2

Dataverzameling

We verzamelen gegevens omdat we iets te weten willen komen. We willen bijvoorbeeld weten of kinderen kunnen optellen en welke begrippen ze beheersen. Soms willen we iets weten van een individu, soms van een bepaalde groep individuen, bijvoorbeeld van een etnische minder- heid. We kunnen individuen onderling vergelijken of hen stuk voor stuk vergelijken met een norm. Dikwijls zijn we niet in de eerste plaats geïnteresseerd in een vergelijking van individuen, maar in een vergelijking van vragen en opgaven. Dan kunnen we ons afvragen of de ene opgave moeilijker is dan de andere, maar ook of vragen bepaalde gewenste eigenschappen hebben. Om dergelijke vragen te beantwoorden, is het meestal nodig op systematische wijze gegevens te verzamelen en data te analyseren.

In dit hoofdstuk komen begrippen ter sprake die in de volgende hoofdstukken worden gebruikt. In paragraaf 2.1 wordt beschreven op welke wijze men van waarnemingen tot data komt. De nadruk ligt er op dat waarnemingen op zichzelf beschouwd niets zeggen, maar dat zij geïnterpreteerd moeten worden. Aansluitend hierop worden er in paragraaf 2.2 diverse schaalniveaus behandeld. We gaan er van uit dat waarnemingen worden gecodeerd in getallen; men noemt dit wel het scoren van de waarnemingen. Schaalniveaus hebben te maken met de eigenschappen die men aan de gebruikte scores kan toekennen. Dat men zich in de praktijk vaak gemakkelijk schikt in assumpties over schaalniveaus, en dat men dit vaak zonder bezwaar kan doen, wordt uiteengezet in paragraaf 2.3. In paragraaf 2.4 komen enige algemene procedures voor het verzamelen van data aan de orde. Twee belangrijke begrippen die bij zulke procedures behoren, zijn betrouwbaarheid en validiteit; zij worden kort behandeld in paragraaf 2.5. In paragraaf 2.6 bespreken we het gebruik van steekproeven van personen. In paragraaf 2.7 gaan we in op het gebruik van proefopzetten; dat zijn procedures om stimuli over personen te verdelen. In paragraaf 2.8 bespreken we de soorten stimuli die voorkomen in de psychometrie, en in paragraaf 2.9 het gebruik van meetmodellen.

2.1 Van waarnemingen tot data

We observeren in het algemeen het gedrag van personen. We beperken ons hier tot het gedrag dat personen vertonen op vragen en opgaven: het gaat om de antwoorden die de personen geven en om de wijze waarop zij een taak volbrengen. Het is van groot belang, vast te stellen dat we observaties nog geen data noemen. Pas als we een interpretatie aan de observaties geven, spreken we van data. Zoals Bezembinder (1970, p. 41) het uitdrukt: "Data zijn relaties tussen objecten, en deze relaties zijn interpretaties van observaties. Kale, niet-geïnterpreteerde observaties, bestaan niet. Maagdelijke data evenmin. De onschuldige observatie is een fictie." Een goed voorbeeld hiervan is te vinden in een artikel van Lord (1953). Een professor geniet het voorrecht, de rugnummers te mogen uitdelen aan de spelers in het rugbyteam. De eerstejaars-studenten beklagen zich: zij zouden wel erg veel lage nummers hebben gekregen. De professor verweert zich tegen de aanklacht door er op te wijzen dat rugnummers slechts etiketten zijn: zij houden slechts de spelers uit elkaar, en de nummers hadden ook letters en plaatjes mogen zijn. Als getuige à charge treedt de statisticus van de universiteit op. Deze voert blijmoedig een t-toets uit voor twee groepen, en stelt vast dat de klagers gelijk hebben. Aan de mededeling dat de rugnummers slechts etiketten zijn, heeft hij geen boodschap: "Die nummers weten immers niet waar zij vandaan komen". We zien dat de studenten de rug- nummers interpreteren als kwalificaties: die rugnummers zouden een ordening in de spelers aanbrengen. De professor ziet de rugnummers als naamkaartjes en hecht geen betekenis aan de numerieke eigenschappen van de rugnummers. De crux van het verhaal is natuurlijk de rol van de statisticus: kan hij wel rugnummers van spelers middelen en hun spreiding bepalen? "Natuurlijk kan ik dat; ik heb het toch zojuist gedaan?" antwoordt de statisticus in het verhaal.

2.2 Schaalniveaus

Het probleem dat is verwoord in het zojuist geparafraseerde artikel van Lord, betreft de toelaatbaarheid van rekenkundige operaties op in getallen weergegeven observaties. Men spreekt wel van het probleem van het schaalniveau. We gaan er van uit dat alle observaties op de een of andere manier zijn omgezet in getallen. Een schaal is een verzameling getallen en tussen die getallen gedefinieerde relaties die een empirische interpretatie hebben. De aan waarnemingen toe- gekende scores zijn getallen die tot

een schaal behoren. Door de met de schaal gegeven empirische interpretatie kan men op grond van de scores empirische uitspraken over de waarnemingen doen. Scores worden geacht van een bepaald schaalniveau te zijn als zij bepaalde transformaties kunnen ondergaan zonder dat de interpretatie van de getallen verandert. Men kan met scores rekenen; het gaat er om vast te stellen welke rekenkundige bewerkingen tot resultaten leiden die geïnterpreteerd kunnen worden in termen van de oorspronkelijke waarnemingen. Hoewel het aantal te onderscheiden schaalniveaus in beginsel heel erg groot is, maakt men doorgaans alleen maar onderscheid in de volgende vijf schaalniveaus: nominaal, ordinaal, interval-, ratio- en absoluut schaalniveau. Deze schaalniveaus zijn opgesomd in volgorde van afnemende vrijheid. Elk volgend schaalniveau in de opsomming laat minder manipulaties met scores toe, maar verschaft meer informatie.

