Statistiek om mee te werken

Noordhoff Uitgevers

Arie Buijs

10e druk

Statistiek om mee te werken

Prof.dr. A. Buijs

Tiende druk

Noordhoff Uitgevers, Groningen/Utrecht

Ontwerp omslag: G2K (Groningen-Amsterdam)
Omslagillustratie: Unsplash/Vladimir Kudinov

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

0 / 17

ISBN 978-90-01-87717-0
NUR 916

Woord vooraf

De rol van wiskunde en statistiek bij diverse opleidingen is dikwijls onderwerp van discussie. Zo zijn er studenten die zich afvragen waarom zij bij hun opleiding allerlei cursussen op het gebied van statistiek en onderzoeksmethoden moeten volgen. Ook docenten en opleidingsmanagers zoeken vaak naar een evenwicht tussen de vakdisciplines van een opleiding en wat men noemt de ondersteunende vakgebieden. Dat alles heeft bij sommige studierichtingen ertoe geleid dat ter bevordering van de slagingspercentages bepaalde moeilijke vakken een minder prominente plaats kregen in het curriculum. De laatste tien jaar is er een duidelijke kentering gekomen in die trend. Met name hebben kritische rapportages over de kwaliteit van afstudeerprojecten ertoe geleid dat een herbezinning heeft plaatsgevonden op vakgebieden die relevant zijn voor onderzoekvaardigheden. Het beheersen van een aantal onderwerpen uit wiskunde en statistiek is essentieel bij tal van opleidingen. Van bedrijfskunde tot biologie, van planologie tot informatica, van scheikunde tot psychologie: overal worden kwantitatieve methoden toegepast. Met dit boek willen we een bijdrage leveren aan de ontwikkeling van vaardigheden op het gebied van de statistiek.

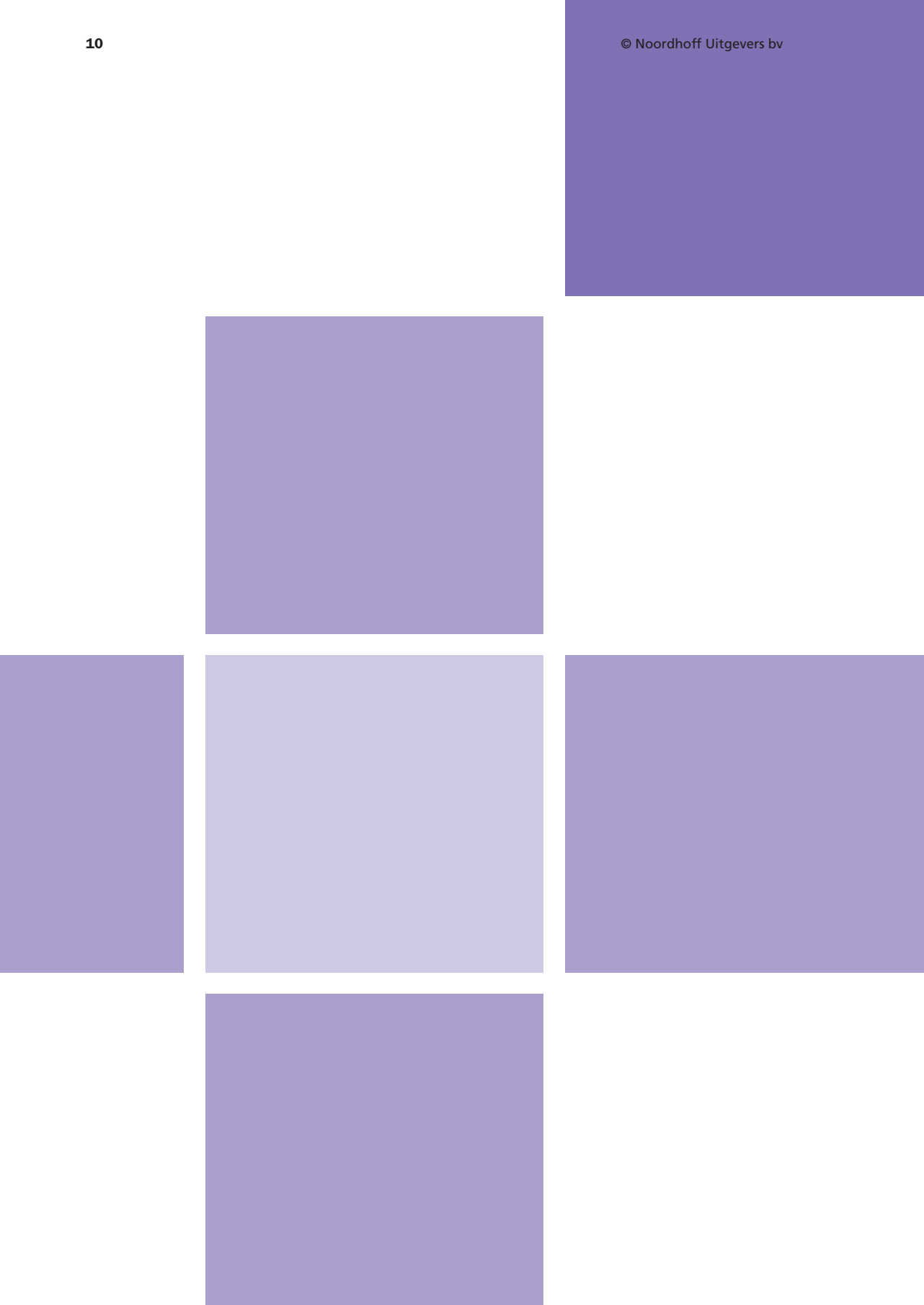
Deze tiende editie van *Statistiek om mee te werken* kent dezelfde indeling in hoofdstukken als de vorige editie. Enkele nieuwe onderwerpen werden toegevoegd, zoals de *normal plot*, uitbreidingen van de Chikwadraattoets, enkele opmerkingen over Big Data, relatief risico en odds, en een toets voor de correlatiecoëfficiënt. Soms werd ook een deel van een tekst herschreven en enkele andere hoofdstukken werden nauwelijks gewijzigd.
Dit boek kan voor vrijwel alle opleidingen aan universiteit en hbo dienen als handboek. Niet alleen omdat het pakket behandelde onderwerpen normaal gesproken voldoende moet zijn voor de statistiekcursussen bij de meeste opleidingen. Ook als naslagwerk zal menigeen hierin de nodige informatie kunnen vinden.
Naast dit theorieboek is er een opgavenboek (met uitwerkingen) beschikbaar. Hierin werden diverse nieuwe opgaven opgenomen.
Tevens is er een selectie van de uitwerkingen van deze opgaven in het opgavenboek opgenomen. De overige uitwerkingen worden via een voucher bij het leerboek beschikbaar gesteld en staan op de bij het boek horende website www.statistiekommeetewerken.noordhoff.nl. Op deze website zijn voor studenten ook diverse databestanden en een lijst met de belangrijkste formules uit het boek beschikbaar. Voor docenten is daar bovenop nog beschikbaar: aanvullend materiaal, extra opgaven en alle uitwerkingen van opgaven uit het opgavenboek.
Ook bij deze editie wil ik weer graag mijn dank uitspreken voor de waardevolle opmerkingen die ik mocht ontvangen van diverse collega's.

Met name ir. Koen de Bont en mr. drs. Jan W. Wijbenga hebben mij in de loop van de jaren veelvuldig geholpen. Ook de uitleg over het gebruik van Excel van drs. Johan Smits vormt een waardevolle bijdrage tot het kunnen toepassen van de statistische technieken. Langer geleden hebben reeds drs. Hans Buhrman en Lea Mol hun bijdrage geleverd, waarvoor mijn dank.
Gaarne nodig ik collega's en studenten uit hun opmerkingen over dit boek aan mij door te geven. En bovenal hoop ik dat het bestuderen van de onderwerpen uit dit boek niet alleen nuttig wordt bevonden, maar ook leuk.

Bilthoven, voorjaar 2017

Arie Buijs

Inhoud



1

Inleiding tot de beschrijvende statistiek

In de hedendaagse samenleving kan men – mede door internet – beschikken over een bijna onuitputtelijke hoeveelheid informatie over allerlei onderwerpen. Om vervolgens deze informatie nuttig te kunnen gebruiken is het vaak nodig dat deze geordend, gepresenteerd en geanalyseerd wordt. We hoeven hierbij maar te denken aan opinieonderzoek, het berekenen van prijsindexcijfers, het toetsen of een nieuw medicijn beter is dan het vorige en het berekenen van verzekeringspremies op basis van ongevallenstatistieken.
Het vakgebied van de statistiek kan voor dit doel een groot aantal hulpmiddelen aanreiken.
In dit hoofdstuk zullen we eerst aandacht besteden aan een aantal algemene aspecten van statistisch onderzoek. Vervolgens bespreken we in paragraaf 1.2 het werken met frequentieverdelingen en laten we in de paragrafen 1.3 tot en met 1.6 een aantal grafische voorstellingen de revue passeren. Tot slot wordt in paragraaf 1.7 een aantal gebruiksmogelijkheden van Excel besproken.
In sommige gevallen zullen methoden worden verduidelijkt aan de hand van de openingscasus.

Woningaanbod van de gezamenlijke makelaars

1

Een pas afgestudeerde bedrijfskundige heeft een eerste baan gevonden bij een groot makelaarskantoor en krijgt als eerste opdracht het woningaanbod van de gezamenlijke makelaars in kaart te brengen. Een uitgebreide inventarisatie levert hem een bestand op van 120 woningen. Van iedere woning is een aantal kenmerken gegeven, zoals de vraagprijs, de woonwijk, het aantal kamers en het bouwjaar.
Zo'n bestand roept allerlei vragen op, zoals:

- Zijn er opvallende prijsverschillen tussen de wijken?
- Hebben oude huizen meer kamers dan nieuwe huizen?
- Hebben huizen met garage meestal een badkamer met ligbad?
- Hoeveel is een huis gemiddeld duurder als het een garage heeft?

En zo kunnen we nog heel lang doorgaan. Om dergelijke vragen te kunnen beantwoorden is het in ieder geval noodzakelijk dat we de beschikbare gegevens op een overzichtelijke manier ordenen. We doen dat door er een gegevensbestand van te maken. Vervolgens kan de computer een handig hulpmiddel zijn om allerlei bewerkingen op deze gegevens toe te passen. Daarvoor zijn allerlei softwarepakketten beschikbaar. In dit boek zullen we ons met name richten op het gebruik van Excel, omdat dit programma op heel veel computers beschikbaar is.

1.1 Een eerste verkenning

In het spraakgebruik kan men het woord 'statistiek' in twee betekenissen tegenkomen. Een eerste betekenis van het woord statistiek heeft betrekking op het *resultaat* van een onderzoek: 'de statistieken' vertellen ons iets over een bepaald verschijnsel. Zo'n resultaat is bijvoorbeeld weergegeven in een tabel of een grafiek. We komen statistieken tegen in de vorm van tellingen die meestal worden opgesteld door een organisatie die dat soort werk doet in opdracht van de overheid, zoals de telling van het aantal verkeersslachtoffers in het kalenderjaar 2017. Zelfs zien we in sommige sportuitzendingen onder de aanduiding 'statistics' allerlei overzichten van prestaties van atleten verschijnen.
In de tweede plaats is statistiek de naam van een vakgebied. Dit vakgebied omvat het verzamelen, ordenen, samenvatten en analyseren van gegevens. Het is de bedoeling dat een aantal aspecten van dit vakgebied aan de orde komt in deze paragraaf.

1.1.1 Drie deelgebieden

Binnen het vakgebied van de statistiek wordt onderscheid gemaakt tussen drie deelgebieden. Dat zijn:
- beschrijvende statistiek
- kansrekening
- wiskundige statistiek

Beschrijvende statistiek

Bij het eerste gebied, de *beschrijvende statistiek*, houden we ons bezig met het verzamelen en verwerken van gegevens. De bedoeling is dat op een (soms grote) hoeveelheid waarnemingsuitkomsten een aantal bewerkingen wordt toegepast zodat de resultaten overzichtelijk worden voor de gebruikers van die gegevens. Hierbij denken we onder meer aan:
- het maken van een tabel of grafiek
- het berekenen van een gemiddelde waarde van de uitkomsten
- het aangeven met spreidingsmaatstaven in welke mate de gegevens onderling verschillen

Zo zal men na het verrichten van vijftig bepalingen bij een laboratoriumproef een beeld willen krijgen van het gemiddelde van de uitkomsten en willen weergeven of er veel onderlinge verschillen zijn bij de resultaten van de vijftig bepalingen. Zoiets kan tot stand gebracht worden door enkele hiervoor geschikte maatstaven te berekenen en door een grafiek te tekenen van de resultaten.
Een ander voorbeeld zou een marktonderzoek kunnen zijn waarbij de voorkeuren van consumenten worden geregistreerd. Door het maken van allerlei tabellen kan men een indruk krijgen van de mening van de ondervraagde personen. Ook kunnen op die manier dwarsverbanden tussen kenmerken weergegeven worden, bijvoorbeeld door te bestuderen of mannen meer voorkeur hebben voor een bepaald merk auto dan vrouwen. Beschrijvende statistiek kunnen we daarom als volgt typeren: door meer overzicht ontstaat meer inzicht.

Kansrekening

Bij het tweede gebied, de *kansrekening*, vormt het opstellen van een theoretische redenering het uitgangspunt. Als we voor een verschijnsel, bijvoorbeeld de uitkomst van een variabele, in gedachten nemen hoe groot de kansen zijn op het waarnemen van een bepaalde uitkomst, dan kunnen

we *vooraf* uitspraken doen over de waarschijnlijkheid dat een experiment in de praktijk een bepaald resultaat laat zien.
Als we bijvoorbeeld bij een bepaald productieproces vooraf weten dat er een kans van 10% is dat een product wordt afgekeurd, dan kunnen we berekenen hoe groot de kans is om zes afgekeurde exemplaren aan te treffen als veertig willekeurig gekozen producten worden gekeurd. Basisprincipe bij kansrekening is dus dat er veronderstellingen worden geformuleerd op grond waarvan men vooraf (dus zonder het experiment uit te voeren) berekeningen kan verrichten.

Wiskundige, inferentiële of verklarende statistiek

Het derde gebied, de *wiskundige*, *inferentiële* of *verklarende statistiek*, vervult een brugfunctie tussen de beschrijvende statistiek en de kansrekening. Op basis van de resultaten van waarnemingen proberen we dan met methoden van schatten en toetsen te komen tot algemene uitspraken over het onderzochte verschijnsel.
Veronderstel bijvoorbeeld dat we bij het controleren van de kwaliteit van producten constateren dat van de 400 onderzochte exemplaren er 60 niet deugen. Wat kunnen we dan zeggen over het productieproces in het algemeen? Op basis van een waargenomen percentage slechte exemplaren (15%, 60 van de 400) kunnen we dan misschien aangeven dat het productieproces in het algemeen bijvoorbeeld tussen 12% en 18% slechte exemplaren voortbrengt. Hoe dit soort berekeningen moet worden verricht, zullen we later in dit boek tegenkomen.

1.1.2 Populatie en steekproef

Wie een bepaald verschijnsel wil bestuderen met behulp van statistische methoden zal duidelijk moeten maken op welke verzameling personen, objecten of elementen het onderzoek betrekking heeft. Zo'n verzameling noemen we de *populatie*.

Populatie

Voor de populatie is het van belang dat deze *operationeel* gedefinieerd is. Dit betekent dat men in de praktijk duidelijk moet kunnen bepalen of een bepaald element wel of niet tot de populatie moet worden gerekend. Dat zoiets lang niet altijd eenvoudig is, blijkt bijvoorbeeld uit de discussies die worden gevoerd over het tellen van het aantal werklozen in Nederland. Hierbij is het soms onduidelijk of iemand wel of niet tot deze populatie behoort. En denk ook eens aan de populatie van alle woningen in Nederland. Wanneer is een gebouw een woning bijvoorbeeld? En kunnen binnen een gebouw meerdere woningen worden onderscheiden? Het is daarom zaak om nauwkeurig af te spreken hoe je kunt vaststellen of een element wel of juist niet tot de beoogde populatie behoort. In dit hoofdstuk zullen we enkele malen spreken over een onderzoek naar de kenmerken van woonhuizen die te koop staan in een bepaalde gemeente. Hiervoor wordt het gezamenlijke aanbod van de plaatselijke makelaars als populatie gedefinieerd. Toch kun je je afvragen of er buiten de plaatselijke makelaars om ook woningen te koop staan, en of sommige huizen wellicht al verkocht zijn en daarom inmiddels niet meer tot de populatie behoren.
Populaties bestaan dus uit elementen. Soms spreken we van populatie-elementen, terwijl deze elementen nog niet eens bestaan. Als voorbeeld kan men denken aan de populatie van alle geproduceerde en nog te produceren apparaten in een fabriek. Door de kwaliteit te controleren van een aantal zojuist geproduceerde apparaten, proberen we uitspraken te doen over de kwaliteit van het productieproces, en dat betreft ook apparaten die in de toekomst nog vervaardigd moeten worden.

Bij de elementen van een populatie wordt doorgaans een (beperkt) aantal kenmerken onderzocht. Deze kenmerken zullen we vaak aanduiden met de term variabele.

Variabele

Bij veel onderzoekingen wordt niet de gehele populatie onderzocht, maar wordt bij een beperkt aantal elementen een waarneming gedaan van een bepaald kenmerk. We nemen dan een *steekproef* uit de populatie.

Steekproef

Om een selectie te maken van populatie-elementen die in de steekproef terechtkomen, zijn allerlei methoden van steekproeven trekken ontwikkeld. Belangrijk is dat ernaar gestreefd wordt dat de steekproef een goede weergave is van de populatie (representativiteit) en de onderzoeker geen subjectieve keuze mag maken bij de selectie van steekproefelementen. Als de onderzochte populatie-elementen volkomen willekeurig worden gekozen, bijvoorbeeld door loting, dan spreken we van een *aselecte steekproef* uit de populatie. Die aanduiding betekent letterlijk dat we niet-selectief geweest zijn bij het samenstellen van de steekproef, dus we hebben niet stiekem bepaalde elementen weggelaten of bepaalde andere er expres in opgenomen omdat dat ons goed zou uitkomen.

