# Samenvatting

## Het doen van verdelingsuitspraken om de grenzen van het weten te preciseren

De vooruitgang in de wetenschap vindt grotendeels plaats door rationele argumenten te combineren met empirische gegevens. In de filosofie en de zuivere wiskunde spelen rationele argumenten de hoofdrol, terwijl men zich in de toegepaste wetenschap bezighoudt met het verzamelen en interpreteren van data. In de wiskundige statistiek tracht men deze aspecten te combineren. Primair gaat het er dan om statistische uitspraken te doen omtrent iets onbekends. Die uitspraken kunnen dan dienen als basis bij verdere beschouwingen, bijvoorbeeld gericht op het nemen van een beslissing. De wiskundig-statisticus tracht dus methoden voor het doen van zulke uitspraken te leveren. Deze methoden moeten niet afhangen van 'the intentions that might be furthered by utilizing the knowledge inferred' [2]. Indien de beschikbare data te beperkt zijn, dan kunnen verschillende uitwerkingen tot verschillende uitspraken leiden. De statisticus zal zich van een specifiek antwoord onthouden als de verschillen 'te groot' zijn. Wanneer zo'n uitspraak wel gegeven wordt, dient hierbij de onzekerheid betrokken te worden, bijvoorbeeld door de uitspraak in de vorm van een kansverdeling te gieten, of door resultaten van diverse uitwerkingen te vermelden.

Een typerend voorbeeld is als volgt. De ornitholoog G.Th. de Roos observeert een populatie Steenlopers (*Arenaria interpres*) op Vlieland. Een deel van deze vogels is geringd, maar dit ring-nummer is niet altijd te lezen, bijvoorbeeld omdat een andere vogel het zicht belemmert. Na hoeveel dagen van observeren mag De Roos er van uit gaan dat alle aanwezige geringde vogels minstens eenmaal geobserveerd zijn? Dit kan beantwoord worden door een verdelingsuitspraak te maken over het aantal nog niet geziene maar wel aanwezige geringde vogels, inclusief een kansuitspraak over de hypothese dat alle geringde vogels gezien zijn. Vanzelfsprekend hangen de resultaten enigszins af van de kanstheoretische veronderstellingen die men maakt en de statistische principes die men hanteert.

Het eerste deel van het proefschrift bestaat uit 'vingeroefeningen' die illustreren dat informatie over het onbekende alleen van waarde kan zijn wanneer bekend is volgens welk mechanisme deze informatie tot stand is gekomen. In de theoretische kansrekening wordt met informatie omgegaan door ernaar te conditioneren. In de statistische praktijk levert dit problemen doordat er vaak onduidelijkheid heerst over de gezamenlijke verdeling van de toevallige grootheden $X$ en $Y$ die zitten achter de waarnemingen

[2] R.A. Fisher, *Statistical Methods and Scientific Inference*, third edition, Macmillan, New York, 1973, p. 107

$x$ en de onbekende $y$. Eerst wordt dit uitvoerig uitgewerkt voor een voorbeeld met een worp met een zuivere dobbelsteen. Zo mag uit de informatie 'er is een even aantal ogen gegooid' niet automatisch de conclusie getrokken worden: 'de kans dat er een zes gegooid is, is één derde'. De wijze waarop de informatiebron werkt, moet meegenomen worden in het statistische model. Vervolgens wordt een gelijksoortig vraagstuk, het twee enveloppen probleem, bekeken. Net als in het eerste voorbeeld speelt de moeilijkheid omtrent de numerieke specificatie van conditionele kansen een grote rol.

Het tweede en belangrijkste deel gaat over de situatie waar men een steekproef $x_1, \ldots, x_n$ heeft uit een verdeling met een kansdichtheid $f$. Men wil deze steekproef gebruiken om een schatting van $f$ te vormen of, wat vrijwel hetzelfde is, om een verdelingsuitspraak ('distributional inference') te doen over $y = (x_{n+1})$. Een nieuwe methode wordt beschreven om de kansdichtheid $f$ te schatten, waarbij 'voorkennis' met betrekking tot $f$ in de beschouwing betrokken wordt. Dit gebeurt door een kansdichtheid $\psi$ als 'initial guess' voor $f$ te specificeren. (Tevens dient de mate van vertrouwen in deze voorkennis te worden vastgelegd.) Op basis van de steekproef $x$, de beginschatting $\psi$ (en de mate van vertrouwen er in) wordt door middel van een multi-modale aanpak, die het midden zoekt tussen de klassieke en de Bayesiaanse statistiek, een schatting $\hat{f}$ van $f$ gemaakt. Wanneer $\psi$ geen onredelijke beginschatting is, dan presteert de bijbehorende dichtheidsschatter, in het algemeen, beter dan de gebruikelijke kernschatters. Dat is geen wonder want de kernschatter maakt geen gebruik van $\psi$. Hoe de vergelijking uitpakt als de kernschatter wordt aangepast aan $\psi$ is nog onbekend.

Om de toepasbaarheid van deze methode te beschouwen, is gekeken naar omvangrijke data over de mate van verontreiniging van oppervlaktewater in Nederland. Eerder onderzoek van deze data heeft uitgewezen dat de concentraties van vervuilende stoffen doorgaans redelijk goed te beschrijven zijn met lognormale verdelingen. Een complicatie is dat de concentratie niet gemeten kan worden wanneer deze onder een bepaald waarnemingsniveau ligt. De dichtheidsschattingstheorie uit dit proefschrift, aangepast aan genoemde complicatie, is gebruikt om de 'voorkennis' van bij benadering lognormale verdelingen aan te passen aan de metingen. Dit levert betere dichtheidsschattingen dan wanneer lognormale verdelingen aan de data worden aangepast.

Deze dichtheidsschattingsmethode blijkt ook zeer geschikt te zijn voor de zogenaamde goodness of fit problematiek. Er wordt aan de hand van de data een uitspraak gedaan over de hypothese $\mathrm{H}_0$: $f = \psi$. De resulterende toetsingsmethoden hebben interessante raakvlakken met bestaande goodness of fit toestingsmethoden, zoals de $\chi^2$-toets, Kolmogorovs goodness of fit methode en de 'smooth tests' van Neyman.

Om te beklemtonen dat het doen van verdelingsuitspraken van belang is in de praktijk, is een extra voorbeeld toegevoegd uit het grensgebied tussen de multivariate analyse en de analyse van tijdreeksen.