STATISTIEK IN DE PRAKTIJK
Theorieboek

David S. Moore
George P. McCabe

5e herziene druk

ACADEMIC SERVICE

Oorspronkelijke titel: *Introduction to the Practice of Statistics, fifth edition.*
First published in the United States by W.H. Freeman and Company, New York and Basingstoke

Meer informatie over deze en andere uitgaven kunt u verkrijgen bij:
Sdu Klantenservice
Postbus 20014
2500 EA Den Haag
Telefoon (070) 37 89 880
www.sdu.nl/service

1e druk 1994
2e druk 1997
3e druk 2001
5e druk, 1e oplage februari 2006
2e oplage april 2007
3e verbeterde oplage juni 2008

Vertaling: Vertaalbureau Transvorm, Florenza Vertalingen, Carola Bouman, Josefien Bruijn, SietskeTol, TheoTromp, Marc Wiersma
Omslagontwerp: Scherphuis | Snijder BNO
Omslagillustratie: Corbis
Zetwerk: Elvenkind, Dordrecht
Druk- en bindwerk: De Groot Drukkerij, Goudriaan

ISBN 90 395 2360 6
NUR 123 / 916

Beknopte inhoud

Inhoud

Voorwoord

Wij zijn verheugd dat zoveel studenten en docenten ontvankelijk waren voor een leerboek dat gegevens en statistisch redeneren centraal stelt. Deze nieuwe editie is niettemin grondiger herzien dan menig andere nieuwe editie, zonder echter de essentie en de stijl van het boek te wijzigen. In dit voorwoord beschrijven we eerst onze globale filosofie en bespreken daarna de veranderingen die in deze nieuwe editie zijn aangebracht.

Statistiek in de Praktijk is een elementaire maar serieuze inleiding in de moderne statistiek, op hbo- en universitair niveau. Het is elementair wat betreft het niveau van de vereiste wiskundekennis en de behandelde statistische procedures. Het is serieus, omdat het onze bedoeling is de lezers te helpen over gegevens na te denken en de beschreven statistische methoden met inzicht te gebruiken. De studenten hoeven slechts praktische kennis te hebben van eenvoudige algebra; dat wil zeggen, ze moeten in staat zijn formules te lezen en te gebruiken zonder dat elke stap verklaard wordt. Statistiek is interessant en nuttig, omdat het een middel is om met behulp van gegevens inzicht te krijgen in reële problemen. Door de voortgaande revolutie in het rekenen en tekenen met computers, wordt het belangrijker en zinvoller de nadruk te leggen op statistische begrippen, en het vanuit de gegevens te verkrijgen inzicht. Wij hebben veel statistische fouten gezien, maar slechts weinig die eenvoudigweg voortkwamen uit een foutieve berekening. Wij vragen daarom de studenten na te denken over de achtergrond van de gegevens, het ontwerp of de proefopzet waarmee de gegevens werden geproduceerd, de mogelijke effecten van uitzonderlijke waarnemingen op het trekken van conclusies en op de redenering die ten grondslag ligt aan de standaardmethoden van inferentie. Gebruikers van de statistiek die zich van meet af aan deze gewoonten eigen maken, zijn goed voorbereid op het leren en gebruiken van meer geavanceerde methoden.

Gegevens

Gegevens zijn getallen met een context. Het getal 10.3 heeft op zich geen betekenis. Wanneer we te horen krijgen dat de baby van vrienden 10.3 pond weegt na de geboorte, sluit dit getal aan op onze achtergrondkennis en dan krijgt het meteen een betekenis. Omdat de context getallen betekenis geeft, zijn onze voorbeelden en opgaven gebaseerd op realistische gegevens in de context van problemen uit de alledaagse werkelijkheid. Het gemiddelde berekenen van vijf getallen is wiskunde, geen statistiek. We hopen dat studenten altijd bij de betekenis van hun berekeningen stil zullen staan en zich niet zullen beperken tot de berekeningen zelf. Een berekening of een grafiek is zelden het antwoord op een statistisch probleem. We raden studenten van harte aan om altijd een korte samenvatting te geven omtrent het probleem in kwestie. Dit helpt bij het ontwikkelen van een gevoel voor de gegevens en van de communicatieve vaardigheden die werkgevers zo waarderen.

Wiskunde

Hoewel statistiek een wiskundige wetenschap is, is het geen tak van de wiskunde en moet het ook niet als zodanig onderwezen worden. Een vruchtbare wiskundige theorie (gebaseerd

op kansberekening, wat *inderdaad* een deelgebied is van de wiskunde) ligt ten grondslag aan sommige delen van de statistiek, maar lang niet aan alle delen. Het onderscheid tussen waarneming en experiment, bijvoorbeeld, is een puur statistisch idee dat door deze theorie wordt genegeerd. Wiskundig opgeleide docenten, die terecht een formule-gebaseerde theorie verwerpen, identificeren conceptueel begrip vaak met wiskundig begrip. Wanneer we echter statistiek onderwijzen, moeten we de nadruk leggen op statistische ideeën en begrippen en inzien dat wiskunde niet de enige drager van conceptueel begrip is. Voor *Statistiek in de Praktijk* moet men vergelijkingen kunnen lezen en gebruiken zonder elk klein detail uit te hoeven spellen. We maken geen gebruik van algebraïsche afleidingen. Omdat dit een boek is over statistiek, is het rijker aan ideeën en vraagt het om meer diepgang dan het lage wiskundige niveau op het eerste oog doet vermoeden.

Calculators en computers

Statistische berekeningen en grafieken worden in de praktijk op een computer uitgevoerd. We hebben enkele onderwerpen opgenomen die de invloed van software in de praktijk weerspiegelen, zoals de interpretatie van normaal-kwantiel-diagrammen en de twee-steekproeven t-procedures met een benaderend aantal vrijheidsgraden. We raden docenten aan om software naar hun keuze te gebruiken of een calculator met grafiekfuncties en functies voor zowel data-analyse als basis-inferentie. Alle studenten moeten een rekenmachine hebben met de mogelijkheden voor statistische berekeningen met tenminste twee variabelen, met functies voor correlatie en de kleinste-kwadraten-regressielijn, en ook voor het gemiddelde en de standaardafwijking. Hoewel niet alle opgaven door studenten met zo'n rekenmachine uitgevoerd kunnen worden, hebben we ervoor gezorgd dat het boek gemakkelijk bruikbaar blijft voor studenten die niet beschikken over computerfaciliteiten.

Beoordeling in de statistiek

Statistiek in de praktijk vereist beoordeling. Het is gemakkelijk de wiskundige aannames ter rechtvaardiging van het gebruik van een bepaalde procedure op te sommen, maar het is niet altijd gemakkelijk te beslissen wanneer die procedure in de praktijk kan worden toegepast. Omdat oordeelkundigheid voortspruit uit ervaring, moet een inleidende cursus duidelijke richtlijnen geven, en geen onredelijke eisen stellen aan de oordeelkundigheid van studenten. Wij hebben richtlijnen gegeven – bijvoorbeeld over het gebruiken van de t-procedures voor verwachtingen en over het vermijden van de F-procedures voor varianties – die wij zelf ook volgen. Ook enkele opgaven in het Opgavenboek vragen oordeelkundigheid en (net zo belangrijk) de vaardigheid om hun keuze met woorden te kunnen onderbouwen. Veel studenten zullen zich liever beperken tot het berekenen, en veel statistiekboeken zullen hen dat dan ook laten doen. Maar wanneer we nu iets meer van hen vragen, zullen ze daar op de lange termijn profijt van hebben.

Onderwijservaring

We hebben *Statistiek in de Praktijk* met succes gebruikt bij verschillende studentengroepen. Bij eerstejaarsstudenten van verschillende disciplines behandelen we de hoofdstukken 1 tot en met 8, en één van de hoofdstukken 9, 10 en 12, met weglating van alle optionele paragrafen.

Bij tweedejaarsstudenten die zich willen specialiseren in verzekeringen of statistiek, voegen we hier de hoofdstukken 10 en 11 aan toe en behandelen we ook de optionele teksten. Op hoofdstuk 4 leggen we dan wat minder nadruk omdat ze later nog een aparte cursus over kansrekening zullen volgen, en we maken intensief gebruik van software. De derde groep bestaat uit gevorderde studenten in de sociale wetenschappen. Deze studenten lezen de hele tekst (hoofdstukken 11 en 13 wat globaler) maar met weinig nadruk op hoofdstuk 4 en sommige delen van hoofdstuk 5. In alle gevallen geldt dat wanneer studenten beginnen met het verwerven en de analyse van gegevens (Deel 1) ze hun vrees voor statistiek overwinnen en een goede basis kunnen leggen voor de bestudering van inferentie.

De vijfde druk

Het herzien van een succesvolle titel houdt meer in dan het veranderen van het druknummer op het titelblad. Er zijn belangrijke verbeteringen ten opzichte van eerdere drukken.

Nieuwe onderwerpen

Hoofdstuk 14 over bootstrap-methoden en permutatietoetsen maakt hersteekproefmethoden inzichtelijk voor studenten die wiskunde en statistiek niet als hoofdvak hebben. Dit hoofdstuk kan gepresenteerd worden na hoofdstuk 7 waar de student kennismaakt met statistische inferentie en *t*-procedures.

In hoofdstuk 17 over statistiek in de kwaliteitsbeheersing bespreken we procesbeheersing en procesmogelijkheden. De nadruk in dit hoofdstuk ligt op concepten en aandacht voor praktijkproblemen in een vakgebied dat vaak overheerst wordt door talloze voorschriften voor grafieken.

Nieuwe opgaven en voorbeelden

Vrijwel alle opgaven zijn aangepast en voorzien van recente gegevens, zodat docenten en studenten beschikken over interessante en eigentijdse problemen. Ook de voorbeelden zijn voor een groot deel herzien of vervangen door nieuwe voorbeelden uit de praktijk.

Nieuwe onderdelen

Nieuw is dat elk hoofdstuk begint met *Statistiek in de praktijk*, een korte biografische schets van een jonge professional die dagelijks met statistiek werkt. Hierin wordt duidelijk hoe een professional statistiek toepast en waarom de statistiek in diverse vakgebieden zo belangrijk is.

Nieuw zijn ook de passages in het theoriegedeelte, aangegeven met een icoontje, waarin gewaarschuwd wordt voor mogelijke valkuilen bij het analyseren van data.

Nieuw zijn ook bij elk hoofdstuk de speciale uitdagende opgaven, gemarkeerd met een icoontje, die een extra beroep doen op het inzicht van gevorderde studenten. Het karakter van deze opgaven is wisselend, bij sommige komt het aan op wiskunde, sommige vergen enig onderzoek, enz.

Ook de collectie applets die beschikbaar is via www.academicservice.nl is aangepast en uitgebreid. Deze applets, ontwikkeld om van te leren, zijn door het hele boek opgenomen. Het icoon in de marge geeft aan waar een applet beschikbaar is om de stof te verduidelijken.

Dankbetuigingen

We zijn er blij mee dat vorige drukken van *Statistiek in de praktijk* ertoe hebben bijgedragen dat het inleidende statistiekonderwijs een richting uitgaat die door de meeste statistici wordt ondersteund. We bedanken de collega's en studenten voor hun nuttige commentaar en hopen dat zij deze nieuwe uitgave weer een stap vooruit zullen vinden.

