Opfriscursus statistiek

G. Cotteleer
K. Gardebroek
H.C.J. Vrolijk
W. Dol

Projectcode 63710

Mei 2003

Rapport 8.03.05

LEI, Den Haag

Het LEI beweegt zich op een breed terrein van onderzoek dat in diverse domeinen kan worden opgedeeld. Dit rapport valt binnen het domein:

- ☐ Wettelijke en dienstverlenende taken
- ☐ Bedrijfsontwikkeling en concurrentiepositie
- ☐ Natuurlijke hulpbronnen en milieu
- ☐ Ruimte en Economie
- ☐ Ketens
- ☐ Beleid
- ☐ Gamma, instituties, mens en beleving
- ☑ Modellen en Data

Opfriscursus statistiek
Cotteleer, C., K. Gardebroek, H.C.J. Vrolijk en W. Dol
Den Haag, LEI, 2003
Rapport 8.03.05; ISBN 90-5242- ; Prijs 21,25 euro (inclusief 6% BTW)
149 p., fig., tab.

Dit rapport sluit nauw aan op het cursusmateriaal dat gebruikt is tijdens de opfriscursus statistiek. Voor zowel mensen die de cursus gevolgd hebben als voor mensen die dit niet gedaan hebben is dit naslagwerk aan te raden. Een uitwerking van de praktijkcases die tijdens de cursusbijeenkomsten aan de orde zijn gekomen, is terug te vinden in dit naslagwerk. De uitwerkingen kunnen door iedereen bekeken worden en bovendien kunnen de praktijkcases uitgevoerd worden aan de hand van de bijbehorende datasets die beschikbaar zijn via de intranetsite van het Domeinteam Data en Modellen.

Dit rapport is niet bedoeld als nieuw statistiekboek, maar meer als een houvast bij statistische exercities. Er is dan ook niet geprobeerd de gehele stof te dekken, er wordt net als in de opfriscursus zelf slechts een beknopt overzicht gegeven van een breed spectrum aan statistische technieken. De behandelde technieken worden kort toegelicht.

Bestellingen:
Telefoon: 070-3358330
Telefax: 070-3615624
E-mail: publicatie@lei.wag-ur.nl

Informatie:
Telefoon: 070-3358330
Telefax: 070-3615624
E-mail: informatie@lei.wag-ur.nl

© LEI, 2003
Vermenigvuldiging of overname van gegevens:
☑ toegestaan mits met duidelijke bronvermelding
☐ niet toegestaan

Op al onze onderzoeksopdrachten zijn de Algemene Voorwaarden van de Dienst Landbouwkundig Onderzoek (DLO-NL) van toepassing. Deze zijn gedeponeerd bij de Kamer van Koophandel Midden-Gelderland te Arnhem.

Inhoud

Blz.

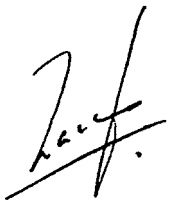
Woord vooraf

Vanuit het onderzoek is er vraag naar een betere onderbouwing van kwantitatieve conclusies. Binnen het LEI wordt veel statistisch onderzoek verricht op basis van beschikbare gegevensverzamelingen, zoals de Landbouwtelling, het Bedrijven-Informatienet van het LEI (het Informatienet), prijsstatistieken en enquêtes die door onderzoekers zelf zijn uitgezet. Het is belangrijk dat deze data op een efficiënte en effectieve wijze worden aangewend en ook volledig wordt benut. Om deze reden is in 2002 op het LEI voor het eerst een 'opfris'-cursus statistiek opgezet, waarin het gebruik van de bovengenoemde databronnen centraal heeft gestaan. Ook is er aandacht besteed aan de opzet van enquêtes. In de cursus lag de nadruk op de praktische toepasbaarheid van bestaande statistische methoden en technieken.

In de cursus is zoveel mogelijk maatwerk geleverd door voorafgaand aan de cursusbijeenkomsten te inventariseren welke behoeften er bij onderzoekers bestonden. Aan de hand van deze inventarisatie is door de cursusleiders een programma opgesteld dat aansloot bij deze behoeften.

Om de kennis opgedaan tijdens de opfriscursus te kunnen behouden is dit naslagwerk gemaakt. In dit naslagwerk is dan ook de cursusstof van de tien cursusbijeenkomsten terug te vinden. Daarnaast is een aantal praktijkcases terug te vinden waarmee onderzoekers zelf aan de slag kunnen. Deze zijn ook tijdens de cursus behandeld. Via de intranetsite van het Domeinteam Data en Modellen zijn de data die bij de praktijkcases horen te benaderen. Dit naslagwerk kan ook geraadpleegd worden door onderzoekers die niet hebben deelgenomen aan de cursus.

De cursusleiders waren Koos Gardebroek, Hans Vrolijk en Wietse Dol. Wij bedanken de onderzoekers die deel hebben genomen aan de cursus en hopen dat zij in de toekomst zinvol gebruik kunnen maken van hun opgedane kennis.



Prof.dr.ir. L.C. Zachariasse
Directeur LEI B.V.

Samenvatting

Dit naslagwerk sluit nauw aan op het cursusmateriaal dat gebruikt is tijdens de opfriscursus statistiek die binnen het LEI voor LEI-medewerkers is georganiseerd. Dit naslagwerk is aan te bevelen voor zowel mensen die de cursus gevolgd hebben als voor mensen die dit niet gedaan hebben. Een uitwerking van de praktijkcases die tijdens de cursusbijeenkomsten aan de orde zijn gekomen, is ook terug te vinden in dit naslagwerk. De uitwerkingen kunnen door iedereen bekeken worden en bovendien kunnen zij zelf uitgevoerd worden aan de hand van de bijbehorende datasets die beschikbaar zijn via de intranetsite van het Domeinteam Data en Modellen.

Dit rapport is niet bedoeld als nieuw statistiekboek, maar meer als een houvast bij statistische exercities. Er is dan ook niet geprobeerd de gehele stof te dekken, er wordt net als in de opfriscursus zelf slechts een beknopt overzicht gegeven van een breed spectrum aan statistische technieken. De behandelde technieken worden kort toegelicht.

Opzet enquête
Uit de inventarisatie die ten behoeve van de opfriscursus statistiek onder LEI-medewerkers is uitgevoerd, bleek dat 70% van de (73) responderende onderzoekers ervaring heeft met het opzetten van enquêtes. Daarnaast vormen enquêtes veelal het startpunt van veel kwantitatief onderzoek. Het Informatienet kan ook opgevat worden als een enquête. De manier waarop enquêtes worden uitgevoerd, is dan ook bepalend voor het verdere verloop van het onderzoek. Het belang van een goede opzet van enquêtes is dan ook groot.

Voordat er gebruik wordt gemaakt van een enquête, is het van belang na te gaan of dit wel het juiste instrument is voor het beantwoorden van de onderzoeksvragen. Als een enquête het juiste instrument is, is het de vraag of er gebruik kan worden gemaakt van secundaire informatie en of het onderzoek een kwalitatieve of een kwantitatieve insteek verdient. Zijn deze vragen beantwoord en blijkt dat er geen relevante secundaire informatie voorhanden is, dan kan er een steekproeftrekking plaatsvinden. De procedure voor het opstellen van steekproeven is als volgt: als eerste wordt een doelpopulatie gedefinieerd, vervolgens wordt het steekproefkader vastgesteld. Dan wordt de steekproeftechniek bepaald en dient de steekproefomvang vastgesteld te worden. Als laatste wordt de steekproef gegenereerd.

Opzet Informatienet
In het voorgaande hoofdstuk is de opzet van enquêtes in het algemeen gesproken. In dit hoofdstuk wordt het Informatienet besproken. Het Informatienet is een enquête die door het LEI is opgezet en waaraan jaarlijks ongeveer 1.500 primaire agrarische bedrijven meedoen. De informatie uit het Informatienet wordt zowel binnen als buiten het LEI breed gebruikt. In het Informatienet is gekozen voor een gestratificeerde steekproef.

Introductie SPSS
Een onderzoek met behulp van SPSS kan bestaan uit de volgende fasen:
- het opstellen van een vragenlijst;

- het afnemen van een vragenlijst;
- het maken van een gegevensbestand;
- het bewerken van een gegevensbestand;
- het analyseren van de gegevens;
- het interpreteren van de gegevens;
- het rapporteren van de gegevens.

Van de bovengenoemde onderzoeksstappen vinden de eerste twee plaats buiten SPSS. Hoewel bij het opstellen van de vragenlijst wel nagedacht moet worden over de verwerking van de gegevens en de mogelijkheden die SPSS hierbij biedt. Je wilt natuurlijk je gegevens zo optimaal mogelijk gebruiken bij de analyse. Bij de derde tot en met de vijfde stap, het maken, bewerken en analyseren van het gegevensbestand, wordt wel gebruikgemaakt van SPSS. Vervolgens vindt het interpreteren en het rapporteren plaats. Dit kan zowel met als zonder gebruikmaking van SPSS. SPSS kent functionaliteiten voor het invoeren, inlezen, bewerken, analyseren en presenteren van data.

Schatters, hypothesen toetsen, verkennen en presenteren data
Je kunt de beschrijvende statistiek vergelijken met het maken van een samenvatting over een literair boek. In plaats van het gehele boek te lezen, kun je met de samenvatting de essentiële informatie overbrengen. Dit is ook de bedoeling bij de beschrijvende statistiek. Door op een juiste manier gebruik te maken van samenvattende statistieken en grafieken kan een goed inzicht in de populatie worden verkregen. Dit inzicht kan ondersteunend zijn bij het nemen van beleidsbeslissingen.

Wanneer je de gehele populatie kunt waarnemen dan kun je de samenvattende statistieken voor deze populatie uitrekenen. Deze samenvattende statistieken worden ook wel populatieparameters genoemd. Wanneer je echter slechts een deel van de populatie kent (een steekproef) dan ben je slechts in staat een inschatting te maken van de populatieparameters. Zodra er geschat wordt dan is er sprake van onzekerheid. Dat wil zeggen: je weet niet voor 100% zeker dat de waarde die je hebt uitgerekend ook de juiste populatieparameter is. Als je gebruik maakt van schatters ben je wel in staat om een betrouwbaarheidsinterval te geven. Dit is een interval waarvan je met een bepaalde zekerheid weet dat de echte populatieparameter hierin valt.

Zodra data gepresenteerd wordt kunnen er allerlei vragen opkomen. Bijvoorbeeld: 'Zijn de inkomsten van dit jaar werkelijk hoger dan die van vorig jaar?' en 'Heeft iets wel invloed op een bepaald proces of niet?'. Omdat de data die gepresenteerd worden vaak het resultaat zijn van een steekproefschatting, moeten deze vragen statistisch getoetst worden voordat er met een bepaalde statistische zekerheid gezegd kan worden of iets wel of niet het geval is.

Technieken om resultaten steekproefschattingen te verbeteren
Regelmatig voert het LEI onderzoek uit waarbij resultaten voor een klein gebied (gemeenten, provincies, landbouwgebieden, kaartvierkanten) of kleine groep gewenst zijn. In veel gevallen worden deze resultaten geproduceerd door beschikbare of berekende bedrijfsgegevens 'op te hogen' naar het gewenste aggregatieniveau. Het is noodzakelijk dat er voldoende waarnemingen voor het gebied zijn om verantwoord te kunnen aggregeren.

Aggregatie van gegevens die betrekking hebben op Informatienet-bedrijven is voor klei-

ne gebieden veelal niet mogelijk op basis van de gebruikelijke procedure die gebruik maakt van wegingsfactoren. In de loop van de tijd zijn daarom op het LEI verschillende methodes toegepast/ontwikkeld die het mogelijk maken om toch bruikbare informatie op een laag ruimtelijk aggregatieniveau te genereren. Het is nuttig een vergelijking te maken van de beschikbare methoden voor het maken van schattingen voor kleine gebieden. De methoden die in aanmerking komen om schattingen te maken op basis van kleine deelgebieden zijn:

- de directe schatter;
- de ratioschatter;
- de regressieschatter;
- de bayesiaanse schatter;
- poststratificatie;
- datafusie en imputatie.

Regressieanalyse

Regressieanalyse is één van de meest gebruikte technieken in kwantitatief onderzoek. Met regressieanalyse probeert men de waargenomen spreiding in een (meetbare) afhankelijke variabele te verklaren met behulp van onafhankelijke verklarende variabelen. Er zijn drie redenen waarom je een regressieanalyse uit kunt voeren. Ten eerste kun je geïnteresseerd zijn in de samenhang tussen de afhankelijke variabele en één of meer verklarende variabelen. Hoe hangen bijvoorbeeld tarweopbrengsten op individuele bedrijven samen met gebruik van meststoffen, gewasbeschermingsmiddelen en het weer. Of: wat is de relatie tussen het aandeel biologische landbouw in Nederland en prijzen van biologische producten, overheidssteun en het plaatsvinden van voedselcrises? Een tweede reden om een regressieanalyse uit te voeren is het voorspellen van een nog niet geobserveerde afhankelijke variabele met behulp van al wel geobserveerde verklarende variabelen. Tot slot kun je met regressieanalyse ook nagaan in hoeverre de afhankelijke variabele verandert als de verklarende variabelen veranderen.

Trendanalyse

In dit hoofdstuk over trendanalyse worden twee onderwerpen behandeld: structurele verandering in regressiemodellen en tijdreeksanalyse met één variabele. Een impliciete veronderstelling die bij regressieanalyse wordt gemaakt, is dat de relatie tussen de afhankelijke en de verklarende variabelen voor alle waarnemingen hetzelfde is. Deze veronderstelling is echter niet altijd geldig. De relatie die in het regressiemodel aanwezig is kan verschillen voor bepaalde groepen van waarnemingen. Dit uit zich in verschillende parameterwaarden voor verschillende groepen. Structurele veranderingen in regressiemodellen die ontstaan in de tijd uiten zich in verschillende parameters voor verschillende periodes.

Het accent in regressiemodellen ligt op structurele relaties tussen variabelen. Met behulp van verklarende variabelen wordt het verloop van de afhankelijke variabele verklaard. Bij tijdreeksanalyse ligt de nadruk op het gedrag van een variabele in het verleden. Door patronen in dit gedrag te modelleren kan het verloop van de variabele naar de toekomst geprojecteerd worden. Tijdreeksanalyse zou je kunnen omschrijven als verstandig extrapoleren door onder andere te zoeken naar trends, patronen en cycli in een variabele. Bij tijdreeksanalyse is de tijdshorizon van de data doorgaans veel langer dan bij regressieanalyse.

Keuzemodellen

Keuzemodellen zijn modellen die gebruikt worden voor de analyse van keuzes. De modellen zijn een verbijzondering van lineaire regressiemodellen. De eenvoudigste modellen gaan uit van twee mogelijke keuzes. De afhankelijke variabele Y_i wordt in lineaire regressiemodellen continu verondersteld (bijvoorbeeld prijs, inkomen, productie). De onafhankelijke variabelen X_{ij} zijn doorgaans ook continu. Daarnaast kunnen er enkele onafhankelijke dummyvariabelen opgenomen zijn in het model. Voorbeelden van dummyvariabelen zijn geslacht (man of vrouw), wel of geen diploma, wel of niet werkloos enzovoorts. In keuzemodellen zijn dummyvariabelen de te verklaren variabelen. Uitbreidingen op de binaire keuzemodellen zijn ook mogelijk, er kunnen bijvoorbeeld meer dan twee mogelijkheden zijn: je hebt bijvoorbeeld de keuze tussen gangbare, geïntegreerde of biologische landbouw.

Hoofdcompontenten en factoranalyse
Hoofdcomponentenanalyse en factoranalyse zijn twee aparte technieken. Toch worden ze doorgaans niet apart onderscheiden (zie bijvoorbeeld SPSS). Dit komt mede doordat de eerste stap in een factoranalyse vaak een hoofdcomponentenanalyse is. En doordat er naast verschillen ook een aantal overeenkomsten in beide technieken is. Zo analyseren beide technieken relaties tussen een (groot) aantal variabelen en definiëren zij gemeenschappelijke onderliggende dimensies (hoofdcomponenten of factoren). Het doel bij beide technieken is het samenvatten van data of datareductie. Toepassingsmogelijkheden zijn bijvoorbeeld: het bepalen van hoofdkenmerken van een bedrijfsstructuur, het bepalen van het imago van een bepaald product of het karakteriseren van management. In alle drie de gevallen gebeurt dit aan de hand van een aantal hoofdkenmerken. Bij deze technieken wordt gezocht naar breed geformuleerde kenmerken die meerdere variabelen omvatten.

Clusteranalyse
Bij clusteranalyse vormen we homogene groepen van observaties uit een dataset van n waarnemingen met p variabelen. Waar we dus bij hoofdcomponenten- en factoranalyse variabelen probeerden samen te vatten, richten we ons bij clusteranalyse op (groeperen/clusteren van) waarnemingen. Elke groep/cluster bevat waarnemingen met dezelfde scores op bepaalde kenmerken. Er worden dus homogene groepen/clusters gevormd. De clusters verschillen op basis van die kenmerken. Het aantal clusters is echter op voorhand niet bekend. Daarin verschilt clusteranalyse van bijvoorbeeld de keuzemodellen. De clustering geschiedt op basis van een numerieke procedure.

Multidimensional scaling en conjunctanalyse
Multidimensional scaling (MDS) en conjunctanalyse (CA) zijn beide methoden waarmee percepties en voorkeuren van consumenten/respondenten kunnen worden gemeten en weergegeven. In tegenstelling tot de andere methoden die in deze cursus aan de orde zijn geweest, kunnen beide methoden worden toegepast op een individuele respondent. MDS en CA zijn recente toevoegingen aan SPSS. In SPSS kan slechts een beperkte verzameling MDS-analyses worden uitgevoerd. CA kan uitsluitend via de syntaxvensters worden uitgevoerd.

Het doel van MDS is te komen tot een ruimtelijke weergave van percepties of preferenties. Een afgeleide doelstelling is het bepalen van de dimensies waarop de vergelijkingen en of voorkeuren zijn gebaseerd.

Conjunctanalyse beschouwt een product als een pakketje attributen. Een televisie heeft

bijvoorbeeld een beeldbuis met een bepaalde omvang, al dan niet een afstandbediening en al dan niet teletekst. Conjunctanalyse probeert vervolgens het nut van attribuutniveaus (part worth oftewel deelnut) en het belang van attributen te achterhalen.

1. Inleiding

Dit naslagwerk sluit nauw aan op het cursusmateriaal dat gebruikt is tijdens de opfriscursus statistiek die binnen het LEI voor LEI-medewerkers is georganiseerd. Gedurende de cursus zijn powerpointsheets verspreid onder de cursisten. De sheets geven echter niet genoeg informatie om ze zonder de bijbehorende presentatie te kunnen begrijpen. Om deze reden is besloten het cursusmateriaal te bundelen en nader toe te lichten in een naslagwerk, zodat de statistische kennis opgedaan of opgefrist tijdens de cursus ook behouden blijft. Daarnaast is het naslagwerk ook aan te bevelen voor mensen die de cursus niet hebben gevolgd. Een uitwerking van de praktijkcases die tijdens de cursusbijeenkomsten aan de orde zijn gekomen, is ook terug te vinden in dit naslagwerk. De uitwerkingen kunnen door iedereen bekeken worden en bovendien kunnen de cases zelf uitgevoerd worden aan de hand van de bijbehorende datasets die beschikbaar zijn via de intranetsite van het Domeinteam Data en Modellen.

Dit rapport is niet bedoeld als nieuw statistiekboek, maar meer als een houvast bij statistische exercities. Er is dan ook niet geprobeerd de gehele stof te dekken, er wordt net als in de opfriscursus zelf slechts een beknopt overzicht gegeven van een breed spectrum aan statistische technieken. Voor mensen die meer over een bepaald onderwerp willen weten is er per onderwerp een literatuuroverzicht gegeven, waarbij ook de moeilijkheidsgraad van deze literatuur wordt aangegeven. Binnen de statistiek is een aantal richtingen te onderscheiden:

- steekproeven;
- kansrekening;
- beschrijvende statistiek;
- verklarende statistiek.

Met behulp van steekproeftechnieken probeert men inzicht te krijgen in de gehele populatie aan de hand van een steekproef, die slechts bestaat uit een deelverzameling van de gehele populatie. Steekproeftechnieken worden besproken in de hoofdstukken 2, 3 en 6. Een belangrijk voorbeeld van een steekproef bij het LEI is het Informatienet. In deze steekproef worden gegevens verzameld van primaire agrarische bedrijven.

Kansrekening is een onderdeel uit de statistiek dat niet of nauwelijks bij het economisch onderzoek van het LEI wordt gebruikt. Voorbeelden van deze richting zijn meer te vinden in de verzekeringsstatistiek. Kansrekening is dan ook niet behandeld tijdens de cursus en komt niet terug in dit naslagwerk. De beschrijvende statistiek probeert een populatie zo goed mogelijk te beschrijven aan de hand van parameters (statistics). Hoofdstuk 5 bespreekt deze richting in de statistiek. Op het LEI wordt de beschrijvende statistiek veel gebruikt om de gegevens die in het Informatienet zijn verzameld te presenteren. De jaarlijkse rapportages (*LEB* en *Actuele ontwikkeling van bedrijfsresultaten en inkomens*) zijn hier voorbeelden van. De verklarende statistiek probeert verbanden tussen variabelen te achterhalen. In de hoofdstukken 7, 8, 9, 10 en 11 wordt deze richting besproken. Veel beleidsvragen kunnen alleen worden opgelost als duidelijk is welke variabelen een bepaald probleem beïnvloeden en hoe de verschillende variabelen die betrekking hebben op het probleem met elkaar samenhangen. Als bekend is hoe de

samenhang is dan kan ook nagegaan worden hoe variabelen zijn te sturen om een gewenste uitkomst te krijgen.

Er zijn veel (introductie)boeken over statistiek geschreven. De meeste problemen die gedurende het onderzoek opgelost moeten worden staan echter niet als kant-en-klare oplossingen in deze boeken vermeld. Dit betekent dat bekend moet zijn welke methode voor welk probleem is toe te passen en dat daarna met veel doorzettingsvermogen en creativiteit het onderzoeksprobleem moet worden opgelost. Het is daarom belangrijk om voorafgaand aan het onderzoek al te beginnen met nadenken over hoe het probleem statistisch het beste aangepakt kan worden.

Wil je bijvoorbeeld een enquête houden, dan zal de manier waarop de vragen geformuleerd worden, bepalen of achteraf bepaalde hypothesen getoetst kunnen worden. De vraag is bijvoorbeeld of je aan de hand van steekproefdata kan aantonen dat de ene groep (statistisch significant) anders reageert dan de andere. Het is dus van belang dat je bij de opzet van de vragenlijst rekening houdt met de beperkingen van de statistische mogelijkheden van de geformuleerde vragen. Anders kun je achteraf niet meer voor het onderzoek belangrijke hypothesen toetsen. Het is daarom belangrijk om de onderzoeksstrategie voorafgaand aan het onderzoek met collega's, enquêteurs en statistici te bespreken.

2. Opzet enquête

2.1 Inleiding

Uit de inventarisatie die ten behoeve van de opfriscursus statistiek onder LEI-medewerkers is uitgevoerd, bleek dat 70% van de (73) responderende onderzoekers ervaring heeft met het opzetten van enquêtes. Daarnaast vormen enquêtes veelal het startpunt van veel kwantitatief onderzoek. Het Informatienet kan ook opgevat worden als een enquête. De manier waarop enquêtes worden uitgevoerd is dan ook bepalend voor het verdere verloop van het onderzoek. Het belang van een goede opzet van enquêtes is dan ook groot. Om deze reden wordt de opzet van enquêtes als eerste onderwerp in de opfriscursus behandeld.

In dit hoofdstuk wordt aandacht besteed aan de opzet van een onderzoek (paragraaf 2.2). Vervolgens wordt in paragraaf 2.3 aandacht besteed aan steekproefprocedures. Paragraaf 2.4 besteedt aandacht aan fouten in steekproefonderzoek en in paragraaf 2.5 komt de representativiteit aan de orde. In paragraaf 2.6 wordt een aantal voorbeelden gegeven van do's en don'ts met betrekking tot het stellen van enquêtevragen. In paragraaf 2.7 wordt afgesloten met literatuurverwijzingen die horen bij de opzet van enquêtes.

2.2 Opzet van een onderzoek

In deze paragraaf zal worden ingegaan op de opzet van een onderzoek. Hierbij spelen de volgende onderdelen een rol: onderzoeksprobleem en de onderzoeksvragen, de keuze voor primaire versus secundaire data en de keuze tussen kwantitatief en kwalitatief onderzoek.

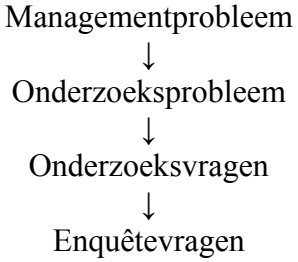
2.2.1 Onderzoeksprobleem en onderzoeksvragen

Allereerst zijn het onderzoeksprobleem en de onderzoeksvragen van een onderzoek van belang. De vraag is hoe deze tot stand komen en hoe deze met elkaar samenhangen. Allereerst dient opgemerkt te worden dat er in dit stuk geen gebruikgemaakt wordt van termen als probleemstelling, doelstelling en vraagstelling omdat deze in de literatuur vaak door elkaar gebruikt worden. Een omschrijving van dé doelstelling in het ene boek kan precies gelijk zijn aan de beschrijving van dé probleemstelling in een ander boek, en omgekeerd. Om deze reden wordt gebruikgemaakt van de termen managementprobleem, onderzoeksprobleem, onderzoeksvragen en enquêtevragen. Onderzoek wordt veelal uitgevoerd als gevolg van een management- of een beleidsprobleem van een ministerie of het bedrijfsleven.

Het door een ministerie of het bedrijfsleven geformuleerde managementprobleem kan echter zeer vaag zijn of moeilijk te onderzoeken. Dit probleem dient dan ook in veel gevallen geconcretiseerd te worden zodat een onderzoeksprobleem ontstaat dat een duidelijke afbakening geeft van wat in het onderzoek wordt gedaan. Deze herformulering van het probleem dient in onderling overleg tussen de opdrachtgever en de onderzoeker tot stand te komen. Het

onderzoeksprobleem dient niet te breed maar ook niet te nauw gedefinieerd te worden. Is deze te breed, dan is het niet sturend voor het onderzoek en is deze te nauw, dan krijgt de opdrachtgever geen antwoord op zijn/haar vraag. Is het onderzoeksprobleem duidelijk geformuleerd en afgebakend dan worden de onderzoeksvragen geformuleerd. Deze dienen zodanig geformuleerd te worden dat een antwoord op de onderzoeksvragen ook een oplossing biedt voor het geschetste onderzoeksprobleem.

