2. Steekproeftrekking en metingen

Wanneer data verzameld moet worden, moet bepaald worden welke subjecten daarvoor gevraagd worden. De steekproef moet representatief zijn voor de populatie. Vervolgens moet bepaald worden wat er gevraagd gaat worden en hoe het wordt gevraagd. Het meten van een variabele moet valide en betrouwbaar zijn. Validiteit verwijst naar meten wat je wilt meten. Betrouwbaarheid heeft betrekking op de stabiliteit. Deze onderwerpen worden in dit hoofdstuk verder besproken.

Meetniveaus
Elk kenmerk dat je kunt meten van een subject noemt men een variabele. Het is een kenmerk dat kan variëren tussen verschillende subjecten in een steekproef of populatie (bijv. geslacht, inkomen, mening). De waarden die een variabele kan aannemen zijn gebonden aan verschillende meetniveaus. Deze meetniveaus zijn belangrijk, omdat deze weer gebonden zijn aan de statistische analyses die met de data gedaan kunnen worden.

Ten eerste onderscheiden we kwantitatieve en categorische variabelen. Kwantitatieve variabelen hebben een meetniveau met numerieke waarden, zoals leeftijd, aantal broers en zussen, inkomen. Categorische variabelen (ook wel kwalitatieve variabelen genoemd) hebben een meetniveau met categorieën, zoals geslacht, burgerlijke staat, lievelingseten. Hier is ook te zien hoe de meetniveaus zijn verbonden aan de statistische analyses: bij kwantitatieve variabelen kun je wel een gemiddelde berekenen (bijv. de gemiddelde leeftijd), en bij categorische variabelen kan dat niet (bijv. het gemiddelde geslacht? Dat kan niet).

We onderscheiden vier meetniveaus: nominaal, ordinaal, interval en ratio. Categorische variabelen zijn nominaal of ordinaal.

Het nominale meetniveau is puur beschrijvend. Neem de variabele geslacht. De mogelijke waarden hierop zijn man en vrouw. Er is geen volgorde waar te nemen, de ene waarde is niet hoger dan de andere. Het is een puur beschrijvend verschil.

Het ordinale meetniveau veronderstelt een bepaalde volgorde. Neem de variabele opleidingsniveau. De mogelijke waarden hierop kunnen zijn vmbo, havo en vwo. Hier is wel degelijk zo dat er een volgorde in zit, omdat havo hoger is dan vmbo, en vwo hoger is dan havo. Dit is daarom een ordinaal meetniveau. Belangrijk hierbij is echter dat de afstanden tussen de waarden niet aan te duiden zijn: je kunt niet aangeven hoe groot het verschil is tussen vmbo en havo, of het verschil tussen vwo en havo. Dit is belangrijk, omdat het een kenmerkend verschil is tussen het ordinaal en het interval niveau.

Het interval meetniveau kent wel meetbare verschillen tussen de waarden. Neem de variabele temperatuur in Celcius. Niet alleen zit er een volgorde in (30 graden is meer dan 20 graden), maar dit verschil is ook duidelijk meetbaar en consistent. Het verschil tussen 10 en 20 graden is even groot als het verschil tussen 15 en 25 graden. Het onderscheid tussen interval en ratio meetniveau ligt in het feit dat interval meetniveau geen nulpunt kent, terwijl ratio dat wel heeft.

Het ratio meetniveau kent dus waarden die numeriek zijn, een bepaalde volgorde hebben, meetbare verschillen hebben, en tot slot een nulpunt hebben. Een voorbeeld is een percentage of inkomen.

Tot slot maken we onderscheid tussen discrete en continue variabelen. Een variabele is discreet wanneer de mogelijke waarden alleen bepaalde, afzonderlijke nummers zijn. Een variabele is continue wanneer de waarden alle mogelijke waarden kan aannemen. Neem bijvoorbeeld de variabelen aantal broers en zussen (een discrete variabele) en gewicht (een continue variabele). Aantal broers en zussen is een discrete variabele omdat de mogelijke waarden 0, 1, 2, 3, etc. kunnen zijn, maar je kunt niet 2,5 broer/zus hebben. Dus niet alle waarden zijn hier mogelijk. Bij gewicht kan dit echter wel. Je kunt daar (in theorie) alle mogelijke waarden op hebben. Je kunt 70 kilo wegen, maar ook 70,1 en 70,5 en 70,52. Het is bij zo'n variabele onmogelijk om alle mogelijke waarden op te schrijven, omdat het te veel mogelijkheden zijn.

Samenvattend:
Variabelen zijn kwantitatief of categorisch. Kwantitatieve variabelen worden gemeten op een interval meetniveau. Categorische variabelen met ongeordende categorieën hebben een nominaal meetniveau. Categorische variabelen met geordende categorieën hebben een ordinaal meetniveau.

Categorische variabelen (nominaal of ordinaal) zijn discrete variabelen. Kwantitatieve variabelen kunnen zowel discreet als continue zijn. In de praktijk is het zo dat kwantitatieve variabelen die veel mogelijke waarden aan kunnen nemen, worden beschouwd als continue variabelen.

Randomisatie
Randomisatie is het mechanisme achter het verkrijgen van een representatieve steekproef. Bij “simpele random steekproeftrekking” (of: aselecte steekproef) heeft ieder subject uit de populatie een even grote kans om in de steekproef terecht te komen. Je kunt het zien alsof je ieder lid uit de populatie een nummer geeft, deze in een bak doet en er vervolgens willekeurig een aantal uittrekt. Deze willekeur is belangrijk, omdat je er zeker van moet zijn dat je data niet biased (vertekend) is. Dit zou de inferentiële statistiek nutteloos maken: je kunt dan niets zeggen over de populatie.

