**Grootschalig veldonderzoek: Inleiding in de data – analyse***Jan Lammers*

**Hoofdstuk 1: Inleiding**

In een grootschalig veldonderzoek worden wordt een groot aantal, naar toeval geselecteerde personen geïnterviewd of geënquêteerd. De vragen die hen worden voorgelegd, maar ook de antwoord mogelijkheden op elke vraag zijn in principe voor alle personen gelijk zodat het mogelijk is de antwoorden met elkaar te vergelijken en op statistische wijze te verwerken.

*Frequentie verdelingen*
De technieken voor statistische analyse van dit soort gegevens is schier onbeperkt. Ze hebben echter 1 ding gemeen: ze zijn gemaakt voor het analyseren van *frequentie verdelingen.* Hierin staat vermeld hoeveel personen de verschillende antwoord mogelijkheden van een bepaalde vraag of van meer vragen tegelijkertijd hebben gekozen. De gegevens van personen (respondenten) staan in een frequentie verdeling gegroepeerd bij elkaar.

*Hypothesen*
Om onze kennis te kunnen verbreden kunnen we trachten een antwoord te vinden op een door ons gestelde vraag of kijken we of een bepaalde verwachting die we er over hadden bevestigd wordt. De combinatie van het uitspreken van verwachtingen en aan de hand van gegevens controleren of de verwachtingen uitkomen zorgt voor uitbreiding van onze kennis. De verwachtingen noemt men *Hypothesen.* Er zijn verschillende manieren om tot een hypothese te komen. Dit kan van te voren al bedacht worden, maar veel vaker worden deze opgesteld tijdens het onderzoek.

*Van hypothesen naar frequentieverdelingen*De gegevens zijn bijeengebracht in een frequentie verdeling en er zijn hypothesen met betrekking tot die gegevens. Nu moet er nog een beschrijving komen van het proces dat deze twee aan elkaar koppelt. Elk grootschalig onderzoek zou je kunnen zien als een keten van onderzoekshandelingen die uiteindelijk doel heeft de gegevens te binden aan de hypothesen. Belangrijke fasen in deze keten zijn: de steekproeftrekking, de dataverzameling en de schaalconstructie.
- steekproeftrekking: welk type personen worden er benaderd? En op welke wijze maak je een selectie uit deze personen?
- Dataverzameling: Er worden concrete vragen opgesteld, wordt bepaald hoe je de vragen voorlegt, en de antwoorden worden verzameld en gecontroleerd op inconsistenties.
- Schaal constructie: Er worden schalen opgesteld om de complexe eigenschappen van de personen te kunnen meten. Vanuit de antwoorden dient de onderzoeker te komen tot 1 score voor elke persoon die aangeeft in welke mate hij die complexe eigenschap bezit.

In dit boekje wordt echter de aandacht het meeste gevestigd op de fase die na voorgenoemde fasen komt: De *analysefase* en hoe deze fase vervlochten is met het gehele onderzoeksproces. Het boekje onderscheid dan ook vier relevantie stappen om van hypothesen naar gegevens te komen: 1. Operationaliseren 2. Schematiseren 3. Prediceren 4. Controleren.
In de eerste drie stappen vindt het naar data toe concretiseren van de hypothese plaats, in de laatste stap vindt de analyse in strikte zin plaats.

De statistische analyse techniek die dit boekje gebruikt heet: de *elaboratietechniek.* Het is geen echte techniek doch belangrijk omdat: het een algemeen kader aanreikt voor frequentie verdelingen (in de vorm van kruistabellen waarin de verdelingen staan) de interpreteren.

**Hoofdstuk 2: Verdelingen en kruistabellen**

In de sociale wetenschappen zegt men tabellen en bedoelt daarmee de kruistabellen met frequenties. In dit boekje wordt dus ook de term tabellen gebruikt in plaats van kruistabellen. In een tabel word een verdeling, maar doorgaans meerdere verdelingen gepresenteerd.

*Verdelingen 2.1*Voor een vb van een verdeling zie blz. 16. De term verdeling slaat op het resultaat van bijvoorbeeld een boer die zaait op het land. De verdeling is dan het resultaat van de spreiding van de zaadjes op het land.

*Trivariate verdeling:* Er is sprake van drie dimensies bij een verdeling (zoals bij de zaadjes, zijn verdeeld in lengte, breedte en hoogte) Men kan ook spreken van die variabelen: Lengte breedte en hoogte en spreekt daarom ook wel van een Trivariate verdeling. Dit is dus een verdeling van onderzoekseenheden waarbij de positie van elke eenheid in de verdeling door drie variabelen bepaald is.

Codeboek: Een lijst van meerdere pagina’s waarin de vragen en antwoorden respectievelijk zijn omgezet in variabelen en codes.
Variabele: Is een eigenschap van een individu die in het onderzoek wordt gemeten.
Code: Een variabele wordt weer onder gescheiden in categorieën en aan elke categorie wordt een code toegekend. Een code wordt meestal een *score* genoemd, wanneer deze betrekking heeft op een individu.
Data: De scores van alle respondenten tezamen vormen de data.
Datamatrix: Hierin staat de data bijeen.
Zie voor vb codeboek blz 18.

De datamatrix bevat de gegevens die uiteindelijk geanalyseerd worden en waarop de conclusies worden gebaseerd. Het codeboek is belangrijk omdat het de sleutel beidt tot het lezen van de data; het legt de verbinding tussen een score en de inhoud van het antwoord op een vraag.

Een voorbeeld van een trivariate verdeling:
1. Geslacht met code 0= jongen en 1= meisje
2. Urbanisatiegraad met code 0= stad en 1=platteland
3. Vervolgonderwijs met code 0= geen vwo en 1= vwo

In de datamatrix staan de scores van de kinderen op deze variabelen. Zie blz. 19.

Het raamwerk van score combinaties ziet er als volgt uit: blz. 20.

Dit raamwerk bestaat uit 8 vakjes die *cellen* worden genoemd. Elke cel staat voor een mogelijke combinatie van de scores.

