KM 40 (1992)
pag. 99-113

WEL EN WEE VAN DE COMPUTER IN DE STATISTIEK

Rede, in verkorte vorm uitgesproken
bij het aanvaarden van het ambt van
hoogleraar in de statistische informatieverwerking
in de Faculteit der Economische Wetenschappen en Econometrie
van de Universiteit van Amsterdam
op woensdag 20 mei 1992

door

Dr. Jelke G. Bethlehem

1992

Universiteit van Amsterdam

Mijnheer Rector Magnificus,

Zeer gewaardeerde toehoorders,

Over praktische statistiek en computers

Statistiek is een tak van wetenschap die op veel terreinen wordt toegepast. Praktizerende statistici kunnen worden gevonden in andere takken van wetenschap, zoals bijvoorbeeld geneeskunde, biologie, psychologie, en sociologie, maar ook bij overheidsinstellingen en in het bedrijfsleven. Statistici moeten goed worden uitgerust voor hun werk en daarbij speelt uiteraard het universitaire onderwijs in de statistiek een belangrijke rol. Als je echter enerzijds kijkt naar wat statistici leren en anderzijds naar wat zij nodig hebben in de praktijk, dan is er sprake van verschillen.

In de eerste plaats is er het verschil tussen theoretische en praktische statistiek. In het onderwijs ligt de nadruk op het theoretische raamwerk van de statistiek, waarbinnen op basis van modelveronderstellingen mooie resultaten kunnen worden afgeleid. In de praktijk is dit raamwerk vaak niet toepasbaar, omdat niet aan de modelveronderstellingen is voldaan. Een aardige illustratie van het verschil tussen theorie en praktijk is het in veel stellingen gehanteerde uitgangspunt dat de waarnemingen worden opgevat als een reeks onafhankelijke, identiek en normaal verdeelde stochastische variabelen, terwijl in de praktijk vaker het 'Dirty Data Theorem' van toepassing is, dat stelt dat waarnemingen verkregen zijn door afhankelijke trekkingen met ongelijke en onbekende trekkingskansen uit bizarre en onspecificeerbare verdelingen, waarbij sommige waarden ontbreken en vele andere waarden onnauwkeurig zijn gemeten.

Een tweede verschil is dat het onderwijs zich wat meer richt op het analyseren van gegevens, terwijl de praktijk met name ook behoefte heeft aan aandacht voor opzet en uitvoering van een onderzoek. Bij het verzamelen van gegevens kan veel mis gaan. Problemen kunnen bijvoorbeeld ontstaan door het ontbreken van gegevens als gevolg van non-respons, maar ook de wel beschikbare gegevens kunnen zijn aangetast door meetfouten. Als de onderzoeker deze problemen niet of onvoldoende onderkent en oplost, dan kan dit de betrouwbaarheid van de uitkomsten van statistisch onderzoek aantasten, en zelfs tot wezenlijk verkeerde conclusies leiden. Het is daarom van vitaal belang dat de statisticus veel aandacht besteed aan deze aspecten van het onderzoek.

Een derde verschil tussen onderwijs en praktijk betreft het gebruik van computers. Wat studenten tijdens hun opleiding leren aan statistiek, is met name 'papieren' statistiek. Daarbij gaat het om het afleiden van stellingen en het met pen en papier oplossen van kleine schattings- en toetsingsproblemen. Praktische statistiek is door de belangrijke rol van de computer veel meer 'elektronische' statistiek. Praktisch statistisch onderzoek kan worden omschreven als het verzamelen, beschrijven, analyseren en interpreteren van gegevens die zijn verkregen door het tellen of meten van eigenschappen van verzamelingen van objecten. Bij dit werk moeten vaak omvangrijke berekeningen worden uitgevoerd. Dat wordt enerzijds veroorzaakt door de omvang van de gegevens, en anderszijds door de complexiteit van de toegepaste statistische technieken. Het is om deze reden dat de computer voor de statisticus een waardevol en onmisbaar hulpmiddel is.

De computer heeft altijd al een grote rol gespeeld in statistisch onderzoek. De capaciteit en mogelijkheden ervan hebben door de tijd heen bepaald wat mogelijk is in de statistiek. We leven nu in een tijd waarin de computer wel een hele grote vlucht heeft genomen. De snelle opkomst van de relatief goedkope microcomputer heeft ertoe geleid dat computers overal in onze samenleving zijn doorgedrongen. Dit heeft ook consequenties gehad voor het werk van de statisticus. De computer voert langdurige en complexe berekeningen sneller uit dan ooit tevoren. Dat heeft niet alleen geleid tot de ontwikkeling van nieuwe reken-intensieve statistische technieken, maar ook tot een nieuwe, andere benadering van de analyse van de verzamelde gegevens, gebaseerd op interactieve grafische exploratie.

Toepassing van de computer in de statistiek kan leiden tot beter onderzoek, maar het is niet allemaal rozegeur en maneschijn. Statistisch onderzoek kan, bewust of onbewust, ook op verkeerde wijze worden uitgevoerd. Als de gegevens op onjuiste wijze worden geïnterpreteerd, of als niet is voldaan aan de veronderstellingen waarop een analyse-techniek is gebaseerd, dan zal de geldigheid van de getrokken conclusies op zijn minst twijfelachtig kunnen zijn. Helaas is het (nog) niet zo dat de computer zelfstandig in staat is de gebruiker te behoeden voor dit soort dwalingen. Voeg daarbij het feit dat tegenwoordig veel niet-statistisch georiënteerde pakketten (spreadsheet, database en presentatie pakketten) als extra optie wat leuke statistische technieken bieden, en alle ingrediënten zijn aanwezig voor verkeerd gebruik van de statistiek. De statisticus dient zich van deze problematiek bewust te zijn.

