Hoofdstuk 1
Statistiek, gegevens en een kritische houding

Inhoud

Wat we gaan behandelen

- Waar gaat statistiek over?
- Hoe wordt statistiek in de praktijk toegepast?
- We bespreken het verband tussen statistiek en gegevens.
- We maken verschil tussen populaties en steekproeven.
- We maken verschil tussen beschrijvende en verklarende statistiek.
- We benadrukken het belang van een kritische houding.

Statistiek in de praktijk

Sociale netwerken en de millenniumgeneratie

Het Pew Research Center, een onafhankelijke organisatie uit Philadelphia, heeft als onderdeel van het Pew Internet & American Life Project (PIALP) meer dan 100 enquêtes over internetgebruik in de Verenigde Staten gehouden. Het PIALP heeft onlangs een reeks rapporten gepubliceerd over tieners en volwassenen in de leeftijd van 18 tot en met 29 jaar, de zogeheten 'millenniumgeneratie'. In een rapport uit 2010 met de titel Social Media & Mobile Internet Use onderzocht het PIALP de houding en het gedrag van de millenniumgeneratie ten opzichte van online sociale netwerken. Hieronder staan de resultaten van enkele aan de tieners gestelde enquêtevragen.

- Internetgebruik
  Op de vraag hoe vaak ze internet gebruiken, antwoordden de tieners:

Enkele keren per dag	36%
Circa eens per dag	27%
3-5 dagen per week	14%
1-2 dagen per week	12%
Om de paar weken	7%
Minder vaak	4%

- Sociale netwerken
  Op de vraag of ze sociale-netwerksites als Facebook of Google+ gebruiken, antwoordden de tieners:

Ja	73%
Nee	27%

- Twitter
  Op de vraag of ze Twitter gebruiken, antwoordden de tieners:

Ja	9%
Nee	91%

- Tekstberichten
  Op de vraag hoe vaak ze tekstberichten op hun mobiele telefoons versturen, antwoordden de tieners:

Dagelijks	54%
Enkele malen per week	10%
Ten minste eenmaal per week	5%
Minder dan eenmaal per week	3%
Nooit	28%

- Gemiddeld aantal telefoongesprekken
  Op een doorsnee dag verzenden en ontvangen tieners gemiddeld 10,7 telefoongesprekken op hun mobiele telefoons.
- Gemiddeld aantal tekstberichten
  Op een doorsnee dag verzenden en ontvangen tieners gemiddeld 112,4 tekstberichten op hun mobiele telefoons.

In de volgende afleveringen van 'Statistiek in de praktijk' in dit hoofdstuk bespreken we in verband met de PIALP-enquête een aantal kernbegrippen uit de statistiek, namelijk:

- Het vaststellen van de populatie, de steekproef en de het trekken van conclusies,
- Het vaststellen van de gegevensverzamelmethode en het gegevenstype,
- Het kritisch beoordelen van de ethiek van een statistisch onderzoek.

Gebaseerd op het rapport 'Social Media & Mobile Internet Use' van het Pew Research Center for the People & the Press.

1.1 De wetenschap statistiek

Waar denk je aan bij het woord statistiek? Aan enquêtes, werkloosheidscijfers van het Centraal Bureau voor de Statistiek of misschien aan het zodanig presenteren van cijfers dat ze een vals beeld schetsen (hierover bestaat een beroemd boekje: *How to lie with statistics*)? Of zie je statistiek alleen maar als een verplicht onderdeel van je opleiding? We hopen dat je na het bestuderen van dit boek het met ons eens bent dat statistiek een nuttige wetenschap is, met een vrijwel eindeloos scala aan toepassingen in het bedrijfsleven, bij de overheid, in de natuurwetenschappen en in de sociale wetenschappen. Ten slotte laten we de sleutelrol zien die statistiek speelt in het kritisch denken – tijdens college, op het werk of in het dagelijks leven. Daarbij zul je ook zien dat je met een kritische houding valse presentaties van resultaten kunt doorprikken.

De *Random House College Dictionary* definieert **statistiek** als 'de wetenschap die zich bezighoudt met het verzamelen, classificeren, analyseren en interpreteren van informatie of gegevens'. Een statisticus is dus niet alleen maar iemand die werkloosheidscijfers berekent of de resultaten van een enquête in een tabel zet. Professionele statistici hebben een opleiding in *de wetenschap statistiek*. Dat wil zeggen dat ze opgeleid zijn in het verzamelen van numerieke informatie, in het analyseren van deze informatie, en in het trekken van conclusies hieruit. Verder bepalen statistici welke informatie bij een gegeven probleem van belang is, en hoe betrouwbaar de conclusies zijn die uit een onderzoek voortkomen.

