

Inleiding tot statistiek en kansrekening

Peter Goos

acco
learn

Voorwoord

Dit boek is het resultaat van een herwerking en compilatie van de boeken “Beschrijvende Statistiek en Kansrekenen” en “Verklarende Statistiek: Schatten en Toetsen”, die ik ontwikkelde voor de opleidingsonderdelen “Statistiek met (bedrijfs)economische toepassingen 1” en “Statistiek met (bedrijfs)economische toepassingen 2” voor studenten Handelsingenieur, Handelsingenieur in de Beleidsinformatica, Sociaal-Economische Wetenschappen en Toegepaste Economische Wetenschappen aan de Faculteit Toegepaste Economische Wetenschappen van de Universiteit Antwerpen. Het belangrijkste doel van de herwerking was om een breder gamma aan toepassingsdomeinen aan bod te laten komen in de voorbeelden en illustraties. Dit boek bevat daarom meer voorbeelden die ook studenten bio-ingenieur, burgerlijk ingenieur, industrieel ingenieur en wetenschappen zouden moeten aanspreken, zonder het oorspronkelijke doelpubliek hierbij uit het oog te verliezen. Dit boek is daarom geschikt voor gebruik in een brede waaier aan universitaire opleidingen. Het boek bevat een aanzienlijk aantal figuren en tabellen, met als doel meer inzicht te verschaffen in de concepten en de formules die doorheen het boek aangehaald worden.

In vergelijking met andere basishandboeken over statistiek komen een aantal extra onderwerpen aan bod. In het gedeelte over beschrijvende statistiek en kansrekening komen nominale en ordinale spreidingsindices, kurtosis of gepiekttheid, de rangcorrelatiecoëfficiënt, en de gamma, de Weibull-, de lognormale en de beta kansdichtheden aan bod. Ook de multinomiale kansverdeling en de multivariate normale kansdichtheid worden behandeld. In het gedeelte over verklarende statistiek wordt ruime aandacht besteed aan niet-parametrische toetsen, zoals de tekentoets, de rangtekentoets en de rangsomtoets. Daarnaast wordt het concept van equivalentietoetsen geïntroduceerd en is er een hoofdstuk over de beginsele van het optimaal ontwerp van experimenten of optimale statistische proefopzet. Bij de niet-parametrische toetsen worden exacte p -waarden uitvoerig belicht, naast de bekendere benaderende p -waarden.

Het boek maakt gebruik van het statistisch softwarepakket JMP (spreek uit als “jump”). 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 bachelor- en masterprogramma’s Handelsingenieur, Handelsingenieur in de Beleidsinformatica, Sociaal-Economische Wetenschappen, Toegepaste Economische Wetenschappen, Bio-Ingenieur, Burgerlijk Ingenieur, Industrieel Ingenieur en Wetenschappen.

Sinds 2025 is JMP Student Edition gratis beschikbaar voor alle studenten en personeelsleden van academische instellingen en dus van universiteiten. Elke lesgever en student kunnen bijgevolg JMP downloaden op hun 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 te vinden in het menu “For Users” onder de noemer “Learn to Use JMP”. Daar is onder meer de online cursus met als titel “Getting Started with JMP: On Demand Course” te vinden.

Het boek veronderstelt dat de studenten een degelijke wiskundige voorkennis hebben, 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.

Bijgevolg combineert dit boek wiskundige en conceptuele diepgang met een introductie van een statistisch pakket dat door veel grote bedrijven en onderzoeksinstellingen gebruikt wordt, wereldwijd. Door deze combinatie onderscheidt dit boek zich van veel moderne tekstboeken statistiek en kansrekening. Het is mijns inziens belangrijk dat studenten zowel aan de slag kunnen met state-of-the-art software als een grondig inzicht hebben in het hoe en het waarom van alle nuttige statistische methoden.

