Voorwoord

Dit boek is het resultaat van een grondige herwerking van de cursustekst "Beschrijvende Statistiek en Kansrekenen", die ik ontwikkelde voor het gelijknamige opleidingsonderdeel voor studenten Handelsingenieur, Handelsingenieur in de Beleidsinformatica, Sociaal-Economische Wetenschappen en Toegepaste Economische Wetenschappen aan de Faculteit Bedrijfswetenschappen en Economie van de Universiteit Antwerpen. In vergelijking met eerder gepubliceerde versies van de cursustekst is de eerste grote wijziging dat een aanzienlijk aantal nieuwe onderwerpen aan bod komen, zoals de nominale en ordinale spreidingsindices, kurtosis of gepiektheid, de rangcorrelatiecoëfficiënt, en de gamma, de Weibull-, de lognormale en de beta kansdichtheden. Ook de multinomiale kansverdeling en de multivariate normale kansdichtheid worden nu behandeld. Daarnaast bevat de herwerkte versie van de cursus een aanzienlijk groter aantal figuren en tabellen, met als doel meer inzicht te verschaffen in de concepten en de formules die doorheen het boek aangehaald worden. Tenslotte maakt dit boek gebruik van het statistisch softwarepakket JMP (spreek uit als "jump"), in tegenstelling tot eerdere versies van de cursus, die steunden op het gebruik van Microsoft Excel.

Het grote voordeel van JMP is dat het gebruiksvriendelijk en krachtig is, uitstekende grafische mogelijkheden biedt (inclusief het maken van kaarten met statistische informatie), en geschikt is voor alle statistisch georiënteerde opleidingsonderdelen in de bachelorprogramma's Handelsingenieur, Handelsingenieur in de Beleidsinformatica, Sociaal-Economische Wetenschappen, Toegepaste Economische Wetenschappen, en in opleidingen Bio-Ingenieur, Burgerlijk Ingenieur en Industrieel Ingenieur. De bedoeling is om hetzelfde pakket doorheen het ganse programma te gebruiken. Sinds 2025 is JMP Student Edition gratis beschikbaar voor alle studenten en personeelsleden van academische instellingen en dus van universiteiten. Elke lesgever en student kan bijgevolg JMP downloaden op zijn of haar eigen laptop en thuiscomputer. De JMP Student Edition is even krachtig en veelzijdig als JMP Pro, de meest uitgebreide versie van het softwarepakket.

Hoe gebruiksvriendelijk JMP ook is, het gebruik ervan zal enige oefening vergen. Om vertrouwd te raken met het softwarepakket kunt u enkele *webcasts* bekijken op www.jmp.com. De belangrijkste *webcasts* zijn misschien wel te vinden in het menu "For Users" onder de noemer "Learn to Use JMP". Daar is ondermeer de online cursus met als titel "Getting Started with JMP: On Demand Course" te vinden.

Het boek veronderstelt dat de studenten een grondige wiskundige voorkennis hebben overgehouden uit het middelbaar onderwijs, inclusief het gebruik van afgeleiden, integralen en matrixalgebra. Een sterkte van het boek is dat alle wiskundige afleidingen gedetailleerd zijn weergegeven. Dit betekent dat alle benodigde tussenstappen in de wiskundige afleidingen zijn weergegeven, ook diegene die voor wiskundigen misschien triviaal zijn. De bedoeling hiervan is om de drempel voor minder wiskundig begaafde studenten niet hoger te maken dan strikt noodzakelijk.

De leerstof in dit boek wordt behandeld tijdens een 35-tal uur hoorcolleges, waarin de concepten uitgelegd worden, en 24 uur werkcolleges, waarin oefeningen gemaakt worden. Daarnaast worden ook sessies georganiseerd over het gebruik van het softwarepakket JMP.

