EKSAMEN I STATISTIK

Eksamensnummer: 84

Opgave 1. Datapræsentation

1.1

d <- read.csv("http://causal.sund.ku.dk/f24/84.csv", header=TRUE, stringsAsFactors=TRUE)

1.2

I denne tabel 1 ønskes det at opstille data ud fra om deltageren indtager sertralin eller ej, idet opgaven tager udgangspunkt i dette medikament i forhold til at det kan give mundtørhed (xerostomi). Jeg synes derfor det kunne være oplagt at opdele tabellen i de to grupper; *"YesS" (sertralin)* og *"NoS" (ingen sertralin).* Herefter bør det for læseren fremstå mere overskueligt at sammenligne med andre variable, hvilket skaber et større overblik over personerne i stikprøven. For dette datasæt er der foretaget 147 observationer hos personer, som har til fordel at undersøge sammenhængen mellem Sertralin, mundtørhed og risiko for caries. I sættet ses 6 variable;

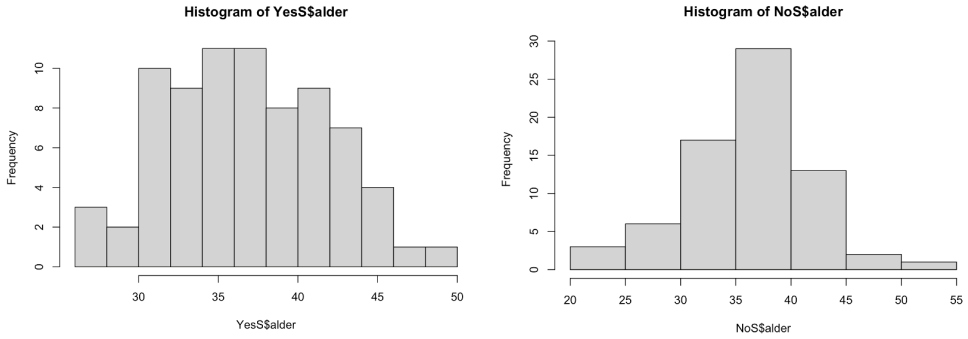
- Alder; som er målt på en diskret kontinuert intervalskala.
- MDI; som er målt på en kategorisk rangskala.
- Sex; som er målt på en kategorisk binomialskala.
- Caries; som er målt på en kategorisk binomialskala.
- Sertralin; som er målt på en kategorisk binomialskala.
- Xerostomi; som er målt på en kategorisk binomialskala.

Sertralin

Binær kategorisk variabel. Variablen fortæller om personen tager sertralin (1: ja) eller ikke (0: nej). Sertralin er et antidepressivt middel. Dette er en spændende variabel at opdele stikprøven i, hvilket specielt ses under MDI variablen, hvor store forskelle ses.

Alder

Variablen alder er numerisk og det undersøges derfor først om hvorvidt variablen er normalfordelt eller ej. Dette gøres ved at lave et histogram og/eller et QQ-plot. Histogrammet til venstre viser fordelingen af alder af personer med forbrug af sertralin, og histogrammet til højre viser fordelingen af alder af personer uden forbrug af sertralin. Sidstnævnte histogram følger nogenlunde den karakteristiske klokkeform, hvorfor data må siges at være nogenlunde normalfordelt. Histogrammet for personer med forbrug af sertralin er dog ikke normalfordelt. Jeg har valgt at bruge IQR og median, idet jeg synes, at det giver et mere overskueligt overblik for læseren.



MDI (Major depression inventory)
MDI er en kategorisk variabel, hvorfor den i tabellen angives med frekvens samt den relative frekvens for personer med angst opdelt i grupperne normal (< 20), let depression (20-24), moderat depression (25-29) og svær depression (> 29) i de to grupper (xerostomi/no xerostomi).

Sex
Binær kategorisk variabel. Ved brug af en opsummeringstabel i R findes frekvensen (N) samt den relative frekvens (%).

Caries
Binær kategorisk variabel. Her viser variablen om hvorvidt personen har haft caries i tænderne (1: ja) eller ingen caries (0: nej). Her anvender jeg ligeledes frekvensen og den relative frekvens, hvilket indsættes i tabel 1.

Xerostomi
Binær kategorisk variabel. Variablen xerostomi omhandler hvorvidt personen lider af mundtørhed (1: ja) eller ikke (0: nej), hvor der i tabellen indsættes frekvens samt den relative frekvens.

Tabel 1.