Er waren dan ook twee belangrijke drijfveren voor het schrijven van dit boek. Eerst en vooral wilde ik een boek schrijven dat verder gaat dan veel bestaand lesmateriaal wat betreft de technische en wiskundige achtergrond. Er is immers een toenemende tendens om technische aspecten en wiskundige afleidingen onder de mat te vegen. Met dit boek roei ik dus tegen de stroom in. Ik ben er immers van overtuigd dat studenten aangemoedigd moeten worden om hun wiskundige kennis toe te passen en dat dit hen helpt om de statistische concepten beter te begrijpen. Het gebruik van dit boek vergt hierdoor uiteraard enige wiskundige voorkennis. Gezien de aandacht die aan wiskunde besteed wordt in het secundair onderwijs en in inleidende vakken wiskunde aan de universiteit zou die voorkennis bij zowat elke student aanwezig moeten zijn. Een tweede drijfveer voor het schrijven van het boek is dat ik ervan overtuigd ben dat studenten moeten leren om de statistische concepten in de praktijk om te zetten. Daarom wordt in dit boek uitvoerig beschreven hoe alle beschrijvende statistieken berekend kunnen worden, hoe alle kansberekeningen en hypothesetoetsen uitgevoerd kunnen worden en hoe alle populatieparameters geschat kunnen worden met behulp van JMP.

Ik ben van mening dat inleidende cursussen statistiek zoveel mogelijk gebruik dienen te maken van gebruiksvriendelijke software. Mits demonstraties in hoorcolleges en het gebruik ervan in werkcolleges kan een dergelijke software studenten enthousiasmeren voor statistiek, terwijl een ongebruiksvriendelijk pakket elk potentieel enthousiasme voor statistiek met grote waarschijnlijkheid in de kiem smoort. De kans dat een student in zijn latere beroeps carrière statistiek gebruikt zal veel groter zijn wanneer het gebruik van statistiek tijdens zijn opleiding een plezier was dan wanneer het een moeizaam traject was.

Bij dit boek horen diverse datasets en educatieve JMP scripts. Deze datasets en scripts zijn beschikbaar via de website van uitgever ACCO.

Ik heb tijdens het schrijven van dit boek tal van bronnen geraadpleegd. Ik wil in dit voorwoord graag de meest belangrijke vermelden. Een bron over niet-parametrische technieken die ik extreem waardevol vond, gezien zijn diepgang en bevattelijke uitleg, is het boek *Nonparametric Statistical Methods* van Hollander, Wolfe en Chicken. Een algemener boek dat ik zeer nuttig vond is *Handbook of Parametric and Nonparametric Statistical Procedures* van Sheskin. Hetzelfde geldt voor *Biostatistical Analysis* van Zar.

Bij de uitgave van dit boek wil ik graag Leonids Aleksandrovs, Kris Annaert, Stefan Becuwe, Evi Breynders, Marco Castro, Filip De Baerdemaeker, Eline De Groof, Tristan De Rijck, Eva Engelen, Jeroen Felis, Gerline Gheysens, Hajar Hamidouche, Thibault Hendrickx, Roselinde Kessels, David Meintrup, Rafaël Michiels, Daniel Palhazi Cuervo, Tuur Panneels, Ida Ruts, Bagus Sartono, Evelien Stoffels, Anja Struyf, Utami Syafitri, Yahri Tillemans, Anil Haydar Topal, Ellen Vandervieren, Katrien Van Driessen, Philippe Van Gerwen, 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. Tevens gaat mijn dank uit naar Kristel Van Rompay en Peter Willemé voor de creatie van enkele tabellen en figuren, en voor de inspiratie bij de uitwerking van enkele voorbeelden. Katrien Van Driessen, Ellen Vandervieren en Stefan Becuwe verdienen een speciaal bedankje voor hun werk aan de *slides*, net als Miet Maertens voor het aanleveren van diverse nieuwe voorbeelden uit de bio-ingenieurssfeer. Ik dank ook Ian Cox, Bradley Jones, Volker Kraft, Brady Brady en Mia Stephens van JMP voor hun hulp met betrekking tot het softwarepakket JMP. Ten slotte wil de auteur ook Willy Gochet van de KU Leuven en 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 nieuwe herwerking van dit 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, onderzoeksinstellingen en de wetenschappen de statistiek meer toepassingen vindt dan de kansrekening, krijgt statistiek in de universitaire opleidingen 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. Statistiek steunt namelijk heel erg op kansrekening. Voor een goed begrip van statistiek is dus een degelijke basiskennis inzake kansrekening noodzakelijk. In dit boek komen daarom zowel kansrekenen als statistiek aan bod.

