STATISTIK ORDINÆR EKSAMEN 2022

Delopgave 1; Datapræsentation

1. d <- read.csv("http://causal.sund.ku.dk/f22/8.csv", header=TRUE, stringsAsFactors=TRUE)

2. I dette datasæt tages der udgangspunkt i 131 svømmere, med følgene 5 variabler;

   Age, hvor man ser på alderen

   Crawl, der fortæller hvorvidt personen træner crawlsvømning (**1; Ja, 0; andre svømmeformer**)

   Pooltime**,** den gennemsnitlig tid brugt i svømmebassinet, vurderet i timer pr uge

   Soda, er den indtagelse af kulsyreholdige læskedrikke, hvor vi har intervallerne;
   - **Never** = Aldrig
   - **>5/Week** = mindre end 5 gange pr uge
   - **<=5/week** = mere eller lig med 5 gange pr uge
   - **NA** = Det har ikke været muligt at indsamle data

   Erosion**,** betegner hvorvidt der er ætsninger på tænderne **(1; Ja, 0; nej)**

   For at kunne opstille min Tabel 1, opdeles datasættet i hhv dem der træner crawlsvømning (1 = ja) og dem der træner andre svømmeformer (0 = Andre svømmeformer).

Crawl:
For at dele datasættet i to, bruges subset komandoen, i R-studio. Dette gøres, for at give et bedre overblik for datasættet.
Crawl defineres som en kategorisk variabel, der fortæller om hvorvidt man træner crawl svømning eller andre svømmeformer.
I datasættet ses der at 84 træner andre svømmeformer, mens 47 træner crawl svømning.

Da der er tale om en kategorisk variabel, kan vi bruge R studio til at beregne den relative frekvens for hver gruppe:

Relativ frekvens for crawl = 0: 0.6412214 = 64,1%

Relativ frekvens for crawl = 1: .3587786 = 35,9%

Ude fra disse to grupper, kan de resterende variabler beskrives, afhængigt af om de er kategoriske eller kontinuerte, således at man til slut kan opstille tabel 1.

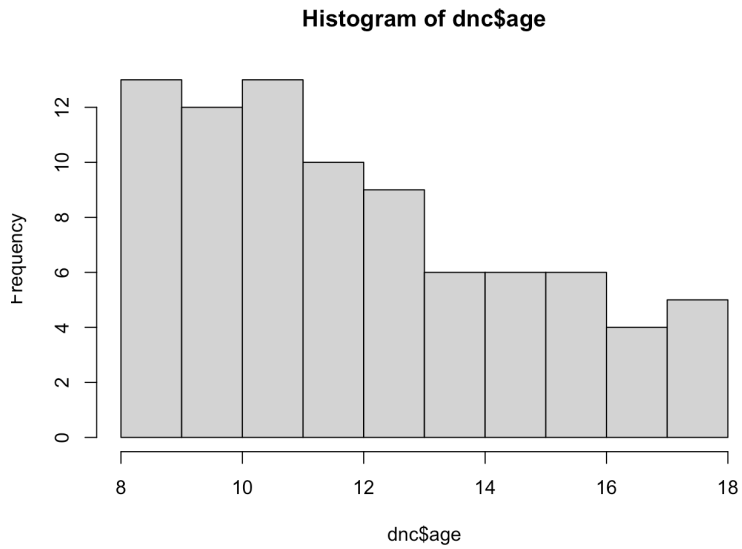
I tilfælde af at variablerne er kontinuerte, beskrives de enten med median og IQR eller middelværdi og spredningen.

I tilfælde af at variablerne er kategoriske, beskrives de med frekvenser og de relative frekvenser.

Age:
Variablen age er en kontinuerte variabel. For at finde ud af om man skal beregne median og IQR eller middelværdi og spredningen, skal der opstilles et histogram, hvorfra der skal vurderes hvorvidt Age er normalfordelt eller ikke.

I tilfælde af at der er tale om en normalfordeling, beregnes middelværdi og spredning.
I tilfælde af at der er tale om en asymetrisk fordeling, beregnes median og IQR.

Histogram over fordelingen af alder, hos personer der træner til andre svømmeformer:

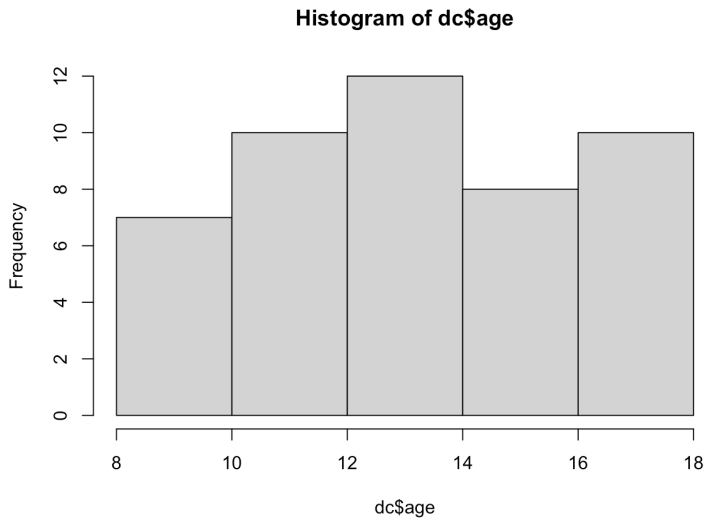


Figur 1: histogram over alder hos personer med andre svømmeformer

Histogrammet ses at være asymetrisk, ved at den blandt andet er forskudt til højre og betegnes derfor ikke som normalfordelt. Derfor beregnes medianen og IQR:

Median: 12
IQR: 4.25

Histogram over fordelingen af alder, hos personer der træner til crawl:



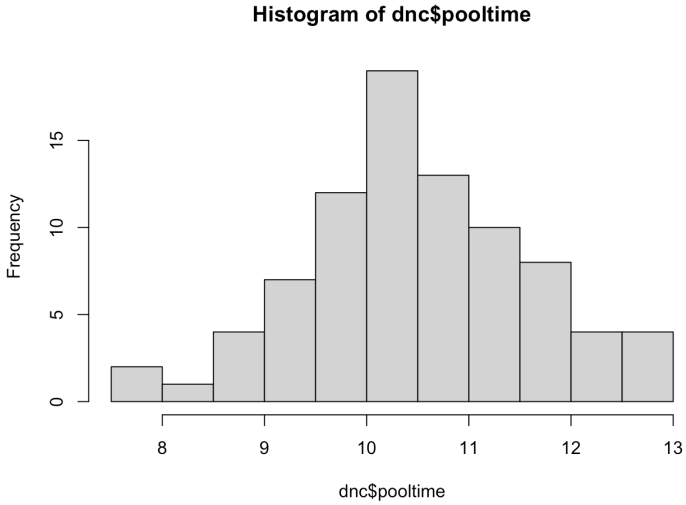
Figur 2; Histogram over fordelingen af alder hos personer der træner til crawl

Histogrammet ses at være asymetrisk, ved at den blandt andet er forskudt til højre og betegnes derfor ikke som normalfordelt. Derfor beregnes medianen og IQR:

Median: 13
IQR: 3.5

Pooltime:
Variablen Pooltime er en kontinuerte variabel, idet at der tages udgangspunkt i antal timer, pr uge. Der opstilles derfor et histogram, hvor der vurderes hvorvidt det er normalfordelt eller ikke.

Histogram over fordelingen af antal timer i poolen pr uge, hos personer der træner andre svømmeformer;

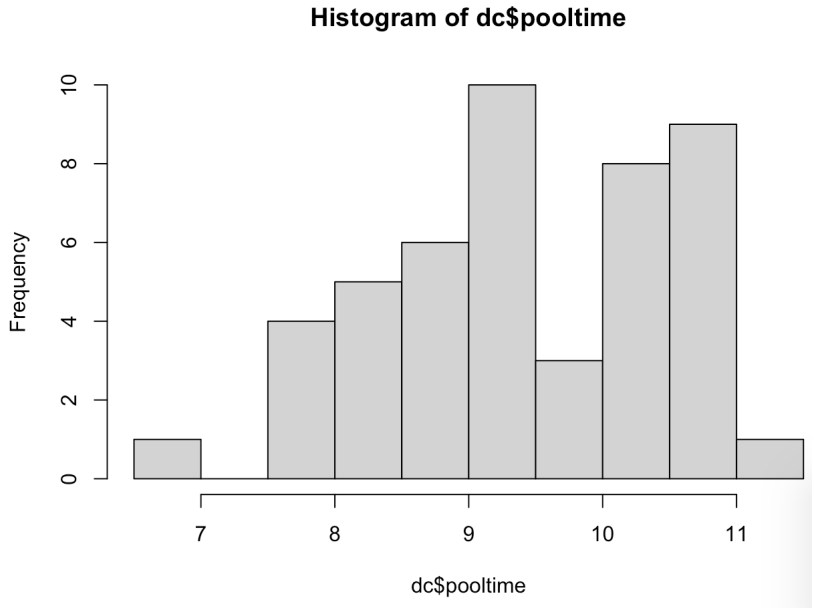


Figur 3: Histogram over fordeling af pooltime, hos personer der træner andre svømmeformer

Histogrammet ses for at være symmetrisk og dermed normalfordelt, hvortil middelværdi og spredningen beregnes:

Middelværdi: 10.50821
Spredning: 1.102852

Histogram over fordelingen af antal timer i poolen pr uge, hos personer der træner til crawl:



Figur 4: Histogram over fordeling af pooltimer hos personer der trænser crawl

Histogrammet ses at være asymetrisk, ved at den blandt andet er forskudt til venstre og betegnes derfor ikke som normalfordelt. Derfor beregnes medianen og IQR:

Median: 9.41
IQR: 1.71

Soda:
Soda er en kategorisk variabel, der fortæller om indtagelsen af læskedrikke. Til dette beregnes frekvensen for de forskellige værdier i R studio.
Der opstilles først en tabel, der fortæller fordelingen af antal personer ved hhv <5/week, >=5/week eller never.

Tabellen tager ikke udgangspunkt i personer under kategorien NA (Not available) og vi beregner derfra frekvensen, på baggrund af personerne tilhørende de 3 overnævnte kategorier. Der beregnes at (131-118) = 13 personer hører indenunder kategorien NA.

Tabel for personer der indtager læskedrik, og træner til andre svømmeformer;

<5/week; 32 personer
>=5/week; 41 personer
Never; 3 personer

Den relative frekvens beregnes;

<5/week; 0.4210526 = 42.1%
>=5/week; 0.5394737 = 53.9%
Never; 0.03947368 = 3.9%

Tabel for personer der indtager læskedrik, og træner til crawl;

5/week; 17 personer
>=5/week; 24 personer
Never; 1 personer

Den relative frekvens beregnes;

<5/week; 0.4047619 = 40,5%
>=5/week; 0.5714286 = 57,1%
Never; 0.02380952 = 2,4%

Erosion;
Erosion er en kategorisk variabel, der fortæller om hvorvidt man har ætsning på tænder. Har man ætsninger (1=ja) og ingen ætsninger svarer til (0=nej). Til dette beregnes frekvensen for de forskellige værdier i R studio.