Als gevolg van de verschillen tussen theoretie en praktijk komt de statisticus na het voltooien van zijn studie in andere wereld terecht. Hij wordt geconfronteerd met praktische moeilijkheden en mogelijkheden, die tijdens zijn studie niet aan de orde zijn geweest. In deze oratie wil ik ervoor pleiten de aankomende statisticus beter toe te rusten voor zijn werk in de praktijk. Dit betekent dat in het onderwijs meer aandacht moet worden besteed aan de praktische aspecten van de statistiek, en in het bijzonder meer aandacht voor de rol van de computer. Het bedrijven van statistiek is tegenwoordig zo verweven met het gebruik van de computer, dat een statisticus zonder computer eigenlijk geen statisticus meer is. Aan de hand van één van de toepassingsgebieden van de statistiek, het enquête-onderzoek, wil ik laten zien dat de computer een steeds grotere rol speelt, en dat dit consequenties heeft voor het werk van de statisticus.

Over enquête-onderzoek

Een enquête-onderzoek wordt uitgevoerd om iets te leren over een populatie van objecten. Veelal betreft dit personen, huishoudens of bedrijven. Aangezien een integraal onderzoek van een populatie vaak kostbaar en tijdrovend is, beperkt men zich in het algemeen tot een steekproef uit deze populatie. Karakteristiek voor enquête-onderzoek is dat gegevens worden verkregen door het stellen van vragen aan mensen. Dit is niet de meest ideale wijze van meten. De kwaliteit van een enquête-onderzoek kan worden afgemeten aan de mate van bruikbaarheid van de uitkomsten. Inzet van computers kan de kwaliteit aanzienlijk verbeteren. Om dit te kunnen uitleggen, wil ik eerst drie kwaliteitsaspecten onderscheiden.

Het eerste kwaliteitsaspect is nauwkeurigheid. De resultaten van een enquête-onderzoek komen beschikbaar in de vorm van schattingen van populatiekenmerken, zoals het gemiddelde

inkomen van een bevolkingsgroep, of het percentage gezinnen met een magnetronoven. Deze schattingen zullen niet gelijk zijn aan de exacte (maar onbekende) waarde van deze kenmerken. Uiteraard zal de statisticus trachten deze afwijking zo klein mogelijk te houden. Tot op zekere hoogte kan hij daar invloed op uitoefenen. Zo kan hij met een geschikte keuze van een steekproefontwerp, steekproefomvang, en schattingstechnieken een grote nauwkeurigheid bereiken. Andere factoren, zoals non-respons en meetfouten, zijn echter veel lastiger onder controle te houden.

Ingevulde vragenlijsten zullen vaak nog fouten bevatten. Om te voorkomen dat onjuiste conclusies uit het onderzoek worden getrokken, zullen deze fouten moeten worden opgespoord en verbeterd. Dit betekent dat na de verzameling van de gegevens nog een uitgebreid bewerkingsproces moet plaatsvinden. Daarbij tracht men zo goed en zo kwaad als dat gaat fouten in de gegevens op te sporen en te verbeteren. Dat is lang niet altijd een eenvoudige zaak, zeker niet als dat achteraf moet gebeuren door de onderzoeker. De respondent is dan immers niet meer beschikbaar om daarbij te helpen.

Het detecteren en corrigeren van fouten kost tijd, en dat brengt ons bij het tweede kwaliteitsaspect. Zeker daar waar het om beleidsondersteunend onderzoek gaat, is actualiteit een kritische factor. Daarom kan worden gesteld dat de kwaliteit van het onderzoek hoger is naarmate de cijfers sneller beschikbaar komen.

Een derde en laatste aspect van kwaliteit is het kostenaspect. Meer geld betekent beter onderzoek. Er kunnen meer gegevens worden verzameld en er kan meer aandacht worden besteed aan controle en correctie. Helaas hebben de meeste onderzoekers geen ongelimiteerd budget tot hun beschikking. Ze moeten het doen met de beperkte fondsen die hun ter beschikking staan. Geld kan maar één keer worden uitgegeven, en daarom is elke kostenbesparing die de nauwkeurigheid of de actualiteit van de uitkomsten niet aantast, belangrijk.

De rol van de computer in enquête-onderzoek

Van oudsher heeft de computer een belangrijke rol gespeeld bij de verwerking van enquête-onderzoek. Er kan zelfs worden gesteld dat grote problemen bij de handmatige verwerking een impuls gaven tot de bouw van de eerste computers. Het was Herman Hollerith die in 1886 de eerste ponskaartenmachine ontwierp om de Amerikaanse volkstelling sneller en beter te kunnen verwerken. Dankzij de inzet van zo'n honderd van deze machines bij de volkstelling van 1890, konden de resultaten al 6 weken na de laatste teldag worden gepubliceerd. De door Hollerith gemaakte machines konden alleen maar tellen en sorteren, en niet echt rekenen. Dat werd pas mogelijk in de veertiger jaren toen de eerste 'echte' computers uitkwamen. De eerste computer voor algemeen gebruik was de UNIVAC in 1951. Hij werd dan ook onmiddellijk aangeschaft door het Amerikaanse Censusbureau dat de volkstelling van 1950 ermee verwerkte.

De ontwikkeling ging voort. Nadat eerst de foutgevoelige elektronenbuizen werden vervangen door de veel betrouwbaardere transistors, vond vervolgens een proces van miniaturisering plaats, dat uiteindelijk leidde tot de chip. Ook al werden de computers beter, en waren ze geschikt voor algemeen gebruik, het werken ermee bleef aanvankelijk in handen van

specialisten. De statisticus was afhankelijk van programmeurs en operateurs voor het specificeren en verwerken van zijn statistische problemen. Pas in de zeventiger jaren kreeg de statisticus zelf toegang tot de computer. Het was het tijdperk van de grote mainframe computers, waarin tientallen, zo niet honderden, terminals gekoppeld konden worden. Door de beschikbaarheid van terminals, en door het beschikbaar komen van algemene statistische pakketten als SAS, SPSS, BMDP, P-STAT e.d. kon de statisticus nu zelf aan het werk.