Het boek onderscheidt zich van veel moderne tekstboeken over beschrijvende statistiek en kansrekenen doordat het een combinatie biedt van theoretische en wiskundige diepgang, gedetailleerde en begrijpbare uitleg bij de geïntroduceerde concepten, tal van praktische voorbeelden, en het gebruik van een gebruiksvriendelijk en toch bijzonder krachtig statistisch pakket.

Bij de uitgave van dit boek wil de auteur graag Leonids Aleksandrovs, Kris Annaert, Stefan Becuwe, Evi Breynders, Filip De Baerdemaeker, Eline De Groof, Thibault Hendrickx, Roselinde Kessels, David Meintrup, Rafaël Michiels, Ida Ruts, Bagus Sartono, Evelien Stoffels, Utami Syafitri, Anil Haydar Topal, Katrien Van Driessen, Ellen Vandervieren, Kristel Van Rompaey, Diane Verbiest en Sara Weyns te bedanken voor hun gedetailleerde opmerkingen en constructieve suggesties, en voor de technische ondersteuning bij het creëren van figuren. De auteur wenst ook Ian Cox, Bradley Jones, Volker Kraft, Brady Brady en Mia Stephens van de JMP divisie in het SAS Institute te bedanken voor hun hulp met betrekking tot het softwarepakket JMP.

Ten slotte wil de auteur ook Prof. Em. Willy Gochet van de KU Leuven en Dr. Anja Struyf van de Universiteit Antwerpen bedanken. De eerste versies van de cursusteksten waarop dit boek gebaseerd is, waren sterk geïnspireerd op de cursusteksten van Willy Gochet. Anja Struyf, co-auteur van het boek "Kansen en Verwachtingen", speelde een cruciale rol bij de tweede herwerking van het boek en gaf in alle voorafgaande jaren ook meerdere suggesties ter verbetering van het boek.

1

WAT IS STATISTIEK?

The world is ready for the truth; the modern age is here; every year another report appears that examines poverty by means of statistical research rather than romantic claptrap.
(uit The Crimson Petal and the White, Michel Faber, p. 334)

Dit inleidend hoofdstuk bevat een algemene omschrijving van de onderwerpen statistiek en kansrekenen. Enkele voorbeelden illustreren het doel en de toepassingsmogelijkheden van beide disciplines, alsook het verschil ertussen. Aangezien binnen de bedrijfswereld, de industrie, het management en de economie de statistiek meer toepassingen vindt dan de kansrekening, krijgt statistiek in de opleidingen Handelsingenieur, Toegepaste Economische Wetenschappen en Sociaal-Economische Wetenschappen veruit de meeste aandacht. Desondanks wordt tijdens deze opleidingen ook de nodige aandacht besteed aan de kansrekening. Beide disciplines kunnen immers niet van elkaar losgekoppeld worden. In dit boek komen zowel kansrekenen als statistiek aan bod.

1.1 Waarom statistiek?

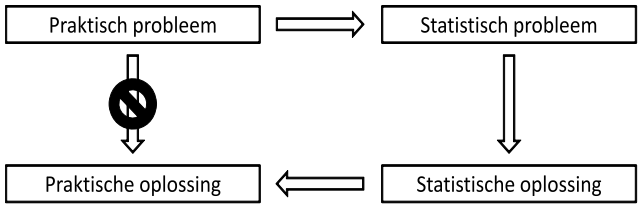
Statistiek is sinds jaar en dag een vak, niet zelden een gevreesd vak, in een groot aantal studierichtingen aan universiteiten en hogescholen. De reden hiervoor is dat heel wat mensen tijdens hun beroepswerkzaamheden vroeg of laat met problemen van gegevensanalyse geconfronteerd worden. Een degelijke statistische achtergrond laat hen niet enkel toe om de gegevens te analyseren en op basis van de analyse concrete beslissingen te nemen, maar het geeft hen ook een voorsprong bij het verzamelen van gegevens.