Het meest dankbaar zijn we alle mensen uit verschillende disciplines en beroepen met wie we hebben samengewerkt om inzicht in gegevens te krijgen. Zij hebben ons door hun ervaring en gegevensmateriaal in staat gesteld dit boek te schrijven. We zijn ervan overtuigd dat een inleiding in de statistiek zich moet richten op gegevens en concepten, en intellectuele vaardigheden moet aanreiken waarmee men steeds complexere situaties kan aanpakken.

We hopen dat de gebruikers en potentiële gebruikers van statistische technieken hiervan profijt ondervinden.

David. S. Moore

George P. McCabe

Woord vooraf bij de Nederlandse vertaling

Sinds de verschijning in 1994 van de Nederlandse vertaling van *Introduction to the Practice of Statics - second edition*, heeft een gestaag groeiende groep van gebruikers in het onderwijs aan hogescholen en universiteiten kennis kunnen maken met een, naar Nederlandse begrippen, vernieuwende aanpak van het vak statistiek. Met name vernieuwend door de praktijkgerichte aanpak van de onderwerpen in dit boek. Los van de verschijning van dit boek is er nu ook in ons land in het wiskunde- en statistiekonderwijs een duidelijke ontwikkeling te bespeuren om deze vakken beter aan te laten sluiten op werkelijke gegevens en situaties uit de praktijk. We zijn er dan ook van overtuigd dat de gebruikersgroep van *Statistiek in de Praktijk* de komende jaren nog aanzienlijk zal toenemen.

In deze druk is een groot aantal verbeteringen aangebracht in de tekst en de presentatie van de leerstof. Ook zijn er nieuwe voorbeelden toegevoegd met een directe link naar de praktijk. Bovendien wordt er gebruik gemaakt van verschillende statistische software, zoals Minitab, Excel en Data Desk, waardoor studenten ook hiermee vertrouwd zullen raken.

Tenslotte willen we drs. Rob Flohr (Stenden Hogeschool Leeuwarden) bedanken voor de verbeteringen die hij in de Nederlandse tekst heeft aangebracht.

Den Haag, de uitgever

Wat is statistiek?

Statistiek is de wetenschap van het verzamelen, ordenen en interpreteren van numerieke feiten, die *gegevens* of *data* worden genoemd. In het dagelijks leven worden wij overstelpt met gegevens. De meeste mensen associëren 'statistiek' met de brokken gegevens die in het nieuws komen: doelpuntgemiddelden bij voetbal, aantallen verkochte geïmporteerde auto's, de laatste opiniepeiling over de steun voor politieke partijen, de gemiddelde hoogste temperatuur voor het seizoen, enzovoort. Advertenties dragen vaak gegevens aan om de superioriteit van het geadverteerde product aan te tonen. Alle deelnemers aan openbare discussies over de economie, het onderwijs en de sociale zorg redeneren vanuit gegevens. Maar het nut van de statistiek strekt zich veel verder uit dan deze alledaagse voorbeelden.

Het Centraal Bureau voor de Statistiek in Nederland en het Nationaal Instituut voor de Statistiek in België publiceren bijvoorbeeld maandelijks de nieuwste cijfers over werkloosheid en inflatie. Economen, financiële analisten en beleidsmakers bij de overheid en in het bedrijfsleven bestuderen deze gegevens ter ondersteuning van de besluitvorming. Artsen moeten de oorsprong en betrouwbaarheid van de in medische tijdschriften gepubliceerde data begrijpen, willen zij hun patiënten de meest effectieve behandeling kunnen bieden. Politici gaan af op uitslagen van opiniepeilingen. Bedrijfsstrategieën zijn gebaseerd op marktonderzoek naar consumentenvoorkeur. Landbouwers bestuderen de gegevens over nieuwe variëteiten op proefvelden. Ingenieurs verzamelen gegevens over kwaliteit en betrouwbaarheid van materialen en producten. De meeste wetenschappers maken gebruik van getallen, en daarmee ook van statistische methoden.

Net zomin als we ons kunnen onttrekken aan het gebruik van woorden, kunnen we dat aan gegevens. Zoals woorden op een bladzijde voor een analfabeet zonder betekenis zijn – of verwarrend voor wie weinig opleiding heeft gehad – zo ook interpreteren data niet zichzelf, maar moeten ze worden gelezen met kennis van zaken. Net als de auteur die zijn woorden kan rangschikken tot het vormen van overtuigende argumenten, of juist een onsamenhangend kletsverhaal, zo ook kunnen data dwingend zijn, of misleidend, of gewoonweg irrelevant. Numeriek geletterd zijn, de bekwaamheid hebben om numerieke argumenten te begrijpen, is voor iedereen belangrijk. In veel beroepen en vakgebieden is het van essentieel belang om zich numeriek te kunnen uitdrukken – auteur te zijn in plaats van slechts lezer. Voor een gedegen opleiding is studie van de statistiek daarom wezenlijk. Wij moeten leren gegevens kritisch en met inzicht te lezen; wij moeten leren gegevens te produceren die duidelijke antwoorden geven op belangrijke vragen; en wij moeten welgefundeerde methoden leren voor het trekken van betrouwbare conclusies op grond van gegevens.

De opkomst van de statistiek

De vroegste oorsprong van de statistiek ligt in de behoefte van heersers om volkstellingen te houden en de belastingsom voor pachters te bepalen. Toen de natuurwetenschappen in de zeventiende en achttiende eeuw sterk in ontwikkeling waren, groeide het belang van zorgvuldige metingen van gewichten, afstanden en andere fysische grootheden. Astronomen

en landmeters kregen in hun streven naar precisie te maken met variaties in hun metingen. Veel metingen moeten wel beter zijn dan één enkele, ook al variëren zij onderling. Hoe kunnen we veel variërende waarnemingen het beste combineren? Voor het analyseren van wetenschappelijke metingen werden toen statistische methoden uitgevonden die nu nog steeds van belang zijn.

In de negentiende eeuw begonnen ook de landbouwwetenschappen en de sociale en gedragswetenschappen gegevens te gebruiken om fundamentele vragen te beantwoorden. Welke relatie bestaat er tussen de lengtes van ouders en die van hun kinderen? Levert een nieuwe variëteit graan een hogere opbrengst dan de oude, en zo ja, onder welke voorwaarden voor regenval en bemesting? Kan iemands IQ of gedrag worden gemeten, zoals we ook lengte of reactietijd meten? Effectieve methoden voor de behandeling van dergelijke vraagstukken kwamen slechts langzaam tot ontwikkeling, en niet zonder controverses.[1]

Terwijl de methoden voor het produceren en begrijpen van data talrijker en verfijnder werden, kreeg het nieuwe vak statistiek in de twintigste eeuw vorm. Ideeën en technieken die waren ontstaan bij het verzamelen van overheidsgegevens, bij de bestudering van astronomische en biologische metingen en bij pogingen erfelijkheid en intelligentie te begrijpen, kwamen samen en vormden een geünificeerde 'wetenschap van gegevens'. Deze wetenschap van gegevens – de statistiek – is het onderwerp van dit leerboek.

De opzet van dit boek

Deel 1 van dit boek, 'Gegevens', gaat over het verwerven en analyseren van gegevens. De eerste twee hoofdstukken handelen over statistische methoden voor het ordenen en beschrijven van gegevens. Deze hoofdstukken gaan van eenvoudige naar meer gecompliceerde gegevens. Hoofdstuk 1 onderzoekt gegevens van één enkele variabele, terwijl hoofdstuk 2 is gewijd aan relaties tussen twee of meer variabelen. We leren hoe we door anderen geproduceerde gegevens moeten onderzoeken, en hoe we eigen gegevens moeten ordenen en samenvatten. Deze samenvattingen zijn eerst grafisch en daarna numeriek, en krijgen waar nodig de vorm van een wiskundig model dat een compacte beschrijving geeft van het globale patroon van de gegevens. Hoofdstuk 3 schetst schema's ('ontwerpen' genoemd) voor het verkrijgen van gegevens ter beantwoording van specifieke vragen. De in dit hoofdstuk gepresenteerde principes zullen u helpen bij het ontwerpen en evalueren van geschikte steekproeven en experimenten.

Deel 2, dat bestaat uit hoofstukken 4 tot en met 8, introduceert de statistische inferentie – formele methoden voor het trekken van conclusies uit correct geproduceerde gegevens. De statistische inferentie gebruikt de taal van de kansrekening om te beschrijven hoe betrouwbaar de conclusies zijn – daarom zijn voor het begrijpen van inferentie enkele basisfeiten uit de kansrekening noodzakelijk. Kansrekening vormt het onderwerp van de hoofdstukken 4 en 5. Hoofdstuk 6, wellicht het belangrijkste hoofdstuk uit het boek, introduceert de redenering achter statistische inferentie. We benadrukken dat effectieve inferentie berust op goede procedures voor het produceren van gegevens (hoofdstuk 3), zorgvuldige bestudering van de gegevens (hoofdstukken 1 en 2) en het begrijpen van de aard van de statistische inferentie, zoals besproken in hoofdstuk 6. De hoofdstukken 7 en 8 beschrijven enkele van de meest gangbare specifieke methoden van inferentie: voor het trekken van conclusies omtrent gemiddelden en fracties uit één en twee steekproeven.

De vijf korte hoofdstukken in Deel 3 behandelen de wat complexere inferentiemethoden omtrent relaties tussen kwalitatieve variabelen, regressie en correlatie, en de variantie-analyse.

Deel 4 bevat aanvullende onderwerpen, waaronder twee geheel nieuwe hoofdstukken: hoofstuk 14 over bootstrapmethoden en hoofdstuk 17 over statistiek in de kwaliteitszorg.

Inzicht verwerven uit gegevens

De praktijk van de statistiek kent vele recepten voor numerieke berekeningen, sommige heel eenvoudig, andere heel ingewikkeld. Terwijl u leert deze recepten te gebruiken, moet u in gedachten houden dat het doel van de statistiek niet het berekenen op zich is, maar het verwerven van inzicht vanuit getallen. Veel van de berekeningen kunnen worden geautomatiseerd met een rekenmachine of een computer, het begrip moet u zelf aanleveren. De hoofdstukken 7 tot en met 13 brengen slechts enkele van de vele specifieke inferentieprocedures. De meer ingewikkelde procedures worden steeds door computers uitgevoerd met gespecialiseerde software. Een grondig begrip van de principes van de statistiek stelt u in staat waar nodig gevorderde methoden te leren. Aan de andere kant zal een fraai ogende computeranalyse, uitgevoerd zonder aandacht voor de basisprincipes, vaak pure onzin opleveren. Probeer bij het lezen van dit boek de principes te begrijpen, evenals de noodzakelijke details van de methoden en recepten.

Noot

1. De opkomst van de statistiek vanuit de natuurwetenschappen en de sociale en gedragswetenschappen wordt in detail besproken door S.M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard-Belknap, Cambridge, Mass. 1986. Veel informatie in de korte historische noten die verspreid in de tekst voorkomen, is aan deze bron ontleend.

Deel 1
Gegevens

Statistiek in de Praktijk

Marktonderzoek: een stem van de consument

JENNIFER KARAS *is manager marktonderzoek bij Shell Oil Products in Houston in de Verenigde Staten.*

Allemaal zijn we consumenten, want we nemen dagelijks beslissingen over wat we wel of niet aanschaffen en bij welke bedrijven we klant zijn. Hoe komen we als consumenten tot onze beslissing? Kiezen we allemaal op basis van dezelfde motivatie en verwachtingen? Als marktonderzoeker hanteer ik statistieken en statistische technieken om te begrijpen wat de consument beweegt. Met dit inzicht draag ik bij aan de besluitvorming van mijn onderneming over de vraag welke soorten benzine en welke vormen van dienstverlening onze tankstations moeten bieden en hoe we deze producten en diensten onder de aandacht van het publiek brengen.