In de tachtiger jaren zien we de opkomst en explosieve groei van de microcomputer. Door de relatief lage prijs, de grote gebruiksvriendelijkheid, en mede als gevolg van de vele PC-privé-projecten, komen de computers in bijna elke huiskamer terecht. De steeds groter wordende markt voor microcomputers leidt ook tot een omvangrijk aanbod van programmatuur. Dit geldt ook voor statistische programmatuur. Er zijn dikke boeken nodig om de honderden programma's op dit gebied te inventariseren. Arme statisticus die het programma moet trachten te vinden dat voor hem het meest geschikt is!

De computer is vandaag de dag op veel plaatsen in het enquête-onderzoek doorgedrongen. Aanvankelijk werd de computer alleen gebruikt voor sorteren en tabelleren. Dit had met name tot gevolg dat men de gegevens sneller kon verwerken. Later kwam daar het controleren van de gegevens bij. De computer kon meer controles, en bovendien ingewikkelder controles, uitvoeren dan de mens met hand en hoofd. Daardoor kon de kwaliteit van de gegevens worden vergroot. Als in de zestiger jaren de statisticus zelf achter de computer kan plaatsnemen, wordt het mogelijk complexe multivariate analyses op de gegevens uit te voeren, en aldus het inzicht in de uitkomsten verder te vergroten.

Computergestuurd enquêteren

Al de genoemde computertoepassingen hebben betrekking op wat er gebeurt met gegevens nadat ze zijn verzameld. In het laatste decennium wordt de computer echter ook steeds vaker ingezet bij het verzamelen van gegevens. Bij het computergestuurd enquêteren wordt de papieren vragenlijst vervangen door een computerprogramma. Dit programma bevat informatie over de volgorde waarin en de voorwaarden waaronder de vragen moeten worden gesteld. De computer stuurt het vraaggesprek. Steeds zet hij de vraag die aan de beurt is op het scherm, waarna het antwoord op de vraag wordt ingetypt. Het ingevoerde antwoord wordt meteen door de computer gecontroleerd, en is iets niet in orde dan verschijnt een melding op het scherm.

Computergestuurd enquêteren integreert drie stappen uit het onderzoekstraject: verzamelen, invoeren en bewerken van gegevens. Dit heeft drie belangrijke voordelen. In de eerste plaats wordt tijdens het interview de enquêteur ontlast (waarbij met de term enquêteur zowel mannen als vrouwen worden aangeduid). Deze hoeft niet steeds de volgende vraag op te zoeken. Dit lijkt misschien een simpele activiteit, maar soms hangt het stellen van een vraag af van de antwoorden op al eerder gestelde vragen, en dan wordt dat een heel gezoek. In de tweede plaats voert de computer tijdens het gesprek een aantal controles uit. Fouten die tijdens het gesprek worden gevonden, kunnen ook tijdens het gesprek worden verbeterd. Dit in tegenstelling tot fouten die achteraf worden gevonden, en die meestal niet of heel moeilijk te corrigeren zijn. Het derde voordeel van het gebruik van de computer is dat na afloop van het gesprek de gegevens meteen in de computer zitten, en naderhand dus niet meer hoeven te worden ingevoerd.

Computergestuurd enquêteren leidt zo tot een vorm van kwaliteitscontrole die geheel in overeenstemming is met de theorie van Deming (1986) over de verbetering van kwaliteit en produktiviteit in industriële produktieprocessen. Deming geeft regels voor de verbetering van het produktieproces. Die regels kunnen ook heel goed worden toegepast in het statistische produktieproces. Eén van die regels stelt dat inspectie achteraf van de produktie geen goede zaak is. Inspectie achteraf met het doel te komen tot kwaliteitsverbetering, vindt plaats in een te laat stadium, is te kostbaar en te weinig effectief. Kwaliteit dient al te worden ingebouwd in de ontwerpfase van het proces, wat vertaald in statistische termen erop neer komt dat controle van de statistische gegevens achteraf geen goede zaak is. Controle moet plaats vinden tijdens het proces, en dat is in dit geval tijdens het enquêteren.

Computergestuurd enquêteren kent twee belangrijke varianten, die algemeen worden aangeduid als CATI en CAPI. CATI staat voor Computer Assisted Telephone Interviewing, oftewel computer-gestuurd telefonisch enquêteren. De enquêteur heeft hierbij de beschikking over een telefoon en een computer. Nadat telefonisch contact is gelegd met een respondent, start de enquêteur het interview, leest de vragen op door de telefoon en typt de verkregen antwoorden in. CAPI staat voor Computer Assisted Personal Interviewing. Het is de geautomatiseerde equivalent van mondeling enquêteren. Net als bij CATI, zit ook hier weer de vragenlijst in de computer, maar het betreft hier een schootcomputer. De enquêteur neemt de schootcomputer mee naar de te bezoeken personen, start daar het enquêteerprogramma, leest de vragen op en typt de antwoorden in. De aldus met CAPI verzamelde gegevens kunnen per telefoon (in combinatie met een modem) worden doorgegeven aan het onderzoeksbureau, of er kunnen diskettes worden opgestuurd.

Nu is alleen het schriftelijk enquêteren nog overgebleven om te automatiseren. De gebruikte terminologie doortrekkend zou deze vorm van gegevensverzameling moeten worden aangeduid met CAMI (Computer Assisted Mail Interviewing). Bij CAMI wordt de vragenlijst op diskette naar de respondenten toegestuurd. De respondent start thuis het enquêteerprogramma en typt de antwoorden in op de door het programma gestelde vragen. Daarna wordt de diskette met de verzamelde gegevens weer teruggestuurd. De haalbaarheid van CAMI valt of staat met de beschikbaarheid van (geschikte) computers bij de respondenten, en daarom zal deze wijze van enquêteren voorlopig nog niet op grote schaal worden toegepast.