Statistiek is de wetenschap van **gegevens**. Zij omvat het verzamelen, classificeren, samenvatten, organiseren, analyseren en interpreteren van numerieke informatie.

In de volgende paragraaf zul je een aantal voorbeelden tegenkomen van statistische toepassingen waarbij het gaat om het nemen van beslissingen en het trekken van conclusies.

1.2 Soorten statistische toepassingen

Statistiek betekent voor de meeste mensen 'beschrijven met getallen'. Maandelijkse werkloosheidscijfers, het percentage mislukte levertransplantaties en het percentage leidinggevende vrouwen in een bepaalde bedrijvensector: dit zijn allemaal voorbeelden van een beschrijving van grote hoeveelheden gegevens. Vaak worden de gegevens geselecteerd uit een grotere verzameling waarvan we de kenmerken willen schatten. We noemen dit proces **steekproef trekken**. Zo zou je bijvoorbeeld de leeftijden kunnen registreren van bezoekers van webwinkels die geïnteresseerd zijn in een bepaald product. Hiermee kun je dan een schatting maken van de leeftijdsverdeling van *alle* potentiële bezoekers die in dat product zijn geïnteresseerd. Die schatting zou je kunnen gebruiken om de reclame van een webshop af te stemmen op de juiste leeftijdsgroep. Je ziet dat het bij statistiek om twee verschillende processen gaat: (1) het beschrijven van gegevensverzamelingen en (2) het trekken van conclusies (schattingen, beslissingen, voorspellingen enzovoort) op basis van een steekproef. Deze twee belangrijkste gebieden waarop de statistiek wordt toegepast, noemen we de **beschrijvende statistiek** en de **verklarende statistiek**.

De **beschrijvende statistiek** gebruikt numerieke en grafische methoden om patronen in een gegevensverzameling te ontdekken, om de informatie in een gegevensverzameling samen te vatten en om deze informatie op een overzichtelijke manier te presenteren.

De **verklarende statistiek** gebruikt steekproefgegevens voor het schatten, het nemen van beslissingen en het voorspellen. De verklarende statistiek wordt ook wel **inductieve statistiek** of **inferentiële statistiek** genoemd.

Hoewel we in de volgende hoofdstukken zowel beschrijvende als verklarende statistiek zullen behandelen, zal het accent in dit boek liggen op de **verklarende statistiek**. Laten we beginnen met het bespreken van een aantal onderzoeken waarin statistische methoden werden toegepast.

Onderzoek 1. Bestverkopende koekjes van scoutingmeisjes[1]

Sinds 1917 hebben Amerikaanse scoutingmeisjes dozen met koekjes verkocht. Momenteel zijn er acht soorten te koop: Thin Mints, Samoas, Tagalongs, Peanut Butter Patties, Do-si-dos, Caramel DeLites, Peanut Butter Sandwiches en Trefoils. Elk van de naar schatting 150 miljoen dozen van de door de scoutingmeisjes verkochte koekjes wordt geclassificeerd naar soort. De resultaten zijn samengevat in figuur 1.1. Uit de grafiek kun je duidelijk aflezen dat Thin Mints het beste wordt verkocht (25%), gevolgd door Samoas (19%) en Tagalongs (13%). Omdat de figuur de omzet van de verschillende soorten koekjes beschrijft, is de grafiek een voorbeeld van beschrijvende statistiek.

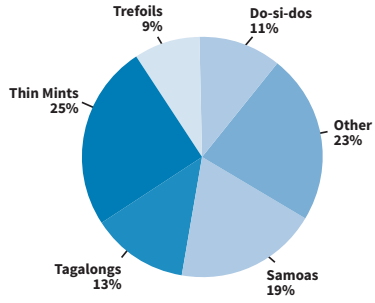
Onderzoek 2. Leidt het spelen van videospellen tot betere visuele waarneming?[2]

Onderzoekers van de Griffin Universiteit (Australië) wilden bepalen of spelers van videospellen beter visueel kunnen waarnemen dan niet-spelers van videospellen. Van een steekproef van 65 mannelijke psychologiestudenten werd elke proefpersoon geclassificeerd als een speler of een niet-speler van videospellen. Vervolgens moesten de twee groepen een aantal taken uitvoeren waaronder de

1 Bron: *www.girlscouts.org*

2 Bron: *Journal of Articles in Support of the Null Hypothesis,* Vol. 6, No. 1, 2009

'attentional blink'-test, de 'field of view'-test, en de 'repetition blindness'-test. Met uitzondering van de attentional blink werden er geen verschillen in prestatie tussen de twee groepen geconstateerd. Uit deze analyse concludeerden de onderzoekers dat 'het spelen van videospellen in beperkte mate onze visuele waarneming beinvloedt'. Hier werd dus verklarende statistiek toegepast.