	Sertralin	**No sertralin**
No. of participants, N (%)	76 (51,7)	71 (48,3)
Sex, N (%) Male Female	 39 (51,32) 37 (48,68)	 28 (39,44) 43 (60,56)
Age, median (IQR)	37,39 (7)	36,83 (7)
MDI, N (%) < 20: normal 20-24: let depression 25-29: moderat depression > 29: Svær depression	 3 (3,95) 14 (18,42) 19 (25,00) 40 (52,63)	 44 (61,97) 22 (30,99) 4 (5,63) 1 (1,41)
Caries, N (%) Ja Nej	 43 (56,58) 33 (43,42)	 11 (15,49) 60 (84,50)
Xerostomi, N (%) Ja Nej	 41 (53,95) 35 (46,05)	 14 (19,72) 57 (80,28)

Opgave 2. Sammenhæng mellem depression, alder og køn

2.1

Her anvender jeg multipel lineær regression, da jeg ved brug af denne analyse kan undersøge, hvordan en afhængig kontinuerlig variabel; *graden af depression (MDI)* afhænger af to uafhængige variable; *alder* og *køn*. I modellen inkluderer jeg både hovedeffekten og interaktionsleddet for at undersøge om disse har signifikant betydning. Såfremt interaktionsleddet ikke har signifikant effekt, da kan modellen reduceres til en model kun med hovedeffekt. De estimerede sammenhænge kan opskrives som følgende.

$$Graden\ af\ depression\ (MDI) = \begin{cases} 26{,}86 - 0{,}05 \cdot alder, & hvis\ sex = male \\ 14{,}64 + 0{,}20 \cdot alder, & hvis\ sex = female \end{cases}$$

Det fremgår af ovenstående model at modellen svarer til to rette linjer med hver deres skæringspunkt *(26,86 og 14,64)* samt hver deres hældning *(-0,05 og 0,20).* Hovedeffekten *(12,22)* og interaktionsleddet *(-0,25)* fra outputtet repræsenterer forskellene i henholdsvis skæringspunkt og hældning mellem de to linjer.

Opsummerende betyder dette, at til alderen 0 er den gennemsnitlige grad af depression (MDI) for mænd 12,22 større end den gennemsnitlige grad af depression (MDI) for kvinder til alderen 0. For mænd falder depressionsgraden 0,05 hvert år, og for kvinder stiger den gennemsnitlige grad af depression med 0,20 hvert år.

2.2
95% konfidensinterval præsenterer intervallet, hvor den sande værdi af en parameter estimeres med 95% sikkerhed. 95% konfidensintervallet for henholdsvis *a* og *b* er *[2.01;27.26]* og *[-0.14;0.53]*.

Jeg er 95% sikker på, at intervallet fra [2.01 ; 27.26] indeholder den sande skæring med y-aksen i populationen for graden af depression (DMI) samt at intervallet [-0.14 ; 0.53] indeholder den sande værdi for hældningskoefficienten i populationen for alderen.

2.3
Nulhypotesen for sammenhæng af alder på depressionsscoren er, at der ikke ses nogen sammenhæng mellem de to parametre, hvorimod den alternative hypotese er at der ses sammenhæng. P-værdien, som er tilknyttet alderskoefficienten, er *0,2511*. Eftersom p-værdien er større end mit signifikansniveau *($p < 0,05\%$),* accepterer jeg nulhypotesen og jeg kan derfor ikke afvise, at der ikke ses nogen effekt af alder på depressionsscoren.

2.4
Ud fra outputtet undersøges p-værdierne, hvoraf p-værdien for hovedeffekten er *0,2648* og p-værdien for interaktionsleddet er *0,3973*, hvor begge p-værdier er større end 5%. Dermed kan nulhypotesen, om at interaktionsleddets skæring med y-aksen er lig med 0 samt at hovedeffektens hældning er lig med 0, ikke forkastes.

Derfor undersøger jeg nu p-værdierne for en model uden interaktionsled. P-værdien for hovedeffekten er nu *0,035991*, hvilket er mindre end mit signifikansniveau *($p < 0,05\%$),* hvorved nulhypotesen om at hovedeffektens hældning er lig med 0 *forkastes*. Der ses derfor en signifikant statistisk sammenhæng, hvorfor jeg vælger at anvende den multiple lineære regression med en hovedeffekt.