Representativiteit

Aselecte steekproef

Veronderstel dat de populatie van te koop aangeboden woningen uit 120 elementen bestaat, en je neemt een steekproef van 20 stuks. Wanneer zou je zo'n steekproef representatief kunnen noemen? Representativiteit betekent dat er zowel oudere als nieuwere huizen in de steekproef zitten, zowel grote als kleine, zowel dure als goedkope, met een redelijke spreiding over de verschillende wijken enzovoort. En dat allemaal naar evenredigheid van de samenstelling van de populatie. Probleem is echter dat je vaak de populatie niet goed kent als je een steekproef neemt, dus je tast enigszins in het duister. Daarom vertrouw je er maar op dat een aselecte steekproef nou juist zorgt – weliswaar op basis van toeval – voor een redelijke mate van representativiteit.

Een van de belangrijke verworvenheden van het vakgebied van de statistiek is dat men met de resultaten van het onderzoek van een *steekproef* bepaalde uitspraken kan doen, die geldig zijn voor de gehele *populatie*. Dat is dus het gebied van de wiskundige statistiek, waarover wij eerder spraken.

1.1.3 Gegevens verzamelen

Gegevens verzamelen

Er zijn tal van manieren te noemen waarop men gegevens kan verzamelen ten behoeve van een onderzoek. Veel hangt af van het soort onderzoek dat gedaan wordt. Zo kan men in een aantal gevallen profiteren van onderzoek dat reeds door anderen is gedaan. We kunnen bijvoorbeeld tal van CBS-statistieken raadplegen. We noemen dit het raadplegen van een *externe bron*, in dit voorbeeld het Centraal Bureau voor de Statistiek (CBS). Ook kan men binnen een organisatie soms een *interne bron* raadplegen. Zo zal een bedrijfsdirecteur die iets wil weten over de leeftijdsopbouw van het personeelsbestand wellicht deze gegevens onmiddellijk beschikbaar krijgen via de salarisadministratie of de afdeling personeelszaken.

We zullen ons echter vooral bezighouden met het verzamelen van gegevens zoals dat door de onderzoeker zelf gebeurt. We onderscheiden daarbij twee hoofdvormen, namelijk de enquête en het experiment.

Enquête

Bij toepassing van een *enquête* of survey maakt de onderzoeker gebruik van een vragenformulier dat aan een aantal proefpersonen of bedrijven wordt voorgelegd. Het maken van een goede vragenlijst is een kunst op zich. De vragen moeten goed aansluiten op de bedoeling van het onderzoek, de vragen moeten ondubbelzinnig zijn en er moet op worden gelet dat de gegeven

antwoorden op een verantwoorde wijze kunnen worden verwerkt. Vaak wordt eerst een proefenquête gehouden om de vragen te testen. Ook moet erop gelet worden dat er niet te veel vragen gesteld worden, omdat proefpersonen in dat geval wellicht minder snel aan de enquête zullen meewerken. Enquêtes worden zowel mondeling als schriftelijk afgenomen. Een belangrijk probleem kan de *non-respons* zijn, wat inhoudt dat benaderde personen weigeren mee te doen aan de enquête. Gevolg van een grote non-respons kan zijn dat de verzameling wel ingevulde formulieren niet meer beschouwd kan worden als een representatieve steekproef uit de populatie. Nadat de enquêteformulieren zijn terugontvangen, kan een begin gemaakt worden met de verwerking van de gegevens. Als het onderzoek van beperkte omvang is, kunnen de gegevens 'met de hand' verwerkt worden tot bijvoorbeeld tabellen. Bij een omvangrijk onderzoek wordt in het algemeen de computer ingeschakeld. De antwoorden op de verschillende vragen worden dan vaak eerst gecodeerd en daarna in een rechthoekig schema geplaatst, de zogenoemde datamatrix. Als de *datamatrix* eenmaal is ingevoerd in de computer, kan een begin worden gemaakt met het toepassen van statistische methoden op het databestand. Er kunnen dan bijvoorbeeld tabellen worden gemaakt voor de gemeten variabelen, er kunnen dwarsverbanden onderzocht worden tussen allerlei variabelen (het 'crossen' van variabelen) en er kunnen nog veel andere bewerkingen op de gegevens worden toegepast.

Non-respons

Datamatrix

VOORBEELD 1.1

In het opgavenboek is een bestand genaamd 'woningen' opgenomen, waarin van 120 huizen die te koop worden aangeboden een achttal kenmerken is vermeld. Deze kenmerken zijn als volgt weergegeven:

- X_1: wijk, er worden vier wijken onderscheiden (1, 2, 3 en 4)
- X_2: aantal kamers (exclusief keuken, hal, sanitaire ruimten)
- X_3: aantal badkamers met ligbad
- X_4: aantal m^2 grondoppervlak van het aangeboden perceel
- X_5: bouwjaar
- X_6: garage (nee = 0, ja = 1)
- X_7: cv (nee = 0, ja = 1)
- X_8: vraagprijs

Voor de eerste tien huizen levert dit de gegevens uit tabel 1.1 op.

TABEL 1.1 Kenmerken woningen

Huisnummer	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	1	3	0	120	1920	0	0	132 000
2	1	4	0	140	1932	0	0	137 500
3	3	4	0	110	1938	0	1	138 000
4	1	3	1	110	1927	0	0	139 500
5	1	3	0	140	1968	0	1	142 000
6	1	3	0	220	1935	0	0	144 000
7	3	3	1	190	1938	0	0	145 000
8	1	4	0	130	1950	0	0	145 000
9	1	4	0	120	1964	0	1	146 500
10	3	3	0	200	1966	0	1	148 000

In totaal zijn er 120 woningen in dit bestand opgenomen (zie opgavenboek hoofdstuk 17).

In tabel 1.1 geeft elke regel een waargenomen huis aan. De voorkolom geeft het nummer van de waarneming aan, terwijl de volgende acht kolommen de uitkomst van een variabele aangeven. Om deze getallen te kunnen begrijpen, is het soms nodig om een overzicht bij de hand te hebben waarin de betekenis van de gebruikte getallen (codes) wordt uitgelegd. Zoiets wordt wel een *codeboek* genoemd.

Codeboek

Het is een goede oefening om voor tabel 1.1 in voorbeeld 1.1 eens te bekijken wat voor het derde huis uit het bestand precies de betekenis is van de opgenomen waarden van de variabelen X_1 tot en met X_8.

Bij het opstellen van een datamatrix is het belangrijk dat ook *ontbrekende gegevens* kunnen worden aangegeven. Vaak gebruikt men hiervoor een code die als uitkomst van een variabele logischerwijs niet kan bestaan. Bijvoorbeeld: leeftijd van een persoon is –1. De computer kan dergelijke codes ook herkennen als ze als zodanig gedefinieerd zijn, zodat ontbrekende gegevens (*missing values*) niet worden meegerekend als bijvoorbeeld een gemiddelde van de waarnemingen moet worden bepaald.

Experiment

Het verzamelen van gegevens door middel van een *experiment* komt vooral voor bij wetenschappelijke toepassingen. Een belangrijk kenmerk van het experiment is dat de onderzoeker zelf een aantal condities creëert waaronder het experiment plaatsvindt. Zo zal bij landbouwkundig onderzoek naar het effect van meststoffen door de onderzoeker een aantal proefakkers ingezaaid worden met gebruik van verschillende hoeveelheden meststof, zodat het effect van deze behandelingen op bijvoorbeeld verschillende rassen van het gewas kan worden bestudeerd.

Een ander voorbeeld van een experiment komen we tegen als het effect van een medische behandeling wordt bestudeerd. We zien dan wel eens dat proefpersonen in twee groepen worden verdeeld, namelijk een groep die de behandeling wel krijgt en een tweede groep die dat niet krijgt (de *controlegroep*). Met de waargenomen gegevens van de twee groepen probeert men dan tot conclusies te komen.

1.1.4 Variabelen

In de statistiek speelt het begrip variabele een belangrijke rol. Wanneer bij een bepaald onderzoek waarnemingen gedaan worden dan zal men geïnteresseerd zijn in een bepaalde eigenschap of karakteristiek van de onderzochte objecten. Een dergelijke eigenschap noemt men een variabele. Afhankelijk van de eigenschap die onderzocht wordt, kan zo'n variabele uitkomsten opleveren die door getallen worden weergegeven. In dat geval kunnen we van een *kwantitatieve* variabele spreken.

Kwantitatieve variabele

Er zijn ook *kwalitatieve* variabelen. Hierbij is de uitkomst niet een getal maar een aanduiding, een kenmerk. Denk bijvoorbeeld aan de religieuze overtuiging van een persoon, de kleur van een auto of een oordeel – goed, matig, slecht – over een docent. Bij het woningenbestand van voorbeeld 1.1 is de wijk een voorbeeld van een kwalitatieve variabele.

Kwalitatieve variabele

Soms zijn de uitkomsten van een variabele uniek bepaald. Zo is in Nederland de hoeveelheid belasting die een persoon moet betalen volledig vastgelegd, zodra zijn belastbaar inkomen gegeven is. We noemen dit wel een *deterministische variabele*.

Deterministische variabele

Dikwijls zijn de uitkomsten van een variabele echter onzeker. Indien de waargenomen waarde van een variabele voortkomt uit een kansexperiment, dat wil zeggen dat de te verschijnen uitkomsten afhankelijk zijn van toeval, dan spreken we van een *kansvariabele*. Voorbeelden zijn de

Kansvariabele

uitkomst van een worp met een dobbelsteen, de leeftijd van een willekeurige persoon die we op straat aanspreken, het gewicht van een vis die we vangen, of het bedrag dat we winnen met een kraslot. In het vakgebied van de statistiek gaat de speciale aandacht uit naar die kansvariabelen. In de praktijk blijkt het nuttig om onderscheid te maken tussen twee typen variabelen, namelijk discrete variabelen en continue variabelen.

Discrete variabele

Een *discrete* variabele (we zullen hiervoor steeds het symbool k gebruiken) is gekenmerkt door de eigenschap dat hij een *eindig* of een *aftelbaar oneindig* aantal verschillende waarden kan aannemen. Een voorbeeld van een eindig aantal verschillende uitkomsten is het resultaat van een worp met een dobbelsteen (een dobbelsteen kan slechts de uitkomsten 1, 2, 3, 4, 5 en 6 tonen). Het aantal aan te nemen getalwaarden is voor dit experiment eindig (namelijk zes mogelijkheden).

Een aftelbaar oneindig aantal mogelijke uitkomsten hebben we bij een experiment waarbij een muntstuk net zolang wordt opgegooid totdat de uitkomst 'kop' verschijnt. Het is hierbij mogelijk dat bij de eerste worp reeds 'kop' verschijnt, maar in theorie kan het oneindig veel pogingen kosten voordat het muntstuk 'kop' laat zien. Als we de variabele k definiëren als het aantal pogingen dat ondernomen moet worden voordat de uitkomst 'kop' verschijnt, dan is het waardebereik van k gegeven door de verzameling {1, 2, 3, ...}. De natuurlijke getallen dus.

Continue variabele

Een variabele wordt *continu* genoemd indien deze ook allerlei tussenliggende waarden kan aannemen. Bekende voorbeelden van continue variabelen zijn *tijd* (bijvoorbeeld de wachttijd tot een bepaalde gebeurtenis optreedt), *afstand* en *gewicht*. Een continue variabele duiden we aan met x.

Het verschil tussen discrete en continue variabelen is in de praktijk kleiner dan in theorie. Het gewicht van een individu is een voorbeeld van een continue variabele. Indien we beschikken over zeer nauwkeurige meetapparatuur, dan zouden wij – in theorie – iemand met een gewicht van 65,9437162. . . kg kunnen aantreffen. In zulke gevallen kan de variabele een *overaftelbaar oneindig* aantal verschillende waarden aannemen. Omdat in de praktijk een dergelijke nauwkeurigheid in het algemeen niet nodig is en omdat alle meetinstrumenten een eindige nauwkeurigheid hebben, wordt er altijd gewerkt met afgeronde getallen. Hierdoor worden continue variabelen in feite omgezet in discrete variabelen.

1.1.5 Indeling in schalen

Schalen

Om te kunnen vaststellen welke rekenkundige bewerkingen zijn toegestaan voor een bepaalde variabele, maken we een indeling in schalen. We onderscheiden vier typen schalen.

Nominale schaal

Bij de nominale schaal meet de variabele een kenmerk dat niet op een voor de hand liggende manier in een getal kan worden weergegeven, bijvoorbeeld kleur, merk, godsdienst of de naam van de krant die iemand leest.

De codenummers zijn doorgaans beter hanteerbaar dan de namen van de wijken als de gegevens in de vorm van een datamatrix in de computer worden opgeslagen. Uiteraard dient de *betekenis* van de codenummers in een codeboek vermeld te worden, zodat de betekenis van zo'n getal altijd kan worden opgezocht.

VOORBEELD 1.2

Bij een onderzoek naar de vraagprijzen van woningen in een stad is een overzicht gemaakt van alle 120 huizen die op een bepaald moment te koop worden aangeboden.
De huizen staan in vier verschillende wijken. Na het verzamelen en tellen van de gegevens zou tabel 1.2 kunnen ontstaan.

TABEL 1.2 Locaties van woningen

Wijk	Codenummer	Aantal
Tuinwijk	1	28
Binnenstad	2	29
Overmaas	3	37
Julianapark	4	26
Totaal		120

Kenmerk van de nominale schaal is dat de getalwaarden geen logische volgorde (ordening) kennen. In ons voorbeeld hadden we de codenummers evengoed in een andere volgorde kunnen toekennen. Het spreekt vanzelf dat we met nominale variabelen geen rekenkundige operaties kunnen uitvoeren. Bewerkingen zoals gemiddelden uitrekenen zijn zinloos. In voorbeeld 1.2 kunnen we niet 'de gemiddelde wijk' berekenen. Zo'n getal zou zonder betekenis zijn.

Ordinale schaal

Wanneer er wel sprake is van een logische volgorde, dan spreken we van een *ordinale* schaal. Een voorbeeld hiervan is de classificatie van restaurants in de Michelin-gids. Hierbij bestaan vier categorieën, die luiden: 3 sterren (uitzonderlijk goed), 2 sterren (zeer goed), 1 ster (goed) en 0 sterren (variërend van slecht tot redelijk). Het aantal sterren kan beschouwd worden als een ordinale schaal, want er is een logische volgorde. Anderzijds is deze variabele niet geschikt voor rekenkundige bewerkingen. Het is immers in het geheel niet zeker dat de onderlinge verschillen dezelfde betekenissen hebben. Een ander bekend voorbeeld van een ordinale schaal vinden we bij enquêtes waarbij vragen beantwoord moeten worden door een van de volgende vijf antwoordmogelijkheden aan te kruisen: 1 = zeer goed, 2 = goed, 3 = matig, 4 = vrij slecht en 5 = zeer slecht. Ook dan is sprake van een logische volgorde, maar is het onduidelijk of het verschil tussen antwoord 1 en 2 hetzelfde is als tussen antwoord 3 en 4.
Bij schalen van dit type is het daarom eigenlijk niet toegestaan om grootheden zoals een rekenkundig gemiddelde uit te rekenen. Overigens wordt hiertegen in de praktijk nogal eens gezondigd, omdat men toch wil kunnen aangeven dat bijvoorbeeld op de ene vraag veel hoger wordt gescoord dan op de andere.

Intervalschaal

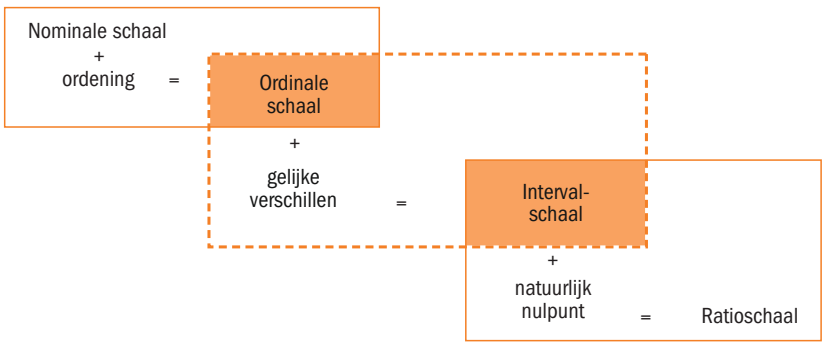
Wanneer het verschil tussen twee uitkomsten een eenduidige betekenis heeft, spreken we van een *intervalschaal*.
Bekende voorbeelden hiervan zijn de temperatuur die we aflezen op een thermometer en de tijd die we aflezen op een klok. Het tijdsinterval tussen

3:00 uur en 5:00 uur is even groot als het tijdsinterval tussen 19:00 uur een 21:00 uur. De verschillen hebben dus dezelfde betekenis. Een intervalschaal kent echter geen natuurlijk nulpunt. Het tijdstip 0 uur op de klok is in feite willekeurig gekozen. We kunnen niet zoiets zeggen als: 'Om 4 uur is het twee keer zo laat als om 2 uur.' Ook bij een thermometer geldt iets dergelijks, omdat je kunt stellen dat het nulpunt door ene heer Celsius vrij willekeurig is gekozen als het vriespunt van water, in plaats van bijvoorbeeld het vriespunt van jonge jenever.
In het databestand 'woningen' van voorbeeld 1.1 kan men het *bouwjaar* als een intervalschaal aanduiden (waarom?).

Ratioschaal
Wanneer er bovendien sprake is van een natuurlijk nulpunt in de schaal, dan spreken we van een variabele met een *ratioschaal*.
Voorbeelden hiervan zijn het gewicht van een individu, de wachttijd in de wachtkamer bij een tandarts, het inkomen van een volwassen Nederlander of de afstand die gesprongen wordt door een skispringer. Bij een variabele van dit type kunnen allerlei rekenkundige bewerkingen op de gegevens worden toegepast.
In het databestand 'woningen' van voorbeeld 1.1 is de vraagprijs een voorbeeld van een variabele met een ratioschaal (waarom?).

Het schema in figuur 1.1 geeft de samenhang tussen de verschillende schaaltypen aan.

