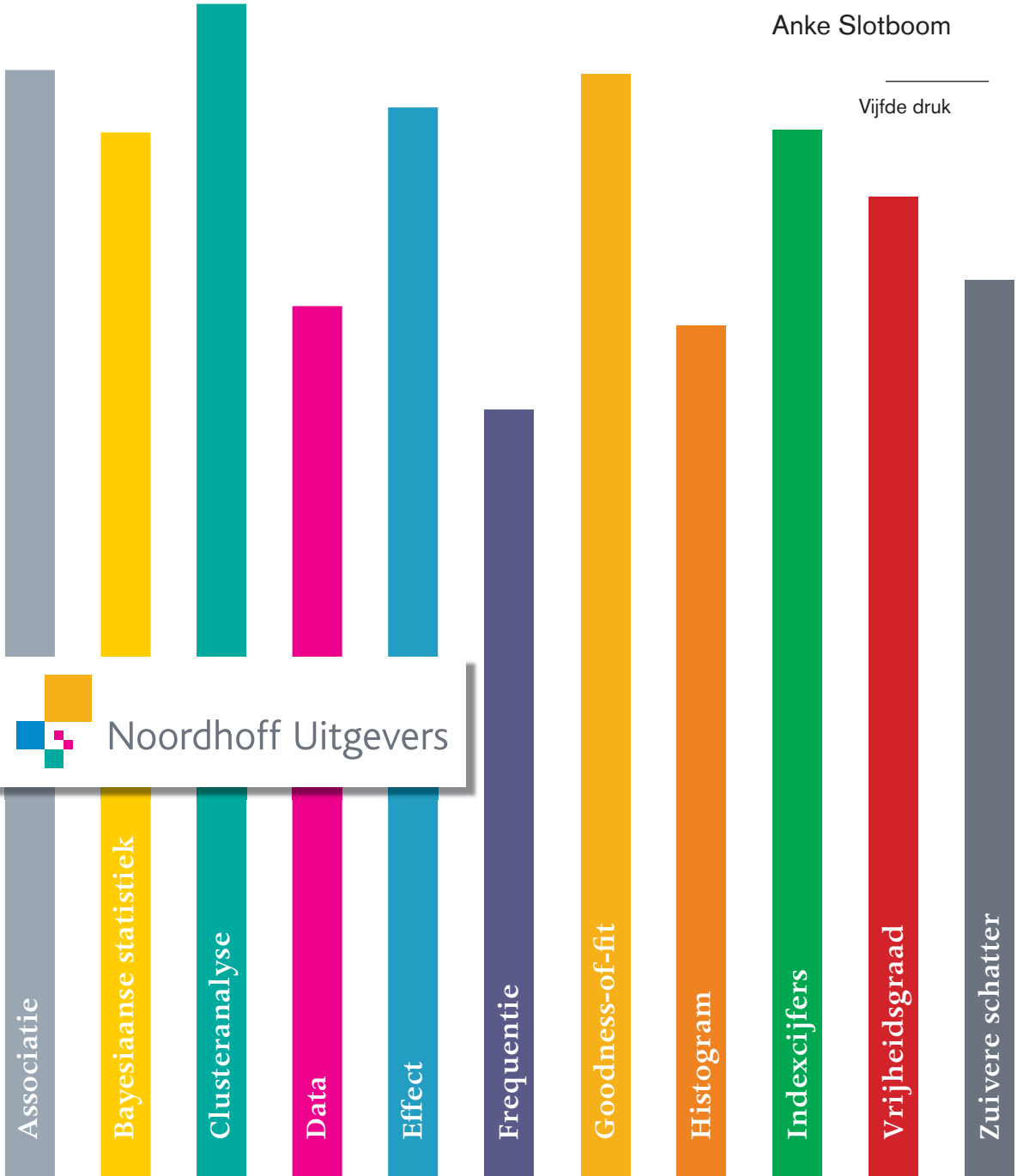
Statistiek in woorden

De meest voorkomende statistische begrippen van A tot Z

Anke Slotboom

Vijfde druk



Noordhoff Uitgevers

Statistiek in woorden

Voor al degenen die statistiek een moeilijk vak vinden

Statistiek in woorden

De meest voorkomende begrippen van A tot Z

Anke Slotboom

Vijfde druk

Noordhoff Uitgevers Groningen

Ontwerp omslag en omslagillustratie: Rocket Industries, Groningen

Eventuele op- en aanmerkingen over deze of andere uitgaven kunt u richten aan:
Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen,
e-mail: info@noordhoff.nl

Deze uitgave is gedrukt op FSC-papier.

