STATISTIK - Sommereksamen 2022
Eksamensnummer: 80

1.Datapræsentation

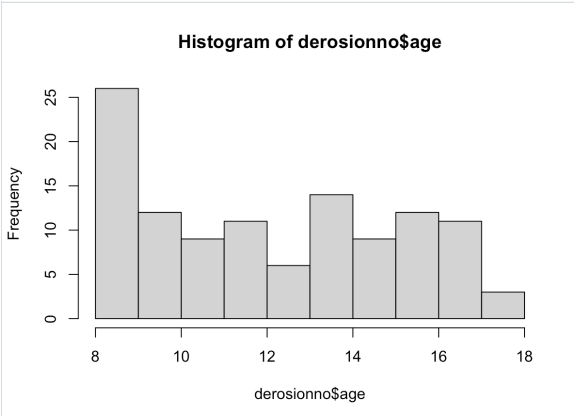
1.1 d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)

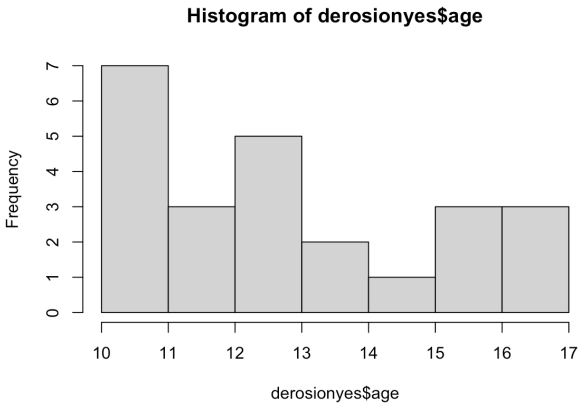
1. 2

Tabel 1 skal være en præsentation af "Baseline data" og man burde derfor have lavet en undersøgelse af studiepopulationen inden de medvirkende begyndte at svømmetræne regelmæssigt. Grundlaget for at undersøge de to grupper inden er, at man ser på om de to grupper er tilstrækkeligt ens til, at det er rimeligt at sammenligne de to gruppen.

Da disse pre-svømningsdata ikke er tilgængeligt i mit datasæt vælger jeg at lave en tabel, hvor jeg undersøger de 5 variable i mit datasæt, som er: "age", "crawl", "pooltime", "soda" og "erosion" Jeg vælger, at tage udgangspunkt i om svømmerne har ætsninger på tænderne eller ej og laver dermed en lidt "bagvendt tabel 1", som kigger på om dem med ætsninger på tænderne har været "Udsat" for det samme:

Age: For at undersøge om svømmernes alder er normalfordelt eller ej, plotter jeg alder og ætsning på tænderne i et histogram: ét for gruppen af svømmer uden ætsninger og ét for gruppen af svømmer med ætsninger

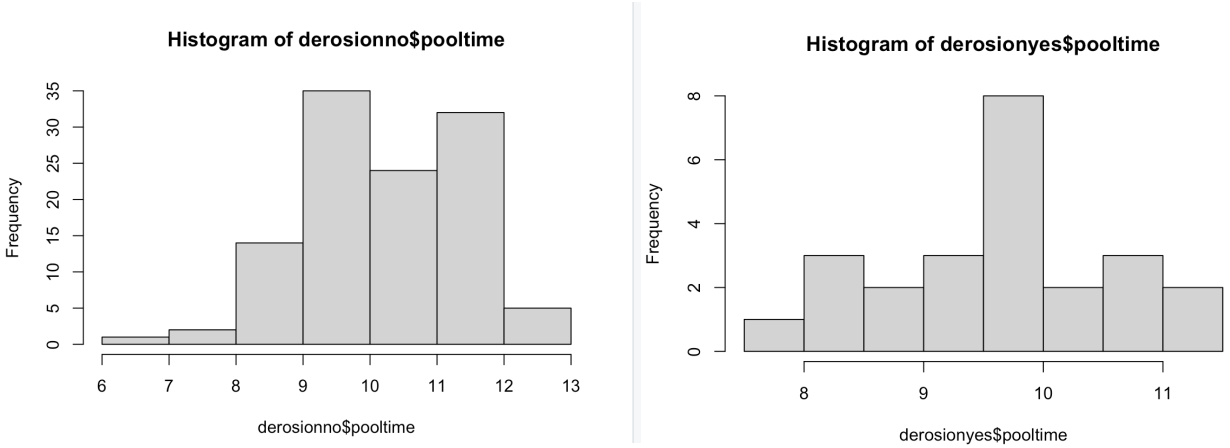




Ud fra histogrammerne ser jeg, at fordelingen i begge grupper er lettere højreskæve. Histogrammet for svømmere med ætsninger på tænderne er kun på 24 observationer. Begge grupper behandles som om, at de er højreskæve og derfor vælger jeg at bruge medianen og IQR.

Det havde også været en mulighed, at behandle alder som en kategorisk variabel. Men eftersom at tabel 1 er en præsentation af data, synes jeg, at det giver bedst mening at arbejde med den på ovenstående måde - da jeg mener at median og IQR giver modtageren et bedre og hurtigere overblik af data.

Pooltime: For at undersøge den bedste illustrering af pooltime-data i tabel 1, undersøger jeg om den gennemsnitlige mængde tid som svømmerne bruger i svømmebassinet er normalfordelt. Jeg laver histogrammer, hvor jeg plotter alder og ætsning på tænderne: ét for gruppen af svømmer uden ætsninger på tænderne og ét for gruppen af svømmer med ætsninger på tænderne.



Her ses det på diagrammerne, at fordelingerne i begge grupper ser normalfordelt pga. klokkeformen og jeg finder derfor middelværdien og spredningen for disse.

Soda og crawl: Begge disse variable bestemmes ved at lave opsummeringstabeller for gruppen af svømmere med ætsninger og for gruppen af svømmer uden ætsninger. Tabellerne viser antallet af person, hvorfor den givne variabel er gældende.

Herefter findes den procentvise andel i den givne gruppe. Da grupperne ikke er lige store, er den procentvise angivelse med for at gøre tabellen mere overskuelig for modtageren, så sammenhængen fremstår tydeligere.