FIGUUR 1.1 Samenhang tussen schaaltypen



Variabelen met een ratioschaal komen we veel tegen in natuurwetenschappelijke toepassingen, terwijl nominale en ordinale schalen veel voorkomen in de gedragswetenschappen.

1.2 Frequentieverdelingen

Wanneer we een groot aantal gegevens hebben verzameld ten behoeve van een bepaald onderzoek dan is het doorgaans noodzakelijk dat deze gegevens nader bewerkt worden. Het op overzichtelijke wijze presenteren van deze gegevens is hierbij belangrijk. Om personen die niet betrokken zijn

geweest bij het onderzoek een idee te geven van de resultaten is het vaak nuttig om de gegevens te verwerken in een tabel of een grafiek. Op deze manier kan een zeker overzicht van de betrokken variabele worden verkregen. Het patroon van de gegevens komt dan tot uiting. In deze paragraaf besteden we aandacht aan het maken van een klassenindeling en het berekenen van absolute en relatieve frequenties. Ook gaan we in op kruistabellen en enkele formele voorschriften voor tabellen. Ten slotte bespreken we de zogenoemde mengverdelingen.

1.2.1 Maken van een klassenindeling

Voordat van een hoeveelheid 'losse' gegevens een tabel of grafiek vervaardigd kan worden, is het noodzakelijk deze gegevens te ordenen. Hierbij wordt de verzameling van mogelijke uitkomsten verdeeld in een aantal intervallen of groepen, die we klassen zullen noemen. Het in klassen verdelen van het totale bereik van de variabele noemt men het maken van een *klassenindeling*.

Klassen

VOORBEELD 1.3A

Voor een bestand van 120 woningen is gegeven hoeveel kamers iedere woning heeft. We zouden dit kunnen weergeven door de frequentieverdeling in tabel 1.3.

TABEL 1.3 Klassenindeling naar aantal kamers per woning

Klasse	Aantal kamers
1	3
2	4
3	5 en 6
4	7 en 8
5	9 en 10
6	11 en 12

We zien hier dat het niet noodzakelijk is dat klassen dezelfde breedte hebben. De eerste twee klassen bevatten slechts één waarde (drie kamers respectievelijk vier kamers) terwijl de overige klassen elk twee waarden bevatten. Dus in klasse 3 worden alle huizen ingedeeld met vijf of zes kamers. Klassen van ongelijke breedte komen we tegen wanneer de verdeling in een bepaald gebied weinig uitkomsten laat zien. Bij het maken van een correcte klassenindeling moeten we erop letten dat rekening gehouden wordt met alle mogelijke uitkomsten van de betrokken variabele. Voor elke uitkomst moet een plaats zijn. Anderzijds moet ervoor gewaakt worden, dat er geen overlappingen plaatsvinden waardoor een bepaalde uitkomst in meer dan één klasse thuishoort. Een klassenindeling die in alle mogelijkheden voorziet maar geen overlappingen kent, noemen we een *categorisch systeem*. Zodra er een deugdelijke indeling in klassen gemaakt is, kan er 'geturfd' worden. Op deze manier kan worden vastgesteld hoe vaak er een waarneming verricht is die behoort tot een bepaalde klasse. Het aantal waarnemingen in een bepaalde klasse noemt men de *frequentie*. De verdeling die aldus voor de klassen ontstaat noemt men een *frequentieverdeling*.

Categorisch systeem

Frequentie

Frequentieverdeling

VOORBEELD 1.3B

Voor de verdeling van het aantal kamers levert het turven ons de frequentieverdeling uit tabel 1.4.

TABEL 1.4 Turftabel aantal kamers

Klasse	Aantal kamers	Turven	Frequentie
1	3	𝍸 𝍸 𝍸 ////	19
2	4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	35
3	5 en 6	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	47
4	7 en 8	𝍸 𝍸 𝍸	15
5	9 en 10	///	3
6	11 en 12	/	1
Totaal			120

Bij het maken van een klassenindeling is het belangrijk nauwkeurig te zijn bij het aangeven van de klassengrenzen. Hierbij dient er geen misverstand mogelijk te zijn over de vraag hoe een uitkomst die samenvalt met een klassengrens moet worden ondergebracht.

VOORBEELD 1.4

Voor de 120 woningen uit het bestand willen we een klassenindeling naar bouwjaar maken. Allereerst moeten we vaststellen wat de hoogste en de laagste waarneming is. Dit blijken de bouwjaren 1910 en 1988 te zijn. Vervolgens kiezen we de klassengrenzen zodanig, dat alle waarnemingen onder te brengen zijn. Het lijkt logisch om hier klassen van 10 jaar breed te kiezen, want er moet altijd op worden gelet dat de grenzen een beetje 'mooi' uitkomen. Dat leidt tot de indeling in tabel 1.5.

TABEL 1.5 Klassenindeling naar bouwjaar

Klasse	Bouwjaar
1	1910 – < 1920
2	1920 – < 1930
3	1930 – < 1940
4	1940 – < 1950
5	1950 – < 1960
6	1960 – < 1970
7	1970 – < 1980
8	1980 – < 1990

Let op de notatie die gebruikt is bij het aangeven van de klassen. Door gebruik te maken van het <-teken is het duidelijk dat de bovengrens niet en de ondergrens wel tot de klasse behoort. Dus een huis dat gebouwd is in 1920 moet geteld worden in de klasse 1920 – < 1930 en niet in de klasse 1910 – < 1920.

Als we een collectie verzamelde gegevens willen weergeven door middel van een frequentieverdeling, dan moeten we zelf een keuze maken voor het aantal klassen dat we willen hebben. Een ruwe richtlijn hiervoor is dat het aantal klassen tussen de vijf en de twintig moet liggen. Hierbij geldt: hoe groter het aantal waarnemingen is, des te groter kan het aantal klassen zijn. Soms wordt ook $\sqrt{n}$ – waarbij n het totaal aantal waarnemingen is – als criterium voor het aantal klassen genoemd. Bij $n = 150$ waarnemingen kiezen we dan voor $\sqrt{150}$, dus ongeveer 12 klassen. Als bij een frequentieverdeling te weinig klassen gebruikt worden, is in het eindresultaat veel informatie niet zichtbaar. Als er te veel klassen zijn, kan de resulterende verdeling onoverzichtelijk worden.
In tabel 1.5 waren de klassen allemaal even breed gekozen. Dat is echter niet altijd wenselijk, met name niet, indien in een bepaald gebied van de verdeling erg weinig waarnemingen voorkomen. We illustreren dit met een voorbeeld.

VOORBEELD 1.5
Voor het woningenbestand maken we een frequentieverdeling van de variabele 'vraagprijs'. We kiezen de klassen 50 000 euro breed, te beginnen vanaf 100 000 euro. Het is duidelijk dat er veel meer huizen zijn met een relatief lage vraagprijs dan met een hoge. Om die reden is het wenselijk om bij de hoge prijzen bredere klassen te kiezen, want anders zouden we een aantal klassen krijgen met geen enkele waarneming. Vandaar dat vanaf een vraagprijs van 500 000 euro, de klassen 250 000 euro breed zijn. De klasseninde-ling die zo ontstaat is weergegeven in tabel 1.6.

TABEL 1.6 Frequentieverdeling van de vraagprijs

Klassengrens × € 1 000	Aantal
100 – < 150	11
150 – < 200	30
200 – < 250	13
250 – < 300	13
300 – < 400	17
400 – < 500	11
500 – < 750	19
750 – < 1 000	5
1 000 – < 1 250	1
Totaal	120

De aantallen in een tabel met ongelijke klassenbreedten kunnen aanleiding geven tot misverstanden. Zo zien we hier dat in de klasse 500 – < 750 maar liefst 19 waarnemingen zitten, terwijl bijvoorbeeld in de klasse 400 – < 500 'slechts' 11 waarnemingen geplaatst konden worden. Toch is de klasse 400 – < 500 eigenlijk 'drukker bezet', omdat de klasse 500 – < 750 tweeenhalf maal zo breed is.
Straks – bij het tekenen van histogrammen – zullen we zien hoe we met het begrip 'frequentiedichtheid' dergelijke verschillen tot uitdrukking kunnen brengen.

1.2.2 Relatieve frequenties

Absolute frequenties

In de voorgaande voorbeelden maakten we kennis met het begrip frequentieverdeling. Nadat een indeling in klassen tot stand gekomen is, kunnen de waargenomen uitkomsten geteld worden. Hierdoor ontstaan de *absolute frequenties*. De som van de frequenties is uiteraard gelijk aan het totaal aantal waarnemingen.

Relatieve frequenties

Wanneer we de frequentie per klasse delen door het totaal aantal waarnemingen, ontstaan *relatieve frequenties*. Relatieve frequenties kunnen van belang zijn bij het vergelijken van verschillende frequentieverdelingen.

VOORBEELD 1.6

Voor de huizenprijzen (zie voorbeeld 1.5) berekenen we de relatieve frequenties. We doen dit door de absolute frequenties voor iedere klasse te delen door 120. De resultaten zijn weergegeven in tabel 1.7.

TABEL 1.7 Relatieve frequenties van de huizenprijzen

Klassengrenzen × € 1 000	Aantal	Relatieve frequentie
100 – < 150	11	0,092
150 – < 200	30	0,250
200 – < 250	13	0,108
250 – < 300	13	0,108
300 – < 400	17	0,142
400 – < 500	11	0,092
500 – < 750	19	0,158
750 – < 1 000	5	0,042
1 000 – < 1 250	1	0,008
Totaal	120	1,000

Relatieve frequentieverdelingen zijn soms handig om twee verdelingen met elkaar te vergelijken. Met name als voor de twee verdelingen het absolute aantal waarnemingen verschilt, dan komt het verschil in opbouw van beide verdelingen op zo'n manier beter tot uiting. Merk op dat de som van de relatieve frequenties gelijk is aan 1.

Procentuele frequentieverdeling

Door alle relatieve frequenties met 100 te vermenigvuldigen, ontstaat de *procentuele* frequentieverdeling. Vaak wordt er gesproken over *fracties* in plaats van relatieve frequenties. Men ziet in tabel 1.7 dat 9,2% van de huizen zich in de laagste prijsklasse bevindt. We zeggen dan ook wel dat de huizen met een vraagprijs tussen 100 000 en 150 000 euro een fractie van 0,092 vormen van alle te koop aangeboden huizen.

1.2.3 Kruistabellen

Het komt regelmatig voor dat men in een tabel het gedrag van twee variabelen wil weergeven. Met name is het dan van belang dat een eventuele relatie tussen de variabelen tot uitdrukking kan worden gebracht. In het volgende voorbeeld laten we een zogeheten *kruistabel* zien, waarin de gegevens van het woningbestand aan de hand van twee variabelen in klassen zijn verdeeld.

De variabelen waarom het hier gaat zijn 'het niet of wel hebben van een garage' en 'de wijk' waarin de woning gelegen is.

VOORBEELD 1.7

De gegevens van het woningbestand leveren tabel 1.8 op.

TABEL 1.8 Kruistabel van het woningbestand

		Wijk				Totaal
		1	*2*	*3*	*4*	
Garage	(nee) 0	21 (31,8%)	20 (30,3%)	17 (25,8%)	8 (12,1%)	66
	(ja) 1	7 (13,0%)	9 (16,7%)	20 (37,0%)	18 (33,3%)	54
Totaal		28	29	37	26	120

Uit de tabel blijkt het volgende:

- In de totaalkolom is rechtstreeks af te lezen dat er 66 huizen zijn zonder, en 54 huizen met garage.
- In de totaalregel onderaan zien we de aantallen huizen per wijk. Kortom: de totaalkolom en de totaalregel geven uitsluitend informatie over één van de twee variabelen.
- In de acht cellen in het middendeel van de tabel staat weergegeven hoeveel huizen twee kenmerken *combineren*. Zo blijken er bijvoorbeeld 18 huizen met garage te koop te staan in wijk 4.
- In de tabel staan ook relatieve frequenties vermeld in de vorm van percentages. Deze hebben in dit geval betrekking op een tabel*regel*. Daaraan kunnen we bijvoorbeeld zien dat van alle te koop staande huizen zonder garage zich 31,8% in wijk 1 bevindt.

In de tweede tabel worden juist de relatieve frequenties, in de vorm van percentages, berekend per *kolom*. Dan krijgen we het beeld uit tabel 1.9.

TABEL 1.9 Kruistabel van het woningbestand

		Wijk				Totaal
		1	*2*	*3*	*4*	
Garage	(nee) 0	21 (75%)	20 (69,0%)	17 (45,9%)	8 (30,8%)	66
	(ja) 1	7 (25%)	9 (31,0%)	20 (54,1%)	18 (69,2%)	54
Totaal		28	29	37	26	120

Op deze manier zien we bijvoorbeeld dat van alle huizen die te koop staan in wijk 4, maar liefst 69,2% voorzien is van een garage.
Als men dergelijke berekeningen uitvoert met een statistisch computerpakket, dan wordt doorgaans als een keuze geboden of men een berekening wil van rij- en/of kolompercentages.

1.2.4 Een bijzonder geval van tabelanalyse: relatief risico

Relatief risico is een begrip dat nogal eens in de medische wetenschap wordt toegepast. Daarbij worden relatieve frequenties met elkaar vergeleken. Het gaat dan om twee populaties die op een bepaald kenmerk verschillen en waarvoor de kans op het krijgen van een bepaalde ziekte wordt vergeleken.

Denk bijvoorbeeld aan personen boven de zestig jaar die in de herfst desgewenst een griepprik kunnen halen bij de huisarts.

TABEL 1.10 Griep en griepprik

Behandeling	**Resultaat**		
	Griep	*Geen griep*	*Totaal*
Griepprik	15 (a)	135 (b)	150
Geen griepprik	24 (c)	60 (d)	84
Totaal	39	195	234

Bron: huisartsenpraktijk Van Kraaij, Gronsveld

Voor de mensen die een griepprik hebben gehaald is de relatieve frequentie om tóch griep te krijgen:

$$\frac{a}{a+b} = \frac{15}{15+135} = 0{,}10$$

Voor de mensen zonder griepprik is de relatieve frequentie

$$\frac{c}{c+d} = \frac{24}{24+60} = 0{,}28$$

Bij relatief risico gaat het om het quotiënt van deze twee cijfers:

$$RR = \frac{0{,}10}{0{,}28} = 0{,}36$$

Vaak wordt bij het relatieve risico de grootste uitkomst boven de deelstreep geplaatst. Dan zouden we hier krijgen:

$$RR = \frac{0{,}28}{0{,}10} = 2{,}8$$

We zeggen dan: mensen zonder griepprik hebben 2,8 keer grotere kans op griep dan mensen die wel een prik halen.
We zijn bij het relatief risico (*RR*) op zoek naar de verhouding tussen de relatieve frequenties van twee verschillende groepen. In feite baseren we onze conclusies op een steekproef. Dus op basis van de griepverschijnselen in deze groep van 234 personen concluderen we dat het nemen van een griepprik de kans op griep een factor 2,8 kleiner maakt.
In de krant wordt dan vaak de suggestie gewekt dat dit voor de hele Nederlandse bevolking zou moeten gelden. Die conclusie is een beetje kort door de bocht want dat bericht is dan gebaseerd op een steekproef van 'slechts' 234 personen.

1.2.5 Enkele formele voorschriften voor tabellen

Tabellen

Er zijn twee belangrijke mogelijkheden om gegevens te presenteren, namelijk met tabellen en grafieken. Zojuist zagen we bij de bespreking van frequentieverdelingen al enkele voorbeelden van tabellen.

Tabellen hebben als doel om op een overzichtelijke manier gegevens te presenteren, vaak ten behoeve van buitenstaanders die niet al te veel kennis hebben van de precieze gegevens die we willen presenteren. Het is daarom van belang om bij een tabel of een grafiek een aantal vermeldingen te doen, zodat zo'n buitenstaander makkelijk kan begrijpen wat de schrijver wil verduidelijken. Daarom formuleren we hier een aantal vereisten waaraan een tabel moet voldoen. Niet met het oogmerk dat altijd aan alle vereisten moet zijn voldaan, maar wel met het doel dat daarmee een aantal richtlijnen beschikbaar is om houvast te geven bij het opstellen van een correcte tabel.

Veld

Een tabel bestaat uit kolommen en regels. De doorsnijding van een kolom met een regel noemt men een *veld* of een cel. Een veld is daarmee een plaats in de tabel waarop een getal kan worden geplaatst. De belangrijkste richtlijnen voor een tabel zijn:

Richtlijnen voor een tabel

1 *Een opschrift*. Boven iedere tabel moet in het kort worden aangegeven wat erin vermeld is. Dit opschrift moet kort en bondig zijn.
2 *Kolomkoppen*. Boven de kolommen van de tabel moet uit een zeer korte aanduiding blijken wat in die kolommen is weergegeven.
3 *Een voorkolom*. In de voorkolom moet omschreven staan wat in de regels van de tabel is af te lezen.
4 *Logische volgorde*. Indien het mogelijk is, moet men de kolommen en de regels in een logische volgorde opstellen.
5 *Nummering*. Bij gecompliceerde tabellen is het nuttig om kolommen en regels te nummeren, zodat in de tekst een gemakkelijke verwijzing kan worden gemaakt naar een bepaald gedeelte van de tabel. Het is aan te bevelen om kolom- en regelnummers tussen haakjes te plaatsen, zodat men deze cijfers niet verwart met de eigenlijke gegevens in de tabel.
6 *Totalen*. Indien de getallen uit de tabel dit zinvol maken, dient men een kolom en/of een regel op te nemen met de totalen.
7 *Speciale aanduidingen*. Voor een aantal bijzondere gevallen moet bij de aanduiding van tabelwaarden gebruik worden gemaakt van de algemeen toegepaste afspraken hiervoor. We kennen de volgende tekens:

.	(punt):	het gegeven is onbekend
*	(ster):	het gegeven is voorlopig
×	(kruis):	het gegeven is geheim
	(blank):	hier kan logischerwijs geen gegeven voorkomen
–	(streepje):	het gegeven is precies gelijk aan nul
0 of 0,0:		het gegeven is na afronding nul (kleiner dan de halve eenheid die werd gebruikt)

Gebruik van de hier genoemde tekens komen we bijvoorbeeld tegen in publicaties van het CBS (Centraal Bureau voor de Statistiek in Nederland).

