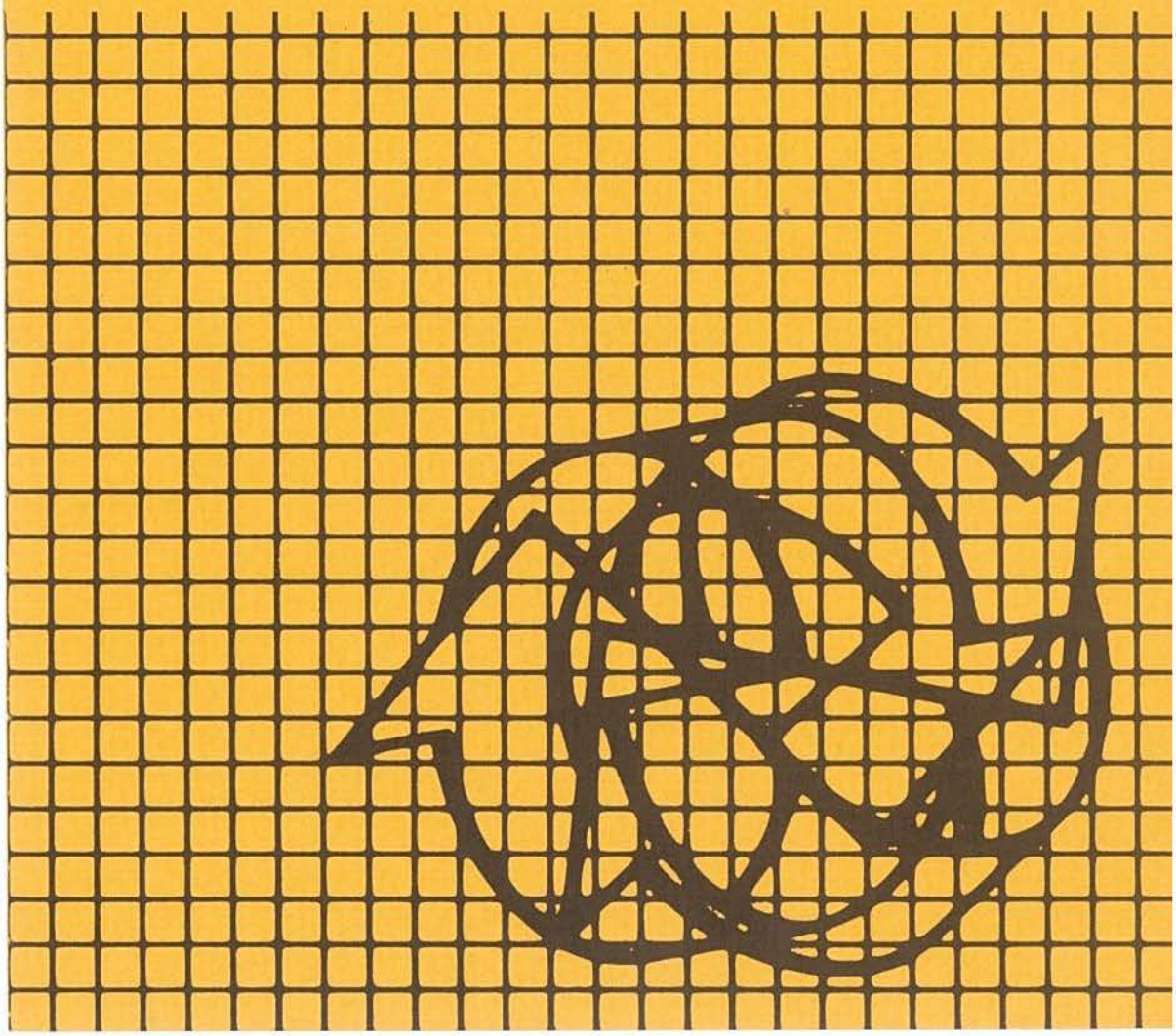
Abstracte inferentie en toegepaste statistiek

CWI

Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science



Abstracte inferentie en toegepaste statistiek

Zowel de natuurlijke theoretische ontwikkeling van de statistiek als dringende vragen uit de praktijk wijzen erop, dat het noodzakelijk is de gebruikelijke basis van de statistiek te verbreden. De structuur en aard van een gegevensverzameling waarop een statistische analyse uitgevoerd moet worden is meestal veel complexer dan het geval waarop de traditionele statistische theorie zich geconcentreerd heeft, het geval namelijk dat de gegevens bestaan uit n onafhankelijke trekkingen uit een enkele populatie van numerieke waarden. Gegevens worden vaak gegenereerd door stochastische processen, die slechts gedeeltelijk kunnen worden waargenomen met een afhankelijk, door toeval beïnvloed, mechanisme. Ook kunnen ze worden gegenereerd door ruimtelijke processen of door stochastische modellen voor geometrische objecten. Niet alleen zijn dergelijke *gegevens* complexer dan in het klassieke geval, maar dit geldt ook voor de *parameters* van de stochastische modellen die beschrijven hoe ze gegenereerd worden, en waarover men conclusies wil trekken. De noodzaak om in zulke situaties een oplossing te bieden heeft er al toe geleid, dat toegepaste statistici nieuwe patronen ontdekt hebben en nieuwe pre-theorie geformuleerd hebben.

