

Statistiek is samenwerken
buiten
de grenzen van het weten

A.G.M. Steerneman¹

Samenvatting

In deze oratie wordt de rol van de statisticus besproken bij het verkrijgen van kennis. Informatie kan beschikbaar zijn in de vorm van boeken, tijdschriften en ervaringsdeskundigheid, maar ook in de vorm van onvolledige en onnauwkeurige gegevens. Statistiek bewijst grote diensten als het erom gaat kennis te halen uit deze laatste vorm van informatie. Dit wordt nader uitgewerkt vanuit een historisch perspectief en aandacht voor de samenwerking tussen klant en statisticus.

¹Vakgroep Econometrie, Faculteit der Economische Wetenschappen, Rijksuniversiteit Groningen, Postbus 800, 9700 AV Groningen. Tel. 050-633807. e-mail: a.g.m.steerneman@eco.rug.nl.

Statistiek is samenwerken buiten de grenzen van het weten

De slogan "werken aan de grenzen van het weten", waarmee de Rijksuniversiteit Groningen zich profileert onderstreept de wetenschappelijke aard van de instelling en het grensverleggende karakter van de onderzoeksactiviteiten. Het woordje "werken" staat voor de aanzienlijke inspanningen die de universiteit daarvoor moet verrichten. Wetenschap bedrijven bestaat immers voor een relatief klein deel uit inspiratie en voor het grootste deel uit transpiratie. Een vergelijking met topsport, waar deze universiteit speciale aandacht voor heeft, ligt voor de hand. Ik zou niet willen beweren dat het verkrijgen van nieuwe kennis een kwestie is van bloed, zweet en tranen. Het is vooral een proces van kennis, intuïtie, toeval, geluk, geld en een systematische manier van werken. Bij topsport heb je te maken met de vorm van de dag, maar veel belangrijker is het systematisch werken aan het verkrijgen van een zo goed mogelijke vorm door middel van bijvoorbeeld uitgekende trainingsschema's en professionele begeleiding. Voor zowel onderzoekers als sporters geeft het bereiken van resultaten een "goed gevoel" en dit betekent weer een nieuwe stimulans. Gedrevenheid, systematisch werken en grensverleggend bezig zijn, zijn punten van overeenkomst tussen onderzoek en topsport.

Statistiek speelt een rol in de systematiek van het doen van empirisch onderzoek. Bij empirisch onderzoek moet er een goede samenwerking zijn tussen de onderzoekers en de statisticus. De rollen van statisticus en onderzoeker zullen overigens vaak verenigd zijn in dezelfde persoon. Statistische technieken worden veelvuldig toegepast om "iets aan de weet te komen" en bewijzen hun nut juist wanneer men de dingen niet meer zeker weet, zoals bij het construeren van modellen, het toetsen van hypothesen, het maken van inschattingen omtrent toekomstige ontwikkelingen en het analyseren van risico's. Ik zou willen zeggen: "Statistiek is samenwerken buiten de grenzen van het weten".

De klant en de statisticus

Wanneer raadpleegt men een statisticus? In het algemeen luidt het antwoord van de statisticus dat het in de praktijk vaak pas gebeurt als het eigenlijk veel te laat is. Het antwoord van degene die statistisch advies wenst, is dat dat pas gebeurt als hij er zelf niet goed meer uitkomt. In dit kader is het wellicht verhelderend op te merken dat veel onderzoekers wel degelijk iets aan statistiek gedaan hebben in hun opleiding. Bovendien is men vaak heel goed in staat om met bijvoorbeeld SPSS zelf statistische analyses te doen. Verder is er goede software waarmee men prachtige grafieken kan maken. Waar het natuurlijk om draait, is de manier waarop dit gereedschap gebruikt wordt. Voor het gemak noemen we degene die statistische ondersteuning

wenst de klant. Laat ik nu alvast vertellen dat een statisticus klantvriendelijk moet zijn.

Zoals men in winkelbedrijven al jaren heeft ervaren, heeft iedere klant zo zijn eigenaardigheden. De ene klant is perfect op de hoogte van wat de winkel hem kan bieden, terwijl de ander nog wegwijst moet worden gemaakt in wat er geboden kan worden en wat niet. Als toegepast statisticus heb ik eigenlijk ook een soort winkel. Zo nu en dan ben ik zelf natuurlijk ook klant. Ik werk bijvoorbeeld samen met klanten die zelf op een professionele manier statistiek toepassen in het kader van hun onderzoek. Zulke klanten komen met problemen die zowel een uitdaging zijn voor henzelf als voor de statistische wetenschap. Meestal geldt voor alle klanten dat hun probleemstelling uitdagend is en interessant. Er is echter een duidelijk onderscheid tussen de "professionele klant" en de klant die een relatief beperkte achtergrond in statistiek heeft. Klanten uit deze laatste categorie willen de statisticus nog wel eens beschouwen als een deskundige die snel zal kunnen vertellen of de manier van werken die hij in zijn gedachten heeft redelijk is, of die binnen vijf minuten kan uitleggen hoe een probleem moet worden aangepakt. Als ik benaderd word voor een klein vraagje, dan weet ik uit ervaring dat het vaak gaat om iets dat veel gecompliceerder is. Bij hoge uitzondering zijn we er binnen vijf minuten uit.

Het voorschrijven van een statistisch recept kan alleen maar goed op basis van een redelijke diagnose. Voor het stellen van zo'n diagnose moet je het een en ander weten over het onderzoek. Zo is de probleemstelling van belang, maar ook welke gegevens er zijn en hoe ze zijn verzameld. Er is inleving in de problematiek vereist. Zoiets neemt echt wel enige tijd. Een statistisch consult is maar zelden een routine-kwestie. Het uiteindelijke advies kan echter best uitdraaien op een recept uit het statistische kookboek. Het kan echter ook gebeuren dat de statisticus zelf een methode moet bedenken die nog niet in de literatuur is beschreven of die nog niet standaard in de bekende software zit.

