Vejledning til Excel-ark til Kappaberegning

Jan Ivanouw

16. december 2008

Om interraterreliabilitet og Kappaberegning

Formålet med Kappaberegning er at vurdere hvor god overensstemmelse der er mellem to personer der bruger det samme vurderingssystem. Det er det som også hedder *interraterreliabnilitet.*

Kappavurdering kan anvendes til hvad som helst hvor to personer skal bruge et kategorisystem til at vurdere det samme. I afspændingspædagogik vurderer vi f.eks. ofte muskeltonusforhold. Vi kan gøre det ved at palpere musklerne, vi kan gøre det ved at undersøge muskulaturens reaktion på stræk, vi kan gøre det ved at vurdere holdning eller åndedræt m.m. I alle disse tilfælde opstiller vi en række vurderingskategorier, og vi har brug for at vide om man kan stole på vurderingerne ud fra systemet. Hvis vi udvikler andre vurderingssystemer, bør vi også vurdere Kappa for de kategorier vi anvender, så vi kan se hvor godt systemet fungerer.

For at få en god vurdering er det nødvendigt at tage hensyn til muligheden for tilfældige sammenfald. Kappametoden strammer fra Jacob Cohen (1962) og er en tilfældighedskorrigeret metode til at vurdere overensstemmelse mellem to (eller flere) personer der vurderer det samme ud fra et kategorisystem. Den oprindelige udgave af Kappa undersøger kun mellem om der er helt nøjagtig overensstemmelse mellem to testere, alt andet regnes for fejl. Der er også udviklet en *vægtet Kappa* (Cohen, 1968) som teger hensyn til hvor stor uenighed der er mellem de to parter.

Hvis Kappa er lav, kan der være to forskellige forklaringer. Den ene er registreringssystemet er mangelfuldt og derfor svært at bruge. Det kan være fordi der er blandet forskellige dimensioner sammen, det kan være fordi der er kategorier der overlapper hinanden, og det kan være der er vigtige ting som ikke fanges af systemet. Selv om systemet i sig selv er godt nok, kan det være at beskrivelsen af systemet (manualen) ikke er tilstrækkelig udbygget og klar. Den anden mulige forklaring på lav Kappa kan være at systemet i sig selv er godt nok, men at de personer der foretager vurderingerne ikke er tilstrækkeligt trænede i at anvende det.

I dette tilfælde har jeg opstillet et system til at beregne Kappa der omfatter 5 kategorier. Jeg har regnet med to grader af overspændt muskeltonus,

en normal kategori og to grader af underspændt muskulatur. (men man kan lige så godt bruge Excel-arket til at beregne noget andet med 5 kategorier).

Instruktion til anvendelse af skemaet

Skemaet bruges ved at to personer først gennemfører en række vurderinger af det samme, f.eks. en række palpationsvurderinger på den samme person ('kaninen'). Den ene person skriver sine vurderinger ned på et papir uden at den anden ser det, og den anden vurderer de samme steder og skriver også sine vurderinger ned.

Derefter anvender man Excel-skemaet. Man vælger det første ledige ark (f.eks. Kappa1) og skriver navnene på de to testere. Derefter indtaster man resultaterne i rammen med de grønne felter. Man krydser de to personers vurderinger, f.eks. 'let overspændt' for første person og 'meget overspændt' fra den anden. Man indfører det i skemaet ved at lægge 1 til tallet i den tilsvarende celle i skemaet. Ved starten står der 0, og når man lægger 1 til, får man selvfølgelig 1. Næste gang man har den samme kombination, lægger man 1 til 1-tallet og skriver 2. De mørkegrønne felter repræsenterer enighed mellem de to personer. De mellemgrønne felter repræsenterer 1 grads forskel mellem personerne, og de lysegrønne viser mere end 1 grads forskel. Farverne har ikke nogen betydning for beregningen, men er kun med for at gøre det lettere at finde rundt i skemaet når man taster ind. Der er 10 Excel-ark til brug for 10 beregninger. Der er endvidere et ark der hedder 'Kappa1-10', som lægger indtastningerne for alle de 10 ark sammen så man få en samlet beregnning.

Når man har indført alle resultaterne, kan man aflæse Kappa, som er et tal der vurderer hvor god overensstemmelse der er mellem de to personer. Kappa ligger mellem -1.00 og +1.00. En Kappa på 1.00 betyder perfekt overenstemmelse mellem de to personer. Kappa på 0 betyder at overensstemmelsen ikke er bedre end hvad man vil kunne opnå ved tilfældigt sammenfald. En Kappa under 0 betyder at overensstemmelsen er værre end man vil kunne forvente ved tilfældigt sammenfald.

Der er endvidere to felter med 95% sikkerhedsgrænser for Kappa (i det røde felt). Imidletid må man forvente at hvis man indsamlede resultater flere gange for de samme to personer, så ville man ikke få nøjagtigt samme Kappaværdi hver gang. Kappa vil nødvendigvis variere noget. Jeg har indlagt sikkerhedsgrænser i skemaet (det er de to felter med lysere rødt). Sikkerhedsgrænserne viser hvor meget man kan forvente at Kappa vil kunne variere for de samme personer i samme situation (de viser inden for hvilke grænser man med 95% sandsynlighed vil finde det mest sande resultat). Afstanden mellem sikkerhedsgrænserne bliver mindre jo flere tal der er i tabellen. Det giver vel også mening: jo flere muskler de to personer undersøger og sammenligner, jo mere sikker kan man være på resultatet giver et sandt billede

af deres enighed. Hvis man omvendt kun nøjes med at lave ganske få sammenligninger, så bliver resultatet selvfølgeligt mere upræcist, hvilket viser sig ved sikkerhedsgrænser der afgrænser et meget bredt interval.