	Swimmer with erosion	**Swimmer without erosion**
No. of participants	113	24
Age, median (IQR)	13 (4.25)	12(5)
Pooltime, mean (standard diviation)	9.62(1.19)	10.25(0.94)
Soda, N (%) Never <5/week >=5/week NA	 2 (8.3) 10 (41.7) 10 (41.7) 2 (8.3)	 10 (8.8) 45 (39.8) 46 (40.7) 12 (10.6)
Crawl, N (%) Yes No	 21 (87.5) 3 (12.5)	 32 (28.3) 81 (71.7)

Tabel 1: Charateristics of study participants

2. Sammenhæng mellem crawlsvømning og alder

2.1

I denne opgave vil jeg gerne undersøge "Hvordan sandsynligheden for crawlsvømning afhænger af alder". Da den uafhængige variabel ("Age") er kontinuert og den afhængige ("crawl") variabel er binær, skal jeg bruge en logistisk regression til at besvare det videnskabelige spørgsmål. Logistisk regressionsmodel med én uafhængig variabel er givet ved følgende ligning: LogOdds(y)=a+b*x Logistisk regression er en statistisk model, hvor log Odds for den binære afhængige variabel beskrives som en ret linje i forhold til den uafhængige variabel. Jeg kan således tænke på logistisk regression som en lineære model men blot på log Odds-skala.

2.2

Modellen opsætte dette ved at lave en generaliseret lineær model, hvorfra jeg får værdierne for modellen, den ser dermed således ud:

LogOdds(crawl)= -2.9318 + 0.19291*age.

For regressionsmodellen finder jeg, at odds-ratioen for alder er givet ved e^b=e^0.19291 = 1.212774. Jeg kan da konkludere følgende: For hver gang alderen øges med ét år, bliver odds for at svømme crawl 1.21 gange større svarende til en stigning i odds på ca. 21%.

P-værdien for effekten af alder på sandsynligheden for at være crawlsvømmer er 0.002245. Nulhypotesen er, at regressionskoefficienten er lig med nul, eller med andre ord at odds-rationen er med ét. Fordi p-værdien er mindre en 0.05, forkaster jeg nulhypotesen og konkluderer, at sammenhængen er statistisk signifikant for alder og crawlsvømning.

Et 95% konfidensinterval for effekten af alder er givet ved [1.07;1.38]. At intervallet ikke indeholder tallet 1, stemmer overens med konklusionen om en signifikant effekt af alder. Jeg konkluderer da følgende: Jeg er 95% sikre på, at intervallet fra 1.07 til 1.38 indeholder den sande odds-ratio for effekten af alder.

2.3

For at bestemme odds-skalaen benyttes følgende ligning: Odds(y)=e^(a+b*x), hvor jeg benytte de variable beregnet ovenfor:

Odds(crawl)=e^(-2.9318 + 0.19291*age)

2.4

Når logistiske regression beskriver en sandsynlighed for en hændelse eller ikke at opleve en hændelse er lig 100%, kan den skrives på følgende måde:

P(y=1) = (E^(a+b*x))/(1+e^(a+b*x))

I dette tilfælde vil jeg gerne opskrive et udtryk for at svømme crawl, som værende hændelsen og udtrykket bliver derfor:

P(crawlsvømning = 1) = (e^(-2.9318 + 0.19291*age))/ (1+e^(-2.9318 + 0.19291*age)))

2.5
Modellen er et udtryk for sandsynlighedsskalaen og derfor sættes 9 inde på pladsen: age.

$$P(\text{crawlsvømning} = 1 \mid \text{age} = 9) = 0.23$$ svarende til 23%

Dette betyder, at sandsynligheden for at en 9-årig svømmer også svømmer crawl er 23%

2.6
Risikorationen beregnes ved at tage den prædiktere sandsynligheden for en 14-årig i forhold til en 12-årig:

$$(P(\text{crawlsvømning} = 1 \mid \text{age} = 14)/ P(\text{crawlsvømning} = 1 \mid \text{age} = 12)$$

$$0.44/0.35 = 1.26$$

Risikoen for at en 14-årig svømmer svømmer crawl er 44% og risikoen for at en 12-årig svømmer svømmer crawl er 35%. Risikoen for at svømme crawl er dermed 1.26 større (dvs. 26% højere) blandt de 14-årige svømmere i forhold til de 12-årige svømmere.

2.7
I opgave 2.2 blev odds beregnet til 1.212774 pr. år og da denne vokser exponentiel, skal dette tal blot sættes i tredje for at beregne odds, når alderen stiger med tre år. Efter 3 år er odds nu 1.78.

2.8
Jeg kan prædiktere alderen, hvor sandsynligheden for crawlsvømning først krydser 25%. Jeg ved fra opgave 2.5 at den er lige under for er 9 årig og derfor eksperimentere jeg videre og det viser sig her at for de 10 årige er den: 0.26 dvs. 26% og det er derfor for alderen 10 år, hvor den først krydser 25%, hvis jeg kigger på alder som hele tal.

Alternativt kunne man ville have fundet den eksakte alder til opgave - ved løsning af denne ligning:

$$(P(\text{crawlsvømning} = 1 \mid \text{age} = x) = 0.25$$

2.9
Intercept (i dette datasæt: -2.93) betegner skæringen med y-aksen på logOdds-skalaen og dette tal er dermed odds for at være crawlsvømmer til alderen 0. Odds er altid et positivt tal og kan derfor teoretisk set ikke være -2.93. Så ud fra dette, er der en indikation, om udtrykket er dårlig til at prædiktere udenfor data, dvs. ekstrapolationen.

P-værdien for dette tal er 0.00483 og er baseret på en nulhypotese om at intercept=0, men da p-værdien er 0.05 forkastes nulhypotesen og man vil derfor teoretisk set ikke tro at intercepts sande værdi er 0. At denne nulhypotese forkastes tyder igen på, at modellen ikke er god til at håndtere områder, der ligger uden for den aldersgruppe, som er observeret i datasættet.

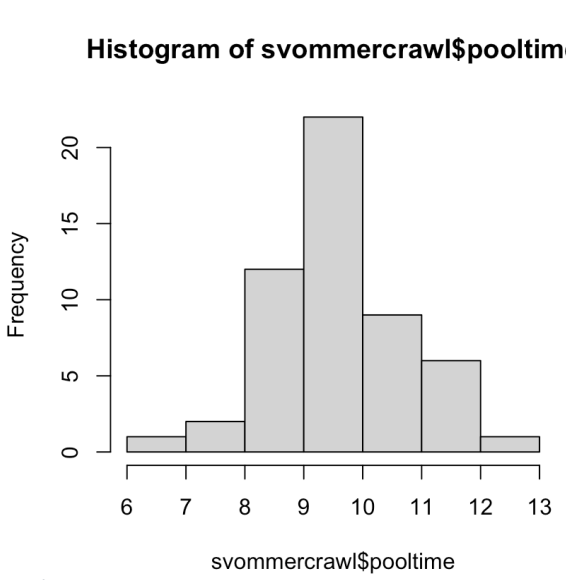
3. Hvad påvirker tiden brugt i svømmebassin?