De stap van onderzoeksvragen naar enquêtevragen wordt alleen gemaakt als een enquête ook werkelijk bij kan dragen aan het beantwoorden van de onderzoeksvragen. En ook het juiste middel hiervoor is. De bovenstaande stappen worden gevisualiseerd in figuur 2.1.



Figuur 2.1 Van managementprobleem tot enquêtevragen

Activiteiten die kunnen worden uitgevoerd om de bovenstaande stappen uit te voeren zijn: interviews, discussies, deskresearch en kwalitatief onderzoek. Om bijvoorbeeld het managementprobleem te kunnen concretiseren tot een onderzoeksprobleem is het noodzakelijk dat er met de opdrachtgever gediscussieerd wordt. Ook is het altijd goed om voordat je zelf aan een onderzoek begint eens te kijken wat er al op dit gebied is gebeurd. Dit kan door middel van deskresearch. Van expertise van anderen kan ook goed gebruikgemaakt worden door middel van interviews om onder andere de onderzoeksvragen of enquêtevragen te definiëren. Verder is het van groot belang dat opgestelde enquêtevragen eerst getest worden voordat zij aan respondenten worden voorgelegd. Als de vragen getest zijn op een kleine groep mensen of mede-onderzoekers kunnen fouten voorkomen worden.

2.2.2 Primaire versus secundaire informatie

Voor het beantwoorden van de onderzoeksvragen, dient eerst gekeken te worden of er gebruikgemaakt kan worden van secundaire informatie. Dat wil zeggen: 'Kunnen we gebruikmaken van de onderzoeksresultaten van een ander onderzoek?' Als dit het geval is dan hoeven we zelf geen informatie te verzamelen. Secundaire informatie is informatie die elders al beschikbaar is en in eerste instantie ten behoeve van een ander onderzoek verzameld is. Als er geen secundaire informatiebron aanwezig is dan biedt primaire informatie wellicht een oplossing. Primaire informatie wordt specifiek verzameld voor het beantwoorden van de huidige onderzoeksvragen. Dit kan bijvoorbeeld door binnen het onderzoek een enquête uit te zetten. Het gebruik van secundaire informatie brengt voor- en nadelen met zich mee. Een nadeel van secundaire informatie is dat de informatie niet altijd precies aansluit op de informatiebehoefte. Voordelen van secundaire informatie zijn dat de gegevens snel beschikbaar zijn en dat de kos-

ten die besteed worden aan het verzamelen van informatie relatief laag zijn. Ook hoeft er relatief weinig tijd besteed te worden aan het verzamelen van secundaire gegevens ten opzichte van primaire gegevens. In de volgende figuur worden de voor- en nadelen van primaire en secundaire informatie schematisch weergegeven.

	Primaire informatie	Secundaire informatie
Doel verzameling	Huidige probleem	Andere problemen
Verzameling	Bewerkelijk	Snel en eenvoudig
Kosten	Hoog	Laag
Tijd	Lang	Kort

Figuur 2.2 Primiare versus secundaire informatie

2.2.3 Kwalitatieve versus kwantitatieve insteek

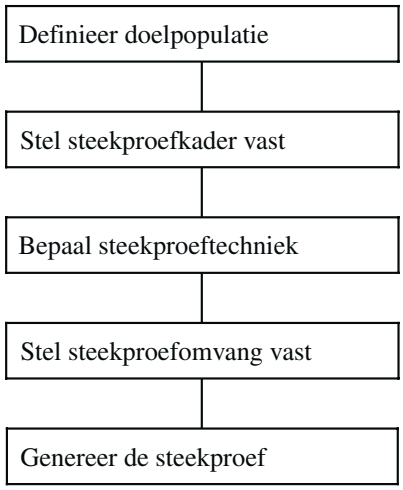
Vervolgens dient de afweging gemaakt te worden of het onderzoek een kwalitatieve of een kwantitatieve insteek verdient. Kwalitatief onderzoek wordt vaak gebruikt voor het verkrijgen van een eerste inzicht op het onderzoeksgebied. 'Wat speelt er zoal', terwijl kwantitatief onderzoek gebruikt kan worden om aan de hand van statistisch onderzoek hypothesen te testen en conclusies te trekken. Een ander belangrijk verschil tussen kwantitatief en kwalitatief onderzoek is het aantal respondenten dat wordt benaderd. Als het gaat om een kwalitatief onderzoek kunnen aan weinig respondenten dieptevragen worden gesteld. Ook open vragen spelen een grote rol in kwalitatief onderzoek. In dit soort onderzoek kunnen open vragen ook geanalyseerd worden. Voor kwantitatief onderzoek dient de datacollectie veel meer vastomlijnd en gestructureerd te zijn. In tabel 2.2 worden de kenmerken van beide onderzoeksmethoden geschetst.

	Kwalitatief	Kwantitatief
Doel	Verkrijgen kwalitatief inzicht in onderliggende redenen en motieven	Kwanitficeren van data en generaliseren van de uitkomsten naar de hele populatie
Steekproef	Klein aantal niet representatieve cases	Groot aantal cases
Datacollectie	Ongestructureerd	Gestructureerd
Data-analyse	Niet statistisch	Statistisch
Uitkomsten	Ontwikkelen van een initieel begrip van het onderwerp	Trekken van conclusies en doen van aanbevelingen

Figuur 2.3 Kwalitatief versus kwantitatief onderzoek

2.3 Steekproefprocedure

In de voorgaande paragraaf is uitgelegd hoe een onderzoek opgezet dient te worden en welke keuzes er gemaakt kunnen worden tussen verschillende onderzoeksvormen. In deze paragraaf wordt dieper ingegaan op steekproefprocedures. Hierbij gaat het om steekproeftechnieken, de steekproefomvang, fouten en problemen die zich voordoen bij steekproeven (bijvoorbeeld non-respons) en representativiteit. De procedure voor het opstellen van een steekproef is als volgt:



Figuur 2.4 Procedure opstellen steekproef

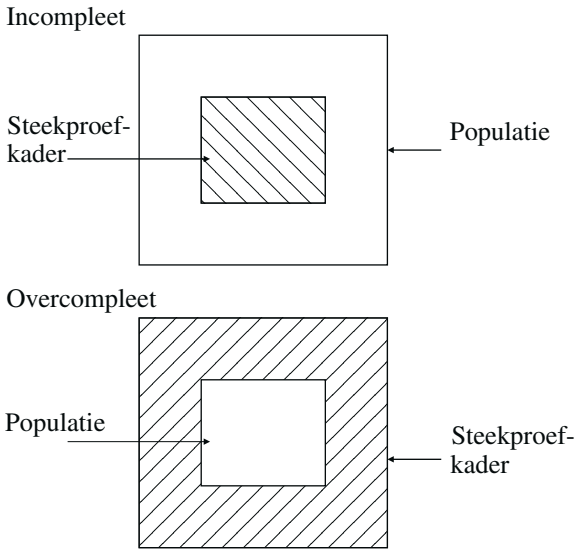
2.3.1 Definieer doelpopulatie

In eerste instantie dient de onderzoeker een doelpopulatie te kiezen. Wil de onderzoeker iets weten over vrouwen die zich in Nederlandse bejaardenhuizen bevinden, of gaat het juist om mannen die in 1999 in Zuid-Amerika schilderactiviteiten hebben uitgevoerd? De doelpopulatie moet zo duidelijk mogelijk gedefinieerd worden. De populatie alle Nederlanders is bijvoorbeeld zeer onduidelijk. Zijn dit alle mensen die op 3 januari 2001 een Nederlands paspoort hadden, of alle mensen die zich op 3 april 2001 in Nederland bevonden, of is dit nog anders bedoeld?

2.3.2 Steekproefkader vaststellen

Als duidelijk is van welke doelpopulatie het onderzoek uitgaat, dient het steekproefkader vastgesteld te worden. Het steekproefkader dient zo goed mogelijk aan te sluiten op de doelpopulatie. Als bijvoorbeeld de doelpopulatie bestaat uit alle mensen die op 1 januari 2002 de Nederlandse nationaliteit hadden en het telefoonboek van 2002 als steekproefkader dient, dan sluiten beiden niet precies op elkaar aan. Mensen die niet in Nederland wonen en mensen die geen vaste telefoonaansluiting hebben worden bijvoorbeeld uitgesloten. Terwijl mensen

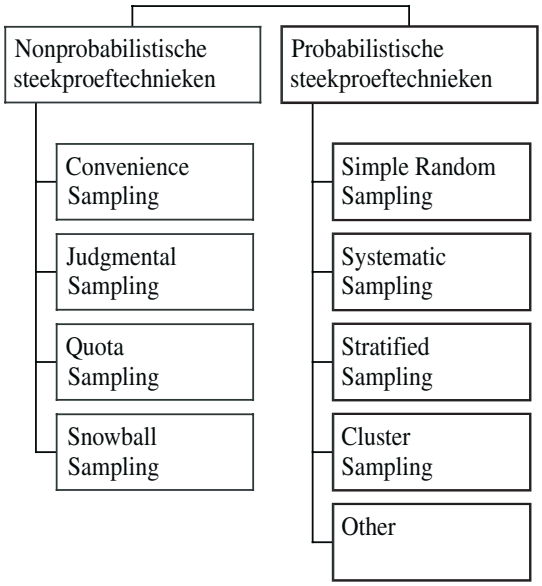
die niet de Nederlandse nationaliteit hebben en wel een vaste telefoonaansluiting kunnen worden meegenomen, terwijl zij niet tot de doelpopulatie behoren. Ook wanneer de onderzoeker op 3 januari 2002 om 10.00 uur 's ochtends voor de Albert Heijn interviews houdt, zal het steekproefkader niet geheel aansluiten bij de gedefinieerde doelpopulatie. Want welke mensen doen er boodschappen bij de Albert Heijn op 3 januari 2002 om 10 uur 's ochtends? Het steekproefkader kan zowel incompleet als overcompleet zijn. In bovenstaande voorbeelden was het steekproefkader incompleet, omdat de doelpopulatie breder gedefinieerd was dan het onderzoekskader. Een steekproefkader kan ook overcompleet zijn. In dat geval zitten er elementen in het steekproefkader die niet in de doelpopulatie gewenst zijn. In de volgende figuur wordt dit gevisualiseerd.



Figuur 2.5 Steekproefkader ten opzichte van de doelpopulatie

2.3.3 Steekproeftechnieken vaststellen

Nadat het steekproefkader is vastgesteld, dient de steekproeftechniek bepaald te worden. Hierbij dient gekozen te worden tussen non-probabilistische en probabilistische steekproeftechnieken. Dat zijn respectievelijk steekproeftechnieken waarbij niet geheel toevallig getrokken wordt en steekproeftechnieken waarbij dit wel het geval is. Bij de eerste groep wordt de keuze van de steekproefelementen niet random bepaald en is de keuze grotendeels afhankelijk van de onderzoeker. Bij de tweede groep vindt random selectie plaats. In figuur 2.4 worden de verschillende steekproeftechnieken en hun onderlinge samenhang weergegeven.



Figuur 2.6 Steekproeftechnieken

Nonprobabilistische steekproeftechnieken

Als eerste worden de nonprobabilistische steekproeftechnieken besproken. Hieronder vallen convenienve sampling, judgmental sampling, quota sampling en snowball sampling. Ook combinaties van één of meerdere genoemde technieken zijn mogelijk.

Convenience sampling is een techniek van steekproeftrekken waarbij gemak voor de onderzoeker een grote rol speelt. De onderzoeker wil zo min mogelijk moeite stoppen in het verzamelen van de gegevens. Deze techniek kan dan ook geschikt zijn om een eerste indruk van een onderwerp te krijgen, maar representatief voor de populatie zal de steekproef niet zijn. Voorbeelden van convenience sampling zijn: het interviewen voor de winkel, het neerleggen van enquêtes bij de receptie van een hotel en ook enquêtes via internet zijn een vorm van convenience sampling.

Verder is *judgemental sampling* een voorbeeld van een nonprobabilistische steekproeftechniek. Dit is een methode die voornamelijk bij kwalitatief onderzoek wordt gehanteerd. Hierbij gaat een onderzoeker op zoek naar mensen waarvan hij/zij bij voorbaat weet dat zij kennis hebben over het desbetreffende onderwerp.

Daarnaast is *quota sampling* ook een steekproeftechniek waarbij niet geheel toevallig steekproefelementen worden getrokken. Bij quota sampling wordt de populatie opgedeeld in groepen aan de hand van kenmerken die voor het onderzoek van belang zijn. Van deze kenmerken wordt vervolgens nagegaan in welke percentages zij voorkomen in de populatie en vervolgens worden dezelfde percentages gehanteerd in de steekproef. Quota sampling kan ook in combinatie met convenience sampling worden uitgevoerd, als deze bijvoorbeeld plaatsvindt voor de winkel. Vooraf heb je bijvoorbeeld aan de hand van je populatie bepaald dat je 40 mannen in de leeftijdscategorie van 20 tot 30 jaar in de steekproef wilt en 20 in de categorie van 30 tot 40 jaar. Als je dan het volledige aantal van een bepaalde categorie hebt geïnter-

viewd, dan ga je alleen nog op zoek naar mensen die in de andere categorie vallen. Op deze manier wordt een soort van representativiteit bewerkstelligd.

Snowball sampling is een methode die vaak wordt toegepast als de doelpopulatie zeer specifiek is. In het voorbeeld van Zuid-Amerikaanse schilders zou deze techniek goed gebruikt kunnen worden. Je zou kunnen beginnen met het interviewen van mensen die winkelen bij een adventure winkel, of de lezers van de Op pad. Als je eenmaal iemand gevonden hebt die dit soort werk gedaan heeft, vraag je aan het eind van het interview of hij/zij nog iemand anders kent die ditzelfde werk gedaan heeft. Op deze manier komt de sneeuwbal van steekproefelementen op gang.

Probabilistische steekproeftechnieken

De probabilistische steekproeftechnieken zijn te verdelen in de technieken: simple random sampling, systematic sampling, stratified sampling, cluster sampling en combinaties van bovenstaande technieken.

Bij *simple random sampling* hebben alle elementen in het steekproefkader dezelfde kans om getrokken te worden. Vooraf wordt de steekproefgrootte vastgesteld. Een illustratie van simple random sampling is dat alle elementen uit het steekproefkader in één grote bak worden gegooid en er vervolgens zoveel elementen worden getrokken als vooraf bepaald.

Systematic sampling is een ander voorbeeld van een probabilistische steekproeftechniek. Bij deze techniek worden eerst alle elementen uit het steekproefkader aan de hand van een bepaald kenmerk gerangschikt, bijvoorbeeld van klein naar groot. Vervolgens wordt er afhankelijk van het aantal te kiezen elementen steeds systematisch een element gekozen. Dus bijvoorbeeld het 25e element, het 125e element, het 225e element enzovoorts.

Stratified sampling is een manier van steekproeftrekken waarbij de elementen uit het steekproefkader eerst ingedeeld worden in groepen aan de hand van kenmerken die voor het onderzoek van belang zijn. Deze groepen worden ook wel strata genoemd. Als bekend is hoeveel elementen uit het steekproefkader in elk van de strata vallen, kan bepaald worden hoeveel elementen er in de steekproef uit elk van de strata getrokken dienen te worden. Het aantal steekproefelementen per stratum kan vervolgens op verschillende manieren bepaald worden. Er is bijvoorbeeld een methode die uitgaat van proportionele aantallen. Dat wil zeggen proportioneel aan het aantal in het streekproefkader. Maar ook is het mogelijk een optimale verdeling vast te stellen, die rekening houdt met de variantie in elk van de strata. Formules voor de proportionele en de optimale verdeling van steekproefelementen over strata worden in paragraaf 3.2 gegeven. Als het aantal gewenste elementen per stratum is bepaald dan worden de steekproefelementen per stratum random getrokken.

De laatste probabilistische techniek is *cluster sampling*. Deze techniek gaat uit van het random trekken van hele groepen. Bijvoorbeeld hele schoolklassen, of hele straten.

2.3.4 Steekproefomvang vaststellen

Naast de steekproeftechniek dient ook de steekproefomvang bepaald te worden. Om deze vast te stellen kan een aantal criteria gebruikt worden zoals intuïtie, statische precisie, kostenbeperkingen en ervaringscijfers. Veelal geldt dat de kosten de beperkende factor zijn voor het aantal steekproefelementen. Hoe meer steekproefelementen, hoe duurder het verzamelen van data. De steekproefomvang kan ook worden vastgesteld aan de hand van intuïtie of ervarings-

cijfers. Een steekproef van 300 elementen geeft veelal een aardig beeld van het te onderzoeken concept. Verder geldt dat statistische technieken eisen stellen aan het aantal elementen. Uit de literatuur is een formule beschikbaar, aan de hand waarvan de steekproefomvang berekend kan worden bij een gegeven betrouwbaarheids-, en nauwkeurigheidsniveau. Een nadeel van deze formule is dat deze betrouwbaarheid en nauwkeurigheid alleen maar gelden voor het vaststellen van het gemiddelde van één van de steekproefvariabelen. Als er meerdere variabelen geschat moeten worden, wat meestal het geval is, is de formule minder eenduidig toepasbaar. Tenzij je een soort gemiddelde aantal vaststelt door de formule voor alle variabelen in te vullen en de steekproefgrootten die je vindt te middelen.

Een ander nadeel van het toepassen van deze formule is dat je van tevoren een idee moet hebben over de standaarddeviatie van je steekproefvariabele. Vaak is dit niet het geval. Dit is soms op te lossen door hier een schatting van te maken, bijvoorbeeld aan de hand van informatie uit eerdere jaren of andere steekproeven. Voor een continue variabele geldt de volgende formule:

$$e = z * \frac{s}{\sqrt{n}} \qquad \rightarrow \qquad n = \left(\frac{z}{e}\right)^2 * s^2$$

Voor een binomiale variabele (een variabele die maar twee waarden aan kan nemen) geldt de volgende formule:

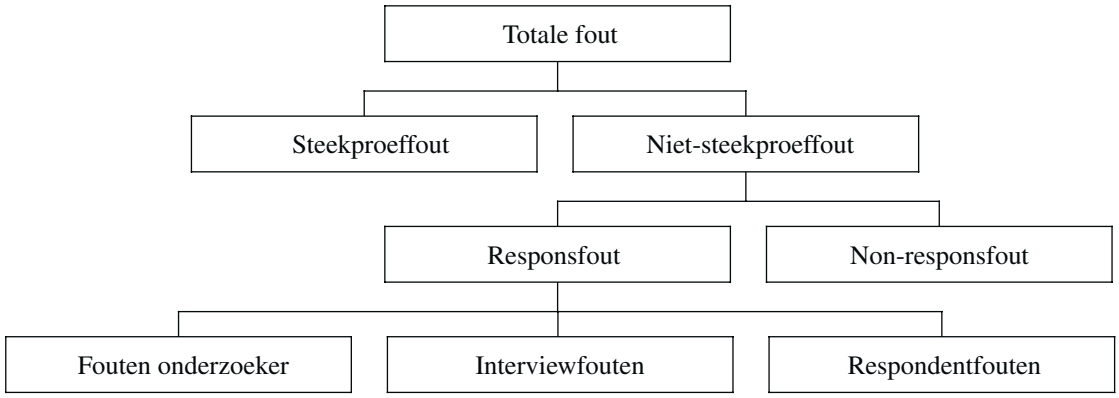
$$e = z * \sqrt{\frac{p * q}{n}} \qquad \rightarrow \qquad n = \left(\frac{z}{e}\right)^2 * p * q$$

n is het aantal steekproefelementen
e is de gewenste nauwkeurigheid
z is de betrouwbaarheid (bij een betrouwbaarheid van 99,7% hoort bijvoorbeeld een z-waarde van 3, bij een betrouwbaarheid van 95,4% een z-waarde van 2, zie tabel 5.1 voor de z-waarden)
s is de (geschatte) standaarddeviatie van de steekproefvariabele
p is de het percentage van de populatie met een bepaald kenmerk
q is (1-p)

2.4 Fouten in steekproefonderzoek

Bij het uitvoeren van steekproefonderzoek kan een aantal fouten optreden. Deze fouten kunnen verschillende oorzaken hebben. Hiervan worden de belangrijkste weergegeven in figuur 2.7.

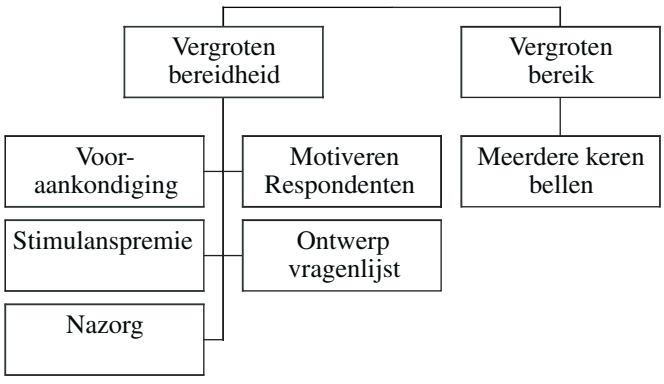
Uit figuur 2.7 blijkt dat de totale fout kan worden opgesplitst in de steekproeffout, ook wel de toevallige fout genoemd en de niet-steekproeffout. Steekproeffouten treden op doordat je niet de gehele populatie, maar slechts een deel van de populatie (de steekproef) bekijkt. De grootte van de steekproeffout is afhankelijk van de homogeniteit van de populatie en het aantal waarnemingen in de steekproef. In de regel geldt: hoe meer waarnemingen en hoe homogener de populatie hoe kleiner de steekproeffout.



Figuur 2.7 Fouten in steekproefonderzoek

Niet-steekproeffouten kunnen weer worden opgedeeld in responsfouten en non-responsfouten. Als we eerst kijken naar de responsfouten dan blijkt dat we deze weer op kunnen delen in fouten van de onderzoeker, interviewfouten en respondentfouten. Onder fouten van onderzoekers vallen fouten in de dataverwerking zoals typefouten. Onder interviewfouten vallen fouten die veroorzaakt worden door een miscommunicatie tussen de interviewer en de geïnterviewde. Hierbij kan sprake zijn van onduidelijke vraagstelling, maar ook kunnen antwoorden van de geïnterviewde verkeerd geïnterpreteerd worden door de interviewer. Respondentfouten kunnen plaatsvinden als de respondent foutieve of onware antwoorden geeft.

Non-responsfouten kunnen ontstaan doordat mensen niet mee willen werken aan enquêtes. Echter, als er sprake is van non-respons hoeft er niet altijd sprake te zijn van een non-responsfout. Er ontstaat bijvoorbeeld geen non-responsfout als responderende mensen niet systematisch afwijken van niet-responderende mensen. Als er echter onderzoek wordt gedaan naar het inkomen van boeren en alle rijke boeren weigeren mee te werken aan de enquête dan is er wel degelijk sprake van een non-responsfout. De enquête geeft dan namelijk een zeer vertekend beeld van de werkelijke situatie.



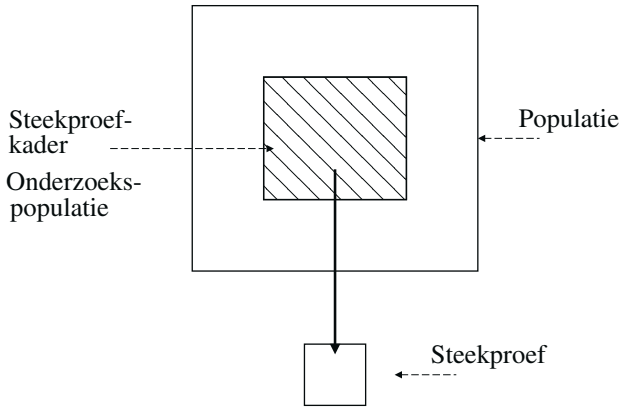
Figuur 2.8 Non-respons

Non-respons heeft zowel effect op de kwaliteit van het onderzoek als op de kosten van het onderzoek. Als bijvoorbeeld de non-respons ingeschat wordt op 50% dan dienen er twee maal zoveel mensen benaderd te worden als het gewenste aantal elementen in de steekproef.

Het bovenstaande schema geeft aan wat er gedaan kan worden om de non-respons zo beperkt mogelijk te houden. De onderzoeker kan zijn/haar bereik vergroten door bijvoorbeeld meerdere brieven te schrijven of meerdere keren te bellen. Maar ook kan de onderzoeker werken aan het vergroten van de bereidheid bij de mogelijke respondenten. Dit kan door de enquête van tevoren aan te kondigen en het belang van de enquête voor de respondent zelf aan te geven. Dit kan ook al motiverend werken. Daarnaast kan een premie uitgereikt worden aan mensen die mee willen werken aan de enquête om zodoende mensen te stimuleren. Een andere mogelijkheid is het bieden van nazorg, door bijvoorbeeld de resultaten van het onderzoek naar de respondent te sturen. Ook kan geprobeerd worden om bij het ontwerp van de vragenlijst rekening te houden met de respondenten. Als je bijvoorbeeld een schriftelijke enquête opzet voor kinderen tussen twaalf en vijftien jaar, dan is een zakelijke opmaak niet aantrekkelijk en zullen de kinderen niet zo snel geneigd zijn de enquête in te vullen.

2.5 Representativiteit

De representativiteit van de steekproef geeft de mate aan waarin de steekproef een goede afspiegeling is van de onderzoekspopulatie. Dat wil zeggen: in hoeverre zijn er significante verschillen tussen de steekproef en de onderzoekspopulatie? Het volgende plaatje geeft de samenhang aan tussen de steekproef, de populatie en het steekproefkader.



Figuur 2.9 Representativiteit

Een algehele uitspraak over de representativiteit van een steekproef is weinig zinvol. Een steekproef zal nooit representatief zijn voor alle mogelijke variabelen. Het verdient daarom aanbeveling om in het kader van een specifiek onderzoek te toetsen in hoeverre de steekproef voor de in dat onderzoek relevante doelvariabelen representatief is. Het is bijvoor-

beeld wel mogelijk na te gaan in hoeverre de steekproef representatief is ten aanzien van het aspect leeftijd.

