Hoe hoog moet *Cohens kappa* zijn bij het indelen in scheef verdeelde categorieën ?

Henk Elffers

Samenvatting

Besproken wordt hoe Cohens maat voor beoordelaarsovereenstemming, kappa, afhangt van de kansen op foutieve classificatie enerzijds, en anderzijds de fractie van voorkomen van het te klasseren kenmerk. Geconcludeerd wordt dat de conventionele standaarden voor kappa te streng zijn in het geval dat deze laatste fractie erg hoog is.

Betrouwbaar indelen in categorieën

Eén van de meest voorkomende taken voor elke empirische sociaal-wetenschappelijke onderzoeker is het indelen van eenheden van analyse in categorieën. Gezinnen indelen in welstandsklassen, gedingen als wel of niet fair beoordelen, teksten als wel of niet racistisch kenschetsen, van incorrecte belastingaangiftes beoordelen of er al of niet sprake is van opzettelijk verkeerd weergeven van de stand van zaken, Rijnmondburgers indelen naar milieusolidariteit, ga zo maar door. Kenmerk van dit soort taken is vaak dat een vrij vaag omschreven concept moet worden toegepast op een concreet geval, en dat het gaat om een operationalisatie van dat concept waarbij de menselijke beoordelaar een onmisbare rol speelt. Juist omdat het meetvoorschrift niet volledig geobjectiveerd is, maar een beoordelaar een niet weg te cijferen rol speelt, dient elke onderzoeker die zich voor zo'n taak gesteld ziet zich

---

Correspondentie:

Henk Elffers, Nederlands Studiecentrum Criminaliteit en Rechtshandhaving NSCR, Postbus 792, 2300 AT Leiden, telefoon: +31 (0)70 5278509, e-mail: elffers@nscr.nl.

Dit artikel is geschreven in het kader van het onderzoekprogramma van de *Onderzoekschool Maatschappelijke Veiligheid*. Een eerdere versie werd gepubliceerd in de bundel *Libris Satiari Nequeo,* aangeboden aan Kees Boender ter gelegenheid van zijn afscheid (Elffers, 2000).

he/20010202

terdege af te vragen: "*Hoe reproduceerbaar is deze indeling van eenheden in een categorieënsysteem*?" Een voor de hand liggende bron van variatie is daarbij uiteraard de persoon van de beoordelaar. Zullen verschillende beoordelaars wel tot dezelfde beoordeling komen? Zoals bekend noemt men een operationalisatie betrouwbaar als de resulterende meting niet teveel afhankelijk is van invloeden 'die er niet toe horen te doen'. We kunnen daarom de vraag naar de ongevoeligheid van zo'n indelingsprocedure ook verwoorden als "*is de beoordelaarsbetrouwbaarheid van de procedure wel voldoende?*"

Er zijn twee manieren in omloop om zich een mening te vormen over de mate waarin de inzet van een beoordelaar tot onbetrouwbaarheid leidt: de *absolute* en de *relatieve* methode. (Op een derde, benaderende, methode, die gebruik maakt van het panelkarakter van de betreffende data, ontwikkeld ten behoeve van een studie naar milieusolidariteit (Boender, 1985), en beschreven in Elffers (1985) gaan we nu niet in). Bij de absolute methode wordt de indeling zoals uitgevoerd door de beoordelaar vergeleken met een indeling volgens een uitwendige maat, bij voorkeur een die van onweersproken kwaliteit is. Als voorbeeld: men kan aan een onderwijzer vragen zijn leerlingen in te delen als al of niet een leerachterstand hebbend, en men kan dat dan vergelijken met wat een gestandaardiseerde schoolvorderingentoets ervan zegt. Dan kan de beoordelende onderwijzer twee soorten fouten maken: hij of zij kan een kind met achterstand ten onrechte niet herkennen, of hij of zij klasseert een kind dat feitelijk geen achterstand heeft ten onrechte wel als een achterstand hebbend. Als een fout van het eerste type wordt gemaakt zeggen we dat de procedure niet *gevoelig* is, als een fout van het tweede type wordt gemaakt noemen we de methodiek *niet-specifiek*. Laten we eens aannemen dat de onderwijzer een fractie $1-\varepsilon$ van de achterstandsgevallen niet als zodanig herkent, en een fractie $1-\zeta$ van de gevallen van 'geen achterstand' foutief als achterstand klasseert (tabel I) .