Tabel over erosion hvis man træner andre svømmeformer:

0 (ingen erosion) = 74
1 (Erosion) = 10

Den relative frekvens beregnes;

0 (ingen erosion) = 0.8809524 = 88.1%
1 (Erosion) = 0.1190476 = 11.9%

Tabel over erosion hvis man træner crawl;

0 (ingen erosion) = 21
1 (Erosion) = 26

Den relative frekvens beregnes;

0 (ingen erosion) = 0.4468085 = 44.7%
1 (Erosion) = 0.5531915 = 55.3%

Tabel 1 kan du opstilles;

	Andre svømmeformer = 0	**Crawl svømning = ja**
Age	Median: 12 IQR: 4.25	Median: 13 IQR: 3.5
Pooltime	Middelværdi: 10.50821 Spredning: 1.102852	Median: 9.41 IQR: 1.71
Soda	<5/week; 0.4210526 = 42.1% >=5/week; 0.5394737 = 53.9% Never; 0.03947368 = 3.9%	<5/week; 0.4047619 = 40,5% >=5/week; 0.5714286 = 57,1% Never; 0.02380952 = 2,4%
Erosion	0 (ingen erosion) = 0.881 = 88.1% 1 (Erosion) = 0.119= 11.9%	0 (ingen erosion) = 0.447 = 44.7% 1 (Erosion) = 0.5531915 = 55.3%

Delopgave 2; Sammenhæng mellem crawlsvømning og alder

1. Det ses at vi har en kontinuerte uafhængig variabel (age), samt en binær afhængig variabel (crawl) og vi gør derfor brug af logistisk regression.

$$LogOdds(Crawl) = a + b \cdot age$$

$$LogOdds(Crawl) = -2.32274 + 0.13490 \cdot age$$

Hvor -2.322 er skæring med y aksen, og 0,134 er hældningen der stiger pr år.

2. For at finde 95% konfidensinterval, gøres der brug af R-studio, kommandoen 'confint', hvortil vi får at 95% konfidensintervallet er mellem [0.009153853; 0.2657367], altså er man 95% sikker på at intervallet mellem 0,0091 og 0,2657 indeholder den sande værdi.

   P-værdien vurderes på baggrund af vores estimeret model, hvor vi får en P-værdi svarende til 0.03827, altså er den under 5% (0,05) og man forkaster derfor nul-hypotesen. Det vil siges at man tilsvarende forkaster hypotesen om at der ikke er en sammenhæng mellem crawl svømning og alder, altså siges der at være en sammenhæng.

3. Udtrykket for den estimerede sammenhæng mellem crawlsvømning og alderen på odds-skalaen defineres som;

$$LogOdds(Crawl) = a + b \cdot age$$

   Hertil indsættes a og b værdien, fundet i opgave 2.1

$$LogOdds(Crawl) = -2.32274 + 0.13490 \cdot age$$

4. Udtrykket for den estimeret sammenhæng mellem crawlsvømning og alderen på sandsynlighedsskalaen defineres som;

$$P(crawl = 1) = \frac{e^{a+b \cdot age}}{1 + e^{a+b \cdot x}}$$

   Hertil indsættes a og b værdien, fundet i opgave 2.1

$$P(crawl = 1) = \frac{e^{-2.322+0.13490 \cdot age}}{1 + e^{-2.322+0.13490 \cdot age}}$$

5. Ved at bruge modellen i opgave 2.4, kan vi prædiktere sandsynligheden for crawlsvømning ved en alder på 9 år:

$$P(crawl = 1) = \frac{e^{-2.322+0.13490 \cdot 9}}{1 + e^{-2.322+0.13490 \cdot 9}}$$

   Ved beregning af dette, fås en værdi på 0.2482626, som svarer til 24,82%.
   Altså er den prædikteret sandsynlighed for at man tager til crawlsvømning 24,82% ved en alder på 9 år.

6. For at estimere risikoratioen for crawlsvømning, ved sammenligning mellem en person på 14 år og en person på 12 år, gør man brug af modellen i opgave 2.4:

$$P(crawl = 1) = \frac{\left(\frac{e^{-2.322+0.13490 \cdot 14}}{1+e^{-2.322+0.13490 \cdot 14}}\right)}{\frac{e^{-2.322+0.13490 \cdot 12}}{1+e^{-2.322+0.13490 \cdot 12}}}$$

Dertil fås følgene resultater:

$$P(crawl = 1) = \begin{cases} 0.3933147 \\ 0.3311031 \end{cases}$$

Der kan konkluderes at risikoen for at gå til crawl ved en alder af 14 år er 39,33% og ved en alder af 12 år ligger den på 33,11%

7. For at estimere hvor mange gange odds for crawlsvømning ændrer sig, når alderen stiger med 3 år, gør man brug af eksponentialfunktionen (exp) til hældningen fundet i opgave 2.1 (Hældning = 0.13490) og opløfter den med 3, således at vi får at;

$$exp(0.13490)^3 = 1.498853$$

Altså kan der konkluderes at når alderen stiger med 3 år, stiger oddsen for at gå til crawlsvømning med 1,498 som svarer til 149,8%.

8. For at prædiktere alderen, for hvornår sandsynligheden for crawlsvømning krydser 25%, tages der udgangspunkt i udtrykket for p(crawl=1), hvortil denne svarer til 25% altså 0,25.

Med vores data vil vi få følgene, hvor age isoleres;

$$0{,}25 = \frac{e^{-2.322+0.13490 \cdot age}}{1+e^{-2.322+0.13490 \cdot age}}$$

Dette beregnes og der fås at alderen hvor sandsynligheden for crawlsvømning krydser 25%, er lig med 9,41 altså er alderen lig med 9 år.

9. Parameteren intercept definerer skæring med y-aksen, hvilket svarer til en værdi på -2.322. Denne talværdi angiver værdien for hvorvidt et barn på 0 år tager til crawlsvømning. Denne værdi er negativ, og derfor vil der konkluderes at et barn på 0 år ikke vil gå til crawlsvømning.

Den tilhørende P-værdi til intercept svarer til 0.00759, hvilket gør at nulhypotesen forkastes. Nulhypotesen fortæller at der ikke er sammenhæng mellem alder og crawlsvømning og denne hypotese forkastes, da p værdien er under 5% signifikantniveauet.