2.6 Do's en don'ts

In deze paragraaf een paar do's en don'ts wat betreft het stellen van enquêtevragen.

Vragen (do's)
1. Gebruik zoveel mogelijk simpele taal;
2. Maak gebruik van bekend en duidelijk vocabulair;
3. Stel korte vragen;
4. Gebruik eenduidige woorden;
5. Stel specifieke vragen.

   'Heeft u in de afgelopen maand…'

Vragen (don'ts)
1. Stel geen dubbele vragen
   'Ken je het merk en denk je dat het…'
2. Stel geen sturende vragen
   'Denk je dat een goede burger…'
3. Stel geen biased vragen. De onderstaande categorieën sturen de respondent bijvoorbeeld in de positieve richting.
   'Slecht - Neutraal - Goed - Zeer goed'
4. Maak geen gebruik van impliciete alternatieven
   'Koop je graag geïmporteerde goederen?'
5. Maak geen gebruik van generalisaties

2.7 Literatuur

Voor het opzetten van enquêtes wordt de volgende literatuur aanbevolen:
- Cochran, W.G. (1977);
- Green, P.E., D.S. Tull en G. Albaum (1988);
- Malhotra, N.K. (1993);
- Zwart, P.S. (1993).

Van deze literatuur wordt Cochran (1977) aangeraden wat betreft het technische deel van het trekken van steekproeven. Hierin staan alle relevante methodieken en de uitleg en afleiding hiervan. Voor het meer marketinggerichte aspect, dus hoe stel je vragen en hoe bouw je enquêtes op zijn Green (1988) en Malhotra (1993) aan te raden. Zwart (1993) is zeer makkelijk te lezen. In dit boek wordt veel informatie gegeven over marketinggerichte aspecten van enquêtes, maar het hoe en waarom van onder andere formules ontbreekt.

3. Opzet Informatienet

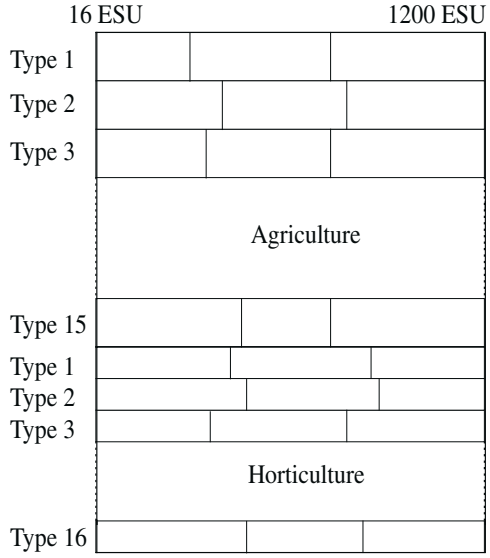
3.1 Inleiding

In het voorgaande hoofdstuk is de opzet van enquêtes in het algemeen gesproken. In dit hoofdstuk wordt het Informatienet (voormalig Bedrijven-Informatienet van het LEI (het Informatienet) besproken. Het Informatienet is een enquête die door het LEI is opgezet en waaraan jaarlijks ongeveer 1.500 primaire agrarische bedrijven meedoen. De informatie uit het Informatienet wordt zowel binnen als buiten het LEI breed gebruikt. In dit hoofdstuk wordt allereerst een globale beschrijving gegeven van het Informatienet, vervolgens worden de kwaliteitsaspecten van de steekproef besproken.

3.2 Globale beschrijving steekproef Informatienet

Bij de globale beschrijving van het Informatienet komen de inrichting van het Informatienet, het voordeel van stratificatie, de toewijzing van capaciteit aan strata, de random keuze van bedrijven en de weging aan de orde.

Inrichting Informatienet
Bij de bedrijfsselectie worden agrarische bedrijven ingedeeld naar bedrijfstypering en naar bedrijfsgrootte. De bedrijfsgrootte wordt uitgedrukt in Europese Grootte Eenheden (EGE's), ook



Figuur 3.1 Inrichting informatienet

wel European Standard Units (ESU's) genoemd. De EGE kan rechtstreeks afgeleid worden uit de Nederlandse Grootte Eenheid (NGE), 16 EGE staat gelijk aan 18,7 NGE. Figuur 3.1 geeft in grote lijnen aan hoe het Informatienet is ingericht.

Binnen elk bedrijfstype worden drie grootteklassen onderscheiden. De ondergrens van de eerste klasse ligt voor alle bedrijfstypen bij 16 EGE en de bovengrens van de grootste klasse ligt voor alle bedrijfstypen bij 1.200 EGE. Verder variëren de klassegrenzen van de grootteklassen. Reden hiervan is dat niet elk bedrijfstype even homogeen is en ook varieert de gemiddelde bedrijfsgrootte over de verschillende typen. De klassegrenzen zijn dan ook zo optimaal mogelijk vastgesteld zodat de variantie tussen elementen in één klasse zo klein mogelijk is, dus de groepen zo homogeen mogelijk zijn.

Stratificatie

De bovenstaande indeling in klassen wordt ook wel stratificatie genoemd, waarbij de verschillende klassen de strata zijn. In het Informatienet is gekozen voor een gestratificeerde steekproef omdat dit nogal wat voordelen biedt. Deze voordelen zijn:

- een grotere betrouwbaarheid van schattingen;
- een efficiëntere aanwending van steekproefeenheden;
- makkelijker om parameters uit individuele strata te schatten;
- de vervangende keus bij non-respons is eenvoudiger;
- de vervangende keus bij uitval is eenvoudigere;
- er is sprake van een grotere representativiteit door rekening te houden met gerealiseerde trekkingskansen.

Het eerste voordeel, de grotere betrouwbaarheid bij schattingen, wordt gerealiseerd doordat de variantie van de gestratificeerde schatter veelal kleiner is dan die van de directe schatter. Dit geldt als de variantie binnen strata klein is vergeleken met de variantie tussen strata. Dit geldt met name voor de stratificatievariabele(n) zelf en voor variabele(n) die sterk samenhangen met de stratificatievariabele(n).

Het tweede voordeel hangt deels samen met het eerste voordeel. Het efficiënter aanwenden van steekproefeenheden wordt namelijk gerealiseerd doordat je met minder steekproefeenheden dezelfde betrouwbaarheid bereikt. Verder kan je de steekproefcapaciteit aanwenden daar waar de steekproeffout relatief groot is.

Een derde voordeel is het vergemakkelijken van individuele strataschattingen. Je bent er zeker van dat er eenheden in elk stratum zitten, dus kan je altijd per stratum een schatting maken.

De vervangende keus bij non-respons is een vierde voordeel. Als er sprake is van non-respons leidt dit bij een gestratificeerde steekproef niet direct tot vertekeningen in de steekproef ten opzichte van de populatie. Van tevoren is bekend hoeveel elementen je in elk stratum wilt trekken. Bij non-respons in een bepaald stratum benader je dan ook een extra element in dat stratum. Als de non-respons in een bepaald stratum heel groot is dan ben je toch verzekerd dat je elementen uit dat stratum in je steekproef hebt. Dit is niet het geval bij een simpele random trekkingsmethode. Dit geldt overigens alleen voor de variabelen waarnaar gestratificeerd wordt. Op het gebied van andere variabelen kan nog steeds vertekening optreden.

Een vijfde voordeel van stratificatie is de vervangende keus bij uitval van elementen uit de steekproef. Bij het Informatienet vallen er bijvoorbeeld jaarlijks bedrijven uit omdat ze niet

meer mee willen werken. In dat geval is direct duidelijk uit welk stratum een vervangend bedrijf moet worden geselecteerd.

Het zesde en laatste voordeel is dat er sprake is van een grotere representativiteit van de steekproef. Reden hiervoor is dat er rekening gehouden wordt met gerealiseerde trekkingskansen. Met behulp van de weging kan je corrigeren voor het onder- of oververtegenwoordigd zijn van elementen in bepaalde strata. Deze grotere representativiteit geldt overigens alleen ten aanzien van de stratificatievariabelen.

Stratificatie in het Informatienet

In de nieuwe opzet van het Informatienet is ervoor gekozen om naar 2 variabelen te stratificeren. Dit zijn de bedrijfsgrootte (uitgedrukt in EGE's) en de bedrijfstypering. In het verleden werd naar vijf variabelen gestratificeerd, namelijk naast de twee genoemde variabelen ook naar leeftijd van de ondernemer, regio en bedrijfsomvang. Naast de voordelen van stratificatie heeft een verregaande stratificatie ook een nadeel, namelijk dat de toepasbaarheid vermindert. Aangezien het Informatienet steeds breder gebruikt wordt binnen het LEI, weegt dit nadeel zwaar en heeft stratificatie naar minder variabelen voordelen.

Toewijzing van steekproefbedrijven aan strata

Het bepalen van het aantal steekproefbedrijven per stratum kan op basis van verschillende methoden. De volgende factoren kunnen een rol spelen:

- relatieve economische belang van een groep (% aantal);
- relatieve belang van een groep (% NGE);
- belang van een agrarische activiteit;
- precisie van de schattingen.

Bij de eerste en tweede factor wordt het aantal steekproefbedrijven per stratum bepaald aan de hand van het relatieve belang van een groep. Bij de eerste factor wordt het belang van een groep vastgesteld aan de hand van het aantal elementen in het steekproefkader. Bij de tweede factor wordt het relatieve belang afgeleid uit de grootte van bedrijven. Dit is het aantal NGE's.

Een andere factor is het belang voor onderzoek of beleid dat aan bepaalde agrarische activiteiten wordt gehecht. Het aantal steekproefelementen per stratum hoeft niet altijd afgeleid te worden van harde getallen. Er kan bijvoorbeeld besloten worden om relatief veel veel melkveebedrijven op te nemen omdat daar veel onderzoek naar gedaan wordt.

Een laatste factor is de precisie van de schattingen. De precisie wordt ook wel de nauwkeurigheid van de schattingen genoemd. Voor de relatie tussen de nauwkeurigheid, de betrouwbaarheid en het aantal steekproefelementen wordt naar paragraaf 2.3.4 verwezen.

Op basis van deze criteria is een verdeling over de bedrijfstypen tot stand gekomen. De verdeling over de klassen binnen een type vindt door middel van optimale allocatie plaats.

In de statistiek wordt onderscheid gemaakt tussen proportionele en optimale verdeling van steekproefelementen over de strata.

De formule voor de optimale verdeling is als volgt:

$$n_h = \frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i} * n$$

De formule voor een proportionele allocatie is:

$$n_h = N_h * \frac{n}{N}$$

n_h is het aantal elementen in stratum h
N_h is het aantal elementen in stratum h in de populatie
n is het totaal aantal steekproef elementen
N is het totaal aantal elementen in de populatie
S_h is de standaarddeviatie in stratum h

Keuze van bedrijven
Het Informatienet is een roterende panel. Een paneldataset is een dataset waarin meerdere elementen over een langere periode gevolgd worden. In het Informatienet is dit het geval, want jaarlijks worden gegevens verzameld van ongeveer 1.500 bedrijven. Deze bedrijven mogen echter in principe niet langer dan 7 jaar deelnemen. Dit maakt dat het gaat om een roterend panel. Een deel van de bedrijven in het Informatienet roteert jaarlijks. Naast het feit dat bedrijven niet langer dan een bepaalde termijn deel mogen nemen, kunnen zij ook afvallen om andere redenen. Zij kunnen bijvoorbeeld zelf aangeven dat zij niet meer mee willen werken. Om deze reden dienen jaarlijks nieuwe deelnemers geworven te worden. Dit gebeurt volgens de regel die zijn weergegeven in figuur 3.2.

Aantal te werven bedrijven
=
Aantal gewenste bedrijven
-
Aantal bedrijven in administratie

Figuur 3.2 Random keuze van bedrijven

De keuze van bedrijven is random, waarbij rekening gehouden wordt met de aantallen bedrijven die per stratum gekozen dienen te worden. Het steekproefkader van het Informatienet is de Landbouwtelling. Hierop wordt in paragraaf 3.3 verder ingegaan. Bij de keuze wordt rekening gehouden met de non-respons. Dit betekent voor het Informatienet dat veelal vier maal zoveel bedrijven worden aangeschreven als werkelijk in de steekproef nodig zijn.

Weging

De weging die in het Informatienet beschikbaar is, is als het ware een weging om een schatting van De Nationale Boerderij te kunnen maken. Bij het vaststellen van de wegingsfactoren wordt gebruikgemaakt van stratificatie-informatie. De wegingsfactor per stratum is het aantal populatie-elementen in stratum h gedeeld door het aantal steekproefelementen in stratum h.

$$W_h = \frac{N_h}{n_h}$$

De wegingsfactoren kunnen gebruikt worden om schattingen te maken die de hele populatie betreffen. Deze schattingen worden echter wel gemaakt aan de hand van de steekproefgegevens. Dit wordt ook wel het ophogen van de steekproefresultaten naar de populatie genoemd. De schattingen per stratum in de steekproef worden dan vermenigvuldigd met de wegingsfactoren per stratum en deze worden gesommeerd over alle strata.

$$\bar{y} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{n_h} = \sum_{h=1}^{L} W_h \bar{y}_h$$

De som van de gewichten is een schatting van het aantal bedrijven in de populatie. Onderzoekers houden in veel gevallen geen rekening met de onzekerheid van deze schatting. Een voorbeeld waarin gebruikgemaakt wordt van de wegingsfactoren is als volgt:

Volgens de Landbouwtelling

N	type a	type b
600	250	350

Volgens de steekproef

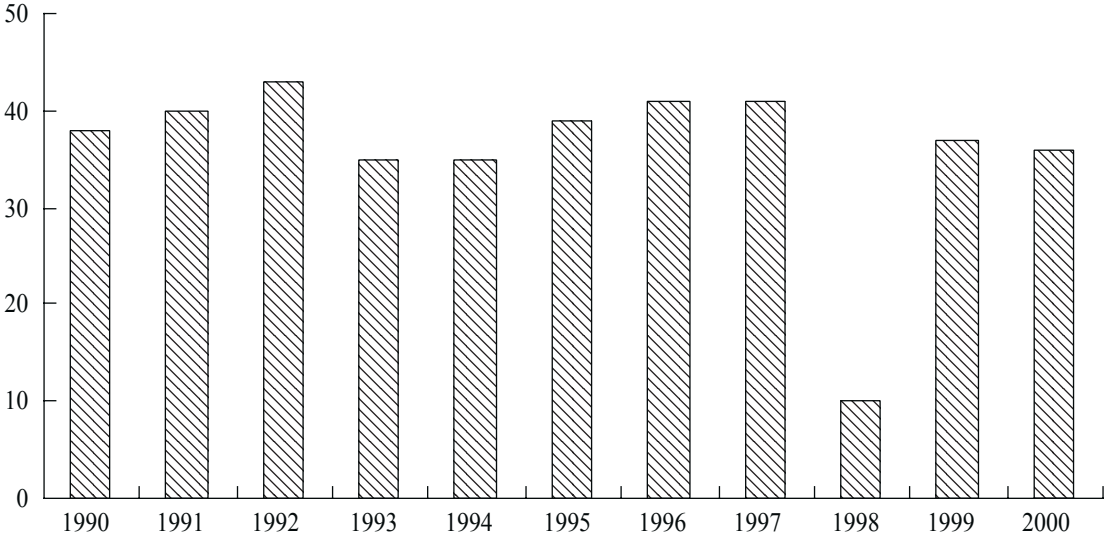
n	type a	type b
12	6	6

De kans (p) op type a is 0,50 als je naar de steekproef kijkt. Op basis van deze kans zou je veronderstellen dat van de 600 elementen in de populatie er 300 van type a zouden zijn. Anders gezegd het gewicht van de bedrijven in de steekproef is 600/12 = 50. De som van de gewichten van de bedrijven in de steekproef van type a is 300. Dan lijkt het alsof je er ver naast zit. Als er nu rekening gehouden wordt met de variantie en de betrouwbaarheid waarbij: var (p) = 0,02 en standaardfout(p) = 0,15. Dan kan je een betrouwbaarheidsinterval afleiden voor p en daaruit blijkt dat p met een betrouwbaarheid van ongeveer 95% in het interval [0,20 - 0,80] ligt. De aantallen van type a die hierbij horen zijn [121 - 479]. Hieruit blijkt dat het betrouwbaarheidsinterval zeer groot is met zo weinig elementen in de steekproef en dat de schatting van 300 niet significant afwijkt van de werkelijke waarde.

3.3 Kwaliteit van de steekproef

De kwaliteit van de steekproef kan worden beoordeeld aan de hand van begrippen zoals: de non-respons, de dekking, de representativiteit en de betrouwbaarheid. In deze paragraaf zullen de verschillende kwaliteitsaspecten aan de orde komen.

Non-respons
In het Informatienet is de non-respons een grote zorg. Deze schommelt namelijk rond de 50 à 70%. Dit is een vrij hoog percentage, maar heeft te maken met de omvang van de gegevensverzameling. Die is namelijk zeer divers. Er worden gegevens verzameld uiteenlopend van mineralenmanagement tot financiën en van duurzame goederen tot aantal dieren. Het kost boeren dan ook tijd en moeite om mee te werken aan het Informatienet. Daarom wordt hen een bedrijfsverslag in het vooruitzicht gesteld. Zij krijgen een financieel verslag, waarin hun bedrijf wordt vergeleken met andere vergelijkbare bedrijven. Aan de hand van deze verslagen kunnen boeren hun marktpositie inschatten en ook eventueel hun management aanpassen. Ook wordt in het bedrijfsverslag informatie over niet financiële zaken gegeven. Ook al wordt boeren een bedrijfsverslag in het vooruitzicht gesteld, toch willen zij vaak niet meewerken. In figuur 3.3 is te zien hoe de non-respons zich gedurende de afgelopen jaren heeft ontwikkeld.



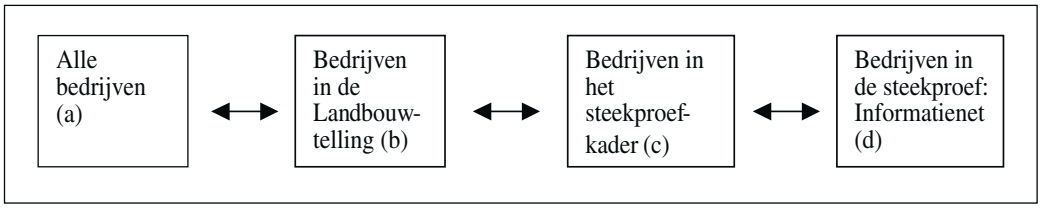
Figuur 3.3 Respons van het Informatienet

Hieruit blijkt dat de respons normaal gesproken 30 tot 50% bedraagt. 1998 was echter een uitzonderlijk jaar. Dit slechte resultaat werd veroorzaakt doordat er bedrijven gekozen dienden te worden uit groepen die meestal niet mee wilden werken, versterkt door de problemen in de varkenshouderij.

Non-respons is een probleem als responderende bedrijven systematisch afwijken van

niet-responderende bedrijven. Door middel van stratificatie kan je al een deel van de non-respons voor belangrijke variabelen opvangen, maar het effect voor variabelen die geen grote correlatie vertonen met de stratificatievariabele is niet te ondervangen en dient onderzocht te worden voordat de onderzoeker aan een onderzoek begint. Een onderzoek naar effecten van non-respons in het Informatienet staat voor 2003 op het programma.

Dekking
De dekking geeft aan in hoeverre de populatie wordt gedekt door de steekproef. Dus voor het Informatienet geeft de dekking aan in hoeverre alle primaire agrarische bedrijven die in Nederland gevestigd zijn kans hebben terug te komen in de steekproef. Figuur 3.4 geeft aan hoe de elementen in de steekproef zich verhouden tot de gehele populatie.



Figuur 3.4 Dekking Informatienet

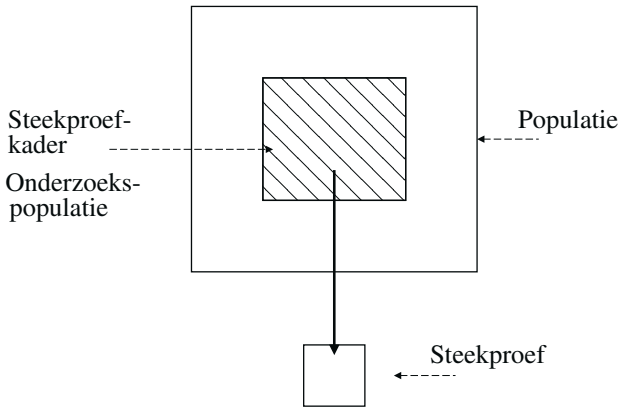
De populatie omvat alle primaire agrarische bedrijven die in Nederland gevestigd zijn (a). Uit deze populatie zijn alle bedrijven die groter zijn dan 3 NGE geselecteerd en weergegeven in de Landbouwtelling (b). Bedrijven die te klein zijn worden niet weergegeven in de Landbouwtelling, omdat zij puur hobbymatig boeren en niet bedrijfsmatig. Uit de Landbouwtelling zijn de bedrijven groter dan 16 EGE en kleiner dan 1.200 EGE gefilterd. De overgebleven bedrijven vormen het steekproefkader van het Informatienet (c). Uit het steekproefkader worden (gegeven de stratificatie en gegeven de bedrijven die overgebleven zijn van het voorgaande jaar) random bedrijven gekozen.

Representativiteit
De representativiteit van de steekproef geeft de mate weer waarin de steekproef belangrijke kenmerken van de populatie weerspiegelt. Om de representativiteit van de steekproef te onderzoeken kunnen kenmerken van de steekproef vergeleken worden met kenmerken van de populatie. In de jaarlijkse rapportage over het Informatienet (De steekproef voor het Bedrijven-Informatienet van het LEI) wordt dit ook gedaan voor een kleine honderd variabelen. In tabel 3.1 wordt een voorbeeld gegeven voor een klein aantal variabelen. Hierbij staat FADN voor het Farm Accountancy Data Network (het Europese Informatienet). Het Nederlandse Informatienet is hier een onderdeel van. FSS staat voor Farm Structure Survey. Dit is te vergelijken met de Nederlandse Landbouwtelling. %SE staat voor de relatieve standaardfout. Dit is de standaardfout van de FADN-schatting gedeeld door de FADN-schatting zelf.

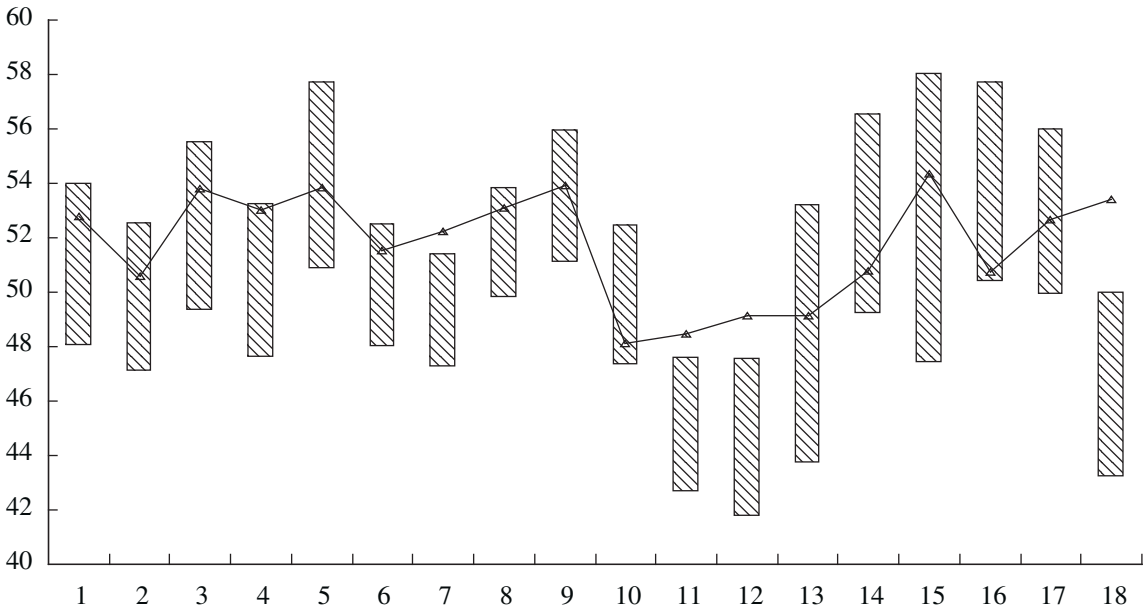
Tabel 3.1 FADN-schatting versus FSS

Variable	FSS	FADN	%SE
Size of farm			
NGE	93,69	96,22	0,77
Size of activities			
Arable	34,95	37,99	1,81
Horticulture (open air)	40,26	34,3	3,79
Horticulture (glass)	57,81	57,86	2,29
Dairy cows	60,3	64,53	1,18
Pigs	2,88	3,97	12,1
Size (hectares)			
total cultivated area	22,94	23,99	1,19
Arable	9,52	10,33	2,02

Representativiteit wordt met betrekking tot het Informatienet nog wel eens foutief gebruikt. Representativiteit moet echter altijd ten aanzien van een bepaalde variabele bekeken worden en is dus afhankelijk van het onderzoek en de variabelen die in dat onderzoek voorkomen. Ga dus altijd na voordat je aan een onderzoek begint of het Informatienet wel representatief is voor jouw onderzoek. Voor de toetsing van de representativiteit moeten de kenmerken van de steekproef vergeleken worden met de kenmerken van de onderzoekspopulatie (zie figuur 3.5).



Figuur 3.5 Representativiteit



Figuur 3.6 Leeftijd van de ondernemer

De bovenstaande figuur geeft de representativiteit van het Informatienet weer ten aanzien van de leeftijd van ondernemers. Op de horizontale as zijn verschillende bedrijfstypes weergegeven en op de verticale as is de leeftijd van ondernemers weergegeven. De driehoekjes geven de gemiddelde leeftijd volgens de Landbouwtelling weer. De balken geven het 95% - betrouwbaarheidsinterval van de schatting op basis van het Informatienet weer. Als de driehoekjes voor een bepaald bedrijfstype niet binnen de balken vallen dan kan je zeggen dat het Informatienet niet representatief is ten aanzien van de leeftijd van de ondernemer voor dat bedrijfstype.