Wanneer je van elke combinatie de streepjes zou optellen krijg je trivariate frequentieverdeling van geslacht, urbanisatiegraad en vervolgonderwijs. Zie tabel blz. 21.

Met een verdeling zijn dus altijd een eenheid, een of meer eigenschappen en de verzameling van alle eenheden weergegeven. De verzameling wordt een populatie genoemd.

Het inzicht dat met een verdeling altijd een eenheid, een of meer eigenschappen van de eenheid en een populatie geassocieerd zijn, is nodig omdat hypothesen altijd uitspraken zijn over eigenschappen van bepaalde eenheden en een bepaald bereik hebben. Een hypothese beweert in wezen iets over de verdeling.

Bivariate verdeling: De plaats van de eenheid wordt bepaald door twee variabelen. Dus praktisch hetzelfde verhaal als bovengenoemde, maar dan bijvoorbeeld met alleen de variabelen geslacht en vervolgonderwijs.
Univariate verdeling: verdelingen waarbij de eenheden maar volgens een variabele verdeelt zijn: bijvoorbeeld geslacht.

De verschillende bovengenoemde verdelingen staan niet los van elkaar.
Vanuit een multivariate (meer dan een trivariate verdeling) verdeling zijn steeds weer ‘lagere’ verdelingen af te leiden.

*Hoe komt een verdeling tot stand? 2.2*We kunnen aan drie typen factoren denken die achtereenvolgens met de eenheid van verdeling, de onderzoeker en het toeval te maken hebben.

1. Factoren die met de onderzoeker te maken hebben, zijn er velen. De gemeenschappelijke naam waarmee men ze aanduidt is: methodefactoren.
Zij kunnen een niet bedoelt effect op de verdeling hebben als zij bijvoorbeeld een code verkeert schrijven (verschrijvingen)

2. De factoren die men toevalsfactoren noemt: Toevallige zaken die zich voordoen en ten gevolge hebben dat er verkeerde frequenties in de verdeling ontstaan. Toevalsfouten kunnen van zowel de onderzoeker als de respondent komen. Over het algemeen zijn toevalsfouten niet heel erg.

3. Er wordt meer aandacht besteedt aan ‘echte’ factoren die met de onderzoekseenheid te maken hebben. Dit noemt men ook wel reële factoren: Ze liggen in de realiteit die we willen bestuderen. Zoals de achterliggende gedachte dat kinderen minder naar het vwo gaan als ze van het platteland komen en als ze een meisje zijn.

De verdeling wordt hierdoor vooral gezien als een product van sociale processen die zich in de werkelijkheid voorgetrokken hebben.

Bij frequente verdelingen wil men meestal niet de absolute, maar de relatieve grootte van de frequenties met elkaar vergelijken. Dit wordt dan gedaan aan de hand van de berekende proporties of percentages. Men spreekt dan ook meestal van een procentuele of percentage verdeling.

*Tabellen 2.3*

Zie hiervoor tabel 1 op blz 27.

De kern van deze tabel is te splitsen in twee partiële verdelingen (met andere woorden te partitionaliseren). Omdat in tabel 1 de trivariate verdeling (geslacht, urbanisatiegraad en vervolgonderwijs) de belangrijkste is wordt deze tabel wel een drie dimensionale tabel genoemd (ook wel een drie weg tabel).

Rond de grijze vlakken (om de kern heen van de tabel), in de marges van de partiele tabellen staan twee bivariate verdelingen: geslacht en vervolg onderwijs en geslacht en urbanisatiegraad. Omdat deze verdelingen in de marges staan worden ze marginale verdelingen oftewel randverdelingen genoemd.

Nadat absolute tabellen als tabel 1 zijn berekend worden de percentages berekend. Dus va absolute gegevens naar relatieve gegevens. Dit kan nog al eens een verschillende uitkomst bieden in de vergelijking. Bij sociaal wetenschappelijk onderzoek is de relatieve uitkomst van meer belang.
Er zit wel een nadeel aan verbonden. Wanneer je kijkt naar de relatieve tabellen op blz 29 zie je dat de percentages over de tabellen heen lastig te vergelijken zijn.

**Hoofdstuk 3: Hypothesen**

Omdat de gegevens van het onderzoek op ervaring van buitenaf berusten, worden ze empirische verdelingen genoemd.

*De hypothese verdeling 3.1*Algemeen gezien is een hypothese een veronderstelling ten aanzien van de realiteit.

Wanneer men een verdeling maakt op basis van een verwachting noemen we dat een hypothetische of verwachte of theoretische verdeling.
En verder is en een populatie, die als het om een hypothese gaat meestal bereik of geldigheidsbereik of domein of omvang wordt genoemd.
Het is de bedoeling om na te gaan of de hypothese empirisch geldig is en door er een verdeling bij te denken met een eenheid, eigenschappen met bepaalde categorieen en een bereik weet men voor zichzelf al welke empirische verdeling men nodig zal hebben om die controle uit te voeren.

*Kenmerken van een hypothese 3.2*

Een hypothese kan gekarakteriseerd worden naar de mate van precisie, determinatie, eenvoud en toetsbaarheid.

Precisie:
Een hypothese wordt precies genoemd naar de mate waarin de erin voorkomende eenheid, eigenschappen en bereik precies omschreven zijn.

Determinatie:
Men spreekt van deterministische hypothesen wanneer het voor elke eenheid zeker is in welk vakje van de verdeling het geplaatst moet worden.
De meeste hypothesen zijn echter niet zo exact. In zo een geval spreekt men van probalistische hypothesen.

Eenvoud:
De eenvoud van een hypothese wordt vooral bepaald door het aantal eigenschappen dat erin voorkomt. De hypothesen zijn gericht op de verdeling van 1 eigenschap die in de gevallen besproken in het boekje ook maar twee categorieen heeft. Hoe meer eigenschappen er in een hypothesen opgenomen worden, hoe complexer.
Het doel is echter van de wetenschap om met zo weinig mogelijk middelen inzicht te verschaffen in de complexe werkelijkheid.