2.5
Jeg opskriver nu den multiple lineære model med en hovedeffekt ud fra estimaterne som følgende:

$$Graden\ af\ depression\ (MDI) = \begin{cases} 20{,}80 + 0{,}11 \cdot alder, & hvis\ sex = male \\ 17{,}77 + 0{,}11 \cdot alder, & hvis\ sex = female \end{cases}$$

Jeg ønsker at forudsige den gennemsnitlige depressionsscore (MDI) ud fra min model for en 30-årig mand som følgende: $Graden\ af\ depression\ (MDI) = 20{,}80 + 0{,}11 \cdot 30 = 24{,}1$.
Den gennemsnitlige depressionsscore for en 30-årig mand vil derfor være 24, som falder i kategorien, *let depression (20-24).*

2.6
Den gennemsnitlige depressionsscore for en 30-årig mand er 24,1 *(jævnfør 2.5).* Jeg kan derfor beregne et 95% referenceinterval for 30-årige mænd ved at aflæse estimatet af spredningen til *8,652 (jævnfør 2.4)* og indsætte i formlen: $95\%\ referenceinterval = middelværdi \pm 2 \cdot SD\ (spredning)$.

$$95\%\ referenceinterval = 24{,}10 + 2 \cdot 8{,}652 = 41{,}40$$
$$95\%\ referenceinterval = 24{,}10 - 2 \cdot 8{,}652 = 6{,}80$$

Jeg er dermed 95% sikker på, at intervallet [6.80 ; 41.40] indeholder den sande hældning for 30-årige mænd i populationen. Det kan derfor konkluderes, at 95% af 30-årige mænd vil have en depressionsscore mellem *7* og *41*, hvilket henholdsvist svarer til *normal* og *svær depression*.

2.7
Jeg ønsker, at undersøge om den gennemsnitlige alder er den samme for mænd og kvinder. Her udføres en uparret t-test, da jeg ønsker at undersøge to grupper, hvor de samme personer ikke indgår i begge grupper, hvilket ikke er tilfældet i den binære kategorisk variabel, *sex*. Den førnævnte variabel er den uafhængige variabel, hvor den afhængige variabel er den kontinuert numeriske variabel, *alder.*

Først laver jeg lidt deskriptiv statistik ved at opdele datasættet i to grupper (male og female) for at finde gennemsnittet af alderen i de to grupper. Gennemsnittet for kvinder i stikprøven er *37,05* år og gennemsnittet for mænd er *37,13* år. Jeg har også visualiseret data ved at lave et histogram for de to grupper, men det kan være svært at få et klart billede af dette ud fra sådan et histogram. Det antages dog, at data er normalfordelt, idet dette også er en af kriterierne for t-testen.

Ud fra den deskriptive statistik ses det, at der ikke er stor forskel på gennemsnitsalderen for kvinder og mænd. Dette er dog et usikkert resultat, hvorfor t-test selvfølgelig laves.

Ud fra analysen i R ses det, at gennemsnittene i de to stikprøver til svarer gennemsnittene som blev udregnet under den deskriptive statistik. Derudover tester den nulhypotesen, som i den uparrede t-test er, at der ikke er forskel på gennemsnitsalderen hos kvinder og mænd. Dette er ækvivalent med at sige, at forskellen i gennemsnitsalder hos mænd og kvinder er lig med nul. Ud fra p-værdien, som er *0,9205*, ses det at denne hypotese ikke kan forkastes. Dette betyder, at der ikke er signifikant forskel mellem gennemsnitsalderen hos kvinder og mænd i populationen.

Den gennemsnitlige alder hos kvinder er 37,05 år og 37,13 år hos mænd. En uparret t-test viste ikke en signifikant forskel i middelværdien (p = 0,9205), og forskellen var i gennemsnit på 0,08 år. Det kan dermed konkluderes, at den gennemsnitlige alder ikke er forskelligt på tværs af køn.

2.8

Sammenhængen mellem graden af depression (MDI) og alder under hensyntagen til køn blev undersøgt med multipel lineær regression.

Jeg har valgt at konkludere på den multiple lineære model med en hovedeffekt (fremfor den med hovedeffekt samt interaktionseffekt), da jeg kom frem til at denne var den mest simple statistiske model. For mænd steg graden af depression (MDI) i gennemsnit med *0,11* hver gang alderen steg med 1 og den forventede grad af depression til alderen 0 var *20,80 (let depression)*. For kvinder steg graden af depression (MDI) i gennemsnit med *0,11* hver gang alderen steg med 1 og den forventede grad af depression til alderen 0 var *17,77 (normal)*. Ved en alder på 0 var der en forskel i depressionsgrad (MDI) på *3,03* for mænd sammenlignet med kvinder. P-værdien for modellen var *0,035991*.

Yderligere kan det konkluderes, at den gennemsnitlige alder ikke er forskelligt på tværs af køn efter udførelse af t-test, idet der ikke ses en signifikant forskel i middelværdien *(p = 0,9205).* Gennemsnitsalderen for mænd i stikprøven var *37,13* år, hvor den for kvinder i stikprøven er *37,05* år.