Marktonderzoekers maken gebruik van statistiek om consumenten in slechts enkele groepen onder te verdelen op grond van vergelijkbare koopwensen en koopgedrag. Zo zijn er bijvoorbeeld zeven hoofdgroepen benzineverbruikers, mensen die uitsluitend afgaan op de prijs of alleen op de brandstofkwaliteit of (ook) op snelheid en gemak. Door deze segmentering van consumenten kunnen ondernemingen bepalen op welke groepen consumenten zij zich richten. Daartoe ontwikkelen ze producten en diensten die nauw aansluiten op de behoeften van die consumenten. Ook maken ze reclame die deze specifieke doelgroepen aanspreekt.

Marktonderzoekers houden zich eveneens op de hoogte van trends op het gebied van opvattingen en gedrag van het publiek. Dagelijks houden we in het hele land marktonderzoek via telefonisch, schriftelijk of persoonlijk contact. Deze onderzoeken geven ons een beeld van het oordeel van het publiek over de kwaliteit van onze prestaties in vergelijking met die van onze concurrentie. Met behulp van statistische technieken analyseer ik de gegevens van deze onderzoeken om inzicht te krijgen in wat de consument beweegt. Dit inzicht vormt de basis voor beslissingen van Shell over de te ontwikkelen soorten brandstof of over de vraag hoe het tankstation van de toekomst eruit gaat zien. Moeten we bij u langs de deur komen voor een tankbeurt? Wilt u onbemande tankstations? Aan u de keus! U bent de consument en door mijn statistische werkzaamheden hebt u een stem in mijn onderneming.

1 Kijken naar gegevens – verdelingen

Inleiding

Statistiek is de wetenschap van de kennisverwerving op basis van gegevens. Gegevens zijn numerieke feiten. Feiten vormen een betere basis voor besluitvorming dan eenvoudigweg raden. Dit wordt geïllustreerd door het volgende voorbeeld.

> **Voorbeeld 1.1**
> De omvang van de pensioenbetalingen van een onderneming aan gepensioneerde medewerkers heeft grote invloed op de financiële situatie van die onderneming. Immers, er moet kapitaal opzij worden gezet om de toekomstige betalingen te dekken. Als de medewerkers op latere leeftijd met pensioen gaan, hoeft er minder geld opzij te worden gelegd. Dat komt omdat het fonds meer jaren kan groeien voordat het tot uitbetalen komt. US Airways veronderstelde dat de piloten met pensioen zouden gaan als ze 60 jaar werden, de leeftijdsgrens voor piloten. Dat was een aanname. Toen de luchtvaartmaatschappij failliet ging, werd met een snelle blik op de gegevens duidelijk dat meer dan de helft van de piloten vervroegd met pensioen was gegaan. Het verschil tussen de aanname en de feitelijke gegevens vormde een flinke schadepost voor de piloten en de Amerikaanse regering (die garant staat voor de pensioenen van failliete ondernemingen).[1]

Voor het onderzoeken van gegevens hebben we meer nodig dan alleen maar getallen. Zo hebben de uitkomsten van een medisch onderzoek weinig betekenis als het doel van dat onderzoek onbekend is en het onduidelijk is in hoeverre bloeddruk, hartslag en andere metingen een rol spelen. Dat wil zeggen, *gegevens zijn getallen binnen een bepaalde context.* Pas als we inzicht hebben in de context kunnen we uit de getallen zinvolle conclusies trekken. Anderzijds zijn onderzoeksmetingen van honderden proefpersonen pas waardevol wanneer ze met statistische hulpmiddelen zijn gerangschikt, weergegeven en samengevat; dit ongeacht de mate van deskundigheid van de medicus die deze gegevens bestudeert. We beginnen onze studie van de statistiek met het ons eigen maken van de kunst van het gegevensonderzoek.

Variabelen

Elke gegevensverzameling bevat informatie over een bepaalde groep *elementen*. De informatie wordt georganiseerd in variabelen.

ELEMENTEN EN VARIABELEN

Elemenenten zijn de objecten die beschreven worden door een gegevensverzameling. Elementen kunnen mensen zijn, maar ook dieren of dingen.
Een **variabele** is een eigenschap van een element. Een variabele kan verschillende waarden aannemen voor verschillende elementen.

Een database van een universiteit met gegevens over studenten, bijvoorbeeld, bevat gegevens over elke ingeschreven student. De studenten zijn de elementen die beschreven worden door de gegevensverzameling. Voor elk element bevatten de gegevens waarden van variabelen zoals de geboortedatum, het geslacht (vrouwelijk of mannelijk), de afstudeerrichting en het cijfergemiddelde. In de praktijk wordt elke gegevensverzameling vergezeld van achtergrondinformatie om de gegevens beter te kunnen begrijpen. Als je een statistisch onderzoek wilt uitvoeren of de gegevens van het werk van iemand anders wilt bestuderen, stel dan bij jezelf de vraag: Waarom? Wie? Wat?

1. **Waarom? Voor welk doel** dienen de gegevens? Proberen we specifieke vragen te beantwoorden? Willen we andere conclusies trekken uit variabelen dan die waar we daadwerkelijk gegevens voor hebben?
2. **Wie?** Welke **elementen** worden door de gegevens beschreven? Op **hoeveel elementen** zijn de gegevens van toepassing?
3. **Wat?** Hoeveel **variabelen** bevatten de gegevens? Wat zijn de **exacte definities** van deze variabelen? In welke **meeteenheden** worden de variabelen uitgedrukt? Gewichten, bijvoorbeeld, kunnen in grammen, ponden of duizenden kilogrammen uitgedrukt worden.

Sommige variabelen, zoals geslacht en afstudeerrichting, delen elementen simpelweg in categorieën in. Andere, zoals hoogte en cijfergemiddelde, nemen numerieke waarden aan waarmee we rekenkundige bewerkingen kunnen uitvoeren.

Het is zinvol een gemiddeld inkomen te berekenen van de werknemers van een bedrijf, het is echter niet zinvol om het 'gemiddelde' geslacht te berekenen. We kunnen echter wel het aantal vrouwelijke en mannelijke medewerkers tellen en daarmee rekenkundige bewerkingen uitvoeren.

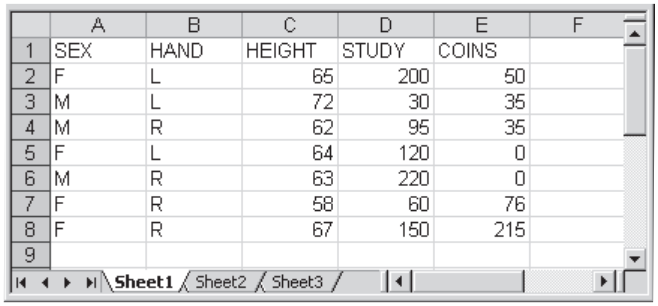
KWALITATIEVE EN KWANTITATIEVE VARIABELEN

Een **kwalitatieve variabele** plaatst een element in één of meer groepen of categorieën.
Een **kwantitatieve variabele** neemt numerieke waarden aan, waarvoor rekenkundige bewerkingen, zoals aftrekken en het bepalen van het gemiddelde, zinvol zijn.
De **verdeling** van een variabele vertelt ons welke waarden aangenomen worden en hoe vaak deze waarden aangenomen worden.

Voorbeeld 1.2

Hier ziet u een klein deel van de gegevensverzameling die betrekking heeft op een grote statistische klasse studenten. De gegevens zijn afkomstig van anonieme antwoorden op een vragenlijst die aan een jaargroep werd voorgelegd.

Elke rij toont de gegevens van een element. Vaak noemen we een rij gegevens een **geval**. Elke kolom bevat de waarden van een variabele voor alle elementen. Er zijn 5 variabelen.



	A	B	C	D	E	F
1	SEX	HAND	HEIGHT	STUDY	COINS	
2	F	L	65	200	50	
3	M	L	72	30	35	
4	M	R	62	95	35	
5	F	L	64	120	0	
6	M	R	63	220	0	
7	F	R	58	60	76	
8	F	R	67	150	215	
9						

Sheet1 Sheet2 Sheet3

Geslacht (mannelijk of vrouwelijk) en links- of rechtshandigheid vormen de kwalitatieve variabelen. De drie resterende variabelen zijn kwantitatief: lengte in inches, het aantal minuten dat op een gewone doordeweekse avond aan de studie wordt besteed. Ook werd gevraagd naar het bedrag aan munten (geen bankbiljetten) dat de student op zak heeft.

De meeste statistische software gebruikt dit formaat om gegevens in te voeren: elke rij betreft een element en elke kolom een variabele. Een dergelijke gegevensverzameling vindt u in een **spreadsheet**. Daarin staan de rijen en kolommen al klaar voor gebruik. Spreadsheets worden doorgaans gebruikt om gegevens in te voeren en te bewerken. De meeste statistische software kan gegevens van spreadsheetprogramma's lezen. Kennis van de context van de gegevens draagt bij aan het begrip van deze data. Deze studenten waren respondenten op een vragenlijst. Een student stelde dat hij op een gewone avond 30.000 minuten studeerde. We weten zeker dat dit een grap is.

Metingen: ken uw variabelen

Tot de context van de gegevens behoort inzicht in de vastgelegde variabelen. Vaak zijn de variabelen in een statistisch onderzoek eenvoudig te begrijpen: hoogte in centimeters, studietijd in minuten enzovoorts. Elk werkterrein kent zijn eigen bijzondere variabelen. Een psycholoog maakt bijvoorbeeld gebruik van de Minnesota Multiphasic Personality Inventory (MMPI). Een deskundige op het gebied van lichamelijke conditie meet 'VO2 max', de hoeveelheid zuurstofverbruik bij training op maximaal vermogen. Beide variabelen worden gemeten met speciale **instrumenten**. VO2 max wordt gemeten tijdens een training waarbij de proefpersoon ademt via een mondstuk dat is verbonden aan een meetapparaat voor het zuurstofverbruik. De scores op de MMPI zijn gebaseerd op een lange vragenlijst die zelf ook als instrument kan worden beschouwd. Deskundigheid op een bepaald werkterrein vereist inzicht in de vraag welke variabelen belangrijk zijn en hoe deze het beste zijn te meten. Aangezien de achtergrond van bepaalde metingen doorgaans een grondige kennis van het betreffende onderzoeksterrein vergt, treden we daarover niet in details.

Zorg ervoor dat elke variabele werkelijk meet wat u wilt dat hij meet. Een ongelukkige keuze van variabelen kan leiden tot bedrieglijke conclusies. Vaak is het **relatieve aantal** keren (percentage) dat iets voorkomt van meer betekenis dan een eenvoudige optelling van die gebeurtenissen.

Voorbeeld 1.3

Volgens een registratiesysteem van fatale ongelukken in de VS waren er in 2002 in totaal 27.102 passagiers betrokken bij ongelukken met een dodelijke afloop.[2] In dat jaar waren slechts 3339 motoren betrokken bij fatale ongelukken. Betekent dit dat motoren veiliger zijn dan auto's? Helemaal niet: er zijn heel wat meer auto's dan motoren. Daarom verwachten we dat er meer auto's zijn betrokken bij dodelijke ongelukken.