Deze ontwikkelingen zijn alleen mogelijk geweest dankzij en vaak geïnspireerd door de ongelooflijke toename in capaciteit, snelheid en mogelijkheden (vooral grafische) van de computer in recente jaren. Dankzij het feit dat men automatisch een weergave kan maken van een curve, een contouren-kaart of een geometrisch object, wordt het denkbaar en uitvoerbaar om dit soort objecten de onderwerpen (dat wil zeggen waarnemingen of parameters of allebei) te laten zijn van een statistische analyse.

De term "abstract inference" is onlangs door U. Grenander als titel van een boek (Wiley, 1981) ingevoerd om de statistische theorie te beschrijven,

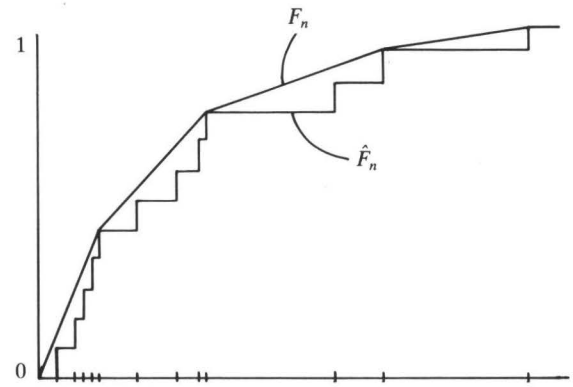
die voor zulke problemen in ontwikkeling is. Het woord "abstract" duidt op het feit dat het om statistische problemen gaat waarbij parameters of waarnemingen in andere ruimten kunnen liggen dan in de gebruikelijke laag-dimensionale Euclidische ruimten. Dit betekent geenszins dat het om puur theoretische problemen gaat. Men is bijvoorbeeld op het CWI op problemen van abstracte inferentie gestuit bij de statistische analyse van verkeersstromen op snelwegen en bij de analyse van overlevingsduren ten behoeve van klinisch kankeronderzoek.

Wij zullen hieronder een eenvoudig voorbeeld bespreken waarin enkele typische aspecten van de theorie die nodig is bij zulke toepassingen naar voren komen. De ontwikkeling van theorie op dit gebied moet momenteel gezien worden als een verkennende excursie op nog grotendeels onbekend terrein. Men bestudeert specifieke problemen in de hoop patronen te ontdekken waaruit gaandeweg een algemene theorie kan worden opgebouwd. Een belangrijk doel is na te gaan, in welke situaties klassieke methoden uit de eindig-dimensionale statistiek nog goed werken. Daarbij valt te denken aan de *methode van meest aannemelijke schatters*, die als schatting van een parameter díe parameterwaarde kiest, waarvoor de feitelijk gedane waarnemingen de grootste kans hadden om te zijn gegenereerd.

Andere aspecten van deze ontwikkeling zijn de verbanden met technieken uit vele andere hoeken van de wiskunde; het veelvuldig voorkomen van stochastische processen; en het feit dat simulatie-experimenten vaak nodig zijn als heuristisch hulpmiddel om aanwijzingen te verkrijgen omtrent de te behalen resultaten of om achteraf hun toepasbaarheid (bijvoorbeeld bij kleine steekproeven) te controleren.