Slutteligt kan det konkluderes, at analyserne tyder på at den uafhængige variabel, *alder*, har en signifikant indvirkning på graden af depression (MDI) ud fra modellen *kun* med hovedeffekten. Samtidig ses den uafhængige binære kategoriske variabel, *sex*, kan spille en rolle, idet mænd til alderen 0 har en højere grad af depression, MDI (3,03).

Opgave 3. Sammenhængen mellem depression og indtagelse af Sertralin

3.1

Til at undersøge hvordan sandsynligheden for at tage Sertralin afhænger af depressionsscoren (MDI). Her er den afhængige variabel en binær kategorisk variabel, *Sertralin*, og den uafhængige variabel en kategorisk variabel, *MDI*. Derfor vil jeg bruge logistisk regression til at belyse sammenhængen.

3.2

Udtrykket for den estimerede sammenhæng mellem indtagelse af Sertralin og graden af depression er opskrevet som følgende på sandsynlighedsskalaen ud fra analysen i R:

$$P(Sertralin = 1) = \frac{e^{-8{,}8770+0{,}3877 \cdot MDI}}{1 + e^{-8{,}8770+0{,}3877 \cdot MDI}}$$

3.3

Resultatet af analysen af den logistiske regressionsmodel kan opskrives som følgende:

$$logOdds\ (Sertralin) = -8{,}8770 + 0{,}3877 \cdot MDI$$

Effekten af den uafhængige variabel kan beskrives som en relativ forskel i odds, *en odds-ratio*, via e^b. Ud fra regressionsmodellen kan det ses, at odds-ratioen for at tage sertralin er givet ved $e^{0{,}3877} = 1{,}473588$.

95% konfidensintervallet for regressionskoefficienten, *b*, er [1.32 ; 1.71], som ovenstående odds-ratio også indgår i.

Nulhypotesen er at der ikke ses nogen effekt af *depressionsgraden, MDI* på forbruget af *Sertralin*. Dette betyder, at såfremt nulhypotesen gælder, da gælder følgende (regressionskoefficienten (b) = 0 og odds-ratio = 1.

P-værdien for effekten af depressionsscoren på sertralin forbrug er $3{,}61 \cdot 10^{-9}$, hvilket er en meget lille p-værdi. Den er også under signifikansniveauet *($p < 0{,}05$),* hvorfor nulhypotesen forkastes. Dette betyder, at depressionsgraden har en signifikant betydning for forbruget af det antidepressive middel, Sertralin.

Da 95% konfidensintervallet ikke omfatter værdien 1 (hvilket svarer til en odds-ratio, der indikerer ingen effekt af den uafhængige variabel), kan det konkluderes, at der er en effekt af depressionsgraden for brugen af Sertralin. Dette understøtter den konklusion, som jeg konkluderede ovenfor i forbindelse med p-værdien.

Sammenhængen mellem depressionsgraden, MDI og forbruget af Sertralin blev analyseret ved logistisk regression. Nulhypotesen forkastes, da p-værdien = $3{,}61 \cdot 10^{-9}$ *er mindre end 5%. Hermed kan det konkluderes, at graden af depression har en statistisk signifikant effekt (p =* $3{,}61 \cdot 10^{-9}$ *og odds-ratio =* $1{,}473588$*) på brug af sertralin. 95% konfidensinterval = [1.32 ; 1.71]. Opsummerende betyder dette, at for hver gang graden af depression øges med 1, bliver odds for brugen af sertralin 1,47 gange større.*

3.4

Jeg prædikterer sandsynligheden for at tage sertralin ved en depressionsscore på 26 ved at indsætte denne i den estimerede model for sandsynligheden:

$$P(Sertralin = 1) = \frac{e^{-8{,}8770+0{,}3877 \cdot 26}}{1 + e^{-8{,}8770+0{,}3877 \cdot 26}} = 0.7690936 = 76{,}91\%$$

Sandsynligheden *(Sertralin = 1| MDI = 26*) for brug af sertralin ved en depressionsgrad på 26 er dermed 76,91%.

3.5

Jeg ønsker at estimere den relative risiko (risikoratioen). Dette findes ved at beregne risikoen for forbrug af sertralin hos to personer, én med depressionsgrad på 26 og én med på depressionsgrad på 18.

$$Risikoratio = \frac{\frac{e^{-8{,}8770+0{,}3877 \cdot 26}}{1 + e^{-8{,}8770+0{,}3877 \cdot 26}}}{\frac{e^{-8{,}8770+0{,}3877 \cdot 18}}{1 + e^{-8{,}8770+0{,}3877 \cdot 18}}} = 5{,}90295$$

Risikoratioen beregnes til 5,90295. Dette betyder, at der er 5,9 gange større risiko for at anvende sertralin for en person med depressionsgrad på 26 sammenlignet med en person med depressionsgrad på 18.