Dat statistiek desondanks door de meeste studenten niet onmiddellijk als nuttig wordt ervaren is veelal te wijten aan het feit dat zij bij het volgen van de cursus nog niet in aanraking gekomen zijn met allerhande praktische beslissingsproblemen waarmee managers, economisten, ingenieurs en onderzoekers dagelijks geconfronteerd worden. Typisch gaan studenten pas bij het uitvoeren van hun thesiswerkzaamheden beseffen dat statistiek ook voor hen nuttig is. De vele voorbeelden in dit boek zijn bedoeld om deze bewustwording een aantal jaar te vervroegen.

Doorgaans worden bij het introduceren van een cursus statistiek een resem citaten uit de kast gehaald in een poging de studenten te motiveren. Een klassiek voorbeeld is "*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*" van de Britse filosoof Herbert George Wells (1866–1946). Meer recent is de uitspraak van de Amerikaanse kwaliteitsgoeroe W. Edwards Deming, aan wie een groot gedeelte van de ronduit spectaculaire economische heropleving van Japan na de Tweede Wereldoorlog wordt toegeschreven. Hij beweerde dat "*Statistics is too important to be left to statisticians. The goal is to have many statistically-skilled workers: engineers, scientists, managers, ...*" Hal Varian, *chief economist* bij Google beweert het volgende: "*I keep saying that the most sexy job in the next 10 years will be statistician. And I'm not kidding.*" De laatste jaren is het makkelijker geworden om grote hoeveelheden data te verzamelen. We leven dan ook in een big data tijdperk. Statistiek is uiteraard ook van cruciaal belang om die grote hoeveelheden te verwerken en om informatie te halen uit grote onoverzichtelijke gegevensbestanden. Dit is waar klassieke statistiek overgaat in datawetenschappen of data science.

Het is trouwens opvallend dat veelgebruikte statistische methoden ontwikkeld zijn door niet-statistici. Sir Ronald Fisher was een geneticus die landbouwonderzoek verrichte en de resulterende onderzoeksgegevens diende te analyseren. Frank Wilcoxon was een chemicus die handige statistische hypothesetoetsen bedacht die nuttig blijken wanneer enkele traditionele basisveronderstellingen voor data-analyse niet voldaan zijn. William Gosset werkte in de brouwerij Guinness toen hij de t-verdeling afleidde, omdat hij die nodig had voor kleine aantallen steekproefgegevens. Deze voorbeelden tonen aan dat statistiek bijzonder nuttig is in uiteenlopende omgevingen, en dat een welopgevoed burger over een gezonde basiskennis van statistiek zou moeten beschikken.

Een goede motivering voor het aanleren van statistische methoden aan economisten is te vinden in het zogenaamde Six Sigma verbeterprogramma. De bedoeling van dit programma, dat populair is in grotere ondernemingen, is om binnen zowel dienstverlenende als industriële ondernemingen concrete problemen met een grote financiële impact op te lossen en op die manier het aantal fouten en defecten te herleiden tot 3.4 per miljoen operaties. De aanpak is gebaseerd op statistische methoden, zoals voorgesteld in Figuur 1.1. De figuur geeft



Figuur 1.1. Gebruik van statistische methoden om problemen op te lossen.

aan dat de traditionele werkwijze om praktische problemen onmiddellijk met praktische oplossingen te verhelpen verlaten wordt. Deze aanpak is immers vaak gebaseerd op gegis, natte-vingerwerk en trial-and-error, met als gevolg dat een definitieve oplossing niet zelden lang op zich laat wachten. Het Six Sigma verbeterprogramma promoot een meer doordachte, wetenschappelijke aanpak van problemen. Eerst worden gegevens verzameld in de zogenaamde meetfase. Met behulp van statistische methoden worden de gegevens vervolgens onderzocht, wat dikwijls leidt tot interessante inzichten en aanbevelingen ter verbetering van bestaande producten, diensten of processen. De Six Sigma aanpak steunt ook op het gebruik van statistische procescontrole en het uitvoeren van experimenten. Op die manier helpt de statistiek de best mogelijke oplossing te vinden voor allerhande praktische problemen.