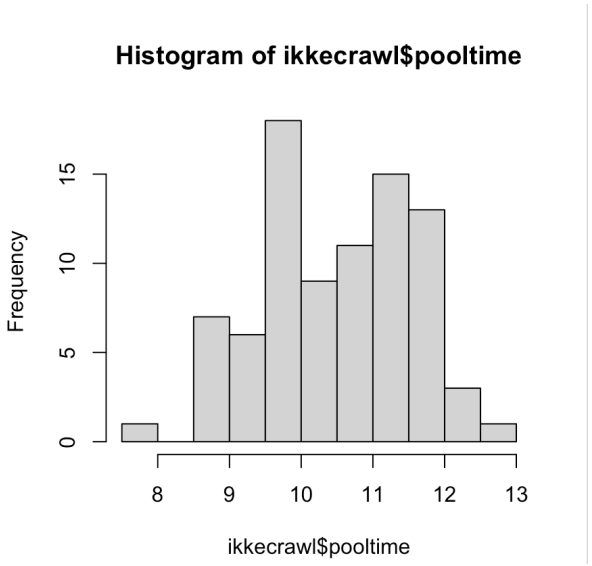
3.1
Da den afhængige variabel ("pooltime") er kontinuert og den uafhængige ("crawl") variabel er binær, skal jeg bruge en t-test for to populationer(uparret) til at besvare det videnskabelige spørgsmål.

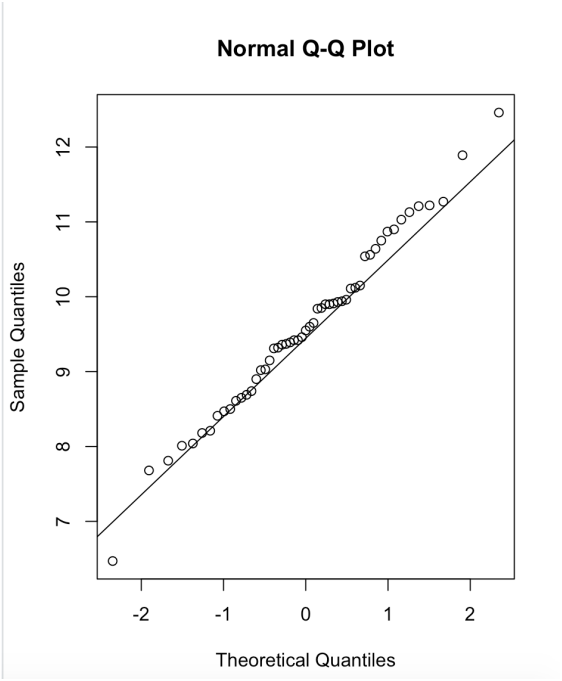
3.2

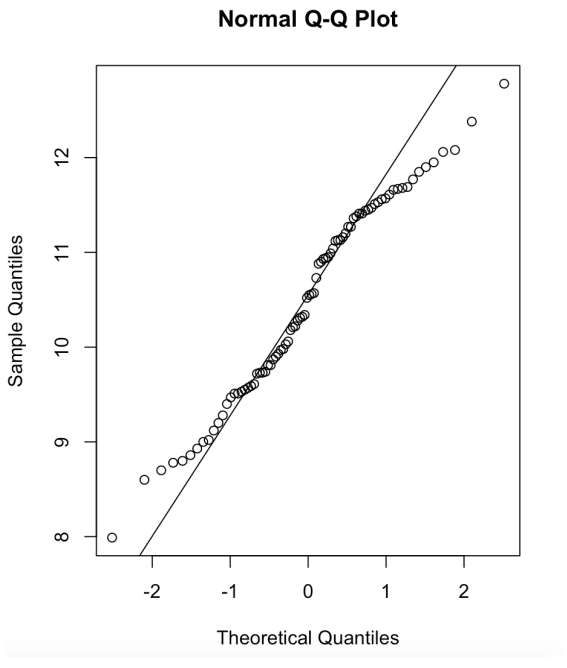
For at undersøge den gennemsnitlige forskel i tid brugt i svømmebassin pr. uge for crawlsvømmere i forhold til ikke-crawlsvømmere deles datasættet op i to grupper: dem som svømmer crawl og dem som ikke gør. Herefter findes gennemsnittet for de to grupper, hvor det for crawlsvømmere er 9.6 time/uge og det for dem der ikke svømmer crawl er 10.3 time/uge - dem der svømmer crawl bruger dermed i gennemsnit 0.7 time (svarende til ca. 42 min.) mindre pr. uge i forhold til dem, som ikke svømmer crawl.

Herefter laves histogrammer for at tjekke om de to grupper af svømmers antal timer pr. uge i svømmebassinet er normalfordelt, da dette er en forudsætning for, at jeg kan lave en t-test. Både gruppen af crawlsvømmere og gruppen af ikke-crawlsvømmere viser sig at have fine klokkeformede histogrammer, derfor antages data som værende normalfordelte. Derudover laves yderligere et Q-Q plot for at teste om data er normalfordelt, som også ser fornuftige ud.
Figurerne til højre viser crawlsvømmer og figurerne til venstre viser ikke-crawlsvømmere









T-testen viser, at den sande værdi for forskellen ligger imellem, at dem der svømmer crawl bruger mellem 1.28 time og 0.51 time mindre pr. uge i svømmebassin end dem der ikke svømmer crawl, dette er angivet på et 95% konfidensinterval.

Når jeg undersøger to uparrede populationer ved en t-test, tester jeg, om middelværdien for de to grupper er den samme. Nulhypotesen i denne opgave er derfor: Der er ikke forskel på den mængde af tid, som crawlsvømmere og ikke-crawlsvømmere bruger i svømmebassinet i løbet af en uge. T-testen for den hypotese er 1.329*10^-5 og eftersom at dette er mindre end 0.05 forkastes nulhypotesen. Jeg kan dermed ikke udelukke, at der ikke er en forskel på den mængde af tid, som crawlsvømmere og ikke-crawlsvømmere bruger i svømmebassinet i løbet af en uge.