Het leuke van mijn vak is dat je enorm veel opsteekt over andere vakgebieden. Soms is het jammer dat je moet constateren dat de klant bij je komt met gegevens die op een dusdanige wijze zijn verzameld dat je er nauwelijks conclusies aan kunt ontleen ten aanzien van de probleemstelling. Als in het kader van een onderzoek gegevens moeten worden verzameld of als er metingen moeten worden gedaan, dan is het aan te raden van meet af aan samenwerking te zoeken met een statisticus. Deze voorbereidende fase bepaalt de kosten van het totale onderzoek maar ook de kwaliteit van de resultaten. Het zal u niet verbazen dat deze twee aspecten met elkaar samenhangen. Ik heb nu wel iets verteld over de statistische consultatie, maar ik heb nog niet zoveel verteld over wat een statisticus nu eigenlijk doet. Dat hij ook iets met statistieken heeft te maken, ligt voor de hand.

Statistieken en de statisticus

Menigeen van u zal wel eens naar een directe reportage van een sportwedstrijd op televisie hebben gekeken. Gedurende zo'n uitzending wordt u veel informatie verschaft over het wel en wee van de spelers, de tijd die er nog gespeeld moet worden, of de scheidsrechter het wel goed gezien heeft, etc. Sommige verslaggevers willen nog wel eens de beroepsgroep van de statistici aanspreken door gebruik te maken van de zinsnede "Voor de statistici onder u ...", waarna een aantal cijfermatige gegevens volgen, zoals het percentage goede eerste services bij tennis of het aantal hoekschoppen bij voetbal. Dergelijke informatie is inderdaad niet voor alle kijkers even interessant en veel statistici zal het trouwens ook maar matig interesseren. Een dergelijke opmerking versterkt het vertekende beeld dat sommigen van een statisticus hebben, namelijk dat het zijn taak is om gegevens te verzamelen, netjes te ordenen, samen te vatten en te presenteren in de vorm van tabellen, grafieken, indexcijfers, etc. In de praktijk heeft hij hier slechts voor een deel mee te maken. Tegenwoordig is er zoveel mensvriendelijke software dat men alleen daarvoor geen statisticus meer nodig heeft.

Het belang van het systematisch verzamelen van allerlei gegevens is enorm in het huidige informatietijdperk. Meer dan tweeduizend jaren geleden was men al bezig op systematische wijze bepaalde gegevens te verkrijgen. Het kerstverhaal bijvoorbeeld speelt zich af tijdens een volkstelling die keizer Augustus liet houden. Eigenlijk heeft dit nog niet veel met statistiek te maken, want het gaat hier vooral om het inventariseren op een systematische wijze.

Volkstellingen zijn er door de eeuwen heen min of meer geregeld gehouden. De laatste volkstelling in Nederland vond plaats in 1971. In het verleden werden volkstellingen nogal eens gebruikt met het doel vast te stellen hoeveel belasting er viel binnen te halen en hoe groot het leger zou kunnen zijn. In Nederland werden volkstellingen ook gebruikt om de bevolkingsregisters van de gemeenten te controleren. In de kranten hebben we recentelijk kunnen lezen dat bijvoorbeeld in Amsterdam de juistheid ervan nogal tegenviel.

In de Verenigde Staten heeft de volkstelling van 1980 aanleiding gegeven tot een proces van de stad en de staat New York tegen het Bureau of the Census, de instantie die de volkstellingen in de Verenigde Staten uitvoert en met de resultaten komt; zie Freedman en Navidi (1986), inclusief de discussie. Er was nogal wat onvrede over de resultaten van de tellingen. De tienjaarlijkse volkstelling deed niet wat het behoorde te doen, namelijk de bevolking meten op een redelijke wijze. De stad en de staat New York waren ervan overtuigd dat ze disproportioneel onderteld waren, omdat de zwarte bevolking en de "Hispanics" moeilijker te tellen zijn dan de blanke bevolking. De resultaten zouden indirect gevolgen hebben gehad voor de

financiële middelen van New York. Het aardige van dit proces was dat de statistische wijze waarop deze schattingen waren verkregen zelf de inzet van het proces vormden. De staat New York heeft dit proces in eerste instantie gewonnen. Op procedurele gronden werd dit vonnis later weer herroepen. Het Bureau of the Census heeft zich overigens wel het een en ander aangetrokken van de discussie over de te gebruiken statistische methodologie.

Het assisteren bij het opzetten van tellingen en metingen kan een belangrijk deel van het werk van een statisticus zijn. Ik zie een statisticus echter vooral als iemand die deskundig is op het gebied van het halen van nieuwe kennis uit getallen op basis van kennis of ervaring zoals die reeds aanwezig is bij andere deskundigen. Dit laatste is erg belangrijk, want op basis van alleen getallen kun je alles bewijzen. Men heeft bijvoorbeeld wel eens een sterk statistisch verband aangetoond tussen maandcijfers van het aantal geboorten en van het aantal waarnemingen van ooievaars. Voor het maken van een goede statistische analyse van een concreet probleem is de samenwerking met de mensen die om advies vragen erg belangrijk.

De statistiek bewijst haar waarde als kennis en ervaring niet voldoende informatie bieden en men met logisch redeneren ook niet verder komt. Men bevindt zich dan aan de grenzen van het weten. Buiten deze grenzen kan men nog wel het een en ander aan informatie hebben of verkrijgen in de vorm van cijfermateriaal, hetgeen uiterst omvangrijk kan zijn, maar soms ook verbijsterend beperkt. Om in het grijze gebied niet te verdwalen is men op samenwerking aangewezen. Men moet er op bedacht zijn dat de beschikbare kennis onjuist kan blijken te zijn.