Samenvattend kan worden gesteld dat met computergestuurd enquêteren het onderzoek sneller kan worden uitgevoerd, en tot betere uitkomsten leidt. In feite is hiermee enquête-onderzoek met papieren vragenlijsten overbodig geworden. De grote voordelen van computer-gestuurd enquêteren hebben ertoe geleid dat een groot instituut als het Centraal Bureau voor de Statistiek (CBS) nu vrijwel al zijn enquêtes onder huishoudens en personen uitvoert met CAPI of CATI. Het CBS heeft hiervoor zelf een computersysteem ontwikkeld dat bekend staat onder de naam Blaise. CBS (1987) bevat meer details over de overgang van de traditonele enquête-verwerking naar de nieuwe technologie. Het Blaise Systeem staat beschreven in Bethlehem et al. (1989).

Correctie voor non-respons

De gegevens die na verzameling worden verkregen, zijn in de meeste gevallen nog niet bruikbaar voor analyse. De belangrijkste oorzaak is non-respons. Als bepaalde selecte groepen minder goed in het onderzoek vertegenwoordigd zijn dan andere groepen, leidt dit tot een vertekening in de resultaten. Het probleem van de non-respons wordt steeds groter, en dus ook de invloed van dit verschijnsel op de kwaliteit van de uitkomsten. Zie ook Bethlehem en Kersten (1986). Voor de effecten van non-respons kan worden gecorrigeerd, en dit gebeurt meestal met een wegingsprocedure. Het komt erop neer dat aan ieder record een gewicht wordt toegekend dat zodanig is bepaald, dat de gewogen verdeling in de steekproef, met betrekking tot bekende kenmerken als leeftijd, geslacht, burgerlijke staat en regio, overeenkomt met de verdeling in de populatie. Wegen kan, indien goed uitgevoerd, niet alleen een vertekening in de uitkomsten wegnemen, maar ook de precisie van de uitkomsten verhogen.

Post-stratificatie is een veel toegepaste wegingstechniek. Hierbij wordt de populatie verdeeld in strata, en per stratum kan dan een gewicht worden berekend. Een probleem bij post-stratificatie is echter altijd weer dat de populatieverdeling van allerlei relevante kenmerken lang niet altijd aanwezig is. Onderzoek heeft geleid tot betere methoden van wegen, zie bijvoorbeeld Bethlehem en Keller (1987). Toepassing van technieken als lineair wegen en multiplicatief wegen vereisen het oplossen van grote stelsels vergelijkingen of omvangrijke iteratieve procedures. Alleen door het beschikbaar komen van de capaciteit van de huidige generatie computers zijn deze technieken mogelijk geworden. Het programma Bascula is een voorbeeld van een computerprogramma waarin deze wegingstechnieken zijn geïmplementeerd, zie Göttgens et al. (1990).

Analyse van de gegevens

Het analyseren van de verzamelde gegevens vormt een zeer belangrijk onderdeel van een statistisch onderzoek. Veel statistiekboeken gaan zelfs alleen maar over analyse. Daarbij beperken ze zich soms ook nog tot de klassieke toetsingstheorie. Deze boeken gaan niet alleen voorbij aan het hele traject dat heeft geleid tot het beschikbaar komen van de gegevens, maar ook onderkennen ze niet dat het in enquête-onderzoek meestal niet gaat om het toetsen van vooraf opgestelde hypothesen. De statisticus is veel meer geïnteresseerd in een verkennende analyse. Hij heeft behoefte aan technieken die hem helpen te zoeken naar strukturen in de gegevens. Voor een deel als gevolg van de opkomst van de computer, is er de laatste jaren een grote reeks nieuwe exploratieve analysetechnieken beschikbaar gekomen.

Voor exploratieve analyse lenen grafische technieken zich vaak beter dan numerieke technieken. Een voorbeeld is het gemiddelde. Het mag dan een goede maat zijn voor de locatie van een verdeling, maar een geschikte grafiek levert in één klap informatie over locatie, spreiding, scheefheid en uitschieters. Het is vooral Tukey (1977) die deze vorm van analyse populair heeft gemaakt. Hij heeft diverse nieuwe grafische technieken (o.a. stem-and-leaf display, box-and-whisker plot) voorgesteld die zich uitstekend laten implementeren in computerprogramma's. Doordat het zo eenvoudig is geworden om met de computer snel allerlei numerieke grootheden en grafieken te produceren, is in feite een geheel nieuwe vorm van gegevensanalyse ontstaan, waarbij de statisticus in een dialoog is gewikkeld met de computer.

De exploratieve statistiek is niet stil blijven staan bij de door Tukey voorgestelde technieken. Met het krachtiger en vriendelijker worden van de computer hebben nieuwe technieken hun intrede gedaan, die slechts kunnen bestaan bij de gratie van de computer. Een eerste voorbeeld hiervan is de multivariate exploratieve analyse van gegevens. In pogingen om de meer-dimensionale aspecten van waarnemingen op scherm of papier zichtbaar te maken, zijn talrijke nieuwe technieken ontwikkeld. Voorbeelden zijn de matrix van spreidingsdiagrammen, de gezichten van Chernoff, de bomen en kastelen van Kleiner en Hartigan, en verder nog sterren en glyphen en windvanen. Zie voor een overzicht bijvoorbeeld Kleiner en Hartigan (1981). Deze techieken gaan uit van de gedachte dat de combinatie van ogen en hersenen een zeer geavanceerd instrument is voor de verwerking van visuele informatie. Het kan zeer goed de essentie van een grote hoeveelheid informatie samenvatten, maar is tevens in staat zich op details te concentreren. Helaas is ons visuele informatiesysteem echter ook in staat ons voor de gek te houden. De talrijke voorbeelden van gezichtsbedrog uit de psychologieboeken zijn daarvan het bewijs. Daarom zal er nog veel onderzoek nodig zijn om te kunnen vaststellen of deze technieken effectief zijn.