3.3

For at beregne et 95% referenceinterval for tid brugt i svømmebassin pr. uge blandt crawlsvømmere bruges den netop tilegnede viden om, at antallet af timerne for crawlsvømmere er normalfordelt. For data som er normalfordelt gælder det, at gennemsnittet +/- 2 gange spredningen giver 95% referenceinterval, og eftersom gennemsnittet allerede er beregnet i ovenstående opgave, kan jeg nøjes med at beregne spredningen og efterfølgende bestemme intervallet.
Jeg får intervallet til at være fra 7.28 til 11.91 timer pr. uge. Fortolkningen er, at 95% af alle dem, som svømmer crawl, vil bruge mellem 7.28 til 11.91 timer pr. uge i svømmebassinet.

3.4

Da den afhængige variabel (“pooltime”) er kontinuert og den uafhængige (“age”) variabel er kontinuert, skal jeg bruge en lineær regression(y=bx+a) til at besvare det videnskabelige spørgsmål.

3.5

Udtrykket for den estimerede sammenhæng beskrives ved:

Antal timer i svømmebassin = 0.1404*alder + 8.3716 dvs. Pooltime=0.1404*age + 8.3716)

Hvor 8.3716 er den estimerede værdi af a(skæringspunktet) og 0.1404 er den estimerede værdi for b (hældningskoefficienten)

Fortolkningen er, at min model viser, at det gennemsnitlige antal timer som man bruger i svømmebassinet ved 0 år er 8.3716 time pr. uge og for hvert år der går, vil tiden man bruger i svømmebassinet stige med 0.1404 time pr. uge i gennemsnit.

Som problematiseret i opgave 2.9 giver det nogle problemer, når jeg prøver at beskrive ekstrapolationen med mit data. Det giver nemlig ingen mening, at en person på 0 år går til svømning 8.3716 time ugentligt (Medmindre man forventer meget babysvømning)

3.6

For at beregne et 95% konfidensinterval for effekten af alder på tid brugt i svømmebassin pr. uge findes standardfejlen for den afhængige og for den uafhængige. Herefter finder 95% konfidensintervallet ved at +/- 2 gange standardfejl fra begge parametre under antagelse af at data er normalfordelt, som omtalt i 3.2.

Jeg kommer frem til følgende: Jeg er 95% sikre på at intervallet [0.078;0.203] indeholder den sande hældning mellem pooltime og age i populationen, og jeg er 95% sikker på, at intervallet [7.567;9.180] indeholder den sande skæring i populationen.

Under disse beregninger tester RStudio 2 nulhypotesen - én for age og én for pooltime, her tester RStudio om værdien for de to parametre er nul.
$H_0: a = 0$ Hvis denne nulhypotese skal være opfyldt vil skæringen med y-aksen være lig nul.
$H_0: b = 0$ Hvis denne nulhypotese skal være opfyldt vil hældningskoefficienten være lig nul.
Jeg undersøger i dette forsøg for effekten af alder på tid brugt i svømmebassin - derfor er jeg mest interesseret i om der er en sammenhæng mellem age og pooltime. Derfor er det $H_0: b = 0$, som jeg interesserer os for. Nulhypotesen for denne er: Der er ikke en sammenhæng imellem age og pooltime. P-værdien for nulhypotese er 1.73*10^-5 og derfor under 0.05. Jeg forkaster derfor nulhypotesen og jeg tror ikke, at der ikke er en sammenhæng imellem age og pooltime.

3.7
Til at prædiktere i denne opgave benyttes den opstillede model, hvor 13 år sættes ind som alder. Herved beregnes antallet af timer pr. uge for en 13-årig til at være 10.2 time.
Referenceintervallet kan findes ved +/- 2 gange spredningen, da jeg fra opgave 3.2 ved at data er normalfordelt. Referenceintervallet er [8.00;12.39] for de 13-årige

3.8
Til denne undersøgelse vil jeg gerne benytte os af multipel lineær regression, hvor vi i dette kursus bruger et specialtilfælde af modellen: yi = a + b1 * x1i + b2*x2i + ei, ei ~ N(0,s)
Dette er udtryk, som beskriver tilfældet med: Én kontinuert afhængig variabel og flere uafhængige variable: én kontinuert og én kategoriske(særtilfældet: binær)
Hvor pooltime er den afhængige variable, age er den kontinuerlige og crawl er binær kategorisk.

Jeg starter med at undersøge, om sammenhængen mellem pooltime og age er påvirket at crawl. Dette er tilfældet med hovedeffekt samt interaktionsled. P-værdien svarende til nulhypotesen at interaktionsleddet er lig 0 er -0.275 og dermed numerisk større end 0.05. Det betyder, at jeg ikke kan forkaste, at de to linjer har samme hældning, og der er derfor ikke tilstrækkelig empirisk evidens for at medtage en interaktion mellem alder og crawl til fordel for model med kun hovedeffekten af crawl. Derfor benytter jeg mig af modellen, som kun indebærer hovedeffekten:

yi = a + b1 * x1i + ei, ei ~ N(0,s)

3.9
Den nyfundne model er derfor givet ved:

Crawlsvømmere(Crawl=1): 6.95948+0.19342*age = pooltime
Ikkecrawlsvømmere (Crawl= 0): 8.17141+0.19342*age = pooltime

I begge udtryk for sammenhængen ganges ages med 0.19342, da dette er hældningen for begge grafer, som argumenteret ovenfor, hvorfor jeg kan tillade os at antage dette.
Hældningskoefficienten er i begge udtryk 0.19342. Det første tal i udtrykket er skæringen med y-aksen.

3.10
Jeg er 95% sikre på, at intervallet fra -1.55 timer til -0.87 timer indeholder den sande forskel i tid i svømmebassinet pr. uge for crawlsvømmere sammenlignet med ikke-crawlsvømmere til enhver alder.

3.11
Nulhypoteserne, som Rstudio tester for os, er, at hver af de pågældende regressionsparametre er lig med 0. P-værdien for nulhypotesen for ingen forskel på crawlsvømmer og ikke-crawlsvømmere er 7.23*10^-11, da dette er meget mindre end 0.05 forkastes nulhypotesen om, at det ikke har betydning om man svømmer crawl eller ej. Jeg konkluderer, at om man svømmer crawl eller ej har en signifikant effekt på tiden man bruger i svømmebassinet på en uge, når alder samtidigt tages i betragtning.

P-værdien for alder er 1.79*10^-10 og da dette ligeledes er mindre end 0.05 forkastes også denne nulhypotese, om at aller ikke har en betydning for antallet af timer brugt i svømmebassinet pr. uge. Jeg konkluderer, at alder også har en signifikant effekt på tiden man bruger i svømmebassinet på en uge, når svømme-arten samtidigt tages i betragtning.

