HOE HOOG BEHOORT EEN GENERALISEERBAARHEIDSKOEFFICIENT TE ZIJN?

Henk Elffers
Geografisch Instituut R.U.Utrecht
Heidelberglaan 2, Utrecht

Paper gepresenteerd
Onderwijsresearchdagen
Nijmegen, 1979

Inleiding

In de generaliseerbaarheidstheorie in de lijn van CRONBACH e.a. (1) voert men, om de kwaliteiten van een meetinstrument te onderzoeken een generaliseerbaarheids (G-)onderzoek uit teneinde, uitgaande van een lineair model voor de skores, voldoende gegevens over diverse variantiekomponenten te verzamelen om bij het plannen van een onderzoek waarin het instrument zal worden toegepast (D-onderzoek) de prestaties van het instrument te kunnen voorspellen. Daartoe berekent men voor elke overwogen opzet van het D-onderzoek een generaliseerbaarheidskoefficient, ρ^2. Evenwel bestaat er geen duidelijkheid over hoe hoog dergelijke koëfficienten nu eigenlijk moeten zijn. Er bestaat wel een, per vakgebied overigens verschillende, communis opinio, maar die schijnt meer gebaseerd te zijn op het feit dat reele studies vaak waarden in die buurt opleveren, dan op een interpretatie van die waarden zelf. In dit paper wordt een interpretatie van de generaliseerbaarheidskoëfficient gegeven voor vergelijkend gebruik van een meetinstrument, in termen van de kans op diverse fouten. Daaraan wordt een norm voor de hoogte van ρ^2 ontleend. Er zij opgemerkt dat men vergelijkbare foutenkansinterpretaties ook voor de gewone korrelatiekoëfficient kan ontwikkelen (ELFFERS (2)).

D-onderzoek

In alle D-onderzoeken voor vergelijkende metingen kunnen we het verschil $\underline{\Delta}_O$ tussen de skores van twee willekeurig gekozen te vergelijken objekten, zoals gemeten volgens de onderzoeksopzet, splitsen in een relevant deel $\underline{\Delta}_R$ en een niet-relevant deel $\underline{\Delta}_N$, $\underline{\Delta}_O = \underline{\Delta}_R + \underline{\Delta}_N$.

VOORBEELD: Bij een kursus voor sociaal werkers gebruikt men een instrument dat 'bedillerigheid' van sociaal werkers in een gesprek met een klient (gespeeld door een van de kursusleiders) bepaalt via observatie (door één uit een team observators). In het G-onderzoek is men uitgegaan van het model voor de skore x van een sociaal werker s in gesprek met klient k geobserveerd door observator o (random variabelen worden onderstreept):

$$\underline{x}(s,k,o) = \mu + S(s) + K(k) + O(o) + SK(s,k) + \underline{e}$$

waarin S, K, O de effekten van de diverse facetten zijn, SK het interaktieeffekt van sociaal werker en klient, $\underline{e}$ de random veronderstelde residuele invloed. (Voor het gemak is hier observatieinteraktie afwezig verondersteld. Dit is niet nodig.)

Eén mogelijk D-onderzoek wil kursisten vergelijken op grond van één gesprek met een willekeurig gekozen klient, geobserveerd door één willekeurig gekozen observator. Het geobserveerde verschil tussen willekeurig gekozen sociaal werkers $\underline{s}_1$ en $\underline{s}_2$ is dan

$$\underline{\Delta}_O = \underline{x}(\underline{s}_1,\underline{k}_1,\underline{o}_1) - \underline{x}(\underline{s}_2,\underline{k}_2,\underline{o}_2) \quad ,$$

het relevante verschil is

$$\underline{\Delta}_R = S(\underline{s}_1) - S(\underline{s}_2) \quad ,$$

en het niet-relevante verschil is dan

$$\underline{\Delta}_N = (K(\underline{k}_1) - K(\underline{k}_2)) + (O(\underline{o}_1) - O(\underline{o}_2)) + (SK(\underline{s}_1,\underline{k}_1) - SK(\underline{s}_2,\underline{k}_2)) + (\underline{e}_1 - \underline{e}_2) \quad .$$

Een ander D-onderzoek wil verschillende kursisten elk met twee klienten laten spreken, willekeurig gekozen, telkenmale geobserveerd door een en dezelfde observator, o. Dan geldt:

$$\underline{\Delta}_O = \tfrac{1}{2}(\underline{x}(\underline{s}_1,\underline{k}_1,o) + \underline{x}(\underline{s}_1,\underline{k}_2,o)) - \tfrac{1}{2}(\underline{x}(\underline{s}_2,\underline{k}_3,o) + \underline{x}(\underline{s}_2,\underline{k}_4,o))$$

$$\underline{\Delta}_R = S(\underline{s}_1) - S(\underline{s}_2)$$

$$\underline{\Delta}_N = \tfrac{1}{2}(K(\underline{k}_1) + K(\underline{k}_2) - K(\underline{k}_3) - K(\underline{k}_4)) + \tfrac{1}{2}(SK(\underline{s}_1,\underline{k}_1) + SK(\underline{s}_1,\underline{k}_2) - SK(\underline{s}_2\underline{k}_3) - SK(\underline{s}_2,\underline{k}_4)) + \tfrac{1}{2}(\underline{e}_1 + \underline{e}_2 - \underline{e}_3 - \underline{e}_4)$$