Een probleem van de genoemde multivariate exploratieve technieken is dat ze uiteindelijk toch op een of andere wijze meer-dimensionale informatie afbeelden op het twee-dimensionale vlak van papier of computerscherm. Dit brengt beperkingen met zich mee bij het interpreteren van de informatie, maar ook daar zijn er nieuwe mogelijkheden. Een nieuwe verzameling tech-nieken, onder de naam 'dynamic graphics', maakt het mogelijk nog meer te profiteren van de mogelijkheden die de computer biedt. Essentie van deze technieken is het gebruik van de elementen beweging en kleur. Toepassing ervan biedt de onderzoeker de mogelijkheid om met toetsenbord of muis een grafiek op het computerscherm direct te manipuleren. Een voorbeeld is het toevoegen van een derde dimensie aan een spreidingsdiagram door de puntenwolk op het scherm te laten roteren. Een ander voorbeeld is de mogelijkheid om één of enkele punten een andere kleur te geven, zodat ze tijdens het draaien goed kunnen worden gevolgd. Deze techniek is ook heel effectief als verschillende aanzichten van een meer-dimensionale puntenwolk tegelijk in vensters op het scherm staan en de consequenties van manipulaties in het ene venster direct zichtbaar worden in alle andere vensters.

Het zou wel eens zo kunnen zijn dat een dergelijke intensieve wisselwerking tussen gegevens en statisticus meer oplevert dan het wat meer passieve staren naar statische plaatjes, maar ook hier is verder onderzoek noodzakelijk. Voor een overzicht van de technieken die dynamic graphics biedt, kan worden verwezen naar Becker et al. (1987).

De steeds groter wordende capaciteit van de computer maakt het ook mogelijk zeer reken-intensieve technieken voor exploratieve analyse te hanteren. Een voorbeeld is de door Cleveland (1979) ontwikkelde techniek voor robuste lokaal gewogen regressie-analyse. Hierbij gaat het om het bepalen van een lijn door een puntenwolk die zo goed mogelijk de aanwezige trend in beeld brengt. Voor elk punt in de puntenwolk moeten daarbij drie gewogen regressie-analyses worden uitgevoerd, waarin alle punten uit de omgeving van dat punt zijn betrokken.

Enquête-onderzoek is er meestal op gericht om op grond van een steekproef uitspraken te doen over karakteristieken van de populatie als geheel. Deze uitspraken hebben in wezen de vorm van schatters van populatieparameters. Het betreft vaak eenvoudige grootheden zoals gemiddelden en percentages. Als voldoende informatie aanwezig is over de wijze waarop de

steekproef tot stand is gekomen, met andere woorden als de verdeling van de schatter bekend is, dan kan de nauwkeurigheid van de schatting worden gekwantificeerd in de vorm van standaardfouten of betrouwbaarheidsintervallen. Het gebruik van complexe steekproefdesigns, het optreden van non-respons, en het corrigeren hiervoor leveren echter problemen op bij het op analytische wijze afleiden van de verdeling van zelfs eenvoudige schatters. Gebruik van de computer kan uitkomst bieden in de vorm bootstrap-technieken. Hierbij worden uit de steekproef een groot aantal nieuwe steekproeven getrokken. Door voor elk van deze steekproeven de betreffende schatting opnieuw te berekenen, wordt informatie verkregen over de verdeling van de oorspronkelijke schatter, en daarom kan deze techniek worden gebruikt voor het bepalen van standaardfouten. Bootstrap-technieken zijn zeer reken-intensieve technieken. Het zal duidelijk zijn dat bootstrappen dus alleen maar kan dankzij de computer en bij uitstek geschikt is voor toepassing met een computer.

Veel minder vaak wordt enquête-onderzoek uitgevoerd voor het toetsen van hypothesen. Zeker als het om wat grotere onderzoeken gaat, is dit ook niet zo zinvol. Als de steekproefomvang maar groot genoeg is, wordt de geringste afwijking van de nulhypothese significant, en daar is de onderzoeker eigenlijk niet geïnteresseerd. Hij wil niet weten of er samenhang bestaat tussen twee verschijnselen, maar hoe sterk die samenhang is. Daar waar hij toch toetst, dient de onderzoeker te beseffen dat in de klassieke toetsingstheorie intensief gebruik wordt gemaakt van modelveronderstellingen waaraan lang niet altijd is voldaan. Diverse computer-intensieve methoden bieden een model-onafhankelijk alternatief. Zo kunnen randomizatie-technieken worden toegepast om te toetsen op onafhankelijkheid tussen twee reeksen waarnemingen. Monte Carlo technieken kunnen worden gebruikt om te toetsen of een reeks waarnemingen afkomstig is uit een nader omschreven verdeling. Tenslotte kunnen de al eerder genoemde bootstrap-technieken worden toegepast bij het toetsen van uitspraken over populatieparameters als het gemiddelde.

Publikatie van de resultaten

Aan het einde van een statistisch onderzoek staat nog een laatste toepassing van de computer. Rapporten en wetenschappelijke publikaties kunnen worden samengesteld met behulp van een tekstverwerker. Doordat een goede onderzoeker alle noodzakelijke programmatuur op dezelfde computer beschikbaar heeft, wordt het erg eenvoudig om tabellen, grafieken en andere resultaten op te nemen in het rapport zonder ze opnieuw te hoeven intypen. Als dan ook nog een goede laserprinter beschikbaar is, kan het rapport ook nog zo worden afgedrukt dat het resultaat onmiddellijk kan worden vermenigvuldigd, dus zonder tussenkomst van zetter en drukker.