Het nominale schaalniveau biedt de onderzoeker grote vrijheid in het manipuleren van scores. De aan observaties toegekende getallen mogen worden vervangen door willekeurige andere getallen mits men zich aan de volgende beperking houdt: aan observaties waaraan gelijke respectievelijk verschillende getallen zijn toegekend, worden na de transformatie wederom gelijke respectievelijk verschillende getallen toegekend. De getallen dienen er slechts toe, als gelijk beschouwde observaties dezelfde scores te geven en als verschillend beschouwde observaties verschillende scores te geven. Daaruit blijkt dat de scores weinig informatie verschaffen. Zij geven slechts aan welke observaties men als gelijk respectievelijk verschillend beschouwt. Het is niet mogelijk te spreken over de mate waarin observaties verschillen. De toegekende getallen fungeren slechts als etiketten of namen; hieraan ontleent het besproken schaalniveau zijn naam. Het is van belang er op te wijzen dat de onderzoeker uiteindelijk bepaalt van welk schaalniveau hij zijn observaties acht. De professor uit het artikel van Lord beschouwt de rugnummers van de studenten als observaties van nominaal niveau: de rugnummers dienen er slechts toe de studenten uit elkaar te houden. In zijn ogen heeft het dan ook geen zin het gemiddelde rugnummer te berekenen: dat getal betekent even weinig als de gemiddelde naam. De studenten in het artikel van Lord zijn een duidelijk andere mening toegedaan. Zij beschouwen de rugnummers als een aanduiding van een ordening onder de studenten. Aan de klagers zouden wel erg veel lage nummers zijn toebedeeld. Die klagers vatten de rugnummers op als van, op zijn minst, ordinaal schaalniveau.

Aan observaties toegekende getallen of scores worden geacht van ordinaal schaalniveau te zijn als zij de een of andere ordening in de observaties weerspiegelen. Zulke getallen mogen worden vervangen door willekeurige andere getallen mits de ordening intact blijft. Dit wordt wiskundig uitgedrukt met de zegswijze dat men op

getallen van ordinaal schaalniveau willekeurige monotone transformaties mag uitvoeren. Voor observaties die geacht worden gemeten te zijn op ordinaal niveau heeft alleen de ordening betekenis. Men kan de observaties bijvoorbeeld onderling vergelijken in termen van groter of mooier; het is echter niet mogelijk te zeggen hoeveel groter of hoeveel mooier de ene observatie is dan de andere.

Men noemt aan observaties toegekende getallen van intervalschaalniveau als men betekenis kan hechten aan verschillen tussen dergelijke getallen. Een bekend voorbeeld van getallen die van intervalniveau zijn, is gegeven door de gangbare schalen voor temperatuur. Een voorwerp heeft een bepaalde temperatuur. Deze temperatuur kan men uitdrukken in graden Celsius maar ook in graden Fahrenheit. Voor dezelfde waarneming heeft men dus twee getallen: dezelfde waarneming is op twee manieren gescoord. De twee getallen kan men tot elkaar herleiden door er een lineaire transformatie op toe te passen. Een lineaire transformatie van x naar y schrijft men als: $y = ax+b$, waarin de getallen a en b willekeurige getallen zijn en a niet gelijk is aan nul. Doordat men zowel a als b vrij kan kiezen, zegt men wel dat men de oorsprong en de eenheid van de schaal vrij kan kiezen.

We illustreren het intervalschaalniveau aan het gebruik van de schalen voor het meten van temperatuur. Als men een bepaalde temperatuur kan beschrijven als x graden Celsius en ook als y graden Fahrenheit, dan bestaat er tussen de getallen x en y de volgende betrekking: $y = 1.8x+32$. Het is van belang er op te wijzen dat bij een lineaire transformatie de verhouding van twee verschillen constant blijft. Zij x het verschil tussen twee op de Celsius- schaal gemeten temperaturen x_1 en x_2, en x' het verschil tussen twee temperaturen x_3 en x_4. Zij de verhouding van de twee verschillen in temperatuur x en x' op de Celsiusschaal gelijk aan r: $r = x/x'$. Als men nu zowel x als x' transformeert naar de Fahrenheitschaal, krijgt men twee getallen y en y'. Daarvoor geldt dat $y = (1.8x_1+32)-(1.8x_2+32) = 1.8(x_1-x_2) = 1.8x$, en $y' = 1.8x'$. De verhouding r' van y en y' is dan gelijk aan x/x', en dus gelijk aan r. Voor getallen die geacht worden van intervalschaalniveau te zijn en dus alleen aan een lineaire transformatie onderworpen mogen worden, blijkt dat verhoudingen van verschillen onder dergelijke transformaties niet veranderen.

Men acht getallen die aan observaties worden toegekend van ratioschaalniveau, als men die getallen aan transformaties kan onderwerpen die de verhoudingen van getallen onverlet laten. De enige transformaties met deze eigenschap zijn de multiplicatieve transformaties: $y= ax$ voor een willekeurig getal a dat niet gelijk is aan nul. Een voorbeeld van meten op ratioschaalniveau is het meten van lengte. Men kan de lengte van een voorwerp uitdrukken in centimeters en in inches; maar ongeacht de keuze van de eenheid kent men het getal 0 toe aan een voorwerp dat 'geen lengte heeft'. De

meting 0 verandert niet door een multiplicatieve transformatie. Aangezien men alleen de schaalfactor *a* vrij kan kiezen, zegt men wel dat bij een ratioschaal alleen de eenheid vrij gekozen kan worden. Merk op dat verschillen tussen getallen die van intervalschaalniveau zijn, zelf van ratioschaalniveau zijn.

