

# Kappa - koefficienten

## Et udtryk for målemetoders pålidelighed

Fysioterapilærer Hans Lund, Skodsborg Fysioterapiskole og instruktionsfysioterapeut  
Merete Wormslev, Amtssygehuset i Herlev

Lund H, Wormslev M (1997) (elektronisk version 2003, 4. november)  
Kappa-koefficienten. *Forskning i Fysioterapi (online, 1. årg.)* s. 1-4:  
URL: <http://www.ffy.dk/sw858.asp>

Artiklen har tidligere været bragt i *Nyt om Forskning* nr. 1, 1997, s. 13-15.

### Indledning

Det er gængs i fysioterapeutisk praksis at bruge bestemte tests i diagnostisk øjemed og ved effektmåling. Men hvor stor betydning kan vi tillade os at tillægge de enkelte test (diagnostisk værdi)? Svaret på dette spørgsmål kan vi afgøre på to måder. Den ideelle måde at vurdere en tests pålidelighed på er ved at sammenligne svarene fra testen med en facitliste (et sandt svar/validitet). I fysioterapi er dette dog sjældent (for ikke at sige aldrig) muligt at gøre. Man er derfor nødt til først ud fra f.eks. pato-anatomien eller den pato-fysiologiske virkningsmekanisme at argumentere for en tests relevans ved en bestemt lidelse. Dernæst må testens pålidelighed bedømmes ud fra dens reproducerbarhed (reliabilitet). Det vil sige, får vi det samme testresultat eller svar ved gentagne målinger, forudsat at det, der måles, ikke har ændret sig. Her vil vi kun beskæftige os med den sidstnævnte metode til at undersøge en tests pålidelighed på.

### Målemetoder

Hvordan det skal gøres er imidlertid afhængigt af, hvilken type af målemetode det drejer sig om. Målemetoder kan deles ind efter den type af skala, som man bruger i målemetoden. Hvis man f.eks. vil undersøge, om der er positiv Laseque eller ikke, taler man om målinger på en *nominal skala*. Det er i dette tilfælde en bi-nomi-

nal skala, idet der kun er svarmulighederne: Positiv Laseque/negativ Laseque.

Der kan også være flere svarmuligheder, f.eks. hvis man vil undersøge om en gruppe mennesker har ryg- eller nakkeproblemer, begge dele eller ingen af delene. Der er fire mulige svar. Kravet er, at den enkelte observation kun kan placeres i en af svarmulighederne (*eksklusive*), og at alle mulige observationer skal kunne placeres i en af svarmulighederne (*exhaustive*). Er man derimod interesseret i at undersøge, hvorledes patienten føler det efter behandlingen, kan dette registreres på en *ordinal- eller rangskala*. F.eks. kan man spørge, om patienten har det bedre, uændret eller værre efter behandlingen. Denne skala kaldes netop en rangskala, fordi den rangordner svarmulighederne.

Et andet eksempel kunne derfor være en registrering af patientens smerter. Her kan patienten få følgende muligheder: Ingen smerter, lette smerter, svære smerter og ubærlige smerter. Eller man kan give patienten et papir med en 10 cm lang streg tegnet på. I den ene ende står der »ingen smerter overhovedet«, og i den anden ende står »de værst tænkelige smerter« (en visuel analog skala forkortet VAS). Patienten sætter en streg, hvor vedkommende mener, at smerten bedst udtrykkes. Så måles afstanden

fra den ende, hvor der stod »ingen smerter« og ud til det punkt patienten, har markeret. Dette tal udtrykker graden af smerte. Men man kan ikke sige, at 4 cm er dobbelt så meget som 2 cm smerte, da markeringen er subjektiv.

Hvis man derimod måler en patients låromfang med et målebånd, eller man måler en patients ganghastighed, sker målingerne på en ratio- eller intervaskala. Mens data målt på en rangskala var karakteriserede ved, at hvert trin ikke nødvendigvis var lige store, er alle trin på en *ratio- eller intervaskala* lige store. Forskellen på en ratioskala og en intervaskala er, at ratioskalaen har et sandt nulpunkt.