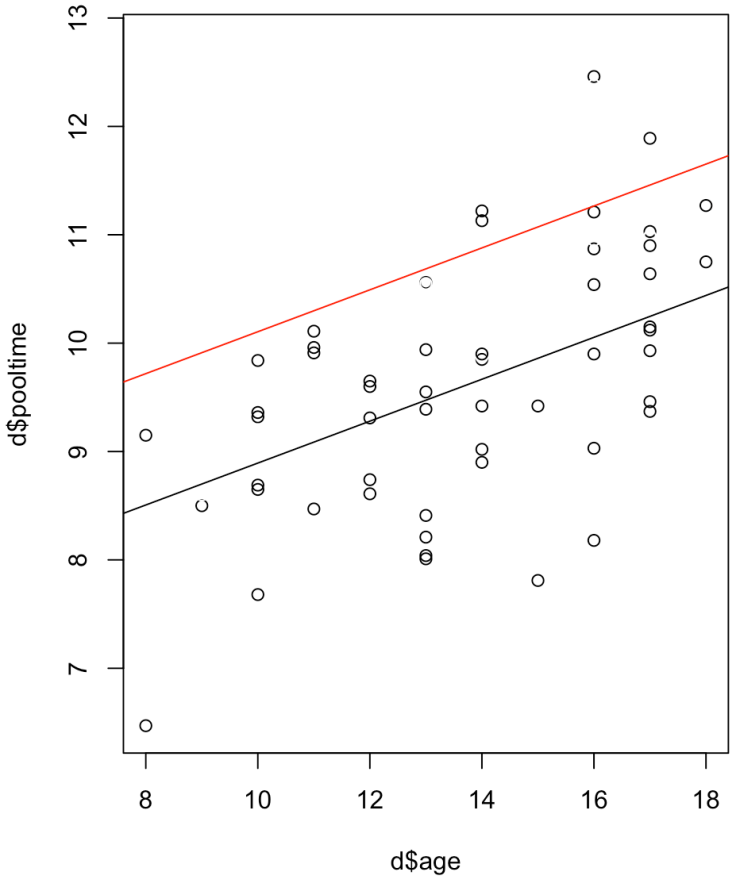
3.12
Min endelige model kan bruges til at prædiktere ved at indsætte den ønskede alder på age´s plads og dermed beregnes antallet af timer i svømmebassin pr. uge. Prædiktionen for en 12-årig crawlsvømmer er at vedkommende bruger 9.3 time i svømmebassinet pr uge.

3.13
Min endelige model kan bruges til at prædiktere ved at indsætte den ønskede alder på age´s plads og dermed beregnes antallet af timer i svømmebassin pr. uge. Prædiktionen for en 12-årig ikke-crawlsvømmer er at vedkommende bruger 10.5 time i svømmebassinet pr uge.

3.14
Da jeg i denne opgave arbejder med et udryk kun med hovedeffekt, betyder dette også at min to linjer er parallelle med andre ord. Disse har samme hældning og vil derfor ændre sig i tid i gennemsnit lige meget. Modellen fortæller os, at for hvert år ældre en svømmer(uanset svømme-art) bliver, da vil vedkommende i gennemsnit bruge 0.19 time mere i svømmebassinet pr. uge end personen gjorde året før

Dette betyder, at efter 4 år vil en svømmer (uanset svømme-art) i gennemsnit bruge 0.77 time mere i svømmebassinet pr. uge ift. hvor meget personen svømmede inden de 4 år.

3.15
Figuren illustrerer, hvordan tid i svømmebassin afhænger af alder og crawl. Hvor den røde linje repræsenterer ikke-crawlsvømmerne og den sorte linje repræsenterer crawlsvømmerne:



4. Indtagelse af kulsyreholdige læskedrikke

4.1
Definitorisk drikker personer, som ikke drikker sodavand, sodavand mindre end 5 gange om ugen. Jeg vælger derfor, at inkludere dem i gruppen af svømmere der drikker sodavand mindre end 5 gange om ugen. For at estimere sandsynligheden laves en opsummeringstabel og opdeler efter grupper, som beskrevet.

Herefter findes sandsynligheden ved at tage gruppen som drikker sodavand mindre end 5 gange om ugen i forhold til alle forsøgspersoner, hvor det har været muligt at indhente deres svar på dette spørgsmål (dvs. at NA-gruppen undlades). Sandsynligheden for at drikke sodavand mindre en 5 gange om ugen estimeres til ca. 54% i populationen (95% konfidensinterval = [0.45; 0.63])

4.2
Odds er givet ved forholdet mellem gruppen af mennesker som oplever dette i forhold til gruppen af mennesker som ikke oplever dette. Odds beregnes derfor til at være 1.1. Fortolkningen af denne odds er, at for hver person som drikker sodavand mere end 5 gange om ugen vil jeg forvente at se 1.1 person som drikker sodavand mindre end 5 gange om ugen.

4.3

Jeg laver en tabel med crawl og soda og for kunne beslutte de relative risiko for disse og samler værdierne for never og <5/week, se under:

	1=crawl	0= ikke crawl	Total
<5/week (mindre) + never	23 + 5=28	32+7=39	67
>=5/week (mere)	20	36	56
Total:	48	75	

Jeg ønsker, at beregne den estimerede risiko og risikodifferensen, dette gøres ved at finde antallet af personer, som oplever dette ud af den totale gruppe. Den beregning laves for begge grupper og jeg finder frem til at:

- Estimerede risiko for at svømme crawl, når man drikke sodavand mindre end 5 gange om ugen: 28/67 = 0.4179104 = 41.8%
- Estimerede risiko for at svømme crawl, når man drikker sodavand 5 gange eller mere om ugen: 20/56 = 0.3571429 = 35.7%

Herefter findes differencen ved at trække de to relative riskoer fra hinanden og finder risikodifferensen til at være 0.0607675 = 6.1 procentpoint. Den relative risiko bestemmes ved at finde forholdet mellem de to estimerede risikoer og bestemmes til 1.17.

Følgende kan derfor konkluderes: Den estimerede risiko for at være crawlsvømmer, når man drikker sodavand mindre end 5 gange om ugen er 41.7% og den estimerede risiko for at være crawlsvømmer, når man drikker sodavand 5 gange eller mere om ugen er 35.7%. Dette er en risikodifferens på 6.1 procentpoint.
Den relativ risiko på 1.17 fortæller, at risikioen for være crawlsvømmer, når man drikker sodavand mindre end 5 gange om ugen er 1.17 gang så høj som risikoen for at crawlsvømmer når man drikke sodavand 5 gange eller mere om ugen.