Delopgave 3; Hvad påvirker tiden brugt i svømmebassin

1. For at undersøge hvorvidt tid brugt i svømmebassin pr uge, afhænger af crawl svømning gør man brug af T-test. Denne statistisk model benyttes, da der er tale om en kontinuerte afhængig variabel (pooltime), samt en binær uafhængig variabel (crawlsvømning). Da der er tale om to grupper, vurderes der at der skal bruges en uparret T-test.

2. På baggrund af den estimeret T-test, ses der en gennemsnitlig tidsværdi hos personer der træner til andre svømmeformer (0 = andre svømmeformer), svarende til 10.50 timer/uge. Gennemsnitstiden for personer der tager til crawl træning (1 = Crawl), svarer til 9,42 timer/uge. Forskellen mellem disse to svarer til 1,08 altså omkring 1 time pr uge. Det vil siges at personer der svømmer andre former end crawl, bruger ugentligt 1 time mere i gennemsnittet, end dem der svømmer crawl.

   T-testen viser et 95% konfidensinterval der ligger mellem 0.687 og 1,482, altså er man 95% sikker på at intervallet mellem [0.687; 1,482] indeholder den sande værdi.

   På baggrund af T-tesetn, bestemmes vores P-værdi til at være 4.496e-07, hvilket ligger under 5% (0,05) signifikantiniveauet og vi forkaster derfor nul hypotesen, der lyder på at der ikke er en sammenhæng mellem pooltiden og hvorvidt man går til crawl eller andre svømmeformer., altså ses der en sammenhæng mellem de to kattegorier.

3. For at beregne 95% referenceinterval for tiden brugt i svømmebassinet pr uge blandt crawlsvømmere, bruges følgene formel;

$$Referenceinterval = middelværdi +/- 2 \cdot spredning\ (SD)$$

   Spredning og middelværdien beregnes i R-studio og der fås følgene;

$$Referenceinterval = 9.422979 - 2 \cdot 1.097929 = 7.227121$$

$$Referenceinterval = 9.422979 + 2 \cdot 1.097929 = 11.61884$$

   Dermed kan der konkluderes at 95% referenceinterval for tiden brugt i svømmebassinet pr uge, vil ligge mellem værdierne [7,22; 11.61].

4. For at undersøge hvordan tid brugt i svømmebassinet pr uge, afhænger af alder gør man brug af en lineær regression. Man benytter denne model, da tiden (pooltime) er en kontinuerte afhængig variabel, der afhænger af alderen som også er kontinuerte.

5. Udtrykket for sammenhængen mellem tid brugt i svømmebasinnet og alderen, kan beskrives med en simpel lineær regression, angivet som:

$$Pooltime = a + b \cdot age + e$$

Hvor a svarer til skæringen med y aksen og b er dens tilsvarende hældning. 'e' defineres som værende fejlledet, der betegner afstanden mellem de observeret punkter og vores rette lineærlinje.

På baggrund af modellen estimeret i opgave 3.4 fås a (hældningen) og b (skæring med y-aksen):

$$Pooltime = 7.72340 + 0.18768 \cdot age + e$$

6. For at beregne 95% konfidensintervallet for effekten af alder på tid brugt i svømmebassin pr uge, bruges kommandoen confint inde i R-studio.
Dertil fås et 95% konfidensinterval mellem [0.1220722; 0.2532923], altså vil den sande værdi ligge mellem dette interval.

P-værdien svarer til 9.36e-08 (Derfor under 5%) og derfor forkastes nul hypotesen, der siger at der ikke er en sammenhæng mellem alder og tid brugt i svømmebassinet pr uge. Altså er der en sammenhæng mellem alder og tid brugt i svømmebassinet.

7. For at bruge modellen til at prædiktere den gennemsnitlige tid brugt i svømmebassin pr uge for en 13-årige, indsættes 13 ind i udtrykket, defineret i 3.5.

$$Pooltime = 7.72340 + 0.18768 \cdot 13$$

$$10.16 = 7.72340 + 0.18768 \cdot 13$$

Altså er den gennemsnitlige tid brugt i svømmebassinet pr uge for en 13-årig cirka lig med 10 timer ugentligt.

For at beregne 95% referenceinterval for tiden brugt i svømmebassinet pr uge, for en 13-årig bruges igen denne følgene formel;

$$Referenceinterval = middelværdi +/- 2 \cdot spredning\ (SD)$$

Spredningen fås ude fra modellen, fra opgave 3.4 (SD= 1,092) og middelværdien er beregnet overfor til en værdi af 10.16:

$$Referenceinterval = 10.16 - 2 \cdot 1.092 = 7.976$$

$$Referenceinterval = 10.16 + 2 \cdot 1.092 = 12.344$$

Dermed kan der konkluderes at 95% referenceinterval for tiden brugt i svømmebassinet pr uge, for en 13-årig vil ligge mellem værdierne [7,22; 11.61].

8. For at undersøge hvordan tid brugt i svømmebassinet, afhænger af både alder og crawlsvømning, gør man brug af multipel lineær regression, da tid er en kontinuerte afhængig variabel, mens crawlsvømning er uafhængig binær og alder er kontinuerte uafhængig variabel.

   For at finde ud af hvilken model der er mest passende til vores datasæt, tages der først udgangspunkt i modellen med hovedeffekt og interaktionsled der defineres i R-studio. Denne model forkastes, idet at vi får en p værdi = 0.26190 som er højre end 5% (0,05) og denne fortæller at der ikke er en sammenhæng, mellem tid brugt i svømmebassinet, alder og crawlsvømningen.

   Vi opstiller nu modellen hvor der alene tages udgangspunkt i hovedeffekten. I dette tilfælde får vi R studio til at opstille modellen med hovedeffekten.

