Statistik for Odontologer - Eksamensopgave 2022

Andreas Kryger Jensen og Claus Thorn Ekstrøm

48 timers hjemmeopgave
1. april kl. 15:00 til 3. april kl. 15:00

— Dette er en individuel eksamen. Det er ikke tilladt at arbejde sammen med andre om løsningen af denne opgave i eksamensperioden. —

Eksamensbesvarelsen skal skrives som en rapport, der besvarer de stillede spørgsmål. Der er følgende strukturelle krav til besvarelsen:

- Besvarelsen af hvert spørgsmål **skal** nummereres i overensstemmelse med nummereringen i dette dokument.
- Gentag **ikke** opgaveteksterne/spørgsmålene i dine besvarelser. Det udløser KUs automatiske plagiatkontrol, og din besvarelse vil dermed automatisk blive markeret som at være afskrevet fra dine medstuderende.
- Rapporten må **ikke** indeholde R-kode men skal formuleres i ord, hvor de relevante resultater fra analyserne i RStudio indsættes.
- Indsæt **ikke** store klumper af output fra RStudio som f.eks. `summary`-output fra regressionsmodeller og lignende, men udtræk de relevante talværdier i forhold til opgavespørgsmålet, og diskutér deres betydninger i ord.
- Den komplette R-kode, som du har brugt til dine analyser, **skal** vedlægges rapporten i form af et appendiks. Se løsningsforslag til øvelsesopgaverne fra SAU 7 og 8 for hvordan dette struktureres.
- Alle analyseresultater skal kunne reproduceres ved at køre din vedlagte R-kode på dit personlige datasæt. Besvarelser, der **ikke** kan genfindes som et resultat af en eller flere R-kommandoer eller simple beregninger, godtages **ikke** som en opgavebesvarelse.
- Hver side i afleveringen **skal** i nederste højre hjørne indeholde fortløbende sidenummer samt det totale antal sider. Der er ingen øvre eller nedre grænse for antal afleverede sider.
- Afleveringen (rapporten samt appendiks) uploades som ét samlet dokument i PDF-format til Digital Eksamen.

— Dette er en individuel eksamen. Det er ikke tilladt at arbejde sammen med andre om løsningen af denne opgave i eksamensperioden. —

Opgavens indhold

Den videnskabelige problemstilling, som du skal beskæftige dig med i denne opgave, tager udgangspunkt i et spørgsmål om risikoen for ætsninger på tænderne blandt personer, der regelmæssigt svømmetræner. Svømmebassiner bliver tilføjet klor af sanitære hensyn enten i form af hypoklorid eller klor i gasform. I større svømmebassiner er klorgas den foretrukne metode, men i modsætning til tilsætning af hypoklorid, som får vandet til at blive basisk, så får denne metode vandet til at blive surt. Dette er et resultat af den kemiske ligevægtsligning $Cl_2 + H_2O \rightleftharpoons HOCl + HCl$, hvor HCl (saltsyre) er et uønsket biprodukt. Det er standard praksis, at surheden af vandet regelmæssigt kontrolleres og justeres gennem tilføjelse af natriumkarbonat (Na_2CO_3), men hvis dette fejler, kan vandet opnå en surhedsgrad, der kan nedbryde tandemaljen.

Ud fra denne forståelse er det en videnskabelige hypotese, at tiden brugt i et svømmebassin med surt vand vil kunne lede til en øget risiko for ætsninger på tænderne, men også at forskellige svømmeformer kan have en indflydelse på denne risiko, da indtagelse af vand i mundhulen under svømmetræning kan afhænge af svømmeformen.

Dit *personlige* datasæt kan indlæses i RStudio ved at køre nedenstående kommando. Det er et **krav**, at din individuelle besvarelse er baseret på netop dit personlige datasæt.

```
d <- read.csv("http://causal.sund.ku.dk/f22/63.csv", header=TRUE, stringsAsFactors=TRUE)
```

Dit personlige datasæt indeholder data fra 131 svømmere med følgende 5 variable:

- `age`: alder i år
- `crawl`: hvorvidt der primært trænes crawlsvømning (`1`: ja, `0`: andre svømmeformer)
- `pooltime`: gennemsnitlige tid brugt i svømmebassin [timer/uge]
- `soda`: indtagelse af kulsyreholdige læskedrikke (`never`: aldrig, `<5/week`: mindre end fem gange pr uge, `>=5/week`: mere end eller lig med fem gange pr uge, `NA`: (Not Available) det har ikke været muligt at indsamle data)
- `erosion`: ætsninger på tænderne (`1`: ja, `0`: nej)

Eksamensopgaven er en besvarelse af de nedenstående opgaver og spørgsmål.