Tabel I: kwaliteit absolute methode *(rijfracties)*		onderwijzer	
		achterstand	geen achterstand
objectieve toets:	achterstand	ε	$1-\varepsilon$
	geen achterstand	$1-\zeta$	ζ

Duidelijk is dat de beoordelingsbetrouwbaarheid van de procedure hoog is als zowel ε (de gevoeligheid) als ζ (de specificiteit) hoog zijn, zeg in de buurt van 1. Men kan erover twisten wanneer ε en ζ hoog genoeg zijn, en wat men meer moet vrezen in een concrete procedure, een lage ε of een lage ζ. In menig geval zal men $\varepsilon, \zeta \geq 0.9$ als zeer bevredigend ervaren.

Vaak zal het zo zijn dat er helemaal geen 'absolute maat' voorhanden is. Stel dat iemand vonnissen moet indelen naar 'voldoende gemotiveerd' of 'onvoldoende gemotiveerd'. Dan is er geen objectieve maatstaf voorhanden. Toch is het van belang te weten of de classificatie door een beoordelaar van vonnissen betrouwbaar genoeg is. In dat geval is het mogelijk daar een indruk van te krijgen door dubbel-beoordelingen uit te voeren: men laat dezelfde zaken door twee (of meer) beoordelaars onafhankelijk van elkaar beoordelen, met als resultaat een tabel als tabel II, waarin de dichtome oordelen van de beoordelaars met de generieke labels 'wel' en 'niet' zijn aangeduid.

Tabel II: aantallen bij dubbelbeoordeling		beoordelaar 2		
		'wel'	'niet'	totaal
beoordelaar 1	'wel'	$\mathbf{n_{ww}}$	$\mathbf{n_{wn}}$	$\mathbf{n_{w+}}$
	'niet'	$\mathbf{n_{nw}}$	$\mathbf{n_{nn}}$	$\mathbf{n_{n+}}$
	totaal	$\mathbf{n_{+w}}$	$\mathbf{n_{+n}}$	$\mathbf{n_{++}}$

Duidelijk is in dit geval dat de procedure weinig kwetsbaar is voor beoordelaarsverschillen als de diagonaaltermen n_{ww} en n_{nn} groot zijn in vergelijking met n_{++}. Men drukt, in navolging van Cohen (1960), in dit geval de beoordelaarsbetrouwbaarheid vaak uit in een coëfficiënt κ (Cohens kappa). Deze is als volgt gedefinieerd. We gaan uit van de feitelijk door beide beoordelaars afgegeven oordelen, n_{w+}, n_{n+}, n_{+w}, n_{+n}. Als beide beoordelaars een oordeel geven dat in het geheel geen relatie heeft met het oordeel van de ander (de procedure is dan extreem onbetrouwbaar), dan zou de celvulling in het binnenste van tabel II proportioneel zijn: we verwachten in dat geval in de 'wel-wel'-cel een even grote proportie van de door beoordelaar 1 als 'wel' geklasseerde vonnissen als we in de 'niet-wel'-cel ten opzichte van de door beoordelaar 1 als 'niet' geklasseerde oordelen, en daarmee ook als in de 'totaal-wel'-cel t.o.v. het 'totaal-totaal'-aantal. In termen van de symbolen uit Tabel II verwachten we dan in de 'wel-

wel'-cel een aantal $(n_{+w} / n_{++}) \cdot n_{w+}$ en in de 'niet-niet'-cel een aantal $(n_{+n}/n_{++}) \cdot n_{n+}$. Ten opzichte van het aantal beoordelingen verwachten we onder de hypothese van 'geen verband' derhalve een fractie κ_e (index e slaat op 'expected')

$$\kappa_e = \{(n_{+w}/n_{++}) \cdot n_{w+} + (n_{+n}/n_{++}) \cdot n_{n+}\}/n_{++} \qquad (*)$$