Om tot een geslaagd huwelijk te komen tussen enerzijds praktijkmensen en anderzijds statistici wordt van beide kanten vanzelfsprekend enige openheid gevergd. Van de kant van ingenieurs, economisten en managers is een degelijke basiskennis van de basisprincipes en -technieken van de statistiek vereist. De statistiek is bijgevolg een onmisbaar facet in de waaier van vaardigheden die van een polyvalente medewerker verwacht worden. Dit verklaart waarom niet enkel in eerste en tweede bachelor Handelsingenieur, Handelsingenieur in de Beleidsinformatica, Toegepaste Economische Wetenschappen en Sociaal-Economische Wetenschappen statistiek gedoceerd wordt, maar ook in de latere jaren van de opleidingen.

In de loop der jaren zijn er specifieke statistische methoden ontwikkeld voor typische ingenieurstoepassingen, en voor typische onderzoeksproblemen uit ondermeer de economie, de psychologie, de biologie en de chemie. Soms spreekt men in die toepassingsdomeinen over *technometrics*, econometrie, psychometrie, biometrie en chemometrie in plaats van statistiek. Soms gebeurt het dat eenzelfde statistische methode bekend staat onder verschillende benamingen in de diverse toepassingsdomeinen. De literatuur over statistische technieken is in de loop der jaren versnipperd geraakt over al die domeinen. In sommige gevallen ging dit gepaard met vakspecifieke terminologie en de ontwikkeling van jargon. Een frequent gebruikt synoniem voor biometrie is biostatistiek. Daarmee bedoelt men de toepassing van statistische methoden in de biologie, de geneeskunde en de bio-ingenieurswetenschappen.

1.2 Definitie van statistiek

Het woord statistiek klinkt iedereen wellicht vertrouwd in de oren. Een statistiek verwijst doorgaans naar numerieke informatie, bijvoorbeeld informatie omtrent

- de bevolking van een land: geboorte- en sterftecijfers, immigraties en emigraties, … (deze informatie noemt men bevolkingsstatistieken),

- de economie: tewerkstellings- en werkloosheidscijfers, investeringen, prijzen, bruto nationaal product (BNP), … (deze statistieken noemt men economische statistieken),
- een bedrijf of sector: verkoopscijfers, resultatenrekening, groei, aanwervingen, afvloeiingen, … (deze cijfers noemt men bedrijfsstatistieken),
- een mens: lengte, bloeddruk, hartslag, gewicht, …,
- een land: oppervlakte, percentage bebossing, bodemsamenstelling, bebouwde oppervlakte, hoeveelheid neerslag, hoeveelheid zonneschijn, …

Meer formeel kan statistiek gedefinieerd worden als het geheel van methodologieën voor het verzamelen, voorstellen, analyseren en interpreteren van data of gegevens. Hieruit blijkt dat de statistiek een heel algemene hulpwetenschap is, waarvoor in nagenoeg elke werkomgeving een belangrijke rol is weggelegd. Toepassingen van de statistiek in de geneeskunde, de economie, de scheikunde en het bedrijfsmanagement zijn legio, maar ook in de literatuurwetenschap, geschiedenis, politieke wetenschappen, criminologie en zelfs musicologie wordt statistiek gebruikt.

Data of gegevens zijn in de moderne maatschappij massaal aanwezig:
- Computerbestanden in bedrijven bevatten verkoopcijfers, kostprijzen, en klantengegevens (zoals adres, bestelde hoeveelheden en frequentie van bestellen).
- De financiële pagina's van kranten bevatten aandelenkoersen, grondstofprijzen en wisselkoersen.
- De federale en regionale overheden publiceren regelmatig data over de bevolking, handel en industrie. Ook niet-gouvernementele organisaties publiceren op geregelde basis data.
- Het internet is een bron van tal van datasets.