3.6

I opgave 3.3 udregnede jeg odds-ratioen til $1{,}473588$. Dette svarer til den relative ændring i odds for at bruge det antidepressive sertralin, når graden af depression stiger med 1. Da odds-ratioen er en multiplikativ effekt, svarer en ændring i depressionsgrad på 5 enheder til en relativ ændring i odds på $1{,}4735588^5 \approx 6.947641$.

3.7

Jeg finder odds ved en sandsynlighed på *80%* ved at indsætte risikoen i nedenstående ligning (jævnfør arket for mellemregninger (SAU7):

$$Odds\ (Y) = \frac{P\ (Y = succes)}{1 - P(succes)} = \frac{0{,}80}{1 - 0{,}80} = 4$$

Ved en sandsynlighed på 80% er odds lig med 4, hvilket svarer til en logaritmisk værdi på 1.386294. Jeg isolerer efter depressionsgraden, MDI i nedenstående ligning vha. *CAS-værktøjet Maple.*

$$1.386294 = -8{,}8770 + 0{,}3877 \cdot MDI$$

$$MDI = 33{,}21$$

Ved brug af modellen forudsiges det, at værdien af *depressionsscoren, MDI* er *33,21*, når risikoen for at tage Sertralin overstiger *80%*. Dette tilsvarer en *svær depression (> 29).*

3.8

Jeg ønsker at finde oddsene for at anvende Sertralin ved en depressionsscore på 0 (MDI = 0), hvor denne værdi indsættes i modellen for den logistiske regression, som omskrives til den opstillede model i Odds skala:

$$logOdds\ (Sertralin) = -8{,}8770 + 0{,}3877 \cdot 0$$

$$Odds(Sertralin) = e^{-8{,}8770+0{,}3877 \cdot 0} = 0{,}0001395622 = 1{,}3 \cdot 10^{-4}$$

Odds for at anvende sertralin ved en depressionsgrad på 0 er dermed svarende til $1{,}3 \cdot 10^{-4}$. Dette betyder, at sandsynligheden for at tage Sertralin ved en depressionsscore på 0 er meget lille, da oddsene er tæt på nul. Dette betyder, at Sertralin formentligt ikke anvendes i nogen tilfælde, når MDI er lig nul.

Opgave 4. Sammenhængen mellem Sertralin, mundtørhed og caries

4.1

For at undersøge om risikoen for caries afhænger af brugen af Sertralin. Den afhængige variabel, *caries* er binær kategorisk *(1:ja / 0:nej).* Den uafhængige variabel, *Sertralin* er ligeledes binær kategorisk. For at undersøge sammenhængen mellem to binære variable anvender jeg en χ^2-test.

Her laver jeg først en matrix tabel over fordelingen af personer, hvor jeg først definerer to nye grupper med *"SertralinYes" og "SertralinNo".* Herefter kan jeg undersøge, hvor mange personer som har caries i de to grupper.

	Caries	Ingen caries
Forbrug af Sertralin	43	33
Intet forbrug af Sertralin	11	60

Nulhypotesen, som R tester, angiver at fordelingen af den afhængige variabel, *caries* er ens på tværs af de forskellige kategorier af den uafhængige variabel, *sertralin* (*forbrug af Sertralin og intet forbrug af Sertralin*). Dette indebærer, at sandsynligheden for de mulige udfald for den afhængige variabel (caries ja/nej) er den samme i begge grupper.

Fra χ^2-testen fås en p-værdi, som er $p = 5{,}964 \cdot 10^{-7}$, som er mindre end signifikansniveauet ($p < 0{,}05$). Dermed kan nulhypotesen, om at der ikke ses nogen sammenhæng, *forkastes*. Dette betyder, at der ses en statistisk signifikant sammenhæng mellem forbruget af Sertralin og udviklingen af caries.

4.2

χ^2-testet afrapporterer kun en p-værdi, men giver ikke indsigt i størrelsen af en eventuel forskel mellem dem, som tager Sertralin og dem som ikke gør. Her kan den relative risiko anvendes til at beskrive denne forskel. Jeg udregner først risikoen for at have caries i de to grupper;

Risiko for at have caries for personer som bruger Sertralin: RR = $\frac{43}{43+33} = 0{,}5657895$

Risiko for at have caries for personer som *ikke* bruger Sertralin: RR = $\frac{11}{11+60} = 0{,}1549296$

Jeg beregner nu den relative risiko (risiko-ratioen): RR = $\frac{0.5657895}{0.1549296} = 3.651914 = 3{,}65$

Den relative risiko er beregnet til 3,65 hvilket betyder, at personer, som bruger det antidepressive Sertralin, har en *3,65 (265%)* gange højere risiko for at udvikle caries sammenlignet med dem, som ikke bruger dette antidepressivum. Dette stemmer godt overens med det indledende teoriafsnit om, hvordan nogle medikamenter (herunder antidepressive) kan øge risikoen for caries ved at forårsage mundtørhed.

