 ; 1,48] og p-værdien for effekten af depressionsscoren på sandsynligheden for indtagelse af Sertralin er $3,01 \cdot 10^{-8}$. Nulhypotesen her er, at depressionsscoren ikke har nogen effekt på sandsynligheden for indtagelse af Sertralin. Det må jo betyde, at regressionskoefficienten er 0, og at odds-ratioen er 1. Med en p-værdi på $3,01 \cdot 10^{-8}$, kan denne nulhypotese forkastes, hvilket betyder, at depressionsscoren har en effekt på sandsynligheden for indtagelse af Sertralin.

Det kan konkluderes, at depressionsscoren har en statistisk signifikant effekt på sandsynligheden for indtagelse af Sertralin ($p = 3,01 \cdot 10^{-8}$ og $\text{odds} - \text{ratio} = 1,32$) med et konfidensinterval på

[1,12 ; 1,48]. Effekten af depressionsscoren er, at for hver gang depressionsscoren øges med 1, vil odds for indtagelse af Sertralin stige med 1,32 gange, da odds-ratio er en multiplikativ effekt.

3.4

For at prædiktere sandsynligheden for at tage Sertralin ved en depressionsscore på 26, indsættes værdien i ovenstående model fra opgave 3.2:

$$P_{\text{sertralin}} = \frac{e^{-6,67337+0,27717 \cdot 26}}{1 + e^{-6,67337+0,27717 \cdot 26}} \approx 0,6301942 = 63,02\%$$

Sandsynligheden for at tage Sertralin ved en depressionsscore på 26 er 63%.

3.5

Her beregnes sandsynligheden for at tage Sertralin ved en depressionsscore på 18, og den sammenlignes med den værdi jeg fik i opgave 3.4.

$$P_{\text{sertralin}} = \frac{e^{-6,67337+0,27717 \cdot 18}}{1 + e^{-6,67337+0,27717 \cdot 18}} \approx 0,1565256 = 15,65\%$$

Nu divideres sandsynligheden for at tage Sertralin ved en depressionsscore på 26 med sandsynligheden ved en depressionsscore på 18 for at finde risiko-ratioen. Der er 4,03 gange større risiko for at tage Sertralin hos en tilfældig person med en depressionsscore på 26, som hos en tilfældig person med en med en depressionsscore på 18.

3.6

I opgave 3.3 fandt jeg frem til, for hver gang depressionsscoren øges med 1, vil odds for indtagelse af Sertralin stige med 1,32 gange. Idet odds-ratio er en multiplikativ effekt estimeres odds til at ændre sig fra 1,32 til $1,32^5 \approx 4,00$. Der kan dermed konkluderes følgende - når depressionsrationen stiger med 5 enheder vil odds for indtagelse af Sertralin være 4,00.

3.7

Når risikoen for at tage Sertralin overstiger 80% er odds 4. Omregningen kan foretages vha. mellemregninger, som findes i kursusrummet under SAU 7.

Her sættes modellen fra opgave 3.2 lig med 4, og ligningen for x løses vha. et CAS-værktøj, for at prædiktere værdien af depressionsscoren, hvor risikoen for at tage Sertralin overstiger 80%.

$$4 = -6,67337 + 0,27717 \cdot x$$



Ligningen løses for x vha. WordMat.

$$x = 38,50839$$

Når risikoen for at tage Sertralin overstiger 80%, er værdien af depressionsscoren ifølge modellen 38,51.

3.8

Odds for at tage Sertralin ved en depressionsscore på nul kan beregnes vha. den logistiske regressionsmodel fra opgaven ovenfor.

$$Odds_{sertralin} = -6,67337 + 0,27717 \cdot 0 = -6,67337$$

Denne værdi svarer til skæring med y-aksen på vores regressionsmodel, altså *intercept*. Denne værdi kan fortolkes som odds for at indtage Sertralin, når depressionsscoren er lig 0, men værdien er negativ. Bemærk at odds er altid et positivt tal, hvorfor denne værdi i teorien ikke er korrekt. Dette viser en af ulemperne ved at prædiktere dato for ekstrapolerede x -værdier.

Opgave 4

4.1

Både brugen af Sertralin og risikoen for caries er binære kategoriske variable. For at undersøge om risikoen for caries afhænger af brugen af Sertralin anvendes χ^2 -testet. Denne test analyserer, hvorvidt der er en signifikant sammenhæng mellem de to kategoriske variable. χ^2 -testet generaliserer stikprøven til populationen, og herfra kan det vurderes om forskellen er signifikant.