Ondernemingen verzamelen uiteraard ook zelf actief gegevens. Dit gebeurt ondermeer door het uitvoeren van experimenten (bijvoorbeeld om nieuwe producten te ontwerpen), in het kader van statistische procescontrole, of door het meten van allerhande eigenschappen van producten, diensten en processen. Kwaliteitsafdelingen van ondernemingen trachten door het voortdurend analyseren van gegevens producten of diensten af te leveren met zo weinig mogelijk defecten en een maximale betrouwbaarheid. Bovendien wordt gepoogd om het bedrijfsproces zodanig te organiseren dat de afvalberg minimaal is, er weinig tot geen inspectie van afgewerkte producten nodig is, en dat de producten en diensten met een minimum aan kosten aan de klantenvereisten voldoen.

Onderzoeksbureaus verzamelen gegevens via enquêtes per telefoon, per post of door straatinterviews. Dergelijke enquêtes worden opgezet om informatie in te winnen over het winkelgedrag van consumenten, over het kiesgedrag van de bevolking of over de publieke opinie betreffende maatschappelijke problemen.

Ten slotte voeren wetenschappelijke onderzoekers aan universiteiten en onderzoeksinstellingen experimenten uit om, bijvoorbeeld, de groei van planten en dieren te bevorderen, om het effect van nieuwe soorten medicatie uit te testen, om de kwaliteit van diepgevroren voedsel of opgeslagen fruit en groenten te optimaliseren, en om de voorkeuren van consumenten voor wijnsoorten te bepalen of de voorkeuren van pendelaars voor diverse transportmodi te bepalen.

Statistiek laat ons toe om gegevens of data te verwerken tot bruikbare informatie. De rol die de statistiek hierin speelt kan het best geïllustreerd worden aan de hand van enkele voorbeelden.

1.3 Voorbeelden

Voorbeeld 1.1. Een luchtvaartmaatschappij voerde een onderzoek uit naar het gedrag van haar passagiers op intercontinentale vluchten en registreerde daarom

- het aantal passagiers met reservatie dat niet opdaagt (de zogenaamde *no-shows*),
- het gewicht aan bagage dat de passagiers meenemen (vaak geldt een limiet van 20 kilogram), en
- de tijd die de passagiers aankomen vóór het officiële vertrekuur van de vlucht (voor intercontinentale vluchten wordt de passagiers gevraagd minimaal twee uur voor vertrek aanwezig te zijn).

De maatschappij registreerde gedurende enkele maanden deze gegevens en maakte hierbij een onderscheid tussen de passagiers die *economy* vliegen en de passagiers in *business class*. De gegevens moesten vervolgens geanalyseerd worden met de bedoeling maatregelen te nemen. Een voorbeeld hiervan kan zijn meer overboekingen toe te laten, dit wil zeggen meer reservaties opnemen dan er plaatsen zijn op het vliegtuig, en/of strenger optreden tegen passagiers die te veel bagage meenemen.

Voorbeeld 1.2. Bij de productie van koffie is de luchtvochtigheid in de productiehal van cruciaal belang voor de kwaliteit van het eindproduct. De vochtigheidsgraad wordt onder controle gehouden door een systeem dat niet feilloos werkt. Daarom worden dagelijks meerdere metingen van het vochtgehalte verricht om na te gaan of het binnen de perken blijft. Deze aanpak valt onder de noemer statistische procescontrole.

Voorbeeld 1.3. Een vulmachine voor flessen heeft doorgaans meerdere vulkoppen, zodat in één beweging meteen een aantal flessen gevuld kunnen worden. Bij een dergelijk vulproces worden elk uur typisch een aantal gevulde flessen zorgvuldig nagewogen om na te gaan of elke vulkop precies de gewenste hoeveelheid in de flessen deponeert. Een andere interessante onderzoeksvraag in deze context is of er verschillen zijn tussen de metingen uitgevoerd door verschillende analisten.