Betrouwbaarheid en nauwkeurigheid

Naast de representativiteit van een steekproef zijn ook de nauwkeurigheid en betrouwbaarheid waarmee je schattingen kan doen van belang. De nauwkeurigheid van je schatting zegt iets over het mogelijke verschil tussen de waarde van je steekproefschatting en de werkelijke waarde in de populatie.

De nauwkeurigheid is afhankelijk van het gewenste betrouwbaarheidsniveau, de homogeniteit van de waarnemingen en het aantal waarnemingen. Is de spreiding groot, dan zijn je waarnemingen heterogeen en is je variantie en daarmee de standaarddeviatie groot. Is de spreiding klein dan is er sprake van homogeniteit en een kleine variantie en standaarddeviatie. De nauwkeurigheid neemt toe met het aantal steekproefelementen in de steekproef.

De variantie van de steekproefschatter bij een enkelvoudige aselecte steekproef is gelijk aan:

$$v(\overline{Y}_D) = \frac{s_y^2}{n}$$

De variantie van de steekproefschatter bij een gestratificeerde steekproef is gelijk aan:

$$v(\bar{Y}_S) = \sum_{h=1}^{H} \frac{N_h}{N} v(\bar{Y}_{Dh})$$

$v(\bar{Y}_D)$ is de variantie van de directe schatter

$v(\bar{Y}_{Dh})$ is de variantie van de directe schatter in stratum h

$v(\bar{Y}_S)$ is de variantie van de gestratificeerde schatter

s_y^2 is de variantie van de variabele y in de steekproef

n is de steekproefgrootte

3.4 Literatuur

Voor informatie over de opzet van het Informatienet wordt de volgende literatuur aanbevolen:

- Dijk, J.P.M van, K. Lodder en H.C.J. Vrolijk (2002).
- Vrolijk, H.C.J. en K. Lodder (2002).

4. Introductie SPSS

4.1 Inleiding

In dit hoofdstuk wordt een introductiebeschrijving van SPSS gegeven. Niet alle mogelijkheden van SPSS worden behandeld, maar alleen de basis. Voor het verwerken van de opdrachten die verder in dit rapport worden gegeven is het van belang wat basiskennis te hebben van SPSS. Voor het gebruik van SPSS heb je kennis nodig van:
- het toepassingsgebied;
- statistiek;
- de mogelijkheden van SPSS;
- de bediening van SPSS.

De statistiek zelf wordt in de hoofdstukken 5 tot en met 11 behandeld en de kennis op het toepassingsgebied is specialistisch en ligt bij de onderzoeker zelf. De mogelijkheden en de bediening van SPSS zullen in dit hoofdstuk aan de orde komen. Ook worden in dit hoofdstuk verwijzingen naar SPSS-literatuur gegeven.

4.2 Onderzoek met SPSS

In deze paragraaf worden de verschillende functionaliteiten van SPSS beschreven. Eerst wordt algemeen ingegaan op het opzetten van onderzoek met SPSS en in welke fasen gebruikgemaakt kan worden van de SPSS functionaliteiten. Een onderzoek met behulp van SPSS kan bestaan uit de volgende fasen:
- het opstellen van een vragenlijst;
- het afnemen van een vragenlijst;
- het maken van een gegevensbestand;
- het bewerken van een gegevensbestand;
- het analyseren van de gegevens;
- het interpreteren van de gegevens;
- het rapporteren van de gegevens.

Van de bovengenoemde onderzoeksstappen vinden de eerste twee plaats buiten SPSS. Hoewel bij het opstellen van de vragenlijst wel nagedacht moet worden over de verwerking van de gegevens en de mogelijkheden die SPSS hierbij biedt. Je wilt natuurlijk je gegevens zo optimaal mogelijk gebruiken bij de analyse. Bij de derde tot en met de vijfde stap, het maken, bewerken en analyseren van het gegevensbestand, wordt wel gebruikgemaakt van SPSS. Vervolgens vindt het interpreteren en het rapporteren plaats. Dit kan zowel met als zonder gebruikmaking van SPSS.

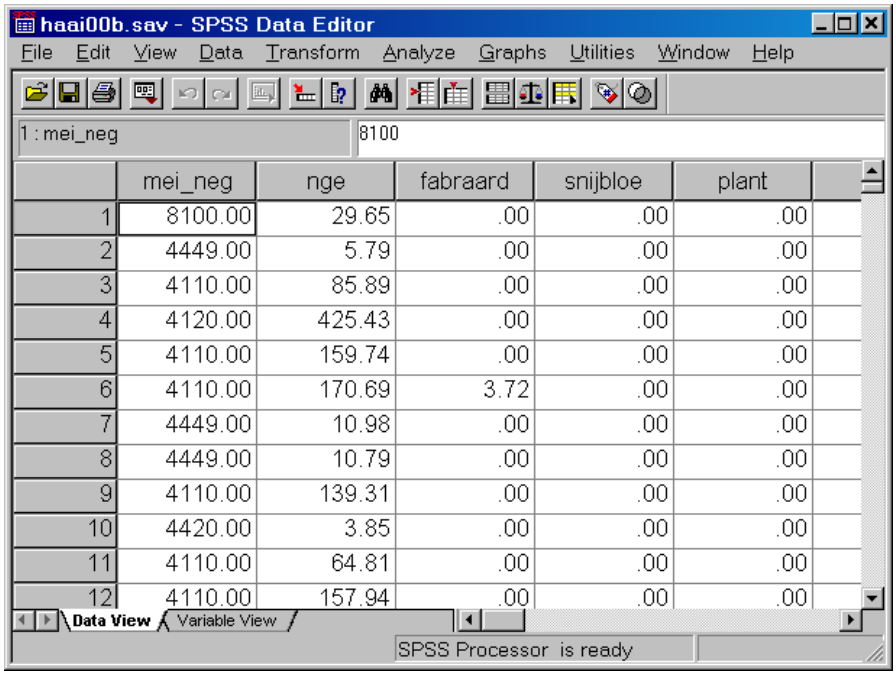
Bij het interpreteren en rapporteren van de resultaten kunnen grafieken en tabellen een

verduidelijkende rol spelen.

4.3 Bestanden in SPSS

SPSS maakt gebruik van databestanden, uitvoerbestanden en syntaxbestanden.

Databestanden
Databestanden waarin gegevens worden ingevoerd of ingelezen krijgen in SPSS de extensie *.sav. In *.sav-bestanden worden gegevens weergegeven in matrixvorm. Hierbij staan de verschillende variabelen in de verschillende kolommen en wordt er per waarneming één rij gebruikt. In figuur 4.1 is een voorbeeld van een *.sav-file gegeven.



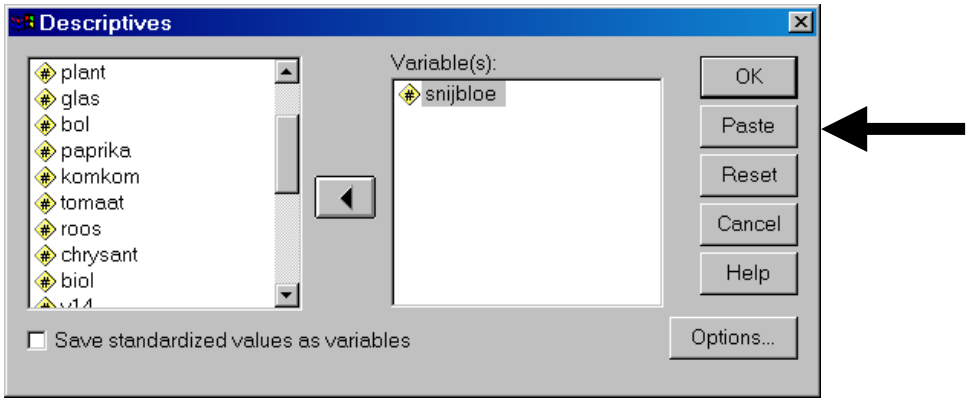
	mei_neg	nge	fabraard	snijbloe	plant
1	8100.00	29.65	.00	.00	.00
2	4449.00	5.79	.00	.00	.00
3	4110.00	85.89	.00	.00	.00
4	4120.00	425.43	.00	.00	.00
5	4110.00	159.74	.00	.00	.00
6	4110.00	170.69	3.72	.00	.00
7	4449.00	10.98	.00	.00	.00
8	4449.00	10.79	.00	.00	.00
9	4110.00	139.31	.00	.00	.00
10	4420.00	3.85	.00	.00	.00
11	4110.00	64.81	.00	.00	.00
12	4110.00	157.94	.00	.00	.00

*Figuur 4.1 Voorbeeld *.sav-file*

In een *.sav-file kan er geswitcht worden tussen data view en variable view. Normaal gesproken wil je de gegevens voor je hebben, dus zet je data view aan. Maar de mogelijkheid variable view is van belang bij het creëren van het bestand. In variable view kan het format van de variabelen en informatie over de variabelen vastgelegd worden. In paragraaf 4.4 wordt bij de functionaliteit invoer van gegevens aandacht besteed aan het gebruik van variable view.

Uitvoer bestanden
Uitvoer bestanden zijn bestanden waar de uitkomsten van gegevensanalyses naartoe geschreven worden. Deze bestanden hebben een *.spo-extensie. Hierin kunnen grafieken en tabellen worden gegenereerd.

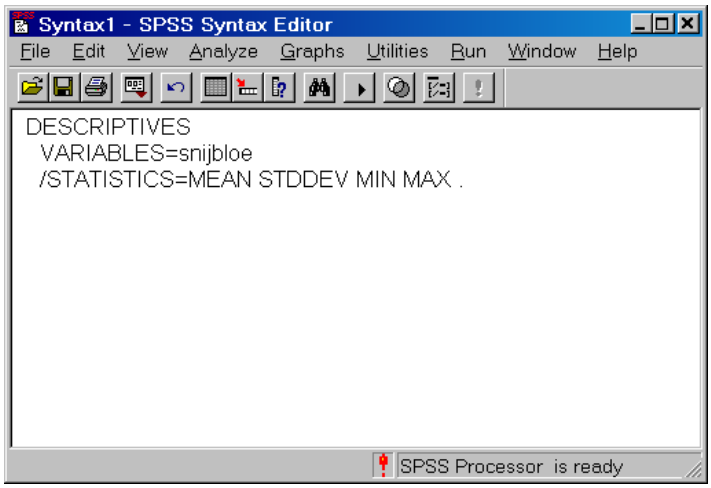
Syntax bestanden
SPSS genereert syntax voor elke opdracht die tijdens de analyse wordt uitgevoerd. Ook wanneer er via een toolbar een optie uit een menu wordt gekozen. Aan de hand van de syntax voert SPSS opdrachten uit. In de loop van de tijd zijn steeds meer mogelijkheden van SPSS via de dialoogvensters beschikbaar gekomen, maar in oude versies van SPSS moet er nog veel syntax met de hand worden ingevoerd. Op dit moment zijn er nog steeds toepassingsmogelijkheden die niet in een menu staan, maar die wel tot de SPSS mogelijkheden behoren. Conjunctanalyse is hier een voorbeeld van. De files waarin de syntax opgeslagen wordt hebben een *.sps extensie. Om een analyse reproduceerbaar te maken, ook in het kader van kwaliteitsborging, is het goed om syntaxfiles te genereren ook als je zelf geen syntax hoeft te schrijven. Dit kan je doen door de optie paste aan te klikken alvorens je een commando uitvoert.



Figuur 4.2 Schermstructuur en paste-commando

De syntax die gegenereerd wordt als in het descriptives-menu voor de variabele snijbloem gekozen wordt ziet er als te zien is in figuur 4.3.

Opdrachten kunnen ook vanuit een syntaxfile worden uitgevoerd. Het selecteren van de bovenstaande syntax en het runnen hiervan, zie het menu run, heeft hetzelfde effect als het drukken op de ok-knop in het venster dat in figuur 4.2 is weergegeven. In SPSS wordt ook automatisch een logboek gegenereerd van de stappen in de analyse die ondernomen worden, dit logboek is te vinden in de spss.jnl-file. In het menu edit-options kan worden nagegaan waar dit bestand wordt vastgelegd.



Figuur 4.3 Syntax gegenereerd

4.4 Functionaliteiten in SPSS

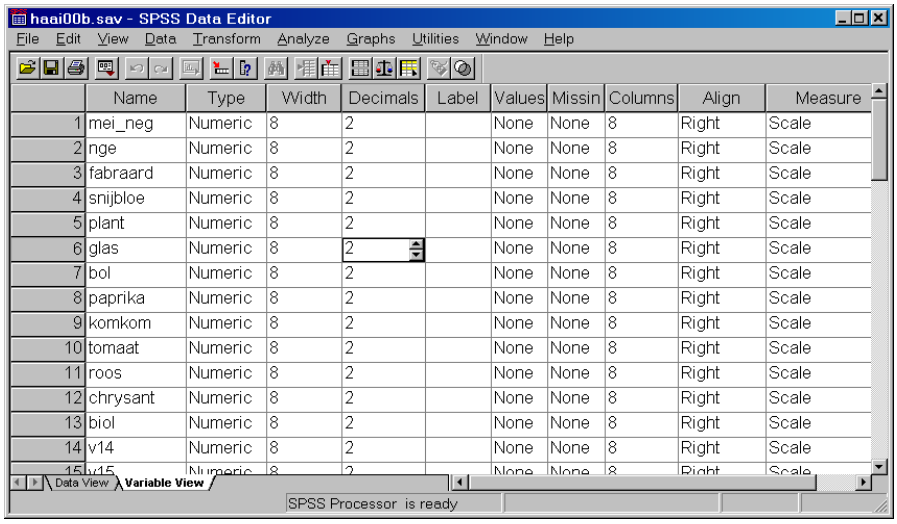
SPSS kent functionaliteit voor:
- het invoeren van gegevens;
- het inlezen van gegevens;
- het bewerken van gegevens;
- het analyseren van gegevens;
- het presenteren van gegevens.

Van de verschillende functionaliteiten worden voornamelijk gegevensinvoer, gegevensbewerking en gegevensanalyse toegelicht.

Invoer van gegevens
Het invoeren van gegevens in SPSS vindt plaats in een databestand, dus een *.sav-file. In de *.sav-file is naast het normale overzicht van de data (de data view) een overzicht ingebouwd, dat variable view heet. Dit overzicht kan het beste gebruikt worden voor het definiëren van het bestand, zodat het format van de data en informatie over de data duidelijk gedefinieerd en vastgelegd zijn.

In figuur 4.4 is een voorbeeld gegeven van variable view. Eén voor één worden de kolommen langsgelopen die voor een bepaalde variabele van belang kunnen zijn:
- in de eerste kolom wordt de naam van de variabele vastgelegd. Hiervoor zijn 8 karakters beschikbaar en de naam mag niet met een getal beginnen;
- in de tweede kolom wordt ingevuld om wat voor soort data het gaat. Gaat het om een getal, een string, of een geldeenheid. Naast deze opties is er nog een aantal mogelijkheden;
- de derde kolom stelt een maximum aan het aantal karakters dat voor de variabele beschikbaar is;

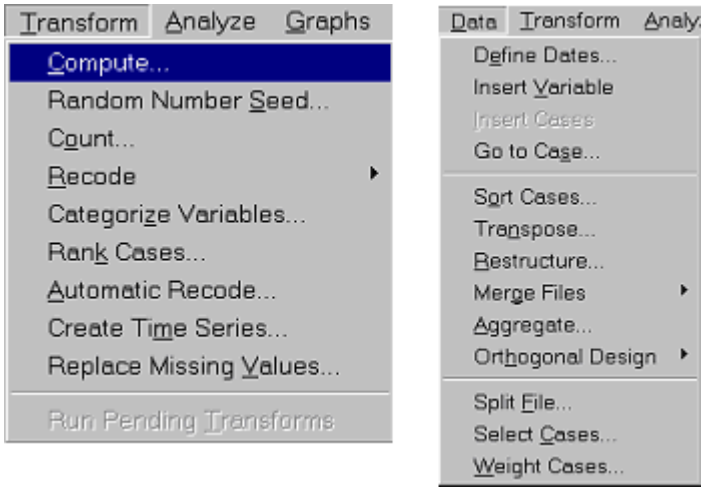
- de vierde kolom geeft het aantal decimalen. Voor een string is deze variabele niet relevant;
- een label kan gedefinieerd worden in de vijfde kolom. Dit is een beschrijving van de variabele zodat de betekenis van de variabele duidelijk is. Dit vergroot de herkenbaarheid en maakt hergebruik makkelijker;
- in de zesde kolom kan ingevuld worden wat de uitleg is die hoort bij bepaalde waarden van de variabele. Dit geldt voornamelijk bij nominale variabelen. Als bijvoorbeeld de kleur groen een waarde 1 krijgt, de kleur rood een waarde 2 en de kleur blauw een waarde 3, dan kan dat als verklaring van de waarden (values) ingevuld worden. Ook bij dummyvariabelen (variabelen die alleen de waarde 0 en 1 kunnen aannemen) kan bij values een beschrijving gegeven worden;
- in de kolom missing kan aangeven worden hoe ontbrekende waarden (missing values) worden omschreven en herkenbaar zijn in de datafile. Dit kan bijvoorbeeld door het getal 9999 hiervoor te reserveren. Behalve als dit getal ook voor kan komen in de echte dataset. De cellen die leeg zijn worden ook als missing values geïnterpreteerd;
- verder is er een kolom waarin het 'column format' aangegeven kan worden. Hiermee kan de breedte van de variabele worden bepaald die geldt bij het afdrukken en weergeven van gegevensbestanden. De breedte van het kolomformaat gaat dus niet over de fysieke breedte van het aantal tekens in een variabele, maar laat meer of minder zien van de inhoud van een cel;
- in kolom 9 kan aangegeven worden aan welke kant de data uitgelijnd moeten worden. Rechts, links of in het midden;
- in de laatste kolom, kolom 10, kan worden aangeven of het om een interval- of ratioschaal, ordinale of nominale variabele gaat.



Figuur 4.4 Definitie van variabelen

Overigens is het niet noodzakelijk voor data-analyse in SPSS dat de data ook in SPSS is ingevoerd. Ook gegevens uit onder andere excelfiles kunnen in SPSS worden ingelezen en geanalyseerd.

Bewerken van gegevens
Gegevensbewerking kan in SPSS uitgevoerd worden met behulp van de menu-opties transform en data, die boven in de menubar te vinden zijn.



Figuur 4.5 Bewerken van gegevens

Gegevensbewerking kan nodig zijn om verschillende redenen die hieronder worden uiteengezet. Lang niet alle bewerkingsmogelijkheden van SPSS worden in dit rapport besproken. Het gaat om een eerste introductie van de mogelijkheden. Voor een uitgebreidere uitleg wordt verwezen naar de literatuurverwijzingen in de laatste paragraaf van dit hoofdstuk.

De eerste reden voor gegevensbewerking is dat gegevens niet de juiste vorm hebben voor een bepaalde analyse. Leeftijd kan bijvoorbeeld voor een bepaalde analyse een handiger uitgangspunt zijn dan geboortejaar. Dan kan gebruikgemaakt worden van de optie *compute* in het *transform* menu. Hierin kan aan de hand van combinaties van bestaande variabelen en rekenregels een nieuwe variabele geconstrueerd worden. De rekenregel voor de nieuwe variabele leeftijd kan zijn: 2002-geboortejaar. Dit gaat altijd goed als je de gegevens op 31 december analyseert. Anders zou je er één jaar naast kunnen zitten.

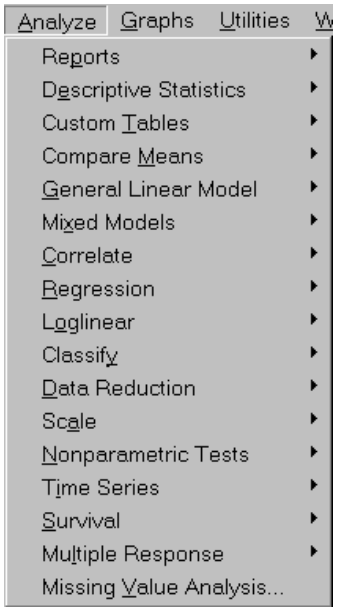
Een andere reden is dat je niet met de gehele dataset wilt werken, maar alleen met een deel van de data. Je wilt bijvoorbeeld een analyse uitvoeren op alleen akkerbouwbedrijven. Dit kan je het beste doen met *select cases* uit het menu *data*. Onder bepaalde voorwaarden, in te vullen bij 'if condition is satisfied', kunnen dan voorkomens (cases) op basis van een bepaalde variabele geselecteerd worden. Onderin kan je kiezen voor het verwijderen, of het filteren van gegevens: als je kiest voor verwijderen kan je de data die je geselecteerd hebt nooit meer terugkrijgen. Wees hier dan ook heel voorzichtig mee! Beter is om te kiezen voor unselected cases are filtered, dan kan het filter gewoon weer verwijderd worden en zijn alle cases weer beschikbaar.

Een andere mogelijkheid is het bewerken van een bepaalde variabele, door bijvoorbeeld van een continue variabele een klasse-indeling te maken, of het berekenen van een nieuwe variabele. Dit kan het beste gedaan worden aan de hand van de optie *recode* in het menu *transform*. Het beste is te kiezen voor *recode into different variable*, anders wordt de oude variabele overschreven.

Een laatste mogelijkheid die behandeld wordt is het sorteren van waarnemingen. Voor het sorteren wordt gebruikgemaakt van de optie *sort cases* in het menu *data*. Er kan in SPSS op basis van één of meerdere variabelen gesorteerd worden, zowel oplopend als afnemend.

Analyseren van gegevens

Ook bij de analyse van de gegevens wordt een aantal standaardtechnieken beschreven. Voor een uitputtend overzicht van alle analysetechnieken wordt verwezen naar de literatuurlijst in paragraaf 4.5. Bij het analyseren van data in SPSS wordt gebruikgemaakt van het menu *analyze*. In figuur 4.6 is een overzicht van de analysetechnieken gegeven die standaard in SPSS beschikbaar zijn. Zoals eerder aangegeven, geeft het onderstaande venster niet een volledig beeld van de mogelijke analysetechnieken. Een aantal technieken is wel beschikbaar, maar de gebruiker dient de opdracht zelf in het syntaxvenster te definiëren. Van de standaardopties worden beschrijvende technieken, kruistabellen en toetsen en technieken waarmee verbanden tussen variabelen kunnen worden onderzocht, besproken.



Figuur 4.6 Analyseren van gegevens

Van de beschrijvende technieken zijn *frequencies* en *explore* de belangrijkste. Met behulp van *frequencies* kunnen de belangrijkste kenmerken van variabelen worden weergegeven. Zowel gemiddelde, modus als mediaan kunnen hier worden weergegeven. Ook maten die de spreiding aangeven kunnen worden weergegeven. *Frequencies* is een optie onder *descriptive statistics* in het menu *analyze*. Aan de hand van *explore* dat ook onder *descriptive statistics* gevonden kan worden, kunnen ook plaatjes en grafieken uitgedraaid worden van de data. Zo krijg je een idee van outliers en van andere patronen in de data. Om een beeld te krijgen van de data die je onder handen hebt is het altijd goed om wat beschrijvende statistieken uit te draaien.

Met behulp van *crosstabs*, ook onder het menu *descriptive statistics* kan een eerste indruk gekregen worden van de samenhang tussen meerdere variabelen. Crosstabs kan het beste gebruikt worden als er niet al te veel verschillende voorkomens zijn van variabelen die je tegen elkaar uitzet, want anders krijg je een zeer grote tabel als output. Dus dit is vooral handig om te gebruiken als er sprake is van variabelen waarbij de cases zijn ingedeeld in klassen, dus bijvoorbeeld nominale variabelen.

Technieken om groepen te beschrijven en verschillen tussen groepen te toetsen zijn de technieken onder *compare means*. Als de mogelijkheid *means* wordt gekozen, worden de gemiddelden van de afhankelijke variabele conditioneel op de waarde van een onafhankelijke variabele weergegeven. Er wordt niet standaard een test uitgevoerd of de gemiddelden ook werkelijk verschillen voor de verschillende waarden van de onafhankelijke variabele. Als er wordt gekozen voor de one sample *t-test*, kan er getest worden of het gemiddelde van een variabele gelijk is aan een vooraf gekozen waarde. De ANOVA-test doet een variantietoets en gaat op die manier na of de gemiddelden van een bepaalde variabele van meer dan twee groepen significant van elkaar verschillen.

Technieken om verbanden tussen variabelen te bekijken en te testen zijn correlatie-analyse en regressieanalyse. Correlatie-analyse kijkt of twee variabelen onderlinge afhankelijkheid vertonen en regressieanalyse bekijkt de samenhang tussen meerdere variabelen. Correlatie- en regressieanalyse zullen inhoudelijk uitgebreid besproken worden in hoofdstuk 6 van dit rapport. In SPSS is correlatie-analyse te vinden onder de optie *correlate* in het menu *analyze*. Regressieanalyse is te vinden onder de optie *regression* in het menu *analyze*. Als je lineaire relaties tussen variabelen wilt onderzoeken dan kan je het beste de optie *linear* kiezen bij *regression*.

4.5 Literatuur

Over SPSS kan de volgende literatuur geraadpleegd worden:

- Huizing, K.R.E. (2002).
- SPSS (2001).

De eerste referentie is gemakkelijk te lezen. De tweede is een zeer uitgebreide SPSS manual. Verder is er zeer veel uiteenlopende SPSS-literatuur in omloop.

5. Schatters, hypothesen toetsen, verkennen en presenteren data

5.1 Inleiding

Je kunt de beschrijvende statistiek vergelijken met het maken van een samenvatting over een literair boek. In plaats van het gehele boek te lezen, kun je met de samenvatting de essentiële informatie overbrengen. Dit is ook de bedoeling bij de beschrijvende statistiek. Door op een juiste manier gebruik te maken van samenvattende statistieken en grafieken kan een goed inzicht in de populatie worden verkregen. Dit inzicht kan ondersteunend zijn bij het nemen van beleidsbeslissingen.