4.4

I denne opgaven findes odds-ratioen ved først at bestemme ratio for crawlsvømmere som unden soda-data svarer: aldrig og for dem som svarer: mindre end 5 gange om ugen, disse værdier sammenholdes med de svømmer, som har registreret et andet svar end dette dvs:

$$Odds\ (never) = never/(>=5/week + <5/week\ (mindre))$$
$$Odds\ (>=5/week) = >=5/week/(never + <5/week\ (mindre))$$

(Her har jeg lavet lignings-kommandoer med rød skrift, da det ellers så for forvirrende ud)
Herefter findes odds-rationen mellem disse ved at finde forholdet mellem dem dvs:

$$OR = Odds(>= 5/week)/Odds(never)$$

Odds for aldrig at drikke sodavand som crawlsvømmer er 0.12 og odds for at drikke sodavand mindre end 5 gange om ugen (men hyppigere end aldrig) er 0.92. Oddsrationen på 7.91 fortæller, at hvis man er crawlsvømmer, er det 7.91 gang mere sandsynligt at drikke sodavand mindre end 5 gange om ugen i forhold til at man aldrig drikker sodavand.

5. Hvad påvirker risikoen for ætsninger på tænder

5.1
Da den afhængige variabel ("erosion") er binær og den uafhængige ("crawl") variabel er binær, skal jeg bruge en chi-i-anden-test til at besvare det videnskabelige spørgsmål.

5.2
Jeg starter med indhente informationer ud af data til at kunne opstille en 2x2 tabel med erosion og crawl, som de to variable:

	Erosion: nej	Erosion: ja
Crawl: nej	81	3
Crawl: ja	32	21

På baggrund af denne tabel ønsker jeg at lave en chi-i-anden-test på. Ved chi-i-anden-test undersøges, om der er en signifikant sammenhæng mellem at svømme crawl og have ætsninger på sine tænder. Dette betyder, at jeg tester en nulhypotese om; at der er ikke er nogen sammenhæng imellem at svømme crawl og have ætsninger på sine tænder. Chi-i-anden-testen viser at p-værdien for nulhypotesen er 2.27*10^-7.
Dette er mindre end 0.05 og på baggrund af p-værdien forkaster jeg dermed nulhypotesen og kan dermed ikke bekræfte, at der ikke er en sammenhæng mellem at være crawlsvømmer og have ætsninger på tænderne.

5.3
På samme vis som i 4.3 finde risikodifferensen ved først at finde den estimerede risiko for de to ting, som jeg ønsker undersøgt:

Risko for crawlsvømmere: crawlsvømmere med erosinoer/totale antal crawlsvømmere: 39.6%
Risiko for ikke-crawlsvømmere: ikke-crawlsvømmere med erosinoer/totale antal ikke-crawlsvømmere:3.6%
Risikodifferense: crawlsvømmer-ikkecrawlsvømmere: 36.05 procentpoint

Tolkningen af dette er, at risikoen for at få ætsninger på tænderne er 36.05 procentpoint større hvis man svømmer crawl sammenlignet med hvis man ikke svømmer crawl.

5.4
Til denne opgave låner jeg de ovenfor estimerede risikoer ætsninger for henholdsvis crawlsvømmere og ikke crawl-svømmer. Til forskellen fra opgave 5.3 hvor jeg fandt forskellen på de to estimerede risikoer, finder jeg nu forholdet imellem disse. Dette gøres ved at dividere og den relative risiko beregnes herved til 11.09, hvilket fortæller os at risikoen for at få ætsninger på tænderne er 11.09 gange så høj, hvis man er crawlsvømmer ift. hvis man er ikke-crawlsvømmer

5.5

Odds-ratio bestemmes på samme måde som i opgave 4.4, dog med disse værdier - disse værdier hentes fra informationerne i opgave 5.2:

$Odds\ (crawlsvømmer))$ = *crawlsvømmer med erosione*/crawlsvømmer uden erosioner: 0.656
$Odds(ikkecrawlsvømmer)$ = *ikkecrawlsvømmer med erosioner*/ikkecrawlsvømmer uden erosioner: 0.037

Herefter findes odds-rationen mellem disse ved at finde forholdet mellem dem dvs:

$$OR = Odds(crawlsvømmer/Odds(ikkecrawlsvømmer): 17.72$$

Odds for have ætsninger på tænderne som crawlsvømmer er 0.656 og odds for at have ætsninger på tænderne som ikke crawlsvømmer er 0.037. Oddsrationen på 17.72 fortæller, at der er 17.72 gange mere sandsynligt at have ætsninger på tænder, hvis man svømmer crawl sammenlignet med hvis man ikke svømmer crawl.

5.6

Da den afhængige variabel ("erosion") er binær og den uafhængige ("pooltime") variabel er kontinuert , skal jeg bruge logistisk regression til at besvare det videnskabelige spørgsmål. Modellen for logistisk regression men én uafhængig variabel er givet ved følgenden:

LogOdds(y)= a + b*x

5.7

Jeg starter med at estimere modellen, dette gøres ved at lave en generaliseret lineær model, hvorfra jeg får værdierne for modellen:

LogOdds(erosion)=3.1793 -0.4757*pooltime

For regressionsmodellen finder jeg, at odds-ratioen for pooltime er givet ved:

e^b = e^-0.4757=0.6214499

Jeg kan da konkludere følgende: For hver gang antallet af timer i svømmebassinet øges med én time ugentligt, bliver odds for at have erosioner 0.62 gange større svarende til et fald i odds på ca. 38%.

P-værdien for effekten af tid brugt i svømmebassinet pr. uge på sandsynligheden for at have ætsninger på tænderne er 0.0188. Nulhypotesen er, at regressionskoefficienten er lig med nul, eller med andre ord at odds-rationen er ét. Fordi p-værdien er mindre end 0.05, forkaster jeg nulhypotesen og konkluderer, at sammenhængen er statistisk signifikant.

