

## Statistiek en meten: wat moet je daarover weten?

Onder redactie van Diana Dolmans, Cees van der Vleuten, Albert Scherpbier en Ineke Wolfhagen

*Bij het lezen van publikaties over onderzoek van onderwijs worden lezers regelmatig geconfronteerd met allerlei statistische begrippen, waarvan de betekenis voor veel lezers niet geheel duidelijk is. Daarom heeft de redactie besloten om een reeks artikelen te publiceren waarin belangrijke onderwerpen op het gebied van statistiek en meten aan de orde worden gesteld. De redactie van deze reeks wordt gevormd door twee gastredacteuren (Diana Dolmans en Cees van der Vleuten) en twee leden van de BMO-redactie (Albert Scherpbier en Ineke Wolfhagen). De keuze van onderwerpen is gebaseerd op veel gebruikte statistische technieken en begrippen. De nadruk ligt op de betekenis en interpretatie hiervan. Op deze manier wordt getracht een bijdrage te leveren aan de verdere professionalisering van docenten binnen het medisch onderwijs. In het vierde artikel van deze reeks staat variantie-analyse centraal.*

### Variantie-analyse

P. Frijns

*Termen die aan bod komen: variantie, binnen-groepenvariantie (MS-binnen), tussen-groepenvariantie (MS-tussen), F-toets, vrijheidsgraden (df)*

Het doel van tal van onderzoeken is te bepalen of er al dan niet een (causale) relatie bestaat tussen twee of meer variabelen. Zo kan iemand geïnteresseerd zijn in de invloed van het aantal college-uren op de hoogte van het tentamencijfer.

Om deze relatie te bestuderen wordt vaak gebruik gemaakt van een opzet waarbij twee groepen met elkaar worden vergeleken: groep 1 krijgt 25 uur college, terwijl aan groep 2 - die qua kenmerken vergelijkbaar is met groep 1 - 5 college-uren worden aangeboden. Met behulp van de tentamencijfers van de personen uit beide groepen kan worden gekeken in hoeverre de twee groepen van elkaar verschillen. Indien het gemiddelde tentamencijfer van groep 1 significant afwijkt van dat van groep 2, wordt gesteld dat het aantal college-uren een

significante invloed heeft op de hoogte van het tentamenresultaat.

De vraag is echter, hoe kan worden vastgesteld of de groepen daadwerkelijk van elkaar verschillen. Of strikter geformuleerd, hoe weten we of de verschillen tussen de groepen groot genoeg zijn om te kunnen spreken van echte verschillen, die niet zijn toe te schrijven aan toeval. Om dit te bepalen wordt in de praktijk geregeld gebruikgemaakt van variantie-analyse.

#### Wat is variantie?

Om de kenmerken van een meting uit te drukken wordt in onderzoeksrapportages en vakliteratuur regelmatig naast het gemiddelde de bijbehorende variantie vermeld. Hierbij wordt de variantie opgevat als een numerieke maat voor de mate waarin de scores afwijken van het gemiddelde.

De reden om beide getallen te vermelden komt voort uit het gegeven dat het gemiddelde

alleen onvoldoende informatie oplevert. Stel bijvoorbeeld dat in een onderzoek de volgende scores worden gevonden:

Groep 1:	3	3	4	4	5	5
Groep 2:	2	1	5	3	7	6

Indien alleen op basis van het gemiddelde zou moeten worden besloten of groep 1 afwijkt van groep 2, dan zou de conclusie 'nee' zijn. Want bij zowel groep 1 als groep 2 is het gemiddelde gelijk aan 4. Toch zijn beide groepen niet identiek. Er bestaan duidelijke verschillen in de verdeling van de scores. Zo zijn er verschillen in het aantal keren dat een bepaalde score voorkomt. Maar ook is bij groep 2 de spreiding van de scores aanzienlijk groter dan bij groep 1.