Toetsbaarheid:
Dit is de mate waarin het mogelijk is om de hypothetische verdeling met een empirische te vergelijken. Bij het opstellen van een hypothese dient men daarom altijd rekening te houden met de beschikbaarheid van data.

*Relationele hypothesen 3.3*

Vaak wordt in een hypothese een verwachting uitgesproken ten aanzien van het verband tussen twee of meer eigenschappen. Het zijn de relationele hypothesen.

Een voorbeeld van een relationele hypothese die gericht is op een verband in een bivariate verdeling is het volgende:

De overgang naar het vwo is voor de kinderen uit de klas van Anny onafhankelijk van het feit of ze al dan niet hard gewerkt hebben.

Deze hypothese zegt iets over de aanwezigheid van een verband. Het verband hoeft niet perfect te zijn. Dit komt omdat men open wil laten dat er naast de in de hypothesen genoemde factoren nog andere factoren zijn die het resultaat kunnen beinvloeden (toevalsfactoren etc.).

Hoewel de hypothese vaak gericht is op de aanwezigheid van een verband, zijn er andere aspecten die in de hypothesen betrokken kunnen worden. We noemen sterkte ( er kan verondersteld worden dat er een zwak, matig of sterk verband is), richting ( de richting waartoe er een verband is) en vorm van een verband.
Vormaspecten treden op wanneer minstens een van de eigenschappen meer dan twee categorieen of waarden heeft. Wanneer een eigenschap twee waarden heeft spreekt men van dichotome eigenschappen.

Wanneer de codes van de categorieen samen overeenkomen als in:
1=hard gewerkt 1=vwo
0= Niet hard gewerkt 0= niet vwo

Men spreekt van een positieve samenhang wanneer de uitkomst 1 samengaat met 1. Dus hard werken en vwo.
Men spreekt over een negatieve samenhang wanneer de uitkomst 1 samengaat met uitkomst 0. Dus Hard gewerkt en niet vwo.

Lineair verband: Bijvoorbeeld dat de kans om naar het vwo te gaan toeneemt in gelijke mate bij meer studie ijver.
Monotoon stijgend is wanneer er wel sprake is van een stijging, maar niet met dezelfde mate.

*Beschrijvende en verklarende (causale) hypothesen*

Relationele hypothesen:
- Beschrijvende hypothesen: Wordt in vermeld of er een verband tussen de eigenschappen bestaat en vervolgens kan er een nadere aanduiding worden gegeven *hoe* het verband eruit ziet: in welke richting het verloopt, hoe sterk het is en welke vorm het heeft.
Verklarende (causale) hypothesen: Zij gaan verder dan het beschrijven en vertellen ook *waarom* er een verband bestaat. Causale hypothesen wijzen op de noodzaak van het samengaan doordat de ene eigenschap de *oorzaak* is van de andere.

Om het extra dat een causale hypothese boven een beschrijvende hypothese heeft naar voren te halen, kiest men andere formuleringen. Bij een causale hypothese gebruikt men termen als: oorzaak, gevolg, heeft invloed op, leidt tot, hangt af van, is een functie van, impliceert en omdat.
Bij het analyseren van gegevens zijn causale hypothesen zeer geliefd.

*Hoe komt een hypothese tot stand? 3.4*

*Invoeling*: is een belangrijke, onmisbare methode om tot een voorstelling te komen van wat er in de realiteit voor bepaalde personen gebeurt.
Door *literatuur* te raadplegen of door *gesprekken* te houden met deskundigen en personen uit het maatschappelijk veld, kan men het beeld dat men zelf heeft, aanvullen. Men kan ook verder gaan en een *voorstudie* houden.

Welke methode of combinatie van methoden men ook hanteert, er dient altijd op een aantal punten geconcentreerd te worden om een hypothese op te stellen.
-Letten op het causale proces waarover men meer empirische kennis wil verkrijgen.
- Jezelf de vraag stellen bij wie of wat het proces zich afspeelt. Op welke eenheid moet men zich concentreren?
-Welke eigenschappen spelen in dat proces een rol? In dit stadium houdt men zich bezig met de conceptuele afbakening van het proces.
- Hierna volgen de vragen die met de volgorde van de eigenschappen in het causale proces te maken hebben.
- Na deze schematische rangschikking komt de vraag aan de orde tussen welke eigenschappen er een direct oorzakelijk verband bestaat en waarom. Bovengenoemde punten hebben alles te maken met de diverse stadia van theorievorming.

Drie redenen waarom het niet nodig is om een voorstelling van het totale causale proces te hebben alvorens men een hypothese opstelt.
1. Hypothesen staan in functie van theorievorming. Bij het opstellen van een hypothese is het dus lang niet altijd mogelijk om een voorstelling te maken van het totale proces in de realiteit. Dit is ook niet nodig. Dus prematuur.
2. Een theorie weergeeft steeds een beperkte voorstelling van de realiteit weer en bevat ook steeds open vragen.
3. In een theorie of hypothese worden vaak alleen de eigenschappen opgenomen die een cruciale rol in het causale proces lijken te spelen.

Samenvattend:
Een hypothese heeft betrekking op een hypothetische verdeling met een eenheid, een of meerdere eigenschappen en een bereik.
Men kan een hypothese over het algemeen karakteriseren naar mate van precisie, eenvoud, determinatie en toetsbaarheid.
Het meeste komen de relationele hypothesen voor.
Beschrijvende relationele hypothese: Er wordt gesteld dat er een verband bestaat tussen eigenschappen. Soms wordt daarnaast ingegaan op andere aspecten als, richting, vorm en sterkte van een verband.
Causale relationele hypothese: Het oorzakelijke karakter van een verband wordt benadrukt. Deze zijn bij onderzoekers het meest geliefd omdat ze anticiperen op het proces in de realiteit op basis waarvan een verdeling tot stand is gekomen.

**Hoofdstuk 4: Analyseren in vier stappen**

We kunnen in het analyse proces vier fasen onderscheiden die we aanduiden met de termen: operationaliseren, schematiseren, prediceren en controleren.