Een betere meting van de gevaren van het gemotoriseerde verkeer is het aantal fatale ongelukken te delen door het aantal voertuigen op de weg. U krijgt dan het relatieve aantal. In 2002 raakten per 100.000 geregistreerde motorvoertuigen er 21 passagiersauto's betrokken bij een dodelijk ongeval. Per 100.000 geregistreerde motoren waren er ongeveer 67 ongelukken met fatale afloop. Het percentage ongelukken met dodelijke afloop ligt bij motoren meer dan drie keer zo hoog als bij auto's. Zoals we zouden kunnen verwachten, zijn motoren veel gevaarlijker dan auto's.

1.1 Weergeven van verdelingen met grafieken

Dankzij statistische hulpmiddelen en ideeën kunnen wij gegevens onderzoeken om hun belangrijkste kenmerken te kunnen beschrijven. Dit onderzoek wordt **exploratieve data-analyse** genoemd. Net als een verkenner die onbekend terrein bewandelt, willen we eerst eenvoudigweg beschrijven wat we zien. Er zijn twee basisstrategieën die ons helpen bij de ordening van onze verkenning van een gegevensverzameling:

- Begin met elke variabele op zich. Onderzoek vervolgens de onderlinge verbanden tussen de variabelen.
- Begin met een diagram of diagrammen. Voeg vervolgens de numerieke samenvattingen toe van de specifieke aspecten van de gegevens.

We zullen deze principes steeds volgen tijdens onze studie. In dit hoofdstuk komen methoden aan de orde voor het beschrijven van een enkele variabele. De verbanden tussen verschillende variabelen worden in hoofdstuk 2 beschreven. Elk hoofdstuk begint met grafische afbeeldingen waaraan vervolgens numerieke samenvattingen worden toegevoegd voor een completere beschrijving.

1.1.1 Grafieken voor kwalitatieve variabelen

De waarden voor een kwalitatieve variabele zijn labels voor de categorieën, zoals 'mannelijk' en 'vrouwelijk'. Bij de **verdeling** van een kwalitatieve variabele worden de categorieën gerangschikt in een lijst en wordt het **aantal** of het **percentage** elementen weergegeven die in elke categorie vallen.

Bij wijze van voorbeeld geven we aan hoe goed volwassenen van rond de dertig jaar zijn opgeleid. Hier volgt de verdeling naar het hoogste onderwijsniveau van mensen van 25 tot 34 jaar in de VS.[3]

Opleiding	Aantal (miljoen)	Percentage
Minder dan middelbare school	4,6	11,8
Met diploma middelbare school	11,6	30,6
Vervolgopleiding	7,4	19,5
Kort HBO	3,3	8,8
Bachelor	8,6	22,7
Master	2,5	6,6

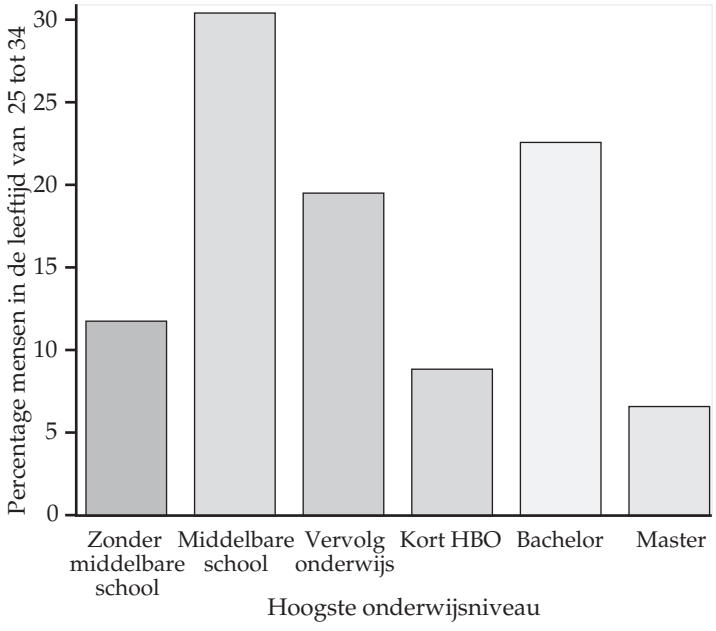
Verbaast het u dat slechts 29,3% van deze jongvolwassenen over minimaal een bachelorsgraad beschikt?

U ziet deze gegevens in de diagrammen van figuur 1.1. Het **staafdiagram** in figuur 1.1 (a) laat op een snelle manier de omvang zien van de zes categorieën onderwijsniveaus. De hoogte van de staven geeft de percentages van de zes categorieën weer. Aan het **taartdiagram** in figuur 1.1 (b) kunnen we zien hoe groot elke groep is ten opzichte van het geheel. Zo beslaat de taartpunt 'bachelor' 22,7% van de taart, omdat 22,7% van de jongvolwassenen niet verder is gekomen dan een bachelorsgraad. We hebben die taartpunt uitgelicht om er de aandacht op te vestigen. Omdat taartdiagrammen geen schaalverdeling kennen, hebben we aan de labels van de taartpunten percentages toegevoegd. *Bij taartdiagrammen is het noodzakelijk dat u alle categorieën, die samen het geheel vormen, opneemt. Gebruik taartdiagrammen alleen als u de verhouding van elke categorie tot het geheel wilt benadrukken.* Staafdiagrammen laten zich eenvoudiger lezen en hebben ook meer mogelijkheden. Met een staafdiagram kunt u bijvoorbeeld een vergelijking maken van de aantallen studenten biologie, bedrijfskunde of politieke wetenschappen aan uw instituut. Met een taartdiagram kunt u een dergelijke vergelijking niet maken, omdat niet alle studenten voor een van deze drie hoofdvakken hebben gekozen.

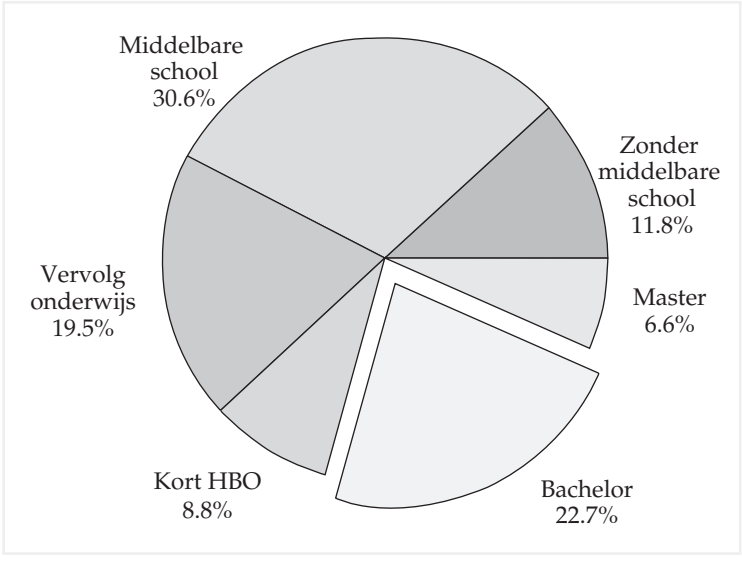
Staaf- en taartdiagrammen helpen de gebruiker om snel een inzicht te krijgen in de verdeling van de variabelen. Ze zijn echter van beperkt nut voor de gegevensanalyse, omdat kwalitatieve gegevens over een enkele variabele, zoals hoogste onderwijsniveau, ook zonder een grafiek eenvoudig zijn te begrijpen. We gaan nu verder met kwantitatieve variabelen. Daarbij vormen grafieken belangrijke hulpmiddelen.

1.1.2 Gegevensanalyse in actie: blijft u even aan de lijn

Veel ondernemingen maken gebruik van klantenservice voor klanten die een bestelling willen plaatsen of op zoek zijn naar informatie. Klanten willen dat hun verzoek zorgvuldig wordt afgehandeld. Ondernemingen willen hun klanten graag goed behandelen, maar ze willen ook voorkomen dat er tijd aan de telefoon wordt verspild. Daarom meten ze de duur van telefoongesprekken en moedigen ze de medewerkers aan die gesprekken kort te houden. Hier volgt een voorbeeld van hoe dit beleid problemen kan veroorzaken.



(a)



(b)

Figuur 1.1 (a) staafdiagram van het voltooide onderwijs van mensen in de leeftijd van 25 tot 34 jaar, (b) taartdiagram van de opleidingsgegevens, met nadruk op de bezitters van een bachelorsgraad.

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

Tabel 1.1 Gespreksduur (seconden) van telefoontjes naar een klantenservice

Voorbeeld 1.4

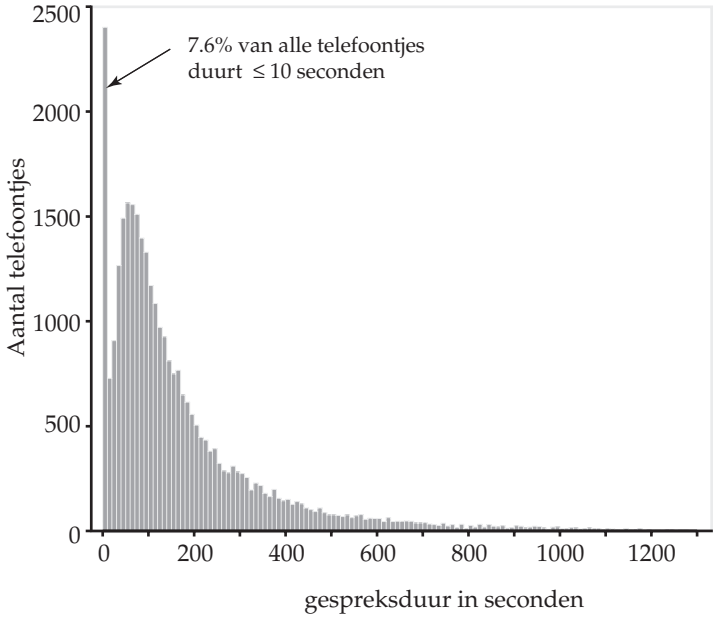
Van een bepaalde maand beschikken we over gegevens over de duur van alle 31.492 telefoontjes naar de klantenservice van een kleine bank. Tabel 1.1 geeft de duur van de eerste 80 telefoontjes weer. De complete gegevensverzameling heet eg01-004.dat. Die vindt u op de website.[4]

Kijk eens bij de gegevens in tabel 1.1. Getallen zeggen niets zonder de achtergrondinformatie. De individuen zijn de telefoontjes die met de klantenservice van de bank zijn gevoerd. De genoteerde variabele is de duur van ieder gesprek. De eenheden zijn seconden. We zien dat de belduur sterk varieert. Het langste gesprek duurde 2631 seconden, dat is bijna 44 minuten. Opvallender nog is dat 8 van de 80 telefoontjes korter dan 10 seconden duurden. Wat is er aan de hand?

Figuur 1.2 is een histogram van de duur van alle 31.492 telefoontjes. Enkele waargenomen gesprekken duurden langer dan 1200 seconden (20 minuten). Die hebben wij niet in de grafiek opgenomen. Zoals verwacht, laat de grafiek zien dat de meeste telefoontjes tussen de 1 en de 5 minuten duren. Enkele duren veel langer als de klant een ingewikkelde vraag heeft. Wat vooral opvalt, is dat 7,6% van alle telefoontjes niet langer duurt dan 10 seconden. Het bleek dat de bank zijn medewerkers erop aansprak als hun gemiddelde belduur te lang was. Daarom hingen sommige medewerkers zomaar op als een klant belde, alleen maar om de gemiddelde belduur omlaag te krijgen. Zowel de klanten als de bank waren daar niet gelukkig mee, dus veranderde de bank zijn beleid. Uit latere gegevens bleek dat er vrijwel geen telefoontjes van minder dan 10 seconden meer voorkwamen. Ons onderzoek van de gegevens van de klantenservice illustreert enkele belangrijke uitgangspunten.