Figuur 1.1: Grafiek van de omzet van koekjes van scoutingmeisjes (MINITAB)

Onderzoek 3. Diergondersteunde therapie voor ziekenhuispatiënten met hartfalen[3]

Een team van het UCLA Medical Center and School of Nursing voerde een onderzoek uit om na te gaan of diergondersteunde therapie de fysiologische reacties van patiënten met hartfalen kan verbeteren. Het team bestudeerde 76 hartpatiënten, willekeurig verdeeld in drie groepen. In de ene patiëntengroep werd elke persoon bezocht door een vrijwilliger vergezeld door een getrainde hond, in een andere groep werd elke persoon uitsluitend bezocht door een vrijwilliger, en de patiënten in de derde groep werden helemaal niet bezocht. De onderzoekers maten de fysiologische reacties van de patiënten (angstniveau, stress en bloeddruk) voor en na de bezoeken. Uit een analyse van de gegevens bleek dat de patiënten met diergondersteunde therapie een aanzienlijk grotere daling hadden in angstniveau, stress en bloeddruk. De onderzoekers concludeerden dat diertherapie een effectieve behandeling kan zijn voor ziekenhuispatiënten met hartfalen. Net als onderzoek 2 is dit onderzoek een voorbeeld van verklarende statistiek. De medische onderzoekers gebruikten gegevens van 76 patiënten om conclusies te trekken over de effectiviteit van diertherapie voor alle hartpatiënten.

Deze onderzoeken bieden drie praktijkvoorbeelden van toepassingen van de statistiek. Je ziet dat de gegevensverzameling wordt beschreven (onderzoek 1), of dat er conclusies over een gegevensverzameling worden getrokken (onderzoeken 2 en 3).

1.3 Basiselementen van de statistiek

Statistische methoden zijn in het bijzonder nuttig voor het bestuderen en analyseren van **populaties** bestaande uit **experimentele eenheden**.

Een **experimentele eenheid** is een object (bijvoorbeeld een persoon, een ding, een transactie of een gebeurtenis) waarvan we gegevens vastleggen.

3 Bron: *American Heart Association Conference, november 2005*

Een **populatie** is een verzameling eenheden (bijvoorbeeld personen, objecten, transacties of gebeurtenissen) die we willen bestuderen.

Voorbeelden van populaties zijn: (1) *alle* werknemers in Vlaanderen, (2) *alle* kiesgerechtigden in Nederland, (3) *alle* personen die een bepaald merk mobiele telefoon hebben gekocht, (4) *alle* auto's die het afgelopen jaar van een bepaalde lopende band zijn gerold, (5) de *totale* voorraad reserveonderdelen van de onderhoudsdienst van de spoorwegen, (6) *alle* dagomzetten van de 'drive-through'-afdeling van een McDonald's restaurant in een bepaald jaar, en (7) de verzameling van *alle* ongelukken op een bepaald stuk snelweg gedurende een vakantieperiode. De eerste drie voorbeelden van populaties zijn verzamelingen van personen; de volgende twee (4, 5) zijn verzamelingen van objecten, de volgende (6) is een verzameling van omzetten, en de laatste (7) is een verzameling gebeurtenissen. Je ziet ook dat elke verzameling alle eenheden van de desbetreffende populatie bevat.

Als we een populatie bestuderen, concentreren we ons op een of meer kenmerken of eigenschappen van de eenheden van die populatie. We noemen zulke kenmerken **variabelen**. Zo kunnen we bijvoorbeeld geïnteresseerd zijn in de variabelen leeftijd, geslacht, inkomen en/of het aantal jaren opleiding van de mensen die op dit moment werkloos zijn in de Europese Unie.

Een **variabele** is een kenmerk of eigenschap van een eenheid uit een populatie.

De term 'variabele' is afgeleid van het feit dat het kenmerk kan variëren over de verschillende eenheden in een populatie.