I nogle tilfælde bliver den øvre sikkerhedsgrænse over 1.00, men det skal blot forstås som 1.00. Tilsvarende kan den laveste sikkerhedsgrænse komme under -1.00, men dette skal tilsvarende forstås som -1.00.

Der er endelig et felt mærket 'Vægtet Kappa'. Ved almindelig Kappaberegning tælles kun fuld overensstemmelse med som et godt resultat (altså at begge f.eks. har noteret 'let overspændt'). Ved vægtet Kappa tages hensyn til hvor stor forskel der er på de to testere. En forskel på en enkelt grad (f.eks. fra 'normal' til 'let overspændt') tæller som mindre fejl end en forskel der er større.

Anvendelsesmuligheder

Skemaet kan bruges til forskellige formål.

- Det kan bruges til forskning (eller empirien i en bacheloropgave).
- Det kan bruges til undervisning således at hele klassens resultater fra forskellige par indtastes så man får et samlet billede af klassens dygtighed.
- Det kan også bruges til træning af de enkelte studerende ved at to studerende arbejder sammen og vurderer en række steder på en eller flere forskellige kaniner og indfører deres resultater i skemaet. Hvis man træner sammen et stykke tid, diskuterer resultaterne og gentager sammenligningerne, kan man forbedre sin enighed.

Anvendelse til undervisning

En mulighed er at alle de studerende deles op i trepersonersgrupper hvor to tester den tredie. Hver gruppe får resultaterne indtastet i sit eget Excelskema. Man kan se hvilke par der har mest vanskeligt ved at blive enige. På det samlede skema (Kappa1-10) kan man se hvor stor enighed man har nået på holdet som helhed.

En anden mulighed er at et bestemt par af studerende tester en tredie på en række muskler og indfører dem i skemaet 'Kappa1'. Herefter diskuterer de forskellene og prøver igen (måske en anden dag). Denne gang anvendes skemaet 'Kappa2' osv. Resultaterne kan bruges til at vurdere fremskridt i reliabiliteten. Til denne anvendelse er *vægtet Kappa* særligt egnet fordi den viser gradvis forbedring i enighed også selvom man (endnu) ikke er helt enige. Det er også her nyttigt at anvende rækken og kolonnen med 'I alt' for at se om man har en systematisk bias - dvs. om man er tilbøjelig til at bruge

f.eks. den ene side af skalaen for meget. Det er det man tidligt i astronomien kaldte *den personlige ligning.*

En tredie mulighed er at samle resultaterne for en bestemt muskel i 'Kappa1' en anden muskel i 'Kappa2' osv. og derved kunne undersøge om nogle muskler ser ud til at være sværere at teste (=sværere at blive enige om).

En fjerde mulighed er at hele klassen tester en række muskler med både stræk- og trykgreb. Resultaterne med strækgreb indføres i 'Kappa1' og resultaterne med trykgreb i 'Kappa12' så man kan undersøge om man når samme grad af enighed med de to typer greb.

Anvendelse til forskning (og bacheloropgaver)

En vis del af de data (eller hvis man har ressourcer til det, alle data) man anvender til sin forskning, checkes ved at man lader to personer teste de samme personer. Kappa angives i opgaven (eller artiklen) til at begrunde reliabiliteten af de anvendte målinger. Det er *Kappa* (uden vægtning) der er interessant til brug for forskning, fordi man her er interesseret i hvor reliabelt det færdige kategorisystem er.

Anvendelse til udvikling af registreringssystemer

Inden for det afspændingspædagogiske område anvendes ofte forskellige vurderingssystemer. Når man udvikler sådan et system, skal man gøre systemet så velbeskrevet og entydig så muligt så man kan anvende det med høj interraterreliabilitet. Altså det system man udvikler, skal kunne anvendes af flere forskellige afspændingspædagoger som gerne skal kunne komme til de samme resultater. Udviklingen af sådan et system sker som regel i flere omgange. Først beskriver man systemet så godt man kan, og så prøver man at lade flere personer bruge det, og man beregner Kappa for deres overensstemmelse. Som regel er den i begyndelsen lav. Man diskuterer derefter hvad der især er vanskeligt at forstå og anvende og laver beskrivelserne om. Derefter prøver man igen. Og således bliver man ved indtil man får acceptable Kappaværdier.

Man arbejder samtidigt med de to mulige forklaringer på lav Kappa. Man udvikler systemet til at være velfungerende og man udvikler en beskrivelse som er klar og tilstrækkeligt omfattende. Samtidig arbejder man på metoder til at undervise de personer der skal bruge systemet.

Man kan godt anvende beregningsarket til andet end muskelpalpationstest. Så længe man (tilfældigvis) bruger kategorisystemer med netop 5 kategorier, kan man anvende beregningsarket (selvom man må kalde kolonnerne og rækkerne noget der passer til formålet). Hvis man har færre eller flere kategorier end 5, må man udvikle et andet beregningsark.

Til denne anvendelse er både beregningsformer nyttige. *Kappa* er nyttig til at beskrive hvor godt man har fået det samlede system til at fungere, mens *vægtet Kappa* er nyttig til at vurdere mulige forbedringer under processen med at udvikle systemet.

Advarsel

Man skal være klar over at Kappaberegningen viser graden af *enighed*. Hvis de to der sammenligner deres resultater begge gør fejl på samme måde, kan de få en god overensstemmelse selvom de faktisk bruger systemet forkert.

Litteratur:

Cohen, J. (1962). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220.