Odds-ratioen udregnes ligeledes for caries ved sammenligning af dem som bruger Sertralin og dem som ikke gør. Først beregnes odds:

Odds for at have caries for personer som bruger Sertralin: OR = $\frac{43}{33} = 1{,}30303$

Odds for at have caries for personer som *ikke* bruger Sertralin: OR = $\frac{11}{60} = 0{,}1833333$

Jeg beregner nu odds-ratioen: OR = $\frac{1{,}30303}{0{,}1833333} = 7{,}107438 = 7{,}11$

En odds-ratio på *7,11* antyder, at personer som bruger Sertralin har *7,11 (eller 611%)* gange større odds for at udvikle caries sammenholdt med dem, som ikke bruger Sertralin. Dette stemmer ligeledes godt overens med den tidligere konklusion for risiko-ratioen.

4.3

For at undersøge om risikoen for mundtørhed (xerostomi) afhænger af brugen af Sertralin. Den afhængige variabel, *mundtørhed* er binær kategorisk *(1:ja / 0:nej).* Den uafhængige variabel, *Sertralin* er ligeledes binær kategorisk. For at undersøge sammenhængen mellem to binære variable anvender jeg en χ^2-test. Igen laver jeg en matrix tabel over fordeling som i opgave 4.1, hvor jeg igen kan bruge de to definerede grupper *("SertralinYes" og "SertralinNo").*

	Xerostomi	Ingen xerostomi
Forbrug af Sertralin	41	35
Intet forbrug af Sertralin	14	57

Nulhypotesen, som testet af R, indikerer at fordelingen af den afhængige variabel, *mundtørhed (xerostomi),* er ens på tværs af de forskellige kategorier af den uafhængige variabel, *sertralin* (*forbrug af Sertralin og intet forbrug af Sertralin*). Med andre ord antyder det, at sandsynligheden for de mulige udfald for den afhængige variabel (mundtørhed: ja/nej) er ens i begge grupper.

Resultatet fra χ^2-testen viser en p-værdi på $p = 3{,}87 \cdot 10^{-5}$, hvilket er mindre end det valgte signifikansniveau *($p < 0{,}05$).* Derfor *forkastes* nulhypotesen, om at der ikke ses nogen sammenhæng mellem *mundtørhed* og forbruget af *Sertralin*. Konklusionen er, at der ses en statistisk signifikant sammenhæng mellem forbruget af Sertralin og forekomsten af mundtørhed.

4.4

Først udregnes den relative frekvens for at have mundtørhed i de to grupper; *("SertralinYes" og "SertralinNo").*

Relativ frekvens for at have mundtørhed for personer, som bruger Sertralin: $\frac{41}{41+35} = 0{,}5394737 = 53{,}9\%$

Relativ frekvens for at have mundtørhed for personer, som *ikke* bruger Sertralin: $\frac{14}{14+57} = 0.197183 = 19{,}7\%$

Jeg beregner nu risikodifferensen ved at trække de to frekvenser fra hinanden, $RD = 0{,}3422906 = 34{,}2\%$. Risikodifferensen er lig med *34,2* procentpoint.

En risikodifferens på *34,2* procentpoint mellem mundtørhed og brugen af sertralin indikerer, at der er en væsentlig forskel i risikoen for at udvikle mundtørhed (xerostomi) mellem dem, der bruger sertralin, og dem som ikke gør. Denne store risikodifferens antyder, at Sertralin kan have en statistisk signifikant variabel i forekomsten af mundtørhed hos dem, som bruger det antidepressivum.

4.5

Igen anvender jeg en χ^2-test for at undersøge om risikoen for caries afhænger af mundtørhed. Den afhængige variabel, *caries* er binær kategorisk *(1:ja / 0:nej).* Den uafhængige variabel, *mundtørhed* er ligeledes binær kategorisk *(1:ja / 0:nej)*. Jeg laver en matrix tabel over fordeling som i opgave *4.1* og *4.3*, men definerer nu to nye grupper *("XerostomiYes" og "XerostomiNo").*

	Caries	Ingen caries
Mundtørhed	38	17
Ingen mundtørhed	16	76

Nulhypotesen, som R tester, indikerer at fordelingen af den afhængige variabel, *caries,* er ens på tværs af de forskellige kategorier af den uafhængige variabel, *mundtørhed (mundtør og ikke mundtør).* Dette betyder, at sandsynligheden for de mulige udfald for den afhængige variabel (caries: ja/nej) er ens i begge grupper.