   Regression for hovedeffekten opstilles:

   Opstilling af regression, for hvis man tager til crawl (Crawl = 1)

   - Hældningen, svarer til b_1
   - Skæringspunktet, svarer til $a_1 + b_2$

   Dertil opstilles regressionen:

   $$Pooltime = a_1 + b_1 \cdot age + b_2 \cdot 1\,(Crawl = 1) + e$$

   På baggrund af vores opstillet model i R-studio, kan man indsætte følgene værdier:

   $$Pooltime = 7.681 + 0.22 \cdot age - 1.336 \cdot 1\,(Crawl = 1) + e$$

   $$Pooltime = 7.681 - 1.336 + 0.22 \cdot age + e$$

   $$Pooltime = 6.345 + 0.22 \cdot age + e$$

Opstilling af regression, for hvis man ikke tager til crawl (andre svømmeformer = 0)

Dertil opstilles regressionen:

$$Pooltime = a_1 + b_1 \cdot age + b_2 \cdot 0\ (andre\ svømmeformer = 0) + e$$

På baggrund af vores opstillet model i R-studio, kan man indsætte følgene værdier:

$$Pooltime = 7.681 + 0.22 \cdot age - 1.336 \cdot 0(andre\ svømmeformer = 0) + e$$

$$Pooltime = 7.681 + 0.22 \cdot age + e$$

Vi kigger nu på p-værdien for denne model med kun hovedeffekt, hvortil der fås en P-værdi på 3.94e-13, hvilket vil sige at vi forkaster nul hypotesen om at der ikke er en sammenhæng og dermed bruger vi denne model med kun hovedeffekt.

9. De endelige modeller til denne kan nu opstilles som:

$$Pooltime = \begin{cases} 6.345 + 0.22 \cdot age + e & (Crawl = 1) \\ 7.681 + 0.22 \cdot age + e, & (Andre\ svømmeformer = 0) \end{cases}$$

For gruppe crawl = 1, vil man kunne se at hældningen er 0,22, som er den gennemsnitlig positive ændring i pooltiden, når alderen stiger med 1 år.
Desuden ses en y værdi på 6,345 som er skæring med y aksen.

For gruppe andre svømmeformer = 0, vil man se at hældningen er den samme som for dem der tager til crawl, altså vil pooltiden stige med 0,22, når alderen stiger med 1 år.
Y værdien for dem der der går til andre svømmeformer end crawl, er højre end for dem der går til crawl og ligger på 7,681, altså har dem der går til andre svømmeformer end crawl en startværdi der er 1.336 højre end dem der går til crawl.

10. For at beregne 95% konfidensinterval, gør vi brug af confint funktionen i R-studio, hvortil vi får værdierne [-1.663; -1.009], hvilket vil siges at man med 95% sikkerhed kan sig at den sande værdi ligger mellem -1.663 og -1.009.

11. I outputtet fra vores endelig model med kun hovedeffekt, ses der at der for age er en p-værdi svarende til 1.28e-13, som er under 5% signifikantniveauet og man forkaster derfor nul hypotesen om at der ikke er en sammenhæng mellem alderen og pool tiden og derfor konkluderer man at der er en sammenhæng mellem disse to variabler.

For dem der går til crawl, fås der en p-værdi på 3.94e-13, hvilket igen er en del under 5% signifikantniveauet og man forkaster derfor nul hypotesen om at der ikke skulle være en sammenhæng mellem crawl og pooltiden. Derfor konkluderes der at der er en sammenhæng mellem pooltiden og om man går til crawl.

12. For at prædiktere den gennemsnitlige tid i svømmebassinet for en person på 12 år der svømmer crawl, gøres der brug af følgene model:

$$pooltime = 6.345 + 0.22 \cdot age + e$$

Hvor der indtastes 12 år, på ages plads:

$$pooltime = 6.345 + 0.22 \cdot 12 = 8.985$$

Altså kan der konkluderes at en person på 12 år, der svømmer crawl, bruger i gennemsnittet omkring de 9 timer i svømmebassinet.

13. For at prædiktere den gennemsnitlige tid i svømmebassinet for en person på 12 år der ikk svømmer crawl, gøres der brug af modellen for dem der svømmer andet end crawl:

$$pooltime = 7.681 + 0.22 \cdot age + e,$$

Hvor 12 indsættes på ages plads:

$$pooltime = 7.681 + 0.22 \cdot 12 = 10.321$$

Altså kan der konkluderes at en person på 12 år, der svømmer andet end crawl, bruger i gennemsnittet omkring de 10 timer i svømmebassinet.
Dermed bruger en person på 12 år i gennemsnittet mere tid i svømmebassinet, hvis personen går til andre svømmeformer end crawl.

14. For at finde den gennemsnitlig ændring i tid i svømmebassinet hen over 4 år for en person der ikke svømmer crawl, tages hældningen (0,22) og ganges med 4 år.
Dette svarer til 0,88, altså stiger tiden i svømmebassinet med 0,88 i løbet af 4 år.

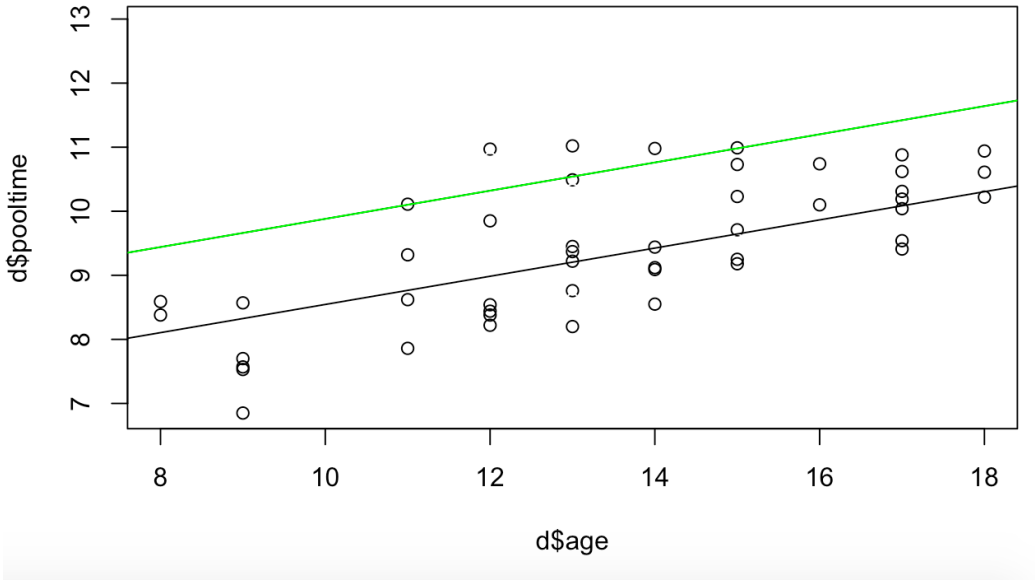
Tilsvarende beregnes tiden over 4 år for en person der svømmer crawl. Hertil tages hældningen (0,22), hvortil den ganges med 4 og der fås samme tal, altså stiger den med 0,88 hen over 4 år.