0 / 13

ISBN (ebook) 978 90 01 84789 0
ISBN 978 90 01 81858 6
NUR 916

Woord vooraf bij de vijfde druk

De bedoeling van dit boek is, om *zonder formules,* te beschrijven wat diverse statistische begrippen en technieken inhouden.
De veelheid aan formules en afleidingen die je in een gewoon statistiekboek aantreft, maakt het in de regel moeilijk om snel te 'zien' wat het kenmerkende van een bepaalde techniek is, of wat een bepaald begrip 'betekent'. Dit boek probeert in die lacune te voorzien. Op een beschrijvende manier wordt aan de hand van eenvoudige (getallen)voorbeelden en figuren uitgelegd wat het idee is achter een bepaalde techniek. Technische details en rekenregels zijn daarbij bewust vermeden.
Natuurlijk kun je zonder formules niet leren een onderzoek te verrichten of statistische analyses te plegen. Daarvoor is dit boek ook niet bedoeld. Het is uitdrukkelijk bedoeld voor gebruik naast een gewoon statistiekboek; dus niet ter vervanging, maar als *aanvulling* op een systematische inleiding in de statistiek.

Voor wie is dit boek bestemd?

De beginnende student die het duizelt door alle verwarrende (vaak Engelse) statistische termen, kan in dit boek snel opzoeken wat bijvoorbeeld het verschil is tussen een percentielrang en een percentielscore. Weet hij op een tentamen ineens niet meer wat een standaardfout is, dan vindt hij in dit boek in woorden uitgelegd wat het begrip inhoudt. Voor de precieze berekening zal hij zijn 'gewone' statistiekboek moeten raadplegen.
Een gevorderde student die voor de verwerking van zijn scriptiegegevens twijfelt tussen een clusteranalyse en een discriminantanalyse, vindt in dit boek wat het kenmerkende van die technieken is.
Een geïnteresseerde leek of een afgestudeerde die in een artikel een variantieanalyse tegenkomt, kan in eenvoudige bewoordingen nalezen wat een variantieanalyse nu ook weer 'doet'.

Bij het schrijven van dit boek heb ik twee soorten gebruikers voor ogen gehad:

1 Ten eerste studenten (aan universiteiten of instellingen voor hbo of mbo) die een cursus statistiek moeten volgen en daar tegenop zien. Zij hebben vaak behoefte aan een begripsmatige, niet-technische beschrijving van de diverse onderwerpen. Voor hen heb ik geprobeerd al die moeilijke en ingewikkelde begrippen in helder en begrijpelijk Nederlands uit te leggen. Gebruikersvriendelijk heb ik dat genoemd … .
2 Ten tweede degenen die ooit in een grijs verleden een cursus statistiek gevolgd hebben, maar bij wie de parate kennis wat is weggezakt. Vaak krijgen zij toch nog met statistiek te maken, hetzij om zelf gegevens te analyseren, hetzij om de artikelen die anderen geschreven hebben te kunnen begrijpen. Zij kunnen dit boek gebruiken als naslagwerk waarin zij in niet-technische termen beschreven vinden wat het wezen is van een bepaalde analysetechniek. Ook de meest gebruikte multivariate technieken en verdelingsvrije methoden worden in dit boek behandeld.

Geprobeerd is duidelijk te maken wat het *idee* is *achter* de techniek; hoe de concrete berekeningen verlopen (met welke formules, welke rekenmethodes, welke computerprogramma's) is bewust achterwege gelaten.