8 *Bronvermelding*. Indien de gegevens van de tabel uit een externe bron voortkomen, is het geven van een bronvermelding een vereiste. Ook als de gegevens binnen het bedrijf verzameld zijn, kan een bronvermelding nuttig zijn.

Een groot aantal van de hier genoemde richtlijnen is terug te vinden in tabel 1.11 van voorbeeld 1.8. Er is een opschrift aanwezig, er zijn kolomkoppen geplaatst, in de voorkolommen zijn de inkomensklassen aangegeven en de inkomensklassen zijn in logische volgorde geplaatst. Ook zien we

een nummering van de kolommen. Hiervan kan men gebruikmaken bij een verwijzing naar de tabel in de lopende tekst. Als in de tekst bijvoorbeeld iets wordt gezegd over de inkomensverdeling van de vrouwen, kan de aanduiding zijn: 'zie kolom 2 van tabel 1.11'. Verder zijn er een totaalregel en een totaalkolom aanwezig. Bij de inkomensklasse '3 000 en hoger' is in kolom 2 door middel van een streepje aangegeven dat er precies nul waarnemingen op dit veld worden geplaatst. Tot slot is door middel van een bronvermelding duidelijk gemaakt hoe en wanneer de gegevens zijn verzameld.
Het is belangrijk de hier geformuleerde richtlijnen in het oog te houden. Men kan er bepaalde slordigheden bij het opstellen van een tabel door voorkomen.

VOORBEELD 1.8
Voor 200 werknemers van een warenhuis is in tabel 1.11 een verdeling gemaakt naar geslacht en is een aantal inkomensklassen geformuleerd.
De tabel heeft twee ingangen: er is namelijk een verdeling naar man/vrouw en een verdeling naar inkomen.

TABEL 1.11 De 200 werknemers van warenhuis Steens, verdeeld naar inkomen en geslacht (ultimo 2018)

Bruto-maandinkomen (in euro)	Geslacht		Totaal
	man	*vrouw*	
	(1)	(2)	(3)
0 – < 1 000	15	8	23
1 000 – < 1 250	32	30	62
1 250 – < 1 500	30	25	55
1 500 – < 2 000	25	10	35
2 000 – < 3 000	12	7	19
3 000 en hoger	6	–	6
Totaal	120	80	200

Bron: Salarisadministratie warenhuis Steens, december 2018

1.2.6 Mengverdelingen

Het lijkt allemaal zo eenvoudig: je verzamelt gegevens, je sorteert deze gegevens en plaatst ze vervolgens in een frequentieverdeling. Daarmee ontstaat vanzelf een beeld van de variabele die we onderzoeken.
Er zijn echter diverse valkuilen en misverstanden die kunnen leiden tot onjuiste conclusies. Een typisch voorbeeld hiervan betreft de zogeheten *mengverdelingen*. In een dergelijk geval hebben we te maken met een variabele die zich vermoedelijk verschillend gedraagt in twee (of meer) subgroepen.
Het makkelijkst kun je dan denken aan een populatie die te verdelen is in mannen en vrouwen. Er kunnen aanzienlijke verschillen zijn tussen de frequentieverdelingen van allerlei kenmerken. Of je nou de lichaamsgewichten neemt van mannen en vrouwen, hun verwachte levensduur, hun sportprestaties of wat dan ook, als je de verzamelde gegevens gaat splitsen voor de subpopulaties mannen en vrouwen zal dikwijls een onderling

verschillende opbouw van de verzamelde gegevens ontstaan. (Zie in het opgavenboek hoofdstuk 1, opgave 1.18 over de beklimmingstijden voor de Alpe d'Huez.) Andersom geldt daarom dat we bij de interpretatie van een reeds beschikbare frequentieverdeling altijd alert moeten zijn op de mogelijkheid dat deze verdeling eigenlijk de samenstelling is van twee afzonderlijke verdelingen die op één hoop zijn gegooid.
Er kan van alles misgaan als we niet goed op dit soort aspecten letten. We lichten dat toe aan de hand van voorbeeld 1.9.

VOORBEELD 1.9
Een onderzoeker wil graag weten in welke mate werknemers in deeltijd werken, of juist voltijds in dienst zijn. Bij een ziekenhuis werd gekeken naar de werktijden van de artsen die aan het ziekenhuis zijn verbonden. Hierbij bleek dat bij dit ziekenhuis 100 artsen werken, namelijk 70 vrouwen en 30 mannen. Van alle 100 artsen is bekend hoe groot de omvang van hun dienstverband is. Voor de 70 vrouwen staat de verdeling in tabel 1.12.

TABEL 1.12 Vrouwen

Leeftijd	Deeltijd	Voltijds
Jonger dan 40	30	20
40 jaar en ouder	14	6

Uit tabel 1.12 kan men concluderen dat van de jongere vrouwelijke artsen 60% in deeltijd werkt. Van de oudere vrouwelijke artsen werkt 70% in deeltijd. Het werken in deeltijd neemt dus toe naarmate vrouwelijke artsen ouder zijn. Van 60% naar 70%.
Voor de 30 mannen staat de verdeling in tabel 1.13.

TABEL 1.13 Mannen

Leeftijd	Deeltijd	Voltijds
Jonger dan 40	0	10
40 jaar en ouder	4	16

In tabel 1.13 zien we dat van de jonge mannen 0% in deeltijd werkt. Voor de veertigplussers geldt dat 20% in deeltijd werkt. Dat is dus een toename van 0% naar 20%. Zowel voor de mannen als voor de vrouwen geldt dus dat het percentage deeltijdwerkers *toeneemt* met de leeftijd.
Men heeft ook een overzicht gemaakt voor alle 100 artsen tezamen. Het resultaat daarvan staat in tabel 1.14.

TABEL 1.14 Alle artsen

Leeftijd	Deeltijd	Voltijds
Jonger dan 40	30	30
40 jaar en ouder	18	22

Uit het totaaloverzicht in tabel 1.14 blijkt dat 50% van de jongeren in deeltijd werkt. Van de veertigplussers is dat slechts 45%. Dus we lezen in de krant: naarmate artsen ouder worden werken ze *minder*(!) in deeltijd.
Wat is hier aan de hand?
Eigenlijk is er voor wat betreft deeltijdwerk een verborgen variabele, dat is namelijk het kenmerk man/vrouw.
In de jongere categorie bleek bij dit ziekenhuis een enorme dominantie van de vrouwen (namelijk 50 van de 60). Bij de ouderen was dat anders, namelijk een fifty-fiftyverdeling man-vrouw.
Dat betekent dat in een gezamenlijke verdeling er een grote dominantie van vrouwen is bij de jongeren en een (relatief) grote dominantie van mannen bij de ouderen.
Bij de mannen zitten opvallend veel voltijdse banen. Dat effect had daarom veel meer gewicht bij de categorie ouderen.

De les is dat we bij het interpreteren van frequentieverdelingen (en daarvan afgeleid tabellen en grafieken) attent moeten zijn op de mogelijkheid dat er nog een verborgen factor is die aangeeft dat we de gegevens eigenlijk moeten splitsen in twee of meer groepen.

1.3 Grafische voorstellingen

We zagen dat tabellen een belangrijk hulpmiddel kunnen zijn om een onoverzichtelijke hoeveelheid gegevens te ordenen en toegankelijk te maken voor de lezer. Een soortgelijke functie heeft de *grafische voorstelling*. De lezer kan vaak in één oogopslag vaststellen wat de belangrijkste conclusies zijn die volgen uit de beschikbare gegevens als we het gedrag van de betrokken grootheden in tekening brengen. Daarnaast kan het bijdragen tot de verlevendiging van een rapport als we een aantal (zinvolle) grafische voorstellingen opnemen.
Bij het opstellen van een grafische voorstelling moet de nodige zorgvuldigheid in acht worden genomen. Als we bijvoorbeeld in een grafiek een onjuiste indeling van de assen maken, kan bij de lezer vrij gemakkelijk een verkeerd beeld van het weergegeven verschijnsel ontstaan.
In dit hoofdstuk zullen we een drietal typen grafische voorstellingen bespreken. Maar eerst gaan we in deze paragraaf in op de voorschriften bij het opstellen van een grafische voorstelling. Daarna bespreken we grafieken in een assenstelsel. In paragraaf 1.4 komen diverse diagrammen aan de orde en in paragraaf 1.5 bespreken we de grafische weergave van bepaalde frequentieverdelingen met behulp van onder andere histogrammen.

1.3.1 Enkele voorschriften

Richtlijnen voor een grafiek

Evenals bij het vervaardigen van tabellen, moet men bij het opstellen van een grafische voorstelling een aantal formele regels in acht nemen. Van belang zijn de volgende onderdelen:

1 *Opschrift*. Boven iedere grafiek moet in het kort worden vermeld, wat men met de grafiek wil weergeven. In beginsel mag er geen tekst komen in de grafiek.

2 *Assen*. Als een grafiek in een assenstelsel is gemaakt, moet langs de horizontale en de verticale as vermeld staan welke variabele door de as wordt weergegeven. Het noemen van de eenheid van telling is hierbij van groot belang.
3 *Teleenheden*. Op regelmatige plaatsen langs de assen dienen getallen te zijn geplaatst, zodat het mogelijk is in de grafiek een waarde af te lezen. We moeten hierbij echter niet overdrijven. Te veel getallen langs de assen maken een grafiek te druk.
4 *Nulpunten*. Bij grafieken in een assenstelsel is het snijpunt van de assen – de oorsprong – het punt waar zowel de horizontaal als de verticaal afgezette variabele de waarde 0 heeft. Als een variabele alleen waarden laat zien in een gebied dat vrij ver van het nulpunt ligt, moet men een onderbreking in de as aanbrengen. In de grafiek laat men dit blijken door het aangeven van een zogenoemde *scheurlijn*. Dat is een zigzaglijntje waaraan de lezer onmiddellijk kan zien dat een gedeelte van de as niet getekend is.
5 *Bronvermelding*. Onder de grafiek moet aangegeven zijn op welke wijze men de gegevens heeft verkregen.
6 *Legenda*. Als in een grafiek verschillende arceringen gebruikt worden, dan dient in een lijstje – de legenda – de betekenis van de arceringen te worden aangegeven.

Scheurlijn

1.3.2 Grafieken in een assenstelsel

Veel grafische voorstellingen komen tot stand door beschikbare gegevens weer te geven in een rechthoekig assenstelsel. Hierbij is een tweetal assen aangegeven, namelijk een horizontale as, die meestal de x-as wordt genoemd, en een verticale as, die doorgaans de y-as wordt genoemd.
Bij de x-as en de y-as dient men te vermelden welke grootheid hierlangs is afgezet. Verder moeten er enkele getallen (niet te veel) langs de assen geplaatst worden, waardoor een lezer gemakkelijk de waarde kan aflezen van een in de grafiek geplaatste uitkomst. Voor het tekenen van grafieken kan het nuttig zijn te beschikken over grafiekpapier, omdat dit een lijnenstructuur heeft, waardoor het eenvoudig wordt om de plaats van een punt in de grafiek te bepalen. Grafieken met een assenstelsel komen we met name tegen bij het weergeven van tijdreeksen en bij het tekenen van spreidingsdiagrammen.

x-as
y-as

Grafieken met een tijdas

Regelmatig komt het voor dat een grafische voorstelling moet worden gemaakt van het verloop van een grootheid (ook wel variabele genoemd) in de tijd. We noemen dit de weergave van een *tijdreeks* of *historische reeks*. Voorbeelden van tijdreeksen zijn: de jaaromzet van een onderneming over de afgelopen tien jaar, de kwartaalwinsten van een bedrijf weergegeven voor vijf achtereenvolgende jaren, de werkloosheid onder de beroepsbevolking zoals die maandelijks is vastgesteld in de afgelopen jaren.
Bij een grafiek van een tijdreeks kiezen we de horizontale as als tijdas. Hierop geven we de tijdstippen of perioden aan waarop de gegevens betrekking hebben. Bij iedere periode zetten we verticaal de waarde van de betrokken variabele af. Hierdoor ontstaan er punten in de grafiek. Het is gebruikelijk dat de punten in de grafiek verbonden worden door middel van lijnstukken (zie voorbeeld 1.10). We spreken daarom wel van een *lijndiagram*. Bij een lijndiagram zien we dus een wat 'hoekig' verloop van de curve. Het is niet toegestaan dit hoekige verloop weg te werken door een ietwat gebogen curve te tekenen. Wie zoiets doet, suggereert meer kennis

Tijdreeks

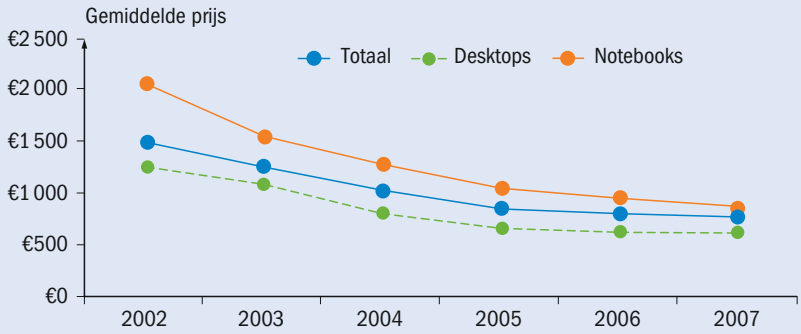
Lijndiagram

te hebben van het verloop van de grafiek dan uit de meetpunten te concluderen valt.
Lijndiagrammen worden vaak gebruikt in jaarverslagen van ondernemingen. Grafieken van bijvoorbeeld de omzetontwikkeling en de winstontwikkeling verstrekken de lezer van een jaarverslag snel een beeld van de gang van zaken. In het volgende voorbeeld zien we drie lijndiagrammen in één tekening geplaatst.

VOORBEELD 1.10

In figuur 1.2 is het prijsverloop van computers (desktops, notebooks) weergegeven voor een aantal jaren.

FIGUUR 1.2 Gemiddelde prijzen computers Nederland, van GfK Panelmarkt in Nederland

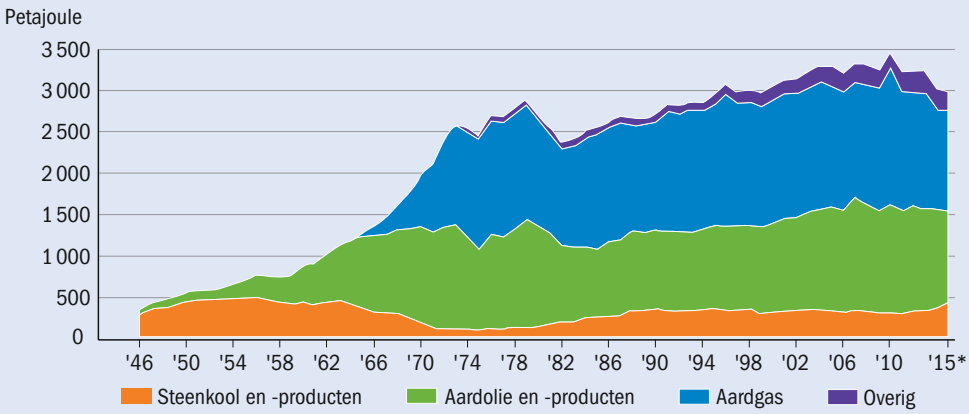


Bron: GfK Jaargids 2008

VOORBEELD 1.11

In de grafiek van figuur 1.3 zien we een zogenoemd gestapeld lijndiagram. De bovenste lijn geeft het totale energieaanbod in Nederland aan. Zichtbaar is gemaakt hoe dit totaal is opgebouwd. Zo kun je zien dat het belang van aardgas de laatste paar jaar terugloopt.

FIGUUR 1.3 Energieaanbod naar energiedrager



Bron: Trends in Nederland 2016

Grafieken met een logaritmische schaalverdeling

Soms wordt het gedrag van een tijdreeks weergegeven door een grafische voorstelling waarbij de schaalverdeling is veranderd van een gewone in een logaritmische. Een dergelijke handelwijze kan zijn nut hebben als er wordt gewerkt met een variabele die in de loop van de tijd een duidelijke groei vertoont. Als een variabele als een exponentiële functie van de tijd kan worden beschouwd, dan vertonen de logaritmen van de waargenomen waarden een lineair verband. We verduidelijken dit aan de hand van het volgende voorbeeld en figuur 1.4.

VOORBEELD 1.12

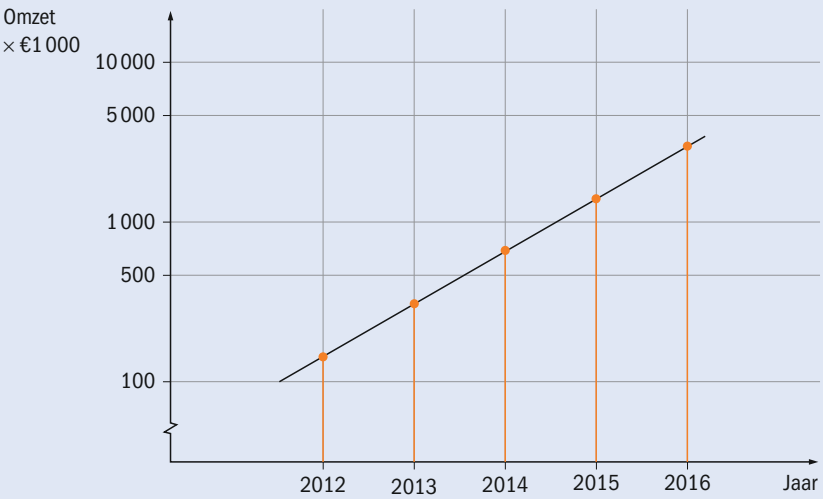
Het verloop van de omzet van een internetwinkel is weergegeven in tabel 1.15.

TABEL 1.15 Omzet internetwinkel

Jaar	2012	2013	2014	2015	2016
Omzet (× €1 000)	250	450	810	1 460	2 620

De omzet vertoont een exponentieel verloop (groei met ongeveer 80% per jaar). Als voor de verticale as een logaritmische schaal wordt gekozen, dan vertoont de reeks gegevens in de tekening een rechte lijn.