- Begrijpt u eenmaal de achtergrond van uw gegevens (elementen, variabelen, meeteenheid), begin dan doorgaans met het **grafisch weergeven van uw gegevens**.
- Kijkt u naar een grafische voorstelling, let dan op het **globale patroon** en op **opvallende afwijkingen** daarvan. Het globale patroon in figuur 1.2 bestaat uit veel telefoontjes van gemiddelde lengte. De lange rechterstaart geeft de langere gesprekken weer. Wat opvalt, is het verrassend aantal zeer korte gesprekken.

We gaan nu verder met de grafische voorstellingen die worden gebruikt voor de beschrijving van verdelingen van kwantitatieve variabelen. We leggen uit hoe u die met de hand kunt construeren, omdat dit u inzicht verschaft in wat de diagrammen laten zien. Echter, het handmatig vervaardigen van diagrammen is zo tijdrovend dat u bijna altijd software nodig hebt voor een effectieve gegevensanalyse, tenzij het slechts om enkele waarnemingen gaat.



Figuur 1.2 De verdeling van de duur van 31.492 telefoontjes naar de klantenservice van een bank (voorbeeld 1.4). De gegevens tonen een verrassend groot aantal bijzonder korte gesprekken. Dat komt vooral omdat medewerkers bewust ophingen om de gemiddelde belduur omlaag te brengen.

1.1.3 Stamdiagrammen

Een stamdiagram (ook wel stam-en-blad diagram genoemd) biedt een snelle manier om de vorm van een verdeling in beeld te brengen, terwijl de feitelijke numerieke waarden in de grafiek worden opgenomen. Stamdiagrammen werken het best voor een gering aantal waarnemingen, alle met waarden groter dan 0.

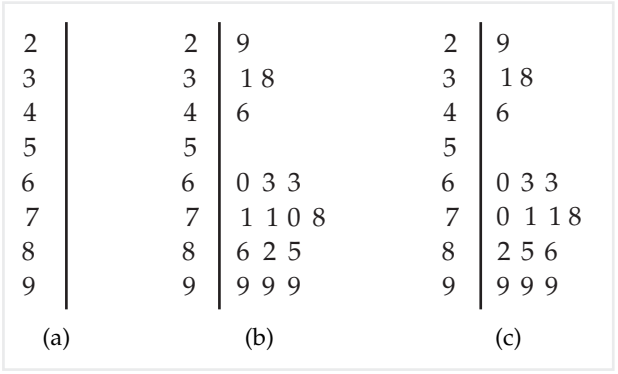
STAMDIAGRAM

U vervaardigt een **stamdiagram** als volgt:

1. Verdeel elke waarneming in een **stam** die bestaat uit alle cijfers behalve de laatste (uiterst rechtse) en een **blad** met het laatste cijfer. Stammen mogen zoveel getallen bevatten als nodig is, maar elk blad bevat slechts een enkel cijfer.
2. Plaats de stammen in oplopende volgorde in een verticale lijst met de kleinste bovenaan. Trek dan een verticale streep aan de rechterkant van deze kolom.
3. Rangschik de bladeren van een rij in oplopende volgorde van links naar rechts vanaf de stam.

Land	Percentage vrouwen	Percentage mannen	Land	Percentage vrouwen	Percentage mannen
Algerije	60	78	Marokko	38	68
Bangladesh	31	50	Saudi-Arabië	70	84
Egypte	46	68	Syrië	63	89
Iran	71	85	Tadzjikistan	99	100
Jordanië	86	96	Tunesië	63	83
Kazachstan	99	100	Turkije	78	94
Libanon	82	95	Oezbekistan	99	100
Libië	71	92	Jemen	29	70
Maleisië	85	92			

Tabel 1.2 Lees- en schrijfvaardigheid in percentages van de bevolking in islamitische landen



Figuur 1.3 Het maken van een stamdiagram van de gegevens uit voorbeeld 1.5. (a) Schrijf de stammen uit, (b) Loop de gegevens door en schrijf elk blad aan de juiste stam. Bijvoorbeeld: de waarden aan stam 8 zijn in de volgorde van de tabel 86, 82 en 85. (c) Rangschik de bladeren aan elke stam in oplopende volgorde vanaf de stam. Stam 8 heeft nu als bladeren 2, 5 en 6.

Voorbeeld 1.5

De islamitische wereld trekt steeds meer de aandacht van Europa en Noord-Amerika. Tabel 1.2 toont van de belangrijkste islamitische landen de percentages mannen en vrouwen van minimaal 15 jaar oud die in 2002 konden lezen en schrijven. We sloegen de landen over met een bevolking van minder dan 3 miljoen. Voor enkele landen, zoals Afghanistan en Irak, waren geen gegevens beschikbaar.[5]

Voor een stamdiagram van de percentages vrouwen die kunnen lezen, gebruikt u het eerste cijfer voor de stam en het tweede cijfer voor het blad. Zo is in Algerije 60% van de bevolking alfabeet. Dit is zichtbaar als blad 0 aan stam 6. Figuur 1.3 laat stap voor stap zien hoe u een dergelijk diagram maakt.

Het globale patroon van het stamdiagram is onregelmatig, zoals vaak het geval is als er slechts enkele waarnemingen zijn. Op het eerste gezicht onderscheiden we twee **groepen** landen. Het diagram roept de vraag op hoe we de verschillen in alfabetisme kunnen verklaren. Waarom hebben bijvoorbeeld de drie Centraal-Aziatische landen (Kazachstan, Tadzjikistan, en Oezbekistan) zeer hoge percentages mensen die kunnen lezen en schrijven?

Als u twee verwante verdelingen met elkaar wilt vergelijken, is een **rug-aan-rug stamdiagram** met gemeenschappelijke stammen zinvol. De bladeren aan weerskanten worden vanuit de gemeenschappelijke stam geordend. Hier volgt een rug-aan-rug stamdiagram dat de verdeling weergeeft van de percentages mannen en vrouwen die kunnen lezen en schrijven in de landen van tabel 1.2.

Vrouwen		Mannen
9	2	
8 1	3	
6	4	
	5	0
3 3 0	6	8 8
8 1 1 0	7	0 8
6 5 2	8	3 4 5 9
9 9 9	9	2 2 4 5 6
	10	0 0 0

De waarden aan de linkerzijde geven de percentages vrouwen aan, zoals in figuur 1.3, maar dan van rechts naar links geordend in oplopende volgorde. De waarden aan de rechterkant geven de percentages bij de mannen aan. Het is duidelijk dat in deze landen het percentage mannen dat kan lezen en schrijven hoger is dan het percentage vrouwen.

Stamdiagrammen werken niet goed bij grote verzamelingen gegevens, waar elke stam aan een groot aantal bladeren plaats moet bieden. Gelukkig zijn er aanpassingen mogelijk van het basis-stamdiagram, die handig zijn bij het weergeven van een middelgroot aantal waarnemingen. Men kan het aantal stammen in het diagram vergroten door elke stam in tweeën te **splitsen**, één met de bladeren 0 tot 4 en de andere met de bladeren 5 tot 9. Als de waargenomen waarden veel cijfers hebben, is het vaak het beste om vóór het maken van het stamdiagram de getallen af te ronden tot een beperkt aantal cijfers. De beslissing om de stammen te splitsen of **af te ronden**, wordt aan de eigen beoordeling overgelaten. Bedenk dat een stamdiagram tot doel heeft de vorm van een verdeling weer te geven. Als een stamdiagram minder dan (ongeveer) 5 stammen heeft, is het normaal gesproken raadzaam de stammen te splitsen, tenzij er weinig waarnemingen zijn. Als er veel stammen zijn met weinig bladeren of slechts één blad, kan men overwegen of afkappen handig is. Hier volgt een voorbeeld waarin deze aanpassingen allebei worden gebruikt.

Voorbeeld 1.6

Laten we terugkeren naar de duur van de 80 telefoongesprekken van klanten uit tabel 1.1. Om van deze verdeling een grafische voorstelling te maken, kappen we eerst de duur van de gesprekken af tot tientallen seconden door het laatste cijfer weg te laten. Zo worden 56 seconden afgekapt tot 5 en 143 seconden tot 14. (We zouden ook kunnen afronden tot de dichtstbijzijnde 10 seconden, maar afkappen werkt sneller als u het met de hand moet doen.)

We gebruiken dan de tientallen seconden als onze bladeren, waarbij de cijfers aan de linkerkant de stammen vormen. Daardoor krijgen we de bladeren van 1 cijfer die nodig zijn voor het stamdiagram. Bijvoorbeeld: 56 afgekapt tot 5 wordt blad 5 aan stam 0; 143 afgekapt tot 14 wordt blad 4 aan stam 1.

Stem-and-Leaf Display: Call Length

```
Stem-and-leaf of C1  N  = 80
Leaf Unit = 10

    0  00000000111233444
    0  555555566677777778889
    1  00001112223344444
    1  57789
    2  0012
    2  7789
    3  2
    3  678

HI 43, 46, 47, 70, 70, 95, 114, 263
```

Figuur 1.4 stamdiagram van de Minitab met de duur van de 80 telefoongesprekken uit tabel 1.1. De software heeft de gegevens afgekapt door het laatste cijfer weg te laten. Het heeft ook de stammen gesplitst en de hoogste waarnemingen afzonderlijk buiten de grafische weergave vermeld.

> Omdat we 80 waarnemingen hebben, splitsen we de stammen. Zo wordt 56 afgekapt tot 5 om blad 5 te worden aan de tweede stam 0, samen met alle bladeren van 5 tot 9. De bladeren 0 tot 4 gaan naar de eerste stam 0. Figuur 1.4 is een met behulp van software vervaardigd stamdiagram van deze gegevens. De software deed automatisch wat wij al voorstelden: het afkappen tot tientallen seconden en het splitsen van de stammen. Om ruimte te besparen, toont de software ook de hoogste waarden als 'HI', in plaats van de stammen helemaal tot en met 26 te laten zien. Het stamdiagram geeft het globale patroon van de verdeling weer, met veel telefoontjes van korte tot gemiddelde lengte en enkele bijzonder lange gesprekken.

1.1.4 Histogrammen

Stamdiagrammen laten de feitelijke waarden van de waarnemingen zien. Deze eigenschap maakt stamdiagrammen onhandig voor grote gegevensverzamelingen. Bovendien verdeelt het door een stamdiagram getoonde beeld de waarnemingen in groepen (stammen) die meer door het getalsysteem worden bepaald dan door beoordeling. Histogrammen hebben deze beperkingen niet. Een **histogram** verdeelt het waardebereik van een variabele in intervallen en toont slechts het aantal of percentage waarnemingen dat in elk interval terechtkomt. Men kan elk gunstig aantal intervallen kiezen, maar men moet altijd intervallen kiezen van gelijke breedte. Het kost meer tijd om histogrammen met de hand te construeren dan stamdiagrammen, en ze laten de feitelijk waargenomen waarden niet zien. Daarom geven we voor kleine gegevensverzamelingen de voorkeur aan stamdiagrammen. De constructie van een stamdiagram kan men het beste via een voorbeeld laten zien. Elk statistisch softwarepakket kan natuurlijk histogrammen maken.