Vaak willen we de waarde van een variabele weergeven door middel van een getal. **Meten** is het proces waarbij we getallen toekennen aan variabelen. We kunnen bijvoorbeeld de voorkeur voor een voedselproduct meten door een consument te vragen een cijfer toe te kennen op een schaal van 1 tot 10 voor de smaak van het product. Of we zouden de leeftijd van de beroepsbevolking kunnen meten door aan elke werknemer te vragen hoe oud hij of zij is. In andere gevallen worden voor het meten instrumenten gebruikt, zoals stopwatches, weegschalen of schuifmaten.

Als de populatie die we willen bestuderen klein is, is het mogelijk om een variabele te meten voor elke eenheid in de populatie. Als we bijvoorbeeld het beginsalaris willen meten van iedereen die het afgelopen jaar een rechtenstudie heeft afgerond aan de Katholieke Universiteit Leuven, is het mogelijk om elk salaris te bepalen. Als we een variabele meten voor iedere eenheid van een populatie, is het resultaat een zogenaamde **census** van de populatie. In de meeste gevallen zullen de populaties waarin we geïnteresseerd zijn echter veel groter zijn, met wellicht vele duizenden of zelfs een oneindig aantal eenheden. Voorbeelden van grote populaties zijn de zeven populaties die we zojuist hebben genoemd. Het zou veel te veel tijd en/of geld kosten om een census te houden voor zulke populaties. Een alternatief is het selecteren en bestuderen van een deelverzameling van de eenheden van die populatie.

Een **steekproef** is een deelverzameling van de eenheden van een populatie.

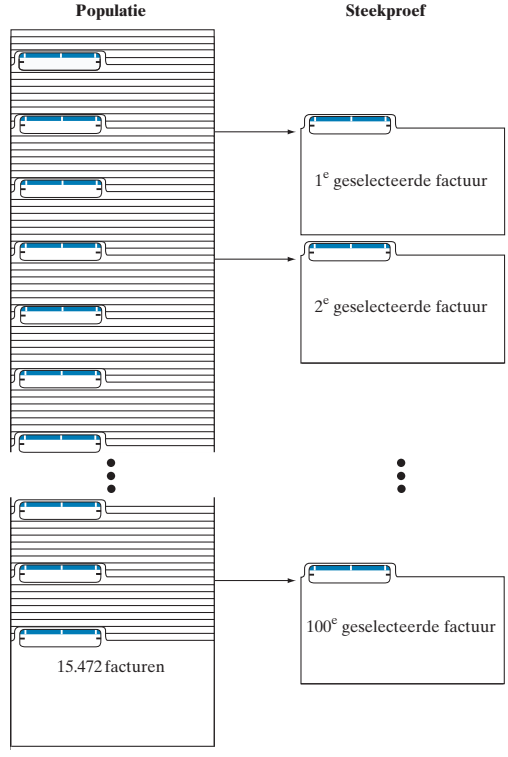
Veronderstel dat een boekhouding wordt gecontroleerd op fouten in de facturen. In plaats van het controleren van alle 15.472 facturen die in een bepaald jaar zijn verstuurd, kan een accountant een

steekproef nemen van slechts 100 facturen (zie figuur 1.2). Hij kan dan de status (fout of geen fout) van elk van de facturen van de steekproef vastleggen (meten).

Nadat de relevante variabelen voor elke eenheid in de steekproef (of populatie) zijn gemeten, worden de gegevens geanalyseerd, met beschrijvende of met verklarende statistische methoden. De accountant zou bijvoorbeeld alleen maar geïnteresseerd kunnen zijn in het beschrijven van het foutpercentage in de steekproef van 100 facturen. Maar waarschijnlijk wil hij de informatie van de steekproef gebruiken om conclusies te trekken over de populatie van alle 15.472 facturen.

Een **statistische conclusie** is een schatting, een voorspelling of een andere generalisatie voor een populatie die gebaseerd is op informatie uit een steekproef.

Dat wil zeggen dat we de informatie uit de steekproef gebruiken om iets te weten te komen over de hele populatie. Zo kan de accountant op basis van de steekproef van 100 facturen het aantal foute facturen in de populatie van 15.472 facturen schatten. De conclusie van de accountant over de kwaliteit van de facturen kan worden gebruikt om te beslissen of het factureersysteem van de firma moet worden veranderd.



