

Startpakke til modulet 'Introduktion til Idræts- og sundhedsforskning' - 2026

Introduktion

Naturvidenskabelig forskning inden for idræt og sundhed bygger på systematisk indsamling og analyse af data, der benyttes til at skabe ny viden om eks. biologiske mekanismer, identificere årsager til sundhed og sygdom eller undersøge effekten af forskellige interventioner. Databehandling og statistik er derfor centrale redskaber i forskningsprocessen, der gør det muligt at beskrive data, teste hypoteser og estimere effekter og sammenhænge.

Undervisningen på det kommende modul introducerer nye metoder og begreber, som bygger videre på grundlæggende metodiske og statistiske begreber, der meget ofte allerede er blevet introduceret i undervisningen på en bacheloruddannelse. Det forventes derfor, at du ved studiestart har et grundlæggende kendskab til de begreber, der introduceres i denne startpakke. Begreberne vil blive anvendt løbende i undervisningen og vil som udgangspunkt ikke blive gennemgået systematisk. Formålet med startpakken er derfor at give alle studerende et fælles fagligt fundament fra start indenfor nogle grundlæggende begreber om data og statistik, uanset hvilken uddannelsesmæssig baggrund man kommer med. Studerende uden tidligere undervisning i forskningsmetode eller statistik anbefales derfor at gennemgå startpakken. Til flere af begreberne er der angivet henvisninger til relevante kapitler i Medical Statistics from Scratch (4. udgave) af David Bowers, som er tilgængelig som e-bog via SDU Bibliotek. Særligt kapitel 1, 4–7, 11–12, 14 og 17.

Til slut i startpakken er der placeret 10 multiple choice-spørgsmål, som giver dig mulighed for at vurdere, om du har opnået et tilstrækkeligt grundlag inden studiestart. Til hvert spørgsmål er der et facit med en kort forklaring

Grundlæggende begreber

Nedenfor er et kort overblik over de begreber, som du forventes at kunne genkende og anvende.

Populationer og stikprøver

Målpopulation: den gruppe, som man i sidste ende ønsker at kunne sige noget om. Den bør afgrænses med personer, sted og tid. Eksempel: alle U19-elitefodboldspillere i Danmark i sæsonen 2026/27.

Kildepopulation: den konkrete gruppe eller ramme, hvorfra deltagerne kan rekrutteres. Den kan fx være spillere i bestemte klubber, medlemmer af et register eller patienter på en klinik.

Studiepopulation: de personer, som faktisk indgår med data i analysen efter inklusion, samtykke, eventuelle eksklusioner og frafald.

Afgrænsningen af populationerne bør foretages før dataindsamlingen og skal være eksplicit i studieprotokollen. Det er særligt vigtigt at skelne mellem den population, som er fagligt relevant, og den population, der praktisk kan rekrutteres. En repræsentativ stikprøve er ikke altid et mål i sig selv, men et middel til at reducere systematiske forskelle mellem studiepopulationen og den målpopulation, som resultaterne ønskes generaliseret til. En stikprøve (studiepopulationen) er et udsnit af en population. Tilfældig udvælgelse kan under de rette betingelser reducere risikoen for at stikprøven ikke er repræsentativ for det spørgsmål man ønsker at undersøge, men selv en tilfældigt udtrukket stikprøve kan afvige fra målpopulationen som følge af manglende deltagelse, utilstrækkelig dækning af kildepopulationen eller efterfølgende frafald. Beskrivelse af rekrutteringsforløbet kan derfor i mange tilfælde være en central del af vurderingen af ekstern validitet. Jo større afstand der er mellem mål-, kilde- og studiepopulation, desto mere forsigtig skal man være med at generalisere resultaterne.

Læs videre: Kapitel 11: "Getting the participants for your study: ways of sampling".

Datatyper, skalaer og fordelinger

Nominal data: kategorier uden naturlig rækkefølge, fx sportsgren, køn eller ja/nej til en skade.

Ordinal data: kategorier med en rækkefølge, fx selv vurderet helbred: dårligt, rimeligt, godt. Afstanden mellem kategorierne behøver ikke være ens.

Metriske data: tal med meningsfulde afstande mellem værdier. Intervaldata mangler et absolut nulpunkt (fx temperatur i °C), mens ratioskalaer har et nulpunkt (fx kropsmasse og

træningstid). Metriske data kan være diskrete (fx antal skader) eller kontinuerte (fx kondital).

Fordeling: hvordan observationerne i en studpopulation fordeler sig. En fordeling kan være omtrent normalfordelt, skæv eller præget af få ekstreme værdier (outliers). Fordelingen har betydning for, hvordan data beskrives og analyseres.