Stigningen er det samme, hvilket vil siges at der ikke er forskel på den gennemsnitlig ændring i tid i svømmebassinet, mellem personer der svømmer og ikke svømmer crawl. Årsagen til dette

er, at de begge har samme hældning, hvilket vil sige at de begge stiger med samme tid, når alderen stiger med 1 år.

15. For at lave en passende figur, plottes modellen i R-studio, hvortil vi har alder på x-aksen og pooltime på y-aksen.
    På baggrund af vores model, kan vi ved brug af de estimeret værdier og kommandoen abline, få to parallele linjer for dem der går til crawl og dem der går til andre svømmeformer.

    Grøn graf = Dem der går til andre svømmeformer
    Sort graf = Dem der går til crawl



Figur 5; Grafer over Pooltime, hvis man går til crawl og hvis man går til andre sportsgrene

Delopgave 4; Indtagelse af kulsyreholdige læskedrikke

1. For at beregne sandsynligheden for den indtaget kulsyreholdige læskedrikke mindre end 5 gange per uge, bruges kommandoen table i R-studio, for at give os en tabel over antal personer der aldrig drikker kulsyreholdige læskedrikke (never), dem der drikker det mindre end fem gange pr uge (<5/week) og dem der drikker mere eller lig med 5 gange pr uge (>=5/week). Desuden ses også en kategori (NA), der definere at det ikke har været muligt at indsamle data og derfor ses der bort for denne kategori.

    Ved at bruge table kommandoen i R-studio, fås at 49 ud af 118 drikker syreholdige læskedrikke mindre end 5 gange pr uge. Der kan dertil beregnes sandsynligheden:

$$\frac{49}{118} = 0.4152542 = 41{,}5\%$$

Altså er sandsynligheden for at indtage kulsyreholdige læskedrikke mindre end 5 gange per uge, omkring de 41,5%.

For at finde 95% konfidensintervallet, bruges kommandoen prop.test i R.studio, der giver en 95% konfidensinterval mellem [0.3264135 ; 0.5096892], altså er man 95% sikker på at intervallet mellem 0.3264 og 0,5096 indeholder den sande værdi.

2. Odds defineres som værende forholdet mellem antal personer der oplever hændelsen, divideret med antal personer der ikke oplever det, altså fås odds som:

Odds (A) = personer der oplever A / personer der ikke oplever A.

$$\frac{49}{(65+4)} = 0.7101449$$

Dertil fås en odds på 0,71 som fortæller at for hver gang en person ikke oplever hændelsen (altså drikker over 5 gange om ugen, eller aldrig drikker syreholdige læskedrikke), er der 0,71 odds for at en person drikker syreholdige læskedrikke mindre end 5 gange om ugen.

3. Risikodifferensen er forskellen i risiko, blandt de to grupper vi har at gøre med. I vores tilfælde, er der tale om følgene to grupper:

Gruppe 1: Dem der drikker mere end 5 eller lig med 5 gange pr uge og svømmer crawl
Gruppe 2: Dem der drikker mindre end 5 gange om ugen og svømmer crawl.

Vi kan udregne denne ved at sige:

$$Risikodifferens = Gruppe\ 1 - Gruppe\ 2$$

Ved at bruge table funktionen inde i R-studio, vil vi få værdier svarende til gruppe 1 og gruppe 2. Vi kan nu beregne risikoen for de to grupper:

Gruppe 1:

$$\frac{24}{(17+24)} = 0.5853659 = 58{,}53\%$$

Gruppe 2:

$$\frac{17}{(17+24)} = 0.4146341 \ = \ 41{,}46\%$$

Risikodifferensen for de to grupper beregnes:

$$0{,}585 - 0{,}414 \ = \ 0.170 \ = \ 17\%$$

Dermed kan man konkludere at risikoen for at en person der svømmer crawl vil drikke læskedrikke mere end 5 gange pr uge er ca 17 procent højre, end risikoen for at en person der svømmer crawl vil drikke syreholdige læskedrikke mindre end 5 gange pr uge.
Hvis de to grupper havde samme risiko for hvor meget de ville drikke syreholdige læskedrikke, ville den sande risikodifferens være lig med 0%.

Den relative risiko (risikoratio), betegnes som forholdet mellem risikoerne i begge grupper, altså kan den formuleres som:

$$\frac{Gruppe\ 1}{Gruppe\ 2} = risikoratio$$

Overfor har vi beregnet risikoerne for de to grupper, som nu sættes ind i formlen:

$$\frac{0{,}585}{0{,}414} = 1.411765 \ = \ 141{,}1\%$$

Altså kan der konkluderes at risikoen for at en person der svømmer crawl, indtager læskedrikke mere eller lig med 5 gange om ugen, er 1,4 gange så høj som risikoen for at svømme crawl og drikke læskedrikke mindre end 5 gange pr uge.
Hvis de to grupper havde samme risiko for hvor meget læskedrikke de drak, ville den sande risikoratio være lig med 1.

4. For at beregne odds for crawlsvømning, blandt personer der aldrig indtager kulsyreholdige læskedrikke i forhold til personer der indtager mindre end 5 gange pr uge, bruges tabellen lavet i opgave 4.4.