Figuur 1.2: Een steekproef van alle facturen van de onderneming

Voorbeeld 1.1 ▶ Volgens het Amerikaanse blad *Variety* van 27 augustus 2009 is de leeftijd van kijkers van televisieprogramma's uitgezonden door CBS, NBC en ABC gemiddeld 51 jaar. Stel dat een leidinggevende van de concurrerende zender Fox veronderstelt dat de gemiddelde leeftijd van Fox-kijkers lager ligt dan 51. Om haar hypothese te toetsen neemt ze een steekproef van 200 Fox-kijkers en bepaalt van elk ervan de leeftijd.

a. Beschrijf de populatie.
b. Beschrijf de variabele waar het om gaat.
c. Beschrijf de steekproef.
d. Beschrijf de conclusie.

Uitwerking

a. De populatie is de verzameling van eenheden waar het de tv-leidinggevende om gaat, namelijk de verzameling van alle Fox-kijkers.
b. De leeftijd (in jaren) van elke kijker is de variabele waar het om gaat.
c. De steekproef moet een deelverzameling zijn van de populatie. In dit geval zijn het de 200 Fox-kijkers die door de leidinggevende geselecteerd zijn.
d. De conclusie zal de informatie uit de steekproef van 200 Fox-kijkers generaliseren naar de populatie van alle Fox-kijkers. De leidinggevende zou de gemiddelde leeftijd van de steekproef kunnen berekenen en daarmee de gemiddelde leeftijd van de populatie kunnen schatten.

Opmerking

Vormen de verzamelde gegevens (in dit voorbeeld de leeftijden van 200 Fox-kijkers) een populatie of een steekproef? Dat is van essentieel belang als je een statistisch probleem wilt aanpakken. ◀

Voorbeeld 1.2 ▶ Er heerst een felle concurrentie tussen de merken Coca-Cola en Pepsi. In hun reclamecampagnes wordt gebruikgemaakt van film- en televisiesterren, rockvideo's, sport en beweringen over de voorkeur van de consument, gebaseerd op smaaktests. Veronderstel dat als onderdeel van een Pepsi-marketingcampagne 1000 coladrinkers een blinde smaaktest doen met twee merken (A en B genoemd). Elke proever wordt gevraagd of hij of zij de voorkeur geeft aan merk A of aan merk B.

a. Beschrijf de populatie.
b. Beschrijf de relevante variabele.
c. Beschrijf de steekproef.
d. Beschrijf de conclusie.

Uitwerking

a. De relevante populatie is de verzameling van alle coladrinkers.
b. De relevante variabele is het kenmerk dat Pepsi wil meten, namelijk de voorkeur van de gebruiker voor een bepaald merk cola.
c. De steekproef bestaat uit de 1000 coladrinkers die uit de populatie van alle coladrinkers worden geselecteerd.
d. De conclusie zal de colavoorkeur in de steekproef van 1000 gebruikers generaliseren naar de populatie van alle coladrinkers. Op basis van de voorkeur in de steekproef kan men schatten welk percentage van alle coladrinkers een voorkeur heeft voor een bepaald merk. ◀

De voorafgaande definities en de voorbeelden laten vier van de vijf elementen van een probleem in de verklarende statistiek zien: een populatie, een of meer relevante variabelen, een steekproef en een

statistische conclusie. Maar met een conclusie alleen zijn we er nog niet. We willen ook de **betrouwbaarheid** weten – met andere woorden: we willen weten hoe goed de conclusie is. Betrouwbaarheid is dus het vijfde element dat bij een probleem uit de verklarende statistiek hoort.

Een **betrouwbaarheidsmaat** is een kwantitatieve uitspraak over de mate van onzekerheid die bij een statistische conclusie hoort.

We besluiten deze paragraaf met een samenvatting van de elementen van de beschrijvende en van de verklarende statistiek en een voorbeeld van een betrouwbaarheidsmaat.

Vier elementen van de beschrijvende statistiek

1. De populatie (of steekproef)
2. Een of meer variabelen
3. Tabellen, grafieken of numerieke hulpmiddelen om een samenvatting te geven
4. Vermelding van de patronen die in de samenvattingen naar voren komen

Vijf elementen van de verklarende statistiek

1. De populatie
2. Een of meer variabelen
3. De steekproef
4. De conclusie over de populatie, gebaseerd op informatie in de steekproef
5. Een betrouwbaarheidsmaat voor de conclusie

Voorbeeld 1.3 ▶ We keren terug naar voorbeeld 1.2, waarin de voorkeur voor een bepaald merk cola werd gemeten met een smaaktest. Hoe zou de betrouwbaarheid van een conclusie over de voorkeur van alle coladrinkers kunnen worden gemeten?