1.1 Waarom statistiek?

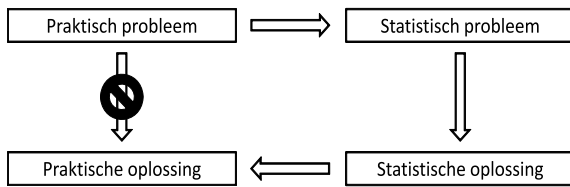
Statistiek is sinds jaar en dag een vak, niet zelden een gevreesd vak, in een groot aantal studierichtingen aan universiteiten en hogescholen. De reden hiervoor is dat heel wat mensen tijdens hun beroepswerkzaamheden vroeg of laat met problemen van gegevensanalyse geconfronteerd worden. Een degelijke 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 lessen nog niet in aanraking gekomen zijn met allerlei praktische beslissingsproblemen waarmee managers, economen, 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 geworden door niet-statistici. Sir Ronald Fisher was een geneticus die landbouwonderzoek verrichtte 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 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 aan dat de traditionele



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

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 allerlei 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, economen en wetenschappers 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 statistiek zo goed als universiteitsbreed gedoceerd wordt, veelal in bacheloropleidingen, maar her en der ook in masterprogramma's.

In de loop der jaren zijn er specifieke statistische methoden ontwikkeld voor typische ingenieurtoepassingen, 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 nati-

- onaal product (BNP), ... (deze statistieken noemt men economische statistieken),
- een bedrijf of sector: verkoopcijfers, 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 klantgegevens (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 allerlei 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 onderzoeksinstituten 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 vertrek 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 verkregen 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 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.

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 gegooide 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 verkregen 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 iedereen 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, het meten van het vochtgehalte van een pakje gemalen koffie, 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.

Deel I

Beschrijvende statistiek

2

DATA EN HUN VOORSTELLING

A microphone in the sidewalk would provide an eavesdropper with a cacophony of clocks, seemingly random like the noise from a Geiger counter. But the right kind of person could abstract signal from noise and count the pedestrians, provide a male/female breakdown and a leg-length histogram ...

(uit Cryptonomicon, Neal Stephenson, p. 147)

Data of gegevens bevatten informatie omtrent één of meerdere kenmerken of eigenschappen van een aantal elementen van een populatie of een aantal objecten gegenereerd door een proces. Omdat die kenmerken of eigenschappen variëren van element tot element, of van object tot object, noemt men deze kenmerken of eigenschappen in het statistisch jargon variabelen. Het soort kenmerken, eigenschappen of variabelen waarover informatie wordt verzameld is uiteenlopend.

2.1 Soorten gegevens en meetschalen

Variabelen worden ingedeeld naargelang hun meetniveau, afhankelijk van de schaal waarop ze worden gemeten. Nominale en ordinale variabelen zijn zogenaamde categorische of kwalitatieve variabelen. Daarnaast zijn er ook nog kwantitatieve variabelen.

2.1.1 Categorische of kwalitatieve variabelen

Nominale variabelen

Met behulp van een nominale variabele kunnen de elementen van een steekproef of van een populatie geïnclassificeerd worden: de waarde van de variabele plaatst een element in een bepaalde klasse of categorie. Voorbeelden van dergelijke variabelen zijn

- appelvariëteit (Jonagold/Pink Lady/Kanzi/...),
- nationaliteit (Belg/Nederlander/...),
- godsdienst (katholiek/protestants/islamitisch/...),
- woonplaats (Antwerpen/Leuven/Gent/...).