Oddsratio er defineret som den relative forskel mellem odds i vores to grupper. Den kan derfor beregnes som

$$\frac{Odds\ Gruppe\ 1}{Odds\ Gruppe\ 2} = Oddsratio$$

Gruppe 1: Dem der aldrig drikker syreholdige læskedrikke, men svømmer crawl.
Gruppe 2: Dem der drikker mindre end 5 gange om ugen og svømmer crawl.

Beregning af Odds gruppe 1:

$$\frac{1}{17} = 0.05882$$

Beregning af Odds gruppe 2:

$$\frac{17}{1} = 17$$

Beregning af odds ratio:

$$\frac{0{,}058}{17} = 0.003458824$$

Derfra er odds ratioen, som er den relative sammenligning af odds mellem de to grupper beregnet til at være 0,00345.

Delopgave 5: Hvad påvirker risikoen for ætsninger på tænder?

1. For at opstille en statistisk model til at undersøge hvordan crawlsvømning påvirker risikoen for ætsninger af tænderne, tages der udgangspunkt i at erosion er en binær afhængig variabel, der afhænger af crawlsvømning der er en uafhængig binær variabel. Til dette bruges X^2-test (chi i anden test).

   For at benytte sig af chi i anden test, starter man med at opstille en matrix, der er benævnt model.5. Denne matrix opstilles på baggrund af om man går til crawl eller ikke, samt hvorvidt man derefter har erosion eller ikke. Derefter opstilles modellen ved brug af chisq.test i R-studio, der definerer en chi i anden test.

2. Der er opstillet en nulhypotese, der lyder på at der ikke er en sammenhæng mellem erosion og hvorvidt man går til crawl eller andre svømmeformer. På baggrund af vores chi i anden test, kan man ved at kigge på p værdien (p = 2.825e-07), se at p værdien er under 5% signifikantniveauet. Derfor forkaster man nul hypotesen og man kan på baggrund af dette konkludere at der er sammenhæng mellem hvorvidt man svømmer crawl eller ikke og erosion.

3. Risikodifferensen er forskellen i risiko, blandt de to grupper vi har at gøre med. I vores tilfælde, er der tale om følgene to grupper:

Gruppe 1: Ætsninger på tænderne hos crawlsvømmer
Gruppe 2: Ætsninger på tænderne hos ikke crawlsvømmer

Gruppe 1:

$$\frac{26}{(26+21)} = 0.5531915 = 55{,}31\%$$

Gruppe 2:

$$\frac{10}{(10+74)} = 0.1190476 = 11.90\%$$

Risikodifferensen for de to grupper beregnes:

$$0{,}553 - 0{,}119 = 0.4341439 = 43{,}41\%$$

Dermed kan man konkludere at risikoen for at en person har ætsninger på tænderne når man er crawlsvømmer er 43% højre, end risikoen for erosion hos en person der svømmer andre svømmeformer en crawl.

4. Den relative risiko (risikoratio), betegnes som forholdet mellem risikoerne i begge grupper, altså kan den formuleres som:

$$\frac{Gruppe\ 1}{Gruppe\ 2} = risikoratio$$

Overfor har vi beregnet risikoerne for de to grupper, som nu sættes ind i formlen:

$$\frac{0{,}5531}{0{,}1190} = 4.646809 = 464\%$$

Dermed kan der konkluderes at risikoen for at en person der svømmer crawl, har erosion, er 4,64 gange højere, end risikoen for at få erosion hos en der træner andre svømmeformer.

5. For at beregne odds ratioen, når man sammenligner odds for ætsninger på tænder mellem crawl svømmere i forhold til ikke crawlsvømmer, ses der på tabellen angivet i opgave 5.3.

Beregning af Odds gruppe 1 (crawlsvømmer):

$$\frac{26}{21} = 1.238095$$

Altså for hver gang der er 1 som træner crawl, ikke har erosion, er der 1,23 odds for at en person der svømmer crawl, har erosion

Beregning af Odds gruppe 2 (ikke crawlsvømmer):

$$\frac{10}{74} = 0.1351351$$

Altså for hver gang der er 1 person som har andre svømmeformer end crawl ikke har erosion, er der 0,135 odds for at en person der ikke svømmer crawl, har erosion.

Beregning af odds ratio:

$$\frac{1.23}{0{,}135} = 9.11$$

Dermed kan der konkluderes at oddsen er 9,11 for at en crawl svømmer får erosion, sammenlignet med en person der ikke svømmer crawl.

6. For at undersøge hvordan erosion afhænger af tid brugt i svømmehallen, gør vi brug af logistisk regression. Denne model benyttes, da der tages udgangspunkt i en afhængig binær variabel (erosion) og en uafhængig kontinuerte variabel (pooltime).

$$LogOdds(erosion) = a + b \cdot pooltime$$

7. Ved brug af tallene fundet i R-studio i delopgave 5.6, kan man aflæse en a og b værdi.
Skæring med y aksen, altså a-værdien aflæses til at være 3.0968, hvortil vi har en hældning og dermed en b værdi, svarende til -0.4071.
Disse værdier indsættes nu i vores logodds model:

$$LogOdds(erosion) = 3.0968 - 0{,}4071 \cdot pooltime$$

På baggrund af 95% konfidensinterval, fås en værdi mellem [-0.7544714; -0.08090175], altså er man 95% sikker på at intervallet mellem -0,754 og -0,08 indeholder den sande værdi.

Nul hypotesen lyder på at der ikke er en sammenhæng mellem tiden brugt i badebassinet og erosion. Denne nulhypotese forkastes i og med at vi har en p-værdi på 0.017, altså er den under 5% signifikantniveauet (0,05).

8. Vi estimerer nu en model, der fortæller sammenhængen mellem tid brugt i bassinet og erosion, når man svømmer crawl.

$$LogOdds(erosion) = -0.001233 + 0.022800 \cdot pooltime$$

Hvor 0.022800 er defineret som værende hældning der stiger som følge af at pooltime stiger, og -0.001233 er skæringen med y-aksen.