Een van de elementen in de publikatie is de tabel. Veel onderzoeksresultaten komen in eerste instantie beschikbaar in de vorm van tabellen. Uiteraard is het mogelijk om tabellen met de hand te maken, maar voor grote bestanden is dit zeer tijdrovend, zo niet onbegonnen werk. Ook hier biedt de computer uitkomst. In veel statistische pakketten zijn tabelleermodulen opgenomen. Veel van deze modulen maken echter alleen analysetabellen. Hierbij gaat het niet zozeer om de vormgeving van de tabel alswel om de inhoud, en de toetsingsgrootheden die onder de tabel worden afgedrukt. In een publikatie zal men echter met name publikatietabellen willen opnemen, en dan ligt de nadruk op nette opmaak en leesbaarheid. Goede, dat wil

zeggen eenvoudige, vriendelijke en krachtige, programmatuur voor het maken van publikatietabellen kan de onderzoeker veel werk uit handen nemen in de eindfase van het onderzoek. Uiteraard gaat het hierbij niet om de uitvoering van geavanceerde statistische technieken, maar wel om het versnellen en vereenvoudigen van in het onderzoek toch noodzakelijke activiteiten.

Een ander belangrijk bestanddeel van de publikatie is de grafiek. Grafieken in publikaties zien er meestal anders uit dan grafieken die worden gebruikt voor exploratieve analyse. Het accent ligt bij presentatie op het overbrengen van informatie. Dit heeft consequenties voor de vormgeving. Het is op voorhand niet duidelijk wie de grafiek zal gebruiken en daarom moet de boodschap zo helder en eenvoudig mogelijk worden aangeboden. De gebruiker van de grafiek mag daarbij niet worden afgeleid door allerlei storende effecten zoals overbodige tierlantijnen, schokkende kleurcombinaties of arceringen die pijn doen aan de ogen. Tufte (1983) geeft een aardig overzicht van wat wel en niet mag in grafieken.

Veel technieken voor de presentatie van statistische gegevens zijn beschikbaar in computerprogramma's. Het zijn met name de 'presentation graphics' in niet-statistische pakketten die uitnodigen tot een minder verantwoord gebruik van deze grafische technieken. Ze bieden de argeloze gebruiker de mogelijkheid om elementaire grafieken op te tutten met fantastische kleuren, drie-dimensionale weergave, uitgesneden punten e.d., waardoor de nadruk meer komt te liggen op de verpakking dan op de inhoud. Nederland is nog niet erg statistiek-minded. Dat is goed te zien aan de wijze waarop de media met grafieken omgaan. Ze lijken vaker door vormgevers en artiesten te zijn gemaakt, dan door statistici, waardoor de waarde van de grafiek als instrument voor het overbrengen van een statistische boodschap in de verdrukking komt. De statisticus zal ontwerpers en gebruikers van grafieken beter moeten voorlichten over wat goed en slecht is. Hij heeft daarbij gelukkig wat meer statistisch georiënteerde pakketten tot zijn beschikking die uitnodigen tot een wat soberder en verantwoorder gebruik. Met de juiste keuze van een techniek in combinatie met goede afdrukapparatuur (plotter, laserprinter) kan snel een mooi en doeltreffend resultaat worden bereiken.

Elektronische publikaties

Met name de brede verspreiding van de microcomputer maakt het nu mogelijk de resultaten van onderzoek ook nog op een andere wijze dan op papier te verspreiden. Doordat geïnteresseerden in de uitkomsten van onderzoek vaak zelf over goede apparatuur en programmatuur beschikken, is er een groeiende vraag om de oorspronkelijke onderzoeksgegevens zelf. Voor kleine onderzoeken kan dat bijvoorbeeld heel goed op diskette, en voor grote onderzoeken zal men wellicht magneetbanden gebruiken.

Bij de publikatie van bestanden dient men zich niet te beperken tot het op band of diskette zetten van alleen de gegevens. Minstens even belangrijk is de dokumentatie van het bestand. Daarin zal goed en duidelijk moeten worden beschreven wat de variabelen voorstellen, welke waarden ze kunnen aannemen, en wat deze waarden betekenen. Ook moet de nodige informatie worden gegeven over het onderzoek in het algemeen, zoals steekproefdesign, en tijdstip en wijze van uitvoering.

Het beschikbaar stellen van de gegevensbestanden aan derden lijkt een aantrekkelijk perspectief. Immers, zo kan een doelmatiger gebruik worden gemaakt van gegevens die in een vast wel kostbaar onderzoek zijn verzameld. Er zijn echter ook nadelen. De onderzoeker die bestanden met gegevens beschikbaar stelt, dient zich daar wel degelijk van bewust te zijn.

Het eerste probleem grijpt terug naar het 'Dirty Data Theorem'. Het is niet zo dat de gegevens kunnen worden opgevat als een nette aselecte steekproef met gelijke kansen uit een populatie. Door het optreden van non-respons hebben de elementen in de populatie ongelijke en onbekende kansen om in de steekproef te komen. Daar kan met een wegingstechniek voor worden gecorrigeerd, maar dat betekent ook dat in elke analyse deze gewichten moeten worden meegenomen. Helaas zal de gewogen verdeling van steekproefgrootheden in veel gevallen niet voldoen aan de voorwaarden die de statististische techniek eraan stelt. Ook een ongewogen analyse biedt geen oplossing, omdat dan op voorhand al zeker is dat de uitkomsten niet kunnen worden gegeneraliseerd naar de populatie waaruit de steekproef afkomstig is. Bij het beschikbaar stellen van onderzoeksgegevens aan anderen zal duidelijk moeten worden gemaakt in hoeverre dit soort zaken een rol spelen.