FIGUUR 1.4 Grafiek van de omzet

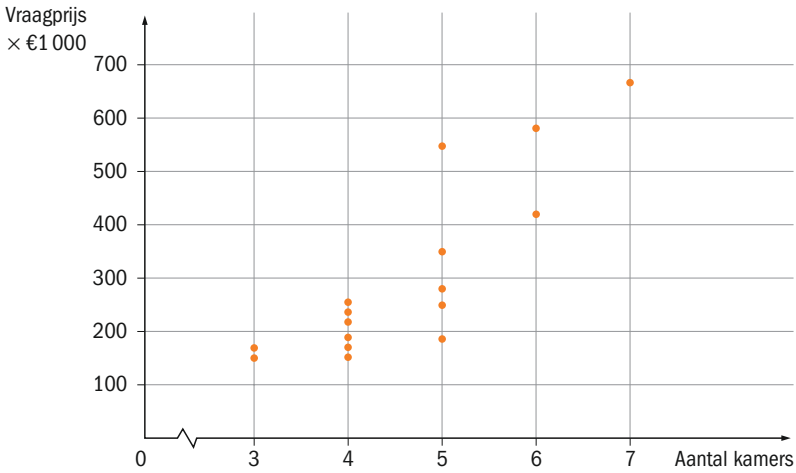


Spreidingsdiagrammen

Een andere toepassing van grafieken in een assenstelsel komen we tegen bij het spreidingsdiagram. Dit gebruikt men in het bijzonder bij rapportage van onderzoek waarbij een verband moet worden gezocht tussen twee variabelen. Men verzamelt dan gegevens die geregistreerd worden als een

x-waarde met een bijbehorende y-waarde. De waargenomen uitkomsten vormen dus getallenparen, zie figuur 1.5.

FIGUUR 1.5 Spreidingsdiagram



VOORBEELD 1.13

Bij makelaar Van Dijk zijn 16 woningen te koop. Van deze woningen staat in tabel 1.16 aangegeven het rangnummer dat deze hebben in het totale bestand, en verder het aantal kamers en de vraagprijs.
Om de samenhang tussen beide variabelen te kunnen onderzoeken, tekenen we een spreidingsdiagram. Dat is een grafiek met een x-as en een y-as waarlangs beide variabelen worden afgezet. Iedere waarneming wordt dan weergegeven door een meetpunt, figuur 1.5. De puntenwolk die aldus ontstaat geeft enigszins de indruk dat bij huizen met een groter aantal kamers doorgaans hogere vraagprijzen voorkomen.

TABEL 1.16 Woningen makelaar Van Dijk

Woning-nummer	Aantal kamers	Vraagprijs	Woning-nummer	Aantal kamers	Vraagprijs
16	3	162 000	58	4	265 000
20	4	168 000	61	5	275 000
25	3	175 000	62	4	275 000
29	4	178 000	74	5	345 000
38	4	189 000	85	6	420 000
41	5	198 000	95	5	549 000
48	4	229 500	103	6	587 500
53	5	247 500	110	7	669 000

Vaak kan men bij onderzoek naar de samenhang tussen bepaalde variabelen een bepaalde rangorde aangeven, waardoor de ene variabele als ‘oorzaak’ en de andere als ‘gevolg’ te typeren is. Als dat het geval is, moet men altijd de ‘oorzaak’-variabele langs de x-as plaatsen en de ‘gevolg’-variabele langs de y-as.

In hoofdstuk 13 (Regressie en correlatie) besteden we uitgebreid aandacht aan dit onderwerp en in hoofdstuk 2 (Maatstaven voor ligging en spreiding) geven we reeds een eenvoudig voorbeeld.

1.4 Diverse diagrammen

We zullen nu enkele voorbeelden geven van grafische voorstellingen die veel gebruikt worden bij het presenteren van gegevens.

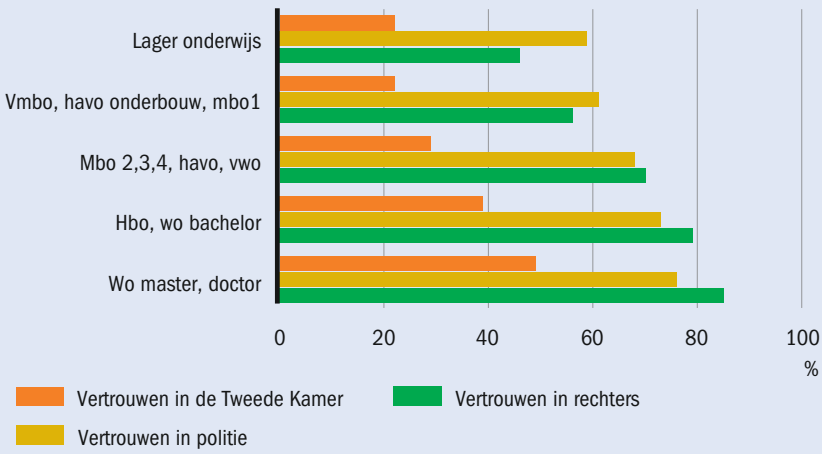
1.4.1 Staafdiagram

Een eenvoudige manier om gegevens te presenteren is het staafdiagram. De lengte van de staaf (of kolom) komt overeen met het gemeten aantal. Er zijn diverse staafdiagrammen mogelijk (horizontaal, verticaal, stapeldiagram). Bij een staafdiagram worden de staven of kolommen doorgaans los van elkaar (niet aaneensluitend) getekend. We geven enkele voorbeelden.

VOORBEELD 1.14

In figuur 1.6 is voor een vijftal opleidingsniveaus aangegeven hoeveel vertrouwen men heeft in de Tweede Kamer, de politie en rechters. De grafiek is te karakteriseren als een horizontaal samengesteld staafdiagram. Merk op, dat in de legenda de betekenis van de gebruikte arceringen wordt vermeld. In z'n algemeenheid kun je concluderen dat hogere opleidingsniveaus meer vertrouwen hebben in deze instanties dan lage opleidingsniveaus.

FIGUUR 1.6 Vertrouwen in autoriteiten naar opleidingsniveau



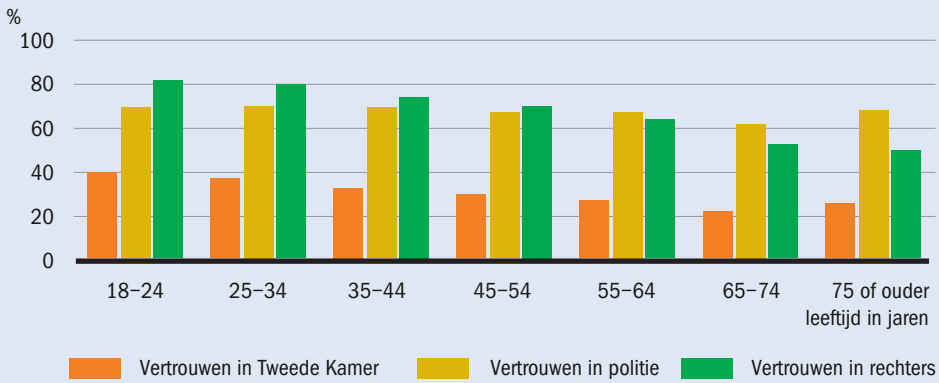
Bron: Trends in Nederland 2015

VOORBEELD 1.15

In figuur 1.7 wordt getoond hoe het vertrouwen in de Tweede Kamer, politie en rechters samenhangt met de leeftijd van de respondenten. Dit is een verticaal samengesteld staafdiagram.

Hoofdlijn is dat naarmate men ouder is er minder vertrouwen is. Ook blijkt dat van de drie categorieën de rechters doorgaans het meeste vertrouwen genieten, hoewel de politie juist het best scoort bij de hoogste leeftijden.

FIGUUR 1.7 Vertrouwen in autoriteiten naar leeftijd



Bron: Trends in Nederland 2015

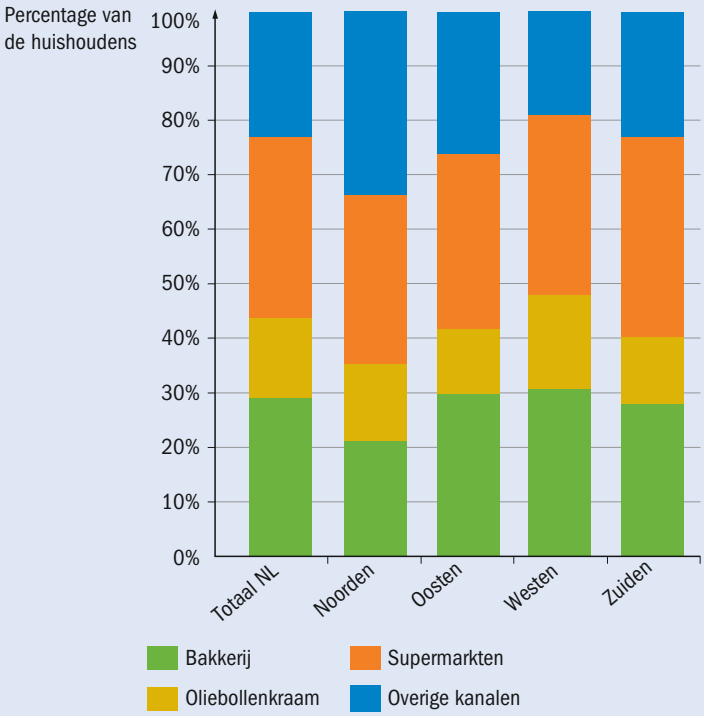
1.4.2 Stapeldiagram

Indien een totale hoeveelheid waarnemingen kan worden onderverdeeld in een aantal groepen, dan kan het stapeldiagram een aardige presentatie van de uitkomsten geven.

VOORBEELD 1.16

In het stapeldiagram van figuur 1.8 wordt tot uitdrukking gebracht dat Nederlanders omstreeks de jaarwisseling op allerlei manieren aan oliebollen en appelbeignets komen. Dit zijn procentuele stapeldiagrammen die betrekking hebben op Nederlanders uit de vier windstreken die niet zélf bakken. We zien dan bijvoorbeeld dat in het westen de oliebollenkraam populairder is dan in andere delen van het land.

FIGUUR 1.8 Waar worden oliebollen / appelbeignets gekocht?



Bron: GfK jaargids 2008

1.4.3 Cirkeldiagram

Het cirkeldiagram kan worden gebruikt bij variabelen waarbij een totaal wordt onderverdeeld in een aantal *disjuncte deelverzamelingen*. De oppervlakte van de cirkel wordt gerekend als 100%. De cirkel wordt verdeeld in een aantal sectoren dat overeenkomt met het aantal categorieën waarin de waarnemingen worden verdeeld. De oppervlakte van een sector moet dan overeenkomen met het aantal waarnemingen dat in een bepaalde categorie valt. De oppervlakte van de cirkel zelf geeft dan het totaal aantal waarnemingen aan.

Disjuncte deelverzamelingen

Niet alleen bij aantallen waarnemingen kan een cirkeldiagram worden gebruikt als grafische voorstelling. Iedere totaliteit die kan worden onderverdeeld in een aantal subgroepen (die samen die totaliteit voortbrengen) kan in beeld gebracht worden door een cirkeldiagram.

Door cirkels met verschillende oppervlaktes te kiezen, kan bovendien, bij vergelijking van twee totalen, het verschil tussen deze totalen tot uitdrukking worden gebracht.

Soms verdient in het betoog een bepaald onderdeel van het cirkeldiagram speciale aandacht. Om een sector een accent te geven, kan men deze een stukje uit de cirkel laten springen, zoals in het volgende voorbeeld.

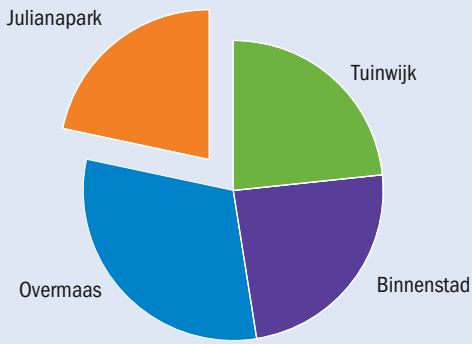
VOORBEELD 1.17

Bij de gezamenlijke makelaars in een gemeente zijn 120 woningen te koop. In een artikel in een wijkkrantje wil men de lezers er speciaal op wijzen hoe groot het aantal woningen is dat te koop staat in de wijk Julianapark. Dat zou kunnen met een cirkeldiagram waarbij de sector van de wijk Julianapark eruit springt. Zie figuur 1.9.
De aantallen zijn weergegeven in tabel 1.17.

TABEL 1.17 Locaties van woningen

Wijk	Aantal
Tuinwijk	28
Binnenstad	29
Overmaas	37
Julianapark	26
Totaal	120

FIGUUR 1.9 Verdeling van het woningaanbod

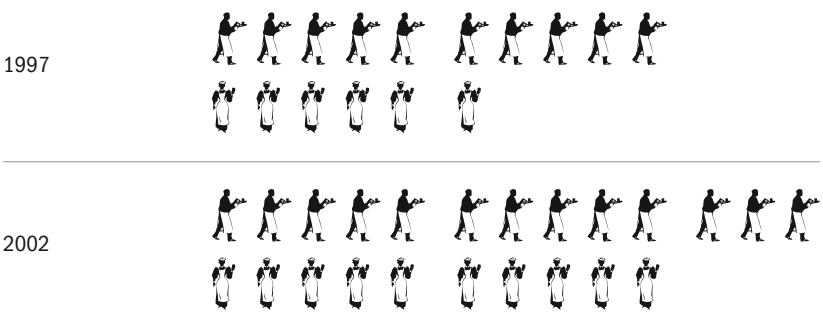


Bron: de lokale makelaars

1.4.4 Beelddiagram

Een andere manier om de uitkomsten van een variabele in grafiek te brengen, vinden we bij het beelddiagram.
Bij het beelddiagram worden tekeningetjes gebruikt om de grootte of omvang van de waargenomen uitkomsten weer te geven (zie figuur 1.10).
Opgemerkt moet worden dat het beelddiagram uitsluitend zijn belang ontleent aan het feit dat het een bepaalde publicatie wat verlevendigt. Het beelddiagram wordt doorgaans gebruikt in situaties waarin men ook een staafdiagram had kunnen toepassen.

FIGUUR 1.10 Aantal mannen en vrouwen in dienst van hotel Continental



(één persoon geeft tien werknemers aan)

Bron: Salarisadministratie van hotel Continental te Rotterdam

1.4.5 Naalddiagram

Erg verwant aan het staafdiagram is het naalddiagram. De lengte van een lijnstuk komt overeen met gemeten aantallen. We noemen het naalddiagram hier afzonderlijk omdat het qua vorm en bedoeling aansluit bij het begrip kansfunctie voor discrete kansvariabelen, dat aan de orde komt in hoofdstuk 4.

VOORBEELD 1.18

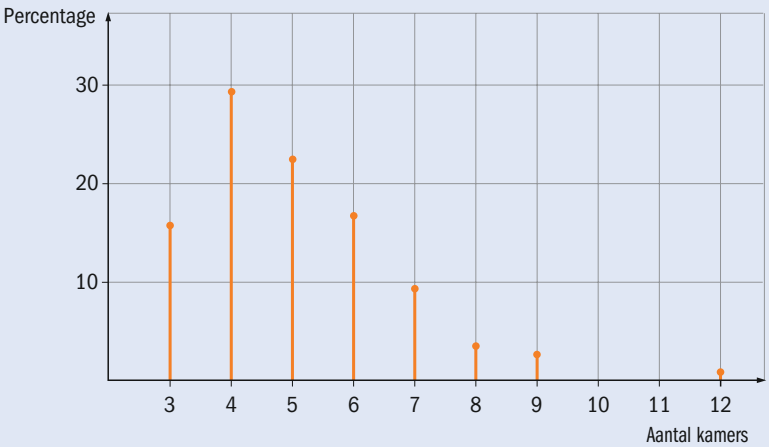
Voor het woningbestand maken we een overzicht van het aantal kamers bij de 120 woningen. Dit leidt tot tabel 1.18 waarvan vervolgens een zogeheten naalddiagram wordt getekend.

TABEL 1.18 Aantal kamers voor 120 woningen

Aantal kamers	Frequentie	Percentage
3	19	15,8
4	35	29,2
5	27	22,5
6	20	16,7
7	11	9,2
8	4	3,3
9	3	2,5
10	0	0
11	0	0
12	1	0,8
Totaal	120	100%

Om een naalddiagram te tekenen, plaatsen we de uitkomsten langs de horizontale as. Bij ieder van deze uitkomsten geven we door middel van een verticaal lijnstuk aan hoe vaak procentueel deze uitkomst is aangetroffen. De grafische voorstelling die aldus in figuur 1.11 ontstaat, noemen we een naalddiagram en is vooral geschikt voor de weergave van een frequentieverdeling met slechts een klein aantal mogelijke uitkomsten.

FIGUUR 1.11 Naalddiagram bij tabel 1.18



Bij voorbeeld 1.18 merken we nog het volgende op: in het naalddiagram zijn de resultaten door middel van percentages weergegeven. Men mag natuurlijk ook de waargenomen aantallen gebruiken zonder omrekening naar procenten. De vorm van de grafiek blijft dan hetzelfde, maar de indeling van de verticale as moeten we in dat geval veranderen.

1.5 Stamdiagram, histogram en frequentiepolygoon

In deze paragraaf komen twee manieren om gegevens te presenteren aan de orde, namelijk het stamdiagram en het histogram. Met beide methoden kan men door middel van de hoogte van de kolommen een beeld geven van de opbouw van een verdeling. Aansluitend op het histogram komt ook de frequentiepolygoon aan de orde.

1.5.1 Stamdiagram

Steel-en-bladdiagram

Een eenvoudige methode om een collectie gegevens weer te geven, is het stamdiagram. Deze wijze van presenteren komen we ook wel tegen onder de naam 'stem-and-leaf-display' of '*steel-en-bladdiagram*'. Het stamdiagram is eigenlijk een tussenvorm van tabellen en grafieken en het vertoont vrij veel verwantschap met het hierna te bespreken histogram. De bedoeling is dat van de verzameling getallen het eerste cijfer wordt afgezonderd. Dit wordt langs de horizontale as geplaatst. De daaropvolgende cijfers van een uitkomst worden boven de waarde van het eerste cijfer geplaatst. Door dit voor alle waarnemingen te doen, ontstaat een soort kolommen. We lichten dit toe met het volgende voorbeeld.