1. Het verbaal formuleren van de operationele hypothese.
2. Het in schema brengen van de operationele hypothese, het tekenen van een pijldiagram.
3. Het opstellen van predicties
4. Het controleren van de predicties.

Deze 4 fasen liggen aan de basis van elke analyse. Ze faseren de toetsing van een hypothese aan de data en men kan ze daarom ook zien als stappen die gezet moeten worden om een analyse uit te voeren.

*Het formuleren van de operationele hypothese 4.1*

Een operationele hypothese: Een naar data omgezette hypothese.
Het operationaliseren van een hypothese zal de drie aspecten: eenheid, eigenschappen en bereik moeten bevatten.
Ecologische fout: Een fout waarbij ten onrechte voor een eenheid (bv een kind) een bepaalde conclusie wordt getrokken op basis van een ander type eenheid (bv een school)

Trekt men conclusies op basis van de steekproef verdeling, dan mogen die onder bepaalde voorwaarden *gegeneraliseerd* worden naar de populatie waaruit die steekproef is getrokken.

*Het maken van een pijldiagram 4.2*

Een niet noodzakelijke stap, maar in de praktijk aan te bevelen omdat met een pijldiagram doorgaans duidelijker dan in woorden kan worden aangegeven hoe de causale relatie tussen de variabelen verondersteld kan worden.

Om het pijldiagram te maken schrijven we eerst de labels van de variabelen op papier en trekken vervolgens pijlen tussen de labels. Dit overeenkomstig met de veronderstelde causale richting.

BV: Hypothese: Een kind gaat eerder naar het vwo door hard te werken.
W= hard werken of niet hard werken en V= vwo of geen vwo
W🡪V

We moeten dit pijldiagram lezen als de weergave van een causaal proces, als iet dynamisch dus. Om de verandering beter uit te laten komen kun je de pijldiagram beter als een mechaniek voorstellen. (zie plaatje blz 51)

Een dergelijk causaal mechaniek is niets anders dan een speelgoedje waarmee we een causaal proces in de realiteit nabootsen. En door ermee te ‘spelen’ beter het proces in de realiteit te kunnen begrijpen.

*Het opstellen van predicties 4.3*

In deze fase worden de consequenties geformuleerd die de hypothese heeft voor de verdeling voor de variabelen die erin worden genoemd.
Predicties: De vanuit de hypothese voorspelde analyse resultaten.
In plaats van predicties spreekt men ook wel van restricties.

*Het controleren van de predicties 4.4*

Als laatste stap voeren we op de betreffende gegevens de analyse uit en gaan we na of er tussen de geprediceerde en de feitelijke analyse resultaten verschillen bestaan.

Of een hypothese empirisch geldig is hangt af van het volgende:
- De hypothese moet voldoende gesloten zijn geweest. De hypothese is voldoende gesloten geweest wanneer alle factoren die in de realiteit van wezenlijk belang zijn voor de structuur van het causale proces ook in de hypothese zijn opgenomen.
- Soms leveren verschillende hypothesen dezelfde predicties op. De hypothesen zijn dan gelijkwaardig. Er moet dan extra empirische of theoretische argumenten worden aangevoerd om jou hypothese te verkiezen boven die gelijkwaardige andere hypothesen.
Kortom: de eventuele operationele mankementen, het mogelijk niet gesloten zijn van een hypothese en de mogelijkheid dat een andere hypothese hetzelfde prediceert, zijn redenen om niet te snel te zeggen dat een hypothese bevestigd is.

**Hoofdstuk 5: De 2-variabelem hypothese**

Een 2-variabelen hypothese spreekt een verwachting uit ten aanzien van de bivariate verdeling van die variabelen. Hebben we te maken met een causale relationele hypothese dan wordt er tussen de twee eigenschappen een causale relatie verondersteld op basis waarvan een verband in de empirische verdeling verwacht wordt.

Een VB van een hypothese met 2 variabelen:
*Kinderen uit het arbeidsmilieu gaan in vergelijking met kinderen uit het middelbaar en hoger milieu na de basisschool minder door naar het vwo.*

Asymmetrisch verband: De bovengenoemde hypothese veronderstelt een causaal effect in 1 richting.

Symmetrisch verband: Een wederkerig causaal verband.
VB: Voor ouders met kinderen op de basisschool bestaat er een verband tussen het bekend zijn met het vwo onderwijs en de mate van ouder participatie.

Schema 1: Overzicht van 2 variabelen hypothese

Hypothese pijldiagram Predicties
1. Asymmetrisch X🡪Y Verband tussen X en Y
verband X🡨Y is ongelijk aan 0

2. Symmetrisch X🡨🡪Y Verband tussen X en Y
verband is ongelijk aan 0

3. Geen verband X-/->Y Verband tussen X en Y
 X<-/-Y is gelijk aan 0
 X<-/->Y

VB van en hypothese dat er geen enkel (causaal) verband bestaat tussen de 2 variabelen:
Voor wat betreft het bekend zijn van de ouders met het vwo onderwijs, maakt het niet uit of hun kind een jongen of een meisje is.

VB van een toetsing van een bivariate hypothese op basis van bovengenoemde hypothese: lees blz 58, 59,60 en 61.
-Operationaliseren van de hypothese
-deze uitdrukken in een pijldiagram
-Opstellen van predicties
- controleren van de predicties

Associatiemaat: Zo druk je de samenhang in een verdeling uit. Het meeste wordt het percentage verschil gebruikt. Genoteerd als D% met de D van Difference.
Empirische verdelingen met een samenhang van exact 0 zijn uiterst zeldzaam. De vraag doet zich dan ook voor of alle afwijkingen van 0 ook als echte afwijkingen van 0 beoordeeld moeten worden. Voor dit probleem zijn 2 oplossingen die in de praktijk worden gehanteerd: Het statistisch significantie criterium en het eigen criterium.