Cohens κ is nu gedefinieerd als het feitelijk gerealiseerde surplus in de diagonaalcellen ten opzichte van het te verwachten aantal, gedeeld door het maximaal mogelijke surplus, dat wil zeggen:

$$\kappa = \{(n_{ww} + n_{nn})/n_{++} - \kappa_e\} / (1 - \kappa_e) \qquad (**)$$

Kappa kan ten hoogste 1 worden, namelijk wanneer beide beoordelaars exact dezelfde vonnissen als wel of niet voldoende gemotiveerd beschouwen (de buiten-diagonaalcellen bevatten dan nullen). Als daarentegen de tabel juist de onder 'geen verband' verwachte aantallen bevat, dan is $\kappa = 0$. Kappa kan zelfs onder nul zakken als de beoordelaars er in zekere mate tegengestelde meningen op na houden. Merk op dat in deze relatieve methode, ook als κ bevredigend groot is, niet wordt aangetoond dat de beoordelaars geen fouten maken, maar als ze fouten maken, dan komen die fouten in behoorlijke mate overeen.

Standaarden voor kappa

Wanneer noemen we een Cohens κ groot genoeg om van een voldoend betrouwbare beoordeling te spreken? Daarvoor zijn wat conventies in omloop. Nogal eens wordt als vuistregel genoemd dat men $\kappa \geq 0.8$ als 'goed' beschouwt, $\kappa \approx 0.7$ als 'redelijk' en $\kappa \leq 0.6$ als 'onvoldoende' kenmerkt, maar Landis & Koch (1977) zijn wat liberaler. Zij noemen reeds een $\kappa \approx 0.4$ 'fair', een $\kappa \approx 0.6$ 'moderate' en een $\kappa \approx 0.8$ 'substantial'. In zekere zin zijn deze maatstaven nogal verrassend, omdat welke κ men als goed kan beschouwen mijns inziens niet onafhankelijk hoort te zijn van de *mate van voorkomen* van het te klasseren kenmerk. Wanneer er sprake is van een scheve verdeling (relatief erg veel of juist erg weinig 'wel'-scores) is het veel moeilijker om een hoge κ te bereiken, dan als het aantal 'wel'-scores in de buurt van de helft ligt. Ik kom daar aanstonds op terug.

Hoe verhouden zich nu de kwaliteitsmaatstaven volgens de absolute methode, in termen van specificiteit en gevoeligheid, met die welke in gebruik zijn bij de relatieve methode voor dubbelbeoordelingen in termen van Cohens κ? Ik zal dat eens nagaan voor een eenvoudig

bijzonder geval, waarin beide beoordelaars in de dubbelbeoordeling even bekwaam zijn, in vergelijking met de absolute beoordeling, en waar bovendien gevoeligheid en specificiteit even groot zijn, dat wil zeggen dat $\varepsilon=\zeta$. We bekijken daarvoor een beoordeling die volgens de absolute methode in een fractie β van de gevallen tot het oordeel 'wel' moet leiden. In dat geval wordt voor beoordelaar i (i=1,2) in tabel III gegeven welke fractie van het totaal der beoordelingen wel en niet overeenkomen met de absolute maatstaf.

Tabel III: fracties beoordelaars/objectief		beoordelaar i	
		'wel'	'niet'
objectieve toets:	'wel'	$\beta\varepsilon$	$\beta(1-\varepsilon)$
	'niet'	$(1-\beta)(1-\varepsilon)$	$(1-\beta)\varepsilon$

Op grond van tabel III kunnen we dan vervolgens afleiden hoe vaak de beoordelaars het eens zullen zijn. Zo kunnen zij beiden 'wel' scoren (linksboven cel) omdat hen een echt 'wel'-geval wordt voorgelegd (dat gebeurt in een fractie β van de gevallen), dat zij beiden terecht herkennen met kans ε, tezamen een kans $\varepsilon^2\beta$, maar ook omdat hen een 'niet'-geval wordt voorgelegd (fractie $1-\beta$), dat zij beiden ten onrechte verkeerd klasseren (met kans $(1-\varepsilon)$ elk), resulterend in een kans $(1-\varepsilon)^2(1-\beta)$. Ook de andere cellen kunnen zo worden bepaald:

Tabel IV: fracties bij dubbelbeoordeling (geval $\varepsilon=\zeta$)		beoordelaar 2	
		'wel'	'niet'
beoordelaar 1	'wel'	$\varepsilon^2\beta + (1-\varepsilon)^2(1-\beta)$	$\varepsilon(1-\varepsilon)$
	'niet'	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)^2\beta + \varepsilon^2(1-\beta)$

Het is een kwestie van eenvoudige doch moeizame algebra om uit te schrijven hoe Cohens κ, gedefinieerd volgens (**) in termen van de cellen uit tabel IV luidt. Het resultaat is:

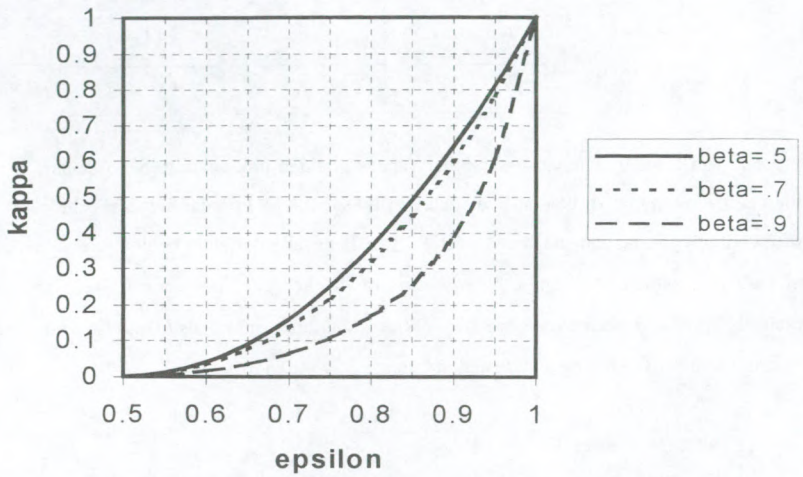
$$\kappa = 1 / \{1 + \varepsilon(1-\varepsilon)(1-2\varepsilon)^{-2}\beta^{-1}(1-\beta)^{-1}\} \qquad (***)^1.$$

---

[1] De algemene formule voor κ luidt als volgt. Zij ε_i de gevoeligheid van de beoordelingsprocedure door beoordelaar i, en ζ_i de specificiteit, voor i=1,2, dan geldt:

Onmiddellijk valt op dat κ niet alleen een functie is van de kwaliteitsmaatstaf ε uit de absolute beoordeling, maar óók van de verdeling van de wel/niet-beoordelingen, β. Het lijkt daarom zaak de vorm van (***) eens aan een nader onderzoek te onderwerpen. Figuur 1 laat zien hoe de relatie tussen κ en ε is, voor enkele waarden van β.