Skalaen har betydning for, hvilke beregninger og sammenligninger der er meningsfulde. For nominale data kan man eks opgøre hyppigheder og andele, men der er ikke grundlag for at beregne en middelværdi. For ordinale data kan rangordenen være relevant, men det kan ikke uden videre antages, at forskellen mellem to nabokategorier er den samme på tværs af skalaen.

Vurderingen af en fordeling af metriske data kan baseres på grafisk fremstilling. Histogrammer og boksplot kan være med til at synliggøre om der er tale om eks normalfordeling, skævhed, og ekstreme observationer. Disse forhold kan have betydning for valg af mål for centraltendens og spredning og dermed for, om antagelserne bag en efterfølgende statistisk model er rimeligt opfyldt.

Læs videre: *Kapitel 1: "First things first – the nature of data"* og *kapitel 4: "Describing data from its shape"*.

Deskriptiv statistik

Deskriptiv statistik bruges til at give læseren et præcist billede af studiepopulationen og af de vigtigste variable. Valget af hvilken deskriptiv statistik man vil opgøre starter altid med at hvilken skala data er på: Er data kategoriske, eller er de metriske? Og er fordelingen nogenlunde normalfordelt eller tydeligt skæv?

Beskrivelse af en stikprøve angiver typisk antal deltagere, relevante baggrundsoplysninger og eventuelt fordelingen på grupper. Kategoriske variable beskrives ofte med antal og procent. Metriske variable beskrives med et mål for centraltendens og et mål for spredning i data.

Middelværdi og standardafvigelse (SD): passer ofte godt til omtrent symmetrisk fordelte metriske data. Middelværdien angiver det gennemsnitlige niveau, mens SD beskriver, hvor meget observationerne varierer omkring middelværdien.

Percentil: er et mål, der beskriver en observations placering i en fordeling (dvs i studiepopulationen). Den 5. percentil er eksempelvis den værdi, hvor 5 % af observationerne er mindre end eller lig med denne værdi, mens 95 % er større.

Median og interkvartilafstand (IQR): er ofte mere informative ved skæve fordelinger eller tydelige outliers. Medianen er den midterste observation (medianen identisk med den 50. percentil), og IQR beskriver afstanden mellem 25.- og 75.-percentilen.

Deskriptiv statistik opgøres typisk i en deskriptive tabel. Den skal gøre det muligt for læseren at vurdere, hvem der er undersøgt, hvordan data fordeler sig, og om der er forhold ved datamaterialet, der bør have betydning for den videre fortolkning. Det bør også fremgå tydeligt, hvor mange observationer der ligger til grund for hvert resultat, og om der forekommer manglende data.

Middelværdi og median besvarer ikke helt det samme spørgsmål. Middelværdien er følsom over for alle observationers størrelse og påvirkes derfor af ekstreme værdier. Medianen angiver derimod den observation, der deler materialet i to lige store dele. Ved vurdering af spredning er standardafvigelsen knyttet til middelværdien, mens IQR knytter sig til de centrale 50 % af observationerne.

Eksempel. Hvis hvilepuls er højreskæv i studiepopulationen, kan “median 58 slag/min (IQR 52–66)” være mere dækkende end “middelværdi 61 slag/min (SD 17)”. Begge beskrivelser kan være korrekte, men de fremhæver forskellige sider af data.

Læs videre: *Kapitel 2–6, især kapitel 5: “Measures of location” og kapitel 6: “Measures of spread”.*

Hyppighedsmål

Hyppighedsmål er også deskriptive statistiske mål, der anvendes til at beskrive, hvor ofte en sygdom, et helbredsudfald eller en eksponering forekommer i en population.

Prævalensproportion: andelen, som har en tilstand på et bestemt tidspunkt eller i en angivet periode. Ved sæsonstart har 18 af 120 håndboldspillere skuldersmerter: prævalensen er $18/120 = 15 \%$.

Kumuleret incidensproportion: andelen blandt personer i risiko ved periodens start, som udvikler et nyt tilfælde i løbet af perioden. Hvis 12 af 100 skadesfri spillere får en ankelskade i en sæson, er den kumulerede incidens 12 %.

Incidensrate: antallet af nye tilfælde pr. samlet tid i risiko. Hvis 30 skader opstår over 15.000 træningstimer, er incidensraten 2,0 skader pr. 1.000 timer. Målet er nyttigt, når deltagerne følges i forskellig lang tid eller har forskellig eksponeringstid.