Statistsich significantie criterium: Wordt veruit het meeste gehanteerd. Hierbij gaat men er van uit dat een verdeling en dus ook de gevonden samenhang mede door een speling van het toeval kan zijn ontstaan. Volgens dit criterium is men pas bereid om een gevonden samenhang als een echte, van 0 afwijkende samenhang te accepteren, als de kans dat men die samenhang of een nog sterkere samenhang gevonden zou hebben, erg klein is. Meestal stelt men de kans op 0,05 of kleiner. > 0,05. Wanneer dit van een bepaalde geconstateerde samenhang gezegd kan worden , noemt men die samenhang statistisch significant.

Eigen criterium: Wanneer men bijvoorbeeld stelt dat een d% kleiner dan 5% te klein is om relevant te zijn. Bij methoden houden we dit op kleiner dan 10%.

Samenvattend kunnen we stellen dat er met betrekking tot een verband tussen twee variabelen drie soorten causale hypothesen zijn te formuleren: Er is een asymmetrische, een symmetrische of helemaal geen causaal verband.

**Hoofdstuk 6: De 3-variabelen hypothese**

De studies beginnen meestal met een bivariate samenhang. Omdat deze samenhang aan het begin staat van verdere analyse, wordt ze de oorspronkelijke samenhang genoemd. Eenmaal die geconstateerd te hebben, wil men vervolgens meer inzicht krijgen en stelt men zich de vraag welke andere eigenschappen in het causale proces waarin de twee variabelen betrokken zijn, een belangrijke rol spelen. Om daar een antwoord op te krijgen worden uitgebreidere hypothesen over het proces opgesteld en wordt de analyse uitgebreid door de hogere verdelingen te onderzoeken dan die van een bivariate.
Men zegt dan dat de oorspronkelijke samenhang wordt geelaboreerd (uitgewerkt) naar andere variabelen.

De verschillende resultaten die deze nadere uitwerking kan hebben, worden typen van elaboratie genoemd. Elke soort hypothese zegt dat de derde variabele op een specifieke manier met de oorspronkelijke relatie te maken heeft.

*Onderscheid conjunctieve en disjunctieve hypothesen 6.1*

De letterlijke betekenis van conjunctief is ‘samengevoegd’ en van disjunctief ‘naast elkaar gevoegd.’

Conjunctieve hypothese: Bij een conjunctief proces (of hypothesen omtrent zo’n proces) gaat het om de meerdere factoren die samen een effect bewerkstelligen. Van elke factor kunnen we dan zeggen dat hij een relatie heeft met de afhankelijke grootheid onder conditie dat de andere factoren een gunstige stand innemen.

Disjunctieve hypothese: Een disjunctief proces verloopt anders. Ook hier is er sprake van meerdere factoren die ergens een invloed op hebben, maar niet conditioneel voor elkaar. Ze zijn niet samen tegelijkertijd nodig om het effect te bereiken.

*De conjunctieve hypothese (de interactie hypothese) 6.2*

Zoals er eerder in het boekje werd besproken wordt er in elke hypothese, bestaande uit drie variabelen, een verwachting uitgesproken ten aanzien van de trivariate verdeling van die variabelen. ( zie voor tabel met trivariate verdeling blz 67)

Drie variabelen ter VB: X= sociaal milieu met categorieen: arbeidersmilieu en middelbaar en hoger milieu
Y= vervolg onderwijs met de categorieen geen vwo en vwo
T= geslacht met de categorieen jongen en meisje

In het algemeen zegt een conjunctieve hypothese dat het verband tussen x en y binnen de ene categorie van t (jongen) anders is dan de andere categorie van t (meisje). Je zou daarom kunnen zeggen dat de categorieen van T condities vormen voor X en Y. Een andere uitdrukking is dat T dit verband specificeert.
Een conjunctieve hypothese wordt daarom ook wel conditionele of specificatie hypothese genoemd, of de interactie hypothese (ontleend aan de statistiek)
Zij geeft aan dat 2 variabelen in interactie met elkaar inwerken op een derde variabele. Oftewel X en T interacteren op Y. Bij onderzoekers is de term interactie hypothese het meeste gangbaar.
kortom: De relatie tussen X en Y verandert onder de conditie van de andere variabele T.

*Drie typen conjunctieve hypothese 6.2.1*

De conjunctieve hypothese dat het oorspronkelijke verband anders ligt in categorieen van variabele T, is nog erg vaag en kan concreter worden aangeduid. Er zijn een drietal concretiseringen die nogal van elkaar verschillen en die achtereenvolgens te maken hebben met de aspecten van aanwezigheid, sterkte en richting van een verband. Aan de hand van voorbeelden worden deze toegelicht.

1. Voor arbeidskinderen geldt dat, als ze in de laatste klas van de basisschool een onderwijzer hebben die sterk leerling gericht is, zij eerder naar het vwo doorstromen dan wanneer ze die niet hebben; voor kinderen uit een ander milieu maakt dit geen verschil.

In deze hypothese wordt gezegd dat er in 1 van de categorieen van T (categorie arbeidersmilieu van variabele sociaal milieu) een verband *aanwezig* is tussen het leerling gericht zijn van de onderwijzers en het doorstromen naar het vwo. De andere categorie van T ( middelbaar en hoger milieu) is dit *niet aanwezig*.
Deze conjunctieve hypothese bakent het gebied dus af waarbinnen het verband tussen 2 variabelen verwacht wordt. Zij beperkt het bereik hiervan. Variable T geeft dit gebied aan.

2. Het verband tussen sociaal milieu en bezoek vwo in de zin van dat er meer kinderen uit het middelbaar en hoger milieu naar het vwo gaan dan uit het arbeidersmilieu is bij meisjes sterker dan bij jongens.

In deze hypothese wordt gesteld dat er voor beide groepen, jongens en meisjes, een verband bestaat tussen sociaal milieu (x) en vervolgonderwijs (y) en dat dit verband van verschillende *sterkte* is. De variabele geslacht (t) geeft de verschillende intensiteiten aan: bij meisjes is het verband sterker dan bij jongens. De hypothese zal des te ongeldiger zijn als blijkt dat de samenhang bij jongens meer afwijkt dan die bij de meisjes.