VOORBEELD 1.19

Bij een onderzoek naar de inkomens van werkende jongeren werden voor 30 jongeren de netto-weekinkomens bepaald. De resultaten zijn weergegeven in tabel 1.19.

TABEL 1.19 Netto-weekinkomens (in euro) voor 30 jongeren

180	269	508	224	440	190	326	378	311	232
212	348	292	293	362	274	196	282	482	266
415	287	345	217	388	210	168	470	317	248

We geven nu eerst de 'stam' aan van de verdeling door te letten op het eerste cijfer van de waargenomen uitkomsten. Hierbij vinden we de getallen 1, 2, 3, 4 en 5. Deze getallen markeren we langs de horizontale as (zie figuur 1.12). Vervolgens plaatsen we boven deze getallen de volgende (twee) cijfers van alle waargenomen uitkomsten. We zien dan dat het getal 180 verschijnt als 80 boven de waarde 1 op de stam. Hierdoor ontstaan in feite kolommen boven de getallen 1, 2, 3, 4 en 5.
Hoe hoger de kolom, des te meer waarnemingen waren er met een begincijfer dat overeenkomt met de kolom.

FIGUUR 1.12 Stamdiagram weekinkomens

	48			
	10			
	17			
	87			
	66			
	82	17		
	74	88		
	93	45		
	92	62		
68	12	48	70	
96	32	11	15	
90	24	78	82	
80	69	26	40	08
1	2	3	4	5

Door deze manier van weergeven ontstaat een aardig beeld van de verdeling van de waargenomen getallen. We kunnen desgewenst ook een groter aantal kolommen maken door iedere kolom (die hier een breedte heeft van €100) te splitsen in een lage helft en een hoge helft. We krijgen dan bijvoorbeeld twee kolommen met begincijfer 2. Bij de eerste kolom worden dan de getallen 200 tot en met 249 afgezet en bij de tweede kolom 250 tot en met 299. In het hiervoor weergegeven stamdiagram hebben we de getallen binnen een kolom ongeordend weergegeven. We kunnen deze cijfers uiteraard ook op volgorde van grootte weergeven. Hierdoor ontstaat de mogelijkheid op een gemakkelijke wijze te inspecteren of de getallen min of meer gelijkmatig gespreid zijn over een klasse.
In het stamdiagram van figuur 1.13 hebben we de verdeling van weekinkomens nogmaals weergegeven. Hierbij hebben we de kolommen gesplitst en bovendien zijn de uitkomsten op volgorde geplaatst.

FIGUUR 1.13 Gesplitst stamdiagram weekinkomens

			93						
		48	92						
		32	87	48					
	96	24	82	45					
	90	17	74	26	88				
	80	12	69	17	78	40	82		
	68	10	66	11	62	15	70	08	
1	1	2	2	3	3	4	4	5	5

Opmerkingen
Soms ziet men bij toepassing van computerprogramma's dat de kolommen van het stamdiagram niet worden opgevuld met getallen maar met sterretjes. In dat geval komt het stamdiagram vrijwel overeen met een frequentieverdeling waarbij geturfd wordt (zie voorbeeld 1.3b).

Een stamdiagram kan ook worden getekend ten opzichte van een verticale as die als stam dient. De kolommen lopen dan uiteraard horizontaal.
Men kan met een stamdiagram de vorm van twee verdelingen met elkaar vergelijken door voor de tweede verdeling kolommen naar beneden te tekenen ten opzichte van dezelfde stam, of door één verdeling links en de andere verdeling rechts van de verticale stam te tekenen.
De getallen die langs de stam staan, kunnen ook uit meer dan één cijfer bestaan.
Afhankelijk van het aantal getallen langs de stam ontstaan weinig of veel kolommen. Om een goed beeld van de verdeling te krijgen, moet vaak een middenweg worden bewandeld.

VOORBEELD 1.20
Een docent aan een hogeschool onderzoekt of de resultaten van een toets statistiek verschillen voor leerlingen met wiskunde B in hun vwo-pakket ten opzichte van leerlingen die dit vak niet hadden. Tabel 1.20 geeft de resultaten.

TABEL 1.20 Toetsuitslagen

Wel wiskunde B					Geen wiskunde B				
78	84	56	92	63	79	62	52	48	64
67	45	60	69	73	82	54	42	56	50
75	85	96	51	62	37	86	68	46	59
78	82	86	72	57	55	71	57	67	49
57	77	64	63	48	65	69	62	32	76

Vervolgens plaatsen we de gegevens in het stamdiagram van figuur 1.14. In dit tweezijdige stamdiagram kunnen we in één oogopslag vaststellen dat de twee verdelingen ten opzichte van elkaar enigszins verschoven liggen.

FIGUUR 1.14 Tweezijdig stamdiagram toetsuitslagen

Wel wiskunde B								Geen wiskunde B							
				6	2	**9**									
			6	5	4	2	**8**	2	6						
	8	8	7	5	3	2	**7**	1	6	9					
9	7	4	3	3	2	0	**6**	2	2	4	5	7	8	9	
			7	7	6	1	**5**	0	2	4	5	6	7	9	
					8	5	**4**	2	6	8	9				
							3	2	7						

1.5.2 Histogram

De bekendste grafische voorstelling waarmee we een frequentieverdeling kunnen uitbeelden, is het histogram. Het histogram wordt in het algemeen toegepast bij frequentieverdelingen waarvan de waargenomen variabele een ratioschaal heeft. Om het histogram te tekenen, wordt de horizontale as verdeeld in een aantal intervallen, die overeenkomen met de klassen uit de frequentieverdeling. Boven elk interval wordt een kolom geplaatst

waarvan de *oppervlakte* overeenkomt met het aantal waarnemingen dat tot de betreffende klasse behoort. De grafische voorstelling die aldus ontstaat, bestaat uit een aantal aaneensluitende kolommen.

Oppervlakte = aantal waarnemingen

Bij het tekenen van de kolommen moet voorzichtigheid worden betracht. We kunnen niet zonder meer het aantal waarnemingen uit de frequentieverdeling gebruiken om de hoogte van de kolom aan te geven, zoals we deden bij het staafdiagram. De klassen kunnen namelijk van verschillende breedte zijn. Omdat de oppervlakte van een kolom wordt berekend als hoogte maal breedte, moet men bij het aangeven van de hoogte van een kolom rekening houden met de breedte van de desbetreffende klasse. Indien er bijvoorbeeld 10 waarnemingen voorkomen in een klasse van 5 eenheden breed, dan moet dit grafisch aangegeven worden door een kolom van twee eenheden hoog (want: $2 \times 5 = 10$, dus oppervlakte kolom = frequentie). Om de hoogte van de kolom te berekenen, moet de frequentie worden gedeeld door de breedte van de desbetreffende klasse. De uitkomsten die op deze wijze worden berekend, noemt men *frequentiedichtheden*.

Frequentiedichtheden

Frequentiedichtheid is frequentie per eenheid van klassenbreedte. In het volgende voorbeeld laten we zien hoe het histogram tot stand komt.

VOORBEELD 1.21

Bij een onderzoek naar het tegoed dat 100 rekeninghouders aanhouden bij een bank is tabel 1.21 ontstaan.

TABEL 1.21 Banksaldi (in euro) van 100 personen

Banksaldo	Frequentie
0 – < 2 000	17
2 000 – < 5 000	18
5 000 – < 10 000	15
10 000 – < 20 000	20
20 000 – < 50 000	30
Totaal	100

We gaan nu een histogram tekenen van de frequentieverdeling.
Omdat de klassenbreedten verschillen, moeten we eerst frequentiedichtheden berekenen. Als eenheid van klassenbreedte kiezen we €1 000. De frequentiedichtheden zijn berekend in tabel 1.22.

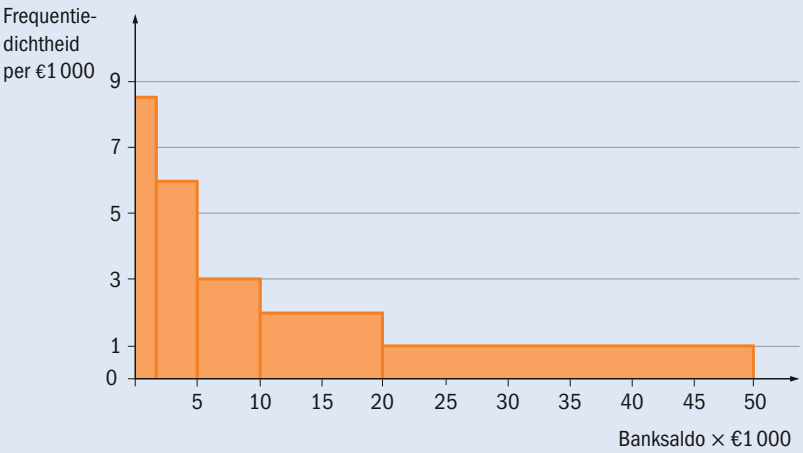
TABEL 1.22 Berekening van de frequentiedichtheid per € 1 000

Banksaldo	Frequentie	Klassenbreedte	Frequentiedichtheid
0 – < 2 000	17	2	8,5 (namelijk 17/2)
2 000 – < 5 000	18	3	6 18/3
5 000 – < 10 000	15	5	3 15/5
10 000 – < 20 000	20	10	2 20/10
20 000 – < 50 000	30	30	1 30/30

Wanneer eenmaal de frequentiedichtheden berekend zijn, is het tekenen van een histogram een eenvoudige zaak, zie figuur 1.15. Als hoogte van een kolom kiezen we telkens de berekende frequentiedichtheid. Automatisch geven oppervlaktes in de tekeningen nu de aantallen aan. Het gebruik van de

absolute frequenties als hoogte van de kolommen zou een onjuist beeld geven van het verschijnsel, omdat kolommen boven brede klassen dan een veel te groot oppervlakte krijgen.

FIGUUR 1.15 Histogram van de banksaldi



Het berekenen van de frequentiedichtheden is niet noodzakelijk als alle klassen dezelfde breedte hebben. Bij even brede klassen komt de vorm van het histogram sterk overeen met het stamdiagram. Zoiets kunnen we nagaan door een histogram te tekenen met de gegevens van voorbeeld 1.20. Als we de klassen €100 breed kiezen, ontstaat een vorm voor het histogram die doet denken aan het geconstrueerde stamdiagram.

1.5.3 Klassengrenzen

In principe is het histogram een grafische voorstelling die wordt gebruikt om de frequentieverdeling van een continue variabele te tonen. Vandaar ook dat de kolommen aaneensluitend worden getekend. Soms is er een probleem met het bepalen en interpreteren van *klassengrenzen*. Deze problemen houden verband met het onderscheid tussen continue en discrete variabelen. We noemen de volgende gevallen:

a Het betreft een 'echte' continue variabele zoals de tijd. Als we er *bovendien* van uit mogen gaan dat de uitkomsten exact zijn gemeten (dus niet afgerond), dan beschouwen we klassengrenzen als exacte grenzen. Een klasse 10,0 – < 15,00 betekent dan dat 10,0000. . . de laagste waarde is die tot de klasse behoort en 14,9999. . . de hoogste waarde.
Vaak hebben we te maken met een geval dat hier dichtbij in de buurt komt. In voorbeeld 1.21 betekent de klasse 5 000 – < 10 000 euro bijvoorbeeld dat het hoogst denkbare banksaldo dat in die klasse wordt geplaatst 9 999,99 is.

b Het betreft een continue variabele waarvan we weten dat de uitkomsten worden gemeten als getallen die op een bepaalde manier zijn afgerond. We kunnen hierbij bijvoorbeeld denken aan de gewichten van personen die in hele kilo's worden weergegeven. Als we dan klassen maken,

bijvoorbeeld 60 tot en met 64 kilogram, moeten we bedenken dat elke gemeten waarde een afgerond getal is. Dus 60 kg staat voor een werkelijk gewicht tussen 59,500 en 60,500 kg. Een klasse 60 tot en met 64 kg moet dan in een histogram worden weergegeven met als echte grenzen 59,500 en 64,500. Op deze manier ontstaat vanzelf een kolom van vijf eenheden (5 kg) breed.

Wanneer verschuiven klassengrenzen?

c Het betreft een discrete variabele die bijvoorbeeld alleen gehele waarden kan aannemen. Als we dan tóch een histogram willen tekenen, dan doen we eigenlijk precies het omgekeerde van afronden: het gehele getal 60 wordt dan beschouwd als het 'gebied' 59,5 – < 60,5, omdat een histogram nu eenmaal met een continue horizontale as werkt. We lichten dit toe in het volgende voorbeeld.

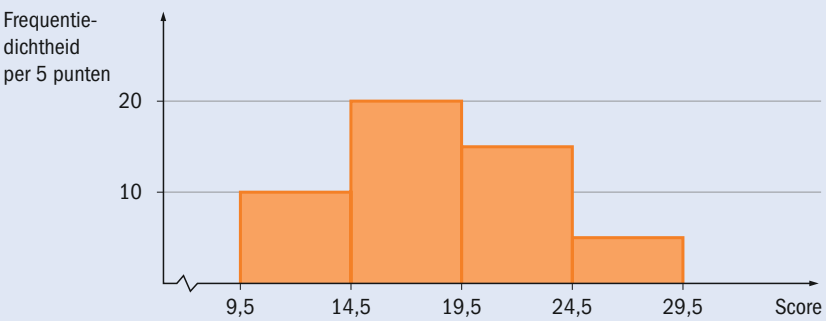
VOORBEELD 1.22

Aan een test wordt door 50 personen deelgenomen. Tabel 1.23 geeft een overzicht van de resultaten. Alle klassen hebben hier dezelfde breedte. Als we als eenheid van klassenbreedte hier 5 kiezen, dan zijn de frequentiedichtheden (die we voor het histogram nodig hebben) gelijk aan de weergegeven frequenties.

TABEL 1.23 Testscores van 50 personen

Scores	Frequenties
10 – < 15 (10, 11, 12, 13 en 14)	10
15 – < 20	20
20 – < 25	15
25 – < 30	5
Totaal	50

FIGUUR 1.16 Histogram van de testscores van 50 personen



Bij het verwerken van deze gegevens in een histogram moeten we bedenken dat het midden van de eerste klasse 12 is (en niet 12,50). Dit is een gevolg van het feit dat de benedengrens 10 wel en de bovengrens 15 niet in de klasse 10 – < 15 valt.

Om van de tabel een 'eerlijk' histogram te tekenen, moeten we daarom de klassengrenzen met $\frac{1}{2}$ verlagen. Zodoende krijgen we als klassengrenzen: $9\frac{1}{2}$, $14\frac{1}{2}$, $19\frac{1}{2}$, $24\frac{1}{2}$ en $29\frac{1}{2}$. Dat leidt tot het histogram van figuur 1.16.

Opmerkingen

- Als we – zoals in voorbeeld 1.22 – de klassengrenzen met $\frac{1}{2}$ dienen te verlagen om een eerlijk histogram te krijgen, dan is het ter wille van de presentatie van de grafiek soms toch wenselijk om gehele waarden langs de assen te zetten. In de grafiek in figuur 1.16 zouden we dan 10, 15, 20, 25 en 30 langs de assen plaatsen. Het dunne streepje waarmee zo het getal 10 aangegeven wordt, dient dan echter correct op de as te verschijnen, dus een halve eenheid rechts van het begin van de kolom.
- Als we enige vrijheid hebben om de klassengrenzen te kiezen, dan heeft het doorgaans de voorkeur om ronde getallen te gebruiken in verband met de leesbaarheid van het geheel. Vandaar dat we in voorbeeld 1.21 (van de banksaldi) de klasse 20 000 – < 50 000 zien en niet bijvoorbeeld 21 470 – < 38 210
- **Open klasse** Een bijzonder soort probleem bij het ontwerpen van een histogram kan nog optreden indien er sprake is van een open klasse. Met een open klasse bedoelen we een klasse waarin niet – zoals gebruikelijk – een benedengrens en een bovengrens zijn gegeven, maar waarbij een van die twee ontbreekt. In het histogram moet dan frequentiedichtheid (= kolomhoogte) 0 worden aangehouden, tenzij op grond van een logische redenering alsnog een klassengrens wordt verzonnen. De open klasse wordt dan weer gesloten.

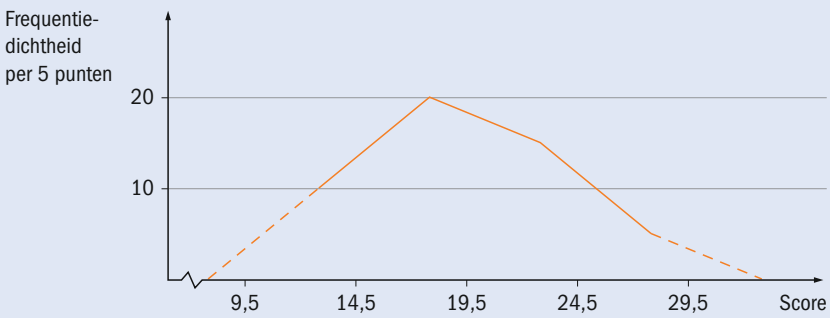
1.5.4 Frequentiepolygoon

Een van het histogram afgeleide grafiek is de frequentiepolygoon. Hierbij worden eerst de klassenmiddens bepaald en daarna worden punten in de grafiek getekend boven de klassenmiddens. De hoogte van deze punten is gelijk aan de hoogte van de kolom in de histogram. De verbindingslijn van deze punten noemt men een frequentiepolygoon.

VOORBEELD 1.23

Voorbeeld 1.22 levert een polygoon zoals weergegeven in figuur 1.17.

FIGUUR 1.17 Grafiek van de testscores van 50 personen



Bij de hier gegeven frequentiepolygoon hebben we in principe de beschikking over vier punten in de grafiek. Bij het klassenmidden 12 hoort de hoogte 10, bij het klassenmidden 17 vinden we de hoogte 20, bij klassenmidden 22 hoort een frequentie van 15 en bij het klassenmidden 27 hoort de waarde 5.