Figuur 1: Cohens kappa als functie van epsilon, voor diverse beta



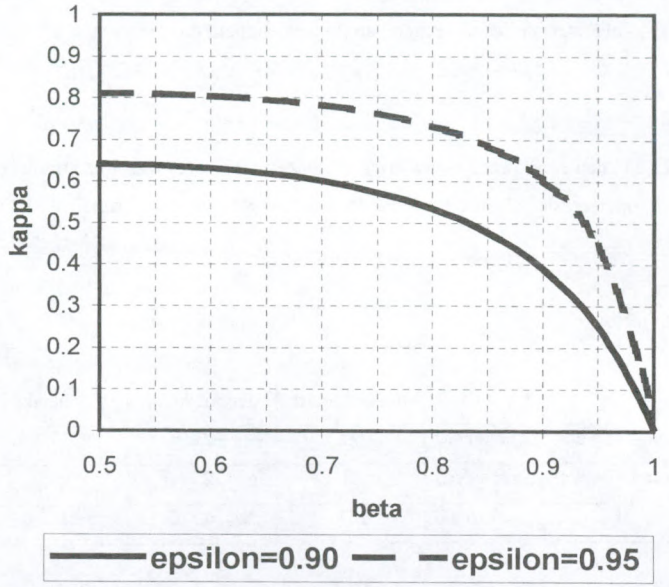
Allereerst is duidelijk dat de parallellie tussen de maatstaven voor voldoende betrouwbaar in een absolute vergelijking ($\varepsilon \approx 0.9$) en in een dubbel-beoordeling ($\kappa > 0.6$) redelijk opgaat voor het geval van β in de buurt van een half, waarbij afwijkingen tot $\beta \approx 0.7$ nauwelijks verschil maken. Maar als we met een verschijnsel te maken krijgen dat scheef verdeeld is (β in de buurt van 1, of, uit symmetrieoverwegingen, β in de buurt van 0), dan is de parallellie veel minder. Zelfs indien de absolute procedure de zeer bevredigende gevoeligheid en specificiteit van $\varepsilon=\zeta=0.95$ heeft, blijft de waarde van κ in dat geval onder de 0.6, de waarde die volgens Landis & Koch als matig, volgens de gangbare vuistregel zelfs als onbevredigend moet worden beschouwd. En indien de absolute procedure wat minder betrouwbaar, maar altijd nog

---

$\kappa = [1 + \{(\varepsilon_1 + \varepsilon_2 - 2\varepsilon_1\varepsilon_2)/(1-\beta) + (\zeta_1 + \zeta_2 - 2\zeta_1\zeta_2)/\beta\}/\{2(1-\varepsilon_1-\zeta_1)(1-\varepsilon_2-\zeta_2)\}]^{-1}$
Deze formule leent zich, door zijn afhankelijkheid van vijf parameters, minder voor eenvoudige weergave à la figuur 1.

heel goed zou zijn, bijvoorbeeld ε=0.90, dan zakt κ zelfs onder de 0.40. De moraal van deze vergelijking tussen de conventionele standaarden bij absolute en relatieve methoden is, dat de heersende standaarden voor κ wat aan de strenge kant schijnen te zijn, en dat het in ieder geval zaak is om de fractie van voorkomen van een positieve beoordeling, β, mede in de beschouwing te betrekken. Als we genoegen nemen met een alleszins fraaie ε in de absolute procedure van 0.9, dan laat figuur 1 zien dat dat dan correspondeert met een $\kappa \approx 0.6$, wanneer we met een niet te scheve verdeling van wel/niet-antwoorden zitten, terwijl voor schevere verdelingen een lagere κ al als bevredigend kan worden beschouwd.

Figuur 2: Cohens kappa* als functie van *beta* voor enkele waarden van *epsilon



Men kan uiteraard figuur 1 ook interpreteren als een methode om, wanneer men dubbel-beoordelingen heeft en daaruit een waarde van κ heeft bepaald, om –onder inachtneming van de geldende waarde van β– daaruit een gevolgtrekking te maken over de waarde van ε die bij een absolute beoordeling zou optreden. Hierbij dienen we ons wel bewust te zijn van het feit dat de betrekking tussen β, ε en κ alleen geldt wanneer beide beoordelaars even goed zijn in het

beoordelen én bovendien het herkennen van het aanwezig zijn van een kenmerk voor beiden even moeilijk is als het herkennen van het ontbreken van dat kenmerk. Onder die omstandigheden kan men figuur 1 gebruiken om af te leiden welke waarde van ε zou gelden, als een absolute standaard in het gegeven geval denkbaar is.

Het is in figuur 1 niet zo gemakkelijk te zien welke β precies met welke κ correspondeert, en daarom presenteren we ook figuur 2, waarin κ als functie van β wordt afgebeeld, voor enkele waarden van ε, uiteraard ook weer gebaseerd op de relatie (***). In deze figuur zien we dat voor waarden van β tussen 0.5 en 0.7 (vanwege de symmetrie in β rond het punt β=0.5 geldt dat dan feitelijk voor $0.3 < \beta < 0.7$), Cohens κ niet zoveel verandert, maar dat het voor grotere β inderdaad zaak is de eisen aan κ niet al te hoog te stellen, willen we niet uit de pas lopen ten opzichte van de absolute standaarden in termen van de gevoeligheid en specificiteit ε .