Een tweede probleem heeft ook te maken met het ontbreken van gegevens. Als in een vragenlijst de antwoorden op enkele vragen ontbreken (bijvoorbeeld doordat de respondent het antwoord weigert of niet weet), dan wordt soms een techniek toegepast waarbij een schatting voor het antwoord op de open plek wordt ingevuld. Dit wordt imputatie genoemd. Het aardige van imputatie is dat het een volledig bestand zonder 'gaten' oplevert, maar daar staat tegenover dat imputatie de eigenschappen van de verdeling van de gegevens kan aantasten. Een eenvoudige voorbeeld van imputatie is het invullen van het gemiddelde van de wel bekende gegevens. Dit heeft tot gevolg dat de op grond van de gegevens geschatte variantie kleiner is dan de werkelijke variantie. Dit wekt een schijn van nauwkeurigheid die in werkelijkheid niet aanwezig is. Bij het beschikbaar stellen van een gegevensbestand aan andere onderzoekers dient erop te worden gewezen of, hoe en waar imputatie is uitgevoerd. In feite zou voor elke waarneming bekend moeten zijn of het een 'echte' is of een 'synthetische'.

Een derde probleem is dat van de privacybescherming. Vaak wordt bij het enquêteren aan de respondenten beloofd dat hun gegevens vertrouwelijk zullen worden behandeld. Dat is terecht, want bij het bedrijven van statistiek gaat het niet om het analyseren van individuele gegevens, maar het onderzoek van de verdeling van geaggregeerde grootheden. Wanneer een onderzoeker een bestand met individuele gegevens publiceert, zal hij er voor moeten zorgen dat de anonimiteit gewaarborgd blijft. Het bestand zal daarom geen identificerende gegevens zoals naam en adres, of SOFI-nummer mogen bevatten. Maar is dat voldoende? Onderzoek heeft aangetoond dat het combineren van veel minder identificerende gegevens, zoals gezinssamenstelling, woonplaats en beroep ook vrij eenvoudig tot identificatie kan leiden. Eenvoudige statistische of database programmatuur is al voldoende om dergelijke vis-acties uit te kunnen voeren. De verstrekker van het bestand dient dus maatregelen te nemen om identificatie en onthulling te voorkomen. De eenvoudigste manier om dat te doen is er voor te zorgen dat er niet teveel informatie in het bestand komt, bijvoorbeeld door het aantal identificerende variabelen te beperken en van de wel opgenomen variabelen het meetniveau laag te houden. Dit soort beveiligingsacties vermindert de informatiewaarde van het bestand, maar dat is de prijs die moet worden betaald voor een veilig bestand. Meer details over de onthullingsproblematiek is te vinden in Bethlehem et al. (1990).

Toekomstperspectief

De computer speelt een steeds grotere rol in het leven van de statisticus. Eigenlijk kan hij niet meer buiten de computer. De vraag die nu rijst, is: kan de computer buiten de statisticus? Met andere woorden, kan de computer het werk van de statisticus volledig overnemen? We komen dan op het terrein van expertsystemen.

Een expertsysteem bestaat uit twee componenten: een kennisbank en een redeneersysteem. In de kennisbank zit alle benodigde kennis over de specifieke problematiek opgeslagen. Het redeneersysteem bevat de procedures waarmee de informatie in de kennisbank kan worden gebruikt om concrete problemen op te lossen. Sinds 1980 zijn er al optimistische toekomstverwachtingen over expertsystemen die de statisticus veel werk uit handen zouden kunnen nemen, maar echte grote expertsystemen zijn er nog steeds niet. Wellicht waren de verwachtingen wat te hoog gespannen. Men is nu wat minder ambitieus bezig. Dit heeft geleid tot kleinere systemen die slechts een een beperkt stukje van de statistiek bestrijken. Zie voor een overzicht Hand (1992).

Een ontwikkeling van de laatste tijd is het gebruik van neurale netwerken in de statistiek. Neurale netwerken kunnen worden gezien als een eerste voorzichtige poging om op zeer vereenvoudigde wijze de werking van de hersenen na te bootsen in een computersysteem. Een neuraal netwerk bestaat uit een systeem van verwerkingselementen die onderling met elkaar zijn verbonden. Aan deze verbindingen zijn gewichten verbonden. In plaats van een netwerk te programmeren, wordt het getraind door er informatie in te stoppen. Op grond van deze informatie en nader te specifieren criteria past het netwerk deze gewichten aan. Na het netwerk voldoende te hebben getraind zal het netwerk accurate informatie kunnen bevatten over de struktuur van de ingevoerde informatie. Netwerken lijken effectief te kunnen worden toegepast op het terrein van de classificatie van waarnemingen. Denk hierbij bijvoorbeeld aan clusteranalyse en discriminantanalyse (zie Murtagh, 1990). Teleurstellend is nog wel dat niet duidelijk wordt waarom deze technieken werken. Ook hier ligt dus nog een uitdaging voor de statisticus.

De computer in het statistiekonderwijs

De rol van de computer in de statistiek is groot, en daarom moet de rol van de computer in het statistiekonderwijs ook groot zijn. Dit kan op verschillende manieren.

In de eerste plaats kan de computer worden ingezet als hulpmiddel in het statistiekonderwijs. Daarbij moet niet in de eerste plaats worden gedacht aan geprogrammeerde statistiekinstruktie. Een veel nuttiger rol is weggelegd bij het zichtbaar maken van allerlei statistische processen. De dynamiek die de computer hierin kan brengen, bevordert inzicht en begrip.

In de tweede plaats moet de computer worden ingezet bij de training in het gebruik van statistische pakketten. Aangezien er honderden statistische pakketten in omloop zijn, is het onmogelijk ze allemaal te leren gebruiken. Dat hoeft ook niet. Belangrijker is dat een statisticus leert hoe hij ze moet gebruiken, en waarop hij moet letten bij het gebruik. De meeste pakketten doen domweg wat ze worden opgedragen, ongeacht of dat zinvol is of niet.