Men acht getallen van absoluut schaalniveau te zijn als er geen transformatie is toegestaan. Wiskundigen zeggen in zo'n geval dat alleen de identiteitstransformatie is toegestaan: elk getal kan alleen maar 'in zichzelf worden getransformeerd'. Van absoluut schaalniveau acht men bijvoorbeeld getallen die een aantal aanduiden. Zoals Bezembinder (1970, p. 73) het uitdrukt: "Een even robuust als rustiek voorbeeld van het gebruik van een absolute schaal levert ons de herder die zijn schaapjes telt".

2.3 Meten per fiat

Het is van belang er op te wijzen dat het toekennen van een schaalniveau aan getallen een activiteit is van de onderzoeker; getallen hebben niet van zichzelf enig schaalniveau. Het onderbrengen van getallen in een bepaald soort schaal is een kwestie van interpretatie. Het is vaak niet eenvoudig, vast te stellen van welk schaalniveau scores zijn. Als de herder dat zou willen, kan hij schapepoten tellen in plaats van schapen: voor hem zijn aantallen kennelijk van ratioschaalniveau. Maar dan moet hij natuurlijk geen schaap met vijf poten in zijn kudde hebben.

In de praktijk houdt men zich niet altijd intensief bezig met de vraag, van welk schaalniveau de verkregen observaties zijn. Dikwijls analyseert men data met methoden die eigenlijk getallen van intervalschaalniveau vereisen zonder dat men heeft onderzocht of zo'n assumptie gerechtvaardigd is. Uit de zinvolheid van de verkregen resultaten leidt men dan alsnog af dat de assumptie gerechtvaardigd is. Veel meetprocedures berusten op vaste af- spraken: men is het er over eens bepaalde zaken op een bepaalde manier te onderzoeken en te analyseren. Daarom spreekt men wel van meten per 'fiat'.

2.4 Procedures voor dataverzameling

De wijze waarop men gegevens verzamelt, en ook de beslissing welke gegevens te verzamelen, hangen af van een groot aantal factoren. Voor een deel zijn deze factoren bepaald door de theorie die men aanhangt, en voor een ander deel door statistische en economische overwegingen. Voor elk onderzoek is nu eenmaal een beperkt budget

beschik- baar en dat moet zo goed mogelijk worden gebruikt. Uit deze overwegingen vloeit voort dat men in elk geval op systematische wijze gegevens moet verzamelen: men zal een welomschreven procedure moeten volgen. Er zijn vele procedures om observaties te verzamelen. Deze procedures kunnen op een aantal manieren worden ingedeeld. De volgende classificaties van procedures voor het verzamelen van gegevens zijn ontleend aan Meerling (1981).

Men kan in de eerste plaats het onderscheid maken tussen directe observatie enerzijds en observatie door middel van een instrument anderzijds. Bij directe observatie nemen we het gedrag van een persoon waar en interpreteren dit gedrag direct bij waarneming. Denk bijvoorbeeld aan het observeren van het gedrag van spelende kinderen. Was die klap nu een goedmoedige por of een echte klap? Bij observatie door een instrument wordt het gedrag van een persoon geobserveerd op een stimulus die door de onderzoeker wordt aangeboden. Het gaat nu om uitgelokt gedrag. Denk aan het antwoord van leerlingen op items in een toets die optelvaardigheid meet of aan een enquête waarin gevraagd wordt naar stemgedrag.

In de tweede plaats kan men procedures onderscheiden naar de bron die de gegevens verschaft. Soms is het de onderzoeker zelf die waarneemt en dan selecteert en interpreteert, zoals de ontdekkingsreiziger in het oerwoud. Maar ook kan het de onderzochte persoon zijn, zoals de bekende Nederlander die de interviewer niet het achterste van zijn tong laat zien. Ook kan het zijn dat de observatie komt van een derde persoon, bijvoorbeeld een on- afhankelijke beoordelaar. Andere bronnen van gegevens zijn dossiers en archieven. Men maakt dan gebruik van gegevens die door anderen op een eerder tijdstip zijn vastgelegd.

In de derde plaats kan men procedures voor het verzamelen van gegevens onderscheiden naar de tegenstelling reactief en niet-reactief. Reactief noemt men de observatieprocedure die het normale gedragspatroon van de proefpersoon verstoort. Men kan hierbij denken aan experimentele behandelingen en in het algemeen aan uitgelokt gedrag. Niet-reactief noemt men procedures waarbij er geen gedrag wordt uitgelokt maar er louter wordt gekeken.

2.5 Betrouwbaarheid en validiteit

Als we het in dit boek hebben over data, hebben we het meestal over antwoorden van personen op items of uitvoeringen van opdrachten. Door deze items of opdrachten, al dan niet gebundeld in een toets, aan personen voor te leggen, hopen we iets te weten te komen over de personen en dikwijls ook over de items en de opdrachten. We

veronderstellen dat de items en de opdrachten operationalisaties zijn van het te onderzoeken gedrag. Het zijn concrete, duidelijk afgebakende stimuli die te zamen alle uitingsvormen bevatten van het te bestuderen gedrag. In hoofdstuk 3 wordt, in het deel over de generaliseerbaarheidstheorie, ingegaan op het idee van alle uitingsvormen van het te bestuderen gedrag. We interpreteren het geobser- veerde gedrag: als we optelitems voorleggen aan een leerling gaan we er van uit dat de antwoorden die de leerling geeft, ons iets zeggen over de optelvaardigheid van die leerling.

We beperken ons tot observaties door een instrument. We willen een interpretatie kunnen geven aan de observaties die verkregen worden door het voorleggen van een stimulus aan een persoon. Het gaat daarbij meestal om gedrag dat we niet direct kunnen observeren; we nemen uitingen van gedrag waar die we interpreteren als manifestaties van niet direct waar te nemen eigenschappen en vaardigheden. Zulke eigenschappen en vaardigheden noemt men wel latente variabelen. Zij zijn begrippen die in een theorie worden gepostuleerd en gedefinieerd.