Soms kan het nuttig zijn labels, codenummers of -letters toe te kennen aan de verschillende klassen of categorieën. Zo kan een Belg de code “1” krijgen, een Nederlander de code “2”, een Fransman de code “3”, en een Duitser de code “4”. Als code voor de woonplaats gebruikt men bijvoorbeeld de postcode. Belangrijk hierbij is op te merken dat deze cijfers geen volgorde en/of hoeveelheid impliceren. Daarom zijn, behalve het berekenen van frequenties en percentages, de meeste rekenkundige bewerkingen op nominale variabelen zinloos.

Ordinale variabelen

Indien een nominale variabele een logische ordening tussen de elementen van een steekproef toelaat, dan is de variabele ordinaal. Een typisch voorbeeld van een ordinale variabele wordt aangetroffen in allerlei enquêtes. Daar wordt de respondenten vaak gevraagd of zij de kwaliteit van een product of dienst als “1: zeer goed”, “2: goed”, “3: matig”, “4: slecht” of “5: zeer slecht” zouden bestempelen. In andere enquêtes wordt de respondenten gevraagd of zij het “1: helemaal oneens”, “2: eerder oneens”, “3: noch eens, noch oneens”, “4: eerder eens” of “5: helemaal eens” zijn met een bepaalde stelling. Andere voorbeelden van ordinale variabelen zijn het aantal Michelinsterren van restaurants, het aantal sterren van hotels, en expertevaluaties over de bitterheid van wijn. Expertevaluaties over bitterheid worden vaak uitgedrukt op de volgende schaal: “1: helemaal niet bitter”, “2: een beetje bitter”, “3: redelijk bitter”, “4: heel bitter” of “5: buitengewoon bitter”.

Een ordinale schaal heeft geen vaste meeteenheid. Dit wil zeggen dat het verschil tussen de verschillende niveaus op de meetschaal niet in een aantal eenheden kan uitgedrukt worden. Het verschil tussen een hotel met drie sterren en één met twee sterren is bijvoorbeeld niet noodzakelijk hetzelfde als het verschil tussen een hotel met twee sterren en één met slechts één enkele ster. Het spreekt vanzelf dat het ook bij ordinale variabelen weinig zinvol is om rekenkundige bewerkingen uit te voeren.

2.1.2 Kwantitatieve variabelen

Een variabele die gemeten wordt op een kwantitatieve schaal kan uitgedrukt worden in een aantal vaste meeteenheden. Voorbeelden zijn lengte, oppervlakte, inhoud, gewicht, duurtijd, aantal bits per tijdseenheid, prijs, inkomen, wachttijd, aantal bestelde goederen, Op kwantitatieve variabelen kunnen vrijwel alle rekenkundige bewerkingen zinvol worden uitgevoerd. Het verschil tussen verschillende waarden van een variabele kan, in tegenstelling tot bij een ordinale variabele, in een aantal eenheden uitgedrukt worden. Binnen de kwantitatieve variabelen wordt nog een onderscheid gemaakt tussen variabelen die gemeten worden op een intervalschaal en variabelen die op een ratioschaal gemeten worden.

Intervalschaal

Een intervalschaal bezit geen natuurlijk nulpunt, dit wil zeggen geen natuurlijke ondergrens. Bij variabelen gemeten op een intervalschaal kunnen verhoudingen niet zinvol berekend worden. Bekende voorbeelden van intervalvariabelen zijn de tijd afgelezen op een klok en de temperatuur uitgedrukt in graden Celsius of Fahrenheit. Het verschil tussen 2 uur en 4 uur is hetzelfde als het verschil tussen 21 uur en 23 uur, maar het is niet zo dat 4 uur twee keer zo laat is als 2 uur. Dit is te wijten aan het feit dat de tijd afgelezen op een klok geen absoluut nulpunt heeft. Hetzelfde geldt voor de temperatuur gemeten in graden Celsius: 20 graden Celsius is niet vier keer zo warm als 5 graden Celsius.

