

Balladen om p-værdier

Såkaldte p-værdier bruges flittigt når der testes videnskabelige hypoteser, selvom man i mere end 100 år har vidst, at de ikke kan stå alene.

Forfatterne



Claus Thorn Ekstrøm professor ved sektion for biostatistik, Københavns Universitet samt formand for Dansk selskab for teoretisk statistik. ekstrom@sund.ku.dk



Per Bruun Brockhoff professor ved Sektion for Statistik og Dataanalyse, DTU Compute perbb@dtu.dk

Læs mere:

ASA's pressemeddelelse om p-værdier: www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

Se også www.sandsynligvis.dk

Nyhedsbreve og de sociale medier gik i selvsving efter the American Statistical Association (ASA) i marts 2016 udsendte en pressemeddelelse omkring brugen og ikke mindst fortolkningen af såkaldte p-værdier. Pressemeddelelsen var ansporet af en videnskabelig (og til tider ret filosofisk) debat om p-værdier, der var blusset op igen, efter at redaktørerne for tidsskriftet *Basic and Applied Social Psychology* i 2015 offentliggjorde, at de ikke længere ville acceptere artikler, der gjorde brug af p-værdier. Hvordan er vi så endt der, at verdens største statistikersammenlutning følger anledning til at understrege, hvad begreberne betyder og ikke betyder her mere end 100 år siden disse begreber blev introduceret?

Intet nyt under solen

ASA's udtalelse indeholder intet nyt eller kontroversielt. Faldgruberne ved p-værdier er velkendte, og statistikere har gjort opmærksom på disse, siden begrebet oprindeligt blev introduceret. Men problemerne er blevet forstærket, fordi p-værdi-begrebet på trods af statistikernes forbehold er kommet til at spille en uforholdsmæssig stor rolle.

En p-værdi udtrykker, hvor godt de observerede data stemmer overens med en statistisk model eller hypotese, og værdien afhænger af de antagelser, man har gjort sig omkring sin model og hypotese.

I videnskabelig forskning vil man ofte gerne vurdere, om en påstand eller hypotese er korrekt eller ej. Med

andre ord vil man – på baggrund af sine data – gerne sige noget om sin forskningshypotese. Det er altså den "omvendte" sammenhæng mellem data og hypotese, og en klassisk misforståelse er, at man kan bruge p-værdien alene til at udtrykke, hvorvidt forskningshypotesen er sand eller ej på baggrund af de observerede data. En bedre fortolkning er, at p-værdien udtrykker "graden af evidens imod hypotesen", og kun under særlig gunstige omstændigheder kan man tilnærmelsesvist fortolke p-værdien som antydning.

Et andet ankepunkt ved p-værdien er, at den afhænger kraftigt af stikprøvestørrelsen. I små datasæt kan det være svært at opnå små p-værdier, selvom forskellen mellem to grupper er stor. Omvendt falder p-værdien automatisk, så selv de mindste forskelle bliver statistisk signifikante, når stikprøvens størrelse vokser. P-værdien siger i sig selv altså ikke noget om, hvorvidt det fundne resultat er biologisk relevant, eller om man har overset noget, som var biologisk interessant.

Mest problematisk er det, at p-værdien har fået så stor betydning for publicering af videnskabelige resultater. Mange tidsskrifter er mere tilbøjelige til at frasortere artikler, der ikke viser resultater med en lille p-værdi (typisk under den arbitrære grænse på 5%). Det betyder, at p-værdiens rolle bliver forstærket, selvom værdien ikke burde stå alene.

Debatten er vigtig – også i gymnasiet

P-værdier er nyttige, men pressemeddelelsen fra ASA er en god anledning til at minde os alle om, hvordan de skal fortolkes, og at man ikke skal basere sin konklusion udelukkende på en p-værdi. Faktisk er der også statistikere som Geoff Cummings, der i en årrække har gjort en dyd ud af at forsøge at ændre holdningen til vurdering af statistiske analyser ved at opfordre til i langt højere grad at fokusere på usikkerheder og konfidensintervaller end på hypotesetest og p-værdier. Samtidig bør man altid så vidt muligt forsøge at se sine forskningsresultater i et større perspektiv end det enkelte studie, der har frembragt disse data. Det er kun fint, hvis ASA's bidrag kan hjælpe til debatten og til at få spredt budskabet om dette.

Herhjemme kunne vi fx kigge på gymnasiernes pensum, hvor statistik er obligatorisk pensum for stx og hxx gymnasieelever med matematik på mindst B-niveau. Ud over deskriptiv statistik er hypotesetest og p-værdier pensum, mens fx konfidensintervaller/fejlestimering ikke nævnes med et ord. Det burde man lave om på! ■

Hvad er en p-værdi?

For en given statistisk model eller hypotese angiver p-værdien sandsynligheden for at observere et datasæt, der passer "dårligere" med modellen/hypotesen end det oprindeligt observerede datasæt. Hvis man fx vil undersøge gennemsnitsvægten i to grupper – hvor den ene gruppe er på slankekur og den anden ikke – så kunne hypotesen være, at kuren ikke virkede, og at gennemsnitsvægten i de to grupper derfor skulle være ens. Finder man en forskel i de to gruppers gennemsnit på fx 6 kg, så vil p-værdien udtrykke sandsynligheden for at finde en forskel i de to gruppers gennemsnit på mindst 6 kg, beregnet under scenariet, at kuren ikke virker. Eller hvis man lader det usandsynlige i et få fx 15 plat ud af 15 møntkast overbevise en om at mønten er en "snydemønt", så har man faktisk tænkt i og anvendt en p-værdi – 15 ens udfald er nemlig særdeles usandsynligt, hvis mønten er fair.