3. Het positief staan van de onderwijzers tegenover het gebruik van dialect op school beinvloed in positieve zin, de schoolprestaties van een kind uit een arbeidersmilieu, maar heeft een averechte werking op de prestaties van een kind uit het middelbaar en hoger milieu.

In deze conjunctieve hypothese worden in de categorieen arbeidersmilieu en middelbaar en hoger milieu tegenover gestelde verbanden verwacht. Er wordt dus benadrukt dat de *richting* van de samenhangen in de twee milieugroepen verschilt. Dit betekent dat in deze milieus het genoemde verband verschillend geïnterpreteerd zal moeten worden.

Nu is het gemakkelijk de pijldiagrammen en predicties van bovenstaande conjunctieve hypothesen op te stellen.

In de praktijk worden kleine verschillen tussen samenhangen in de verschillende condities niet relevant geacht. Men hanteert hierbij het statistisch significantie criterium. Op grond van toeval kunnen samenhangen in subgroepen verschillen en die verschillen nemen we niet serieus. Dus: NIET statistisch significant= < 5% = op grond van toeval.

Zie het overzicht van de conjunctieve hypothesen met drie variabelen op blz 70. Of kijk dit even na in de collegesheets van de syllabus.

Men moet goed realiseren wat in zo een pijldiagram staat. Er is op de eerste plaats een pijl getrokken van x naar y; deze geeft in de werkelijkheid aan dat er een invloed verwacht wordt van x op y.
De andere pijl die vanuit T gaat (dit is belangrijk!!) naar de pijl tussen de x en de y. Dit is de kern van het statistisch begrip van interactie: een relatie verandert (tussen x en y) onder conditie van een andere variabele (t).
De pijl vanuit T gaat dus niet naar x of y. In een conjunctieve hypothese wordt daarom doorgaans niets gezegd over een verband tussen t en x of over een verband tussen t en y.
Bij hypothese 3 kan daarom een mechaniek diagram tekenen te vinden op blz 71.
Het sociaal milieu (t) beinvloed dus de relatie tussen x en y hier.
Men moet er goed op letten dat interactie wil zeggen dat niet een variabele, maar een relatie tussen variabelen afhankelijk is van andere variabelen.

Bij maatschappelijke processen zijn steeds twee interpretaties mogelijk. Men kiest meestal voor de interpretatie die het meeste voor de hand ligt. Men kiest hier meestal voor de variabele waarvan de categorieen relatief het meest stabiel zijn. ( De persoon kan in deze variabele niet zo gemakkelijk van categorie wijzigen, geslacht, leeftijd etc.)

Zie voor voorbeeld van een toetsing van conjunctieve hypothese blz 73 en 74.

Samenvattend kunnen we zeggen dat in een conjunctieve hypothese een invloed verondersteld wordt van een derde variabele T op het verband tussen 2 andere variabelen (x en y meestal) Die invloed kan op drie verschillende wijzen worden geconcretiseerd: In een van de categorieen van T kan het verband wel voorkomen en in de andere niet, de sterkte van het verband kan varieeren binnen de categorieen van T en aks derde mogelijkheid kan de richting van het verband in 1 van de categorieen van T positief zijn en in de andere negatief.
Volgens het boekje dient deze vorm van hypothese voorrang te krijgen op de disjunctieve hypothese. Volgens onze leraren is het juist weer andersom......

*De Disjunctieve hypothese 6.3*

Variabelen beïnvloeden variabelen en niet relaties. (verschil met conjunctieve).
Het algemene karakter van een disjunctieve hypothese wordt uitgelegd en een onderwerp wordt aan de orde gesteld dat met de toetsing van een disjunctieve hypothese te maken heeft 🡪 Het elimineren van de invloed van een variabele.

Het algemene karakter van een disjunctieve hypothese:

VB. Als een kind uit het arbeidersmilieu komt, is het in vergelijking met kinderen uit hogere of middelbare milieus minder in staat in abstracte begrippen te denken en het minder kunnen denken in abstracte begrippen heeft tot gevolg dat het minder snel op het vwo terecht zal komen.

In deze hypothese worden 2 relaties gesuggereerd die zich na elkaar binnen 1 causaal proces afspelen. Ze kunnen in het volgende pijl diagram worden weergegeven:
x= Sociaal milieu T= denken in abstracte begrippen Y= vwo vervolg onderwijs:
X🡪T🡪Y De relatie tussen X en T is onafhankelijk van de relatie tussen T en Y.

In dit proces beinvloeden variabelen elkaar: X beinvloed T en T beinvloed Y. De relaties zelf zijn niet van andere variabelen afhankelijk.
In Disjunctieve hypothesen worden relaties verwacht die onafhankelijk zijn van de waarden van andere variabelen.
Bij een disjunctieve hypothese mag geen statistische interactie aanwezig zijn.
Dus bij een trivariate verdeling van variabelen mag de relatie tussen x en y niet varieren over de categorieen van T.
Een trivariate verdeling is af te leiden uit een bivariate verdeling als een relatie tussen twee variabelen niet afhangt van de derde variabele.

Een disjunctieve hypothese wordt dus gefalsificeerd wanneer er veel statistische interactie voorkomt. Dit betekent dus dat een relatie in subgroepen, hoe dan ook te formeren, niet sterk mag varieeren.

Een manier om vooraf te controleren of er van een statistische interactie sprake is, is eenvoudig nagaan of de sterkte van de betreffende relatie in bepaalde subgroepen verschilt.

Assumptie van geen interactie: In plaats van materiaal te analyseren wordt vaak simpelweg aangenomen dat er geen interactie aanwezig is. Hiervoor moeten dan gegronde redenen voor aanwezig zijn.

Voor alle disjunctieve hypothesen die in het boekje worden behandeld geldt dat zij een causaal proces veronderstellen dat in 1 richting verloopt.