De vraag is op welke wijze dit verschil in een numerieke maat kan worden vastgelegd. In principe zijn er verschillende methoden om de spreiding in scores uit te drukken. Per groep kunnen de individuele scores vergeleken worden met het groepsgemiddelde. Dit kan bijvoorbeeld door van elke score het groepsgemiddelde af te trekken. Voor de groepen in ons voorbeeld levert dit het volgende resultaat op:

Groep 1:	-1	-1	0	0	1	1
Groep 2:	-2	-3	1	-1	3	2

Aangezien op alle individuele scores het groepsgemiddelde in mindering is gebracht, geldt voor elke groep dat de som van de afwijkingen per definitie gelijk is aan 0. Kortom, ook met deze op het eerste gezicht bruikbare werkwijze kan het verschil niet in één getal worden weergegeven. Het probleem bij deze werkwijze is dus dat de afwijkingen tegen elkaar worden weggemiddeld. Om dit 'wegmiddelen' tegen te gaan kunnen twee strategieën worden gevolgd: (1) het absoluut maken van de verschillen en (2) het kwadrateren van de verschillen.

De methodiek van het absoluut maken van de afwijkingen houdt in dat zowel de positieve als de negatieve afwijkingen worden weerge-

geven als een positief getal. In ons voorbeeld levert dit het volgende resultaat op:

Groep 1:	1	1	0	0	1	1
Groep 2:	2	3	1	1	3	2

Als voor elke groep de som van de afwijkingen wordt bepaald, wordt inderdaad een verschil gevonden: bij groep 1 is de som van de afwijkingen gelijk aan 4, terwijl bij groep 2 de som gelijk aan 12 is.

Het kwadrateren van de verschillen tussen individuele scores en het groepsgemiddelde is de tweede methode om het wegmiddelen te voorkomen. Voor ons voorbeeld betekent dat:

Groep 1:	1	1	0	0	1	1
Groep 2:	4	9	1	1	9	4

Als vervolgens de afwijkingen worden gesommeerd, blijken de verschillen tussen beide groepen (nog) groter te zijn: bij groep 1 is de som gelijk aan 4, terwijl bij groep 2 de som gelijk is aan 28. Het getal dat wordt verkregen als bij deze laatste methodiek de som van de verschillen wordt gedeeld door het aantal scores binnen de groep wordt de steekproefvariantie genoemd. De methode die leidt tot de variantie wordt met de term variantie-analyse aangeduid.

### Hoe toets je de verschillen tussen groepen?

In de vorige paragraaf is ingegaan op de wijze waarop een groep kan worden gekarakteriseerd in termen van gemiddelde en variantie. Bij de beschrijving is uitgegaan van twee groepen met eenzelfde gemiddelde. In de praktijk wordt variantie-analyse echter vaak gebruikt om te bestuderen in hoeverre twee groepen van elkaar verschillen voor wat betreft het gemiddelde. Om hierover een uitspraak te kunnen doen, moet worden gekeken naar de spreiding van de scores rondom het groepsgemiddelde alsmede naar de mate waarin de scores van beide groe-

pen elkaar overlappen (dezelfde scores in beide groepen voorkomen).

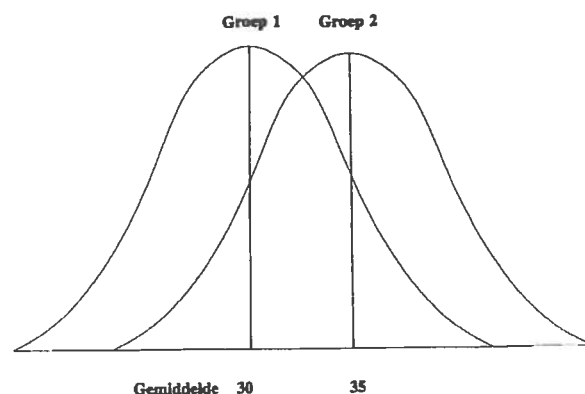
Indien een grote overlap tussen de scores van beide groepen wordt geconstateerd (figuur 1), wordt verondersteld dat het gevonden verschil tussen de groepsgemiddelden op toeval berust. Naarmate het verschil tussen de gemiddelden toeneemt en de mate van overlap tussen de scores van beide groepen afneemt (figuur 2), stijgt de kans dat er een significant verschil tussen beide groepen bestaat. De vraag is echter wat is een kleine en wat is een grote overlap? Oftewel, op welke wijze kan worden bepaald of het verschil tussen de gemiddelden significant is of op toeval berust?