Voorbeeld 1.4. Grootwarenhuizen verzamelen dankzij de klantenkaarten massa's gegevens. Zaken die typisch geregistreerd worden zijn

- het gespendeerde bedrag per winkelbeurt, al dan niet opgesplitst in categorieën (voeding, kleding, …),
- het aantal verkochte artikelen,
- de betalingswijze (contant, debetkaart, kredietkaart en maaltijdcheque).

Onderzoekers maken gebruik van statistische methoden om deze gigantische hoeveelheid informatie samen te vatten en op een overzichtelijk manier voor te stellen.

Voorbeeld 1.5. Financiële analisten zijn ondermeer geïnteresseerd in de graad van risico van het beleggen in een bepaald aandeel. Daartoe houden zij gedurende jaren de maandelijkse rendementen van dat aandeel bij. Hierbij wordt niet enkel met koerswijzigingen, maar ook met uitgekeerde dividenden rekening gehouden. Bovendien worden de maandelijkse rendementen van de globale markt, bijvoorbeeld de BEL20-index, bijgehouden. Indien het aandeel gemiddeld procentueel sterker stijgt of daalt dan de markt, dan noemt men het aandeel risicovol. In het andere geval spreekt men van een aandeel met weinig risico. Via statistische methoden kunnen verbanden tussen de rendementen van het aandeel en de globale markt onderzocht worden.

1.4 Onderwerp van de statistiek

In de voorbeelden uit de vorige paragraaf worden telkens één of meerdere vragen onderzocht over een **populatie** van objecten of elementen of over een **proces** dat objecten of elementen genereert.

De gegevens over de populatie of het proces worden bekomen door één of meerdere eigenschappen of karakteristieken van hun elementen te registreren. Deze eigenschappen of karakteristieken worden **variabelen** genoemd. Deze naam geeft aan dat de waarde van de eigenschap varieert van element tot element. Daarom wordt de statistiek soms de studie van de variabiliteit genoemd.

Het is meestal onpraktisch om alle elementen uit een populatie of gegenereerd door een proces in een studie op te nemen. In die gevallen werkt men slechts met een deel van de elementen: de **steekproef**. Het is niet altijd eenvoudig om steekproefgegevens op een correcte manier te verzamelen. Aan het verzamelen van gegevens moet bij elk statistisch onderzoek dan ook de nodige aandacht besteed worden. In deze context wordt soms de afkorting GIGO[1] gebruikt. Dit staat voor *garbage in, garbage out* en slaat op het feit dat zelfs de meest geavanceerde statistische methoden weinig tot geen betrouwbare informatie kunnen halen uit gegevens van slechte kwaliteit. Tegenwoordig worden we overspoeld door grote aantallen steekproefgegevens, maar vaak zijn die data van povere kwaliteit. Statistische modellen die getraind worden op gegevensbestanden van povere kwaliteit, vaak onder de noemer artificiële intelligentie, zijn dan eveneens van povere kwaliteit.

Voorbeeld 1.6. Bij het peilen naar het kiesgedrag bij Belgische verkiezingen is de populatie erg duidelijk te omschrijven: alle kiesgerechtigde burgers van het land. Variabelen die in deze context geregistreerd kunnen worden zijn geslacht, beroep, politieke overtuiging en leeftijd.

[1]Deze afkorting parodieert de afkortingen FIFO (first in first out) en LIFO (last in first out), die in accounting gebruikt worden bij het boeken van artikelen uit voorraad.

Voorbeeld 1.7. Het opgooien van een dobbelsteen is een proces. Een mogelijke steekproef bestaat erin de dobbelsteen vijftig keer op te gooien. Variabelen die bij elke worp of waarneming geregistreerd kunnen worden zijn het aantal gegooide ogen of het al dan niet even zijn van het gegooid aantal ogen.