Het elimineren van de invloed van een variabele.
In een onderzoek wil men vaak de invloed van de ene variabele op de andere vaststellen en men wil er dan bovendien helemaal zeker van zijn dat die invloed juist aan die ene variabele moet worden toegeschreven en niet aan een derde variabele. De invloed van de derde variabele wil men dan elimineren. Men spreekt dan ook wel van het uitschakelen of controleren of constant houden van een variabele.

Bij de disjunctieve hypothesen die in het boekje worden besproken zijn telkens drie variabelen in het geding. 1 variabele is de afhankelijke variabel. We zullen deze de gevolg-variabele Y noemen. De andere 2 variabelen, x en t, kunnen beiden al dan niet of in verschillende maten een invloed hebben op Y. Om erachter te komen in welke mate variabele x onafhankelijk van de variabele t, een invloed heeft, moeten we de invloed van T elimineren. Lees voor VB blz 80, 81 en 81 (aanrader) Belangrijk is dat er bij een D% uitkomst bij een bivariate verdeling een derde variabele van invloed kan zijn op de uitkomst van D%. Dit is op dat moment de verborgen variabele T (kennis van abstracte begrippen). Hierbij rijst de vraag in welke mate de differentiele deelname aan het vwo (d%) kan worden toegeschreven aan verschillen in sociaal milieu wanneer de kennis van abstracte begrippen verwijderd zijn??

In een drie dimensionale tabel worden twee partiele tabellen toegevoegd die beiden een ‘stand’ van T (meer of minder kennis van abstracte begrippen) toegevoegd. De partiele percentage verschillen in tabel 11 op blz 82 laat zien dat deze afwijken van de D% van de oorspronkelijke relatie (tabel tien de bivariate verdeling). De partiele percentage verschillen laten zien in welke mate x en y met elkaar gerelateerd zijn NA eliminatie van de invloed van T (door de tabel te partialiseren in twee standen van T haal je de invloed van T weg op de variabelen x en y). Je ziet dat in de oorspronkelijke relatie in tabel 10 de D% nog -24 was en in de partiele tabellen de invloed van het sociaal milieu tot praktisch 0 is gereduceerd 3 en 0.

De gecontroleerde samenhang in 1 getal uitgedrukt
Bij bovenstaand verhaal ontstaat een probleem omdat je twee differences krijgt door de twee partiele tabellen: namelijk D%=3 en D%=0. Om deze gecontroleerde samenhang te kunnen vergelijken met de ongecontroleerde samenhang van de oorspronkelijke relatie D%=-24, dien je de gecontroleerde samenhang in 1 getal vast te leggen.
Een oplossing hiervoor is het nemen van een soort gemiddelde van de twee percentageverschillen in de twee partiele tabellen van de tabel 11. Het leidt tot een maat die bekend staat onder gestandaardiseerde partiele regressie effect (niet belangrijk) oftewel het *partiele percentage verschil* ( wel belangrijk).
Omzetting naar 1 getal wordt pas echt nodig als je de invloed van een variabele, gecontroleerd voor een andere variabele, kwantitatief wilt uitdrukken en de relaties in de partiele tabellen toch redelijk verschillen. Het partiele percentage verschil bij tabel 11 zou dus zijn (3+0)/2=1.5.

Een nadeel ervan is echter dat alle partiele relaties dan even zwaar meetellen, terwijl dat om 2 redenen ongewenst kan zijn:
-Om de reden van ongelijke subgroepen. In tabel 11 is de eerste partiele tabel berust op 22 kinderen en de tweede op 28 kinderen. De partiele relatie in de tweede groep berust dus op waarnemingen. Deze dient dus eigenlijk zwaarder meegeteld te worden dan de eerste groep.

-de tweede reden hangt samen met de variatie in variabele X. Er is spreiding nodig in de subgroepen van de variabele x. Als bijvoorbeeld 20 kinderen van de 22 uit de eerste partiele tabele uit middelbaar of hogere milieus zouden komen en maar 2 uit de arbeiders milieus geeft dit geen eerlijke weergave in de partiele relatie. Je kan dan dus niet spreken van een samenhang. Het beste is dus als in de subgroepen evenveel deelnemers zitten. Dus in beide sociale milieu groepen evenveel kinderen. De variatie naar sociaal milieu zou dan maximaal zijn.

Voor de maat van spreiding gebruikt men wel het produkt van de twee proporties waarin de categorieen voorkomen. Dus in de eerste partiele tabel van tabel 11 is de spreiding naar sociaal milieu (18/22)\*(4/22)=0.15 en in de tweede partiele tabel (7/28)\*(21/28)=0.19. De spreiding is bij 0,19 wat groter en deze dient dan ook wat zwaarder meegewogen te worden om een eerlijk partieel percentage verschil mee te laten wegen
Een weging naar zowel de grootte van de subgroep als naar de variatie van x in de subgroep, wordt bereikt door de grootte met de variatie te vermenigvuldigen:
De eerste partiele samenhang: 22\*0,15=3,3 en voor de tweede 28\*0,19=5.3. Je zou dit de absolute gewichten w1 en w2 kunnen noemen. 22 is de grootte van de groep en 0,15 is de spreiding, 3.3 is de absolute hoogte van de frequenties.

Om de frequenties afhankelijk te maken van het absolute berekend men de relatieve gewichten. Dit gaat als volgt:
Voor partiele relatie 1: w1/(w1+w2)= en w2/(w1+w2)
Hieruit volgt dat de som van de absolute gewichten 3.3+5.3=8.6 =(w1+w2)
De relatieve gewichten zijn dus: 3.3/8.6=0.38 en 5.3/8.6=0.62.

Met deze gewichten worden de partiele samenhangen van D%=3 en D%= 0 gewogen. De gewogen som hiervan (0.38\*3)+(0.62\*0)=1,1 staat voor het partiele percentage verschil. Dit noteer je als volgt: D%(yx.t)=1.1dit getal kun je nu vergelijken met het getal van de oorspronkelijke relatie D%(yx)=-24

Dus (relatief gewicht van T1\*D%T1)+(relatiefgewicht van T2\*D%T2)=D%(YX.T)

De samenhang tussen sociaal milieu en vervolg onderwijs blijkt sterk terug te vallen wanneer we de invloed van de kennis van abstracte begrippen elimineren.