Opzet van het boek

Enerzijds is dit boek uitgebreider, anderzijds minder uitgebreid dan de meeste inleidende statistiekboeken. Minder uitgebreid, omdat er geen formules, afleidingen of bewijzen in staan. Ook de basisbegrippen uit de kansberekening zijn niet behandeld, evenmin als begrippen uit de proefopzettenleer of de testtheorie. Dit betekent dat er geen aandacht wordt besteed aan (bijvoorbeeld) operationaliseren, verschillende manieren van steekproeftrekking of Cronbach's α.
Uitgebreider, omdat ook (niet-lineaire) multivariate analysetechnieken beschreven zijn, die in inleidende statistiekboeken zelden aan bod komen.

Duidelijkheid en begrijpelijkheid hebben bij het schrijven van dit boek steeds voorop gestaan. In een statistiekboek zonder formules zal de 'mathematische werkelijkheid' noodzakelijkerwijs soms wat vereenvoudigd worden. Bij gebruik naast een 'gewoon' statistiekboek hoeft dit mijns inziens geen bezwaar te vormen; integendeel zelfs.

Graag wil ik prof. dr. L. J. Th. van der Kamp bedanken voor zijn aanmoediging en de waardevolle suggesties die hij gedaan heeft. Ook alle anderen die door hun steun en belangstelling hebben bijgedragen tot het ontstaan van dit boek: bedankt!
Opmerkingen van gebruikers en suggesties ter aanvulling of verbetering van dit boek zie ik graag tegemoet.

Leiden, juni 2012
Anke Slotboom

Studiewijzer

Om het boek als naslagwerk toegankelijk te maken, is gekozen voor een *alfabetische* opzet. In ongeveer 120 rubrieken wordt een groot aantal kernbegrippen uit de statistiek behandeld. De rubrieken zijn los van elkaar te lezen; het is niet de bedoeling dat je bij de A begint en met Z eindigt, maar je kunt kriskras door het boek heengaan om precies dat onderwerp na te lezen waarover je iets wilt weten. Bij praktisch geen enkel onderwerp is uitgegaan van voorkennis. Enige herhaling is daardoor onvermijdelijk. Om herkenning te bevorderen is bovendien veel gewerkt met dezelfde voorbeelden. De getallenvoorbeelden zijn in een aantal gevallen bewust vereenvoudigd om duidelijk te maken wat er aan de hand is. Omdat er geen formules of rekenregels gegeven worden, zul je in dit boek geen oefeningen aantreffen.

Voor de alfabetische indeling van onderwerpen heb ik voortdurend keuzes moeten maken in de trant van: zal ik 'standaardafwijking' in een aparte rubriek behandelen, of onderbrengen bij de rubriek over variantie? Opsplitsen in allemaal aparte rubrieken vereenvoudigt weliswaar het opzoeken, maar brengt voortdurende herhalingen met zich mee. Praktische overwegingen (hoe lang wordt een rubriek? krijg ik niet te veel herhalingen?) hebben voor een groot deel de keuzes bepaald. Sommige rubrieken behandelen één begrip of analysetechniek (bijvoorbeeld de rubriek over gemiddelde of die over variantieanalyse). Soms wordt in één rubriek een aantal logisch samenhangende begrippen behandeld (in de rubriek over percentielen komen bijvoorbeeld ook fractielen en decielen aan de orde). Een aantal rubrieken, ten slotte, biedt een globaal overzicht van onderwerpen die elders in het boek uitgebreider beschreven staan (bijvoorbeeld de rubriek correlatiematen – een overzicht; of de rubriek over multivariate analyse, waarin alle multivariate technieken op een rijtje gezet worden). Deze overzichtsrubrieken zijn op turquoise bladzijden afgedrukt.