Statistische gegevens

In dit informatietijdperk is er aan getallen geen gebrek. Er wordt van alles gemeten aan mensen, bedrijven, chemische processen, sterren, economieën, skeletten, verzekeringspolissen, literatuur, de atmosfeer, etc. En hoe kan het ook anders, veel gegevens worden zo opgeslagen dat ze heel gemakkelijk met computers kunnen worden geanalyseerd. Getallen zijn op zich spijkerhard en zo worden ze ook vaak gebruikt mede door het aanwenden van computers. Voor een statisticus zijn niet alleen getallen erg belangrijk, maar ook het verhaal over de afkomst en de zachtheid van de betreffende gegevens. De beschikbare gegevens zijn vaak steekproefsgewijs verzameld. Het databestand is daardoor onvolledig. De meetapparatuur is soms niet zo nauwkeurig. Individuen hebben de neiging vragen over hun inkomen niet correct te beantwoorden. Soms ontbreken er antwoorden. Het kan ook gebeuren dat een enquêteur per ongeluk een verkeerde categorie aanstreept, etc. Zo wordt je telkens geconfronteerd met de vraag welke informatie je nog kunt halen uit een bestand dat onvolledig is en waarvoor bovendien

geldt dat de informatie onnauwkeurig is.

Statistiek kenmerkt zich door de wijze waarop er wordt omgegaan met databestanden die onvolledig en onnauwkeurig zijn. Wat kun je nog bereiken als je te maken krijgt met zulke gegevens? Het antwoord is dat uit zulke gegevens nog heel wat valt af te leiden. Een stukje geschiedenis, gebaseerd op Westergaard (1969), Kendall (1970) en Egon S. Pearson (1978), kan het een en ander illustreren. Een voor de hand liggend vertrekpunt is het woord statistiek zelf.

De oorsprong van het woord statistiek

Het woord statistiek heeft het woord staat in zich en het werd gebruikt voor wat vroeger de vergelijkende beschrijving van staten heette, hetgeen door de Duitsers Staatenkunde werd genoemd en later Statistik door Gottfried A. Achenwall. Het woord statistiek schijnt oorspronkelijk van het Italiaanse woord statistica te komen. De geschiedenis van de Staatenkunde gaat terug tot Aristoteles. Wat we tegenwoordig onder statistiek verstaan vindt haar oorsprong ergens anders. Inhoudelijk gezien moet "Political Arithmetic" als de voorloper van de statistiek worden beschouwd.

De oorsprong van het vak statistiek

Londen rond 1660 wordt tegenwoordig als de bakermat van de hedendaagse statistiek beschouwd. In 1662 verscheen er namelijk een boek dat algemeen wordt bestempeld als het eerste boek van statistische aard. Dit boek werd geschreven door de koopman John Graunt (1620-1674) en het had als titel "Natural and political observations upon the bills of mortality". In die tijd was Londen al een redelijk grote stad met enkele honderdduizenden inwoners. Londen had al verscheidene keren te maken gehad met epidemieën van pest, waaraan heel veel mensen waren overleden. De parochies deden geregeld aankondigingen over het aantal dopen, het aantal begrafenissen en het aantal sterfgevallen van pest. U kunt zich voorstellen dat men in die tijd zeer veel belangstelling had voor die informatie. Ook werd er gerapporteerd over doodsoorzaken in het algemeen. Iedere parochie had een beëdigd persoon in dienst die doodsoorzaken moest vaststellen. Het zal u niet verbazen dat de lijst van mogelijke doodsoorzaken verre van goed gedefinieerd was en dat deze beëdigde personen vaak niet in staat waren de juiste doodsoorzaak vast te stellen. Sommige ziekten werden zelfs buiten beschouwing gelaten. Graunt had heel goed in de gaten dat de cijfers die verstrekt werden verre van volledig en nogal onnauwkeurig waren. Graunt twijfelde bijvoorbeeld sterk aan het geregistreerde aantal sterfgevallen ten gevolge van syfilis. Zo waren er nog een aantal zaken waar hij zo zijn vraagtekens bij plaatste. Hoewel de gegevens waarover hij beschikte niet

zo best waren, was hij toch in staat bepaalde wetmatigheden in de gegevens te constateren. Zijn manier van omgaan met zulke gegevens gaf een hele nieuwe impuls en inspireerde andere auteurs. Graunt schreef maar één boek, maar hij werd er beroemd mee. Hij werd zelfs gekozen tot lid van de pas opgerichte Royal Society. Het belang van zijn werk werd eigenlijk meteen ingezien. Eén van zijn vrienden was William Petty die de term "Political Arithmetic" heeft bedacht. Graunt lijkt de eerste te zijn die opmerkte dat er relatief meer jongens dan meisjes werden geboren. Wat Graunt in die tijd deed was daadwerkelijk werken buiten de grenzen van het weten.

Zonder overdrijving kan men stellen dat Graunt de uitvinder is van de sterftetafel. Hij gaf per leeftijdscategorie het gemiddelde aantal doden per 100 personen van de totale bevolking. Zulke tabellen werden later opgesteld voor mannen en vrouwen afzonderlijk. Graunt was overigens niet zo erg duidelijk over de verschillen tussen sterftecijfers van mannen en vrouwen. Al spoedig speelden zulke sterftetafels een belangrijke rol bij levensverzekeringen. Juist in Nederland, waar "gewone lijfrenten" in die tijd vaak gebruikt werden, was men erg geïnteresseerd in sterftetafels. We komen hier een aantal beroemdheden tegen. Graunt had via correspondentie met Lodewijk Huygens duidelijk invloed op onderzoek dat diens broer Christiaan Huygens in 1669 uitvoerde. Twee andere personen die een enorme invloed hebben gehad op de constructie van sterftetafels en het gebruik ervan voor lijfrenten waren Johannes Hudde en de beroemde staatsman Johan de Witt. Wat essentieel is, is dat iemand als Graunt met een grote diepgang redeneringen opbouwde over verre van perfecte gegevens. Voor mij is dat het wezen van de statistische wetenschap.