F.eks. har temperaturskalaen celsius ikke et sandt nulpunkt, da nul er fastsat som vands frysepunkt. Det betyder, at man ikke kan sige, at 20° er dobbelt så varmt som 10°. Man kan derimod godt sige, at en mand på 180 cm er dobbelt så høj som en dreng på 90 cm.

Disse skalaer er af betydning, når vi skal undersøge en målemetodes pålidelighed. Vi skal nemlig have defineret, hvilken skala vore mål er angivet i, før vi beslutter os for, hvilken analysemetode vi skal anvende til at undersøge pålideligheden. Har man en målemetode, som registrerer data på en ratio- eller interval skala, kan man bruge flere forskellige statistiske metoder. Disse kunne f.eks. være korrelations-analyser, varians-analyser eller andre. Ofte vil en fysioterapeut bruge målemetoder på enten nominal, rang- eller ordinal skalaer. For at teste disse målemetoders pålidelighed kan man med fordel anvende *Kappa-koefficienten* (herefter: Kappa) (1,2,3).

### Observatør-variation

Lad os antage, at vi måler noget på en bi-nominal skala. Det kunne f.eks. være positiv eller negativ Laseque, positiv eller negativ Rhomborg, positiv eller negativ Trendelenburg osv. Som fysioterapeuter udfører vi den slags test flere gange hver dag. Men hvor sikre kan vi være på, at vi selv måler på samme måde hver gang (*intra-observatør-variationen*), eller at jeg måler ligesom min kollega inde bag det andet forhæng og får det samme resultat (*inter-observatør-variationen*)?

Variationen i resultaterne kan ud over observatør-variationen også skyldes tekniske (teknik-va-

riation) eller tidsmæssige (tids-variation) faktorer, men i fysioterapisammenhænge skyldes forskellen i test-svar oftest observatør-variationen. Intuitivt ved vi, at der kan være forskel på, hvad vi finder ved en bestemt test, og hvad min kollega finder ved den samme test. Måske kan vi endda gå med til, at der også er forskel på det, vi selv finder første gang, vi undersøger, og anden gang vi undersøger - uden at der er sket nogen ændring i situationen.

Måden at finde ud af det på synes da også ganske simpel. Hvis vi f.eks. gerne vil undersøge inter-observatør-variationen, kan vi jo først lade en fysioterapeut teste en gruppe patienter og derefter en kollega. Er der stor forskel, er der altså stor inter-observatør-variation, men er der lille forskel, synes testen altså at være ganske pålidelig. Vi ville måske endda acceptere, at vi kun var enige i 80 % af tilfældene. Der vil jo altid være en vis usikkerhed på enhver måling. Det er jo også derfor, vi ikke stiller en diagnose ud fra bare én test.

Men hvis vi undersøger pålideligheden på denne måde, har vi ikke taget højde for, at vi tilfældigvis kunne være enige. At den enighed, vi fandt, ikke var betinget af vor dygtighed og af testens pålidelighed, men simpelthen var udtryk for et tilfældigt sammentræf. Det betyder, at vi på en eller anden måde må trække de mulige tilfældige ens svar fra resultatet. Det er netop dette Kappa-beregningen er udformet til.

Lad os se på et eksempel. To fysioterapeuter skal undersøge pålideligheden af deres evne til at vurdere, hvorvidt en gruppe hofte-arthrose patienter har nedsat indadrotation i hoften eller ikke. 100 hoftepatienter møder op og bliver først testet af den ene og derefter af den anden. For at undgå, at den første undersøgelse skal påvirke den anden undersøgelse, skiftes de to fysioterapeuter til at undersøge patienterne først.

Det eneste fysioterapeuterne altså skal svare på er: har disse hofte-arthrose patienter nedsat indadrotation i hoften - JA/NEJ. Den ene fysioterapeut må selvfølgelig ikke se, hvad den anden finder og omvendt. Efter nogen tid er alle 100 patienter undersøgt og følgende resultat er opnået (se tabel 1).