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Tabel 1.3 Scores in IQ-test voor 60 aselect gekozen tienjarige leerlingen

Voorbeeld 1.7
Waarschijnlijk hebt u wel eens gehoord dat de verdeling van scores op IQ-testen ongeveer 'klokvormig' is. Laten we eens kijken naar enkele feitelijke IQ-scores. Tabel 1.3 geeft de IQ-scores weer van 60 aselect gekozen tienjarige leerlingen van een school.[6]

1. Verdeel de spreidingsbreedte van de gegevens in klassen van gelijke breedte. De scores in tabel 1.3 variëren van 81 tot 145, dus kozen wij als onze klassen

$$\begin{gathered} 75 \leq \text{IQ-score} < 85 \\ 85 \leq \text{IQ-score} < 95 \\ \vdots \\ 145 \leq \text{IQ-score} < 155 \end{gathered}$$

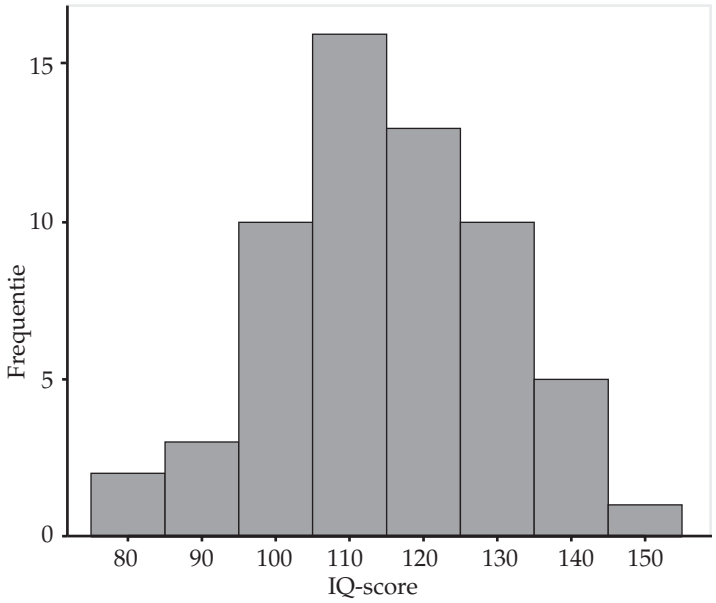
   Zorg ervoor dat u de klassen precies zo definieert dat elke score precies binnen een klasse valt. Een leerling met een IQ van 84 valt in de eerste klasse, maar een leerling met 85 komt terecht in de tweede.

2. Tel het aantal scores in iedere klasse op. Deze optellingen noemen we **frequenties**. Een tabel met frequenties voor alle klassen noemen we een frequentietabel.

Klasse	**Aantal**	**Klasse**	**Aantal**
75 – 84	2	115 – 124	13
85 – 94	3	125 – 134	10
95 – 104	10	135 – 144	5
105 – 114	16	145 – 154	1

3. Teken het histogram. Om te beginnen zet u op de horizontale as de schaal uit voor de variabele waarvan u de verdeling wilt weergeven. Dat is de IQ-score. De schaal loopt van 75 tot 155, want dat is het bereik van de door ons gekozen klassen. De verticale as geeft de schaal van het aantal leerlingen aan. Elke kolom staat voor een klasse. De basis van de kolom dekt de klasse, en de hoogte van de kolom is de klassenfrequentie. Er is geen horizontale ruimte tussen de kolommen, tenzij een klasse leeg is; dan is de hoogte van de kolom nul. Figuur 1.5 is ons histogram. Het ziet er ongeveer klokvormig uit.

Grotere gegevensverzamelingen worden doorgaans weergegeven in de vorm van frequentietabellen als het niet praktisch is de afzonderlijke waarnemingen te tonen. Naast de frequentie

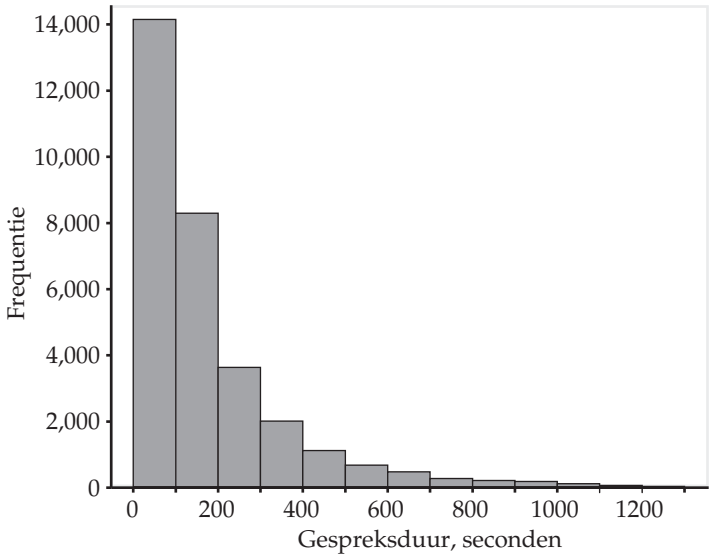


Figuur 1.5 Histogram van de IQ-scores van 60 tienjarige leerlingen (voorbeeld 1.7).

(het aantal) voor elke klasse kan de interesse ook uitgaan naar de fractie of het percentage waarnemingen in iedere klasse. Een histogram met percentages lijkt precies op een frequentiehistogram zoals in figuur 1.5. Zet eenvoudigweg de percentages af op de verticale schaal. Gebruik histogrammen met percentages (relatieve frequenties) voor de vergelijking van verscheidene verdelingen met verschillende aantallen waarnemingen.

Onze ogen reageren op de *oppervlakte* van de kolommen in een histogram. Aangezien de klassen allemaal dezelfde breedte hebben, wordt de oppervlakte bepaald door de hoogte en zijn alle klassen eerlijk vertegenwoordigd. Er bestaat niet zoiets als één juiste keuze van de klassen in een histogram. Te weinig klassen geven een 'wolkenkrabber' diagram, met alle waarden in een paar klassen met lange kolommen. Te veel klassen geeft een 'pannenkoek' diagram, met klassen met één of geen waarnemingen. Geen van deze situaties zal de vorm van de verdeling duidelijk tonen. Het kiezen van geschikte klassen is een kwestie van de juiste beoordeling. Statistische software kiest doorgaans de klassen voor u. De keuze die de software maakt is meestal goed, maar u kunt deze desgewenst aanpassen.

Denk eraan dat het uiterlijk van een histogram verandert als u de klassen verandert. Figuur 1.6 is een histogram van de belduur van de klantenservice die u ook vindt weergegeven in figuur 1.2. Het werd zonder bijzondere aanwijzingen van de gebruiker door software vervaardigd. Het standaard histogram van de software geeft de globale vorm van de verdeling weer, maar verbergt de uitschieters van zeer korte telefoontjes door alle telefoontjes van minder dan 100 seconden in de eerste klasse onder te brengen. We verkregen figuur 1.2 door te vragen om kleinere klassen nadat tabel 1.1 bij ons de indruk had gewekt dat hele korte telefoontjes wel eens een probleem konden vormen. Software automatiseert het maken van grafieken, maar



Figuur 1.6 Het standaard histogram vervaardigd met software voor de gegevens van de belduur (voorbeeld 1.4). Deze keuze van de klassen verbergt het grote aantal zeer korte telefoontjes dat wel door het histogram van dezelfde gegevens in figuur 1.2 wordt getoond.

kan niet over uw gegevens nadenken. De histogramfunctie in het *One-Variable Statistical Calculator*-applet op de website stelt u in staat het aantal klassen te wijzigen door met de muis te slepen. Zo kunt u eenvoudig zien hoe de keuze van de klassen het histogram beïnvloedt.

Hoewel histogrammen op staafdiagrammen lijken, zijn de details en de toepassingen verschillend. Een histogram toont de verdeling van de frequenties of relatieve frequenties van de waarden van een enkele variabele en een staafdiagram vergelijkt de omvang van de verschillende categorieën. De horizontale as van een staafdiagram hoeft geen maatschaal te hebben, want hierbij worden alleen de categorieën onderscheiden die worden vergeleken. Teken staafdiagrammen met een blanco ruimte tussen de kolommen om de categorieën te scheiden en histogrammen zonder blanco ruimte om aan te geven dat alle waarden van de variabele zijn gedekt. Sommige software die oorspronkelijk niet bedoeld is voor statistiek, zoals spreadsheet programma's, tekenen histogrammen alsof deze staafdiagrammen zijn, met ruimte tussen de kolommen. Men kan de software zo instellen dat deze de ruimte verwijdert en een correct histogram tekent.

1.1.5 Onderzoeken van verdelingen

Het maken van een diagram is geen doel op zich, maar het moet ons helpen de gegevens beter te begrijpen. Na het maken van een diagram of grafiek moet je je altijd afvragen, 'Wat zie ik?' Na het weergeven van een verdeling kunnen we de belangrijke kenmerken er als volgt uithalen.

HET ONDERZOEKEN VAN EEN VERDELING

Kijk in een diagram of grafiek naar het **globale patroon** en naar opvallende **afwijkingen** van dat patroon.
Je kunt het globale patroon van een verdeling beschrijven door middel van zijn **vorm**, **centrum** en **spreiding**.
Een belangrijk type afwijking is een **uitschieter**, een individuele waarde die buiten het globale patroon valt.

In paragraaf 1.2 zullen we leren hoe we het centrum en spreiding numeriek kunnen beschrijven. Voorlopig zullen we het centrum van een spreiding beschrijven door zijn *mediaan*, de waarde waarvoor geldt dat de helft van de waarnemingen een lagere waarde heeft en de helft een hogere waarde. We kunnen de spreiding van een verdeling beschrijven door het bereik te bepalen tussen de *laagste en hoogste waarden*. Om de vorm van de verdeling beter te kunnen zien, moet je een stamdiagram op zijn kant zetten, zodat de hogere waarden rechts komen te liggen. Een aantal punten waarop gelet moet worden bij het beschrijven van de vorm zijn:

- Heeft de verdeling één top of verschillende **toppen**? Een verdeling met één top wordt **unimodaal** genoemd.
- Is zij bij benadering symmetrisch of is zij naar één kant scheef? Een verdeling is **symmetrisch** als de waarden die lager of hoger zijn dan het centrum elkaars spiegelbeeld zijn. Zij is **scheef** naar rechts als de rechterstaart (hogere waarden) veel langer is dan de linkerstaart (lagere waarden).

Sommige variabelen hebben vaak verdelingen van een voorspelbare vorm. Veel biologische metingen van exemplaren van dezelfde soort en geslacht hebben symmetrische verdelingen, denk bijvoorbeeld aan de afmetingen van een vogelbek, of de lengte van jonge vrouwen. Hoeveelheden geld zijn echter qua verdeling doorgaans scheef naar rechts. Zo vallen veel huizen qua prijs in de middenklasse, maar de weinige zeer dure herenhuizen geven de verdeling van de huizenprijzen een rechtsscheve verdeling.

Voorbeeld 1.8

Wat heeft het histogram van de IQ-scores (figuur 1.5) ons te zeggen? Vorm: De verdeling is bij benadering symmetrisch met een enkele top in het centrum. We verwachten niet dat echte gegevens perfect symmetrisch zijn. Daarom stellen we ons tevreden als beide zijden van het histogram qua vorm en omvang globaal vergelijkbaar zijn. **Centrum**: U ziet in het histogram dat het middenpunt niet ver van 110 ligt. Een blik op de feitelijke gegevens leert dat het middenpunt 114 is. **Spreiding**: De spreiding loopt van 81 tot 145. Er zijn geen uitschieters of andere sterke afwijkingen van het symmetrische, unimodale (eentoppige) patroon.