En zo kan men nog vele D-onderzoeken beschrijven, waarbij telkenmale geldt dat $\underline{\Delta}_R$ en $\underline{\Delta}_N$ ongekorreleerd zijn, en dat $\sigma^2(\underline{\Delta}_O)$, $\sigma^2(\underline{\Delta}_R)$ en $\sigma^2(\underline{\Delta}_N)$ uit het G-onderzoek kunnen worden afgeleid (vergelijk ELFFERS & TAVECCHIO (3)). Willen we nu de bedillerigste kursisten eruithalen, of in het algemeen objekten onderscheiden, dan moeten we ten minste elk willekeurig tweetal met grote kans goed rangordenen. Dat dient eigenlijk te geschieden op grond van hun $\underline{\Delta}_R$, maar we zullen slechts $\underline{\Delta}_O$ kunnen observeren. Wat is nu de kans dat het dan fout gaat ?

Naieve foutenkans

Eerst bepalen we de kans op een fout bij naief gebruik van het instrument. Met naief gebruik bedoelen we dat we elk tweetal objekten, gemeten volgens de D-onderzoeksopzet, rangordenen zodra hun $\underline{\Delta}_O$ van nul verschilt. Dit is naief omdat miskend wordt dat een betrouwbare ordening op grond van $\underline{\Delta}_R$ zou moeten geschieden, en $\underline{\Delta}_O$ daarvan zeker afwijkt. Als maat voor wanprestatie van deze manier van doen bij deze opzet van het D-onderzoek bepalen we nu de kans dat een willekeurig getrokken tweetal te meten objekten verkeerd geordend wordt. Er wordt zo'n fout gemaakt als $\underline{\Delta}_O$ en de onbekende $\underline{\Delta}_R$ een verschillende kant opwijzen, dus als hun teken verschillend is, dus als $\underline{\Delta}_O \cdot \underline{\Delta}_R < 0$.
De kans op zo'n fout noemen we de naieve foutenkans n.e.p.

$$n.e.p = P(\underline{\Delta}_O \cdot \underline{\Delta}_R < 0)$$

Een flink gedeelte van de fouten die door de naieve foutenkans worden gekwantificeerd valt ook door de betrouwbaarste instrumenten niet te vermijden, namelijk die waar $\underline{\Delta}_R$ slechts zeer weinig van nul verschilt. Bovendien is het veelal niet zo erg dat objekten die zo weinig verschillen niet op de juiste wijze gerangordend worden. Daarom dienen we de grens voor een aanvaardbare foutenkans niet te streng te stellen, bijvoorbeeld rond de 15 %.

Scheidend vermogen

Sophisticated gebruik van het instrument vermijdt fouten wegens minieme relevante verschillen door te stellen: orden twee objekten alleen als je er redelijk zeker van kunt zijn dat ze echt in die richting verschillen, d.w.z. stel een drempelwaarde $d(\alpha)\cdot\sigma(\underline{\Delta}_0)$ vast, zodanig dat alleen objekten die niet minder dan de drempelwaarde verschillen geordend worden; in het andere geval doen we geen uitspraak. (De drempelwaarde hangt van $\sigma(\underline{\Delta}_0)$ af, omdat $\underline{\Delta}_0$ slechts groot of klein genoemd kan worden in vergelijking met zijn standaarddeviatie.) Kies deze zogenaamde scheidingsdrempel nu zo, dat objekten met $\underline{\Delta}_0$ precies op de drempel met vaste, kleine kans α verkeerd geordend worden; d.w.z. de scheidingsdrempel wordt bepaald door:

$$P(\underline{\Delta}_0 \cdot \underline{\Delta}_R < 0 \mid |\underline{\Delta}_0| = d(\alpha)\cdot\sigma(\underline{\Delta}_0)) = \alpha$$

Te denken valt aan α = 5 % of 10 %.
Deze procedure leidt ertoe dat geregeld paren objekten niet geordend zullen kunnen worden. Een kwaliteitsmaat voor een D-opzet is nu in hoeveel van de gevallen deze sophisticated procedure met beperkte fout nog tot een beslissing komt, dat wil zeggen

$$P(|\underline{\Delta}_0| \geq d(\alpha)\cdot\sigma(\underline{\Delta}_0))$$

een grootheid die we het α-scheidend vermogen $dp(\alpha)$ zullen noemen. Een redelijke waarde voor het scheidend vermogen is bijvoorbeeld 50 %.