De snelste manier om een bepaalde term op te zoeken, is via de *index* van *trefwoorden* achter in het boek. In deze index zijn de aparte rubrieken vet gedrukt. Omdat sommige begrippen samen in één rubriek behandeld worden, is het aantal trefwoorden in de index veel groter dan het aantal alfabetische rubrieken. Wordt een trefwoord uit de index pas halverwege een bepaalde rubriek genoemd, dan wordt dat woord in **rood** aangegeven.
Bij een alfabetische indeling is de samenhang tussen diverse begrippen niet altijd direct duidelijk. Zo veel mogelijk is ieder begrip in een breder kader geplaatst ('clusteranalyse is een **multivariate techniek**') en worden verbanden gelegd met andere begrippen of technieken ('nauw verwant met het histogram is het **kolommendiagram**').
Begrippen in **turquoise**, worden ergens anders in het boek uitgebreider behandeld. Begrijp je zo'n term dus niet direct, wanhoop dan niet: op een andere plaats kun je er méér over lezen.

Veel statistiekboeken en (wetenschappelijke) artikelen zijn in het Engels geschreven. Bijlage 2 geeft een Engels-Nederlandse woordenlijst van de meest gangbare statistische begrippen. In bijlage 3 vind je een schematisch overzicht van de belangrijkste (toetsings)technieken.

Inhoud

Z

A posteriori vergelijkingen

Met a posteriori vergelijkingen kun je, na een significant resultaat bij een variantieanalyse, bekijken welke groepsgemiddelden significant van elkaar verschillen.

Bij een **variantieanalyse** wordt een aantal groepsgemiddelden tegelijkertijd met elkaar vergeleken. Een **significant** resultaat geeft alleen aan dát er verschillen tussen de groepsgemiddelden bestaan. Vanwege de aard van de toetsprocedure kun je echter niet zeggen wélke verschillen significant zijn. Een variantieanalyse is dus een zeer *globale* procedure om na te gaan óf er iets aan de hand is, maar geeft niet aan wát er precies aan de hand is.

Voorbeeld

Stel dat je vier groepen (I tot en met IV) hebt, met respectievelijk de gemiddelden 20, 24, 30 en 32, en dat een variantieanalyse uitwijst dat deze gemiddelden ergens significant van elkaar verschillen. Het is nu niet duidelijk wáár de significante verschillen zitten. Ongetwijfeld zal het grootste verschil, dat tussen groep I (20) en groep IV (32), significant zijn. Maar verschillen bijvoorbeeld groep I (20) en groep III (30) ook significant van elkaar?

Als je bij een variantieanalyse een significant resultaat vindt, zou je daarna met behulp van afzonderlijke **t-toetsen** alle gemiddelden paarsgewijs met elkaar kunnen vergelijken. Dat levert echter een aantal problemen op. Ten eerste is het qua rekenwerk nogal omslachtig, omdat het aantal toetsen snel toeneemt met het aantal groepen in de variantieanalyse. Bij vier groepen zijn er zes mogelijke paarsgewijze vergelijkingen, bij tien groepen zijn dat er al 45! Belangrijker echter is het verschijnsel dat men **kanskapitalisatie** noemt: wanneer een reeks t-toetsen wordt uitgevoerd, neemt de kans op **fouten van de eerste soort** snel toe. Weliswaar blijft de kans op een type-I fout voor elke toets afzonderlijk gelijk aan het gekozen **significantieniveau** α (meestal 5%), maar de kans op ten minste één type-I fout *voor alle toetsen gezamenlijk* is recht evenredig met het aantal groepen. De **Bonferroni-ongelijkheid** stelt dat de kans dat één van die toetsen tot een type-I fout zal leiden, gelijk is aan $J \times \alpha$, waarbij J het aantal verschillende groepen in de variantieanalyse is.

Bij a posteriori vergelijkingen, ook wel **meervoudige (paarsgewijze) vergelijkingen** genoemd, worden extra waarborgen ingebouwd om kanskapitalisatie tegen te gaan. Deze waarborgen komen erop neer dat een verschil groter moet zijn dan in een gewone t-toets om significant te zijn. Volgens de Bonferroni-methode bijvoorbeeld, wordt het significantieniveau voor elke toets afzonderlijk niet op α, maar op α/J gesteld. Er bestaan verschillende procedures voor meervoudige vergelijkingen, die alle gebaseerd zijn op het idee van t-toetsen voor verschillen tussen gemiddelden. De procedures verschillen van elkaar in de hoogte van de extra drempel die ze inbouwen

voor significantie. Een aantal bekende procedures zijn de **Newman-Keuls-toets**, de **Duncan multiple range-toets**, de **Fisher's LSD-toets** en de **Scheffétoets**. De meest gebruikte toets is de **HSD-toets** van Tukey. HSD staat hierbij voor *h*onestly *s*ignificant *d*ifference.