Stel de volgende fictieve situatie. Een arts is geïnteresseerd in het verband tussen het al dan niet hebben gehad van een hartinfarct en de frequentie van de hartslag. Om dit te onderzoeken voert hij de volgende studie uit. Uit de groep patiënten in Nederland die een hartinfarct hebben gehad trekt hij at random vijf personen. Daarnaast trekt hij at random vijf personen uit de groep Nederlanders die nog nooit een hartinfarct hebben gehad. Vervolgens wordt bij alle betrokkenen de frequentie van de hartslag gemeten.

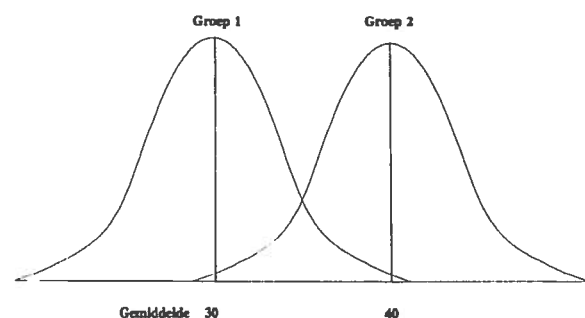
Om te bestuderen of er een significant verschil bestaat tussen beide groepen wordt gebruikgemaakt van variantie-analyse. In bovenstaand voorbeeld kunnen twee soorten varianties worden onderscheiden: (1) de spreiding van de gemeten hartslag binnen elke groep en (2) de spreiding van de gemiddelde score van elke individuele groep ten opzichte van het totaal gemiddelde gebaseerd op alle scores uit de twee groepen. Dit betekent dat de totaal geobserveerde spreiding in hartslag bestaat uit de spreiding in hartslag binnen de groepen en de spreiding in hartslag tussen de groepen. Oftewel:

$$SS(\text{totaal}) = SS(\text{binnen}) + SS(\text{tussen})$$

We zijn niet alleen geïnteresseerd in de verschillen tussen de twee steekproeven. We zouden ook op basis van de steekproefgegevens



Figuur 1. Grote overlap tussen de scores van groep 1 en 2



Figuur 2. Kleine overlap tussen de scores van groep 1 en 2

een algemeen geldende uitspraak willen doen over het verschil tussen beide populaties waaruit de steekproeven zijn getrokken. Om dat te kunnen doen moet eerst op basis van de steekproefgegevens een schatting worden gemaakt van de populatie-variantie (zie tekstkader). Pas daarna kan worden getoetst of het verschil tussen de groepen significant is danwel op toeval berust.

Bij het toetsen van het verschil tussen de twee groepen gaat het om de vraag of de variantie tussen de groepen groter is dan de variantie binnen de groepen. Of strikter geformuleerd: is de *tussen-groepenvariantie* groot genoeg ten opzichte van de *binnen-groepenvariantie* om van een significant verschil te kunnen spreken? Hiertoe wordt naar de volgende verhouding gekeken:

$$\frac{\text{tussen-groepvariantie}}{\text{binnen-groepvariantie}} = \frac{\text{MS(tussen)}}{\text{MS(binnen)}}$$

Deze verhouding wordt de F-ratio genoemd. Het zal duidelijk zijn dat deze verhouding groter is naarmate de groepsgemiddelden meer van elkaar verschillen en/of de spreiding van de scores binnen de groepen kleiner is.