Wanneer je de gehele populatie kunt waarnemen, dan kun je de samenvattende statistieken voor deze populatie uitrekenen. Deze samenvattende statistieken worden ook wel populatieparameters genoemd. Wanneer je echter slechts een deel van de populatie kent (een steekproef) dan ben je slechts in staat een inschatting te maken van de populatieparameters. Zodra er geschat wordt dan is er sprake van onzekerheid. Dat wil zeggen: je weet niet voor 100% zeker dat de waarde die je hebt uitgerekend ook de juiste populatieparameter is. Als je gebruikmaakt van schatters, ben je wel in staat om een betrouwbaarheidsinterval te geven. Dit is een interval waarvan je met een bepaalde zekerheid weet dat de echte populatieparameter hierin valt.

Zodra data gepresenteerd wordt kunnen er allerlei vragen opkomen. Bijvoorbeeld: 'Zijn de inkomsten van dit jaar werkelijk hoger dan die van vorig jaar?' en 'Heeft iets wel invloed op een bepaald proces of niet?'. Omdat de data die gepresenteerd wordt vaak het resultaat zijn van een steekproefschatting, moeten deze vragen statistisch getoetst worden voordat er met een bepaalde statistische zekerheid gezegd kan worden of iets wel of niet het geval is.

5.2 Inhoud

In deze paragraaf wordt een introductie of opfrissing gegeven van beschrijvende statistieken. Als eerste worden schatters en parameters beschreven in paragraaf 5.2.1. Vervolgens komen kansverdelingen aan bod in paragraaf 5.2.2. Paragraaf 5.2.3 gaat in op betrouwbaarheidsintervallen. In paragraaf 5.2.4 wordt het toetsen van hypothesen besproken en als laatste worden non-parametrische toetsen beschreven in paragraaf 5.2.5.

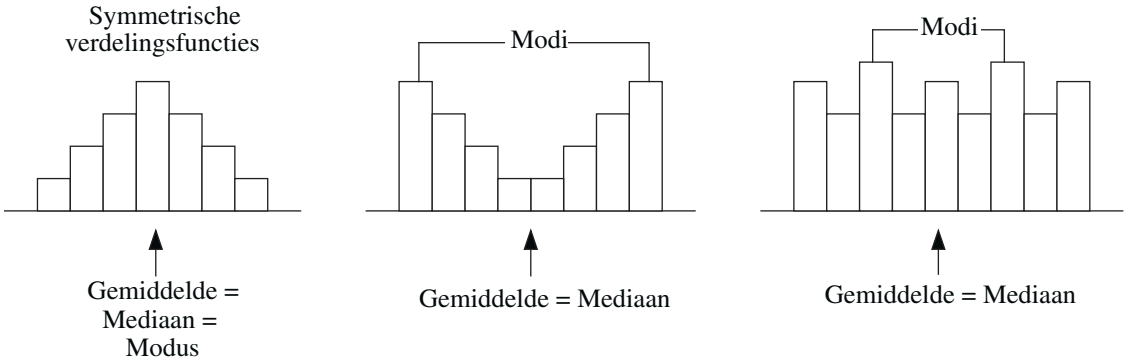
5.2.1 Schatters en parameters

Om een goede 'samenvatting' te maken van de populatie worden vaak de volgende parameters/schatters gebruikt:

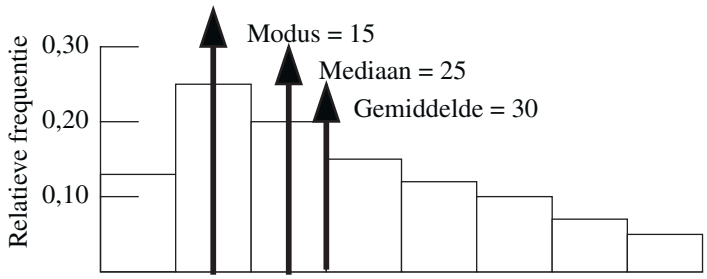
- gemiddelde;
- mediaan;
- modus;

- standaarddeviatie;
- variantie;
- bereik;
- quartielen enzovoorts;
- kurtosis;
- …(SPSS geeft nog veel meer parameters/schatters)

Om een indruk te krijgen van de waarden van de populatie/steekproef worden vaak frequentieverdelingen gemaakt. De vorm van deze frequentieverdelingen wordt vaak al verklaard door het verschil tussen het gemiddelde, de mediaan en de modus. De figuren 5.1 en 5.2 geven daarvan voorbeelden.



Figuur 5.1 Verschillen van modus, mediaan en gemiddelde geven de vorm van een frequentieverdeling weer



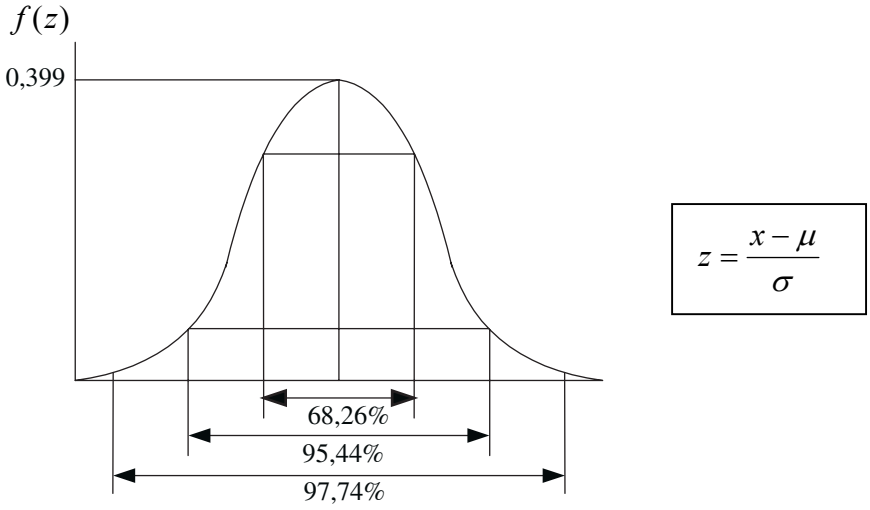
Figuur 5.2 Modus, mediaan en gemiddelde voor een scheve verdeling met een staart naar rechts

5.2.2 Kansverdelingen

Een populatie wordt volledig beschreven door zijn momenten. Van de momenten zijn het gemiddelde en de variantie de bekendste. Frequentieverdelingen zijn eenvoudig om te zetten in kansverdelingen. Binnen de statistiek is een aantal kansverdelingen belangrijk:
- normale verdeling;

- t-verdeling;
- chi-kwadraatverdeling;
- F-verdeling.

Normale verdeling
Zoals later zal blijken zijn dit de kansverdelingen die een rol spelen bij het toetsen van hypothesen en de verklarende statistiek. De normale verdeling is handig omdat alle berekeningen die je moet doen te herleiden zijn tot een standaard normale verdeling. Dit is een verdeling met een gemiddelde van 0 en een variantie van 1: $N(0,1)$. Een vereiste hiervoor is echter wel dat het gemiddelde, of de verwachte waarde als het gaat om een kansverdeling, en de variantie bekend zijn. In figuur 5.3 zie je de vorm van de (standaard) normale verdeling. Z-waarden en de bijbehorende kansen van een standaard normale verdeling zijn terug te vinden in tabel 5.1.



Figuur 5.3 Normale verdeling

De cumulatieve normale verdeling wordt weergegeven door:

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{z^2/2} dz$$

De bijbehorende tabel 5.1 kan op de volgende manier gebruikt worden: de kans dat z tussen 0 en 1,5 in ligt is 0,433. Zie ook onderstaande tabel bij de coördinaten (1,5). Dus P(z>0 en z < 1,5) = 0,433).

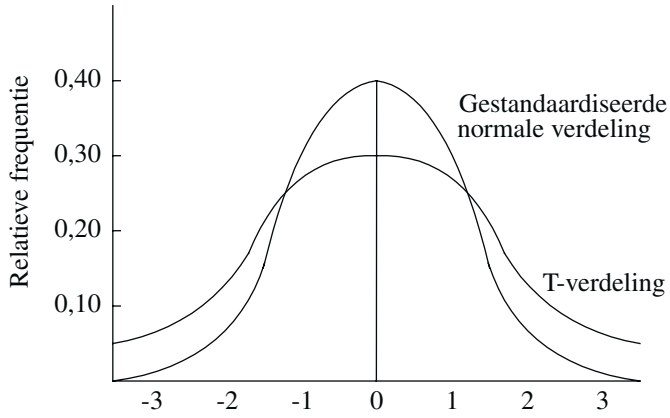
Tabel 5.1 Standaard normale verdeling

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,000	0,004	0,008	0,012	0,016	0,020	0,024	0,028	0,032	0,036
0,1	0,040	0,044	0,048	0,052	0,056	0,060	0,064	0,068	0,071	0,075
0,2	0,079	0,083	0,087	0,091	0,095	0,099	0,103	0,106	0,110	0,114
0,3	0,118	0,122	0,126	0,129	0,133	0,137	0,141	0,144	0,148	0,152
0,4	0,155	0,159	0,163	0,166	0,170	0,174	0,177	0,181	0,184	0,188
0,5	0,192	0,195	0,199	0,202	0,205	0,209	0,212	0,216	0,219	0,222
0,6	0,226	0,229	0,232	0,236	0,239	0,242	0,245	0,249	0,252	0,255
0,7	0,258	0,261	0,264	0,267	0,270	0,273	0,276	0,279	0,282	0,285
0,8	0,288	0,291	0,294	0,297	0,300	0,302	0,305	0,308	0,311	0,313
0,9	0,316	0,319	0,321	0,324	0,326	0,329	0,332	0,334	0,337	0,339
1	0,341	0,344	0,346	0,349	0,351	0,353	0,355	0,358	0,360	0,362
1,1	0,364	0,367	0,369	0,371	0,373	0,375	0,377	0,379	0,381	0,383
1,2	0,385	0,387	0,389	0,391	0,393	0,394	0,396	0,398	0,400	0,402
1,3	0,403	0,405	0,407	0,408	0,410	0,412	0,413	0,415	0,416	0,418
1,4	0,419	0,421	0,422	0,424	0,425	0,427	0,428	0,429	0,431	0,432
1,5	0,433	0,435	0,436	0,437	0,438	0,439	0,441	0,442	0,443	0,444
1,6	0,445	0,446	0,447	0,448	0,450	0,451	0,452	0,453	0,454	0,455
1,7	0,455	0,456	0,457	0,458	0,459	0,460	0,461	0,462	0,463	0,463
1,8	0,464	0,465	0,466	0,466	0,467	0,468	0,469	0,469	0,470	0,471
1,9	0,471	0,472	0,473	0,473	0,474	0,474	0,475	0,476	0,476	0,477
2	0,477	0,478	0,478	0,479	0,479	0,480	0,480	0,481	0,481	0,482
2,1	0,482	0,483	0,483	0,483	0,484	0,484	0,485	0,485	0,485	0,486
2,2	0,486	0,486	0,487	0,487	0,488	0,488	0,488	0,488	0,489	0,489
2,3	0,489	0,490	0,490	0,490	0,490	0,491	0,491	0,491	0,491	0,492
2,4	0,492	0,492	0,492	0,493	0,493	0,493	0,493	0,493	0,493	0,494
2,5	0,494	0,494	0,494	0,494	0,495	0,495	0,495	0,495	0,495	0,495
2,6	0,495	0,496	0,496	0,496	0,496	0,496	0,496	0,496	0,496	0,496
2,7	0,497	0,497	0,497	0,497	0,497	0,497	0,497	0,497	0,497	0,497
2,8	0,497	0,498	0,498	0,498	0,498	0,498	0,498	0,498	0,498	0,498
2,9	0,498	0,498	0,498	0,498	0,498	0,498	0,499	0,499	0,499	0,499
3	0,499	0,499	0,499	0,499	0,499	0,499	0,499	0,499	0,499	0,499

T-verdeling

De t-verdeling lijkt op de normale verdeling (zie figuur 5.4). De t-verdeling heeft echter meer massa in de staarten van de verdeling. De grootte van deze massa wordt bepaald door het aantal vrijheidsgraden. De vrijheidsgraden geven de mate van onzekerheid die je hebt over de betrouwbaarheid van je schatter. Voor bepaalde combinaties van vrijheidsgraden en kansen zijn t-waarden uit een tabel te halen (zie tabel 5.2). De t-verdeling wordt weergegeven door:

$$F(t) = \int_{-\infty}^{t} \frac{\left(\frac{v-1}{2}\right)!}{\left(\frac{v-2}{2}\right)!\sqrt{\pi n}\left(1+\frac{t^2}{v}\right)^{(v+1)/2}}\,dt$$



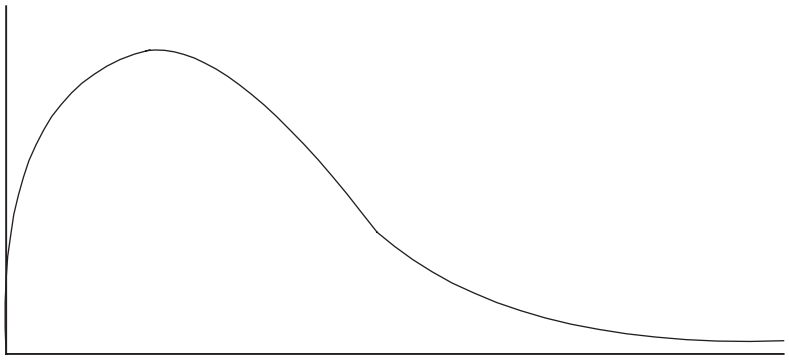
Figuur 5.4 Normale verdeling en t-verdeling

Tabel 5.2 T-tabel met kansen in de rechterstaart

df\p	0,4	0,25	0,1	0,05	0,025	0,01	0,005	0,0005
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	636,619
2	0,289	0,816	1,886	2,920	4,303	6,965	9,925	31,599
3	0,277	0,765	1,638	2,353	3,182	4,541	5,841	12,924
4	0,271	0,741	1,533	2,132	2,776	3,747	4,604	8,610
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	6,869
6	0,265	0,718	1,440	1,943	2,447	3,143	3,707	5,959
7	0,263	0,711	1,415	1,895	2,365	2,998	3,499	5,408
8	0,262	0,706	1,397	1,860	2,306	2,896	3,355	5,041
9	0,261	0,703	1,383	1,833	2,262	2,821	3,250	4,781
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	4,587
11	0,260	0,697	1,363	1,796	2,201	2,718	3,106	4,437
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	4,318
13	0,259	0,694	1,350	1,771	2,160	2,650	3,012	4,221
14	0,258	0,692	1,345	1,761	2,145	2,624	2,977	4,141
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	4,073
16	0,258	0,690	1,337	1,746	2,120	2,583	2,921	4,015
17	0,257	0,689	1,333	1,740	2,110	2,567	2,898	3,965
18	0,257	0,688	1,330	1,734	2,101	2,552	2,878	3,922
19	0,257	0,688	1,328	1,729	2,093	2,539	2,861	3,883
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,850
21	0,257	0,686	1,323	1,721	2,080	2,518	2,831	3,819
22	0,256	0,686	1,321	1,717	2,074	2,508	2,819	3,792
23	0,256	0,685	1,319	1,714	2,069	2,500	2,807	3,768
24	0,256	0,685	1,318	1,711	2,064	2,492	2,797	3,745
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,725
26	0,256	0,684	1,315	1,706	2,056	2,479	2,779	3,707
27	0,256	0,684	1,314	1,703	2,052	2,473	2,771	3,690
28	0,256	0,683	1,313	1,701	2,048	2,467	2,763	3,674
29	0,256	0,683	1,311	1,699	2,045	2,462	2,756	3,659
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,646
inf	0,253	0,674	1,282	1,645	1,960	2,326	2,576	3,291

Tabel 5.2 geeft verschillende t-waarden weer. Uitgaande van negentien vrijheidsgraden is de kans 0,025 dat t groter of gelijk is aan 2,093. Het aantal vrijheidsgraden (df van degrees of freedom) is te vinden op de verticale as en de kans is weergegeven op de horizontale as. Bovenstaande kan ook worden omschreven als P(t19≥2,093)=0,025).

Chi-kwadraatverdeling
De chi-kwadraatverdeling kun je zien als de som van onafhankelijke gekwadrateerde normale verdelingen. Tabel 5.3 geeft voor deze verdeling voor bepaalde vrijheidsgraden en kansen de bijbehorende chi-kwadraatwaarden. In figuur 5.5 is het verloop van de verdelingsfunctie weergegeven.



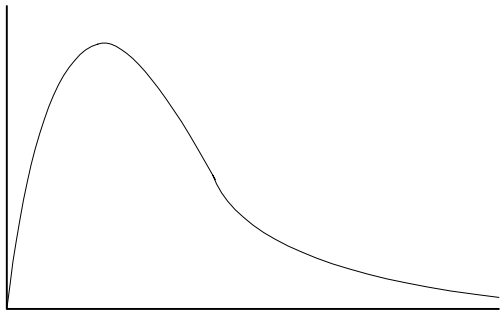
Figuur 5.5 Chi-kwadraatverdeling

De chi-kwadraatverdeling wordt weergegeven door:

$$F(\chi^2) = \int_0^{\chi^2} \frac{\chi^{(v-2)/2} e^{-\chi/2} d\chi}{2^{v/2}\left[(v-2)/2\right]}$$

Tabel 5.3 geeft de kritische waarden van de chi-kwadraatverdeling weer. Wederom uitgaande van 19 vrijheidsgraden is de kans 0,05 dat chi-kwadraat groter of gelijk is aan 30,1. Het aantal vrijheidsgraden (df van degrees of freedom) is te vinden op de verticale as en de kans is weergegeven op de horizontale as. Bovenstaande kan ook worden omschreven als P(χ219≥30,1)=0,05.

F-verdeling
De F-verdeling is de breuk van twee chi-kwadraatverdelingen. Figuur 5.6 geeft het verloop van de F-verdeling weer. In tabel 5.5 zijn de F-waarden terug te vinden.



Figuur 5.6 Verloop F-verdeling

Tabel 5.3 Chi-kwadraatverdeling met de rechterstaartwaarden

df\p	0,995	0,99	0,975	0,95	0,9	0,75	0,5	0,25	0,1	0,05	0,025	0,01	0,005
1,0	0,00	0,00	0,00	0,00	0,02	0,10	0,45	1,32	2,71	3,84	5,02	6,63	7,88
2,0	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3,0	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4,0	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5,0	0,41	0,55	0,83	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6,0	0,68	0,87	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7,0	0,99	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8,0	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,95
9,0	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10,0	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11,0	2,60	3,05	3,82	4,57	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,72	26,76
12,0	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13,0	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14,0	4,07	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15,0	4,60	5,23	6,26	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16,0	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17,0	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18,0	6,26	7,01	8,23	9,39	10,86	13,68	17,34	21,60	25,99	28,87	31,53	34,81	37,16
19,0	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20,0	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00
21,0	8,03	8,90	10,28	11,59	13,24	16,34	20,34	24,93	29,62	32,67	35,48	38,93	41,40
22,0	8,64	9,54	10,98	12,34	14,04	17,24	21,34	26,04	30,81	33,92	36,78	40,29	42,80
23,0	9,26	10,20	11,69	13,09	14,85	18,14	22,34	27,14	32,01	35,17	38,08	41,64	44,18
24,0	9,89	10,86	12,40	13,85	15,66	19,04	23,34	28,24	33,20	36,42	39,36	42,98	45,56
25,0	10,52	11,52	13,12	14,61	16,47	19,94	24,34	29,34	34,38	37,65	40,65	44,31	46,93
26,0	11,16	12,20	13,84	15,38	17,29	20,84	25,34	30,43	35,56	38,89	41,92	45,64	48,29
27,0	11,81	12,88	14,57	16,15	18,11	21,75	26,34	31,53	36,74	40,11	43,19	46,96	49,64
28,0	12,46	13,56	15,31	16,93	18,94	22,66	27,34	32,62	37,92	41,34	44,46	48,28	50,99
29,0	13,12	14,26	16,05	17,71	19,77	23,57	28,34	33,71	39,09	42,56	45,72	49,59	52,34
30,0	13,79	14,95	16,79	18,49	20,60	24,48	29,34	34,80	40,26	43,77	46,98	50,89	53,67

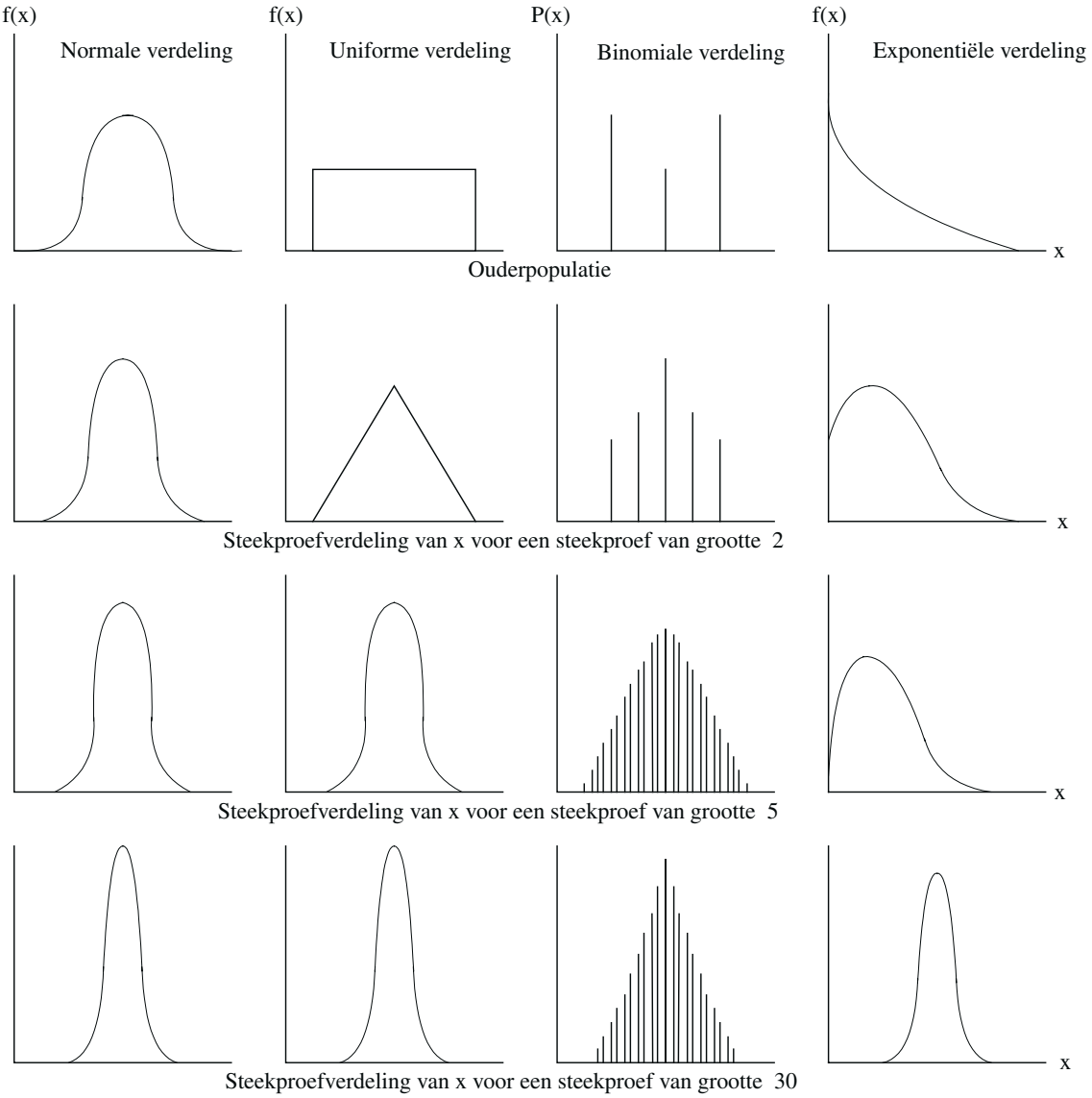
Tabel 5.4 *F-verdeling*

1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	243,91	245,95	248,01	249,05	250,10	251,14	252,20	253,25	254,31
18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39

In de tabel zijn de kritische waarden van de F-verdeling weergegeven bij een betrouwbaar-

heidsniveau van 95% (α= 0,05). De kans dat F groter of gelijk is aan 2,20 is 0,05 uitgaande van 15 en 20 graden van vrijheid. Bovenstaande kan ook worden omschreven als (P (F≥2,20) = 0,05 (met 15 en 20 graden van vrijheid)). De vrijheidsgraden v1 en v2 zijn in de tabel respectievelijk uitgezet op de horizontale en de verticale as.

Binnen de statistiek nemen we snel aan dat iets normaal verdeeld is. De vraag bij hoeveel steekproefwaarnemingen een steekproefgemiddelde normaal verdeeld is, wordt beantwoord door figuur 5.7. Stel je trekt een steekproef van omvang n uit een kansverdeling en berekent het steekproefgemiddelde. Je herhaalt dit trekkingsproces heel vaak en kijkt welke frequen-



Figuur 5.7 De verdeling van het gemiddelde voor diverse populatieverdelingen voor steekproef omvang 2, 5 en 30

tie/kansverdeling het steekproefgemiddelde oplevert. Als je dit doet voor verschillende kansverdelingen en voor verschillende steekproefgroottes, dan kun je kijken hoe de verdeling van het steekproefgemiddelde eruit gaat zien bij de diverse combinaties.

Random Variabele	Situatie	Resulterende verdelingsfunctie	Gemiddelde	Variantie
$\bar{x}$	Populatie normaal σ onbekend steekproefgrootte n	$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma / \sqrt{n}}$	0	1
$\bar{x}$	Populatie normaal σ onbekend steekproefgrootte n*	$t_v = \frac{\bar{x} - \mu_{\bar{x}}}{s / \sqrt{n}}$	0	v/(v-2) met v = n-1 vrijheidsgraden
$\bar{x}$	Populatie onbekend σ bekend n > 30	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$	0	1
s^2	Populatie normaal steekproefgrootte n	$\chi^2_v = \frac{(n-1)s^2}{\sigma^2}$	σ^2	$\frac{2\sigma^4}{v}$ met v = n-1 vrijheidsgraden
s_1^2 s_2^2	Populatie normaal steekproefgrootte van n1 en n2	$F_{v_1 v_2} = s_1^2 / s_2^2$	-	- met v1=n1-1 en v2=n2-1 vrijheidsgraden

Figuur 5.8 Samenvatting van de steekproefkansverdelingen

Als n > 30, $\bar{x}$ zal waarschijnlijk normaal verdeeld zijn en s benadert σ; dus $(\bar{x} - \mu)(s/n)$ benadert N(0,1).