Analysen af χ^2-testen viser en p-værdi på $p = 9{,}65 \cdot 10^{-10}$, hvilket er mindre end det valgte signifikansniveau *($p < 0{,}05$).* Dermed kan nulhypotesen, om at der ikke ses nogen sammenhæng, *forkastes*. Dermed ses en statistisk signifikant sammenhæng mellem mundtørhed og forekomsten af caries.

4.6

Nu ønskes det at opdele datasættet i to grupper: personer med mundtørhed og personer ude mundtørhed *(som i 4.5)*. Dette vil tillade mig at undersøge sammenhængen mellem sertralin og caries uden hensyntagen til mundtørhed. Jeg vil fokusere på de to binære variable, sertralin og caries i analysen.

Jeg har tre kategoriske binære variable. En af disse variabler udelades, idet mundtørhed anvendes til at definere det nye datasæt *(XerostomiNo).* Eftersom jeg igen skal analysere to binære variable, kan jeg igen anvende χ^2-testet.

En χ^2-test kan være en god statistisk metode til at undersøge sammenhængen mellem de to binære variable, hvor p-værdien kan give indsigt i eventuelle signifikante sammenhænge mellem grupperne. I R laver jeg en krydstabuleringstabel, som viser antallet af observationer for hver kombination af værdierne for *sertralin* og *caries* i datasættet for personer uden mundtørhed.

	Ingen caries	Caries
Ingen brug af sertralin	48 (52%)	9 (9,8%)
Brug af sertralin	28 (30%)	7 (7,6%)

Af ovenstående tabel, hvor der samlet er *92* observationer (relativ frekvens i parentes) ses det, at der er 48 observationer, hvor patienten hverken bruger sertralin eller har caries. Kun *7* personer har caries samtidigt med forbrug af sertralin, hvilket svarer til en relativ frekvens på *7,6%* af personerne uden mundtørhed. Ud fra ovenstående tabel ses der ikke den store statistiske sammenhæng mellem forbruget af sertralin sammenholdt med forekomsten af caries for personer uden mundtørhed. Dette undersøges nu yderligere ved statistiske analyser.

P-værdien, som analysen i R, viste var *0,815*. Denne p-værdi er langt over det valgte signifikansniveau *($p = 0,05$),* hvorfor nulhypotesen, om at der ikke findes en sammenhæng mellem sertralin og forekomsten af caries, ikke kan forkastes.

I opgave 4.1 blev det konkluderet, at der var en signifikant statistisk sammenhæng mellem brugen af sertralin og forekomsten af caries ($p = 5{,}964 \cdot 10^{-7}$). Denne signifikante sammenhæng blev imidlertid fundet i hele datasættet, hvilket inkluderer både mundtøre personer og personer uden mundtørhed. Opsummerende illustrerer opgave *4.6* en vigtig pointe i statistik, hvor flere analyser på tværs af datasættet og dets variable er nødvendige for at drage en korrekt konklusion.

I opgave 4.1 konkluderedes det, at der var en sammenhæng i populationen i forhold til udvikling af caries ved brug af sertralin. I 4.6 kan det nu konkluderes, at der *kun* ses en signifikant statistisk sammenhæng for udviklingen af caries ved brug af sertralin for mundtøre personer i datasættet.

4.7

Opgave 4 undersøger sammenhængen mellem de tre kategoriske binære variable; *sertralin (ja/nej)*, *caries (ja/nej)* samt *mundtørhed/xerostomi (ja/nej)*. I de følgende opgaver fandtes deres indbyrdes sammenhæng.

I opgave *4.1* udførte jeg en χ^2-test, hvorfra jeg dragede konklusionen, at der ses signifikant statistisk sammenhæng mellem forbruget af Sertralin og forekomsten af caries. I opgave *4.2* kunne førnævnte konklusion yderligere bekræftes, idet gruppen som bruger sertralin har *3,65* gange højere

risiko samt *7,11* gange større odds for at udvikle caries sammenlignet med dem, som ikke tog dette antidepressivum.

Herefter undersøgtes sammenhængen mellem risikoen for at have mundtørhed og brugen af sertralin. Dette blev ligeledes undersøgt ved en χ^2-test, hvor det konkluderedes at der ses en statistisk signifikant sammenhæng mellem de to variable. Denne konklusion bekræftes yderligere af opgave *4.4*, hvor risikodifferensen var på *34,2* procentpoint mellem mundtørhed hos personer, som bruger sertralin og dem som ikke gør.