De toetsing van het significantie-niveau geschiedt met behulp van de *F-toets*. Deze F-toets volgt een bepaalde statistische verdeling. Hiermee kan worden vastgesteld of een F-waarde groot of klein is. Om uiteindelijk het significantie-niveau te bepalen moet in de F-tabel bij de bijbehorende *vrijheidsgraden* (*degrees of freedom* = *df*) worden gekeken (zie voorbeeld in tekstkader). Indien de gevonden F-ratio groter is dan de F-ratio die in de tabel vermeld wordt, dan is er sprake van een significant verschil.

Het is gebruikelijk de resultaten van de variantie-analyse en de bijbehorende F-toets weer te geven in een samenvattende tabel. Een dergelijke overzicht is weergegeven in tabel 1.

**Tabel 1.** Samenvattende tabel

	SS	df	MS	F-ratio	p
Tussen	65	1	65.0	7.65	.05
Binnen	118	12	8.5		
Totaal	183	13			

Naast een samenvattende tabel wordt ook regelmatig in de tekst van een artikel naar de resultaten van de F-toets verwezen. Normaliter gebeurt dit in de vorm van  $F(1,8)=7.65, p<.05$ ). Deze getallen moeten als volgt worden gelezen:  $F(1,8)=7.65$  betekent dat de gevonden F-ratio gelijk is aan 7.65 en dat het aantal vrijheidsgraden voor groep 1 gelijk is aan 1 en het aantal vrijheidsgraden voor groep 2 gelijk is aan 8. De toevoeging  $p<.05$  geeft aan dat het

verschil tussen de groepen significant is op een 5%-niveau, uitgaande van de gevonden F-ratio met de bijbehorende vrijheidsgraden.

## Meerdere variabelen

In bovenstaand voorbeeld is er sprake van één onafhankelijke variabele (wel/geen hartinfarct gehad) en één afhankelijke variabele (hartslag). Het zou echter heel goed mogelijk zijn dat de onderzoeker in het vorige voorbeeld tevens de invloed van medicijn x op de frequentie van de hartslag wil bestuderen. Om dit te bestuderen dient hij bijvoorbeeld bij drie personen van elke groep medicijn x toe.

Bij deze opzet kunnen vier groepen worden onderscheiden: (1) de groep 'gezonde' personen die medicijn x niet krijgen toegediend, (2) de groep 'gezonde' personen die wel medicijn x krijgen toegediend, (3) de groep patiënten die een hartinfarct hebben gehad en die medicijn x niet krijgen toegediend en (4) de groep patiënten die een hartinfarct hebben gehad en die medicijn x wel krijgen toegediend.

Op basis hiervan kan een aantal aspecten worden bestudeerd. Allereerst, wat is de invloed van het wel of niet gehad hebben van een hartinfarct op de frequentie van de hartslag. Dit kan worden onderzocht door groep (1) te vergelijken met groep (3). Ten tweede kan worden bestudeerd wat de invloed van medicijn x is op de hartslag van 'gezonde' personen. Door de resultaten van groep (1) en groep (2) met elkaar te vergelijken kan op deze vraag een antwoord worden verkregen. Een derde aspect dat kan worden bestudeerd, is de invloed van medicijn x op de frequentie van de hartslag van patiënten die een hartinfarct hebben gehad. Een antwoord op deze vraagstelling wordt verkregen door de groepen (3) en (4) met elkaar te vergelijken. Ten slotte kan worden onderzocht wat de invloed is van zowel het wel of niet hebben gehad van een hartinfarct als medicijn x op de frequentie van de hartslag. Dit aspect kan worden bestudeerd door de resultaten van alle groepen met elkaar te vergelijken. In dit

laatste geval spreken we van een zogenaamd interactie-effect; de invloed van zowel de 'gezondheidstoestand' als medicijn x. Kortom, bij een dergelijke opzet kunnen drie aspecten worden bestudeerd: (1) Wat is de invloed van de gezondheidstoestand van de persoon op de hartslag, gegeven de medicatie? (2) Wat is de invloed van medicijn x op de hartslag, gegeven de gezondheidstoestand? En (3) wat is de invloed van de interactie tussen medicatie en gezondheidstoestand op de hartslag?