<table>
<tr><td colspan="6">Probleem waarbij gemiddelden een rol spelen (bijv. testen van geobserveerde $\bar{x}$ versus aangenomen μ.</td></tr>
<tr><td colspan="3">Populatie σ_x bekend (of aangenomen)</td><td colspan="3">Populatie σ_x onbekend</td></tr>
<tr><td colspan="2">Niet-normale ouderpopulatie</td><td>Normale ouderpopulatie</td><td colspan="2">Niet-normale ouderpopulatie</td><td>Normale ouderpopulatie</td></tr>
<tr><td>N klein a)
Niet oplosbaar</td><td>n groot a)
Eindige populatie zonder terugleggingen
$z = \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-}{N-}}}$</td><td>Oneindige populatie of populatie met terugleggingen
$z = \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}}$</td><td>$z = \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}}$</td><td>n klein a)
Niet oplosbaar</td><td>n groot a)
$t \cong \frac{\bar{x} - \mu_x}{s / \sqrt{n}}$
(centrale limiet stelling)</td><td>$t = \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}}$</td></tr>
</table>

Figuur 5.9 Gebruik van kansverdelingen bij steekproeven (onderzoeksvragen die betrekking hebben op gemiddelden)

a) 'Groot' en 'klein' is afhankelijk van de gewenste nauwkeurigheid en de vorm van de ouderpopulatie. Voor (bijna) symmetrische verdelingsfuncties, n≥10 kan al voldoende zijn. In bijna alle gevallen is n≥30 toereikend voor 'groot'.

Voor steekproeven met een omvang van 30 waarnemingen lijkt de normale verdeling al een goede benadering. Dit geldt zelfs voor niet-symmetrische verdelingen. Je kunt daarom concluderen dat het steekproefgemiddelde normaal verdeeld is. Wanneer je minder dan 30 waarnemingen hebt, dan kun je de t-verdeling gebruiken als een benadering. Als het gemiddelde normaal verdeeld is, dan is de variantie een kwadraat van normale verdelingen en daarom een chi-kwadraatverdeling. Als je twee populaties met elkaar vergelijkt, zal je gebruikmakend van bovenstaande al snel de F-verdeling hebben gevonden. De F-verdeling zou bijvoorbeeld uitkomst kunnen bieden in het geval je de uitkomsten van het Informatienet van twee jaar met elkaar wilt vergelijken. De bovenstaande resultaten zijn terug te vinden in de figuur 5.8 en 5.9.

<table>
<tr><td colspan="6">Probleem waarbij varianties een rol spelen (bijv. testen van geobserveerde s2 versus aangenomen σ2.</td></tr>
<tr><td colspan="3">Eén steekproef</td><td colspan="3">Twee steekproeven</td></tr>
<tr><td colspan="2">Niet-normale ouderpopulatie</td><td>Normale ouderpopulatie</td><td colspan="2">Niet-normale ouderpopulatie</td><td>Normale ouderpopulatie</td></tr>
<tr><td>n klein a)
Niet oplosbaar</td><td>N groot a)
$\chi^2 \cong \frac{(n-1)s^2}{\sigma_x^2}$</td><td>$\chi^2 = \frac{(n-1)s}{\sigma^2}$</td><td>n1 en n2 klein a)
Niet oplosbaar</td><td>n1 en n2 groot a)
$F \cong \frac{s_1^2}{s_2^2}$</td><td>$F = \frac{s_1^2}{s_2^2}$</td></tr>
</table>

Figuur 5.10 Het gebruik van kansverdelingen bij steekproeven (onderzoeksvragen die betrekking hebben op varianties)

a) 'Groot' en 'klein' is afhankelijk van de gewenste nauwkeurigheid en de vorm van de ouderpopulatie. Voor (bijna) symmetrische verdelingsfuncties, n≥10 kan al voldoende zijn. In bijna alle gevallen is n≥30 toereikend voor 'groot'.

5.2.3 Betrouwbaarheidsintervallen

Wanneer je het steekproefgemiddelde uitrekent, dan kun je met de t-verdeling of de normale verdeling een betrouwbaarheidsinterval maken. Voor de variantie wordt gebruikgemaakt van de chi-kwadraatkansverdeling om een betrouwbaarheidsinterval te maken. Een samenvatting van dit alles is te vinden in figuur 5.11.

Onbekende parameter	Populatiekarakteristieken en andere omschrijving	Teststatistieken om tot de beste steekproefschatter te komen	Eindpunt voor een 100 (1-α)% betrouwbaar-heidsinterval
Gemiddelde μ	Populatie N(μ, σ2) of steek-proefgrootte n≥30, σ bekend	$z = \frac{(\bar{x} - \mu)}{(\sigma / \sqrt{n})}$	$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n})$
Gemiddelde μ	Poplatie N(μ, σ2), σ onbekend	$t_{vd.f.} = \frac{(\bar{x} - \mu)}{(s / \sqrt{n})}$ waarbij v=n-1 d.f. = vrijheidsgraden	$\bar{x} \pm t_{(\alpha/2,v)}(s / \sqrt{n})$
Variantie σ2	Populatie N(μ, σ2)	$\chi^2_{vd.f.} = \frac{(n-1)s^2}{\sigma^2}$ waarbij v=n-1 d.f. = vrijheidsgraden	$\frac{(n-1)s^2}{\chi^2_{(\alpha/2,v)}}$ en $\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,v)}}$
Proportie p	Herhaalde onafhankelijke proefneming, npq≥3, $\hat{p}$ =x/n	$z = (\hat{p} - p) \Big/ \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Figuur 5.11 Toetsen en betrouwbaarheidsintervallen

5.2.4 Hypothesen toetsen

Bij het LEI worden vaak tabellen gepubliceerd met de uitkomsten van de afgelopen 4 jaren. Hiermee wordt gesuggereerd dat deze uitkomsten direct met elkaar vergeleken kunnen worden. Dit is onjuist! Omdat de getallen over de afgelopen 4 jaren allemaal het gevolg van steekproeven zijn, en dus schatters zijn van de populatieparameter, moeten er hypothesen geformuleerd en getoetst worden. Het is niet genoeg om alleen naar de waardes van de 4 steekproefjaren te kijken. Voor goed onderzoek moet je van tevoren nadenken welke hypothesen je wilt toetsen en hoe betrouwbaar de uitspraak moet zijn. Dit bepaald onder andere de steekproefgrootte en de toets die je uitvoert. Statistiek is niet met zekerheid iets zeggen maar beredeneerd gokken en daarmee kansen op fouten accepteren. Bij het toetsen van hypothesen komen twee soorten fouten voor:

- α is de fout van de eerste soort (type-I-fout);
- β is de fout van de tweede soort (type-II-fout).

In figuur 5.11 wordt de samenhang tussen de twee soorten fouten weergegeven. Figuur 5.12 geeft een overzicht van de toetsen die we kunnen uitvoeren.

Actie	De werkelijke situatie kan zijn:	
	H0 is waar	H0 is niet waar
Accepteren van H0	Goede beslissing	Foute beslissing (Type II fout)
Accepteren van H1	Foute beslissing (Type I fout)	Goede beslissing

De vier mogelijke beslissingsuitkomsten bij het testen van hypothesen

Actie	De werkelijke situatie kan zijn:	
	H0 is waar	H0 is niet waar
Accepteren van H0	1 - α (Betrouwbaarheidslevel)	β
Accepteren van H1	α	1 - β (Power van de test)
Som	1,00	1,00

De kans op de verschillende beslissingsuitkomsten bij het testen van hypothesen.

Figuur 5.12 α- en β-fouten

5.2.5 Non-parametrische toetsen

Alle toetsen die tot nu toe beschreven zijn, gaan er vanuit dat de data minstens een intervalschaal hebben. Vaak wordt aangenomen dat de normale verdeling geldt. Wat moet je doen als de data niet-intervalgeschaald is en hooguit de ordinaliteit van de data bekend is? En wat als er geen normaliteit verondersteld kan worden omdat er bijvoorbeeld te weinig waarnemingen zijn? Dan kan je non-parametrisch toetsen. Voorbeelden van niet parametrische toetsen zijn:

- Mann-Whitney U-toets. Hiermee kan getoetst worden of twee steekproeven uit dezelfde populatie komen. We veronderstellen hier geen t-verdeling van de steekproefgemiddelden;
- Wald-Wolfowitz Runs-toets. Deze toets gaat na of twee steekproeven uit dezelfde populatie komen. Deze toets is minder krachtig dan Mann-Whitney, maar handig als Mann-Whitney minder geschikt is;
- Chi-kwadraat-toets. Voordat deze toets kan worden uitgevoerd wordt een set van waarnemingen aan de hand van een attribuuteigenschap geclassificeerd in c-groepen. Vervolgens wordt getoetst hoe goed een set van waargenomen frequenties een set van theoretische frequenties benadert;
- Kolmogorov-Smirnov-toets. Deze toets vergelijkt de theoretische en de steekproef cumulatieve frequentieverdeling.

Onbekende parameter	Populatiebeschijving	Test statistiek
Gemiddelde μ	Populatie N(μ, σ2) of steek-proefgrootte n≥30, σ bekend	$z = \frac{(\bar{x} - \mu_0)}{(\sigma / \sqrt{n})}$
Gemiddelde μ	Populatie N(μ, σ2), σ onbekend	$t_{vd.f.} = \frac{(\bar{x} - \mu)}{(s / \sqrt{n})}$ waarbij v=n-1 d.f. = vrijheidsgraden
Verschil tussen μ1 en μ2	Beide populaties normaal, of n1 en n2≥25 en σ_1^2 en σ_2^2 bekend	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$
Verschil tussen μ1 en μ2	Beide populaties normaal en σ_1^2 en σ_2^2 onbekend maar wel gelijk aan elkaar	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]}{(n_1 + n_2 - 2)}\left(\frac{n_1 + n_2}{n_1 n_2}\right)\right)}}$
Verschil tussen μ1 en μ2 (paren gematched)	Beide populaties normaal	$t = \frac{\overline{D} - \Delta}{s_p / \sqrt{n}}$
Proportie p	Herhaalde onafhankelijke proefneming, npq≥3,	$z = \frac{(x/n) - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$
Verschil tussen p1 en p2 = 0	n1 en n2 zijn beiden 'groot'	$z = \frac{\left[(x_1 / n_1) - (x_2 / n_2)\right]}{\sqrt{\left(\frac{x_1 + x_2}{n_1 + n_2}\right)\left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)\left(\frac{n_1 + n_2}{n_1 n_2}\right)}}$

Figuur 5.13 Toetsen van hypothesen

5.3 Literatuur

- Harnett, D.L. (1982).
- Thomas H. en R. J. Wonnacott (1990).
- Slotboom, A. (1996).
- Brink, W.P. van den en Koele, P. (1985).
- Brink, W.P. van den en Koele, P. (1986).
- Brink, W.P. van den en Koele, P. (1987).

Er zijn veel eerstejaars statistiekboeken waarin de basisconcepten van de statistiek worden behandeld. In dit hoofdstuk is veelvuldig gebruikgemaakt van Harnett. Denk niet dat jouw probleem kant-en-klaar staat uitgelegd in een statistiekboek. Je zult zelf moeten herkennen welke methoden op jouw probleem toepasbaar zijn en welke voor- en nadelen een methode heeft. Wat nodig is voor goed onderzoek is een dosis doorzettingsvermogen en creativiteit.

6. Technieken om resultaten steekproefschattingen te verbeteren

6.1 Inleiding [1]

Regelmatig voert het LEI onderzoek uit waarbij resultaten voor een klein gebied (gemeenten, provincies, landbouwgebieden, kaartvierkanten) of kleine groep gewenst zijn. In veel gevallen worden deze resultaten geproduceerd door beschikbare of berekende bedrijfsgegevens 'op te hogen' naar het gewenste aggregatieniveau. Het is noodzakelijk dat er voldoende waarnemingen voor het gebied zijn om verantwoord te kunnen aggregeren.

Aggregatie van gegevens die betrekking hebben op Informatienet-bedrijven is voor kleine gebieden veelal niet mogelijk op basis van de gebruikelijke procedure die gebruik maakt van wegingsfactoren. In de loop van de tijd zijn daarom op het LEI verschillende methodes toegepast/ontwikkeld die het mogelijk maken om toch bruikbare informatie op een laag ruimtelijk aggregatieniveau te genereren. Het is nuttig een vergelijking te maken van de beschikbare methoden voor het maken van schattingen voor kleine gebieden.

6.2 Schattingsmethoden

Verschillende methoden komen in aanmerking om schattingen te maken op basis van kleine deelgebieden. Er wordt een samenvatting gegeven van alle methoden die uit het onderzoek naar kleine deelgebieden (Vrolijk et al., 2002) naar voren zijn gekomen. Voor elk specifiek probleem dient een afweging gemaakt te worden welke methode het beste kan worden gebruikt. In dit hoofdstuk worden de volgende schattingstechnieken beschreven; de directe schatter, de ratioschatter, de regressie schatter, de Bayesiaanse schatter, poststratificatie, datafusie en imputatie.

Directe schatter
Het is gebruikelijk om schattingen van gemiddelden en totalen te maken op basis van de waarden van de variabele zoals die in de steekproef zijn waargenomen. Het gemiddelde is de directe schatter van het daadwerkelijke steekproefgemiddelde. De directe schatter is gelijk aan:

$$\bar{Y}_D = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$\bar{Y}_D$ is de directe schatter van $\bar{Y}$

[1] Dit hoofdstuk is gebaseerd op de LEI-rapportage 8.02.05: *Schatten van kenmerken van kleine deelgebieden* (Vrolijk et al., 2002).

n is het aantal steekproefelementen

y_i is de i-de waarneming in de steekproef

De variantie van deze schatter is gelijk aan:

$$v(\overline{Y}_D) = \left(\frac{1-f}{n}\right)\left[\frac{\sum_{i=1}^{n}(y_i - \overline{Y}_D)^2}{n-1}\right]$$

$v(\overline{Y}_D)$ is de variantie van de directe schatter

$(1-f)$ is gelijk aan $\frac{N-n}{N}$

N is het aantal elementen in de populatie

Ratioschatter

Indien een hulpvariabele beschikbaar is die in grote mate correleert met de doelvariabele, dan kan deze hulpvariabele worden gebruikt voor het maken van betrouwbaardere schattingen. Als men bijvoorbeeld een schatting wil maken van de totale melkproductie, kan gebruik worden gemaakt van het gegeven dat de melkproductie op een bedrijf sterk zal correleren met het aantal koeien op dat bedrijf. Bij het gebruik van de ratioschatter geldt wel de voorwaarde dat het gemiddelde of totaal van deze hulpvariabele voor de hele populatie bekend is. Voor het aantal koeien is dit het geval, op basis van de Landbouwtelling kan het totaal aantal koeien worden vastgesteld. De reden waarom deze indirecte schatting betrouwbaarder kan zijn dan een directe schatting is dat de verhouding tussen twee variabelen stabieler kan zijn dan de variabelen afzonderlijk. De melkproductie op verschillende boerderijen kan sterk uiteenlopen. Een directe schatting zou dan ook een hoge variantie laten zien. De melkproductie zal echter sterk afhankelijk zijn van het aantal koeien. De verhouding productie per koe zal een kleinere spreiding laten zien dan de spreiding in de melkproductie of het aantal koeien zelf over de steekproefbedrijven. Indien men op basis van andere bronnen gegevens heeft over het totaal aantal koeien dan kan een veel nauwkeurigere uitspraak over de totale melkproductie in Nederland worden gedaan.

Een bijkomend voordeel van het gebruik van ratioschatters is dat de representativiteit wordt verhoogd. Stel dat in de steekproef vooral kleine bedrijven zijn opgenomen. Doordat in de indirecte schatter van de totale melkproductie rekening wordt gehouden met het aantal koeien op de steekproefbedrijven ten opzichte van het aantal koeien in Nederland wordt automatische gecorrigeerd voor de omvang van de bedrijven. De verhouding melkproductie per koe wordt vermenigvuldigd met het uit een andere bron bekende aantal koeien. Een directe schatter zou in dit geval tot een onderschatting van de totale melkproductie leiden.

De ratioschatter wordt weergegeven door:

$$\overline{Y}_R = R\overline{X} = \frac{\overline{Y}_D}{\overline{X}_D}\overline{X}$$

$\overline{Y}_R$ is de ratioschatter van $\overline{Y}$

$\overline{X}_R$ is de ratioschatter van $\overline{X}$

$\overline{X}$ is het populatiegemiddelde van X

De variantie van de ratioschatter is gelijk aan:

$$v(\overline{Y}_R) = \left(\frac{1-f}{n}\right)\left[\frac{\sum_{i=1}^{n}(y_i - Rx_i)^2}{n-1}\right]$$

$v(\overline{Y}_R)$ is de variantie van de ratioschatter

Regressieschatter

Regressieschatters maken net als ratioschatters gebruik van extra informatie van een hulpvariabele die sterk is gecorreleerd met de doelvariabele. Wanneer er wel een verband bestaat tussen deze variabelen maar wanneer deze niet door de oorsprong gaat, kan beter gebruik worden gemaakt van een regressieschatter dan van een ratioschatter. Bij een relatie tussen aantal koeien en de melkproductie is het aannemelijk dat de relatie door de oorsprong gaat. Een veestapel van nul koeien zal immers leiden tot een melkproductie van nul liter. In andere situaties is de aanname van een verband door de oorsprong minder waarschijnlijk. Indien bijvoorbeeld een verband wordt verondersteld tussen het nettobedrijfsresultaat en het aantal koeien dan zal dit verband niet door de oorsprong gaan. Door de vaste lasten zal op een gespecialiseerd bedrijf een veestapel van nul koeien leiden tot een negatief bedrijfsresultaat. In dergelijke situaties is het gebruik van een regressieschatter aan te raden. Regressieschattingen komen in hoofdstuk 7 uitgebreid aan de orde. Voor de volledigheid wordt in dit hoofdstuk de regressieschatter ook weergegeven.

$$\overline{Y}_L = \overline{Y}_D + b_1(\overline{X} - \overline{X}_D) = b_0 + b_1\overline{X}$$

$\overline{Y}_L$ is de regressieschatter van $\overline{Y}$

b_1 is gelijk aan $\dfrac{\sum_{i=1}^{n}(y_i - \overline{Y}_D)(x_i - \overline{X}_D)}{\sum_{i=1}^{n}(x_i - \overline{X}_D)^2}$

De variantie van de ratioschatter is gelijk aan:

$$v(\overline{Y}_L) = \left(\frac{1-f}{n}\right)\left[\frac{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{n-1}\right]$$

$v(\overline{Y}_L)$ is de variantie van de regressieschatter

Bayesiaanse schatter

De Bayesiaanse schatter is een lineaire combinatie van de regressieschatter en de directe schatter. Als het verband tussen doel- en hulpvariabele niet alleen opgaat binnen een klein deelgebied, maar geldt voor de gehele populatie, verdient de Bayesiaanse schatter de voorkeur boven een directe of een regressieschatter. Bayesiaanse analyse maakt beter gebruik van de informatie die vooraf beschikbaar is over de te schatten grootheden dan bijvoorbeeld een directe schatter of een regressieschatter. De directe schatter gebruikt enkel de individuele eigenschappen van de doelvariabele binnen een bepaald klein deelgebied en de lineaire regressieschatter maakt enkel gebruik van de relatie tussen de doelvariabele met andere variabelen (verklarende variabelen) die sterke correlatie vertonen met de doelvariabele. De regressieschatter houdt geen rekening met het feit dat individuele eigenschappen kunnen gelden voor de doelvariabele binnen verschillende kleine deelgebieden die niet terugkomen in de verklarende variabele(n).

Voor onderzoek op het LEI zou deze techniek gebruikt kunnen worden wanneer onderzoeksvragen een bepaalde regio aangaan en verondersteld kan worden dat een bepaalde relatie tussen hulp- en doelvariabele voor het gehele land geldt. De Bayesiaanse schatter op het deelgebied h ziet er als volgt uit:

$\overline{Y}_{Bh} = B\overline{Y}_{Lh} + (1-B)\overline{Y}_{Dh}$

$\overline{Y}_{Bh}$ is de Bayesiaanse schatter van $\overline{Y}$ op deelgebied h

$\overline{Y}_{Lh}$ is de regressieschatter van $\overline{Y}_h$

$\overline{Y}_{Dh}$ is de directe schatter van $\overline{Y}_h$

B is de wegingsfactor (voor het bepalen van deze wegingsfactor wordt verwezen naar Vrolijk et al. (2002)

Poststratificatie

In het geval een dataset een groot aantal toepassingen heeft, dat wil zeggen dat een groot aantal variabelen als doelvariabele gebruikt wordt, heeft poststratificatie de voorkeur boven stratificatie vooraf (Sarndal, 1992). Bij een gestratificeerde steekproefopzet worden de strata definitief ingevoerd. Dit leidt tot een reductie in de variantie voor de daarbij gehanteerde doelvariabelen, de stratificatievariabelen. Deze opzet kan echter minder efficiënt zijn voor vele andere doelvariabelen. De combinatie van een aselecte steekproef en poststratificatie kan de totale efficiency verbeteren. Bij de analyse van gegevens kan gebruik worden gemaakt van de kennis en intuïtie van de onderzoeker om bij het onderzoek passende poststratificaties vast te stellen.

Stel dat in het type glastuinbouw twee typen productiesystemen bestaan die van grote invloed zijn op het energieverbruik. Indien men kennis heeft over de verdeling van deze systemen in de populatie (bijvoorbeeld op basis van de Landbouwtelling), dan kan men deze kennis gebruiken om een betere schatting te maken. Stel dat men weet dat 30% van de bedrijven productiesysteem A gebruikt en 70% systeem B. Omdat de steekproef niet is gestratificeerd op basis van dit kenmerk, kan het voorkomen dat in de steekproef 50% van de bedrijven systeem A en 50% systeem B gebruikt. In een onderzoek naar het energieverbruik kan het zinvol zijn te corrigeren voor deze verhouding. Poststratificatie leidt ertoe dat het gewicht van bedrijven met systeem A iets lager wordt (bedrijven met systeem A zijn oververtegenwoordigd in de steekproef) en bedrijven met systeem B iets hoger wordt (bedrijven met systeem B zijn ondervertegenwoordigd) bij het maken van schattingen omtrent het energieverbruik.

De poststratificatieschatter ziet er als volgt uit:

$$\bar{Y}_P = \sum_{h=1}^{H} W_h \bar{Y}_{Dh}$$

$\bar{Y}_P$ is de poststratificatieschatter van $\bar{Y}$

W_h $\frac{N_h}{N}$ dit is het gewicht dat toegekend wordt aan groep h. Deze wordt vastgesteld aan de hand van bekende gegevens (zoals verhoudingen van groepen ten opzichte van elkaar) uit de populatie. Let op: deze definitie van gewicht verschilt van de definitie die in paragraaf 3.2 wordt gehanteerd.

De variantie van de poststratificatieschatter is gelijk aan:

$$v(\bar{Y}_P) = \left(\frac{1-f}{n}\right)\sum_{h=1}^{H} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{H}(1-W_h)s_h^2$$

s_h^2 is gelijk aan $\frac{\sum_{i=1}^{n}(y_{ih} - \bar{Y}_{Dh})^2}{n-1}$

Datafusie en imputatie

Datafusie is een methode om gegevens vanuit verschillende bronnen te integreren en samen te voegen. Binnen het LEI kunnen op die manier Informatienet- en Landbouwtellingsgegevens worden geïntegreerd. De kenmerken in de Landbouwtelling zijn bekend voor alle agrarische bedrijven groter dan circa 3 nge. Daarnaast is in het Informatienet een gedetailleerde administratie beschikbaar van een kleine 1.500 bedrijven. Voor het overgrote deel van de bedrijven in de Landbouwtelling is deze gedetailleerde administratie niet beschikbaar. Om toch uitspraken te kunnen doen over kenmerken die gelden voor de populatie op het kleine deelgebied, gaat men op zoek naar bedrijven (waarvan wel een administratie beschikbaar is) welke op basis van kenmerken in de Landbouwtelling sterk op het bedrijf lijken waarover men een uitspraak wil doen. Een bedrijf dat sterk op een ander bedrijf lijkt op basis van de beschikbare variabelen in de Landbouwtelling zal naar alle waarschijnlijkheid ook lijken op dat andere be-

drijf voor variabelen die niet beschikbaar zijn, ervan uitgegaan dat de beschikbare en de niet-beschikbare variabelen in grote mate met elkaar gecorreleerd zijn.

De methode kan bijvoorbeeld worden toegepast indien men een uitspraak wil doen over een regio waarvoor men over weinig directe waarnemingen beschikt. In een regio zullen bedrijven zitten van verschillende typen. Om alle typen afzonderlijk te schatten zijn veel waarnemingen nodig. Door middel van datafusie en imputatie gaat men op zoek naar bedrijven die een grote gelijkenis vertonen met de bedrijven in de te bestuderen regio. Men zoekt voor elk bedrijf in de regio naar een bedrijf in het Informatienet dat er sterk op lijkt gegeven de kenmerken in de Landbouwtelling. Vervolgens wordt de veronderstelling gemaakt dat de te schatten kenmerken van het bedrijf ook hetzelfde zullen zijn, ervan uitgegaan dat de gebruikte kenmerken in de Landbouwtelling gecorreleerd zijn met de kenmerken uit het Informatienet. De gegevens van het gelijkende bedrijf in het Informatienet worden dus van toepassing verklaard op het bedrijf in de te bestuderen regio waar men deze gegevens niet direct heeft waargenomen. Op basis van deze (geïmputeerde) gegevens kunnen vervolgens bepaalde statistieken voor de regio worden berekend.