Drie disjunctieve hypothesen:

Tussen drie variabelen kunnen zich verschillende disjunctieve hypothesen voordoen. Zie de eerste drie pijldiagrammen op blz 86:
1. Hypothese: Interpretatie, verklaring en ketenmodel
2. Hypothese: Explanatie, schijnrelatie en wegverklaring
3. Hypothese: Hybride, gedeeltelijke verklaring van relatie (xy), gedeeltelijke schijnrelatie (xy)

1. Het interpretatieve type
Dit type is erg belangrijk en komt als denkmodel veel voor.
Variabele X heeft alleen maar via variabele T invloed op Y

Dus VB:
Arbeidskinderen gaan minder dan kinderen van de hogere of middelbare milieu (x) door naar het vwo (y) omdat ze minder abstracte begrippen kennen (t)

Pijldiagram:
(x) soc. Milieu🡪 (t) denken in abstr. Begr. 🡪 (y) vwo vervolg onderwijs

Welke predicties zijn hier nu uit af te leiden voor de verdeling van de variabelen?
twee predicties zijn voldoende. Om die af te leiden kun je het beste de pijldiagram weer omzetten in een causaal mechaniek: blz 87. Wanneer X draait, draait T en daardoor draait Y weer. Y varieert dus mee met X door T.
De eerste predictie is dan ook: als we een tabel bekijken waarin X is afgezet tegen Y, Y met X mee varieert. Er moet dus een samenhang zijn tussen X en Y.
Wanneer deze predictie uitkomt weet je nog niet of dit komt doordat de invloed van X op Y via T verloopt. Dit controleer je door de clausale mechaniek “los te koppelen” Als je nu aan X draait, draait T niet en daardoor Y niet.
De tweede predictie is wanneer we geen verandering in T toelaten ( dus in de eerste en tweede partiele tabel van tabel 11) de samenhang tussen X en Y 0 zal zijn.
Partiele relaties (xy) waarin voor T gcontroleerd is, moeten derhalve, als het proces tussen x en y uitsluitend via T verloopt, gelijk zijn aan 0. (zie schema blz 86). Deze twee predicties worden gewoonlijk gesteld als er sprake is van een interpretatie disjunctieve hypothese.
Om dit te controleren zou je dus moeten nagaan of er daadwerkelijk een relatie bestaat tussen de variabelen T en Y. Om na te gaan of de relatie (ty) positief is zal er een oorspronkelijke tabel hiervan opgemaakt dienen te worden. Zie blz 89.

Replicatie= wanneer de relaties in partiele tabellen nagenoeg gelijk zijn aan de relatie in de oorspronkelijke (OR= oorspronkelijke relatie) tabel. Dus D%(t1) en D%(t2)=D%OR
Het is een regel bij deze vorm van hypothese dat afzonderlijke predicties tot dezelfde conclusie leiden.
De relatie van (XT) onderzoeken gaat niet. Dat betekent dat je Y zou moeten elimineren en dit gaat niet aangezien Y een afhankelijke variabele is. Y heeft dus geen invloed en is alleen maar het gevolg van een andere variabele.

Dit type wordt het interpretatie of verklaringstype genoemd omdat variabele T het verband tussen X en Y volledig verklaart (interpreteert); met T wordt de reden aangegeven waarom x en y iets met elkaar te maken hebben.
variabele T wordt daarom een interveniërende variabele genoemd.

2. Het schijn relatie type

Men spreekt van een schijnrelatie tussen twee variabelen wanneer de variabelen wel in de empirische zin, maar niet in causale zin samenhangen. De relatie tussen het aantal kinderen dat handschoenen op een dag draagt (x) en het aantal blaadjes dat die dag aan de bomen zit (y) is hier een voorbeeld van. Je kan vast stellen dat hoe minder blaadjes aan de bomen hoe meer kinderen er handschoenen dragen. Maar er is geen sprake van een causaal verband dat er meer blaadjes van de bomen vallen omdat er meer kinderen handschoenen dragen.

Een VB pijldiagram
(x) in het ziekenhuis liggen 🡨 (t) ziek zijn 🡪 (y) sterven
T zet zowel x als y in beweging, maar x zet y niet in beweging door t.
omdat variabele T in causale zin vooraf gaat aan x en y wordt t een antecederende variabele genoemd. Dus in de OR de relatie tussen x en y wel aanwezig. Na elimineren van T dus in de PR relatie tussen x en y niet aanwezig.

3. Het hybride type
Voorafgaande hypothesen zijn weinig realistisch. Meestal hebben hypothesen beide typen in zwakkere vorm in zich. Dit noemt men soms het hybride type. Een wezen dat in twee heterogene elementen is samengesteld.

VB. Dat arbeiderskinderen minder naar het vwo doorstromen dan andere kinderen kmt voor een deel doordat hun ouders minder bekend zijn met het vwo.
Hierbij is er de relatie (xy) zowel als de relatie (xy.t)
Doorstroming naar het vwo (y) is afhankelijk van het sociaal milieu (x) en van het sociaal milieu (x) en of de ouders bekend zijn met het vwo (t).

Dus X🡪Y en X🡪T🡪Y
Er zijn vanuit x dus twee impulsen, via t en rechtstreeks, die y in beweging zetten.

Je kan hier spreken van het gedeeltelijk interpretatie type (minder strenge variant van type 1) en van het gedeeltelijke schijn type (minder strenge variant van type 2).

Zie blz 100 en 101 voor een vb van een toetsing van een disjunctieve hypothese.

Wanneer je de predicties toets en ze blijken uit te komen, neem je de hypothese niet aan. Je kunt hem beter: ‘niet verwerpen’.

Wanneer het partiele percentage verschil =0 is de predictie dus uitgekomen.

Voor een kleine algemene samenvatting van de verschillende hypothesen: zie blz 105 en 106