Relatie met de generaliseerbaarheidskoëfficient

Het is eenvoudig in te zien dat naieve foutenkans en scheidend vermogen worden bepaald door de verhouding $Q = \sigma^2(\underline{\Delta}_R) / \sigma^2(\underline{\Delta}_N)$, de kwaliteitsratio. Immers bij grote $\sigma^2(\underline{\Delta}_N)$ overschaduwt $\underline{\Delta}_N$ $\underline{\Delta}_R$ vaak, te meer als $\underline{\Delta}_R$ zelf dicht bij 0 ligt, d.w.z. zelf kleine variantie heeft.
Als we aannemen dat $(\underline{\Delta}_R, \underline{\Delta}_N)$ bivariaat normaal verdeeld is (geen gratuite aanname, maar in de kontekst van een lineair model zeker voor verschilskores niet al te ongebruikelijk; in ieder geval geeft het een indruk), blijken beide kansen geheel bepaald te zijn door Q en α.

Q is uit het G-onderzoek af te leiden, en daar geldt dat $\rho^2 = Q / (Q + 1)$, kunnen we de resultaten ook uitdrukken in ρ^2. Enige algebra (ELFFERS & TAVECCHIO(3)) laat zien dat (Φ is de verdelingsfunktie van de standaardnormale verdeling)

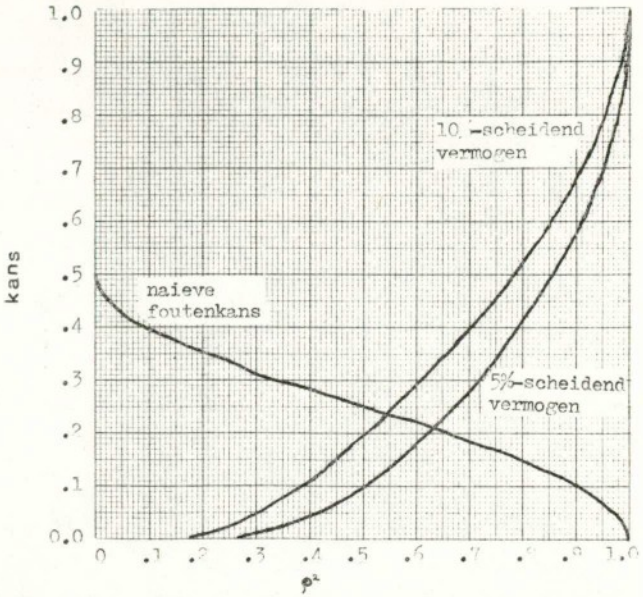
$$n.e.p = \arctan(Q^{-\frac{1}{2}})/\pi = \arctan((\rho^{-2} - 1)^{\frac{1}{2}})/\pi =$$
$$= \tfrac{1}{2} - \arcsin(\rho)/\pi$$
$$d(\alpha) = -\Phi^{-1}(\alpha)\cdot Q^{-\frac{1}{2}} = -\Phi^{-1}(\alpha)\cdot(\rho^{-2} - 1)^{\frac{1}{2}}$$
$$dp(\alpha) = 2\Phi(\Phi^{-1}(\alpha)\cdot Q^{-\frac{1}{2}}) = 2\,\Phi(\Phi^{-1}(\alpha)\cdot(\rho^{-2} - 1)^{\frac{1}{2}})$$

Minimaal aanvaardbare waarden voor de generaliseerbaarheids-koëfficient

Het is aan te raden telkenmale bij de beoordeling van een D-opzet de naieve foutenkans of het scheidend vermogen te bepalen. Evenwel kunnen we ons ook in het algemeen afvragen: welke waarden van Q of ρ^2 leiden tot aanvaardbare waarden voor deze kriteria ? Dat is natuurlijk afhankelijk van het doel van het D-onderzoek, en blijft ook gedeeltelijk een kwestie van smaak. Bestudering van figuur 1 en tabel I leidt mij tot de aanbevelingen in tabel II.

Q	ρ	ρ^2	n.e.p	d(.1)	dp(.1)
2	.82	.67	.20	.91	.36
4	.89	.80	.15	.64	.52
9	.95	.90	.10	.43	.67

Tabel I : Enige waarden van Q, ρ, ρ^2, n.e.p, d(.1) en dp(.1) voor normaal verdeelde ($\underline{\Delta}_R$, $\underline{\Delta}_N$)



Figuur 1: Naieve foutenkans en scheidend vermogen als funktie van ρ^2 voor normaal verdeelde $(\underline{\Delta}_R, \underline{\Delta}_N)$

Q uit	ρ^2 uit	kwaliteitsoordeel
[0, 2)	[0, .67)	slecht
[2, 4)	[.67, .80)	net voldoende
[4, 9)	[.80, .90)	redelijk
[9, ∞)	[.90, 1]	goed

Tabel II : Kwaliteitsoordeel over een D-opzet als funktie van Q of ρ^2

Literatuur

(1)L.J.Cronbach,G.C.Gleser,H.Nanda & N.Rajaratnam: The dependability of behavioral measurements (New York,1972)

(2)H.Elffers: On interpreting the product moment correlation coefficient. (Aangeboden aan Statistica Neerlandica).

(3)H.Elffers & L.W.C.Tavecchio: Variance components in test generalizability research: which, when, why? (zal verschijnen als VOR-publikatie, Vereniging voor Onderwijsresearch).