Prævalens og incidens skal ikke opfattes som udskiftelige mål. Prævalensen afspejler både, hvor ofte nye tilfælde opstår, og hvor længe tilstanden varer. En høj prævalens kan således skyldes høj incidens, lang varighed eller begge dele. Incidensmål beskriver derimod forekomsten af nye tilfælde og er særligt relevante ved undersøgelser af risiko og ætiologi.

Ved beregning af kumuleret incidens skal nævneren omfatte personer, der er uden udfaldet og dermed i risiko ved periodens start. Ved beregning af incidensrate er tidsdimensionen central: Personer bidrager med tid, så længe de er under observation og i

risiko for udfaldet. Det giver mulighed for at håndtere forskellig opfølgningstid, men forudsætter en omhyggelig definition af både hændelse og person-tid.

Læs videre: *Kapitel 7: "Incidence, prevalence, and standardisation"*.

Sandsynlighed, hypotesetest og signifikansniveau

Når vi undersøger en stikprøve, vil resultaterne altid være påvirket af tilfældig variation. Sandsynlighedsteori beskriver, hvordan denne variation opstår, og gør det muligt at kvantificere usikkerheden omkring vores estimer. Hvis den samme stikprøveudtagning blev gentaget mange gange, ville estimer som et gennemsnit eller forskellen mellem to grupper variere fra stikprøve til stikprøve. Den centrale grænseværdisætning (Central Limit Theorem) viser, at fordelingen af disse estimer ofte vil nærme sig en normalfordeling, når stikprøven er tilstrækkeligt stor, også selv om de enkelte observationer ikke er normalfordelte. Dette danner grundlaget for statistisk inferens og gør det muligt at beregne konfidensintervaller og p-værdier, som anvendes til at kvantificere usikkerhed og teste hypoteser om forskelle mellem grupper eller sammenhænge mellem variable.

Statistisk inferens: er den del af statistikken, der anvendes til at drage konklusioner om en population på baggrund af en stikprøve (studiepopulationen). Ved hjælp af sandsynlighedsteori kvantificeres usikkerheden omkring de estimerede forskelle eller sammenhænge, hvilket gør det muligt at beregne konfidensintervaller og teste hypoteser.

Nulhypotese (H_0): en formel antagelse om ingen forskel eller ingen sammenhæng i populationen, fx at et træningsforløb ikke ændrer gennemsnitlig systolisk blodtryk.

P-værdi: sandsynligheden for at observere et resultat mindst lige så ekstremt som det fundne, hvis nulhypotesen er sand. En p-værdi er ikke sandsynligheden for, at nulhypotesen er sand.

Signifikansniveau (α): den på forhånd valgte grænse for, hvor stor risiko man vil acceptere for en falsk positiv konklusion. Ofte bruges $\alpha = 0,05$. Er $p < 0,05$, kaldes resultatet traditionelt statistisk signifikant. Statistisk signifikans siger ikke i sig selv noget om, hvor stor eller relevant en forskel eller sammenhæng er.

Hypotesetestning: er en statistisk metode, hvor resultater fra en stikprøve anvendes til at vurdere, om en observeret forskel eller sammenhæng kan forklares ved tilfældig variation alene, eller om data giver grundlag for at forkaste en på forhånd specificeret nulhypotese.

Læs videre: *Kapitel 12: "The idea of probability" og kapitel 17: "Testing hypotheses about the difference between two population parameters"*.

Konfidensinterval

Konfidensinterval (CI): angiver usikkerheden omkring et estimat fra stikprøven (studiepopulationen), fx et gennemsnit, en prævalens eller en forskel mellem to grupper. Et 95 % CI kan også forstås ud fra en tanke om, at den samme undersøgelse gentages mange gange: I omkring 95 % af undersøgelserne vil det beregnede konfidensinterval omfatte den sande populationsværdi.

En estimeret forskel i systolisk blodtryk på -4 mmHg med et 95 % konfidensinterval på $[-7; -1]$ mellem interventions- og kontrolgruppen peger på, at interventionen er forbundet med et lavere systolisk blodtryk. Punktestimatet er en reduktion på 4 mmHg, mens konfidensintervallet viser, at den sande gennemsnitlige reduktion, som er forenelig med de observerede data, kan være mellem 1 og 7 mmHg.

Læs videre: Kapitel 14: "Estimating the value of a single population parameter – the idea of confidence intervals".

Type 1- og type 2-fejl

Type 1-fejl: at forkaste en nulhypotese, som faktisk er sand – en falsk positiv konklusion. Sandsynligheden styres af signifikansniveauet α ; ved $\alpha = 0,05$ accepterer man højst 5 % risiko pr. test.