Først laves en opsummeringstabel for at få et overblik over de kategoriske variable:

| | Sertralin, Ja | Sertralin, Nej |
|-------------------|---------------|----------------|
| Caries, Ja | 46 | 11 |
| Ingen caries, Nej | 32 | 32 |

Ud fra denne tabel bruges *matrix-funktionen* i R-studio til at udføre tabellen igen og efterfølgende definere den, hvorefter der nu kan laves χ^2 -testet. Nulhypotesen for χ^2 -testet er, at der ikke er nogen sammenhæng mellem de to variable - at brugen af Sertralin og risikoen for caries er uafhængige af hinanden. Efter udførelsen af testet fås en p-værdi på 0,000863, hvilket er mindre end signifikansniveauet på 0,05. Derfor kan nulhypotesen afvises, og jeg kan konkludere, at der er signifikant sammenhæng mellem brugen af Sertralin og risikoen for caries.

4.2

For at udregne den relative risiko (RR, også kaldet risikoratioen) og odds-ratioen (OR) for caries, når man sammenligner dem, der tager Sertralin med dem, der ikke gør bruges følgende formler:

$$RR = \frac{\text{Risiko for caries for SertralinJa}}{\text{Risiko for caries for SertralinNej}}$$

$$OR = \frac{\text{Odds for caries for SertralinJa}}{\text{Odds for caries for Sertralin Nej}}$$

For at estimere den relative risiko for caries blandt de to grupper, dem som tager Sertralin og dem som ikke gør, udregnes andelen i stikprøven, det vil sige de relative frekvenser. De udregnes og indsættes i ovenstående formel:

$$RR = \frac{\frac{46}{46+32}}{\frac{11}{11+32}} \approx 2,31$$

For at estimere odds-ratioen for caries i de to grupper, divideres dem som har caries med dem som ikke har caries i hhv. gruppen der tager Sertralin og gruppen som ikke gør, hvorefter udtrykket indsættes i formelen for OR:

$$OR = \frac{\frac{46}{32}}{\frac{11}{32}} \approx 4,11$$

Den relative risiko, RR, er større end 1, hvilket betyder, at dem der tager Sertralin har en højere risiko for caries sammenlignet med dem, som ikke tager Sertralin. De har 2,31 gange eller 131% højere risiko for caries.

Odds-ratioen, OR, er ligeledes større end 1, og dette indikerer også, at dem der tager Sertralin har større odds for caries sammenlignet med dem, som ikke gør. De har 4,11 gange eller 311% større odds for caries. Ved at sammenligne både den relative risiko og odds-ratioen, fås dermed et udvidet forståelse for risikoen for caries i de to grupper fra datasættet.

4.3

For at undersøge, om risikoen for mundtørhed (xerostomi) afhænger af brugen af Sertralin udføres samme test som i opgave 4.1, χ^2 -testet, idet begge variable er binære kategoriske variable. Denne test vælges når man undersøger sammenhængen mellem to kategoriske variable, og herfra kan man vurdere, om der er forskel i observationerne mellem de pågældende variable.

Her ses en opsummeringstabel for at få et overblik over de kategoriske variable:

| | Sertralin, Ja | Sertralin, Nej |
|-----------------|---------------|----------------|
| Mundtørhed, Ja | 51 | 6 |
| Mundtørhed, Nej | 27 | 37 |

χ^2 -testet udføres i R-studio, præcis på samme måde som i opgave 4.1. Nulhypotesen for χ^2 -testet er, at der ikke er nogen sammenhæng mellem de to variable, det vil sige, at brugen af Sertralin og mundtørhed ikke afhænger af hinanden. Efter udførelsen af testet fås en p-værdi på $1,66 \cdot 10^{-7}$, hvilket er langt under signifikansniveauet på 0,05. Derfor kan nulhypotesen forkastes, og det kan konkluderes, at der er en statistisk signifikant sammenhæng mellem brugen af Sertralin og mundtørhed.

4.4

Risikodifferensen (RD) er forskellen i risiko blandt de to grupper, dem der tager Sertralin og dem som ikke tager Sertralin. For at udregne risikodifferensen, udregnes den relative frekvens for begge grupper, hvorefter værdierne trækkes fra hinanden:

$$RD = \frac{51}{(51 + 27)} - \frac{6}{(6 + 37)} \approx 0,51$$

Ud fra ovenstående udregning kan det konkluderes, at risikodifferensen for mundtørhed er positiv, hvilket betyder, at risikoen for mundtørhed er højere blandt dem der tager Sertralin. Helt præcis ses det, at risikodifferensen er 51 procentpoint, når man sammenligner dem, der tager Sertralin med dem, som ikke gør.