De verdeling van de belduur in figuur 1.6, anderzijds, is sterk scheef naar rechts. Het middenpunt, de belduur van een gangbaar telefoontje, is ongeveer 115 seconden, of iets minder dan 2 minuten. De spreiding is heel groot, van 1 seconde tot 28.739 seconden.

De enkele lange telefoongesprekken zijn uitschieters. Ze staan buiten de lange rechterstaart van de verdeling, hoewel we dit niet kunnen zien aan figuur 1.6, dat de grootste waarnemingen buiten

beschouwing laat. Het langste telefoontje duurde bijna 8 uur; dat komt mogelijk eerder door een technische storing dan door een feitelijk gesprek met een klant.

1.1.6 Omgaan met uitschieters

Bij gegevensverzamelingen die kleiner zijn dan die van de belgegevens van de klantenservice, kunt u uitschieters opsporen door te kijken naar waarnemingen die vallen buiten het algemene patroon (eronder of erboven) van een histogram of stamdiagram. *Het opsporen van uitschieters is een kwestie van inschatten. Kijk naar punten die duidelijk buiten de puntenwolk staan, en dus niet alleen naar de meest extreme waarnemingen in een verdeling.* Zoek naar een verklaring voor elke uitschieter. Soms wijzen uitschieters op vergissingen bij het vastleggen van de gegevens. In andere gevallen kan de perifere waarneming worden veroorzaakt door een technisch mankement of ongebruikelijke omstandigheden.

Voorbeeld 1.9

De fabricage van een elektronisch onderdeel vereist de bevestiging van zeer fijne draden aan een schijf halfgeleidermateriaal. Bij een zwakke verbinding kan het onderdeel het laten afweten. Hier volgen enkele metingen van de breeksterkte (in Amerikaanse ponden) van 23 verbindingen.[7]

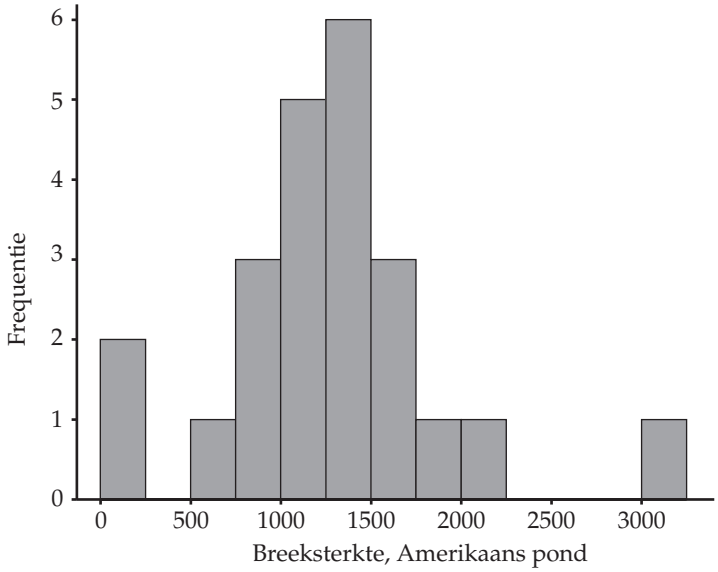
0	0	550	750	950	950	1150	1150
1150	1150	1150	1250	1250	1350	1450	1450
1450	1550	1550	1550	1850	2050	3150	

Figuur 1.7 is een histogram van deze gegevens. We verwachten dat de breeksterkte van als identiek bedoelde verbindingen bij benadering als een symmetrisch patroon wordt weergegeven, met enige toevalsvariatie tussen de verbindingen. Figuur 1.7 toont inderdaad een symmetrisch patroon waarbij het centrum ligt bij ongeveer 1250 Amerikaans pond – maar het toont ook drie uitschieters die buiten dit patroon liggen, twee laag en een hoog.

De technici konden alle drie de uitschieters verklaren. De twee laagst gelegen uitschieters hadden een breeksterkte van 0 omdat de verbinding tussen de draad en de schijf niet tot stand was gekomen. De hoogste uitschieter van 3150 pond was een meetfout. Bij verder gegevensonderzoek kunnen deze drie uitschieters buiten beschouwing blijven. Wel werd al direct duidelijk dat de variatie in breeksterkte te groot is: 550 tot 2050 pond als we de uitschieters buiten beschouwing laten. Het proces van de bevestiging van de draden aan een schijf moet worden verbeterd met het oog op consistentere resultaten.

1.1.7 Tijdgrafieken

Als gegevens over een bepaalde periode zijn verzameld, is het steeds aan te raden om de waarnemingen in chronologische volgorde te plaatsen. *Weergaven van de verdeling van een variabele die de chronologische volgorde buiten beschouwing laten, zoals stamdiagrammen of histogrammen, kunnen bedrieglijk zijn als er sprake is van een systematische verandering in de loop van de tijd.*



Figuur 1.7 Histogram van een verdeling met zowel lage als hoge uitschieters (voorbeeld 1.9).

TIJDGRAFIEK

Een **tijdgrafiek** van een variabele zet elke waarneming uit tegen de tijd waarop zij was gemeten. Plaats altijd de tijd op de horizontale as van uw grafische voorstelling en de gemeten variabele op de verticale as. Door de gegevenspunten met behulp van lijnen te verbinden, benadrukt u de veranderingen in de loop van de tijd.

Voorbeeld 1.10

Tabel 1.4 geeft een overzicht van de hoeveelheid water die jaarlijks van 1954 tot 2001[8] uit de Mississippi in de Golf van Mexico stroomde. De eenheden zijn in kubieke kilometers water: de Mississippi is een grote rivier. Beide diagrammen in figuur 1.8 laten deze gegevens zien. Het histogram in figuur 1.8(a) toont de verdeling van de uitgestroomde hoeveelheid. Het histogram is symmetrisch en unimodaal, waarbij het centrum op bijna 550 kubieke kilometers ligt. Wellicht zijn we geneigd te denken dat de gegevens slechts betrekking hebben op toevallige jaarlijkse fluctuaties in het rivierpeil rond het gemiddelde op langere termijn.

Figuur 1.8(b) is een tijdgrafiek van dezelfde gegevens. Het eerste punt ligt boven 1954 op de 'jaar'-as op een hoogte van 290, de hoeveelheid water die in 1954 uit de Mississippi wegstroomde. De tijdgrafiek vertelt een interessanter verhaal dan het histogram. U ziet flink wat verschillen van jaar tot jaar, maar het is ook duidelijk dat het tijdsverloop een stijgende **trend** te zien geeft. Dat wil zeggen, er is op langere termijn een toename zichtbaar van de hoeveelheid weggestroomd water. De lijn op de grafiek is een 'trendlijn'. Deze wordt berekend met behulp van de gegevens en maakt de trend zichtbaar. Deze trend laat een klimaatverandering zien: er is in toenemende mate sprake van regenval en overstromingen in Noord-Amerika.

Jaar	Uitstroom	Jaar	Uitstroom	Jaar	Uitstroom	Jaar	Uitstroom
1954	290	1966	410	1978	560	1990	680
1955	420	1967	460	1979	800	1991	700
1956	390	1968	510	1980	500	1992	510
1957	610	1969	560	1981	420	1993	900
1958	550	1970	540	1982	640	1994	640
1959	440	1971	480	1983	770	1995	590
1960	470	1972	600	1984	710	1996	670
1961	600	1973	880	1985	680	1997	680
1962	550	1974	710	1986	600	1998	690
1963	360	1975	670	1987	450	1999	580
1964	390	1976	420	1988	420	2000	390
1965	500	1977	430	1989	630	2001	580

Tabel 1.4 Jaarlijkse uitstroom van de Mississippi, kubieke kilometers water

Veel interessante gegevensverzamelingen zijn **tijdreeksen**, metingen van een variabele in regelmatige, opeenvolgende tijdvakken. De overheid publiceert dikwijls economische en sociale data als een tijdreeks, bijvoorbeeld het maandelijks werkloosheidscijfer en het Bruto Nationaal Product per kwartaal. Weerberichten, de vraag naar elektriciteit en metingen aan eenheden vervaardigd tijdens een productieproces, zijn andere voorbeelden van tijdreeksen.

Een grafische weergave van een tijdreeks kan de voornaamste eigenschappen ervan zichtbaar maken.

Boven de basis: Splitsing van tijdreeksen*

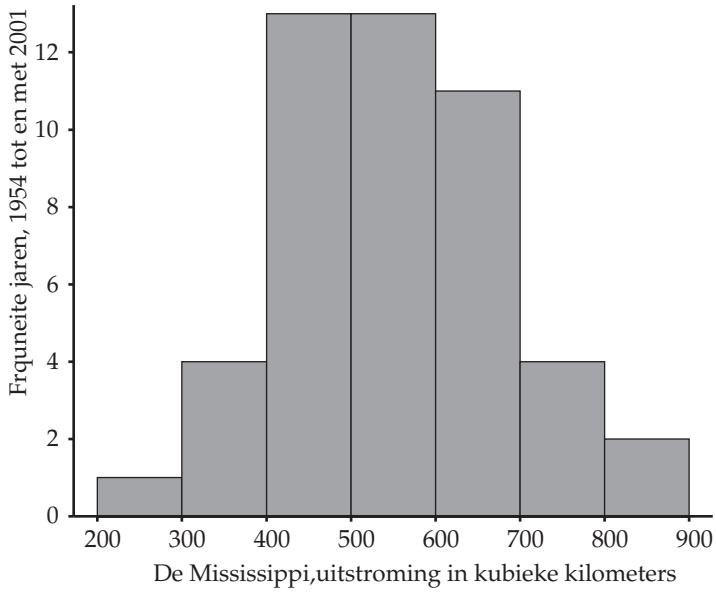
Onderzoekt u een tijdgrafiek, kijk dan eerst naar het globale patroon en vervolgens naar markante afwijkingen van dat patroon. Hier volgen twee belangrijke categorieën globale patronen van tijdreeksen.

SEIZOENSVARIATIE EN TREND

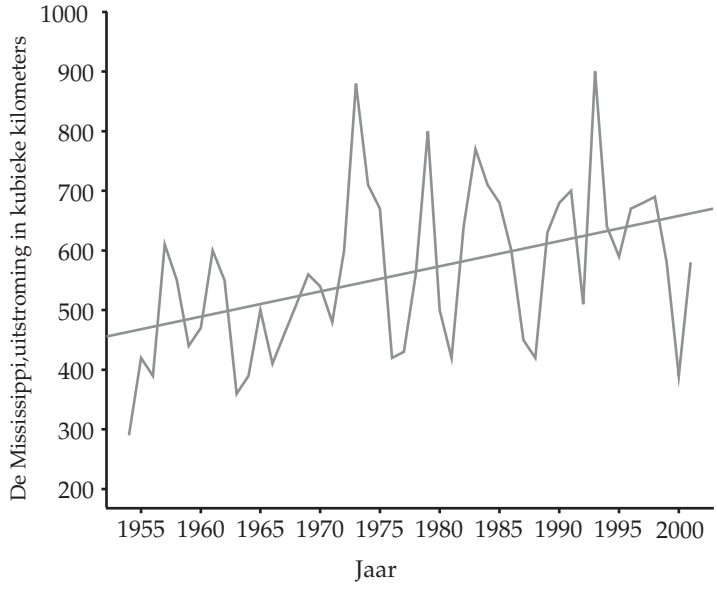
Een patroon in een tijdreeks dat zich steeds herhaalt op bekende regelmatige tijdsintervallen wordt een **seizoensvariatie** genoemd.
Een **trend** in een tijdreeks is een aanhoudende lange termijn stijging of daling.