Waarom statistiek statistiek heet

Graunt wordt beschouwd als de grondlegger van de hedendaagse statistiek. Deze tak van wetenschap werd, zoals reeds gezegd, oorspronkelijk Political Arithmetic genoemd. U zult zich wel afvragen waarom het dan toch statistiek is gaan heten. De Staatenkunde werd in de loop van de tijd steeds numerieker van aard, maar het systematisch verzamelen van feiten was nog steeds het centrale thema. Halverwege de negentiende eeuw kon men voor het eerst een bepaalde mate van toenadering constateren. De Schot Sir John Sinclair stal het woord Statistik van de Duitsers en gebruikte het in plaats van de term Political Arithmetic. Dit gebeurde ergens tussen 1790 en 1798. Hij publiceerde in die periode "The statistical account of Scotland". Voor wat de geschiedenis van de wetenschap statistiek betreft, heeft deze naamsverandering wel voor de nodige onduidelijkheid gezorgd.

Statistiek als wetenschap

De statistiek als wetenschap houdt zich bezig met de vraag hoe men conclusies kan trekken uit gebrekkige statistische gegevens en op welke wijze men die conclusies dient te presenteren. De statisticus rekent het tot zijn verantwoordelijkheid aan te geven wat de kwaliteit van zijn uitspraken is. Hij kan bijvoorbeeld een interval bij een voorspelling aangeven, waarvan hij meedeelt dat de werkelijke uitkomst voor 1994 daarbuiten zal liggen met een risico van bijvoorbeeld hoogstens 5%. Hoe breder het interval is des te onzekerder is de voorspelling zelf, die meestal halverwege het interval zal liggen. Als men een hoger risico accepteert dan die 5%, dan zal de statisticus met een kleiner voorspelgebied kunnen komen.

Statistiek maakt gebruik van de veronderstelling dat de gegevens tot stand zijn gekomen volgens een bepaald toevalsmechanisme. Uitgaande van kennis over dit mechanisme is het in principe mogelijk tot bepaalde conclusies te komen. In de statistiek maakt men handig gebruik van de wetten van het toeval. Bij het steekproeftrekken stopt men dit toeval er zelf in door individuen bijvoorbeeld op aselechte wijze te kiezen uit een hele grote populatie, waarin men geïnteresseerd is. Als de steekproef groot genoeg is, dan is het steekproefgemiddelde een zeer betrouwbare schatter voor het gemiddelde in de populatie. In de praktijk zijn die steekproeven relatief heel klein. Alle mensen ondervragen zou onmogelijk zijn vanwege de kosten en de beschikbare tijd. In de statistische procesbeheersing heeft men vaak genoeg aan hele kleine steekproeven die met een zekere regelmaat worden herhaald.

De ontwikkelingen op het gebied van de kansrekening en de wiskunde zijn van zeer groot belang voor de statistische wetenschap geweest en zullen ook van groot belang blijven. Men kan bijvoorbeeld denken aan de rol van stochastische differentiaalvergelijkingen (zie bijvoorbeeld Øksendal (1991)) in populatie-groeimodellen en in financiële modellen. Verder speelt de kansrekening een belangrijke rol in de verzekeringswiskunde. Ik zal verder niet ingaan op de kansrekening, maar wil wel het belang van dit vakgebied voor de statistiek onderstrepen.

De statistiek als wetenschap bedient zich van wiskunde, kansrekening en computers bij het ontwerpen en evalueren van procedures volgens welke men cijfermatige gegevens kan analyseren en interpreteren, hypothesen kan toetsen, modellen kan schatten, of voorspellingen kan maken. Het gewenste type uitspraak wordt bovendien nog voorzien van een procedure volgens welke men de onzekerheden tot uitdrukking kan brengen. Centraal staat telkens een kans theoretisch model met onbekende parameters volgens welke de statistische gegevens geacht worden te zijn gegenereerd. In de werkelijkheid kan het best zo zijn dat de gegevens niet volgens een toevalsmechanisme tot stand zijn gekomen. Dit neemt niet weg dat een dergelijke

modelveronderstelling voor de praktijk vaak een bevredigende benadering is.

De vooruitgang van de statistiek wordt voor een belangrijk deel bepaald door probleemstellingen die uit toepassingsgebieden te voorschijn komen. De statistische theorie zelf roept natuurlijk ook hele interessante vraagstellingen op. Voor mij is die steeds terugkerende link met de praktijk heel stimulerend. Zo is het een voordeel dat statistici daadwerkelijk werkzaam zijn binnen de diverse faculteiten, hetgeen in Nederland overwegend het geval is. Immers, zo behoudt het vak bronnen van inspiratie en de impulsen voor nieuwe ontwikkelingen. De keerzijde is echter wel dat de statistische discipline in Nederland ook kwetsbaarder is, omdat er een zekere versnippering is. In de Verenigde Staten heeft iedere zichzelf respecterende universiteit een Department of Statistics, soms in combinatie met Mathematics. Een dergelijke structuur maakt de groep van statistici minder kwetsbaar. Ik wil daarmee niet beweren dat een dergelijke structuur ook in Nederland gewenst is.

In de Verenigde Staten en in Groot-Brittannië heeft het academische beroep van statisticus maatschappelijk en wetenschappelijk gezien een duidelijker profiel dan in Nederland. In Nederland begint dit te komen door de komst van de studierichting bedrijfs- en industriële statistiek aan de Universiteit van Amsterdam en de studierichting statistiek aan de Rijksuniversiteit Groningen. Voorts spant de VVS (Vereniging voor Statistiek en Operationele Research), een vereniging van ongeveer 1400 leden, zich in het vak te promoten.