Bij elke procedure voor het vergaren van data zijn twee begrippen van belang. In de eerste plaats is het belangrijk te weten wat we meten; dit is de vraag naar de validiteit van de procedure en van het instrument. Het afnemen van een instrument moet leiden tot een interpreteerbare observatie van het gedrag van de leerling op de vragen en de opdrachten. De geïnterpreteerde reactie geeft binnen het kader van de theorie aan, welke conclusies we kunnen trekken. Als we een leerling een optelopgave geven, interpreteren we een goed antwoord als: de leerling beschikt over voldoende optelvaardigheid om het in de opgave weergegeven probleem op te lossen.

In de tweede plaats is het belangrijk dat we een zo nauwkeurig mogelijke observatie hebben; dit is de vraag naar de betrouwbaarheid van de procedure en het instrument. Indien we een meting zouden kunnen herhalen onder identieke omstandigheden zouden we dezelfde meting moeten krijgen. Er zullen in praktijk echter altijd verstorende invloeden gelden. Zo is de eis van identieke omstandigheden meestal niet te vervullen: het aanbieden van een item zou al een leereffect kunnen hebben.

In de psychometrie besteden we aandacht aan personen, aan stimuli en aan de reacties van personen op stimuli. Analyse van de data moet antwoord geven op de gestelde onderzoeksvragen. Het moet dan mogelijk zijn individuen en groepen individuen met elkaar te vergelijken, en ook stimuli en groepen stimuli. We kunnen vaststellen dat de ene leerling beter kan optellen dan een andere, en dat de ene groep beter kan optellen dan een andere. Stimuli, bijvoorbeeld items, kunnen met elkaar worden vergeleken: het ene item is moeilijker dan het andere.

Dikwijls wil men het gedrag van een enkel persoon bestuderen. Voorbeelden daarvan zijn te vinden in de psychodiagnostiek en in het gebruik van toetsen voor het meten van

vorderingen op school. Maar even zo vaak stelt men geen belang in het individu. Zo tracht de psychonomie algemeen geldende wetten te vinden die psychologische functies beschrijven: hoe ziet een oog, hoe grijpt een hand. En in het onderwijs wil men vaak groepen personen op hun prestaties in een vak onderscheiden. Een belangrijk gebied waar groepen personen een rol spelen, is dat van het ontwikkelen van meetinstrumenten. Als een psycholoog de van een persoon verkregen responsen op een meetinstrument wil kunnen interpreteren, moet hij er staat op kunnen maken dat het instrument de tussen personen bestaande verschillen kan blootleggen. En als een leraar de vorderingen van een bepaalde leerling in de tijd wil kunnen volgen, moet hij er op kunnen rekenen dat het gebruikte instrument in staat is, werkelijk opgetreden veranderingen vast te stellen. Hier is de betrouwbaarheid van het instrument in het geding. De klassieke testtheorie, die in hoofdstuk 3 wordt behandeld, is een meettheorie waarin een kwantitatief begrip betrouwbaarheid is gedefinieerd. Om deze maat te schatten, heeft men waarnemingen nodig van groepen personen. Veel psychometrie houdt zich dan ook bezig met groepen personen. Daarbij komt men voor het probleem te staan dat men in een onderzoek veelal niet alle personen kan betrekken waar men iets over te weten wil komen. Men zal dan zijn toevlucht moeten nemen tot het trekken van steekproeven van personen. Een vergelijkbaar probleem, zeker bij het ontwikkelen van meetinstrumenten, is dat men vaak beschikt over veel kandidaatstimuli waarvan men de eigenschappen wil leren kennen; men kan echter niet alle stimuli aan elk der personen voorleggen. Men zal dan zijn toevlucht moeten nemen tot procedures om stimuli aan personen toe te wijzen. De combinatie van het trekken van steekproeven van personen en het verdelen van stimuli over de personen heet een proefopzet.

2.6 Steekproeven

Een steekproef van personen is een selectie van personen uit een duidelijk omschreven groep personen waar men belang in stelt. Deze laatste groep heet populatie, en dient zo gedefinieerd te zijn dat men van elke persoon kan vaststellen of hij tot de populatie behoort. Voorbeelden van populaties zijn: alle mensen met een leeftijd tussen vijftien en vijfenzestig jaar, en alle leerlingen uit groep acht van de basisschool in Nederland. Uit de voorbeelden blijkt dat het niet eenvoudig is een populatie te definiëren. Het zal immers vaak voorkomen dat een persoon slechts gedurende een beperkte tijd deel uitmaakt van een populatie. Wie de basisschool verlaat, verlaat tevens de zojuist als voorbeeld gegeven populatie. Men maakt daarom wel onderscheid tussen twee soorten

populaties: de doelpopulatie en de bemonsterde populatie. De bemonsterde populatie wordt ook wel aangeduid als het steekproefkader. De doelpopulatie is niet de groep maar de soort personen waar men belang in stelt. De bemonsterde populatie is de groep personen waar men een steekproef uit trekt. Bij de gegeven voorbeelden van doelpopulaties kan men de volgende bemonsterde populaties definiëren: alle mensen in Nederland die op 1 januari 1980 een leeftijd hebben tussen vijftien en vijfenzestig jaar, en alle leerlingen in Nederland die op 15 september 1990 in groep acht van de basisschool zitten. De statistiek verschaft de middelen om uit gegevens van een steekproef kansuitspraken te doen over eigenschappen van de bemonsterde populatie. In hoeverre men uit deze uitspraken iets kan concluderen over de doelpopulatie, is niet louter een kwestie van statistiek. Daarbij zijn kennis, ervaring en theoretische inzichten onontbeerlijk (Cornfield & Tukey, 1956). Voor het maken van generalisaties zijn twee statistische begrippen van belang: de representativiteit van een steekproef en de nauwkeurigheid van op steekproeven gebaseerde schattingen van kenmerken van de populatie. In het vervolg beperken wij ons tot het trekken van steekproeven uit de bemonsterde populatie, die we kortheidshalve populatie zullen noemen.