Ratioschaal

Een ratioschaal heeft wel een absoluut nulpunt. Verhoudingen kunnen daarom bij gegevens die gemeten worden op een ratioschaal zinvol berekend worden. Een lengte van 6 cm is dubbel zoveel als een lengte van 3 cm omdat de lengteschaal een absoluut nulpunt heeft. Analooq is een bestelling van 6 producten dubbel zo groot als een order van 3 producten. De temperatuur in Kelvin heeft wel een absolute ondergrens, zodat deze temperatuur wel op een ratioschaal gemeten wordt. De reden hiervoor is dat 0 Kelvin de koudst mogelijke temperatuur is en dus een absolute ondergrens voor de temperatuur is.

Discrete versus continue variabelen

Een discrete variabele kan slechts een eindig of oneindig aftelbaar aantal verschillende waarden aannemen, terwijl een continue variabele een continuüm van waarden kan aannemen. Voorbeelden van discrete variabelen zijn het aantal passagiers op een lijnvlucht, het aantal kinderen in een gezin of het aantal biggen geworpen door een zeug. Voorbeelden van continue variabelen zijn lengte, duurtijd, gewicht en body mass index.

In de praktijk zullen alle waarnemingen van een continue variabele discreet zijn: een continue lengte wordt in de praktijk gemeten tot op een bepaalde nauwkeurigheid (bijvoorbeeld een millimeter) en aldus discreet gemaakt. Toch zullen we die lengte dan bestuderen alsof het een continue variabele betrof.

2.1.3 Hiërarchie van meetschalen

Uit het bovenstaande blijkt duidelijk dat er een hiërarchie zit in de meetschalen. De hoogste of meest informatieve meetschaal is de ratioschaal, gevolgd door de intervalschaal, en de ordinale en nominale meetschalen. Gegevens die op een bepaalde schaal gemeten worden, kunnen omgevormd worden tot gegevens van een lagere meetschaal. Gegevens gemeten op een ratioschaal (bijvoorbeeld lengte) zijn natuurlijk intervalgeschaald (het verschil tussen 6 cm en 3 cm is hetzelfde als het verschil tussen 15 cm en 12 cm, ordinaal (lengtes kunnen zinvol geordend worden) en nominaal (lengtes kunnen ingedeeld worden in klassen). Omgekeerd kunnen nominale gegevens nooit omgevormd worden tot ordinale of kwantitatieve gegevens. Daarom zullen in het vervolg van het boek alle technieken die toepasbaar zijn op nominale gegevens, automatisch toepasbaar zijn op ordinale en kwantitatieve gegevens. Alle technieken die toepasbaar zijn op ordinale gegevens kunnen tevens zinvol gebruikt worden bij kwantitatieve gegevens. Tussen gegevens gemeten op een intervalschaal en gegevens gemeten op een ratioschaal wordt zelden een onderscheid gemaakt.

2.1.4 Meetschalen in JMP

Het softwarepakket JMP maakt een onderscheid tussen nominale, ordinale en kwantitatieve variabelen. Het pakket verwijst naar de meetschaal als “Modeling type”, en gebruikt de Engelstalige termen “Nominal”, “Ordinal” en “Continuous” voor respectievelijk nominaal, ordinaal en kwantitatief.

2.2 De datamatrix of gegevenstabel

Gegevens worden vaak voorgesteld in een matrix of tabel waarbij de rijen de elementen of waarnemingen van een steekproef voorstellen en de kolommen de verschillende gemeten variabelen. Een rij in een datamatrix of gegevenstabel wordt soms observatievector genoemd.

Voorbeeld 2.1. Figuur 2.1 bevat gegevens uit een onderzoek van Testaankoop (nummer 448, november 2001) over een aantal eigenschappen van Spaanse rode wijnen. De steekproef bevat 70 wijnen. Figuur 2.2a toont de verschillende symbolen die het pakket JMP gebruikt om de verschillende meetschalen, “Nominal”, “Ordinal” en “Continuous”, aan te duiden. De variabele “Merknaam” is een nominale variabele. De variabele “Oordeel” is een ordinale variabele. De overige variabelen zijn kwantitatief. De meetschaal van een variabele kunt u in JMP veranderen door met de rechtermuisknop op de naam van een kolom te klikken en vervolgens “Column Info” te selecteren. Een snelle methode om de meetschaal aan te passen, is klikken op het symbool voor de meetschaal bij de naam van de variabele, zoals geïllustreerd in Figuur 2.2b.