In de Voorbeelden 1.2 en 1.3 kunnen alle tijdstippen waarop het productieproces in werking is als de populatie beschouwd worden. Op een beperkt of eindig aantal tijdstippen kunnen metingen of waarnemingen omtrent de variabelen luchtvochtigheid (Voorbeeld 1.2) of gewicht (Voorbeeld 1.3) verricht worden. De metingen vormen samen de steekproef. Voor de financiële analist uit Voorbeeld 1.5 wordt de steekproef gevormd door een beperkte reeks rendementen en marktindices. In Voorbeeld 1.4 is de populatie waarnaar de interesse van de onderzoeker uitgaat de verzameling van alle klanten van het grootwarenhuis. Een mogelijke steekproef bestaat uit alle klanten die het warenhuis in de maand mei bezocht hebben en van hun klantenkaart gebruik gemaakt hebben.

De in een steekproef verzamelde gegevens kunnen op allerlei manieren overzichtelijk voorgesteld worden met behulp van tabellen en grafieken. Daarnaast laat ook het berekenen van een aantal kenmerkende waarden of statistieken, zoals bijvoorbeeld een gemiddelde, het toe om een overzichtelijk beeld van een verzameling gegevens samen te stellen. Het voorstellen van steekproefgegevens valt onder de noemer **beschrijvende** of **descriptieve statistiek**. Dit onderdeel komt aan bod in de Hoofdstukken 2 en 3.

Het beschrijven van de steekproefgegevens is in veel gevallen slechts een eerste stap in een onderzoek. Een tweede onderdeel omvat het analyseren en interpreteren van de steekproefgegevens. Deze analyse en interpretatie zijn vereist om een antwoord te vinden op een aantal vooraf gestelde vragen over de populatie of het proces, om gestelde hypothesen te testen, of om de waarde of kwaliteit van een voorgesteld statistisch model te toetsen. De antwoorden en conclusies die hierbij bekomen worden, worden veralgemeend naar de populatie of het proces. Deze veralgemening wordt **inferentie** genoemd, wat meteen de benaming **inferentiële statistiek** verklaart. Een andere vaak gebruikte benaming is **verklarende statistiek**.

De veralgemening van conclusies betreffende een reeks steekproefgegevens naar een gehele populatie of naar een proces is meteen de zwakke plek van de statistiek: op basis van steekproefgegevens kunnen nooit met zekerheid uitspraken gedaan worden over de populatie of het proces in kwestie. Aan de gedane uitspraken kan wel een betrouwbaarheid meegegeven worden indien bij het verzamelen van de steekproefgegevens statistisch verantwoorde methoden gebruikt werden. Deze graad van betrouwbaarheid wordt uitgedrukt met behulp van een kans, zodat een basiskennis van de kansrekening vereist is om statistische methoden te kunnen begrijpen en te kunnen toepassen.

1.5 Kansrekening

Nog meer dan het begrip statistiek klinken de woorden kans en waarschijnlijkheid vertrouwd in de oren. Zo heeft iedereen intuïtief een goed idee van de betekenis van een kans van 1/4 bij deelname aan een gokspel. Een dergelijke kans kan dan ook door vrijwel ieder-

een gebruikt worden om af te wegen of men al dan niet aan het spel zal deelnemen. De berekening van zo'n kans kan evenwel voor grotere moeilijkheden zorgen.

De kansrekening bestudeert processen of experimenten waarbij de uitkomst onzeker is. De begrippen proces en experiment dienen hier in hun breedste betekenis geïnterpreteerd te worden. Voorbeelden zijn het gooien van een dobbelsteen, de prijs van een aandeel bij het sluiten van de Nasdaq, de hypothecaire rentevoet, de vraag naar laptopcomputers van een bepaald merk, het percentage defecte producten in een productielijn gedurende een bepaalde periode, het aantal bezoekers op een website of het lukraak trekken van een winnaar uit alle deelnemers aan een tombola.

Het verschil tussen de kansrekening en de statistiek bestaat erin dat de kansrekening populaties en processen rechtstreeks bestudeert, terwijl de statistiek dit doet via steekproefgegevens. De kansrekening vertrekt hierbij telkens van een aantal veronderstellingen, aannames of assumpties omtrent de populatie of het proces. Een aantal voorbeelden verduidelijken dit.