Het belang van statistiek

Ontwikkelingen op andere vakgebieden kunnen een stimulans zijn voor de statistiek. Omgekeerd kan statistiek veel betekenen voor andere vakgebieden en maatschappelijke ontwikkelingen. Ik kan het niet laten dit te illustreren met een voorbeeld uit de geschiedenis, waarbij ik mij baseer op Rubens (1993). Een aantal van u zal wel eens gehoord hebben van Florence Nightingale, die beschouwd wordt als de grondlegster van de Verpleegkunde. Ze werd geboren op 12 mei 1820 in Florence naar welke plaats ze genoemd is. Ze is overleden in 1910. Ze is van goede, Engelse komaf. Haar vader nam een deel van haar onderwijs voor zijn rekening. Wiskunde, kansrekening en statistiek vormden een aanzienlijk deel van haar verdere opleiding. Haar ouders zagen dat niet zo zitten en eigenlijk werd van Florence verwacht dat ze zou trouwen, kinderen krijgen, etc. Ze voelde zich geroepen tot het verzorgende beroep van verpleegster en wenste dat ook een opleiding te volgen. Haar ouders waren hier heel erg op tegen. Een vrouw met een beroep: zoiets ging tegen het gebruikelijke rolpatroon in. Zij maakt een aanzienlijke carrière en wordt directrice van een zorginstelling. Vader

Nightingale krijgt genoeg van het geruzie in de familie over wat Florence doet en besluit haar jaarlijks een rente van 500 pond te verstrekken. Dit vergroot haar onafhankelijkheid. Ze wordt naar de Krim uitgezonden, waar een oorlog tussen Engeland en Frankrijk enerzijds en Rusland anderzijds was uitgebroken. Door haar contacten hoog in de politieke regionen kreeg ze het verzoek de leiding op zich te nemen van een groep verpleegsters die werd uitgezonden naar het oorlogshospitaal. Dit was een rechtstreeks gevolg van de grote verontwaardiging in Engeland als blijkt dat bijna de helft van de soldaten overlijdt door gebrek aan logistiek en vooral medische zorg en niet direct ten gevolge van vijandelikheden. Bovendien blijken de Fransen dit probleem niet te hebben. Florence Nightingale gebruikt de ideeën van de Belgische statisticus Adolphe Quetelet. Van meet af aan brengt ze de mortaliteit van de soldaten in kaart met behulp van grafieken. Statistiek wordt door haar beschouwd als een onmisbaar instrument bij het voeren van beleid. Florence Nightingale wist met behulp van statistiek duidelijk inzicht te verschaffen in de mortaliteit in de gewone Britse kazerne. Deze was in vredetijd veel te hoog zelfs als de soldaten geen oefeningen hadden. Zij vecht voor afwatering en sanitaire voorzieningen en schrijft als het niet anders kan naar de minister of Koningin Victoria. Men kan onmogelijk om haar gedetailleerde gegevens heen. Zoals Rubens(1993) schreef: "Als zij in 1910 sterft, is de verpleegkunde veranderd, maar is ook de statistiek definitief binnengedrongen in de epidemiologie en de beleidsinstrumenten.

Naar mijn stellige overtuiging is de statistische wetenschap geworden wat het is mede door de stimulans van mensen met een interesse die niet in eerste instantie de statistiek betrof.

Statistiek en onzekerheid

Men kan het woord toeval opvatten als een eufemistische uitdrukking voor onzekerheid. De Vesting, een vereniging van studenten Econometrie van deze universiteit, organiseerde op 26 mei 1993 een symposium onder de treffende titel "Meer zekerheid met statistiek". Inderdaad, Statistiek kan helpen meer zekerheid te verschaffen over onzekerheid, maar kan nooit alle onzekerheden wegnemen. Een statisticus houdt beroepshalve van onzekerheden, privé ligt dat natuurlijk anders. Het gevoel van onzekerheid kan natuurlijk een soort sensatie tweeweg brengen. Sommigen genieten ervan risico's te nemen. Toch streeft iedereen naar een bepaalde mate van zekerheid. We zijn bereid voorspellingen van deskundigen te accepteren om een verhoogd gevoel van zekerheid te krijgen. Vaak gaat het zover dat we met voorspellingen omgaan alsof het tamelijk zekere uitspraken over de toekomst betreffen. Het verschil tussen realisatie en voorspelling wil nog wel eens tegenvallen. We hoeven alleen maar de weerberichten te volgen. Natuurlijk willen we weten waarom het anders is gelopen, want er kan een

leer-effect optreden. De teleurstelling over het tegenvallen van een voorspelling is emotioneel gezien begrijpelijk, maar is deels ook te wijten aan het feit dat we niet verder informeren naar hoe zeker een deskundige is van zijn voorspellingen. Misschien overdrijf ik enigszins, maar de manier waarop sommigen omspringen met macro-economische prognoses doet vermoeden dat men zich niet altijd bewust is van de soms zeer grote onzekerheden daarin. In sommige maatschappelijke discussies houdt men juist heel veel rekening met onzekerheden.

Een statisticus behoort niet alleen als erom wordt gevraagd, maar ook - en misschien juist dan - als er niet om wordt gevraagd, aan te geven hoe betrouwbaar zijn uitspraken zijn. Dit heeft eigenlijk te maken met de ethiek van het vak. Het is helaas niet altijd mogelijk zowel in theorie als in de praktijk om betrouwbaarheidsuitspraken van hoge kwaliteit te doen. De intentie moet echter zijn dat er naar gestreefd wordt een beeld te geven van de kwaliteit van de onzekere conclusies hoe goed of hoe kwaad het ook gaat.