4.5

For at undersøge, om risikoen for caries afhænger af mundtørhed, benytter jeg igen χ^2 -testet, som fra opgaverne 4.1. og 4.3. Jeg gør brug af netop denne test fordi det er den mest simple test, når man ønsker at undersøge den statistiske sammenhæng mellem to binære kategoriske variable.

Opsummeringstabel for de to binære variable mundtørhed og caries:

| | Mundtørhed, Ja | Mundtørhed, Nej |
|-------------|----------------|-----------------|
| Caries, Ja | 42 | 15 |
| Caries, Nej | 15 | 49 |

Nulhypotesen for χ^2 -testet er, at der ikke er nogen signifikant sammenhæng mellem mundtørhed og caries. Efter udførelsen af χ^2 -testet fås en p-værdi på $9,6 \cdot 10^{-8}$. Denne p-værdi er igen langt under signifikansniveauet på 0,05, hvorfor nulhypotesen forkastes. Det kan derfor konkluderes, at der er en statistisk signifikant sammenhæng mellem mundtørhed og caries.

4.6

I denne opgave gentager jeg analysen fra opgave 4.1, hvor jeg undersøgte om risikoen for caries afhænger af brugen af Sertralin, som var den afhængige variabel. I denne opgave kigger jeg nu kun på de personer, der ikke lider af mundtørhed. Det vil sige, at udover de to binære variable, der var i opgave 4.1, er der nu en tredje faktor man skal tage højde for. I denne situation kan jeg derfor med fordel bruge den logistiske funktion, der tillader mig at analysere sammenhængen mellem en binær variabel (risikoen for caries) og to uafhængige variable (brugen af Sertralin og mundtørhed). Her foretrækker jeg den logistiske funktion frem for χ^2 -testet, da den giver mulighed for at undersøge cofaktorer.

Ved hjælp af *glm-funktionen* i R-studio fås en logistisk regression, som jeg kan analysere. I denne analyse fik jeg en p-værdi på 1,001. En p-værdi så høj som denne indikerer, at der ikke findes nogen betydelig forskel eller effekt i datasættet. Nulhypotesen i denne undersøgelse er, at der ikke er nogen sammenhæng mellem brugen af Sertralin og risikoen for caries blandt dem, der ikke lider af mundtørhed. Med denne høje p-værdi, er der ikke tilstrækkelig evidens for at forkaste nulhypotesen, og der er ikke bevis for at antage den alternative hypotese. Dette betyder, at der ikke er fundet en signifikant sammenhæng mellem indtagelse af Sertralin og risikoen for caries blandt dem uden mundtørhed.

Disse resultater kan sammenlignes med resultaterne fra opgave 4.1. I opgave 4.1 fandt jeg frem til, at der er signifikant sammenhæng mellem brugen af Sertralin og risikoen for caries gennem analyse med χ^2 -testet og samme svar blev etableret i opgave 4.2, men her tog jeg ikke gruppen uden mundtørhed med i betragtning. Det vil sige, at brugen af Sertralin ikke har nogen signifikant effekt på risikoen for caries hos personer uden mundtørhed.

4.7

Ved hjælp af χ^2 -testet i opgave 4.1 kom jeg frem til den konklusion, at der er statistisk signifikant sammenhæng mellem brugen af Sertralin og risikoen for caries ($p = 0,000863$). I opgave 4.2 konkluderede jeg, at den gruppe som indtog Sertralin har 2,31 gange (131%) højere risiko for caries, og 4,11 gange (311%) større odds for caries.

I opgave 4.3 fandt jeg efterfølgende ud af, at der er en statistisk signifikant sammenhæng mellem brugen af Sertralin og mundtørhed ($p = 1,66 \cdot 10^{-7}$). Ligeledes blev det fastslået i opgave 4.5, at der er bemærkelsesværdig sammenhæng mellem mundtørhed og caries ($p = 9,6 \cdot 10^{-8}$).

Disse konklusioner stemmer godt overens med teorien om, at de personer, der indtager stoffet Sertralin mod depression i højere grad dør med mundtørhed, som er den faktor, der rent faktisk bidrager til udvikling af caries - og at det ikke er stoffet Sertralin der *i sig selv* giver caries. Så selvom indtagelse af Sertralin ikke *direkte* medfører caries, så kom jeg i denne opgave frem til, at det *indirekte* kan medføre caries, idet en hyppig bivirkning af stoffet er mundtørhed.