I opgave *4.5* blev sammenhængen mellem caries og mundtørhed ligeledes undersøgt vha. χ^2-test. Her sås igen en statistisk signifikant sammenhæng mellem mundtørhed og forekomsten af caries.

De indbyrdes relationer mellem de tre binære variable i opgave *4* er dermed blevet analyseret. Heraf kan det afledes, at personer med forbrug af sertralin lider af mundtørhed. Det er netop mundtørheden, som er en faktor for udviklingen af caries. Det er derfor ikke det antidepressive stof sertralin, som direkte fører til caries *(jævnfør opgave 4.6)*. Opsummerende kan sertralin have en indirekte indvirkning på udviklingen af caries, idet en bivirkning heraf kan være mundtørhed, som det også kendes fra klinikken. I opgave 3.3 udledtes det yderligere, at depressionsgraden har en signifikant betydning for forbruget af det antidepressive middel, *Sertralin*. Dermed vil en øget depressionsgrad ligeledes føre til en øget risiko for forekomsten af mundtørhed – og dermed udviklingen af caries.

Appendix

```r
#Opgave1
#1.1
d <- read.csv("http://causal.sund.ku.dk/f24/84.csv", header=TRUE, stringsAsFactors=TRUE)
#1.2
table(d$sertralin)
76/147*100
71/147*100
YesS<-subset(d,sertralin==1)
NoS<-subset(d,sertralin==0)
#Age
hist(YesS$alder)
hist(NoS$alder)
mean(YesS$alder)
mean(NoS$alder)
IQR(YesS$alder)
IQR(NoS$alder)
#MDI
table(YesS$MDI)
3/76*100
14/76*100
19/76*100
40/76*100
table(NoS$MDI)
44/71*100
22/71*100
4/71*100
1/71*100
#Sex
table(YesS$sex)
37/76*100
39/76*100
table(NoS$sex)
43/71*100
28/71*100
#Caries
table(YesS$caries)
43/76*100
33/76*100
table(NoS$caries)
11/71*100
60/71*100
#Xerostomi
table(YesS$xerostomi)
41/76*100
35/76*100
table(NoS$xerostomi)
```

```
14/71*100
57/71*100
#Opgave2
#2.1
m1<-lm(d$MDI~d$alder*d$sex)
summary(m1)
#2.2
confint(m1)
#2.4
m2<-lm(d$MDI~d$alder+d$sex)
summary(m2)
#2.5
20.80+0.11*30
#2.6
24.10+2*8.652
24.10-2*8.652
#2.7
male<-subset(d,sex=="male")
mean(male$alder)
hist(male$alder)
female<-subset(d,sex=="female")
mean(female$alder)
hist(female$alder)
t.test(d$alder~d$sex)
#Opgave3
#3.2
m3<-glm(d$sertralin~d$MDI,family=binomial)
summary(m3)
#3.3
exp(0.3877)
confint(m3)
exp(0.2751644)
exp(0.5361457)
#3.4
exp(-8.8770+0.3877*26)/(1+exp(-8.8770+0.3877*26))
#3.5
(exp(-8.8770+0.3877*26)/(1+exp(-8.8770+0.3877*26)))/(exp(-8.8770+0.3877*18)/(1+exp(-8.8770+0.3877*18)))
#3.6
1.4735588^5
#3.7
0.80/(1-0.80)
log(4)
#3.8
exp(-8.8770)
#Opgave4
#4.1
```

```r
#4.1
SertralinYes<-subset(d,sertralin==1)
SertralinNo<-subset(d,sertralin==0)
table(SertralinYes$caries)
table(SertralinNo$caries)
m4<-matrix(c(43,11,33,60),nrow=2,ncol=2)
m4
chisq.test(m4)
#4.2
43/(43+33)
11/(11+60)
(43/(43+33))/(11/(11+60))
43/33
11/60
(43/33)/(11/60)
#4.3
table(SertralinYes$xerostomi)
table(SertralinNo$xerostomi)
m5<-matrix(c(41,14,35,57),nrow=2,ncol=2)
m5
chisq.test(m5)
#4.4
41/(41+35)
14/(14+57)
(41/(41+35))-(14/(14+57))
0.5394737-0.1971831
#4.5
XerostomiYes<-subset(d,xerostomi==1)
XerostomiNo<-subset(d,xerostomi==0)
table(XerostomiYes$caries)
table(XerostomiNo$caries)
m6<-matrix(c(38,16,17,76),nrow=2,ncol=2)
m6
chisq.test(m6)
#4.6
table(XerostomiNo$sertralin, XerostomiNo$caries)
m7<-matrix(c(48,28,9,7),nrow=2,ncol=2)
m7
chisq.test(m7)
```