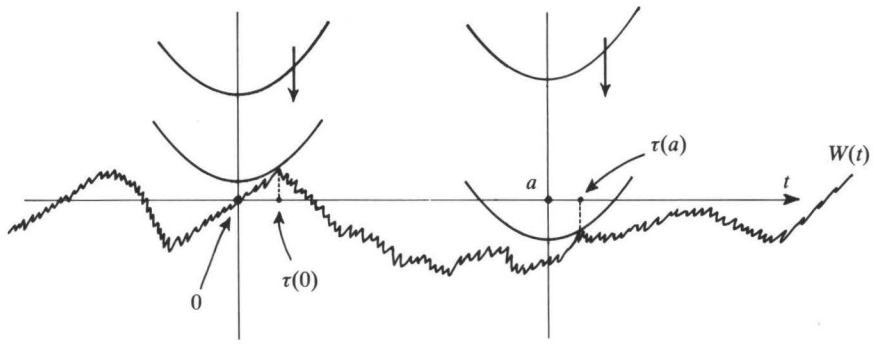
Laten we nu een voorbeeld bekijken, namelijk het schatten van een monotone kansdichtheid f (een "abstracte" parameter). De waarnemingen zijn van het klassieke soort. Wel zullen we zien dat in dit voorbeeld, zoals in vele andere, de theorie van stochastische processen via een achterdeur binnensluipt. Stel dat aselecte trekkingen gedaan worden uit een populatie van numerieke waarden x, zodanig dat hoe groter x, hoe kleiner de kans is om deze waarde aan te treffen. Bijvoorbeeld zou x de in een jaar waargenomen hoogste waterstand op een plaats aan de Nederlandse kust kunnen voorstellen, mits deze groter is dan een zekere kritieke waarde (die we voor het gemak 0 zullen nemen). Aan de hand van n waarnemingen wil men de kansdichtheid $f(x)$, $x \geqslant 0$, schatten: dus voor kleine h is $f(x) \times h$ de kans dat een waarneming in het interval x tot $x + h$ ligt. De enige aanname die gemaakt wordt is dat de dichtheid dalend is in x. De statisticus wordt nu niet gedwongen a priori (vaak nogal implausibele) parametrische veronderstellingen te maken over de onderliggende verdeling van de waarnemingen, terwijl toch de zoveel zwakkere informatie, die de eis van monotonie geeft, gebruikt wordt om een goede schatting van de dichtheid te krijgen.

De meest aannemelijke schatter van een monotoon dalende dichtheid is eenvoudig te beschrijven. De functie F_n, die in een punt x als waarde de fractie waarnemingen $\leq x$ heeft, wordt de empirische verdelingsfunctie genoemd. De meest aannemelijke schatter van f is nu de afgeleide $\hat{f}_n$ van de kleinste concave functie $\hat{F}_n$ groter dan F_n, en wordt dus verkregen door als het ware een touw over de functie F_n te spannen (zie figuur 1).



Figuur 1: Empirische verdelingsfunctie F_n en zijn concave majorant $\hat{F}_n$ voor $n = 12$. De schatting $\hat{f}_n$ van de onbekende monotone dichtheid f wordt gegeven door de afgeleide (helling) van $\hat{F}_n$.

In 1969 werd door Prakasa Rao het opmerkelijke resultaat bewezen, dat de kansverdeling van de schattingsfout $\hat{f}_n(x) - f(x)$ in een vast punt x (na standaardisatie) bij een groot aantal waarnemingen bij benadering gelijk is aan de kansverdeling van de locatie van het maximum van de tweezijdige Brownse beweging minus een parabool, dat wil zeggen de locatie $t = \tau(0)$ van het maximum van $W(t) - t^2$ waarbij $W(t)$ de standaard Brownse beweging is (een van de meest bekende stochastische processen); zie figuur 2.



Figuur 2: $\tau(a)$ is de locatie van het punt waar de parabool $(t - a)^2$, die langs de lijn $t = a$ naar beneden glijdt, de Brownse beweging $W(t)$ treft.

Dit resultaat geeft belangrijke informatie over de nauwkeurigheid van $\hat{f}_n(x)$ als schatter van $f(x)$. Onlangs is op het CWI een veel inzichtelijker bewijs hiervoor gevonden, dat ook nieuwe resultaten over de Brownse beweging levert. Het bleek daarbij dat het gedrag voor grote n van de globale afstand $\int|\hat{f}_n(x) - f(x)|dx$ van schatter $\hat{f}_n$ tot onderliggende dichtheid f beschreven kan worden met behulp van het stochastische sprongproces $\tau(a)$, waarbij $\tau(a)$ de locatie t van het maximum van het proces $W(t) - (t - a)^2$ is (zie weer figuur 2). Dit sprongproces kan weer gekarakteriseerd worden door oplossingen van bepaalde warmtevergelijkingen.

Grote simulatie-experimenten zijn uitgevoerd om te laten zien dat deze geheel onverwachte asymptotische resultaten al bij redelijke steekproefgroottes duidelijk te zien zijn. Ook is zeer veel rekenwerk nodig geweest om de bovengenoemde warmtevergelijkingen numeriek op te lossen.

Samenvattend, een eenvoudig gesteld probleem blijkt voor zijn oplossing gereedschap uit een aantal zeer verschillende gebieden van de wiskunde nodig te hebben. De klassieke methode van de meeste aannemelijkheid blijkt hier goed te werken, maar heeft een heel ander gedrag dan men uit de eindig-dimensionale statistiek gewend is.

november 1983 Richard Gill