2.6.1 Representativiteit van steekproeven

Een noodzakelijke voorwaarde voor het op valide wijze kunnen generaliseren van de waarnemingen in een steekproef naar eigenschappen van een populatie, is dat de steekproef representatief is voor de populatie. De steekproef dient een goede weergave te zijn van de populatie. In beginsel kan men zich het begrip representativiteit als volgt voorstellen. De personen die deel uitmaken van de populatie kunnen op een veelheid van kenmerken worden onderscheiden. Deze kenmerken hebben een gezamenlijke verdeling in de populatie. Dezelfde verdeling van de kenmerken wil men graag terugzien in de steekproef. Als men, bijvoorbeeld, een algemene schets wil geven van de praktijk van een huisarts in Nederland, kan men niet volstaan met een steekproef van huisartsen uit Amsterdam. Daarmee kan men ten hoogste een beschrijving maken van de praktijk van een huisarts in een grote stad.

In de praktijk is het niet goed mogelijk, alle kenmerken van een populatie in beschouwing te nemen. In de eerste plaats kent men niet alle mogelijke kenmerken van een populatie. En in de tweede plaats acht men bepaalde eigenschappen niet van belang voor het onderzoek. Zo kan men zich voorstellen dat het er niet toe doet welke

kleur de auto van een huisarts heeft. Evenzo kan men zich voorstellen dat de omvang van een praktijk wel een belangrijk kenmerk is. Als men een kenmerk van een populatie in een onderzoek betrekt, kan blijken dat het kenmerk niet van belang is voor de onderzoeksvraag. In dat geval kan men vaak het bij de analyse van de gegevens gehanteerde model vereenvoudigen. Ernstiger is het buiten beschouwing laten van een kenmerk dat wel van belang is. In dit geval spreekt men van een specificatiefout. Specificatiefouten kunnen leiden tot verkeerde conclusies. Men zal zich bij het kiezen van de in een onderzoek te betrekken kenmerken van een populatie moeten laten leiden door een theorie. Men beperkt zich bij het vaststellen van de representativiteit van een steekproef tot de eigenschappen van een populatie die op grond van theoretische kennis van belang worden geacht voor het onderzoek.

2.6.2 Nauwkeurigheid

Veelal zal men op grond van een steekproef een schatting maken van een kwantitatief kenmerk van een populatie. Zo'n kenmerk noemt men een parameter van de populatie. De uit de steekproef berekende grootheid wordt een schatting van de parameter genoemd. Het voorschrift waarmee uit gegevens van een steekproef een schatting van een parameter wordt berekend, noemt men een schattingsvoorschrift of kortweg een schatter. Nu kan men vaak uit een populatie op veel manieren een representatieve steekproef trekken. Men zal dan ook, bij het gebruik van steeds dezelfde schatter, bij elke steekproef een andere schatting van de parameter kunnen vinden. Het is te hopen dat deze verschillende schattingen niet teveel uiteenlopen. Een maat voor de variatie in de schattingen is de standaardafwijking van alle mogelijke schattingen. Deze standaardafwijking heet de standaardfout van de gebruikte schatter. Bij elke schatting die wordt gerapporteerd, behoort de standaardfout vermeld te worden. Het behoeft geen betoog dat een standaardfout niet zonder meer beschikbaar is; immers, om hem te berekenen zou men moeten beschikken over alle mogelijke steekproeven. Veel standaardfouten worden dan ook geschat met behulp van hulpmiddelen uit de mathematische statistiek en de kansrekening. De statistiek leert dat veel standaardfouten omgekeerd evenredig zijn met de wortel van het aantal personen in de steekproef. Om een standaardfout te halveren, moet men dan ook in het algemeen een vier keer zo grote steekproef trekken.

2.6.3 Aselecte steekproeven

De eenvoudigste steekproef is de aselecte steekproef. Zo'n steekproef ter grootte n bestaat uit n personen uit de bemonsterde populatie. Men kan op veel manieren zo'n steekproef samenstellen; dat wil zeggen dat men allerlei n-tallen uit de populatie kan kiezen. Als elk van die n-tallen dezelfde kans heeft om getrokken te worden, spreekt men van het trekken van een aselecte steekproef ter grootte n. Aan de hand van statistische en economische criteria kan men de vereiste omvang van de steekproef bepalen. Zulke criteria zijn bijvoorbeeld: de kans op onjuiste uitspraken en de kosten van het vergaren van responsen. De aselecte steekproef is om veel redenen aantrekkelijk. Zo is de kans groot dat de steekproef een goede representatie biedt van de populatie. Als, bijvoorbeeld, een populatie voor de helft uit vrouwen bestaat, dan is de kans erg klein om bij aselect getrokken steekproeven een steekproef te verkrijgen met louter vrouwen er in. Van belang is dat het bepalen van schatters en standaardfouten bij aselecte steekproeven doorgaans redelijk eenvoudig is.

Aan de aselecte steekproef kleven echter wel enige bezwaren. Het voornaamste bezwaar is dat er geen rekening wordt gehouden met heterogeniteit in de populatie. De populatie bestaat dikwijls uit deelgroepen personen die onderling meer op elkaar lijken dan personen uit verschillende deelgroepen. Aan het verschijnsel van homogeniteit van deelgroepen wordt aandacht geschonken in paragraaf 2.6.6. Als er sprake is van homogene deelgroepen, kan men gebruik maken van een gestratificeerde steekproef.