Voorbeeld

Uit Elffers (1991:137) nemen we een (samenvatting) van een tabel over waarin de oordelen van twee onafhankelijk opererende belastingambtenaren over een groep van 165 aangiften IB wordt gegeven.

Tabel V: voorbeeld aangifte IB		belastingambtenaar 2		
		niet gecorrigeerd	wel gecorrigeerd	totaal
belasting-ambtenaar 1	niet gecorrigeerd	97	22	119
	wel gecorrigeerd	21	25	46
	totaal	118	47	165

Voor deze tabel geldt, hetgeen op grond van (**) te berekenen valt, dat $\kappa = 0.36$ en dat een schatting voor β, de fractie wel gecorrigeerden, op grond van de marges, uitkomt op β=0.73. (We gaan hier nu niet in op het vraagstuk hoe β te schatten als beide beoordelaars een heel verschillend aantal gecorrigeerden zouden aanwijzen. Binnen onze veronderstelling dat beide beoordelaars gelijke specificiteit en gevoeligheid hebben mag dat niet voorkomen.). κ voldoet daarmee niet aan conventionele maatstaven, zoals de auteur ook besprak. Evenwel is β extremer

dan we in bovenstaande aanbeveling hebben verwoord. Uit figuur 1 blijkt dat voor deze waarden van κ en β een ε van iets boven de 0.8 hoort (als we mogen aannemen dat beide beoordelaars even goed zijn, en fouten de ene of de andere kant even vaak voorkomen, hetgeen bij belastingaangiften zeker niet vanzelf spreekt). De conclusie van Elffers (1991), gebaseerd op κ alleen, zonder te kijken naar β, was misschien we wat te streng, maar in het licht van de gevonden schatting van ε (die beduidend onder de standaard van 0.9 ligt) toch ook na deze verfijndere analyse alleszins verdedigbaar.

Conclusie

Voor het nagaan van beoordelaarsbetrouwbaarheid van een procedure om eenheden van analyse in twee categorieën in te delen wordt veelal Cohens κ bepaald op grond van de analyse van onafhankelijke dubbelbeoordelingen. Men doet er wijs aan niet direct te schrikken van lage κ's –in termen van de gebruikelijke conventies– wanneer het verschijnsel dat wordt geklasseerd erg scheef verdeeld is. Dan kan een relatief lage κ voorkomen, ook wanneer de beoordelingsprocedure in bevredigende mate gevoelig en specifiek is. Met behulp van de gegeven grafieken en formules kan, onder zekere voorwaarden, de relatie tussen gevoeligheid en specificiteit, Cohens kappa en het voorkomen van het kenmerk worden gelegd.

Literatuur

Boender, K. (1985) *Milieuprotest in Rijnmond. Sociologische analyse van milieusolidariteit onder elites en publiek.* Rijswijk: Sijthoff Pers.

Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37-46.

Elffers, H. (1985), Over de betrouwbaarheid van de meetmethode van milieusolidariteit. In: K. Boender, *Milieuprotest in Rijnmond. Sociologische analyse van milieusolidariteit onder elites en publiek.* Rijswijk: Sijthoff Pers, 1985, 654-658.

Elffers, H. (1991). *Income tax evasion, theory and measurement.* Deventer: Kluwer.

Elffers, H. (2000). Betrouwbaar indelen in categorieën. Hoe hoog moet Cohens kappa zijn? In: Elffers, H. (red.) *Libris satiari nequeo. Afscheidsbundel voor Kees Boender..* Rotterdam: Onderzoekschool Maatschappelijke Veiligheid, 2000, 21-28. ISBN 90-5886-005-1.

Landis, J.R. en G.G. Koch (1977): The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

Ontvangen: 1 december 2000
Geaccepteerd: 6 februari 2001