1 Datapræsentation

1. Besvar dette spørgsmål ved at skrive filnavnet på dit personlige datasæt på samme måde, som det er angivet ovenfor i kommandoen, der viser, hvordan du skal indlæse det i RStudio. Det skal du gøre, så det tydeligt fremgår af din aflevering, at det er det korrekte datasæt, som du anvender til løsningen af din individuelle eksamensopgave.

2. Lav en præsentation af studiepopulationen i form af en "Tabel 1" på baggrund af observationerne i stikprøven. Argumentér for, hvorfor du har valgt at strukturere tabellen på den måde, som du har gjort, samt hvorfor du har valgt at beskrive de enkelte variable i tabellen på den måde, som du har gjort.

2 Sammenhæng mellem crawlsvømning og alder

1. Brug en relevant statistisk model til at undersøge, hvordan sandsynligheden for crawlsvømning afhænger af alder. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, samt en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål.
2. Angiv den estimerede effekt for alder i modellen, dens tilhørende 95% konfidensinterval, den i outputtet tilhørende p-værdi, og forklar den tilhørende nulhypotese og betydningen i ord med henblik på formidling af en videnskabelig konklusion.
3. Skriv udtrykket for den estimerede sammenhæng mellem crawlsvømning og alder på odds-skalaen.
4. Skriv udtrykket for den estimerede sammenhæng mellem crawlsvømning og alder på sandsynlighedsskalaen.
5. Brug modellen til at prædiktere sandsynligheden for crawlsvømning ved en alder på 9 år.
6. Brug modellen til at estimere risikoratioen for crawlsvømning, når man sammenligner en person på 14 år med en på 12 år.
7. Brug modellen til at estimere, hvor mange gange odds for crawlsvømning ændrer sig, når alder stiger med 3 år.
8. Brug modellen til at prædiktere alderen, hvor sandsynligheden for crawlsvømning først krydser 25%.
9. I modellens output findes et estimat for parameteren navngivet `(Intercept)`. Forklar dette estimats betydning i ord, og hvordan det skal fortolkes på odds-skalaen. Kig desuden på den tilhørende p-værdi, og forklar hvilken nulhypotese, der bliver testet, hvordan den skal fortolkes, samt den tilhørende videnskabelige konklusion.

3 Hvad påvirker tiden brugt i svømmebassin?

1. Brug en relevant statistisk model til at undersøge, hvordan tid brugt i svømmebassin pr uge afhænger af crawlsvømning. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, samt en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål.
2. Angiv den gennemsnitlige forskel i tid brugt i svømmebassin pr uge for crawlsvømmere i forhold til ikke-crawlsvømmere, 95% konfidensintervallet for denne forskel, den relevante nulhypotese for forskellen med tilhørende p-værdi, samt en beskrivelse af den videnskabelige konklusion.
3. Beregn et 95% referenceinterval for tid brugt i svømmebassin pr uge blandt crawlsvømmere, og giv en fortolkning af dette interval.
4. Brug en relevant statistisk model til at undersøge, hvordan tid brugt i svømmebassin pr uge afhænger af alder. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, samt en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål.

5. Skriv udtrykket for den estimerede sammenhæng, og giv en fortolkning af parameterestimaterne.
6. Beregn et 95% konfidensinterval for effekten af alder på tid brugt i svømmebassin pr uge, og skriv en fortolkning i ord. Afrapportér desuden den af RStudio testede nulhypotese, dens fortolkning, den tilhørende p-værdi samt den videnskabelige konklusion.
7. Brug modellen til at prædiktere den gennemsnitlige tid brugt i svømmebassin pr uge for 13-årige samt et 95% referenceinterval for de 13-årige.
8. Brug en relevant statistisk model til at undersøge, hvordan tid brugt i svømmebassin samtidigt afhænger af både alder **og** crawlsvømning, og undersøg samtidig, om sammenhængen mellem tid brugt i svømmebassin og alder også er påvirket af crawlsvømning. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål, samt hvordan du er kommet frem til dit endelige resultat.
9. Skriv et udtryk for de estimerede sammenhænge baseret på din endelige model fra forrige spørgsmål, og giv en fortolkning af modellens parametre.
10. Beregn et 95% konfidensinterval for effekten af crawlsvømning i din endelig model, og giv en fortolkning i ord.
11. Betragt p-værdierne i outputtet fra din endelige model, fortolk deres nulhypoteser i ord, angiv de tilhørende p-værdier, og beskriv de videnskabelige konklusioner.
12. Brug din endelige model til at prædiktere den gennemsnitlige tid i svømmebassin for en person på 12 år, der svømmer crawl.
13. Brug din endelige model til at prædiktere den gennemsnitlige tid i svømmebassin for en person på 12 år, der *ikke* svømmer crawl.
14. Beregn ud fra din model den gennemsnitlige ændring i tid i svømmebassin hen over 4 år for en person, der *ikke* svømmer crawl. Er den det samme for en person, der svømmer crawl? Hvis ja: Forklar hvorfor. Hvis nej: Forklar hvorfor ikke.
15. Lav en passende figur baseret på din model, der illustrerer, hvordan tid i svømmebassin afhænger af alder og crawlsvømning.