Type 2-fejl: at undlade at forkaste en falsk nulhypotese – en falsk negativ konklusion. Risikoen betegnes β .

Læs videre: Kapitel 17: "Testing hypotheses about the difference between two population parameters".

Selvtest (multiple choice)

Vælg det bedste svar. Se facit på sidste side.

1. *Hvilken population består af de deltagere, der faktisk indgår med data i analysen?*

- A. Målpopulationen
- B. Kildepopulationen
- C. Studiepopulationen

2. *Et studie vil sige noget om alle danske motionsløbere. Deltagerne rekrutteres fra tre lokale løbeklubber. Hvad udgør medlemmerne af de tre løbeklubber i denne sammenhæng?*

- A. Kildepopulationen
- B. Studiepopulationen
- C. Målpopulationen

3. *Hvilken variabel er ordinal?*

- A. Antal træningstimer pr. uge
- B. Selvrapporteret helbred: dårligt, rimeligt, godt
- C. Typen af sportsgren

4. *Hvad vil ofte være den mest robuste statistik til at beskrive en tydeligt højreskævet metrisk variabel?*

- A. Middelværdi og standardafvigelse (SD)
- B. Median og interkvartilafstand (IQR)
- C. Kun minimum og maksimum

5. *Ved sæsonstart har 18 ud af 120 spillere knæ smerter. Hvilket hyppighedsmål beskriver dette?*

- A. Incidensrate

- B. Kumuleret incidensproportion
- C. Prævalensproportion

6. *Hvad indgår i nævneren ved beregning af en incidensrate?*

- A. Antal personer ved periodens start
- B. Samlet person-tid i risiko
- C. Antal eksisterende tilfælde

7. *Hvilken påstand om p-værdien er korrekt?*

- A. Den angiver sandsynligheden for, at nulhypotesen er sand.
- B. Den angiver, hvor vigtig en effekt eller sammenhæng er.
- C. Den angiver sandsynligheden for at observere en forskel eller sammenhæng mindst lige så stor som den observerede, hvis nulhypotesen er sand.

8. *Hvad vil typisk gøre et konfidensinterval smallere?*

- A. En større stikprøve
- B. Større variation i data
- C. Færre deltagere

9. *Hvad er en type I-fejl?*

- A. At overse en reel effekt eller sammenhæng.
- B. At konkludere, at der er en effekt eller sammenhæng, som ikke findes.
- C. At beregne et for bredt konfidensinterval.

10. *Et ikke-signifikant resultat i et lille studie betyder bedst set, at:*

- A. Der helt sikkert ikke er nogen forskel eller sammenhæng.
- B. Nulhypotesen er bevist.
- C. Data ikke giver tilstrækkeligt grundlag for at forkaste nulhypotesen.

Facit og kort forklaring

1. C. Studiepopulationen. Studiepopulationen består af de deltagere, der faktisk indgår med data i analysen.
2. A. Kildepopulationen. Kildepopulationen er den population, hvorfra deltagerne rekrutteres. Målpopulationen er alle danske motionsløbere, mens studiepopulationen er de personer, der faktisk deltager.
3. B. Selvrapporteret helbred: dårligt, rimeligt, godt. Ordinale variable består af kategorier med en naturlig orden, men hvor afstanden mellem kategorierne ikke kan kvantificeres.
4. B. Median og interkvartilafstand (IQR). Median og IQR påvirkes mindre af ekstreme værdier og er derfor velegnede til højreskæve fordelinger.
5. C. Prævalensproportion. Prævalens beskriver, hvor stor en andel af en population der har en sygdom eller tilstand på et bestemt tidspunkt.
6. B. Samlet person-tid i risiko. Incidensraten beregnes som antallet af nye tilfælde divideret med den samlede person-tid, hvor deltagerne har været i risiko for at udvikle udfaldet.
7. C. Den angiver, hvor sandsynligt det er at observere en forskel eller sammenhæng som den fundne, hvis nulhypotesen er sand. P-værdien er et mål for, hvor forenelige de observerede data er med nulhypotesen. Den angiver ikke sandsynligheden for, at nulhypotesen er sand.
8. A. En større stikprøve. En større stikprøve giver typisk mere præcise estimater og dermed smallere konfidensintervaller.
9. B. At konkludere, at der er en effekt eller sammenhæng, som ikke findes. En type I-fejl opstår, når nulhypotesen forkastes, selv om den i virkeligheden er sand (falsk positiv).

10. C. Data ikke giver tilstrækkeligt grundlag for at forkaste nulhypotesen. Et ikke-signifikant resultat betyder ikke, at nulhypotesen er sand, men at data ikke giver tilstrækkelig evidens til at forkaste den.