Uitwerking

Het percentage in de steekproef van coladrinkers dat een voorkeur heeft voor Pepsi zal niet precies overeenkomen met het percentage met een voorkeur voor Pepsi in de gehele populatie. Als bijvoorbeeld 56% van de 1000 gebruikers de voorkeur geeft aan Pepsi, dan zal niet precies 56% van alle coladrinkers de voorkeur geven aan Pepsi. Niettemin kunnen we met statistische methoden grenzen afleiden waartussen het werkelijke percentage voor de populatie zich vrijwel zeker zal bevinden. In dit geval zouden die grenzen 51% en 61% kunnen zijn. Het interval (0,51; 0,61) noemen we dan de uitkomst van een betrouwbaarheidsinterval. Meer hierover in hoofdstuk 7. ◀

Statistiek in de praktijk (vervolg)

Vaststellen van de populatie, de steekproef en de conclusie

Beschouw uit de PIALP-enquête uit 2010 in het bijzonder de vraag of men sociale-netwerksites als Facebook gebruikte. De experimentele eenheid is dan de persoon die de vraag beantwoordt, en de gemeten variabele is het antwoord ('ja' of 'nee').

Aan het PIALP-onderzoek namen circa 800 tieners deel. Natuurlijk omvat dat aantal niet alle tieners in de Verenigde Staten. Daarom zijn de 800 antwoorden een steekproef uit de denkbeeldige antwoorden van de veel grotere populatie van alle Amerikaanse tieners.

Eerdere enquêtes wezen uit dat in 2006 en in 2008 respectievelijk 55% en 65% van de Amerikaanse tieners een sociale-netwerksite gebruikte. Dit zijn beschrijvende samenvattingen die informatie geven over de populariteit van sociale netwerken in de afgelopen jaren. Omdat 73% van de geënquêteerde tieners in 2010 een sociale-netwerksite gebruikte, concludeerde het Pew Research Center dat steeds meer tieners elk jaar sociale-netwerksites bezoeken. Dat wil zeggen, de onderzoekers gebruikten de beschrijvende statistieken uit de steekproef om een conclusie te trekken over het gebruik van sociale netwerken door de huidige populatie van Amerikaanse tieners.

1.4 Soorten gegevens

Je hebt nu geleerd dat statistiek betrekking heeft op gegevens, en dat gegevens worden verkregen door de waarden van een of meer variabelen te meten van de eenheden in de steekproef (of populatie). Gegevens kunnen in twee algemene categorieën worden ingedeeld: **kwantitatieve gegevens** en **kwalitatieve gegevens**.

Kwantitatieve gegevens zijn gegevens die worden geregistreerd op een van nature voorkomende numerieke schaal.[4] Voorbeelden van kwantitatieve gegevens zijn:

1. De temperatuur (in graden Celsius) waarop een eenheid uit een steekproef van 20 stukjes hittebestendig plastic begint te smelten.
2. Het huidige werkloosheidspercentage voor elk van de landen van de EU.
3. De scores van een steekproef van 150 MBA-kandidaten op de GMAT, een gestandaardiseerde toets voor toelating tot een business graduate school in de Verenigde Staten.
4. Het aantal vrouwelijke leidinggevenden in elke onderneming in een steekproef van 75 productiebedrijven.

Kwantitatieve gegevens zijn meetwaarden die worden geregistreerd op een van nature voorkomende numerieke schaal.

4 Kwantitatieve gegevens kunnen we onderscheiden in gegevens op intervalschaal en op ratioschaal. Bij een ratioschaal is er sprake van een logisch, natuurlijk nulpunt: 0 minuten, 0 kilometer, 0 gram, 0 personen. Bij een intervalschaal is het nulpunt willekeurig gekozen: 0 graden temperatuur volgens Celsius of Fahrenheit, het jaar 0, het tijdstip 0:00 op de klok. Bij beide schalen mogen gegevens worden opgeteld en afgetrokken. Vermenigvuldigen en delen is echter alleen maar zinvol toepasbaar bij de ratioschaal. Een temperatuur van 20 graden is niet tweemaal zo hoog als een temperatuur van 10 graden. 4 uur op de klok is niet tweemaal zo laat als 2 uur op de klok, maar wel: 4 uur wachten is tweemaal zo lang als 2 uur wachten en 4 kg is tweemaal zoveel als 2 kg.