De verbindingslijn van deze punten geeft ons in principe de polygoon. Om de curve te laten doorlopen tot de x-as zijn stippellijnen in de grafiek getekend. Deze komen tot stand door naast de bestaande klassen nog een tweetal klassen te bedenken, één ter linkerzijde en één ter rechterzijde, waarvan de bijbehorende frequenties de waarde 0 hebben. Zo'n denkbeeldige klasse kiezen we even breed als de naastliggende bestaande klasse. In dit voorbeeld levert deze gedachte een klassenmidden van 7 op (van de klasse 5 tot en met 9) en een klassenmidden van 32 (van de klasse 30 tot en met 34). Omdat de hoogte van het histogram bij deze punten 0 bedraagt, kan de polygoon nu volledig worden getekend (zie de stippellijn in figuur 1.17).

1.6 Cumulatieve frequenties

De gegevens van een frequentieverdeling kunnen ook worden weergegeven door een cumulatieve verdeling. Hierbij gaat het niet meer om de aantallen per klasse maar om de waargenomen aantallen *beneden* een bepaalde grenswaarde.
We gaan eerst in op de cumulatieve frequentieverdeling en daarna bespreken we het Paretodiagram.

1.6.1 Cumulatieve frequentieverdeling

Een cumulatieve frequentieverdeling kunnen we vormen door allereerst de bovengrens van alle klassen te bepalen en vervolgens het aantal waarnemingen te tellen dat beneden elke grens ligt.

VOORBEELD 1.24

De gegevens uit voorbeeld 1.22 leveren de cumulatieve frequenties zoals in tabel 1.24.

TABEL 1.24 Cumulatieve frequentieverdeling van de testscores van 50 personen

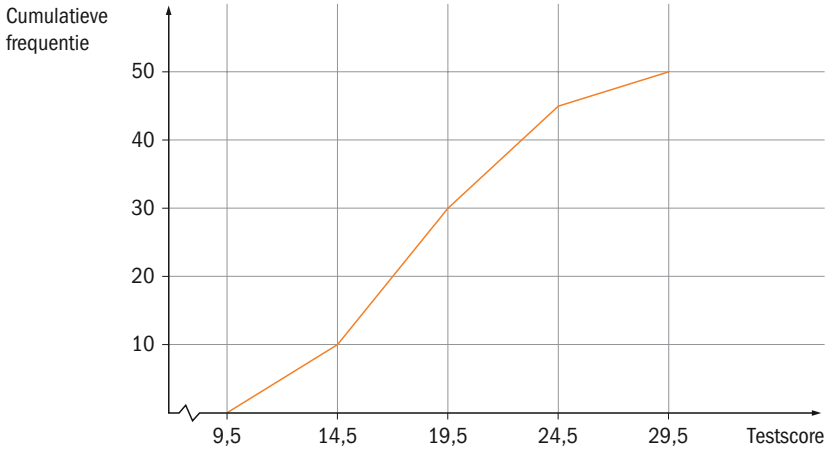
Score	f_i	Bovengrens	Aantal beneden de grens
10 – < 15	10	15 (of 14,5)	10
15 – < 20	20	20 (19,5)	30
20 – < 25	15	25 (24,5)	45
25 – < 30	5	30 (29,5)	50
Totaal			50

We kunnen dit resultaat ook in een grafiek verwerken. We krijgen dan een cumulatieve frequentiepolygoon (soms genoemd 'kleiner-dan-ogive') zoals weergegeven in figuur 1.18.

Cumulatieve frequentiepolygoon

Merk op dat frequentie*dichtheden* geen rol spelen bij de constructie van een cumulatieve frequentieverdeling.
Door cumulatieve frequenties te delen door het totale aantal waarnemingen, ontstaat een *relatieve cumulatieve frequentieverdeling*. Een dergelijke verdeling kan soms belangrijk zijn bij het vergelijken van een tweetal verdelingen.

FIGUUR 1.18 Cumulatieve frequentiepolygoon van de testscores van 50 personen



VOORBEELD 1.25

We berekenen in tabel 1.25 de relatieve cumulatieve frequenties van de gegevens uit het vorige voorbeeld. Omdat het totaal aantal waarnemingen 50 bedraagt, ontstaan de relatieve cumulatieve frequenties door de cumulatieve frequenties door 50 te delen.

TABEL 1.25 Relatieve cumulatieve frequenties van 50 testscores

Score	f_i	Bovengrens	Cumulatieve frequentie	Relatieve cumulatieve frequentie
10 – < 15	10	15 (of 14,5)	10	0,20
15 – < 20	20	20 (19,5)	30	0,60
20 – < 25	15	25 (24,5)	45	0,90
25 – < 30	5	30 (29,5)	50	1,00

1.6.2 Paretodiagram

Het *Paretodiagram* is een bijzondere variant op het staafdiagram. Het kan – net als het gewone staafdiagram – worden gebruikt voor het weergeven van frequenties bij een nominale variabele, dus een variabele waarvan een uitkomst een kenmerk is, maar niet noodzakelijk een getal.

Bij een Paretodiagram worden de staven van groot naar klein naast elkaar gezet. Daardoor zie je in de grafiek aan de linkerkant de uitkomsten die het vaakst zijn waargenomen. Meer naar rechts staan staafjes die aangeven dat deze kenmerken maar zelden worden genoteerd.

Vervolgens is het de bedoeling dat er een cumulatieve grafiek ontstaat van de frequenties. Hierbij worden eerst de punten in de grafiek aangegeven boven het midden van alle kolommen. Boven de tweede kolom tellen we de frequentie van kolom 2 bij de frequentie van kolom 1. En zo gaan we door: we tellen steeds de nieuwe frequentie bij de eerder bereikte stand. Deze cumulatieve grafiek eindigt rechtsboven uiteraard op het niveau 100%.

Paretodiagrammen worden nogal eens toegepast in het gebied van kwaliteitsmanagement. Doel is dan om factoren op te sporen die belangrijk zijn bij klachten over het functioneren van een product. Met het Paretodiagram zie je in één oogopslag welke factoren het belangrijkst worden gevonden. Als je deze als eerste aanpakt, wordt de klant vermoedelijk heel wat tevredener.

VOORBEELD 1.26

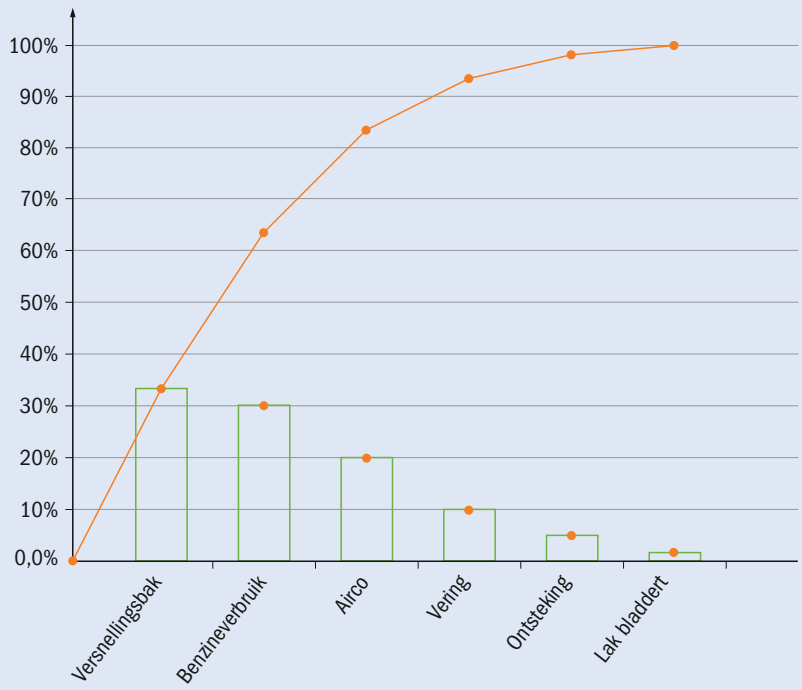
De auto-importeur van de Niswoo Suv de luxe maakt zich zorgen over de klachten die worden gemeld door kopers tijdens de garantieperiode. In tabel 1.26 worden de soort en het aantal klachten weergegeven.

TABEL 1.26 Verdeling klachten naar type en aantal

Type klacht	Aantal klachten	Percentage
Airco	24	20 %
Benzineverbruik	36	30 %
Lak bladdert	2	1,7 %
Ontsteking	6	5 %
Vering	12	10 %
Versnellingsbak	40	33,3 %
Totaal	120	100 %

We geven de verdeling aan met een paretodiagram, zie figuur 1.19.

FIGUUR 1.19 Paretodiagram



In de grafiek is gemakkelijk te zien welke de drie belangrijkste factoren zijn. Als de importeur op basis hiervan prioriteiten zou moeten aangeven, wordt duidelijk dat meer dan 83% van de klachten zou kunnen worden voorkomen indien de problemen met versnellingsbak, benzineverbruik en airco kunnen worden opgelost.

1.7 Weinig data, veel data of Big Data?

De laatste jaren zien we een nieuwe trend met de naam Big Data. Wat wordt daar eigenlijk onder verstaan? En wat kunnen we ermee? We geven hier wat context van Big Data vanuit het perspectief van het vakgebied statistiek.

Ooit was de focus van het vak statistiek vooral gericht op het verzamelen van heel veel gegevens die vervolgens in kaart werden gebracht door ze te ordenen zodat een beeld kan ontstaan van het grote geheel. Het meest aansprekende voorbeeld zijn misschien volkstellingen die al in de tijd van de Romeinen werden gehouden. Ruim twee eeuwen geleden werden volkstellingen ingevoerd in de USA en allerlei Europese landen. Men zou een volkstelling kunnen aanduiden als het verzamelen en analyseren van *veel* data.

Dankzij de opkomst van de kansrekening kwam een andere aanpak in beeld, namelijk het gebruik van steekproeven om een beeld te vormen van het grote geheel. Op basis van gegevens over enkele honderden proefpersonen keurt men soms een nieuw medicijn goed of men controleert steekproefsgewijs of een productieproces nog goed is ingesteld. Typisch het werken met *weinig* data. Destijds was men gedwongen om de hoeveelheid waarnemingen een beetje binnen de perken te houden, omdat het rekenwerk 'met de hand' soms bijna onuitvoerbaar is als databestanden erg groot zijn.

Alles werd anders met de komst van computers, die het mogelijk maakten om met grote bestanden allerlei analyses uit te voeren. Dus toen konden we weer werken met *veel* data. Het is de stormachtige technologische ontwikkeling van de laatste twintig jaar geweest waardoor een geheel nieuwe trend is ontstaan. Deze wordt aangeduid met de term Big Data. Eigenlijk speelt die ontwikkeling op twee fronten.

1 Er wordt steeds meer informatie gegenereerd. Op allerlei manieren wordt heden ten dage heel veel informatie verzameld. In diverse auto's zitten allerlei slimme sensoren die gegevens kunnen registreren over het rijgedrag van de automobilist. Als je dat voor miljoenen automobilisten dag-in dag-uit doet, dan ontstaan er gigantisch grote hoeveelheden gegevens. Als verzekeringsmaatschappijen die data zouden analyseren dan kunnen zij precies bepalen aan welk type personen zij premiekorting kunnen geven. Ook de bonuskaart van Albert Heijn is een voorbeeld van een immense datahoeveelheid die dagelijks verder toeneemt. Met slimme methoden van data-analyse kan men proberen allerlei verbanden te ontdekken.
2 De geheugencapaciteit en de verwerkingscapaciteit van computers is exponentieel gegroeid. Dat betekent dat de analyse en de opslag van immens grote bestanden technologisch gezien geen probleem meer vormt. Een van de belangrijkste aspecten van Big Data is dat het niet

alleen gaat om databestanden die goed georganiseerd in een computer zijn opgeslagen in de vorm van een datamatrix. Er is ook een bijna oneindig grote hoeveelheid informatie die via het internet vindbaar is. We krijgen dan ook te maken met ongestructureerde data zoals tekstberichten, foto's en allerlei communicatie op de social media. De uitdaging is om deze informatie zódanig te bewerken dat het mogelijk is daarmee analyses uit te voeren.

Vijf kenmerken

Het is gebruikelijk om Big Data te karakteriseren aan de hand van enkele kenmerken die vaak worden aangeduid als de (Engelstalige) V's. In eerste instantie waren dat er drie, namelijk:

a *Volume*, dat verwijst naar de hoeveelheid informatie (Veel),
b *Velocity*, dan hebben we het over de snelheid (Vaart) waarmee informatie ons kan bereiken,
c *Variety*, omdat de informatie allerlei verschijningsvormen (Variatie) kan hebben, zoals foto's, e-mails, traditionele databestanden en informatie van sensoren.

Aan dit lijstje worden soms nog toegevoegd de termen *Veracity* (betrouwbaarheid) en *Value* (wat is de waarde van al die informatie, wat doen we ermee als organisatie).

Big data is vooral technologie voor dataverzameling, dataselectie en dataopslag. Om die klus uit te kunnen voeren zijn veel goed-opgeleide data-analisten nodig. Anno 2017 was data-analist de meest gevraagde functie op de Nederlandse arbeidsmarkt voor afgestudeerden aan het hbo en de universiteiten.

Na de fase van dataverzameling is de volgende stap het maken van een zinvolle analyse van die data. Dan wordt het vakgebied statistiek belangrijk. Dit hoofdstuk gaat over beschrijvende statistiek. Verderop maken we kennis met heel wat methoden – zoals schatten, toetsen, correlatie en regressie – die van onschatbare waarde zijn voor de data-analist die ontdekkingen wil doen met het analyseren van Big Data.

1.8 Werken met Excel

Het programma Excel beschikt over diverse mogelijkheden om tabellen en grafieken te maken. Binnen het programma kan men op diverse manieren toegang krijgen tot bepaalde statistische technieken.

In de eerste plaats kan men in het lint op het tabblad 'Formules' de knop 'Functie invoegen' vinden. Hier worden diverse functiecategorieën genoemd. De categorie 'Statistisch' levert een grote hoeveelheid mogelijkheden op. Het gebruik hiervan wordt in volgende hoofdstukken toegelicht.

Op het tabblad 'Invoegen' bevindt zich ook de groep 'Grafieken'. Hiermee start men een stappenplan om een grafiek op te bouwen. Er is een grote collectie beschikbaar en de meeste standaardgrafieken, zoals cirkeldiagram en staafdiagram, behoeven geen nadere toelichting. Men leert deze het gemakkelijkst gebruiken indien men met wat data eens een uurtje gaat experimenteren.

- Op het tabblad 'Gegevens' is een faciliteit genaamd 'Gegevensanalyse'. Hier is een aantal statistische analysetechnieken beschikbaar.
- Op het tabblad 'Invoegen' is een keuzemogelijkheid 'Draaitabel'.

Aan de twee laatstgenoemde faciliteiten geven we in het kader van dit hoofdstuk aandacht, namelijk bij het werken met 'Histogram' en 'Draaitabelrapport'. Het histogram biedt de mogelijkheid om tellingen te doen binnen een op te geven waardebereik en het resultaat daarvan vervolgens als een tabel of een grafiek weer te geven.

FIGUUR 1.20 Het bestand woningen.asc

	A	B	C	D	E	F	G	H	I
1	nr.	wijk	kamers	badkamer	perceelopp	bouwjaar	garage	c.v.	vraagprijs
2	1	1	3	0	120	1920	0	0	132000
3	2	1	4	0	140	1932	0	0	137500
4	3	3	4	0	110	1938	0	1	138000
5	4	1	3	1	110	1927	0	0	139500
6	5	1	3	0	140	1968	0	1	142000
7	6	1	3	0	220	1935	0	0	144000
8	7	3	3	1	190	1938	0	0	145000
9	8	1	4	0	130	1950	0	0	145000
10	9	1	4	0	120	1964	0	1	146500
11	10	3	3	0	200	1966	0	1	148000

Met behulp van de functie 'Draaitabel' kan men snel scores voor een bepaalde variabele tellen, maar men kan ook kruistabellen maken en binnen zo'n kruistabel gemiddelden of totalen laten berekenen. We bespreken beide zaken aan de hand van het gegevensbestand 'woningen', dat reeds in dit hoofdstuk is besproken. Dit bestand is verkrijgbaar op diskette en men kan de ascii-gegevens direct in Excel inlezen.

1 Start Excel en kies 'Bestand' → 'Openen'. Zorg dat alle bestanden worden weergeven en open het bestand 'woningen.asc', zie figuur 1.20. Excel start automatisch de Wizard 'Tekst importeren' en herkent de gegevens. Doorloop de stappen en sla het bestand ten slotte op als woningen.xls. Let op dat nu wordt gekozen voor een Excel-werkmap.
2 Als eerste rij in het werkblad willen we de namen van de variabelen invoegen (via het menu 'Invoegen' → 'Rijen'). Voeg een rij in en vermeld hier de namen van de variabelen.

1.8.1 Histogram

Excel biedt in het lint op het tabblad 'Gegevens' een knop 'Gegevensanalyse'[1] een verzameling analysehulpmiddelen.
Voor het maken van een klassenindeling is de optie 'Histogram' beschikbaar. We passen dit toe op de gegevens van voorbeeld 1.4.

1 Voeg in de werkmap met de gegevens van de woningen een nieuw werkblad toe (via het menu 'Invoegen' → 'Blad invoegen'). Dit werkblad gebruiken we om de klassenindeling op te stellen en de frequentietabel af te drukken.