Evaluatie schattingstechnieken

Methode Criteria	Directe schatter	Ratio-Schatter	Regressie-Schatter	Bayesiaanse schatter	Poststratificatie	Fusie
Betrouwbaarheid te berekenen	++	++	++	-	+	--
Betrouwbaarheid bij kleine aantallen	-	+	+	Nvt	-	Nvt
Zuiverheid	++	-/+	-/+	-/+	+	--
Indicatie GFI aannames	Nvt	+	+	+	Nvt	-
Validiteit bij kleine aantallen	-	+	++	++	++	??
Onderbouwing	Steek-Proeven	Model en Steekproef	Model en Steekproef	Model en steekproef	Steek-proeven	Afstands-maten
Eenvoud	++	-	--	--	-	+
Bewerkelijkheid	++	-	--	--	-	+
Flexibiliteit	++	-	-	-	-	++
Wetenschappelijke acceptatie	++	-/+	-/+	-	+	-
Meerdere doelvariabelen	+	-	-	-	+	++
Gebruik extra info	Geen	Gemiddelde of totaal van hulpvariabele	Gemiddelde van hulpvariabele	Gemiddelde van hulpvariabele en directe schatter op alle deelgebieden	Verdeling in de populatie	Kenmerken in populatie
Meerdere hulpvariabelen	Nvt	-	+	+	+	+
Nominale of ordinale hulpvar	Nvt	-	-	-	++	++
Interval of ratio hulpvar	Nvt	++	++	++	-	++
Reproduceer-baarheid	++	+	+	+	+	-

Figuur 6.1 Evaluatie van de beschreven methoden

6.3 Literatuur

De aanbevolen literatuur op het gebied van verbeteren van steekproeftechnieken is:
- Dol, W. (1991).
- Vrolijk, H.C.J., W. Dol en G. Cotteleer (2002).

Het stuk dat het beste leesbaar is op het gebied van kleine deelgebieden is Vrolijk et al. (2002). De theorie wordt in dit rapport toegelicht aan de hand van voorbeelden. Dit maakt het stuk makkelijk leesbaar en begrijpelijk. Het proefschrift van Dol (1991) gaat veel dieper in op de theorie en is moeilijker leesbaar voor beginners, maar aan te raden voor mensen die hun theoretische kennis willen verdiepen.

7. Regressieanalyse

7.1 Inleiding

Regressieanalyse is een van de meest gebruikte technieken in kwantitatief onderzoek. Met regressieanalyse probeert men de waargenomen spreiding in een (meetbare) afhankelijke variabele te verklaren met behulp van onafhankelijke verklarende variabelen. Er zijn drie redenen waarom je een regressieanalyse uit kunt voeren. Ten eerste kun je geïnteresseerd zijn in de samenhang tussen de afhankelijke variabele en één of meer verklarende variabelen. Hoe hangen bijvoorbeeld tarweopbrengsten op individuele bedrijven samen met gebruik van meststoffen, gewasbeschermingsmiddelen en het weer. Of: wat is de relatie tussen het aandeel biologische landbouw in Nederland en prijzen van biologische producten, overheidssteun en het plaatsvinden van voedselcrises? Een tweede reden om een regressieanalyse uit te voeren is het voorspellen van een nog niet geobserveerde afhankelijke variabele met behulp van al wel geobserveerde verklarende variabelen. Tot slot kun je met regressieanalyse ook nagaan in hoeverre de afhankelijke variabele verandert als de verklarende variabelen veranderen.

Dit hoofdstuk is als volgt opgebouwd. In paragraaf 7.2 wordt de relevante theorie van regressieanalyse besproken. Onderwerpen als specificatie en aannames, schatten, kwaliteit van het model en toetsen van hypothesen komen hier aan de orde. In paragraaf 7.3 wordt een casestudie uitgewerkt, waarbij uitgelegd wordt hoe je met SPSS een regressieanalyse kunt uitvoeren. Paragraaf 7.4 ten slotte bespreekt bruikbare achtergrondliteratuur.

7.2 Theorie

In deze paragraaf wordt in een notendop de theorie van regressieanalyse besproken. De volgende onderdelen komen achtereenvolgens in de paragrafen 7.2.1 tot 7.2.7 aan de orde: modelspecificatie en aannames, schatten, kwaliteit van het geschatte model, toetsen van hypothesen, voorspellen, aannames die niet kloppen en uitbreidingen op het basis model.

7.2.1 Modelspecificatie en aannames

Voordat je met regressieanalyse aan de slag gaat, is het goed eerst helder te krijgen wat de doelstelling van je onderzoek is en welke specifieke onderzoeksvragen je wilt beantwoorden. Wanneer dat duidelijk is kun je beter beoordelen of regressieanalyse een bruikbare methode is voor je specifieke onderzoek. Let daarbij op de hierboven genoemde redenen om een regressieanalyse uit te voeren.

Vervolgens moet je je afvragen welke gegevens beschikbaar zijn. Kun je met deze gegevens een regressieanalyse uit voeren? Hoe groot is het aantal waarnemingen? Wat is de aard van de gegevens? Betreft het cross-sectie data (aantal bedrijven, huishoudens of individuen met maar één waarneming op hetzelfde moment verricht), tijdreeksdata (meerdere waarne-

mingen in de tijd voor een eenheid zoals een prijsreeks; zie hoofdstuk 7) of panel data (meerdere bedrijven, huishoudens met meerdere waarnemingen in de tijd). Elk type data heeft zijn specifieke voor- en nadelen en dient op een eigen wijze behandeld te worden. Belangrijk is dat de veronderstelde afhankelijke variabele werkelijk afhankelijk is van de onafhankelijke variabelen. Onafhankelijke variabelen dienen echter onafhankelijk te zijn van de afhankelijke variabele. Om dit vooraf te bepalen kun je gebruiken maken van bestaande theorie of van boerenwijsheid. Let hierbij ook op mogelijke simultaniteit. Sommige variabelen worden namelijk tegelijkertijd bepaald en dan is het moeilijk om onderscheid te maken tussen afhankelijke en onafhankelijke variabelen. Denk hierbij aan de relatie tussen evenwichtsprijzen en -hoeveelheden in een vrije markt of aan de relatie tussen de productie van varkens en het gebruik van varkensvoer. Ten slotte mogen er geen onderliggende relaties tussen verklarende variabelen bestaan. Nadat je de afhankelijke en onafhankelijke variabelen hebt bepaald stel je een model op. In algemene termen ziet het meervoudige lineaire regressiemodel er als volgt uit:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

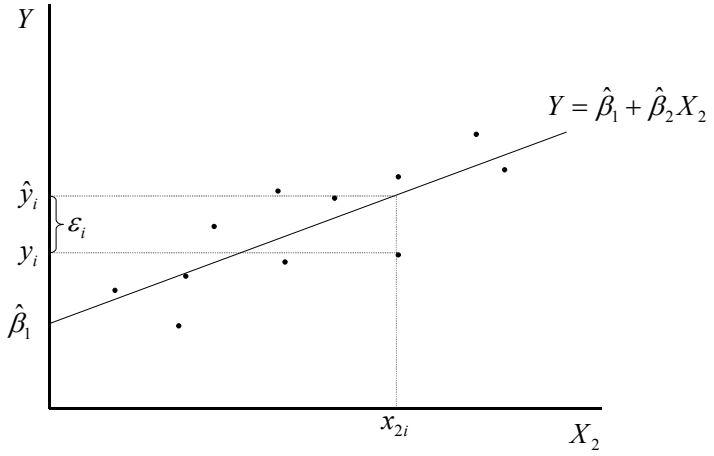
Y_i is de afhankelijke, endogene of te verklaren variabele
X_{ji} is de verklarende of exogene variabele
β_j is parameter j (j = 1..k)
ε_i is een storingsterm
i is de indicator die de waarnemingen aanduidt (i =1..N)
j is de indicator die de variabelen aanduidt (j = 1..k)

De parameter β_1 is een interceptparameter en de overige parameters β_j zijn hellingsparameters. De storingsterm ε_i wordt opgenomen om verschillen tussen waargenomen waarden en modelwaarden voor Y_i op te nemen. Dit is nodig omdat een model nooit perfect in staat is de waargenomen waarden te verklaren. Een verkorte manier van opschrijven is mogelijk met een matrixnotatie, waarbij data en parameters worden samengevat in vectoren en matrices. Om vectoren en matrices te kunnen herkennen zijn zij altijd vet gedrukt.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Voor een eenvoudig model met één verklarende variabele en een intercept is het mogelijk het model in figuur 7.1 weer te geven.

De stippen geven de werkelijk waargenomen waarden aan voor Y en X_2. De lijn geeft de geschatte vergelijking weer, waarbij $\hat{\beta}_1$ het geschatte intercept is, en $\hat{\beta}_2$ de geschatte hellingsparameter. Het intercept kan men beschouwen als een basisniveau voor Y_i (als $X_2 = 0$) en de hellingsparameter geeft weer hoeveel Y_i verandert als de waarde van X_2 verandert. Het verschil tussen een werkelijke waarde voor Y_i en een voorspelde waarde op basis van het model wordt



Figuur 7.1 Eenvoudig lineair model met 1 verklarende variabele (X2)

weergegeven met εi. Er geldt een aantal aannames voor het correct uitvoeren van een regressieanalyse:

- samenhang tussen *Y* en *X* is lineair en gegeven zoals in het model.
  Let er echter op dat niet-lineaire relaties vaak getransformeerd kunnen worden. Voorbeelden hiervan zijn:

  $Y_i = B \cdot X_{2i}^{\beta_2} X_{3i}^{\beta_3} \varepsilon_i$ wordt $\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \ln \varepsilon_i$

  $Y_i = \exp\left[\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}\right] \varepsilon_i$ wordt $\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ln \varepsilon_i$

  $Y_i = \left(1/\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i\right)$ wordt $1/Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

  Ook het volgende voorbeeld is lineair in de parameters:
  $Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 X_{3i} X_{4i} + \beta_4 \left(X_{5i} / X_{6i}\right) + \varepsilon_i$;
- er bestaat geen exacte lineaire relatie tussen bepaalde verklarende variabelen *X*. Dus, $X_{2i} = aX_{3i} + bX_{4i}$ is fout omdat er geen storingsterm in het model is opgenomen;
  de storingstermen zijn normaal verdeeld met gemiddelde 0 en een constante variantie σ^2: $\left(\varepsilon_i \sim N\left(0, \sigma^2\right)\right)$;
- de verklarende variabelen *X* zijn niet gecorreleerd met de storingstermen εi;
- de storingstermen εi zijn onderling niet gecorreleerd.

In paragraaf 7.2.6 worden verschillende toetsen beschreven waarmee kan worden nagegaan of voldaan is aan de bovenstaande voorwaarden. Daarnaast worden in deze paragraaf ook oplossingen aangedragen voor problemen die zich voordoen als niet wordt voldaan aan één of meerdere aannames.

7.2.2 Schatten

Om het lineaire model te schatten wordt doorgaans gebruikt gemaakt van de gewone kleinste kwadraten methode, ook wel Ordinary Least Squares (OLS) genoemd. Daarbij gaat de aandacht uit naar de storingstermen

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{\beta X}$$

Deze willen we zo klein mogelijk hebben. OLS kiest een set parameters die de som van deze storingstermen in het kwadraat minimaliseert. Daartoe wordt bovenstaande uitdrukking voor de storingstermen in het kwadraat genomen en een minimum gezocht met behulp van de eerste orde afgeleiden. Los deze op voor de parameters. De oplossing in matrixvorm is:

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

De werkelijke parameter β_j (in de totale populatie) is een constante. De parameter schatting $\hat{\beta}_j$ is een trekking uit de kansverdeling van mogelijke schattingen. Als de hierboven gemaakte aannames 1-5 gelden, geeft OLS zuivere (unbiased) en efficiënte schattingen. Zuiver houdt in dat je mag verwachten dat de schatting en de werkelijke waarde gelijk zijn. Efficiënt betekent dat de schatter de kleinst mogelijke variantie heeft. Dit is namelijk belangrijk bij het toetsen van hypothesen.
De variantie van de schattingen $\hat{\beta}_j$ is gegeven in de matrixnotatie:

$$\text{var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

De standaardfout van een geschatte parameter $\hat{\beta}_j$ is:

$$se(\hat{\beta}_j) = \sqrt{\text{var}(\hat{\beta}_j)}$$

Een zuivere schatting voor de variantie van de storingstermen is:

$$s^2 = \frac{\sum \hat{\varepsilon}_i^2}{N-k} \qquad \left(= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{N-k}\right)$$

7.2.3 Kwaliteit van het geschatte model

Om de kwaliteit van het geschatte model te bepalen, kijken we in hoeverre de spreiding (variatie) in de afhankelijke variabele *Y* verklaard wordt door het model. Daarbij geldt dat de spreiding van Y gelijk is aan de in het model verklaarde spreiding en de niet in het model ver-

klaarde spreiding. Of in wiskundige termen:

$$\underbrace{\sum\left(Y_i-\bar{Y}\right)^2}_{\text{Total Sum of Squares (TSS)}} = \underbrace{\sum\left(\hat{Y}_i-\bar{Y}\right)^2}_{\text{Explained Sum of Squares (ESS)}} + \underbrace{\sum\left(Y_i-\hat{Y}_i\right)^2}_{\text{Residual Sum of Squares (RSS)}}$$

Let erop dat sommige tekstboeken (bijvoorbeeld Pindyck en Rubinfeld, 1998) ESS en RSS omgedraaid worden (Regression Sum of Squares en Error Sum of Squares). De hier gebruikte afkortingen zijn echter het meest gebruikelijk. Een maat voor de kwaliteit van het model is nu de fractie door het model verklaarde variatie:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum\left(\hat{Y}_i-\bar{Y}\right)^2}{\sum\left(Y_i-\bar{Y}\right)^2} = 1-\frac{\sum\hat{\varepsilon}_i^2}{\sum\left(Y_i-\bar{Y}\right)^2} = 1-\frac{RSS}{TSS}$$

De R^2 ligt altijd tussen 0 en 1 en geeft een indicatie van de kwaliteit van het model. Bij een waarde 0 verklaart het model helemaal niets en 1 geeft aan dat het model perfect in staat is om Y te verklaren. Hierbij hebben we wel aangenomen dat het model goed is gespecificeerd.

Als dat niet het geval is, om wat voor redenen dan ook, geeft deze maat een vertekend beeld. Een ander probleem is dat R^2 nooit afneemt als we variabelen blijven toevoegen aan het model. Het aantal verklarende variabelen (relevant en niet-relevant) in het model speelt dus geen rol. Om deze tekortkoming op te lossen kun je ook de voor vrijheidsgraden aangepaste R^2 gebruiken. Deze legt een straf op aan het toevoegen van variabelen die weinig verklarende waarde hebben in het model:

$$\bar{R}^2 = 1-\frac{RSS/(N-k)}{TSS/(N-1)}$$

Andere maten die iets zeggen over de kwaliteit van het model en die ook een straf opleggen aan het toevoegen van niet relevante verklarende variabelen zijn het Schwartz-criterium,

$$SC = \ln\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{N}\right)+\frac{k}{N}\ln N$$

en het Akaike-informatiecriterium

$$AIC = \ln\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{N}\right)+\frac{2k}{N}$$

Voor deze maten geldt: hoe kleiner hoe beter.

7.2.4 Toetsen van hypothesen

Er zijn 2 hoofdredenen om hypothesen te toetsen. De eerste reden is meer technisch van aard en betreft het beoordelen van de kwaliteit van de geschatte parameters. Vragen die hierbij een rol spelen zijn:
- verschilt een parameter significant van nul;
- zijn alle (helling)parameters gelijk aan nul;
- krijg je dezelfde schattingen als je slechts een deel van de waarnemingen gebruikt?

De andere hoofdreden (die overigens nauw samenhangt met de eerste) betreft het beantwoorden van specifieke onderzoeksvragen. Deze worden beantwoordt door toetsten uit te voeren op parameters die onze interesse hebben. Voorbeelden zijn:
- heeft de rentestand een significant effect op de investeringen;
- is de vraag naar biologische producten prijs-inelastisch;
- wordt akkerbouwproductie gekenmerkt door constante schaalopbrengsten?

Bij het toetsen van hypothesen stel je eerst de hypothese op die je gaat toetsen. Dit noemen we de nulhypothese (H_0). Voorbeelden zijn:

$H_0 : \beta_2 = 1$ of

$H_0 : \beta_2 + \beta_3 + \beta_4 = 1$

Daarnaast is er de alternatieve hypothese H_1 die aangeeft dat H_0 niet waar is in strikte zin. Je kunt hypothesen met betrekking tot individuele parameters toetsen (*t*-toets) of je kunt toetsen uitvoeren voor meerdere parameters (*F*-toets). Omdat beide veelvuldig voorkomen en omdat er verschillende manieren zijn om de toetsten uit te voeren worden beiden hieronder besproken.

Toetsen voor individuele parameters.

De te toetsten hypothese in een model $Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$ in algemene zin is:

$H_0 : \quad \beta_j = \beta_j^0$,waar β_j^0 de toetswaarde aangeeft (bijvoorbeeld 0 of 2,6).

H_1 : Niet H_0

In principe zijn er 3 gerelateerde manieren om de toets uit te voeren, namelijk:
- door middel van een betrouwbaarheidsinterval;
- afgeleide vuistregel op basis van t-waarde;
- p-waarde (overschrijdingskans).

ad 1

Een betrouwbaarheidsinterval is een gebied waarin de werkelijke waarde met xx% zekerheid (bijvoorbeeld 95 of 99%) ligt. Betrouwbaarheidsintervallen voor individuele parameters zijn gebaseerd op de *t*-verdeling (omdat de geschatte variantie van de residuen gebruikt wordt; zie hoofdstuk 5). Vervolgens volgen we een drietal stappen.

- Bereken de teststatistiek (gestandaardiseerde t):

$$t_{N-k} = \frac{\left(\hat{\beta}_j - \beta_j^0\right)}{s_{\hat{\beta}_j}}$$

N is het aantal waarnemingen
k is aantal parameters
N-k is aantal vrijheidsgraden
$\hat{\beta}_j$ is de geschatte waarde voor parameter β
$s_{\hat{\beta}_i}$ is de bijbehorende standaardfout

- Kies een significantieniveau. Met hoeveel procent zekerheid wil je een uitspraak doen (95 of 99%)? Omdat t_{N-k} een t-verdeling heeft met N-k vrijheidsgraden, zoek je een kritieke waarde t_c met dit aantal vrijheidsgraden en het gewenste significantieniveau op in een tabel voor de t-verdeling (zie paragraaf 5.2.2). Let hierbij op of de toets eenzijdig of tweezijdig is. Bij een tweezijdige toets kan de parameter groter of kleiner dan de testwaarde zijn. Bij een eenzijdige toets is de parameterwaarde groter of kleiner dan de testwaarde (bijvoorbeeld $H_0: \quad \beta_j \leq \beta_j^0$ versus $H_1: \quad \beta_j > \beta_j^0$).
Voorbeelden van kritieke waarden zijn:
t_{10} (0,05) = 2,228 (2-zijdig)
t_{10} (0,01) = 3,169 (2-zijdig)
t_{10} (0,01) = 2,764 (1-zijdig)
t_{15} (0,10) = 1,753 (2-zijdig)
t_{15} (0,05) = 1,753 (1-zijdig)

- Stel nu het betrouwbaarheidsinterval op:

$$P\left(-t_c < t_{N-k} < t_c\right) = 0{,}95$$

Hierbij geeft P(a) de kans weer op gebeurtenis a. In de bovenstaande formule staat de kans op de gebeurtenis dat de teststatistiek tussen twee kritieke waarden ligt gelijk is aan 0,95. Dit kan ook anders omschreven worden:

$$P\left(-t_c < \frac{\hat{\beta}_j - \beta_j^0}{s_{\hat{\beta}_j}} < t_c\right) = 0{,}95$$

of

$$P\left(\hat{\beta}_j - t_{N-k;0.025} \cdot s_{\hat{\beta}_j} < \beta_j^0 < \hat{\beta}_j + t_{N-k;0.025} \cdot s_{\hat{\beta}_j}\right) = 0{,}95$$

H_0 wordt niet verworpen voor alle waarden β_0 die in het betrouwbaarheidsinterval liggen. Ligt een waarde β_0 niet in het betrouwbaarheidsinterval, dan verwerpen we H_0 en is met 95% zekerheid vastgesteld dat de parameter niet gelijk is aan de testwaarde.

Een eenvoudige weergave van het betrouwbaarheidsinterval is:

$$\hat{\beta}_j \pm t_{N-k;0.025} \cdot s_{\hat{\beta}_j}$$

ad 2

In plaats van een betrouwbaardheidsinterval op te stellen wordt vaak gebruikgemaakt van *t*-waarden om de hypothese: $\beta_j = 0$ te testen. Deze t-waarden worden doorgaans door het gebruikte schattingsprogramma (SPSS, SAS, E-views, LIMDEP) gegeven. Als de gestandaardiseerde *t* (test waarde) > t_c wordt de nulhypothese dat β_j gelijk is aan 0 verworpen. Zie de analogie met het betrouwbaarheidsinterval hierboven.

ad 3

Voor het betrouwbaarheidsinterval kies je een bepaalde zekerheid waarmee je een uitspraak wilt kunnen doen voor de uitspraak dat de parameter gelijk is aan een bepaalde waarde (bijvoorbeeld 95 of 99%). Hoe meer zekerheid, hoe groter het betrouwbaarheidsinterval, hoe moeilijker de nulhypothese verworpen kan worden. De overschrijdingskans (*p*-waarde) is de kans op een waarde voor de teststatistiek die net zo extreem of nog extremer is dan de berekende waarde. Met andere woorden hoe groter de *p*-waarde, hoe moeilijker we een nulhypothese kunnen verwerpen. Wanneer een gegeven *p*-waarde kleiner is dan een kritische waarde (bijvoorbeeld 1-0,95 of 1-0,99) wordt de nulhypothese verworpen.

Algemeen toetsprincipe gebaseerd op de F-toets

Met behulp van de som van de gekwadrateerde storingstermen (RSS) kunnen we een algemene *F*-toets uitvoeren. Deze toets heeft als voordelen dat ze eenvoudig is, algemeen toepasbaar is (toetsen voor 1 of meerdere parameters) en ook gebruikt kan worden voor bijvoorbeeld toetsen op structurele verandering. Het principe van deze toets is dat een nulhypothese niet verworpen wordt als er een gering verschil is tussen de *RSS* van het niet-gerestricteerde model (RSS_U) en de *RSS* van het gerestricteerde model (RSS_R). Wanneer dat namelijk het geval is, heeft de restrictie weinig gevolgen voor de kwaliteit van het model. De test wordt uitgevoerd in een aantal stappen:

- schat een model zonder restricties en bereken RSS_U;
- schat het model opnieuw, met te testen restrictie(s) opgelegd in het model. Bereken RSS_R;
- stel toetsingsgrootheid op:

  $$F_{test} = \frac{(RSS_R - RSS_U)/q}{RSS_U/(N-k)}$$

  q is het aantal opgelegde restricties;
- als H_0 waar is dan heeft F_{test} een F-verdeling met q en N-k vrijheidsgraden;
- als $F_{test} > F_{q,N-k;0,05}$ verwerp je de nulhypothese en kan je concluderen dat het gerestricteerde model significant verschilt van het ongerestricteerde model.

Een nadeel van de hierboven besproken *F*-toets is dat een model twee keer geschat moet worden. Eén keer zonder restricties en één keer met de restricties. Dit hoeft bijvoorbeeld niet

als gebruik wordt gemaakt van een *t*-toets. Aan de andere kant worden schattingen doorgaans uitgevoerd met behulp van een computer en een softwarepakket waardoor dit nadeel grotendeels opgeheven wordt. Het opleggen van de restrictie(s) in het gerestricteerde model vergt soms wel de nodige creativiteit.

7.2.5 Voorspellen

De OLS-methode geeft op basis van n waarnemingen in de steekproef een set van parameterschattingen (weergegeven met de vector $\hat{\beta}$) en een schatting voor de variantie van de storingstermen, s^2. Vervolgens kan met behulp van de geschatte parameters en waarden voor de verklarende variabelen voorspellingen gedaan worden. Hierbij geeft j aan om welke voorspelling het gaat ($j = n+1.. N$). De beste voorspeller voor y_j is:

$$\hat{y}_j = \boldsymbol{x}_j \hat{\boldsymbol{\beta}}$$

De voorspelfout is gelijk aan:

$$y_j - \hat{y}_j = y_j - \boldsymbol{x}_j \hat{\beta}$$

Deze voorspelfout is normaal verdeeld met verwachting 0 en variantie:

$$\text{var}\left(y_j - \boldsymbol{x}_j \hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(1 + \boldsymbol{x}_j \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_j\right)$$

We kunnen ook voorspellingsintervallen opstellen. Een 95%-voorspellingsinterval voor y_{N+1}:

$$\boldsymbol{x}_j \hat{\boldsymbol{\beta}} \pm t_{N-k;0.975} \cdot s\sqrt{1 + \boldsymbol{x}_j \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_j}$$

Met een kans van 95% ligt de werkelijke waarde voor y_j in dit gebied.

7.2.6 Wat als de aannames niet kloppen?

Vooraf is een vijftal aannames gemaakt met betrekking tot het te schatten model. Het kan echter zijn dat aan deze aannames niet voldaan wordt. Zo kan het zijn dat de aannames met betrekking tot de storingstermen ε_i of de verklarende variabelen X niet gelden. Ook kan het zijn dat er problemen zijn met de te schatten parameters zelf (deze zijn bijvoorbeeld niet constant over een bepaalde periode). Op dit laatste probleem wordt ingegaan in hoofdstuk 8. Hier wordt een aantal problemen met betrekking tot ε_i en X besproken.

Heteroscedasticiteit
In aanname 3 (paragraaf 7.2.1) werd verondersteld dat de storingstermen εi een constante variantie hebben. In geval van cross-sectie gegevens kan het echter voorkomen dat de variantie niet constant is maar varieert met waarnemingen. Grote bedrijven hebben bijvoorbeeld een grote variantie en kleine bedrijven een kleine variantie. In dat geval is er sprake van hete-

roscedasticiteit. Het gevolg van heteroscedasticiteit is dat de OLS-schattingen nog wel zuiver zijn, maar niet meer efficiënt. Dit betekent dat de standaardfouten van de parameters niet correct zijn en dat geeft uiteraard problemen bij het toetsen van hypothesen. Of er sprake is van heteroscedasticiteit kun je toetsen met een White test. Deze werkt als volgt:

- schat het model en bereken alle $\hat{\varepsilon}_i$;
- voer een regressie uit van $\hat{\varepsilon}_i^2$ op een constante, variabelen en gekwadrateerde variabelen. Bereken vervolgens de R^2;
- onder de nulhupothese van homoscedasticiteit geldt dat $N \cdot R^2 \sim \chi^2(l)$, waar l het aantal variabelen in stap twee is (de constante wordt hierbij niet meegerekend);
- als $N \cdot R^2 > \chi^2(l)$ verwerp je de nulhypothese van homoscedasticiteit. Corrigeren voor heteroscedasticiteit doe je met behulp van Weighted Least Squares (WLS). Vaak bestaat er een relatie tussen de variantie van de storingstermen en een bepaalde verklarende variabele: $\sigma_i^2 = \sigma^2 \cdot x_{ji}$. Bij WLS deel je alle variabelen door de wortel van de variabele die de heteroscedasticiteit veroorzaakt en pas je vervolgens OLS op de getransformeerde data toe.