95% konfidensintervallet er givet ved [-0.5142579; 0.5599882], hvilket vil siges at vi er 95% sikker på at dette interval indeholder den sande værdi.

Den testede nulhypotese, siger at der ikke er en sammenhæng mellem pooltime og erosion, når man svømmer crawl. Modellen har givet en p-værdi på 0.933, som er større end 5% (0,05) og dermed forkastes nul-hypotesen ikke og der konkluderes derfor at der ikke er en sammenhæng mellem erosion og pooltime, når man går til crawl.

9. Igen, indsættes værdierne ind i Logodds:

$$LogOdds(erosion) = -0.8150 \ - 0.1135 \ \cdot pooltime$$

Hvor -0.1135 definerer hældningen og -0.8150 definerer skæring med y-aksen.

Den tilhørende 95% konfidensinterval er givet ved værdierne [-0.7278372; 0.4936247], hvilket vil siges at vi er 95% sikker på at dette interval indeholder den sande værdi.

Den testede nul-hypotese fortæller at der ikke er en sammenhæng mellem pooltime og erosion, når man træner andre, svømmeformer en crawl. Da vores p-værdi er 0.712 (og dermed over 5%), forkastes nul hypotesen ikke.

10. I delopgave 5.7, konkluderes der at der er en sammenhæng mellem pooltime og erosion, idet at nulhypotesen forkastes da P-værdien ses til at være under 5% signifikantniveauet.

I delopgave 5.8 undersøges hvorvidt crawlsvømning har en indflydelse på erosionens sammenhæng med pooltime. Der vises hertil en p-værdi over 5% signifikantniveauet, hvilket svare til at nulhypotesen ikke forkastes, og der dermed konkluderes at der ikke er en sammenhæng mellem pooltiden og erosion, hvis man er crawl svømmer.

I delopgave 5.9 undersøges der hvorvidt personer der har andre svømmeformer end crawl, har en indflydelse på sammenhængen mellem erosion og pooltime. Nul-hypotesen fortæller at der

ikke er en sammenhæng og denne forkastes ikke, da p-værdien blev fundet værende over 5% signifikantniveauet. Derfor ville man konkludere at der ikke er en sammenhæng.

For at svare på den stillede hypotese i opgavebeskrivelsen, kan man på baggrund af delopgave 5.8 og 5.9 konkludere at de forskellige svømmeformer (crawl eller ikke crawl) ikke har en indflydelse på risikoen for at få erosion, idet at vi ikke forkaster vores nul hypoteser, efter opstilling af en logistisk regression og der dermed kan konkluderes at der ikke er en sammenhæng.

Appendix:

```
d <- read.csv("http://causal.sund.ku.dk/f22/8.csv", header=TRUE, stringsAsFactors=TRUE)
```

Opgave 1.2

```
dc<-subset(d,crawl=="1")
dnc<-subset(d,crawl=="0")

table(d$crawl)
84/(47+84)
47/(47+84)

hist(dnc$age)
median(dnc$age)
IQR(dnc$age)

hist(dc$age)
median(dc$age)
IQR(dc$age)

hist(dnc$pooltime)
mean(dnc$pooltime)
sd(dnc$pooltime)

hist(dc$pooltime)
median(dc$pooltime)
IQR(dc$pooltime)

table(dnc$soda)
32/76
41/76
```

3/76

table(dc$soda)
17/42
24/42
1/42

table(dnc$erosion)
74/84
10/84

table(dc$erosion)
21/47
26/47

Opgave 2.1

model.1 <- glm(d$crawl ~ d$age, family = binomial)
summary(model.1)
confint(model.1)

Opgave 2.5
exp(-2.322+0.13490*9)/(1+exp(-2.322+0.13490*9))

Opgave 2.6
exp(-2.322+0.13490*14)/(1+exp(-2.322+0.13490*14))
exp(-2.322+0.13490*12)/(1+exp(-2.322+0.13490*12))

opgave 2.7
exp(0.13490)^3

Opgave 3.1;
t.test(d$pooltime~d$crawl)

Opgave 3.3:
sd(dc$pooltime)
mean(dc$pooltime)

Opgave 3.4:
model.2 <- lm(d$pooltime ~ d$age)
summary(model.2)

Opgave 3.6

```
confint(model.2)
```

Opgave 3.8

```
model.3<- lm(d$pooltime~ d$age * d$crawl)
summary(model.3)

model.4<- lm(d$pooltime~ d$age + d$crawl)
summary(model.4)
```

Opgave 3.10

```
confint(model.4)
```

Opgave 3.15

```
plot(d$age,d$pooltime,col=d$crawl)
abline(6.345,0.22)
abline(7.681,0.22,col="green")
```

Opgave 4.1

```
table(d$soda)
prop.test(49,(65+49+4))
```

Opgave 4.3

```
table(dc$soda)
```

Opgave 5.1

```
dc<-subset(d,crawl=="1")
dnc<-subset(d,crawl=="0")
sum(dc$erosion=="0")
sum(dc$erosion=="1")
sum(dnc$erosion=="0")
sum(dnc$erosion=="1")
model.5 <- matrix(c(21,26,74,10), nrow = 2, ncol = 2)
table(model.5)
chisq.test(model.5)
```

Opgave 5.3

```
table(dc$erosion)
table(dnc$erosion)
```

Opgave 5.6

```
model.6 <- glm(d$erosion ~ d$pooltime, family = binomial)
summary(model.6)
confint(model.6)
```

Opgave 5.8

```
model.7 <- glm(dc$erosion ~ dc$pooltime, family = binomial)
summary(model.7)
confint(model.7)
```

Opgave 5.9

```
model.8 <- glm(dnc$erosion ~ dnc$pooltime, family = binomial)
summary(model.8)
confint(model.8)
```