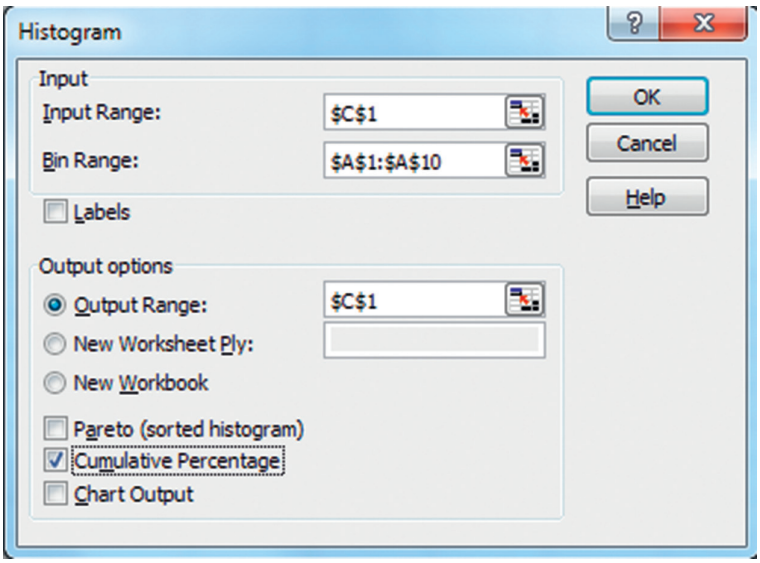
1 Als de opdracht 'Gegevensanalyse' niet op het tabblad 'Gegevens' wordt vermeld, moet u de invoegtoepassing Analysis ToolPak activeren met behulp van 'Bestand' → 'Opties' → 'Invoegtoepassingen'. Gebruik de helpfunctie van Excel voor meer informatie over het installeren van de Analysis ToolPak.

2 Plaats de gewenste klassengrenzen in kolom A van dit nieuwe werkblad, te beginnen bij cel A2. Zie figuur 1.21.
3 Kies in het menu 'Gegevens' → 'Gegevensanalyse' de optie 'Histogram'.

FIGUUR 1.21 Klassengrenzen

	A
1	
2	1910
3	1920
4	1930
5	1940
6	1950
7	1960
8	1970
9	1980
10	1990

FIGUUR 1.22 Gegevens vermelden



Je moet nu gaan vertellen waar het lijstje met klassengrenzen staat (= Verzamelbereik) en waar de gegevens staan (= Invoerbereik. Let op: deze gegevens staan op het eerste werkblad). En verder geef je aan waar je de resultaten afgedrukt wilt zien (= Uitvoerbereik). Zie figuur 1.22. Merk op dat de klassengrenzen door Excel worden gebruikt als bovengrens en dat gewerkt wordt met een indeling *inclusief* die bovengrens. Dat is dus niet conform de gebruikelijke aanpak zoals besproken in subparagraaf 1.2.1. Het resultaat is te zien in figuur 1.23.

Ook aardig is om de faciliteit 'Grafiek maken' aan te klikken. Er verschijnt dan een wat miserabel grafiekje. Door met het pijltje op een hoekpunt van de grafiek te gaan staan kan deze worden uitgerekt, hetgeen leidt tot een wat mooiere collectie kolommen.
We kunnen Excel ook toepassen op de vraagprijzen van de huizen. De gegevens van voorbeeld 1.5 leveren met Excel figuur 1.24 op.
Omdat er hier geen waarnemingen zijn die precies gelijk zijn aan een klassengrens komt deze tabel overeen met tabel 1.6.
Waarschuwing: de procedure 'Histogram' geeft ons helaas niet een grafiek zoals we die kennen onder deze naam. De hier gepresenteerde grafiek is eigenlijk een staafdiagram. Alleen in het speciale geval van een indeling waarbij alle klassen dezelfde breedte hebben is de vorm correct als men dan ook nog de breedte van de tussenruimten van de staven op 0 instelt. Een bijkomend probleem is dat de horizontale as een categorieas is in plaats van een metrische as. We gaan hier verder niet op in.

FIGUUR 1.23 Frequentieverdeling van bouwjaar

Bin	*Frequency*	*Cumulative %*
1910	0	0,00%
1920	1	10,00%
1930	1	20,00%
1940	4	60,00%
1950	1	70,00%
1960	0	70,00%
1970	3	100,00%
1980	0	100,00%
1990	0	100,00%
More	0	100,00%

FIGUUR 1.24 Frequentieverdeling van vraagprijzen

	A	B	C
1		*Frequency*	*Cumulative %*
2	100000	0	0,00%
3	150000	11	9,17%
4	200000	30	34,17%
5	250000	13	45,00%
6	300000	13	55,83%
7	400000	17	70,00%
8	500000	11	79,17%
9	750000	19	95,00%
10	1000000	5	99,17%
11	1250000	1	100,00%
12	More	0	100,00%

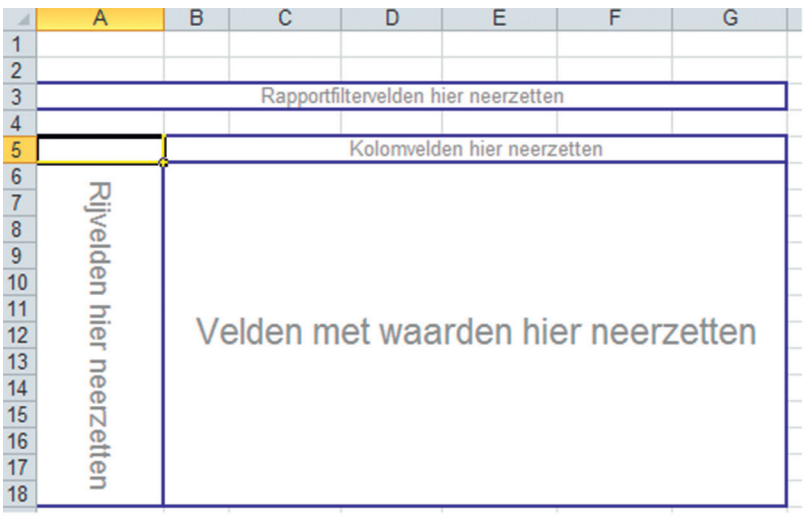
1.8.2 Draaitabellen

De optie 'Draaitabellen' biedt allerlei mogelijkheden om de gegevens te groeperen. We illustreren dat aan de hand van het woningenbestand.

1 Kies op het tabblad 'Invoegen' de knop 'Draaitabel' om de Wizard 'Draaitabellen' te starten.
2 Bij de eerste stap kunnen we de optie 'Een Microsoft Excel-lijst of -database' als keuze laten staan.
3 Selecteer op het blad Woningen de kolommen B t/m I als bereik in de tweede stap.
4 In stap 3 kunnen we de variabele kiezen waarvan een telling wordt gewenst (zie figuur 1.25).
5 Plaats de variabele wijk in het vak *Rij* en ook in het vak *Gegevens*. Het resultaat levert ons een frequentietabel van de variabele wijk (zie figuur 1.26). We zijn dan dus uitsluitend geïnteresseerd in één variabele.
6 Laat in de vierde stap de begincel van de draaitabel op een nieuw blad plaatsen en klik op 'Voltooien' om de draaitabel te maken.
7 Let goed op het hokje linksboven in de tabel. Daar kun je allerlei keuzes maken. Als je kiest voor 'aantal', dan ontstaat een frequentietabel. Maar ook 'gemiddelde' zou een goede optie zijn. Door bijvoorbeeld in de voorkolom te kiezen voor wijk, in de bovenrij voor garage en in het middenveld (bij 'gegevens') voor prijs, kan men een overzicht krijgen van de gemiddelde prijs in de vier wijken, uitgesplitst naar huizen mét en zonder garage. En als je dan vervolgens nog de variabele badkamer sleept bovenop het woord wijk, dan kun je nog verder uitsplitsen.

Met de werkbalk kan men snel allerlei instellingen aanpassen of zelfs opnieuw de wizard activeren om bijvoorbeeld een andere variabele te kiezen of een andere samenvattingsfunctie te kiezen. Dubbelklikken we op de cel wijk in de draaitabel, dan verschijnt het dialoogvenster Draaitabelveld (zie figuur 1.27). Hier kan men onder andere de richting van de draaitabel wijzigen of aangeven dat bepaalde categorieën in de draaitabel verborgen moeten worden.

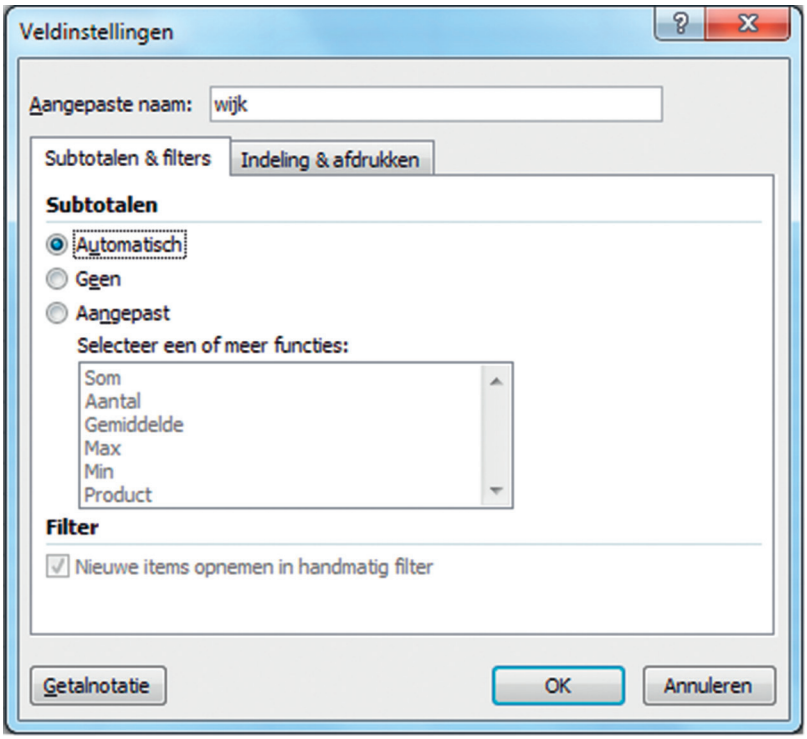
FIGUUR 1.25 Stap 3 van de Wizard Draaitabel



FIGUUR 1.26 Frequentietabel van de variabele wijk

Aantal van nr	
wijk	Totaal
1	28
2	29
3	37
4	26
Eindtotaal	120

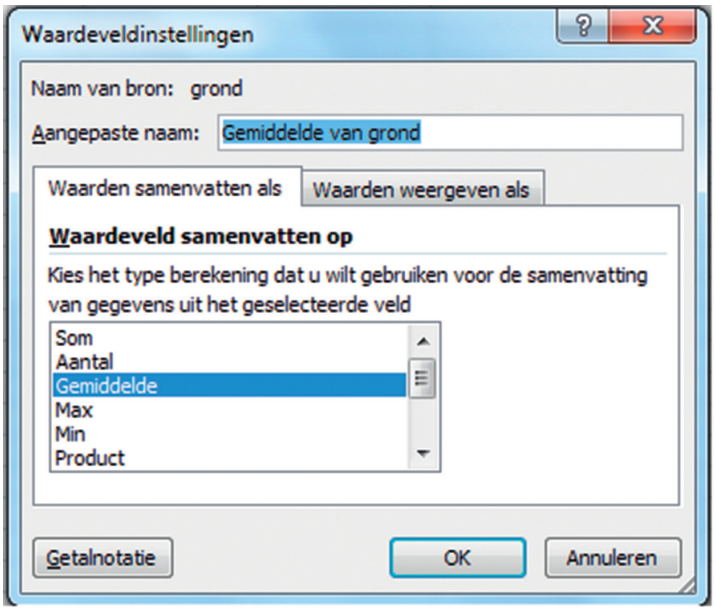
FIGUUR 1.27 Draaitabel



FIGUUR 1.28 Draaitabel

	Gegevens	
wijk	Aantal van nr	Gemiddelde van grond
1	28	288
2	29	316
3	37	566
4	26	930
Eindtotaal	120	520

FIGUUR 1.29 Perceelgrootte per wijk



Om de gemiddelde grondoppervlakte van de aangeboden percelen per wijk te bepalen, starten we opnieuw de wizard met de tweede knop op de werkbalk (zorg dat de cursor ergens in de draaitabel staat!). De wizard laat ons weer stap 3 zien en we plaatsen 'Perceelopp' in het vak 'Gegevens' en dubbelklikken hierop. In het dialoogvenster 'Draaitabelveld', zie figuur 1.29, kunnen we aangeven dat we het bronveld 'Perceelopp' willen samenvatten per gemiddelde. Het resultaat staat in figuur 1.30 weergegeven, waarbij ons opvalt dat de percelen in wijk 4 gemiddeld veel groter zijn dan in wijk 1. (Opmerking: dubbelklikken op een cel in de totaalkolom bij een wijk levert een overzicht van de gegevens van die wijk.)

Door in stap 3 van de wizard draaitabel een variabele in het vak 'Rij' en een variabele in het vak 'Kolom' te plaatsen, krijgen we een kruistabel, zie figuur 1.30.

FIGUUR 1.30 Aantal garages per wijk

Aantal van wijk	wijk				
garage	1	2	3	4	Eindtotaal
0	21	20	17	8	66
1	7	9	20	18	54
Eindtotaal	28	29	37	26	120

Het is ook mogelijk om in de cellen van de tabel een andere variabele en samenvattingsfunctie te kiezen. Zijn we bijvoorbeeld geïnteresseerd in de gemiddelde vraagprijs, dan maken we figuur 1.31.

FIGUUR 1.31 Vraagprijs per wijk afhankelijk van garage

Gemiddelde van prijs	wijk				
garage	1	2	3	4	Eindtotaal
0	178476	243350	251000	422500	246394
1	266143	278222	469575	664778	476380
Eindtotaal	200393	254172	369149	590231	349888

Door een tijdje te 'spelen' met het bestand kan men vertrouwd raken met deze faciliteit.

1+ Lorenzcurve

Lorenzcurve of concentratiecurve

Een bijzonder soort grafiek om een cumulatieve verdeling weer te geven, is de Lorenzcurve of concentratiecurve. Deze curve wordt in de praktijk vooral toegepast om een beeld te geven van de verdeling van inkomens en vermogens over individuen. Ook kan deze curve worden gebruikt als concentratiemaatstaf voor een bedrijfstak. We willen dan in een grafiek tot uitdrukking brengen in welke mate bijvoorbeeld de totale omzet van een bedrijfstak terechtkomt bij de 5%, 10%, . . . grootste bedrijven. In een Lorenzcurve worden percentages tegen elkaar uitgezet. Hierbij worden de waarnemingen geordend van laag naar hoog, waardoor gegevens worden berekend zoals: de 10% gezinnen met de laagste inkomens verdienen 3% van het totale inkomen van alle gezinnen. We illustreren dit met een voorbeeld.

VOORBEELD 1.27

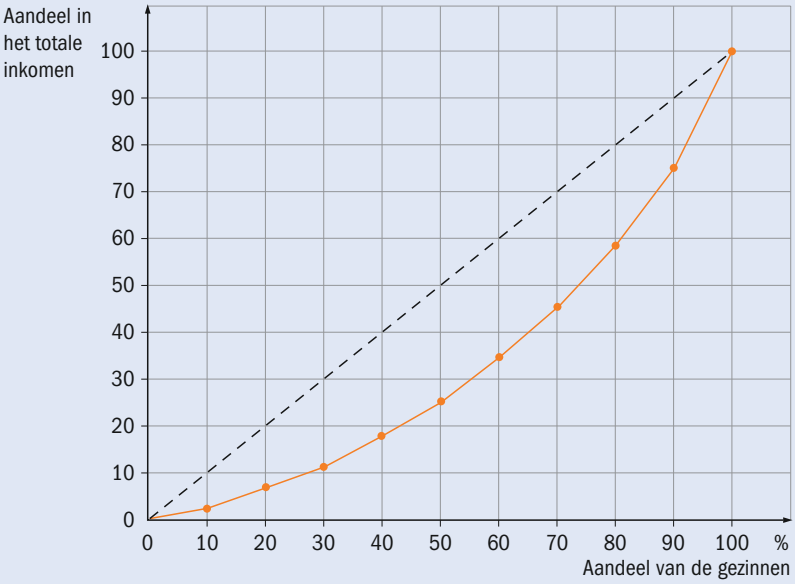
In de gemeente X zijn van 10 000 gezinnen de gezinsinkomens bepaald. Nadat de gezinnen op volgorde van inkomen zijn geplaatst, resulteert tabel 1.27.

TABEL 1.27 Gezinsinkomens in gemeente X

% van de gezinnen	% van het totale inkomen
laagste 10	3
20	7
30	12
40	18
50	25
60	35
70	46
80	59
90	75
100	100

Het resultaat is te zien in de grafiek van figuur 1.32. In de grafiek is de 45°-lijn getekend. Als bij de onderzochte verdeling volstrekte gelijkheid van inkomens zou gelden, dan zou de Lorenzcurve precies samenvallen met deze 45°-lijn. Naarmate de ongelijkheid groter is tussen de uitkomsten, zal de Lorenzcurve dieper 'doorzakken'. De oppervlakte tussen de Lorenzcurve en de diagonaal wordt wel gebruikt als maatstaf voor ongelijkheid.

FIGUUR 1.32 Lorenzcurve van inkomens



Samenvatting

- Bij het beoefenen van statistiek is het vrijwel altijd van levensbelang om op een nauwkeurige wijze de populatie te omschrijven waarnaar men onderzoek doet. Dikwijls onderzoekt men een populatie door er een steekproef uit te trekken. Men kiest vaak voor een aselecte steekproef in de hoop dat men hiermee een goed beeld krijgt van de te onderzoeken populatie.

- In de statistiek spelen variabelen een belangrijke rol. Variabelen kunnen op verschillende manieren worden onderscheiden. We denken hierbij aan het onderscheid:
  - kwantitatieve variabelen – kwalitatieve variabelen
  - deterministische variabelen – kansvariabelen
  - discrete variabelen – continue variabelen

  Variabelen kunnen op verschillende schalen worden gemeten, namelijk de nominale schaal, de ordinale schaal, de intervalschaal en de ratioschaal.

- Belangrijk bij de beschrijvende statistiek is het opzetten van een frequentieverdeling. Hierbij moet op een verantwoorde manier een klassenindeling worden gemaakt.

- Er zijn diverse manieren om een verdeling door middel van een grafiek weer te geven. Belangrijk daarbij is dat zo'n tabel of grafiek te begrijpen is voor buitenstaanders. Door een aantal formele richtlijnen voor de constructie van tabellen en grafieken te hanteren, kan een en ander worden bevorderd.

- Het histogram is wellicht de belangrijkste grafische voorstelling van een frequentieverdeling. Zorgvuldigheid moet worden betracht bij de keuze van de klassengrenzen en er moet op worden gelet dat wordt gewerkt met frequentiedichtheden.