2.6.4 Gestratificeerde steekproeven

Men maakt gebruik van gestratificeerde steekproeven als men onderkent dat de populatie bestaat uit deelgroepen die in veel opzichten van elkaar verschillen. Vaak wil men, naast uitspraken over de gehele populatie, uitspraken doen over deze deelgroepen. Die deelgroepen, strata genoemd, kunnen zoveel verschillen dat men elk stratum op een aparte manier moet benaderen. Zo maakt men bij bevolkingsonderzoeken vaak onderscheid tussen de strata urbaan of stedelijk enerzijds en ruraal of landelijk anderzijds. Niet alleen leven personen in beide strata op verschillende wijze, ook brengt elk stratum zijn eigen wijze van onderzoeken met zich mee. Te denken valt aan de verschillen in afstand en reistijd tussen twee personen in de stad en die tussen twee personen op het land. De aselecte steekproeftrekking beschouwt personen als de eenheden waarvan men een steekproef trekt. De gestratificeerde steekproef- trekking bestaat uit het trekken van een steekproef uit elk der strata.

Dikwijls is het om administratieve en logistieke redenen niet mogelijk steekproeven van personen te trekken. Zo komt het vaak voor dat men wel beschikt over een lijst met adressen van gemeenschappen maar niet over adressen van personen. Bij gemeenschappen kan men denken aan huishoudens en scholen. In zo'n geval trekt men een aselecte steekproef van gemeenschappen en onderzoekt dan alle in een gemeenschap aangetroffen personen, of trekt weer een steekproef van personen uit elke gemeenschap. In het laatste geval spreekt men van getrapte steekproeftrekking.

2.6.5 Getrapte steekproeven

Als men een bevolkingsonderzoek wil doen in een omvangrijke regio, verdeelt men vaak de regio in deelgebieden en trekt dan een steekproef van deelgebieden. De deelgebieden vormen nu de eenheden van de steekproef. Deelgebieden worden doorgaans 'clusters' genoemd. Alle personen uit een deelgebied of cluster worden onderzocht, of een steekproef van personen. De onderzoekers kunnen een deelgebied in een keer bezoeken, wat reistijd en kosten bespaart. Ook kan men denken aan leerlingen die gegroepeerd zijn in klassen en klassen die weer gegroepeerd zijn in scholen. Leerlingen uit dezelfde klas lijken in veel opzichten op elkaar omdat ze in dezelfde omstandigheden verkeren. Als men de reacties van een leerling op een instrument kent, kan men vaak al een redelijk goede voorspelling maken van de reacties van de klasgenoten. Men zou dan ook kunnen volstaan met het trekken van een steekproef uit elke klas. Om logistieke redenen is dat vaak niet mogelijk. Een school stelt bijvoorbeeld een lesuur en een gehele klas ter beschikking; dan is het niet praktisch om een steekproef van leerlingen uit de klas te trekken. Zonder hogere kosten kan men alle leerlingen uit de klas in het onder- zoek betrekken.

Diverse vormen van steekproeftrekken kunnen desgewenst gecombineerd worden. Zo kan men in elk stratum van een gestratificeerde steekproef een getrapte steekproef trekken.

2.6.6 Intraklassecorrelatie

De onderlinge gelijkenis van personen uit hetzelfde cluster van een getrapte steekproef, ook wel homogeniteit van het cluster genoemd, kan men uitdrukken in een bepaalde maat die de intraklassecorrelatiecoëfficiënt wordt genoemd. In deze paragraaf spreken we over de getrapte steekproef. De intraklassecorrelatiecoëfficiënt is gedefinieerd als

de proportie van de variantie van een variabele in een populatie die is toe te schrijven aan het effect van de clusters. Aan deze definitie ligt een uit de variantie-analyse bekende decompositie van scores ten grondslag. Elke score wordt geschreven als de som van een algemeen gemiddelde, een clustereffect, en een residu.

Het is van groot belang, te weten hoe groot de intraklassecorrelatiecoëfficiënt in een steekproef is. Natuurlijk zal deze grootheid veelal geschat moeten worden; vaak kan men er voor teruggrijpen op eerder onderzoek. Het voert te ver, in dit hoofdstuk in te gaan op het schatten van de intraklassecorrelatiecoëfficiënt. Wel willen we de lezer een indruk geven van de invloed die deze coëfficiënt heeft op het vaststellen van de omvang van de te trekken steekproef. We veronderstellen daartoe dat we het gemiddelde van een kenmerk in een populatie willen schatten met een bepaalde nauwkeurigheid. Een relatieve maat voor de nauwkeurigheid van een schatter is de precisie. De precisie van een schatter is de verhouding van de standaardfout van de schatter en de standaardafwijking van de variabele in de populatie. Zonder de waarden van de standaardfout en de standaardafwijking te kennen, kan men bijvoorbeeld toch als eis formuleren dat de standaardfout ten hoogste een tiende is van de standaardafwijking van de variabele. De precisie wordt aangeduid met het symbool π; de intraklassecorrelatie met het symbool ρ. Merk op dat een kleine respectievelijk grote waarde van π overeenkomt met een grote respectievelijk kleine precisie. Een eenvoudig voorbeeld moge het begrip precisie verduidelijken. Veronderstel dat men het gemiddelde van een variabele wil schatten met een precisie van 0.10. De standaardafwijking van de variabele is niet bekend. Het is bekend dat de standaardfout van een geschat gemiddelde gelijk is aan de standaardafwijking van de variabele gedeeld door de wortel uit het aantal personen in de steekproef. De standaardfout duiden we aan met het symbool SE. Omdat we gesteld hebben dat π gelijk is aan 0.10, kunnen we schrijven: $SE/\sigma = 0.10$. Hieruit volgt dat $SE = 0.10\sigma$. Omdat in het onderhavige geval geldt dat $SE = \sigma/\sqrt{n}$, krijgen we de vergelijking $\sigma/\sqrt{n} = 0.10\sigma$. Als we deze vergelijking oplossen, vinden we dat de steekproef moet bestaan uit $n = 100$ personen om het gemiddelde te schatten met de gewenste precisie.