Omdat veel economische tijdreeksen een sterke seizoensvariatie laten zien, voeren overheidsinstanties dikwijls een bijstelling uit met betrekking tot deze variatie, alvorens de economische

* In de boven-de-basisparagrafen worden aanvullende onderwerpen kort behandeld. Met behulp van uw software kunt u enkele van deze onderwerpen bestuderen. Zo zijn de in figuur 1.9 tot en met 1.11 weergegeven resultaten afkomstig van Minitab statistische software.

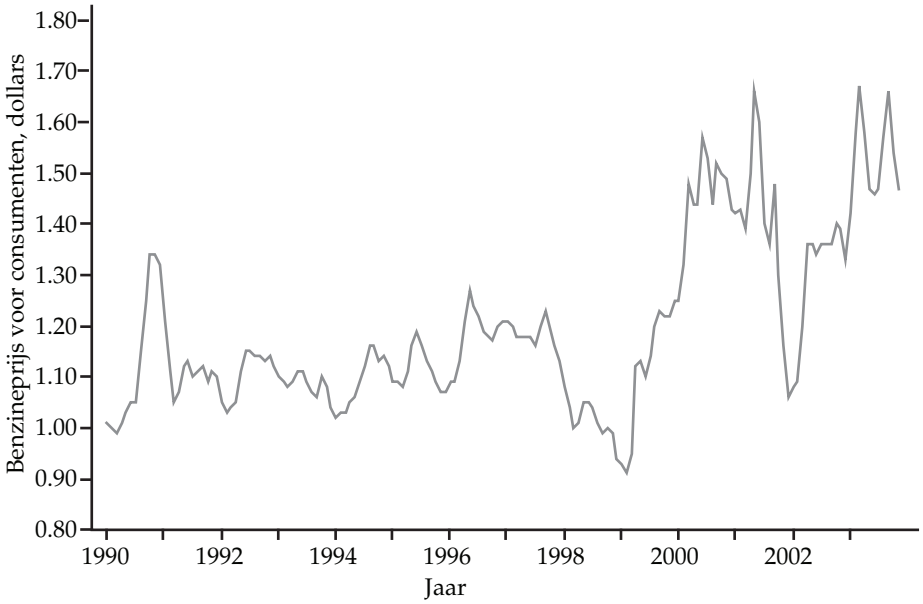


(a)



(b)

Figuur 1.8 (a) Histogram van de hoeveelheid water die uit de Mississippi wegstroomde in de 48 jaar van 1954 tot en met 2001. Gegevens van tabel 1.4. (b) Tijdgrafiek van de hoeveelheid water die de Mississippi uitstroomde in de jaren 1954 tot en met 2001. De lijn toont een stijgende trend van uitgestroomd rivierwater, een trend die niet zichtbaar is in het histogram in figuur 1.8(a).



Figuur 1.9 Tijdgrafiek van de gemiddelde maandprijs van gewone benzine van 1990 tot en met 2003 (voorbeeld 1.11).

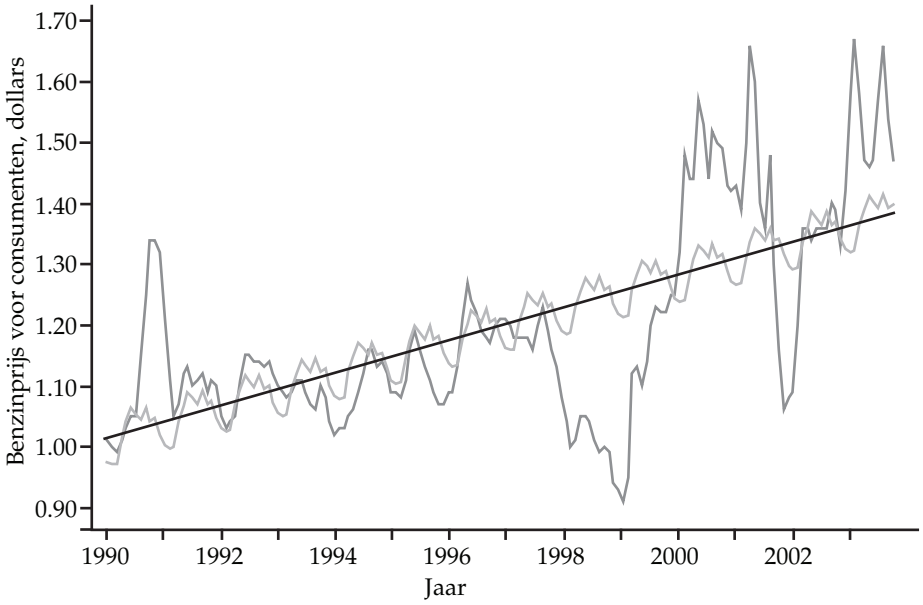
data vrij te geven. Men zegt dan dat de data *gecorrigeerd zijn voor seizoensinvloeden*. **Seizoenscorrectie** helpt misinterpretatie voorkomen. Een stijging in het werkloosheidscijfer van december tot januari betekent bijvoorbeeld niet dat de economie zwakker wordt. De werkloosheid stijgt bijna altijd in januari wanneer het met tijdelijk werk voor Kerstmis is gedaan en het buitenwerk afneemt wegens de kou. Het voor het seizoen gecorrigeerde werkloosheidscijfer kondigt alleen een stijging aan als de werkloosheid van december tot januari sterker dan normaal omhooggaat.

Voorbeeld 1.11

Figuur 1.9 is een tijdgrafiek van de gemiddelde detailhandelsprijs per maand van gewone benzine in de jaren 1990 tot 2003.[9] De prijzen zijn niet gecorrigeerd voor het seizoen. U ziet de pieken in de prijzen in 1990 door de invasie van Irak in Koeweit; de sterke daling in 1998 toen door een economische crisis in Azië de vraag naar brandstof terugliep; en snelle prijsstijgingen in 2000 en 2003 door de instabiliteit in het Midden-Oosten en de productiebeperkingen van de OPEC-landen. Deze afwijkingen zijn zo groot dat het globale patroon nauwelijks zichtbaar is.

Niettemin is er een duidelijke trend naar stijgende prijzen. Een groot deel van deze trend is toe te schrijven aan de inflatie, de stijging van het algemene prijsniveau gedurende deze jaren. Ook laat een nadere blik op het diagram zien dat er sprake is van seizoensvariaties: een jaarlijks terugkerende stijging en daling. Amerikanen rijden meer auto tijdens het zomerseizoen, dus stijgt de prijs van benzine steeds in de lente en daalt weer in de herfst als de vraag daalt.

Statistische software helpt ons een tijdreeks te onderzoeken door gegevens onder te verdelen naar systematische patronen, zoals trends en seizoensfluctuaties, en naar de residuen die



Figuur 1.10 Tijdgrafiek van benzineprijzen met trendlijn en seizoensvariatie. Dit zijn de globale patronen die door de software zijn afgeleid uit de gegevens.

achterblijven na het verwijderen van de patronen. In figuur 1.10 zijn de trend en seizoensfluctuaties toegevoegd aan de tijdgrafiek van de benzineprijzen. De zwarte lijn laat de stijgende trend zien. De seizoensfluctuaties zijn zichtbaar als lichte lijn die jaarlijks met geregelde tussenpozen stijgt en daalt. Dit is een gemiddelde van de seizoensfluctuaties over alle jaren in de oorspronkelijke gegevens, zoals automatisch door software wordt gegenereerd.

De trend en seizoensfluctuaties in figuur 1.10 vormen het globale patroon van de gegevens. Figuur 1.11 is een grafische voorstelling van de residuen die overblijven als we zowel de trend als de seizoensfluctuaties aftrekken van de oorspronkelijke gegevens. Dat wil zeggen, figuur 1.11 legt de nadruk op de afwijkingen van het patroon. In het geval van benzineprijzen zijn de afwijkingen groot (tot wel 30 dollarcent hoger of lager). Het is duidelijk dat trend en seizoensfluctuaties in het geheel niet geschikt zijn om de benzineprijzen nauwkeurig te voorspellen.

Samenvatting

Een gegevensverzameling bevat informatie over een verzameling **elementen**. Elementen kunnen mensen, dieren of dingen zijn. De gegevens voor één element vormen als geheel een **geval**. Voor elk element geven de data waarden voor één of meer **variabelen**. Een variabele beschrijft bepaalde eigenschappen van een element, zoals de lengte, het geslacht of het salaris van een persoon.

Figuur 1.11 De residuen die overblijven wanneer we de maandelijkse benzineprijzen hebben gecorrigeerd op de trend en de seizoensfluctuaties.

Sommige variabelen zijn **kwalitatief** en andere **kwantitatief**. Een kwalitatieve variabele plaatst een element in een categorie. Een kwantitatieve variabele neemt numerieke waarden aan die bepaalde eigenschappen van een element meten, zoals de lengte in centimeters of het jaarsalaris in euro's.

Exploratieve data-analyse gebruikt grafieken en numerieke samenvattingen om de variabelen te beschrijven in een gegevensverzameling en de relaties tussen de variabelen.

De **verdeling** van een variabele toont de waarden die deze aanneemt en hoe vaak dit gebeurt.

Staafdiagrammen en **taartdiagrammen** geven de verdelingen van kwalitatieve variabelen weer. Deze grafieken maken gebruik van de frequenties of de relatieve frequenties van de categorieën.

Stamdiagrammen en **histogrammen** geven de verdelingen van kwantitatieve variabelen weer. Stamdiagrammen verdelen elke waarneming in een **stam** en een **blad** van één cijfer. Histogrammen geven de frequenties of relatieve frequenties weer van de klassen van waarden.

Bij het onderzoeken van een verdeling wordt er gekeken naar de **vorm**, **het centrum** en **de spreiding** en naar duidelijke **afwijkingen** van het globale patroon.

Sommige verdelingen hebben eenvoudige vormen, zoals **symmetrisch** en **scheef**. Het aantal **toppen** is een ander aspect van de globale vorm. Niet alle verdelingen hebben een eenvoudige globale vorm, vooral als er slechts enkele waarnemingen zijn.

Uitschieters zijn waarnemingen die buiten het globale patroon van een verdeling vallen. Probeer altijd naar uitschieters te zoeken en probeer ze te verklaren.

Wanneer de waarnemingen van een variabele over een bepaalde tijd zijn genomen, maken we een **tijdsgrafiek** waarin de tijd horizontaal is uitgezet en de waarden van de variabelen verticaal. Een tijdsgrafiek kan **trends** weergeven of andere veranderingen over een bepaalde tijd.

1.2 Verdelingen beschrijven

Geïnteresseerd in een sportauto? Maakt u zich bezorgd dat deze te veel benzine verbruikt? De Environmental Protection Agency neemt de meeste sportwagens op in zijn categorieën 'tweezitters' of 'mini-auto's. Tabel 1.5 geeft het benzineverbruik (per mijl) in de stad en op de snelweg voor de auto's in deze categorieën.[10] We gaan de tweezitters vergelijken met mini-auto's en het stadsverbruik met het verbruik op de snelwegen. We kunnen beginnen met diagrammen, maar numerieke samenvattingen maken de vergelijkingen specifieker.

Tweezitters			Mini-auto's		
Model	**Stad**	**Snelweg**	**Model**	**Stad**	**Snelweg**
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborgini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborgini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

Tabel 1.5 Benzineverbruik (mijl per gallon) voertuigen van modeljaar 2004