Data kan verzameld worden aan de hand van enquêtes, experimenten en observaties. Bij al deze methoden kan randomisatie een rol spelen.

Er zijn verschillende typen enquêtes, zoals telefonische enquêtes, persoonlijke vragenlijsten. Deze komen allemaal met hun eigen representativiteitsproblemen. Deze worden later nog besproken.

Het doel van experimenten is de reacties meten en vergelijken van subjecten onder bepaalde condities. Deze condities zijn waarden van een variabele die de reactie kunnen beïnvloeden. De onderzoeker kan bepalen welke subjecten aan welke condities worden blootgesteld. Dat is waar randomisatie een rol speelt. Hij moet op basis van willekeur de groepen indelen.

Bij observaties meet de onderzoeker waarden van bepaalde variabelen, zonder de situatie te beïnvloeden. Op basis van willekeur zou bepaald kunnen worden wie er wordt geobserveerd.

Steekproefvariabiliteit en mogelijke vertekening
Ook al trek je meerdere volledig random steekproeven, dan nog zijn deze verschillend en wijken ze allebei anders af van de populatie. De afwijking tussen de steekproef en de populatie op een bepaalde variabele noemt men de steekproeffout. Bijvoorbeeld: in de populatie staat 66% achter het beleid van de regering, maar in de steekproef is dat 68%. De steekproeffout is in dat geval 2%. Bij verschillende steekproeven heb je ook verschillende steekproeffouten. Echter, wanneer randomisatie wordt gebruikt is de steekproeffout bij steekproeven van meer dan 1000 subjecten meestal beperkt tot ±3%. Dit noemen we de foutmarge.

Naast de steekproeffout zijn er nog andere factoren die de resultaten uit een random steekproef kunnen laten variëren. Er worden er hier drie besproken: de steekproef bias, de response bias en de non-response bias.

De steekproefbias is wanneer het niet mogelijk is om vast te stellen dat alle leden uit de populatie een even grote kans hebben om in de steekproef te komen. Een voorbeeld hiervan is wanneer mensen worden opgeroepen om mee te doen aan een onderzoek. Je krijgt dan alleen vrijwilligers. Maar deze vrijwilligers kunnen op belangrijke variabelen verschillen van de mensen die zich niet aanmelden. Zij vertekenen dan de steekproefdata. Dit zie je vaak bij polls op televisie: wat voor mensen zien zulke oproepen en doen daaraan mee?

De response bias is wanneer vragen slecht worden gesteld of in een ongelukkige volgorde. Een voorbeeld is het opwekken van sociaal wenselijke antwoorden, door vragen als: “Bent u het er ook mee eens dat…?”. Respondenten willen het liever niet oneens zijn met de onderzoeker en zullen eerder instemmen, terwijl ze dat eigenlijk niet willen.

De non-response bias heeft betrekking op uitval en missing data. Sommige respondenten vallen halverwege uit om wat voor reden dan ook. Deze mensen kunnen op belangrijke variabelen verschillen van de overblijvers. Dit kan de data vertekenen, zelfs bij een random steekproef.

Niet-aselecte steekproeftrekking
Het doen van een volledige aselecte steekproef is niet altijd mogelijk. Bovendien is het soms beter om dat niet te doen. Er moet dan een niet-aselecte steekproef getrokken worden. Er zijn verschillende methoden om dat te doen.

Systematische steekproef
Bij een systematische steekproeftrekking worden de subjecten die in de steekproef moeten komen op systematische wijze gekozen. Bijvoorbeeld door elk tiende huis in een straat te selecteren.

Gestratificeerde steekproef
Een gestratificeerde steekproef verdeelt de populatie in groepen, ook wel strata genoemd. Vervolgens wordt uit elke strata willekeurig een aantal subjecten gekozen die samen de steekproef gaan vormen. Zo'n steekproef kan proportioneel en disproportioneel zijn. Bij een proportionele gestratificeerde steekproef zijn de proporties in de strata gelijk aan de proporties in de populatie. Bijvoorbeeld wanneer in de populatie 60% man is en 40% vrouw, dan moet dat in de steekproef ook zo zijn. Soms is het echter beter om een disproportionele gestratificeerde steekproef te doen. Stel je voor dat je een steekproef van 100 subjecten hebt, en dat in de populatie slechts 10% vrouw is. Dan zou je anders ook maar 10 vrouwen in je steekproef hebben. Zo'n aantal is alleen te klein om representatief te zijn en dan kun je niks zeggen over de populatie. Het is dan beter om voor een disproportionele gestratificeerde steekproef te kiezen.

Clustersteekproef
Bovenstaande steekproeven vereisen echter dat je toegang hebt tot de gehele populatie. Maar in de realiteit is dat niet altijd zo. Dan kan je beter een cluster steekproef doen. Hierbij verdeel je de populatie onder in clusters (bijvoorbeeld stadsblokken), en vervolgens kies je er willekeurig een cluster uit.

Getrapte steekproef
Een getrapte steekproef bestaat uit meerdere steekproeftrekkingen. Bijvoorbeeld wanneer er eerst willekeurig een aantal provincies worden gekozen. Vervolgens worden daar willekeurig een aantal steden in gekozen. Daarin worden willekeurig een aantal straten gekozen.