Appendix

OPGAVE 1

OPGAVE 1.1

```
d <- read.csv("http://causal.sund.ku.dk/f24/85.csv", header=TRUE, stringsAsFactors=TRUE)
```

OPGAVE 1.2

```
xerostomiJa <- subset(d, xerostomi=="1")
```

```
xerostomiNej <- subset(d, xerostomi=="0")
```

```
57/121*100
```

```
64/121*100
```

```
table(xerostomiJa$sex)
```

```
table(xerostomiNej$sex)
```

```
25/121*100
```

```
32/121*100
```

```
43/121*100
```

```
21/121*100
```

```
hist(xerostomiJa$alder)
```

```
hist(xerostomiNej$alder)
```

```
mean(xerostomiJa$alder)
```

```
mean(xerostomiNej$alder)
```

```
IQR(xerostomiJa$alder)
```

```
IQR(xerostomiNej$alder)
```

```
table(xerostomiJa$MDI)
```

```
6/57*100
```

```
7/57*100
```

```
8/57*100
```

```
36/57*100
```

```
table(xerostomiNej$MDI)
```

```
21/64*100
```

```
10/64*100
```

```
6/64*100
```

```
27/64*100
```

```
table(xerostomiJa$sertralin)
```

```
table(xerostomiNej$sertralin)
```

```
51/57*100
```

```
6/57*100
```

```
27/64*100
```

```
37/64*100
```

```
table(xerostomiJa$caries)
42/57*100
15/57*100
table(xerostomiNej$caries)
15/64*100
49/64*100
```

OPGAVE 2

OPGAVE 2.1

```
Depression1<-lm(d$MDI~d$alder*d$sex)
```

```
summary(Depression1)
```

```
28.84+6.83
```

```
0.09-0.39
```

OPGAVE 2.2

```
confint(Depression1)
```

OPGAVE 2.3

Ingen tilhørende kommando.

OPGAVE 2.4

```
Depression2<-lm(d$MDI~d$alder+d$sex)
```

```
summary(Depression2)
```

```
36.6606-6.1745
```

OPGAVE 2.5

```
30.49-0.13*30
```

OPGAVE 2.6

```
26.59-2*10.56
```

```
26.59+2*10.56
```

OPGAVE 2.7

```
t.test(d$alder~d$sex)
```

```
34.11321-32.95588
```

OPGAVE 2.8

Ingen tilhørende kommando.

OPGAVE 3**OPGAVE 3.1**

Ingen tilhørende kommando.

OPGAVE 3.2

```
Depression3<-glm(d$sertralin~d$MDI, family=binomial)
summary(Depression3)
```

OPGAVE 3.3

```
exp(0.27717)
confint(Depression3)
exp(0.1909797)
exp(0.3894747)
```

OPGAVE 3.4

```
(exp(-6.67337+0.27717*26)/(1+exp(-6.67337+0.27717*26)))
```

OPGAVE 3.5

```
(exp(-6.67337+0.27717*18)/(1+exp(-6.67337+0.27717*18)))
0.6301942/0.1565256
```

OPGAVE 3.6

```
1.32^5
```

OPGAVE 3.7

Ingen tilhørende kommando.

OPGAVE 3.8

Ingen tilhørende kommando.

OPGAVE 4**OPGAVE 4.1**

```

sertralinJA<-subset(d, sertralin=="1")
sertralinNej<-subset(d, sertralin=="0")
table(sertralinJA$caries)
table(sertralinNej$caries)
d4 <- matrix(c(46, 32, 11, 32), nrow = 2, ncol = 2)
d4
chisq.test(d4)

```

OPGAVE 4.2

```

(46/(46+32))/(11/(11+32))
(46/32)/(11/32)

```

OPGAVE 4.3

```

table(sertralinJA$xerostomi)
table(sertralinNej$xerostomi)
d43 <- matrix(c(51, 27, 6, 37), nrow = 2, ncol = 2)
d43
chisq.test(d43)

```

OPGAVE 4.4

```

(51/(51+27))-(6/(6+37))

```

OPGAVE 4.5

```

table(xerostomiJa$caries)
table(xerostomiNej$caries)
d45 <- matrix(c(42, 15, 15, 49), nrow = 2, ncol = 2)
d45
chisq.test(d45)

```

OPGAVE 4.6

```

m46<-glm(caries~sertralin, data=xerostomiNej, family=binomial)
summary(m46)
exp(0.00126)

```

OPGAVE 4.7

Ingen tilhørende kommando.