Als nu in een getrapte steekproef elk der clusters bestaat uit m personen en elk getrokken cluster in zijn geheel wordt beschouwd, dan kan men afleiden dat men c clusters in de steek- proef moet hebben waarbij c gelijk is aan: $\pi^{-2}m^{-1}\{1+(m-1)\rho\}$. De afleiding van dit resultaat is te vinden in Cochran (1977). De formule geldt alleen als de populatie heel erg groot is; wij geven haar alleen voor illustratieve doeleinden. Als de intraklassecorrelatie gelijk is aan 1, blijkt c gelijk te zijn aan π^{-2}. Het doet er niet meer toe hoe groot een cluster is: als men er een waarneming uit heeft gedaan, heeft men ze immers allemaal. Als echter de intraklassecorrelatie gelijk is aan 0, blijkt c

gelijk te zijn aan $\pi^{-2}m^{-1}$. In dat geval is het aantal te trekken clusters omgekeerd evenredig met de omvang van elk der clusters.

In de praktijk neemt men vaak intraklassecorrelaties waar tussen 0.05 en 0.20. Bij wijze van voorbeeld is in tabel 2.1 voor verschillende combinaties van clustergrootte, precisie en intraklassecorrelatie aangegeven hoeveel clusters men in de steekproef moet hebben om een gemiddelde te schatten met de gegeven precisie.

Tabel 2.1

Aantal te trekken clusters bij gegeven precisie, intraklassecorrelatie en clustergrootte

	π					
	0.05		0.075		0.10	
ρ	$m=4$	$m=20$	$m=4$	$m=20$	$m=4$	$m=20$
0	100	20	45	9	25	5
0.05	115	39	52	18	29	10
0.10	130	58	58	26	33	15
0.15	145	77	65	35	37	20
0.20	160	96	72	43	40	24
0.25	175	115	78	52	44	29

Uit de tabel blijkt dat het aantal te trekken clusters toeneemt als de intraklassecorrelatie toeneemt. Dat komt doordat een relatief grote intraklassecorrelatie betekent dat elke persoon in een cluster relatief weinig nieuwe informatie aandraagt: als men er een heeft geobserveerd, kan men al vrij goed voorspellen wat andere observaties uit dezelfde cluster zullen opleveren. Ook blijkt uit de tabel dat het aantal te trekken clusters toeneemt als π afneemt en dus de precisie toeneemt. Dat komt overeen met de eerder genoemde eigenschap van een standaard- fout, kleiner te worden als het aantal observaties groter wordt. Tenslotte blijkt dat men, bij dezelfde intraklassecorrelatie en precisie, minder clusters nodig heeft naarmate de clusters groter zijn. Dit effect neemt af naarmate de intraklassecorrelatie toeneemt, om de eerder al genoemde reden van verlies aan informatieve waarde van elke waarneming.

2.7 Proefopzetten

Zoals gezegd, is het vaak niet mogelijk een persoon alle stimuli voor te leggen waar men belang in stelt. Ook hier leggen tijd en geld hun beperkingen op. Men moet dan

procedures bedenken waarmee men zo goed mogelijk de informatie inwint die men wil hebben. Zulke procedures worden toewijzingsprocedures of proefopzetten genoemd. We beperken ons hier tot enige algemene beschouwingen. Veronderstel dat, bijvoorbeeld vanwege een beperkt budget of vanwege de beperkte tijd waarin men over een persoon kan beschikken, het totale aantal te verzamelen responsen vastligt. De vraag rijst dan op welke wijze men de aantallen personen en stimuli in het uit te voeren onderzoek moet kiezen. Als de stimuli op de een of andere wijze op elkaar lijken, waardoor men uit responsen op de ene stimulus een redelijk goede voorspelling kan maken van responsen op de andere stimulus, heeft het niet veel zin alle stimuli aan personen voor te leggen. Men beperkt dan het aantal aan te bieden stimuli, en trekt een grotere steekproef van personen.

Omdat het meestal niet mogelijk is alle personen alle stimuli aan te bieden, rijst de vraag hoe men de stimuli over de personen moet verdelen. Doorgaans verdeelt men de te onder- zoeken stimuli in een aantal elkaar uitsluitende groepjes stimuli en de personen in elkaar uit- sluitende groepjes personen. Aan elk groepje personen wijst men een van de groepjes stimuli toe; men spreekt van multiple matrix sampling. Het verdient aanbeveling de verdeling van groepjes stimuli over groepjes personen evenwichtig te houden: alle stimuli en alle personen moeten ongeveer evenveel te doen hebben. Enerzijds voorkomt men hiermee dat sommige personen veel meer werk moeten verrichten dan andere; anderzijds bewerkstelligt men ermee dat grootheden die met statistische methoden worden geschat, niet erg uiteenlopen in de met schattingen nu eenmaal gepaard gaande standaardfouten. Daarom maakt men in de psycho- metrie veel gebruik van onvolledige proefopzetten. Dat zijn proefopzetten waarin stimuli zodanig aan personen worden aangeboden dat niet elke persoon alle stimuli voorgelegd krijgt.

Men kan vaak met vrucht gebruik maken van aanwezige kennis om stimuli toe te wijzen aan personen. Op theoretische gronden of op grond van eerder onderzoek stelt men vast dat de reacties van bepaalde personen op bepaalde stimuli op voorhand goed te voorspellen zijn. Het is dan zonde van de moeite en het geld zulke stimuli toch aan die personen aan te bieden. Zo kan men besluiten items die men op voorhand erg gemakkelijk acht, niet voor te leggen aan leerlingen die men op voorhand heel knap vindt: men durft de veronderstelling wel aan dat zulke leerlingen zulke items goed zullen beantwoorden.