In het vervolg van dit hoofdstuk worden voornamelijk zogenaamde univariate en bivariate voorstellingen van variabelen behandeld. Een univariate voorstelling heeft betrekking op één variabele, terwijl een bivariate voorstelling betrekking heeft op twee variabelen. Op

Spaanse wijnen - JMP Student Edition

File Edit Tables Rows Cols DOE Analyze Graph Student Tools View Window Help

Spaanse wijnen

Scripts Filter Views

	Merknaam	% alcohol	% gemeten	Prijs	Oordeel
1	Castillo de Almansa	12,5	12,9	4,88	U
2	Coto de Imaz	12,5	12,8	8,55	U
3	El Meson	13	13,1	5,85	U
4	Mas Collet Capcanes	13,5	14,1	8,18	G/U
5	Los Condes	12,5	12,7	4,19	G/U
6	Sangre de Toro	13,5	13,4	6,2	G/U
7	Vina Cobranza	12,5	13,4	8,16	G/U
8	Rincon de Navas	12	13,3	10,14	G/U
9	Coronas	13	12,9	6,82	G
10	Jumilla-Monastrel	13,5	14,4	2,55	G
11	Guerra	13,5	13,2	8,06	G
12	Baron de Ley	13	13,2	9,12	G
13	Hecula	13,5	13,8	9,02	G
14	Senorio de los llanos	12,5	12,6	4,93	G
15	Carchelo	13	13,2	4,96	G
16	Castillo Maluenda	13	12,6	5,26	G
17	Cabernet Sauvignon	13,5	13,5	5,83	G
18	Vega Saucó	13,5	14	5,85	G
19	Arco Viejo	12,5	13,1	5,92	G
20	Marques de Caceres	13	13,2	6,92	G
21	Valdubon	13	13,2	10,06	G
22	Vina Albina	13	13,4	11,53	G
23	Faustino V	13	13,1	11,63	G
24	Ribon	13	13,5	12,27	G
25	Clos de Torribas Pinord	12	12,3	4,86	R/G
26	Castillo de Nelida	12	12,3	3,1	R/G
27	Senorio de los llanos	12,5	12,6	3,59	R/G
28	El Coto	12,5	12,7	5,68	R/G
29	Vina del Recuerdo	12,5	12,5	5,92	R/G
30	Camparron	13,5	13,7	8,55	R/G
31	Dionisos	12,5	13,1	9,97	R/G
32	Vina Salceda	13	13,3	12,07	R/G
33	El Hidalgo	12,5	12,9	2,21	R/G
34	Hoya de Cadenas	12,5	12,5	4,02	R/G

Columns (5/0)

Filter

- Merknaam
- % alcohol
- % gemeten
- Prijs
- Oordeel

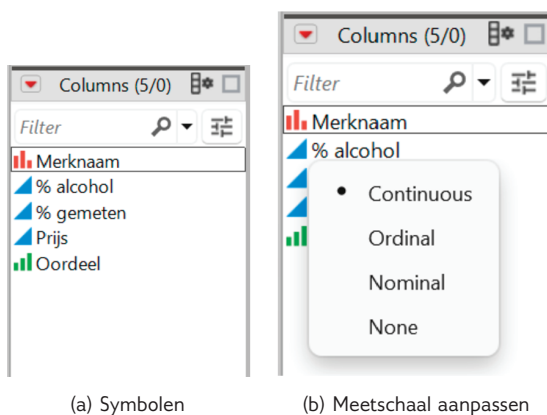
Tags

Rows

All rows 70
Selected 0
Excluded 0
Hidden 0
Labeled 0

evaluations done

Figuur 2.1. Gedeelte van de datamatrix of gegevenstabel in verband met Spaanse rode wijnen.



Figuur 2.2. Symbolen gebruikt door het pakket JMP voor de verschillende meetschalen.