4 Indtagelse af kulsyreholdige læskedrikke

1. Estimer sandsynligheden for at indtage kulsyreholdige læskedrikke mindre end fem gange pr uge samt dens tilhørende 95% konfidensinterval, og skriv en fortolkning af resultaterne.
2. Beregn odds for at indtage kulsyreholdige læskedrikke mindre end fem gange pr uge, og skriv en fortolkning af resultatet.
3. Beregn både risikodifferensen samt den relative risiko for at svømme crawl, der sammenligner personer, der indtager kulsyreholdige læskedrikke mindre end fem gange pr uge, med personer, der indtager kulsyreholdige læskedrikke mere end eller lig med fem gange pr uge, og skriv en fortolkning af de to estimater.

4. Beregn oddsratioen, der sammenligner odds for crawlsvømning blandt personer, der aldrig indtager kulsyreholdige læskedrikke i forhold til personer, der indtager kulsyreholdige læskedrikke mindre end fem gange pr uge, og fortolk værdien.

5 Hvad påvirker risikoen for ætsninger på tænderne?

1. Brug en relevant statistisk model til at undersøge, hvordan crawlsvømning påvirker risikoen for ætsninger på tænderne. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, samt en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål.
2. Hvad er din videnskabelige konklusion ud fra analysen, og hvad er din statistiske argumentation for at komme frem til denne konklusionen?
3. Beregn risikodifferensen, når man sammenligner ætsninger på tænderne blandt crawlsvømmere i forhold til ikke-crawlsvømmere, og giv en fortolkning i ord.
4. Beregn den relative risiko (risikoratioen), når man sammenligner risikoen for ætsninger på tænderne blandt crawlsvømmere i forhold til ikke-crawlsvømmere, og giv en fortolkning i ord.
5. Beregn oddsratioen, når man sammenligner odds for ætsninger på tænderne blandt crawlsvømmere i forhold til ikke-crawlsvømmere, og giv en fortolkning i ord.
6. Brug en relevant statistisk model til at undersøge, hvordan tid brugt i svømmebassin påvirker risikoen for ætsninger på tænderne. Dit svar på dette spørgsmål skal være hvilken statistisk model, du har anvendt, samt en argumentation for, hvorfor du har valgt netop denne til at besvare det videnskabelige spørgsmål.
7. Skriv udtrykket for din estimerede model på odds-skalaen, og giv en fortolkning af den estimerede effekt af tid brugt i svømmebassin sammen med et 95% konfidensinterval, den testede nulhypotese, p-værdien, og den tilhørende videnskabelige konklusion.
8. Gentag den forrige analyse (hvordan tid brugt i svømmebassin påvirker risikoen for ætsninger på tænderne), men hvor du nu *kun* udfører den blandt personer, der svømmer crawl. Fortolk den estimerede sammenhæng mellem ætsninger på tænderne og tid brugt i svømmebassin, angiv et tilhørende 95% konfidensinterval for effekten, beskriv den testede nulhypotese, p-værdien og den videnskabelige konklusion.
9. Gentag igen analysen, men hvor du nu *kun* udfører den blandt personer, der *ikke* svømmer crawl. Fortolk den estimerede sammenhæng mellem ætsninger på tænderne og tid brugt i svømmebassin, angiv et tilhørende 95% konfidensinterval for effekten, beskriv den testede nulhypotese, p-værdien og den videnskabelige konklusion.
10. Sammenlign din konklusion om sammenhængen mellem ætsninger på tænderne og tid brugt i svømmebassin fra spørgsmål 7. med de to konklusioner fra spørgsmål 8. og 9., hvor du kiggede på samme sammenhæng men blot blandt crawlsvømmere og ikke-crawlsvømmere hver for sig. Stemmer konklusionerne fra disse tre spørgsmål overens? Hvis ja: Forklar hvorfor. Hvis nej: Forklar hvorfor ikke.