Men kan vaststellen dat onvolledige proefopzetten eerder regel dan uitzondering zijn in psychometrisch onderzoek, op grond van de geschetste overwegingen en omdat in praktijk budgetten voor onderzoek beperkt zijn.

2.8 Stimuli

Stimuli kunnen vele vormen aannemen, van ongestructureerde vragenlijsten tot welomschreven opdrachten en toetsen die bestaan uit een aantal met elkaar samenhangende items. Welke soort stimuli men gebruikt, is natuurlijk afhankelijk van het soort probleem dat men bestudeert. Stimuli worden geacht operationalisaties te zijn van het te onderzoeken gedrag, ze moeten valide zijn. Zo ligt het voor de hand leerlingen optelopgaven voor te leggen indien men wil weten in hoeverre leerlingen getallen kunnen optellen.

In de praktijk is het operationaliseren van gedrag in stimuli geen eenvoudige zaak. In het onderwijs maakt men veel gebruik van items: vragen die door leerlingen beantwoord moeten worden. Maar ook komt het voor dat door personen vertoonde gedragingen door een of meer beoordelaars of keurmeesters worden beoordeeld. Voorbeelden daarvan zijn het kunstrijden op de schaats, het Eurovisie Songfestival en de verkiezing van Miss World. De beoordelaars beschikken over een beoordelingsschema of beoordelingsmodel; voor Miss World bevat dit model een lijst met ideale maten. In het beoordelingsmodel staat vermeld welke interpretatie aan een waarneming moet worden gegeven.

Omdat het construeren van goede stimuli erg moeilijk is, zal men doorgaans niet met een enkele stimulus volstaan als operationalisatie van het te onderzoeken gedrag. Er is dus reden genoeg om meer stimuli aan te bieden; door vaker stimuli van hetzelfde soort aan te bieden, voert men als het ware een meting herhaaldelijk uit. Men verhoogt op deze manier de betrouwbaarheid van de meting. Daarbij veronderstelt men dat niet de reactie op elke stimulus van belang is maar dat het waargenomen responspatroon betekenis heeft. De veel gehoorde uitroep "Deze vraag meet toch geen intelligentie!" snijdt dan ook geen hout; slechts de combinatie van antwoorden heeft betekenis. Die betekenis ontleent een responspatroon aan een meetmodel.

2.9 Meetmodellen

Door gebruik te maken van een meetmodel kan men een responspatroon betekenis geven, dat wil zeggen interpreteren. Een voorbeeld van een meetmodel is de Guttmanschaal (Guttman, 1950). Dit model veronderstelt dat het mogelijk is items te ordenen naar moeilijkheidsgraad
en personen naar vaardigheidsniveau. Ook veronderstelt het model dat de moeilijkheidsgraden en de vaardigheidsniveaus op dezelfde schaal zijn uitgedrukt; personen en

items liggen op dezelfde schaal. Daarmee is ook een relatie gegeven tussen elk der personen en elk der items. Personen die op de schaal rechts van het item liggen, zullen het item juist beantwoorden; de andere personen geven een fout antwoord. Als juiste antwoorden worden gecodeerd met een 1 en foute antwoorden met een 0, en men de items rangschikt van gemakkelijk naar moeilijk en de personen van dom naar knap, zal men het volgende kunnen vaststellen. Aangezien elke persoon het juiste antwoord geeft op de items die links van hem liggen en het foute antwoord op de items die rechts van hem liggen, kunnen er alleen maar de volgende antwoordpatronen voorkomen: allemaal enen, allemaal nullen, of een aantal enen die gevolgd worden door een aantal nullen. Natuurlijk weet men niet of er aan de veronderstellingen van het meetmodel is voldaan. Het meetmodel krijgt zin doordat men van de andere kant begint. Men probeert, als men de antwoorden van personen op items heeft geregistreerd, de items en de personen zo te rangschikken dat de resulterende antwoordpatronen de door het meetmodel vereiste structuur hebben. Als dat lukt, heeft men een verklaring van het vertoonde gedrag gevonden. Die verklaring is gegeven in de veronderstellingen van het meetmodel. In dit voorbeeld van een meetmodel laten we een aantal belangrijke kwesties onbesproken. Zo zal men in de praktijk altijd antwoordpatronen vinden die niet de door het model vereiste samenstelling hebben. Men kan dan het model voor onhoudbaar verklaren. Maar ook kan men het meetmodel omwerken tot een probabilistisch of kansmodel: men eist dan alleen maar dat de kans op van het model afwijkende antwoordpatronen een zekere waarde niet overschrijdt. Zulke probabilistische meetmodellen komen in dit boek uitgebreid aan de orde.

Een verzameling stimuli, te zamen met een door een meetmodel verschaft interpretatie- kader, noemt men een meetinstrument. Een vragenlijst die naar een aantal socio-economische eigenschappen van personen vraagt, behoeft geen meetinstrument te zijn. Men kan de groep personen naar een aantal concrete zaken classificeren en daarmee volstaan. Zo'n inventarisatie kan een praktisch nut dienen maar levert zonder een model geen kennis en inzicht op.

Bij een meetinstrument is er doorgaans sprake van een niet direct waar te nemen eigenschap maar van een latente variabele: de moeilijkheidsgraad van een vraag of het vaardigheidsniveau van een persoon. Als iemand veel van de hem voorgelegde optelitems goed beantwoordt, concludeert men daaruit dat hij beschikt over een grote mate van optelvaardigheid. Het is van belang er op te wijzen dat een psychometrisch meetmodel niet noodzakelijkerwijze een psychologische theorie weergeeft. Zelfs als een Guttmanschaal blijkt te passen bij een tabel met antwoordpatronen, weet men nog niet waarom sommige items gemakkelijker zijn dan andere. De gevonden rangschikking van items en personen kan echter van groot nut zijn bij het formuleren van een theorie.