De procedure is simpel: uitgaande van het gewenste significantieniveau, de spreiding van de scores binnen de groepen, het aantal groepen en de groepsgrootte, wordt een getal HSD berekend. Significant is ieder verschil tussen de groepsgemiddelden dat groter is dan HSD.

Voorbeeld

Men kan de verschillen tussen alle groepsgemiddelden overzichtelijk weergeven in een tabel van de volgende vorm:

		Groep I 20	**Groep II** 24	**Groep III** 30	**Groep IV** 32
Groep I	gemiddelde 20	...	4	10	12
Groep II	24		...	6	8
Groep III	30			...	2
Groep IV	32				...

Stel dat HSD in dit voorbeeld gelijk is aan 5,88. We zien dan dat het verschil tussen groep I en groep III (10 punten) significant is, want 10 punten is meer dan 5,88. Eveneens significant zijn de verschillen tussen groep I en groep IV (12 punten), tussen groep II en groep III en tussen groep II en IV. De groepen III en IV haalden dus hogere scores dan de groepen I en II. Niet significant zijn echter de onderlinge verschillen tussen III en IV, en evenmin die tussen I en II.

A posteriori vergelijkingen worden ook wel **post-hoc vergelijkingen** genoemd (post-hoc = a posteriori = achteraf, dat wil zeggen *na* een variantieanalyse).
Als je *van tevoren* al bepaalde veronderstellingen of hypothesen hebt over de verschillen tussen de groepen, dan kun je *in plaats van* een variantieanalyse gerichtere vragen stellen door middel van **a priori vergelijkingen.**

A priori vergelijkingen

Met a priori vergelijkingen kun je gerichte vragen stellen over de verschillen tussen groepen. De techniek wordt gebruikt in plaats van een variantieanalyse.

Bij een **variantieanalyse** wordt een aantal groepsgemiddelden tegelijkertijd met elkaar vergeleken. Een **significant resultaat** geeft alleen aan dát er verschillen tussen de groepsgemiddelden bestaan, maar de toetsprocedure geeft niet aan welke verschillen significant zijn. Een variantieanalyse is dus een zeer *globale* procedure om na te gaan 'of er iets aan de hand is'.
Soms heb je bij voorbaat (a priori betekent letterlijk 'van tevoren') al *gerichte* vragen of veronderstellingen over de verschillen tussen de groepen. Een variantieanalyse geeft daarop niet direct een antwoord. Het is dan handiger om géén variantieanalyse te doen, maar een toetsprocedure te gebruiken die deze gerichte vragen beantwoordt.

Voorbeeld

Stel dat je de effectiviteit van verschillende methoden van meningsbeïnvloeding wilt bestuderen. Je laat proefpersonen een vragenlijst invullen over een controversieel onderwerp. Vervolgens vorm je vier groepen:

- Groep I krijgt een positieve lezing te horen van een deskundige op het betrokken terrein.
- Groep II krijgt een positieve film over het onderwerp te zien.
- Groep III krijgt zowel de film als de lezing.
- Groep IV, de controlegroep, krijgt geen 'behandeling' en vult alleen de vragenlijsten in.

Na afloop moet iedereen de vragenlijst opnieuw invullen. Het verschil tussen de eerste en de tweede invulling wordt als index genomen voor de mate van meningsverandering.
Het ligt voor de hand dat je als onderzoeker hier een aantal gerichte vragen hebt, bijvoorbeeld:

1 Verschilt de controlegroep van de groepen die wél een behandeling kregen?
2 Is het effect van een film anders dan het effect van een lezing?
3 Heeft een film *plus* lezing méér effect dan een film apart of een lezing apart?