Et 95% konfidensinterval for effekten af tid brugt i svømmebassin er givet ved [0.408;0.410]. Jeg konkluderer følgende: Jeg er 95% sikre på, at intervallet fra 0.408 til 0.410 indeholder den sande odds-ratio effekten af tid brugt i svømmebassinet pr. uge. At intervallet ikke indeholder tallet 1, stemmer overens med konklusionen om en signifikant effekt af antal timer i brugt i svømmebassinet pr. uge

5.8

Jeg bruger samme fremgangsmåde som i ovenstående opgave, dog har jeg i denne opgave underinddelt datasættet, således at jeg nu kun undersøger crawlsvømmerne og finder følgende udtryk:

LogOdds(erosion) = 0.31741-0.07707*pooltime

For regressionsmodellen finder jeg, at odds-ratioen for pooltime er givet ved:

e^b = e^-0.07707 = 0.925825

Jeg kan da konkludere følgende: For hver gang antallet af timer i svømmebassinet øges med én time ugentligt, bliver odds for at have erosioner 0.925825 gange større svarende til et fald i odds på ca. 7.5%.

P-værdi for effekten af antal timer i svømmebassinet i løbet af en uge på sandsynligheden for ætsninger på tænderne aflæses til 0.754. Nulhypotesen er, at regressionskoefficienten er lig med nul, eller med andre ord at odds-rationen er ét. Fordi p-værdien er større end 0.05, forkaster jeg ikke nulhypotesen og konkluderer, at der ikke er sammenhæng mellem antallet af timer crawlsvømmerne bruger i svømmebassinet i løbet af en uge og om de har erosioner på tænderne.

Et 95% konfidensinterval for effekten af tid brugt i svømmebassin er givet ved [0.5636118;1.503304]. Jeg konkluderer da følgende: Jeg er 95% sikre på, at intervallet fra 0.408 til 0.410 indeholder den sande odds-ratio effekten af timer brugt om i ugen i svømmebassinet. At intervallet indeholder tallet 1, stemmer overens med konklusionen om: at antal timer i svømmebassinet om ugen ikke har en signifikant effekt på om crawlsvømmerne har ætsninger på tænderne.

5.9

Jeg bruger samme fremgangsmåde som i ovenstående opgave, dog har jeg i denne opgave underinddelt datasættet, således at jeg nu kun undersøger crawlsvømmerne og finder følgende udtryk:

LogOdds(erosion) = -0.7265-0.2479*pooltime

For regressionsmodellen finder jeg, at odds-ratioen for pooltime er givet ved:

e^b = e^-0.7265 = 0.4835986

Jeg konkluderer følgende: For hver gang antallet af timer i svømmebassinet øges med én time ugentligt, bliver odds for at have erosioner 0.4835986 gange større svarende til et fald i odds på næsten. 52%.

P-værdi for effekten af antal timer i svømmebassinet i løbet af en uge på sandsynligheden for ætsninger på tænderne aflæses til 0.661. Nulhypotesen er, at regressionskoefficienten er lig med nul, eller med andre ord at odds-rationen er ét. Fordi p-værdien er større end 5%, forkaster jeg ikke nulhypotesen og konkluderer, at der ikke er sammenhæng mellem antallet af timer ikke-crawlsvømmerne bruger i svømmebassinet i løbet af en uge og om de har erosioner på tænderne.

Et 95% konfidensinterval for effekten af tid brugt i svømmebassin er givet ved [0.238168;2.43397]. Jeg konkluderer da følgende: Jeg er 95% sikre på, at intervallet fra 0.238168 til 2.43397 indeholder den sande odds-ratio effekten af timer brugt om i ugen i svømmebassinet.. At intervallet indeholder tallet 1, stemmer overens med konklusionen om: at antal timer i svømmebassinet om ugen ikke har en signifikant effekt på om ikke-crawlsvømmerne har ætsninger på tænderne.

5.10

Jeg finder ikke en signifikant sammenhængen imellem tiden, som svømmerne bruger i svømmebassinet og om de har ætsninger på tænderne, når jeg inddeler svømmerne i om de svømmer crawl eller ej. Når jeg kigger på alle svømmerne er der en signifikant sammenhæng mellem tiden brugt i svømmebassinet pr. uge og om svømmerne har ætsningerne, dette er den afledte effekt af det som jeg observerede i opgave 3.3 og 5.3. I opgave 3.3 observerer en signifikant forskel på tiden som crawlsvømmere og ikke-crawlsvømmere brugte i svømmebassinet i løbet af en uge. I opgave 5.3 kunne jeg konstatere at crawlsvømmere har signifikant højere odds for at have ætsninger på tænderne i forhold til ikke-crawlsvømmere.

Når jeg i opgave 5.7 ser en signifikant sammenhæng mellem tid i svømmebassinet pr. uge og ætsninger på tænderne, er dette sandsynligvis en effekt af at crawlsvømmer, som har signifikant flere ætsninger på tænderne, også er dem som bruger signifikant kortest tid i svømmebassinet pr. uge af de to grupper. Denne konklusion som 5.7 giver (desto kortere tid i vand, desto flere erosion) er derfor mere et kendetegn på, at det er crawlsvømmer end den egentlige tid i svømmebassinet ugentligt. Dette er et tilfælde hvor to sammenhænge korrelerer.

Konklusionerne fra opgave 7,8 og 9 kan derfor godt stemme overens - selvom de umiddelbart virker modstridende. Datasættet indikerer nemlig, at det centrale for, om at man som svømmer har ætsninger på tænderne eller ej - afhænger af om man svømmer crawl eller ej - og ikke om hvor mange timer, man bruger ugentligt i svømmebassinet.

Appendix:

1. Datapræsentation

1.1 – Ingen tilhørende R-kode

1.2 - Table 1

```
#tabel 1
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
derosionyes<-subset(d,erosion==1)
derosionno<-subset(d,erosion==0)
#Age
hist(derosionno$age)
hist(derosionyes$age)
median(derosionno$age)
median(derosionyes$age)
IQR(derosionno$age)
IQR(derosionyes$age)
#POOLTIME
hist(derosionno$pooltime)
hist(derosionyes$pooltime)
mean(derosionno$pooltime)
mean(derosionyes$pooltime)
sd(derosionno$pooltime)
sd(derosionyes$pooltime)
#soda
table(derosionno$soda)
113-(45+46+10)
12/113
10/113
45/113
46/113
table(derosionyes$soda)
2/24
10/24
24-(10+10+2)
#crawl
table(derosionno$crawl)
32/113
81/113
table(derosionyes$crawl)
21/24
3/24
```

2. Sammenhæng mellem crawlsvømning og alder

2.1 — Ingen tilhørende R-kode

2.2

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
dto<-glm(d$crawl~d$age,family=binomial)
summary(dto)
exp(0.19291)
confint(dto)
exp(0.07216019)
exp(0.3208125)
```

2.3 — Ingen tilhørende R-kode

2.4 — Ingen tilhørende R-kode

2.5