Bij het toetsen van elk van deze drie vragen wordt dezelfde redenering gevolgd als bij het vorige voorbeeld: de totale variantie wordt geschat door middel van de binnen-groepenvariantie en de tussen-groepenvariantie, waarna de F-ratio wordt bepaald. Hierbij dient wel te worden opgemerkt dat hetgeen onder binnen-groepenvariantie en tussen-groepenvariantie wordt verstaan verschilt per vraagstelling. Het voert echter te ver om in dit artikel de verschillende varianties te berekenen. Voor diegenen die geïnteresseerd zijn in dergelijke berekeningen wordt in het tekstkader een voorbeeld verder uitgewerkt. Voor verdere details moet worden verwezen naar de boeken die zijn opgenomen in de lijst met aanbevolen literatuur.

## Opdrachten

1. Wat betekent het als de tussen-groepenvariantie groter is dan de binnen-groepenvariantie?
2. Waarom wordt de binnen-groepenvariantie ook wel error-variantie genoemd?

## Aanbevolen literatuur

- Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics. Norwalk/San Mateo: Appleton & Lange, 1990.
- Hays WL. Statistics. New York: Holt-Saunders, International Editions, 1981.
- Henkel RE. Test of significance. London/Beverly Hills: Sage Publications, 1984.
- Slotboom A. Statistiek in woorden. De meest voorkomende termen en technieken. Groningen: Wolters-Noordhoff, 1987.

## DE AUTEUR

*P. Frijns is hoofd van het Bureau Onderwijsontwikkeling van het Koninklijk Instituut voor de Marine te Den Helder. Daarvoor was hij werkzaam als AIO bij de vakgroep Onderwijsontwikkeling en Onderwijsresearch van de Rijksuniversiteit Limburg en als onderwijskundige bij de invoering van een probleemgestuurd curriculum bij de Faculteit der Bouwkunde aan de TU Delft.*

## Correspondentie-adres:

*P. Frijns, Bureau Onderwijsontwikkeling, Koninklijk Instituut voor de Marine, Het Nieuwe Diep 8, 1781 AC Den Helder.*

### Een voorbeeld

Stel dat de onderzoeker in het voorbeeld de volgende resultaten heeft verkregen (zie tabel 2). Om het verschil tussen beide groepen op significantie te toetsen wordt allereerst een schatting van de variantiecomponenten gemaakt. Vervolgens wordt de F-ratio berekend en het significantie-niveau bepaald.

**Tabel 2.** Hartslagfrequentie bij patiënten die een hartinfarct hebben gehad en 'gezonde' personen (per minuut)

	Gezondheidstoestand	
	Patiënten die een hartinfarct hebben gehad	Gezonde personen
Persoon 1	75	81
Persoon 2	70	79
Persoon 3	80	80
Persoon 4	77	78
Persoon 5	73	82
Gemiddelde	75	80

### Schatting van de variantiecomponenten

De schatting van de binnen-groepen- en de tussen-groepenvariantie verloopt in drie stappen. Voor de variantie binnen groepen gaat het als volgt. Allereerst wordt per groep de variantie berekend van de individuele scores ten opzichte van het eigen groeps-gemiddelde. Hiertoe wordt het verschil tussen elke individuele score en het groeps-gemiddelde berekend en gekwadeerd, en worden de individuele resultaten gesommeerd (= Sum of Squares = SS). Tenslotte worden de gesommeerde scores van de afzonderlijke groepen opgeteld, resulterend in de SS(binnen). Toegepast op ons voorbeeld:

patiënten die een hartinfarct hebben gehad:

$$(75 - 75)^2 + (70 - 75)^2 + (80 - 75)^2 + (77 - 75)^2 + (73 - 75)^2 = 58$$

'gezonde' personen:

$$(81 - 80)^2 + (79 - 80)^2 + (80 - 80)^2 + (78 - 80)^2 + (82 - 80)^2 = 10$$

$$SS(\text{binnen}) = 58 + 10 = 68$$

De tweede stap in het schatten van de binnen-groepenvariantie, is het bepalen van het aantal vrijheidsgraden. Het aantal vrijheidsgraden geeft het aantal scores aan dat in principe vrij gekozen kan worden gegeven de situatie. Dus als het gemiddelde van  $N$  waarnemingen bekend is, dan kan  $N-1$  scores vrij worden gekozen, aangezien de laatste score bepaald is door de som van de andere scores en het gemiddelde. In ons geval ligt bij elke groep het gemiddelde vast en kunnen er dus  $5 - 1 = 4$  scores vrij worden gekozen. In totaal zijn er bij het design  $2 \times (5-1) = 8$  vrijheidsgraden.

De laatste stap is het daadwerkelijk schatten van de binnen-groepenvariantie. Door  $SS(\text{binnen})$  te delen door het aantal vrijheidsgraden wordt de zogenaamde  $MS(\text{binnen})$  (= mean sum of squares) verkregen. Voor het voorbeeld geldt dat  $MS(\text{binnen}) = 68/8 = 8.5$ . Dit is het eerste gedeelte van de schatting van de totale populatie-variantie. Nu moet  $MS(\text{tussen})$  nog worden bepaald.

De variantie tussen groepen wordt op gelijksoortige wijze geschat. Allereerst wordt het totaal gemiddelde over alle scores berekend. Vervolgens wordt het verschil tussen elk groepsgemiddelde en het totaal gemiddelde berekend en het verschil gekwadeerd (=  $SS(\text{tussen})$ ). De gevonden variantie is kunstmatig klein, doordat de gemiddelde scores per definitie dichter bij elkaar liggen dan de geobserveerde individuele scores. Vandaar dat de berekende variantie wordt vermenigvuldigd met het aantal observaties per groep om tot een meer zuivere schatting te komen. Vervolgens wordt op identieke wijze als bij de binnen-groepenvariantie het aantal vrijheidsgraden bepaald. Tenslotte wordt  $MS(\text{tussen})$  bepaald door  $SS(\text{tussen})$  te delen door het aantal vrijheidsgraden (df). Voor het beschreven voorbeeld ziet dit er als volgt uit:

Totaal gemiddelde	=	$(75 + 80)/2 = 78$ (eigenlijk 77.5)
$SS(\text{tussen})$	=	$5[(75 - 78)^2 + (80 - 78)^2] = 5 \times 13 = 65$
df	=	$(2 - 1) = 1$
$MS(\text{tussen})$	=	$65/1 = 65$

### Toetsing significantie-niveau

De eerste stap bij het bepalen van het significantie-niveau is het berekenen van de F-ratio. Deze is gelijk aan  $MS(\text{tussen})$  gedeeld door  $MS(\text{binnen})$ . In ons voorbeeld is de F-ratio gelijk aan:  $65/8.5 = 7.65$ . Om te kijken in hoeverre dit betekent dat er een significant verschil tussen de groepen bestaat wordt in de F-tabel gekeken bij de cel behorende bij 1 en 8 vrijheidsgraden. Als de F-ratio groter is dan het getal in de tabel, dan is het verschil significant op het respectievelijke niveau. Als we kijken bij de tabel van het 5%-niveau, dan blijkt dat het gevonden verschil wel significant is op 5%-niveau. Het verschil blijkt echter niet op 1%-niveau significant te zijn.