Je wilt dus verschillende (combinaties van) groepsgemiddelden met elkaar vergelijken, met name bij:

- vraag 1: groep IV tegenover de groepen I, II en III;
- vraag 2: groep I tegenover groep II;
- vraag 3: groep III tegenover groep I en groep II.

Om de eerste vraag te beantwoorden, gooi je als het ware de scores van de groepen I, II en III op één hoop en vergelijkt het gemiddelde daarvan met de gemiddelde

score van groep IV. Voor vraag 3 vergelijk je groep III met het gemiddelde van de groepen I en II samen.

Een dergelijke combinatie van verschillende groepsgemiddelden tegenover elkaar heet een *vergelijking* (Engels: comparison) of een **contrast**.
Per vraag wordt één vergelijking gevormd. De drie vergelijkingen worden met drie afzonderlijke **t-toetsen** getoetst.
Als er veel verschillende t-toetsen op dezelfde gegevens worden uitgevoerd, neemt de kans op **fouten van de eerste soort** snel toe; een verschijnsel dat bekend staat als **kanskapitalisatie**. Om het gevaar van kanskapitalisatie tegen te gaan, moet de gevolgde procedure aan een aantal eisen voldoen.
In ieder geval moeten de vragen zó geformuleerd worden dat ze elkaar niet overlappen, dat wil zeggen dat ze geen dubbele informatie opleveren. Zou je groep I met groep II vergelijken, groep I met groep III en groep II met groep III, dan is één van die vragen overbodig: uit de verschillen I-II en I-III is het verschil II-III namelijk af te leiden. Niet-overlappende vragen worden onafhankelijk of **orthogonaal** genoemd. Bij J groepen zijn er maximaal J-1 onafhankelijke vergelijkingen mogelijk. Er bestaan procedures voor het opstellen van stelsels orthogonale vergelijkingen.
Daarnaast wordt soms het **significantieniveau** van de afzonderlijke t-toetsen aangepast. De **Dunn multiple comparison** toetsprocedure voor a priori vergelijkingen, ook wel de **Bonferroni-toets** genoemd, gebruikt een aangepast significantieniveau.

Associatie

Samenhang tussen twee kwalitatieve variabelen duidt men aan met de term associatie. Twee variabelen hangen samen – anders gezegd: zijn geassocieerd – als het kennen van de waarde van de ene variabele informatie verschaft over de andere variabele.

Voorbeeld

Stel dat een pedagoog wil weten of de leerprestaties in de eerste klas van de middelbare school samenhangen met het type basisschool dat bezocht is. Hij selecteert daartoe 45 kinderen van elk van drie types basisscholen: confessioneel, openbaar en 'neutraal bijzonder' (Vrije Scholen, montessorischolen en dergelijke). Het gemiddelde rapportcijfer van het paasrapport in de eerste klas van de middelbare school dient als aanduiding van de leerprestaties. Stel dat de verdeling van de leerprestaties van de groepen er als volgt uitziet:

		Type basisschool		
		confessioneel	openbaar	neutraal bijzonder
Leerprestaties	goed	30	30	30
	middelmatig	10	10	10
	slecht	5	5	5
		45	45	45

Van de 45 leerlingen van confessionele scholen behaalden er 30 'goede' leerprestaties, tien middelmatige en vijf slechte. Merk op dat de verdeling van leerprestaties hier bij alle groepen gelijk is.

Als je op basis van deze gegevens gevraagd wordt te voorspellen hoe de leerprestatie van een willekeurig kind van een confessionele basisschool is, kun je het beste gokken: goed. Dit geldt echter evenzeer voor een leerling van een openbare school of van een Vrije School. Er bestaat hier *geen verband* tussen schooltype en leerprestaties: als je weet welk type basisschool een kind bezocht heeft, zegt dat nog niets over zijn leerprestaties op de middelbare school. Voor ieder schooltype is het het meest waarschijnlijk dat de leerprestaties goed zijn. Men zegt dan dat leerprestaties **onafhankelijk** zijn van schooltype. Er is geen associatie tussen beide variabelen, ofwel de associatie is nul.