Het verzekeren heeft, zoals ik reeds heb aangegeven, een lange traditie in het uitbaten van onzekerheden. Het gegeven dat zoveel gezinnen, bedrijven en natuurlijk ook verzekeringsinstellingen zelf zich verzekeren onderstreept het maatschappelijke en het economische belang ervan. De verzekeringswiskunde levert technologie die wordt gebruikt om uit te rekenen hoe hoog de premie van een verzekeringspolis moet zijn. Dit vereist kennis omtrent de gezamenlijke kansverdeling van het optreden van een gebeurtenis die tot een claim leidt en de hoogte van die claim. Verzekeringswiskunde is dus een zeer belangrijk vak. Voor zover ik het kan bekijken is het verzekeren er niet eenvoudiger op geworden. In dit kader hoeven we alleen maar even te denken aan de zware verliezen bij Lloyd's. Naar mijn overtuiging vertonen de kansverdelingen die in het verleden tamelijk stabiel waren en waarop veel polissen en premiestellingen zijn gebaseerd tegenwoordig een eigen dynamiek. De gebruikelijke modellen geven een onderschatting van de huidige risico's. Dit leidt tot verhogingen van de premies. Verzekeringsstatistiek speelt een rol in het analyseren van de dynamiek in de risico's. Welke wetmatigheden en welke achtergrondkenmerken kan men in dit proces onderkennen op basis van onder andere gegevens over aantallen claims en de hoogtes ervan. De risico's van stormschade en autodiefstal zijn duidelijk toegenomen en wat te denken van de recente overstromingen. Ik denk dat de rol van de statistiek in het verzekeringswezen de komende jaren zal groeien omdat de aanwezige kennis over risico's uitgebreid zal worden met kennis die zit verborgen in de cijfers over claims, weersomstandigheden, criminaliteit, enz.

Toepassingen en statistiek

Statistiek speelt in maatschappij en wetenschap een min of meer onopvallende rol. Statistiek is zelf niet bepaald spraakmakend. Het gaat om de resultaten die mede met behulp van statistische technieken kunnen worden bereikt. Het dienstverlenende karakter moge blijken uit het gebruik van termen zoals statistische consultatie, methodologiewinkel en kwantitatieve ondersteuning. Eén van de mooie kanten van het vak statistiek is dat het op zoveel uiteenlopende terreinen wordt toegepast. Om u een idee te geven van de reikwijdte van het vak zou ik willen putten uit ervaringen die ik heb opgedaan uit contacten met klanten, ervaringen met stages van studenten en ervaringen met groepsopdrachten van studenten in het kader van het doctorale keuzevak toegepaste statistiek dat ik verzorg. Ik noem een aantal onderwerpen:

- constructie van de stamboom van de mens;
- calibratie van chromatografie systemen;
- effect van een bepaald medicijn op de levensduur van nierkankerpatiënten;
- de ouderdom van quasars;
- het UK lezersonderzoek;
- kostenontwikkeling in de gezondheidszorg in de regio Groningen;
- statistische procesbeheersing bij PTT Post;
- evaluatie verkeerscirculatieplan Gemeente Groningen.

De ruime toepasbaarheid van statistiek is één van de charmes van het vak. Als statisticus maak ik op deze wijze kennis met zeer uiteenlopende en stuk voor stuk interessante probleemstellingen.

Statistiek en kwaliteit

In de sfeer van de toepassingen bespeur ik een sterk groeiende belangstelling voor statistische procesbeheersing. Bekende namen op dit vakgebied zijn bijvoorbeeld W.A. Shewhart, W.E. Deming, J.M. Juran en G. Taguchi. De sterke concurrentie van Japan wordt voor een deel verklaard uit het gebruik van statistiek bij het streven naar kwaliteitsverbetering van produkten. In eerste instantie ging het vooral om kwaliteitscontrole, waarbij het vooral gaat om het keuren van partijen gereed produkt. Afkeuren van partijen is natuurlijk duur. Het kan extra werk betekenen in de vorm van reparaties

en als dat niet kan dan volgt er vernietiging of verkoop tegen een veel lagere prijs.

Geleidelijk aan groeide het besef dat het mogelijk moest zijn om afkeuring van partijen te voorkomen door de delen waaruit het productieproces is opgebouwd te beheersen. Als alle onderdelen binnen het proces goed functioneren, dan moet het eindproduct wel goed zijn. Door het gebruik van statistische procesbeheersing, hetgeen op de werkvloer tot uitdrukking komt in onder andere het gebruik van regelkaarten in te vullen door de operators, wordt op basis van steekproeven van relatief kleine omvang de kwaliteit van het proces bewaakt in de loop van de tijd. Zo komt men tijdig storingen op het spoor of kan men kleine afwijkingen ten opzichte van een optimale instelling van de machine achterhalen.

Kwaliteit kan men ook proberen te verhogen door eisen te stellen aan de kwaliteit van de geleverde grondstoffen. Vroeger controleerde men die meestal bij ontvangst en had de leverancier de partij soms ook al gecontroleerd. Tegenwoordig gaat men er vaker toe over in het contract met een leverancier vast te leggen aan welke eisen grondstoffen moeten voldoen en hoe de leverancier deze kwaliteit kan garanderen met behulp van een systeem van kwaliteitsbeheersing. Op deze wijze is het niet meer nodig binnenkomende partijen te controleren. De leverancier wordt dan wel gecontroleerd op het functioneren van dit systeem.