**Tabel 1.** Eksempel på observatør-variation

		Fysioterapeut 1		
Fysioterapeut 2		JA - nedsat indadrotation	NEJ - ingen nedsat indadrotation	
	JA - nedsat indadrotation	31	6	37
	NEJ - ingen nedsat indadrotation	12	51	63
		43	57	100

Tabellen viser, at Fys 1 fandt 43 med nedsat indadrotation og 57 med normal indadrotation i hoften. Fys 2 fandt derimod kun 37 med nedsat indadrotation og 63 med normal indadrotation i hoften. Tallene med kursiv midt i tabellen viser deres indbyrdes enighed og uenighed. De var f.eks. enige om, at de samme 31 patienter havde nedsat indadrotation i hoften, og at de samme 51 patienter havde normal indadrotation i hoften.

På den anden side mente Fys 1, at 12 havde nedsat indadrotation, mens Fys 2 mente, at de samme 12 havde normal indadrotation. Og mens Fys 1 mente, at 6 havde normal indadrotation, mente Fys 2 at de havde nedsat indadrotation. De var altså enige i 82 af tilfældene (31+51), hvilket svarer til 0,82 (82%) enighed.

Dette tal kaldes den observerede overensstemmelse og angives som  $p_o$ . Men vi har ikke taget højde for de gange, hvor de tilfældigvis var enige.

Hvis vi skal tage højde for tilfældigheden, må vi beregne sandsynligheden for at en bestemt hændelse kan indtræffe sådan rent tilfældigt. For at beregne sandsynligheden for dette anvendes multiplikationsreglen for sandsynligheder: Sandsynligheden for at få udfaldet A og udfaldet B i samme forsøg er lig med sandsynligheden for at få udfaldet A gange med sandsynligheden for at få udfaldet B, såfremt sandsynligheden for at få udfaldet B er uafhængig af, om udfaldet A er indtruffet eller ej.

Hvis vi ser på resultatet, ser vi at Fys 1 fandt 43 med nedsat indadrotation (31+12). Hyppighe-

den var altså  $(31+12)/100 = 0,43$  for Fys 1 og  $(31+6)/100 = 0,37$  for Fys 2. Hyppigheden af det overensstemmende fund af nedsat indadrotation er derfor  $0,43 \cdot 0,37 = 0,159$ . På samme vis er hyppigheden af det overensstemmende fund af normal indadrotation  $0,63 \cdot 0,57 = 0,359$ . Man kan derfor forvente, at sandsynligheden for at få dette resultat i alt er  $0,159 + 0,359 = 0,518$ . Dette tal kaldes for den forventede tilfældige overensstemmelse (angives som  $p_c$ ).

### Kappa

For at udregne graden af pålidelighed, hvor vi har taget højde for den tilfældige overensstemmelse ( $p_c$ ), skal man bruge en formel for Kappa:

$$Kappa = p_o - p_c / 1 - p_c$$

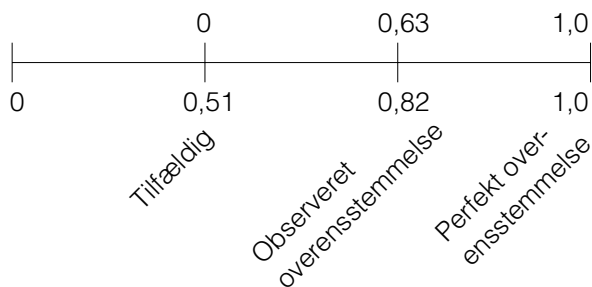
I tælleren kan man se, hvor meget større den observerede overensstemmelse ( $p_o$ ) er i forhold til den tilfældige overensstemmelse ( $p_c$ ). Hvis nu begge fysioterapeuter er helt enige, bliver den observerede overensstemmelse ( $p_o$ ) 1. I nævneren angives derfor, hvor meget den observerede overensstemmelse ( $p_o$ ) kan blive større end den tilfældige overensstemmelse ( $p_c$ ). Kappa kan variere mellem +1 og -1. I dette eksempel her blev Kappa:

$$0,820 (p_o) - 0,518 (p_c) / 1 - 0,518 (p_c) = 0,625 = \mathbf{0,63}$$

Det vil sige, at vi ikke fandt 82 % enighed, men kun 63 % enighed, når vi altså tog højde for den tilfældige overensstemmelse. Er det så godt? Som udgangspunkt må man sige, at jo nærmere 1, jo bedre. Landis & Koch (1977) foreslår følgende inddeling af Kappa-koefficienterne:

< 0,00	Poor
0,00-0,20	Slight
0,21-0,40	Fair
0,41-0,60	Moderate
0,61-0,80	Substantial
0,81-1,00	Almost perfect

Betydningen af Kappa-koefficienten kan forklares således: I stedet for blot at se på graden af overensstemmelse, ser man på forskellen på den tilfældige overensstemmelse og 100 % overensstemmelse. På strengen er markeret graden af overensstemmelse:



Kappa-koefficienten på 62,9 % (0,63) udtrykker således, at forskellen mellem den observerede overensstemmelse (0,82) og den tilfældige overensstemmelse (0,51) kun er 63 % af forskellen mellem perfekt overensstemmelse (1,0) og tilfældig overensstemmelse (0,51).

I faktaboksen (tabel 2) kan man se, hvordan man skal indsætte sine resultater. Hvis man vil undersøge intra-observatør-variationen, dvs. ens egen evne til at finde det samme svar i sine test, indsætter man blot dit første testsvar som observatør A og sit andet testsvar som observatør B.

Der er imidlertid nogle begrænsninger med Kappa. For det første kan den viste udregning, kun bruges på data, som er på en bi-nominal skala. For det andet kan man ikke anvende Kappa, hvis man undersøger to fysioterapeuters overensstemmende resultater hos en gruppe personer, hvor man f.eks. næsten ikke finder nogen positive svar. Altså hvor antallet af syge i gruppen er meget lille (prævalensen er lille). For det tredje vil Kappa-koefficienten ikke kunne sige noget om, hvor stor en del af gruppen, der er berørt af uoverensstemmelsen. Det vil sige, man kan ikke på baggrund af størrelsen af Kap-

pa sige noget om, hvilke konsekvenser uoverensstemmelsen mellem to observatører har.

Det er også muligt at beregne Kappa-koefficienten for data på en rang-skala, f.eks. ingen smerte, let smerte, meget smerte, ubærlig smerte, men det vil føre for vidt i denne artikel. Nedenfor har vi angivet nogle referencer, hvor interesse-rede kan læse videre.

## Referencer

1. T. Gjørup & A. Mørup Jensen: Kappakoefficienten - et mål for reproducerbarhed af nominale og ordinale data, *Nordisk Medicin*, 1986 101: 90-94
2. T. Gjørup: Klinisk vurdering af diagnostiske undersøgelsesmetoder, *Lægeforeningens Forlag 1988*, Doktordisputats, Københavns Universitet
3. Henrik Wulff : *Rationel klinik- grundlaget for diagnostiske og terapeutiske beslutninger*, Munksgaard 1987, s. 50-55
4. J.R.Landis & G.G.Koch: The measurement of observer agreement for categorical data, *Biometrics*, 1977, 33:159-174

## Tabel 2.

- a = Det antal, hvor begge observatører er enige om, er positive.  
 b= Det antal, som observatør A mener, er negative, men som observatør B mener er positive.  
 c= Det antal, som observatør A mener er positive, men som observatør B mener er negative.  
 d= Det antal, som begge observatører er enige om, er negative.  
 n= Det samlede antal observationer.

### Observatør A

	Positiv	Negativ	I alt
Positiv	a	b	a+b
Negativ	c	d	c+d
	a+c	b+d	N

$$p_o = \text{den observerede overensstemmelse} = (a+d)/n$$

$$p_c = \text{den forventede tilfældige overensstemmelse} = [(a+b)(a+c) + (c+d)(b+d)]/n^2$$

$$\text{Kappa} = p_o - p_c / 1 - p_c$$