Ze zullen er geen probleem van maken om een correlatiecoëfficiënt tussen twee nominale variabelen uit te rekenen, en ook zullen ze niet controleren of is voldaan aan de modelveronderstellingen. Doet de statisticus dat ook niet dan zal het resultaat in feite neerkomen op 'garbage in, garbage out'. Deze term dateert al uit de zeventiger jaren, en heeft helaas nog niets aan waarde ingeboet.

In de derde plaats kan de computer worden ingezet als instrument bij onderzoek naar nieuwe methoden en technieken in de statistiek. De praktische waarde van de voorgestelde vernieuwingen zal zich moeten bewijzen in concrete toepassingen en met name ook in simulatie-experimenten. Dus ook de wat meer theoretisch georiënteerde statisticus ontkomt niet aan de computer.

Een goede statisticus is een all-rounder, die niet alleen thuis is in de theoretische en praktische aspecten van de statistiek, maar ook ervaring heeft op een breed spectrum van computer-toepassingen. Hierbij moet niet alleen worden gedacht aan statistische programmatuur, maar ook aan spreadsheetprogramma's, databasepakketten, presentatieprogramma's, en misschien ook nog wel desktop publishing. Kortom, een statisticus zonder computer is geen statisticus.

Tenslotte

Aan het eind van mijn voordracht gekomen, wil een aantal personen bedanken. In de eerste plaats zijn dat het bestuur van de Universiteit van Amsterdam, en alle betrokkenen van de Faculteit der Wiskunde en Informatica en van de Faculteit der Economie en Econometrische Wetenschappen die zich hebben ingezet voor mijn benoeming.

Een deeltijdhoogleraarschap is niet mogelijk zonder de medewerking van mijn andere werkgever, het Centraal Bureau voor de Statistiek. Ik wil de directie dank zeggen voor de medewerking die zij verleende aan mijn benoeming. Het is mijn overtuiging dat de combinatie van werkzaamheden bij de universiteit en het CBS voor beiden zinvol zal zijn. Enerzijds kunnen studenten kennis nemen van de praktische aspecten van de statistiek van het CBS, waardoor ze beter worden toegerust voor hun toekomstige werk. Anderzijds biedt de universiteit mogelijkheden voor onderzoek dat worden gevoed door problemen uit de praktijk.

Een bijzonder woord van dank gaat uit naar Wouter Keller. Op veel terreinen hebben wij samengewerkt en werken wij nog samen. Dank zij veel van zijn stimulerende ideeën hebben we met zijn allen op het CBS mooie dingen gerealiseerd, mede waardoor het nu een gerenommeerd instituut in de wereld is.

Ik wil de hooggeleerde Hemelrijk danken voor het wekken van mijn belangstelling voor de statistiek. Zijn boeiende colleges hebben mij op het goede spoor gezet, en ik denk nog regelmatig terug aan de vele wijze lessen die daarin opgesloten zaten.

Dit alles zou uiteraard niet mogelijk zijn geweest zonder de steun en aanmoedigingen van mijn ouders. Daarvoor ben ik hen dankbaar.

Beste Ingrid, Maarten en Hayo. Het hebben van twee banen in verschillende steden geeft een hoop onrust in het gezin. Jullie vangen dat prima op en dat waardeer ik zeer.

Tot slot wil ik de medewerkers van de afdeling Statistische Informatica van het CBS sterkte toewensen op de dagen dat ik afwezig ben. Het zijn allemaal goede vrienden en prima collega's die elkaar helpen daar waar dat nodig is. Ik ben er van overtuigd dat het werk daar gewoon doorgaat alsof er niets aan de hand is. Daarvoor spreek ik mijn waardering uit.

Ik dank u voor uw aandacht.

Literatuur

Becker, R.A., Cleveland, W.S. and Wilks, A.R. (1987), Dynamic Graphics for Data Analysis, *Statistical Science* 2, blz. 355-395.

Bethlehem, J.G, Hundepool, A.J., Schuerhoff, M.H. en Vermeulen, L.F.M. (1989), Blaise 2.0 / Een eerste kennismaking, *CBS-rapport*, Centraal Bureau voor de Statistiek, Voorburg.

Bethlehem, J.G. en Keller, W.J. (1987), Linear Weighting of Sample Survey Data, *Journal of Official Statistics* 3, pp. 141-154.

Bethlehem, J.G., Keller, W.J. en Pannekoek, J. (1990), Disclosure Control in Microdata, *Journal of the American Statistical Association* 85, blz. 38-45.

Bethlehem, J.G. en Kersten, H.M.P. (1986), *Werken met non-respons*, Dissertatie Universiteit van Amsterdam. Ook verschenen als: *Statistische Onderzoekingen* M 30, Staatsuitgeverij, 's-Gravenhage.

CBS (1987), *CBS Select 4, Automation in Survey Processing*, Centraal Bureau voor de Statistiek, Voorburg.

Cleveland, W.S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association* 74, blz. 829-836.

Deming, W.E. (1986), *Out of the Crisis*, Cambridge University Press, Cambridge, Mass.

Göttgens, R.M.J., Vellen, H.W.J.M., Odekerken, M.E.P. en Hofman, L.P.M.B. (1990), Bascula, versie 1.0, Een weegpakket onder MS-DOS, Gebruikshandleiding, *CBS-rapport*, Centraal Bureau voor de Statistiek, Voorburg.

Hand, D.J. (1992): AI in Statistics, *Proceedings of the Conference on New Techniques and Technologies in Statistics*, Bonn, To be published.

Kleiner, B. en Hartigan J.A. (1981), Representing Points in Many Dimensions by Trees and Castles, *Journal of the American Statistical Association* 76, blz. 260-276.

Murtagh, F. (ed) (1990), *Proceedings of the Workshop on Neural Networks for Statistical and Economic Data*, Munotec Systems, Dublin.

Tufte, E. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Chesire, Conn.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.