Uitval tijdens het productieproces kan mede beperkt worden door het ontwerp van de produktielijn en het ontwerp van het produkt zelf te optimaliseren. Er wordt meer en meer gestreefd naar een design van het produkt zodanig dat er tijdens de fabricage bijna geen problemen meer zijn. Dit betekent dat de experimentele fase van een produkt zeer belangrijk wordt. Ik had de naam Taguchi reeds laten vallen. Deze Japanner wist de engineers ervan te overtuigen dat het gebruik van statistische technieken uit de theorie over experimentele proefopzetten heel nuttig is in de ontwerpfase. Vanuit kostenoverwegingen is het enorm belangrijk met zo weinig mogelijk experimenten een zo'n goed mogelijk resultaat te behalen. Hier kan de statistiek enorm dienstbaar zijn. De theorie van experimentele proefopzetten wordt niet alleen bij het ontwerpen van produkten toegepast, maar vindt ook heel veel toepassing op landbouwkundige proefstations en bij marktonderzoek.

Bij de statistische procesbeheersing is statistiek natuurlijk enorm belangrijk. In de praktijk gaat de meeste tijd zitten in het rijp maken van de organisatie om van hoog tot laag instemming te verkrijgen voor een systeem van statistische kwaliteitsbeheersing. Dit moet zover gaan dat iedereen ervan overtuigd is dat het een goed idee is. Het zal u duidelijk zijn dat voor het bereiken van "commitment" het woord statistiek zo weinig mogelijk moet worden gebruikt. Centraal moet staan hoe het gebruik van regelkaarten ten dienste kan staan van de organisatie als geheel en alle betrokkenen

in het bijzonder. Waar de statistiek een rol speelt, is bij het opstellen van de regelkaarten, het vaststellen wanneer en hoe vaak er een steekproef moet worden getrokken en hoe groot deze moet zijn. Het heeft voordelen de operator bij de machine de regelkaarten te laten bijhouden al dan niet met gebruik van een computer. Op deze wijze kan iemand namelijk zien hoe de zaak ervoor staat en dit vergroot de betrokkenheid op de werkvloer. Het een en ander vergt veel voorbereiding. Er moeten bijvoorbeeld interne opleidingen worden verzorgd. Het gaat om het kwaliteitsdenken op zich en regelkaarten vormen daarbij een instrument.

Statistiek en informatie

Het introduceren van regelkaarten op de werkvloer heeft duidelijk te maken met de uitspraak "meten is weten", welke goed past in dit informatietijdperk. Met de moderne computerfaciliteiten en de enorme mogelijkheden om heel veel informatie op te slaan begint ook het fenomeen overinformatie te ontstaan. Men kan dit tegenkomen op verschillende manieren. Het kan gebeuren dat er zoveel informatie beschikbaar is dat men het niet meer kan overzien. De gebruiker komt dan in een situatie terecht die valt te vergelijken met een toestand van weinig beschikbare informatie. De individualisering van de samenleving brengt met zich mee dat er enorme bestanden nodig zijn om per individu allerlei kenmerken vast te leggen. Het is enorm veel wat er bijvoorbeeld per student aan gegevens wordt bewaard in verband met de studiefinanciering. De mogelijkheden van de computer leiden ertoe dat die mogelijkheden ten volle worden benut. Echter het manipuleren van dergelijke bestanden is door de complexiteit een zeer tijdrovende zaak. Om de produktie niet te verstoren kan het gebeuren dat men voor statistische analyses de kans krijgt om eenmalig een kleine steekproef te trekken uit zo'n bestand op basis waarvan de meest uiteenlopende beleidsvragen moeten worden beantwoord.

Met databestanden is er altijd wel iets dat de zaak behoorlijk kan compliceren. Bij een aantal recente ervaringen heb ik kunnen meemaken dat aanzienlijke gegevensbestanden veel onjuistheden bevatten. Het bleek mogelijk veel fouten te corrigeren. Naar mijn gevoel wordt er te weinig aandacht besteed aan het screenen van gegevens en het zorgvuldig invoeren en verzamelen ervan. De aard van de fouten was van dien aard dat het gevolgen voor conclusies kan hebben gehad. Waar men ook nogal eens mee te maken kan krijgen is het ontbreken van allerlei gegevens. Het komt voor dat er zoveel gegevens op de diverse karakteristieken van een steekproefelement ontbreken dat een zeer klein percentage van de steekproef volledig gemeten is. Het is helaas niet altijd mogelijk zo'n situatie te voorkomen, maar het maakt een degelijke analyse vrijwel onmogelijk. Met veel kunst en vliegwerk is er nog wel eens iets van te maken, maar van de

betrouwbaarheid moet men echter niet teveel verwachten.

Een andere ervaring is dat er voor een jaar goede daggegevens beschikbaar zijn, maar dat de oudere jaargegevens worden vernietigd. Om de tijdreeks voor een bepaalde variabele te analyseren heeft men wel de gegevens voor een aantal jaren nodig, want op basis van gegevens voor één jaar is het haast onmogelijk om bijvoorbeeld seizoeneffecten te schatten.

De beschikbaarheid van informatie en statistische software maakt het steeds eenvoudiger voor iedereen om zelf allerlei analyses te doen. Het gevolg ervan is probeergedrag: er zijn heel wat pogingen gedaan om met het software pakket tot bevredigende resultaten te komen. Dit soort analyses zijn zeer riskant. De meeste statistische software is ontworpen voor situaties, waarbij men van te voren heeft vastgesteld wat en hoe er geanalyseerd moet worden. Zo'n softwarepakket geeft een kwaliteitsoordeel voor de gevonden uitkomsten. Als zo'n pakket rekening zou houden met probeergedrag, dan zouden de kwaliteitsoordelen er veel ongunstiger uitzien.