```
(exp(-2.93138+0.19291*9))/(1+ exp(-2.93138+0.19291*9))
```

2.6

```
(exp(-2.93138+0.19291*14))/(1+ exp(-2.93138+0.19291*14))
(exp(-2.93138+0.19291*12))/(1+ exp(-2.93138+0.19291*12))
0.4425942/0.350587
```

2.7

```
1.212774^3
```

2.8

```
(exp(-2.93138+0.19291*9))/(1+ exp(-2.93138+0.19291*9))
(exp(-2.93138+0.19291*10))/(1+ exp(-2.93138+0.19291*10))
#Sandsynlighed 9 år: 0,23 - 10 år: 0,26

#Alternativ løsning: Løs som ligning
(exp(-2.93138+0.19291*x))/(1+ exp(-2.93138+0.19291*x)=0.25
```

2.9 – Ingen tilhørende R-kode

3. Hvad påvirker tiden brugt i svømmebassin?

3.1 – Ingen tilhørende R-kode

3.2

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
svommercrawl<-subset(d,crawl==1)
ikkecrawl<-subset(d,crawl==0)
mean(svommercrawl$pooltime)
mean(ikkecrawl$pooltime)
```

```
hist(svommercrawl$pooltime)
hist(ikkecrawl$pooltime)
qqnorm(svommercrawl$pooltime)
qqline(svommercrawl$pooltime)
qqnorm(ikkecrawl$pooltime)
qqline(ikkecrawl$pooltime)
#svømmer crawl gns. = 9.59434 og ikke gns. 10,29012
9.59434-10.29012
#Til minutter:
  0.69578*60
t.test(svommercrawl$pooltime,ikkecrawl$pooltime)
```

3.3

```
sd(svommercrawl$pooltime)
9.59434+2*1.157084
9.59434-2*1.157084
```

3.4 —Ingen tilhørende R-kode

3.5

```
plot(d$age,d$pooltime)
model<-lm(d$pooltime~d$age)
summary(model)
abline(8.3716,0.1404)
#8,3716 er skærringen m y-aksen, og 0,1404 er hældningen
```

3.6

```
confint(model)
```

3.7

```
0.1404*13+8.3716
10.1968+2*1.098
10.1968-2*1.098
```

3.8

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
#Med hovedeffekt og interaktionsled
ga<-lm(d$pooltime~d$age*d$crawl)
summary(ga)
# Da hvis crawl ja = (8.44020-2.05275)+(0.17100+0.06441)*age
# Da hvis crawl er nej:8.44020+0.17100*age = pooltime. rød
plot(d$age,d$pooltime, col=d$crawl)
abline((8.44020-2.05275),(0.17100+0.06441))
abline(8.44020,0.17100,col="red")
```

```
#Kun med hovedeffekt
pl<-lm(d$pooltime~d$age+d$crawl)
summary(pl)
# Da hvis crawl ja:(8.17141-1.21193)+0.19342*age = pooltime
# Da hvis crawl er nej:  8.17141+0.19342*age = pooltime. rød
plot(d$age,d$pooltime, col=d$crawl)
abline((8.17141-1.21193), 0.19341)
abline(8.17141,0.19341,col="red")
```

3.9

```
(8.17141-1.21193)
```

3.10

```
confint(pl)
```

3.11—Ingen tilhørende R-kode

3.12

```
(8.17141-1.21193)+0.19341*12
#Ja til crawl:9.2804 timer pr. uge
```

3.13

```
8.17141+0.19341*12
#nej til crawl:10.49233 timer pr. uge
```

3.14

```
0.19342*4
```

3.15

R-kommando fra udregninger i 3.8 kørt igen, dog nu brugbart :)

```
plot(d$age,d$pooltime, col=d$crawl)
abline(8.17141-1.21193, 0.19341)
abline(8.17141,0.19341,col="red")
```

4. Indtagelse af kulsyreholdige læskedrikke

4.1

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
table(d$soda)
#55+12 er mindre end 5 gange ugentligt, de resterende 56 er mere end 5 gange ugentligt.
(55+12)/(55+12+56)
prop.test((55+12),(55+12+56))
```

4.2

```
#odds = personer som oplever/personer som ikke oplever
```

```
(55+12)/56
```

4.3

```
table(d$soda, d$crawl)
23+5
28+20
32+7
39+36
28+39
20+36
28/67
20/56
0.4179104-0.3571429
0.4179104/0.3571429
```

4.4

Tal henter fra tabel i 4.3

```
#never=5, mindre=23, mere=20
5/(23+20)
23/(5+20)
0.92/0.1162791
```

5. Hvad påvirker risikoen for ætsninger på tænderne?

5.1—Ingen tilhørende R-kode

5.2

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE,
stringsAsFactors=TRUE)
table(d$crawl,d$erosion)
m <-matrix(c(81,32,3,21),nrow = 2,ncol=2)
m
chisq.test(m)
```

5.3

```
#risikodifferens
21/(32+21)
3/(81+3)
0.3962264-0.03571429
```

5.4

```
0.3962264/0.03571429
```

5.5

```
#data er hentet fra figur lave i 5.2
21/32
3/81
0.65625/0.03703704
```

5.6 —Ingen tilhørende R-kode

5.7

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE,
stringsAsFactors=TRUE)
mm<-glm(d$erosion~d$pooltime, family = binomial)
summary(mm)
plot(d$erosion~d$pooltime)
exp(-0.4757)
exp(3.1793)
1-0.62
confint(mm)
exp(-0.8918654)
exp(-0.8962603)
```

5.8

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
dcrawlyes<-subset(d,crawl==1)
mcm<-glm(dcrawlyes$erosion~dcrawlyes$pooltime, family = binomial)
summary(mcm)
# log odds(erosion) = 0.31741-0.07707*pooltime
exp(-0.07707)
#P-værdi aflæses til 0.754

confint(mcm)
exp(-0.5733895)
exp(0.4076656)
#Interval [0.5636118 til 1.503304]
```

5.9

```
d <- read.csv("http://causal.sund.ku.dk/f22/80.csv", header=TRUE, stringsAsFactors=TRUE)
dcrawno<-subset(d,crawl==0)
mim<-glm(dcrawno$erosion~dcrawno$pooltime, family = binomial)
summary(mim)
# log odds(erosion) = -0.7265-0.2479*pooltime
#P-værdi aflæses til 0.661

exp(-0.7265)
#e^b = 0.4835986

confint(mim)
exp(-1.434779)
exp(0.8895236)
#Interval [0.238168 til 2.43397]
```

5.10 – Ingen tilhørende R-kode