Van onafhankelijkheid is altijd sprake als de verdeling binnen iedere kolom [men noemt dit de **voorwaardelijke (frequentie)verdeling** van leerprestaties, gegeven een bepaald schooltype] voor alle kolommen gelijk is.

Voorbeeld

Vergelijk bovenstaande hypothetische gegevens met de volgende (eveneens verzonnen) gegevens:

		Type basisschool		
		confessioneel	openbaar	neutraal bijzonder
Leerprestaties	goed	30	10	5
	middelmatig	10	30	10
	slecht	5	5	30
		45	45	45

Hier bestaat wel een verband of associatie tussen schooltype en leerprestaties. Voor kinderen die een confessionele basisschool bezocht hebben is de voorspelling van de leerprestaties anders dan voor kinderen die een openbare of een neutraal-bijzondere basisschool bezochten. Als je weet welk schooltype een kind bezocht heeft, verschaft dit informatie over de te verwachten leerprestaties op de middelbare school. De associatie is echter niet perfect: voor een kind van een openbare basisschool kun je het beste 'middelmatige' leerprestaties voorspellen; bij 15 van de 45 kinderen gaat die voorspelling echter niet op.

In een **kruistabel** zoals hierboven (de rijen corresponderen met de leerprestaties en de kolommen met het schooltype), is er sprake van een verband tussen beide variabelen als de frequentieverdelingen per rij en per kolom van elkaar verschillen.

Een dergelijk resultaat kan natuurlijk bij toeval verkregen zijn, bijvoorbeeld omdat de steekproeven niet representatief waren. Men kan **toetsen** of er 'echt' een verband bestaat tussen beide variabelen. Hierbij wordt berekend hoeveel kinderen men in ieder vakje van de tabel zou verwachten als er geen verband tussen beide variabelen zou bestaan. Men berekent dus de – bij onafhankelijkheid – *verwachte* aantallen. Vervolgens wordt beoordeeld in hoeverre de in het onderzoek *gevonden* (geobserveerde) aantallen hiervan afwijken. Hoe groter de afwijkingen, des te meer aanwijzing vormt dit dat beide variabelen *niet* onafhankelijk zijn. Bij deze **χ^2(chi-kwadraat)-toets voor associatie** wordt gebruik gemaakt van de **χ^2-verdeling** om de gevonden waarde op **significantie** te beoordelen.

We hebben hierboven gezien dat de associatie tussen leerprestatie en schooltype niet perfect was: informatie over het bezochte schooltype verschaft wel enige informatie over de te verwachten leerprestaties, maar de voorspelling gaat niet altijd op.

Voorbeeld

A

Een perfect verband – en daarmee een 100% juiste voorspelling – zou er bestaan als de gegevens er bijvoorbeeld als volgt uit zouden zien:

		Type basisschool		
		confessioneel	openbaar	neutraal bijzonder
Leerprestaties	goed	45	0	0
	middelmatig	0	0	45
	slecht	0	45	0
		45	45	45

Hier is sprake van een perfecte associatie. De associatie is 1 (of 100%).

Er bestaan verschillende indexen die aangeven hoe *sterk* het verband tussen twee kwalitatieve variabelen is. De meeste van deze **associatiematen** kunnen variëren tussen 0 (geen verband) en 1 (perfect verband). Waarden tussen 0 en 1 duiden op enige samenhang. Hoe hoger de waarde, des te *sterker* is de samenhang en des te beter is de voorspelbaarheid.
Een bekende associatiemaat is φ′, de **phi-coëfficiënt van Cramér**. Een andere veel gebruikte index is de **contingentiecoëfficiënt C**, die echter als nadeel heeft dat de maximale waarde ervan niet 1 is. De waarde van C is afhankelijk van de grootte (= aantal rijen en kolommen) van de kruistabel. Dat maakt C lastig interpreteerbaar.