Voorbeeld 1.8. Indien het bestudeerde proces het opgooien van een dobbelsteen is, dan kunnen we met behulp van de kansrekening de kans proberen te berekenen om minstens 20 keer een zes te gooien wanneer we de dobbelsteen 100 keer opgooien. Deze berekening is enkel mogelijk indien een belangrijke veronderstelling gemaakt wordt omtrent de gebruikte dobbelsteen, namelijk dat de dobbelsteen eerlijk is, of, met andere woorden, dat de dobbelsteen volledig homogeen en totaal symmetrisch is, zodat het even waarschijnlijk is om een één te gooien als een twee, een drie, een vier, een vijf of een zes.

Een mogelijk statistisch onderzoek over de dobbelsteen zou kunnen zijn het nagaan van de eerlijkheid van de dobbelsteen. Hiertoe kan de dobbelsteen een (groot) aantal keer opgegooid worden om op die manier de nodige steekproefgegevens te verzamelen. Vervolgens kan een statistisch onderbouwde conclusie getrokken worden omtrent de hypothese dat de dobbelsteen eerlijk is.

Voorbeeld 1.9. Bij een industrieel vulproces kan, gegeven een aantal instellingen van de vulmachine en een aantal veronderstellingen inzake de nauwkeurigheid van de machine, de kans berekend worden dat een fles te weinig gevuld zal zijn. Een andere mogelijkheid is om de kans te berekenen dat er bij een levering van 1000 kratten ten hoogste 5% van de flessen zullen zijn met een te kleine inhoud.

Een statistisch onderzoek van hetzelfde vulproces bestaat typisch uit het regelmatig wegen van een aantal flessen (steekproef) om na te gaan of de gemiddelde inhoud van de flessen niet te groot of te klein is, en of de variabiliteit of veranderlijkheid in de inhoud van fles tot fles niet te groot is.

Voorbeeld 1.10. De kansrekening maakt bij het bestuderen van het kiesgedrag van de Belgische bevolking de veronderstelling dat 30% voor partij A zal stemmen, 25% voor

partij B, 20% voor partij C, en 25% blanco of voor een aantal kleinere partijen. De kansrekening kan in dat geval berekenen dat van elke 500 kiezers er gemiddeld 150 voor partij A zullen opteren, 125 voor partij B, 100 voor partij C en 125 blanco zullen stemmen of voor andere partijen kiezen.

De statistiek zal daarentegen een statistische voorspelling maken aan de hand van een steekproef van bijvoorbeeld 2000 kiezers. Bij deze voorspelling kan ook een foutenmarge gegeven worden.

Belangrijk is dus dat bij statistiek gewerkt wordt met een beperkte hoeveelheid steekproefinformatie (met andere woorden, met een beperkt aantal gegevens) en dat uitspraken over populaties en processen daarom fout kunnen zijn. Dit is de zwakte van statistiek. Idealiter zijn de kansen op fouten natuurlijk klein. De kans op fouten kan kleiner gemaakt worden door op oordeelkundige wijze veel kwaliteitsvolle gegevens te verzamelen. Het is evenwel een feit dat je nu eenmaal pech kunt hebben bij het verzamelen van gegevens en dat je gegevens je tot een verkeerde conclusie leiden, ook al voer je een gepaste statistische analyse uit op correcte wijze.

Kansberekening heeft ook een zwakke plek: de veronderstellingen omtrent het bestudeerde proces of de bestudeerde populatie kunnen fout zijn, waardoor de conclusies ongeldig worden.

1.6 Software

Bij kansrekening en statistiek zijn vaak heel wat berekeningen nodig. Ook is het belangrijk om overzichtstabellen te maken van alle gegevens in een steekproef, of om grafische voorstellingen te maken van de gegevens. Het gebruik van een computer en van gespecialiseerde statistische software is dan noodzakelijk. Zoals aangegeven in het Voorwoord gebruiken we in dit boek het statistisch softwarepakket JMP.