Een ander fenomeen is dat van het interpreteren van ruis. Als men getallen zou krijgen die volstrekt ruis zijn en als men probeergedrag toelaat, dan is de kans bijna één dat men een bevredigend model vindt. Een aardig experiment is een aantal jaren geleden door Breiman en Freedman (1983) gedaan. De uitkomst had natuurlijk moeten zijn dat er geen structuur in de gegevens zit. Na het uitmelken van de gegevens krijgt men een model waarvan het risico groot kan zijn dat het niets zegt over de werkelijkheid. Vandaar dat ik er altijd voor pleit om met de klant te overleggen welke variabelen kunnen worden weggelaten en of hij een volgorde van belangrijkheid aan kan geven. Op deze wijze kan uitgebreid experimenteren worden voorkomen. Men kan ook 70% van de steekproef opofferen om eventueel na veel proberen een bevredigend model te vinden. Vervolgens bekijkt men hoe goed het model bevalt voor de overgebleven 30%. Het verschijnsel van aanzienlijke overschatting van de kwaliteit na het uitgebreid manipuleren van de gegevens staat bekend onder de term kanskapitalisatie. In het Engels heet het datamining of datapeeping.

Kanskapitalisatie heeft duidelijk te maken met de gebruiker die te goeder trouw een bepaalde statistische techniek een aantal keren loslaat op hetzelfde databestand. In de praktijk zijn veel mensen zich van dit fenomeen niet bewust. Een ander gevaar dat zo'n gebruiker loopt heet het piekverschijnsel of de vloek van de dimensionaliteit, in het Engels: peaking phenomenon of curse of dimensionality. Dit verschijnsel is voor het eerst ontdekt bij de patroonherkenning. Bij deze term kan men denken aan herkenning van handschriften, spraak, landschappen, en dergelijke. Laten we even stilstaan bij het herkennen van bijvoorbeeld letters. Per letter valt er veel te meten. In een systeem waarbij we denken aan het sorteren van brieven op postcode, komt het erop neer dat het verschrikkelijk snel moet. Dit betekent dat men het met zo weinig mogelijk karakteristieken moet doen

ook al is de computer enorm snel. Men kwam erachter dat de kwaliteit van het herkennen, zoals het percentage goed herkende letters, verbeterde als het aantal karakteristieken toenam. Dit is logisch, zult u zeggen. Ging men door met het toevoegen van karakteristieken, dan gebeurde er iets gekks. De kwaliteit begon na enige tijd te dalen en het kon zelfs gebeuren dat het geheel ontaardde en de kwaliteit veel slechter werd dan die voor een klein aantal karakteristieken. Er is een eenvoudige economische term die beschrijft wat er gebeurt. Een extra karakteristiek betekent het gebruik van meer informatie maar men krijgt te maken met zoiets als afnemende meeropbrengsten. Het introduceren van extra variabelen betekent inderdaad meer informatie, maar het brengt ook extra onzekerheden met zich mee ten aanzien van nieuwe, onbekende parameters. Op een gegeven moment is de winst in informatie niet groot genoeg om de additionele onzekerheid te compenseren. Een groot model wekt de indruk de werkelijkheid goed te beschrijven, maar de voorspelkwaliteit kon wel eens tegenvallen. Men moet dus zuinig zijn met het aantal te gebruiken variabelen.

De fenomenen kanskapitalisatie en vloek van de dimensionaliteit kan men beschouwen als keerzijden van dezelfde medaille. Een toepasser die zuinig wil zijn met het aantal variabelen in zijn model, is geneigd een aantal combinaties van variabelen te proberen en krijgt dan te maken met het risico van kanskapitalisatie. Het vervelende is dat de argeloze gebruiker zich dit allemaal niet realiseert. In een sombere bui denk ik wel eens dat er met behulp van statistiek ook de nodige onzin is bewezen. Men moet dus niet proberen zoveel mogelijk uit de gegevens te wringen. Integendeel, men moet eerst zoveel mogelijk uit de klant zien te krijgen. Mogelijk kan het ook helpen als men eerst vastlegt hoe men de beschikbare gegevens denkt te analyseren en wat men van de diverse analyses verwacht, voordat de computer wordt aangezet.

De geschetste problemen van de argeloze gebruiker hebben alles te maken met de geweldige mogelijkheden van de pc. Toen de automatisering haar intrede nog niet had gedaan, haalde men het niet in zijn hoofd om veel statistische analyses uit te proberen op zijn gegevens. De berekeningen moesten immers met de hand geschieden. Om dezelfde reden bleef het aantal in de analyses te betrekken karakteristieken beperkt. De positieve kanten van de computer zijn u wel bekend, daarom heb ik de geweldige voordelen daarvan niet aan de orde gesteld. Ik heb gemeend van de gelegenheid gebruik te moeten maken wel de nadelen van een al te argeloos gebruik te bespreken.

Referenties

- Breiman, L., D. Freedman (1983), *How many variables should be entered in a regression equation?*; Journal of the American Statistical Association, vol. 78, 131-136.
- Freedman, D.A., W.C. Navidi (1986), *Regression models for adjusting the 1980 census*; Statistical Science, vol. 1, 3-39 (inclusief discussie).
- Kendall, M.G. (1970), *Where shall the history of statistics begin?*; pagina's 45-46 uit Pearson, E.S., M.G. Kendall, editors (1970), *Studies in the history of statistics and probability*; Charles Griffin & Company Limited, Londen (oorspronkelijk verschenen in Biometrika, vol. 47 (1960), 447-449).
- Øksendal, B. (1991), *Stochastic differential equations*; third edition, Springer-Verlag.
- Pearson, E.S., editor (1978), *The history of statistics in the 17th & 18th century against the changing background of intellectual, scientific and religious thought; Lectures by Karl Pearson given at University College Londen during the academic sessions 1921-1933*; Charles Griffin & Company Limited, Londen.
- Rubens, R. (1993), *Florence Nightingale, grondlegster van de verpleegkunde; Haar verbondenheid met Quetelet en het wetenschappelijk denken*; Geschiedenis der Geneeskunde, jaarg. 1, 56-64.
- Westergaard, H. (1969), *Contributions to the history of statistics*; herdruk van de editie van 1932, Augustus M. Kelley Publishers, New York.