Autocorrelatie

In aanname 5 (paragraaf 7.2.1) werd verondersteld dat de storingstermen ε_i onderling niet gecorreleerd zijn. Bij tijdreeksgegevens is dat echter vaak wel het geval als bepaalde ontwikkelingen over tijd niet door het model worden verklaard en deze logischerwijs in de storingsterm terechtkomen. Wanneer storingstermen onderling gecorreleerd zijn noemt men dit autocorrelatie. Het gevolg is opnieuw dat de OLS-schattingen nog wel zuiver zijn, maar niet meer efficiënt. Toetsen op autocorrelatie kan met behulp van de de Durbin-Watson (DW-) toets. De toetsingsgrootheid voor deze test is:

$$DW = \frac{\sum_{t=2}^{N} \left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{N} \hat{\varepsilon}_t^2}$$

Kritieke waarden voor deze test vindt men in speciale DW-tabellen. In het algemeen geldt dat $DW < (>)\ 2$ voor positieve (negatieve) autocorrelatie en dat $DW \approx 2$ duidt op geen autocorrelatie. Corrigeren voor autocorrelatie kan met behulp van de Generalised Least Squares (GLS)-methode. Hiervoor is echter wel een schatting van de covariantiematrix nodig.

Misspecificatie

Ook met de verklarende variabelen X kan het een en ander mis zijn:

- relevante variabelen ontbreken in het model;
- er zijn niet-relevante variabelen opgenomen in het model;
- er is een verkeerde functievorm verondersteld (lineair, lineair met interactie-termen, in logs, niet-lineair of……).

Een algemene specificatietest voor deze problemen is de Ramsey RESET test. Deze test kan uitgevoerd worden in 2 stappen:

- schat een model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$ waarbij Z een vector is met machten van $\hat{\boldsymbol{y}}$ (Bijvoorbeeld $\boldsymbol{Z} = \left[\hat{\boldsymbol{y}}^2 \ \hat{\boldsymbol{y}}^3 \ \hat{\boldsymbol{y}}^4\right]$);
- test of alle parameters α_j tezamen gelijk zijn aan 0 met behulp van een F-test. Indien dit het geval, is dan kan de nulhypothese (geen misspecificatie) niet verworpen worden.

Multicollineariteit
Er zijn echter ook andere problemen mogelijk met de verklarende variabelen X. Zo kunnen ze onderling (lineair) afhankelijk zijn (in dat geval is aanname 2 uit paragraaf 7.2.1 ongeldig). In dat geval werkt de OLS-methode niet. Wanneer variabelen bijna afhankelijk zijn is er sprake van multicollineariteit. Dit uit zich in hoge standaardfouten. Een eenvoudige controle is om de partiële correlatiecoëfficiënten van de verklarende variabelen te onderzoeken op hoge waarden.

Simultaniteit
Wanneer verklarende variabelen gecorreleerd zijn met storingstermen is er ook een probleem (aanname 4 uit paragraaf 7.2.1 geldt dan niet). Mogelijke oorzaken hiervan zijn meetfouten in data of simultaniteit van variabelen (evenwichtsprijs en -hoeveelheid in een vrije markt). Het gevolg is dat schattingen niet meer zuiver zijn. Dit kan opgelost worden met de methode van Instrument Variabelen (IV) of Two Stage Least Squares (2SLS).

7.2.7 Uitbreidingen van het basismodel

Op het hier besproken basismodel zijn vele uitbreidingen mogelijk. Zo is het mogelijk om een stelsel van meerdere lineaire vergelijkingen te schatten. Bijvoorbeeld een stelsel van vraagvergelijkingen naar inputs van individuele landbouwbedrijven. Het is ook mogelijk om vergelijkingen te schatten voor variabelen die in onderlinge samenhang (simultaan) bepaald worden. Wanneer de afhankelijke variabele slecht een beperkt range van mogelijke waarnemingen heeft (bijvoorbeeld alleen 0 en 1) zijn er andere methoden bruikbaar (zie hiervoor hoofdstuk 8). Een andere mogelijke uitbreidingsrichting is het in acht nemen van de eigenschappen van de data. Paneldata, cross-sectie data en tijdreeks data hebben hun eigen voor- en nadelen die in acht genomen moeten worden. In hoofdstuk 7 wordt met name aandacht geschonken aan tijdreeksgegevens. Zie verder ook de aanbevolen literatuur in paragraaf 5.4. Voor de aanpak van panel data wordt verwezen naar Reinhard et al. (2001).

7.3 Casestudie

In een studie van Polman et al. (1999) zijn betaalde grondprijzen onderzocht voor Nederlandse melkveehouders. Een dataset met waarnemingen voor 222 melkveehouders over de periode 1992-1995 (in totaal 286 waarnemingen) is beschikbaar via de intranetsite van het Domeinteam Data en Modellen. De volgende variabelen zijn aanwezig:

year	is het jaar van observatie
num	is het bedrijfsnummer (1 tot en met 222)
p	is de grondprijs per hectare (in 10.000 Euro's)
w	is de korte termijn winst (saldo) per hectare (in 10.000 Euro's)
ehs	is de regio variabele, percentage van grond in de regio in ecologische hoofdstructuur
vinex	is de regio variabele, percentage van grond in regio bestemd als VINEX locatie

1. Start het programma SPSS op en open de datafile (*File, Open ▸, Data... ; selecteer 'landdat.sav'* vanuit de juiste directory in het geopende window).

2. Voordat we een regressieanalyse uit gaan voeren is het goed om eerst een globaal beeld van de gegevens te krijgen. Met behulp van Analyze, Descriptive Statistics ▸, Descriptives... wordt een window geopend waarin een aantal variabelen geselecteerd kan worden. Selecteer *p*, *w*, *ehs* en *vinex* en bereken het gemiddelde, de standaardafwijking, de minimumwaarde en de maximumwaarde van elke variabele door bij Options... de juiste instellingen te kiezen. SPSS opent nu een Output window met een tabel waarin de gevraagde gegevens staan.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
P	286	.5899	4.3382	1.868869	.677372
W	286	.0900	.9782	.343060	.124035
EHS	286	.9900	17.2930	6.855982	3.889371
VINEX	286	.0028	14.3245	1.135037	1.976067
Valid N (listwise)	286				

Met de beschikbare data is het volgende model opgesteld om de grondprijs per hectare (*p*) te verklaren door middel van de variabelen *w*, *ehs* en *vinex* en een constante:

$$p_i = \beta_1 + \beta_2 w_i + \beta_3 ehs_i + \beta_4 vinex_i + \varepsilon_i \text{ t}$$

3. Doe een regressieanalyse in SPSS om de parameters van dit model te schatten. Kies Analyze, Regression ▸, Linear.... Vervolgens selecteer je de juiste afhankelijke variabele en de verklarende variabelen. Om een constante in het model op te nemen kies je bij Options... Include constant in equation (standaard al aangevinkt). De overige instellingen kunnen blijven zoals ze zijn.

Variables Entered/Removed [b]

Mode	Variabl Entere	Variabl Remove	Metho
1	VINEX, EH [a]	.	Ente

a. All requested variables

b. Dependent

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.510[a]	.260	.252	.585646

a. Predictors: (Constant), VINEX, W, EHS

ANOVA [b]

Mode		Sum Squar	df	Mean	F	Sig.
1	Regressi	34.04	3	11.34	33.08	.00 [a]
	Residu	96.72	282	.34		
	Tota	130.76	285			

a. Predictors: (Constant), VINEX,

b. Dependent

Coefficients [a]

Mode		Unstandardiz Coefficien		Standar ze Coefficie ts	t	Sig.
		B	Std.	Bet		
1	(Consta	.94	.12		7.74	.00
	W	2.26	.28	.41	8.07	.00
	EH	4.932E-	.00	.02	.53	.59
	VINE	.10	.01	.29	5.61	.00

a. Dependent

In bovenstaande output zie je vier tabellen. De eerste tabel geeft aan welke variabelen opgenomen zijn in het model en wat de afhankelijke variabele is. De tweede tabel geeft de R^2 en de voor vrijheidsgraden aangepaste $R^2 (= \overline{R}^2)$. De derde tabel is de ANOVA tabel. ANOVA staat voor ANalysis Of VAriance. In deze tabel kun je zien wat de totale variatie is in p (TSS), hoeveel van de variatie door het model verklaard wordt (ESS, Regression sum of squares) en welk deel niet verklaard is door het model, oftewel de variatie in de storingstermen (RSS, residual sum of squares). Ook worden de vrijheidsgraden (*df*) gegeven. In de laatste tabel worden de parameterschattingen, de standaardfouten, t-waarden en de p-waarden gegeven.

4. Je kunt R^2 en $\overline{R}^2$ ook zelf uitrekenen met de in paragraaf 5.2 gegeven formules.
R^2 = 1-RSS/TSS = 1 - 96,721/130,767 = 0,260.
$\overline{R}^2 = 1 - \frac{RSS/(N-k)}{TSS/(N-1)}$ =1 - (96,721/282)/(130,767/285) = 0,252. Beide waarden zijn hetzelfde als de door SPSS berekende waarden. Deze waarden zijn echter laag en geven

aan dat slechts 25% van de variatie in de grondprijzen wordt verklaard door de opgenomen verklarende variabelen.
Toetsen voor individuele parameters worden doorgaans uitgevoerd met een *t*-toets (zie hoofdstuk 5 en paragraaf 7.2.4). SPSS toetst standaard voor elke parameter of deze significant verschilt van 0 (zie hiervoor de vierde tabel van de output). Het is echter ook mogelijk bij het schatten een betrouwbaarheidsinterval voor elke parameter te vragen. Schat het model nogmaals maar nu met betrouwbaarheidsintervallen (Analyze, Regression ▸, Linear... knop Statistics...en vink dan Confidence intervals aan).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.945	.122		7.749	.000	.705	1.184
	W	2.261	.280	.414	8.071	.000	1.710	2.812
	EHS	4.932E-03	.009	.028	.537	.591	-.013	.023
	VINEX	.101	.018	.295	5.612	.000	.066	.137

a. Dependent Variable: P

5. Om te toetsen met behulp van een t-toets of de parameter voor ehs (β_3) gelijk is aan 0, kijken we of 0 in het betrouwbaarheidsinterval voor deze parameter ligt. Dat is het geval. De nulhypothese dat $\beta_3 = 0$ wordt dus niet verworpen bij 95% betrouwbaarheid. Een soortgelijke test voor de parameter van *vinex* (β_4) leert dat de hypothese dat deze parameter gelijk is aan 0 met 95% zekerheid wordt verworpen, aangezien 0 niet in het betrouwbaarheidsinterval ligt. We kunnen ook *t*-waarden vergelijken met kritische waarden. De parameter voor *ehs* blijkt een kleine waarde te hebben (in ieder geval kleiner dan de kritische waarde). Voor de parameter van *vinex* geldt het omgekeerde. Variabelen met een significantie (*p*-waarde) groter dan 0,05 verschillen niet significant van 0. Om de hypothese of w (β_2) gelijk is aan 2,5 te toetsen kijken we of 2,5 in het betrouwbaarheidsinterval ligt voor deze parameter. Dit blijkt zo te zijn. Deze hypothese wordt dus niet verworpen. Met dergelijke hypothesen kun je uitspraken van experts testen.

6. Is de hypothese $\beta_3 = \beta_4$ hetzelfde als de hypothese $\beta_3 = \beta_4 = 0$? Nee, de tweede hypothese is restrictiever omdat de waarde ook nog eens vastgesteld wordt op het niveau 0, terwijl in het eerste geval deze waarde kan variëren. Om de hypothese $\beta_3 = \beta_4 = 0$ te testen, voeren we een *F*-toets uit. De RSS_U is hierboven al gegeven, namelijk. 96,721. Vervolgens schatten we een model met de te testen restrictie opgelegd. De RSS_R van dat model is 108,728. De toetsingsgrootheid is:

$$F_{test} = \frac{(RSS_R - RSS_U)/q}{RSS_U/(N-k)} = \frac{(108{,}728 - 96{,}721)/2}{96{,}721/282} = \frac{6{,}004}{0{,}343} = 17{,}505$$

Deze waarde blijkt groter te zijn dan de $F_{2,282;0,05}$ waarde uit een F-tabel (= 3,00), dus we verwerpen deze hypothese met 95% zekerheid.

7. Bereken met SPSS voorspelde waarden en residuen (storingstermen) voor elke individuele waarneming. Geef ook een voorspellingsinterval voor elke waarneming (in het Linear Regression window kies Save... en vink Unstandardized Predicted Values, Unstandardized Residuals en Individual Prediction Intervals aan). Deze waarden worden in de onderstaande testen gebruikt.

8. Vervolgens testen we op heteroscedasticiteit met behulp van de White test.
Maak een aantal nieuwe variabelen (Transform, Compute...):

*e*2	het kwadraat van de residuen (de in opgave 7 aan de data toegevoegde residuals)
*w*2	het kwadraat van w
*ehs*2	het kwadraat van ehs
*vinex*2	het kwadraat van vinex

Doe nu een regressie met *e*2 als afhankelijke variabele en een constante, *w*, *ehs*, *vinex*, *w*2, *ehs*2 en *vinex*2 als verklarende variabelen. De output is als volgt:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.201[a]	.040	.020	.502550

a. Predictors: (Constant), VINEX2, W, EHS, VINEX, EHS2, W2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.193	.210		.919	.359
	W	.487	.885	.119	.550	.582
	EHS	5.815E-03	.027	.045	.217	.829
	VINEX	2.883E-02	.040	.112	.720	.472
	W2	-.587	.959	-.132	-.612	.541
	EHS2	-4.49E-04	.002	-.058	-.287	.775
	VINEX2	2.203E-03	.004	.091	.594	.553

a. Dependent Variable: E2

We nemen de R^2 van deze regressie (0,04) en vermenigvuldigen deze met het aantal waarnemingen. Dit geeft 286*0,04 = 11,44. Dit blijkt kleiner te zijn dan $\chi^2(6)$ bij een 95% significantie niveau, zodat we de H_0 van homoscedasticiteit niet verwerpen. Met andere woorden we kunnen concluderen dat er geen sprake is van heteroscedasticiteit.

9. In deze opgave testen we de modelspecificatie met behulp van een Ramsey Reset test. We gebruiken de voorspelde waarden voor *p* en creëren *p2* en *p3* (2e en 3e macht van *p*). Vervolgens schatten we het model opnieuw met deze extra termen toegevoegd:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34.161	5	6.832	19.802	.000[a]
	Residual	96.606	280	.345		
	Total	130.767	285			

a. Predictors: (Constant), P3, EHS, VINEX, W, P2

b. Dependent Variable: P

In de Ramsey RESET test toetsen we vervolgens of de parameters voor *p2* en *p3* gezamelijk gelijk zijn aan nul. Dit doen we met een *F*-test. De RSS_u is hierboven gegeven: 96,606. De RSS_R van het originele model (het model waarin deze extra termen niet zijn toegevoegd) is: 96,721. Hieruit zien we al dat het toevoegen van de extra termen slechts leidt tot een geringe daling in de som van de gekwadrateerde residuen. De test waarde is

$$F_{test} = \frac{(RSS_R - RSS_U)/q}{RSS_U/(N-k)} = \frac{(96{,}721 - 96{,}606)/2}{96{,}606/280} = \frac{0{,}058}{0{,}345} = 0{,}168 < F_{2,280;0,05}\ (= 3{,}00)$$

zodat we de H_0 (geen misspecificatie) niet verwerpen.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.535	1.315		.407	.685
	W	-2.135	11.040	-.391	-.193	.847
	EHS	-4.53E-03	.026	-.026	-.176	.861
	VINEX	-9.95E-02	.501	-.290	-.198	.843
	P2	.845	2.319	1.810	.364	.716
	P3	-.116	.356	-.823	-.326	.745

a. Dependent Variable: P

7.4 Literatuur

Er is een groot aantal boeken dat regressieanalyse behandelt. De meeste boeken gaan veel uitgebreider in op de hierboven behandelde onderwerpen. Ook de genoemde uitbreidingen kan men daarin terug vinden. Een kleine en diverse greep uit het aanbod van literatuur:

- Pindyck, R.S. en Rubinfeld (1998).
- Verbeek, M. (2000).
- Maddala, G.S. (2000).
- Greene, W.H. (2000).

Pindyck en Rubinfeld (1998) is een zeer toegankelijk boek. Het gaat onder andere uitgebreid in op analyse van tijdreeksgegevens. Een toegankelijk boek voor gevorderde gebruikers is Verbeek (2000). Dit boek is zeer up to date, behandelt de hele breedte van econometrische technieken en heeft onder andere en een prima hoofdstuk over het omgaan met panel data. Maddala (2000) wordt veel gebruikt op universiteiten. Het behandelt duidelijk, maar soms wat te bondig het gehele basisterrein van de econometrie. Voor zeer ervaren onderzoekers die niet genoeg hebben aan een basisboek is Greene (2000) één van de standaardwerken. Voor de doorsnee gebruiker is dit boek echter te formeel.

8. Trendanalyse

8.1 Inleiding

In dit hoofdstuk over trendanalyse worden twee onderwerpen behandeld: structurele verandering in regressiemodellen (zie hoofdstuk 7) en tijdreeksanalyse met één variabele. Aangezien beide onderwerpen verschillen van aard worden ze respectievelijk in paragraaf 8.2 en 8.3 behandeld. In paragraaf 8.4 wordt een casestudie uitgewerkt waar uitgelegd wordt hoe je met SPSS zelf trendanalyses uit kunt voeren. Tot slot behandelt paragraaf 8.5 bruikbare achtergrondliteratuur.

8.2 Structurele verandering in regressiemodellen

In hoofdstuk 7 is regressieanalyse besproken. Een impliciete veronderstelling die in dat hoofdstuk werd gemaakt is dat de relatie tussen de afhankelijke en de verklarende variabelen voor alle waarnemingen hetzelfde is. Deze veronderstelling is echter niet altijd geldig. De relatie die in het regressiemodel aanwezig is kan verschillen voor bepaalde groepen van waarnemingen. Dit uit zich in verschillende parameterwaarden voor verschillende groepen. In deze paragraaf bespreken we veranderingen in een regressiemodel die ontstaan in de tijd. Met andere woorden we onderzoeken of de parameters in een regressiemodel hetzelfde zijn voor verschillende periodes. Aan de hand van deze modellen is het mogelijk een antwoord te geven op een breed scala van onderzoeksvragen, zoals:

- heeft de invoering van fosfaatrechten geleid tot een verandering in de investeringen van varkenshouders;
- hoe houd ik rekening met de invoering van het melkquotum in 1984;
- is de relatie tussen tarweopbrengsten en het gebruik van gewasbeschermingsmiddelen de laatste jaren veranderd?

Eerst gaan we in op het toetsen van dergelijke veranderingen in de tijd. Vervolgens wordt besproken hoe met dergelijke veranderingen om te gaan in een regressiemodel.

Toetsen op structurele verandering

Het probleem van structurele verandering is dat de parameters β_j in een regressie model $Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$ niet hetzelfde zijn voor alle waarnemingen. Er zijn daarbij drie mogelijkheden:

- het intercept verschilt voor verschillend periodes maar de hellingsparameters niet;
- het intercept verschilt niet maar de hellingsparameter(s) wel;
- het intercept en de hellingsparameter(s) verschillen beide.

Om boventaande gevallen te toetsen gebruiken we de *F*-toets op basis van de verschillen in de som van de gekwadrateerde residuen (zie hoofdstuk 7). Het idee achter deze toets is als volgt: wanneer de parameters werkelijk verschillen voor meerdere periodes, zal een model met de restrictie dat de parameters constant zijn leiden tot grotere storingstermen dan het model waarin rekening gehouden wordt met verschillen in parameters. De som van de gekwadrateerde residuen (*RSS*) is dan uiteraard ook groter dan in een model met niet-constante parameters. Wanneer de veronderstelling van constante parameters in de tijd niet beperkend is zal er een gering verschil zijn in de *RSS* van beide modellen. In de literatuur wordt deze toets ook vaak aangeduid als de Chow-test. Om deze toets uit te voeren dienen de volgende stappen doorlopen te worden (voor het geval van twee verschillende periodes):

- orden de waarnemingen over de tijd en bepaal het vermoedelijke breekpunt (punt vanaf waar de parameters een andere waarde aannemen) en verdeel de dataset in twee subsamples (voor en na het breekpunt);
- bepaal wat je wilt testen. Bijvoorbeeld:
  $H_0 : \beta_{1,\ periode\ 1} = \beta_{1,\ periode\ 2}$
  $H_0 : \beta_{1,1} = \beta_{1,2}, \beta_{2,1} = \beta_{2,2}, \ldots, \beta_{k,1} = \beta_{k,2}$;
- schat het model voor beide sub-samples apart en bereken per model de som van de gekwadrateerde residuen. Dit levert een schatting op van de van de *RSS* van beide modellen. De totale RSS_U van het niet-gerestricteerde model wordt berekend door de *RSS* van beide sub-samples op te tellen;
- schat het model in één keer voor alle waarnemingen samen. Dit geeft de *RSS* van het gerestricteerde model (RSS_R);
- stel de toetsingsgrootheid op en vergelijk met de juiste kritische *F*-waarde (let daarbij goed op het aantal vrijheidsgraden. De F-waarden zijn te vinden in tabel 5.4). De toetsinggrootheid en het aantal vrijheidsgraden is daarbij afhankelijk van wat we toetsen (case 1, 2 of 3 hierboven):

$$F_{test} = \frac{RSS_R - RSS_U}{RSS_U / (N-k-1)} \sim F(1, N-k-1) \qquad \text{(intercept verschilt)}$$

$$F_{test} = \frac{(RSS_R - RSS_U)/k-1}{RSS_U / (N-2k)} \sim F(k-1, N-2k) \qquad \text{(hellingsparamters verschillen)}$$

$$F_{test} = \frac{(RSS_R - RSS_U)/k}{RSS_U / (N-2k)} \sim F(k, N-2k) \qquad \text{(alle parameters verschillen);}$$

- als $F_{test} > F_{q,p;0,05}$ verwerp je de nulhypothese.

In het bovenstaande stappenplan gaan we uit van een toets waarbij de parameters van twee sub-samples worden vergeleken. Uiteraard is het mogelijk dit uit te breiden naar meerdere sub-samples. Het toetsprincipe blijft hetzelfde.

Modelleren van structurele verandering

Wanneer uit bovenstaande toetsen blijkt dat één of meer parameters verschillen voor bepaalde periodes kun je daar rekening mee houden in het regressiemodel. Dit doen we met behulp van dummyvariabelen. Een dummyvariabele heeft slecht 0 en 1 als mogelijke waarden en kan daarom gebruikt worden om verschillende periodes te onderscheiden. Bijvoorbeeld:

$$d_1 = \begin{cases} 0 & \text{voor periode 1, bijv. 1970 - 1983 of april - september} \\ 1 & \text{voor periode 2, bijv. 1984 - 1999 of oktober - maart} \end{cases}$$

Op deze manier kun je verschillende intercepten en hellingsparameters modelleren. In algemene zin ziet het model er dan als volgt uit:

$$Y_i = \beta_1 + (\beta_2 - \beta_1)d_1 + \beta_3 X_{3i} + (\beta_4 - \beta_3)d_1 X_{3i} + \varepsilon_i$$

Voor periode 1 is d_1 gelijk aan nul, zodat het intercept gelijk is aan β_1 en de hellingsparameter aan β_3. Voor periode 2 is d_1 gelijk aan 1, zodat het intercept β_2 wordt en de hellingsparameter gelijk aan β_4. In de praktijk worden de verklarende variabelen X_k gesplitst voor beide periodes door 2 aparte variabelen aan te maken. De eerste met de waarden voor X_k tot aan het breekpunt en vanaf daar met alleen nullen en de tweede met eerst nullen voor X_k tot aan het breekpunt en vanaf daar met de waarden voor X_k.

Waneer we meerdere periodes onderscheiden, dienen we ook meerdere dummyvariabelen te gebruiken. Let er echter op dat we voor M periodes, slechts $M-1$ dummyvariabelen nodig hebben. Dit is omdat het intercept al een bepaalde basisniveau aangeeft en hiermee een periode voor zijn rekening neemt.

8.3 Tijdreeksanalyse met één variabele

In de vorige paragraaf bespraken we structurele verandering in regressiemodellen. Het accent in dergelijke modellen ligt op structurele relaties tussen variabelen. Met behulp van verklarende variabelen wordt het verloop van de afhankelijke variabele verklaard. Bij tijdreeksanalyse ligt de nadruk op het gedrag van een variabele in het verleden. Door patronen in dit gedrag te modelleren kan het verloop van de variabele naar de toekomst geprojecteerd worden. Tijdreeksanalyse zou je kunnen omschrijven als verstandig extrapoleren door onder andere te zoeken naar trends, patronen en cycli in een variabele. Bij tijdreeksanalyse is de tijdshorizon van de data doorgaans veel langer dan bij regressieanalyse.

Tijdreeksanalyse is een interessant alternatief voor regressieanalyse als voorspellen niet mogelijk is op basis van verklarende variabelen. Dit kan verschillende redenen hebben:

- de parameters in een regressiemodel verschillen niet significant van nul;
- de verklarende variabelen zijn niet beschikbaar om mee te schatten;
- de verklarende variabelen zijn niet beschikbaar om te projecteren naar de toekomst.

Eenvoudige deterministische extrapolatie

Een eenvoudige tussenvorm van regressieanalyse en tijdreeksanalyse is deterministische extrapolatie. Deterministisch in de zin dat het random karakter van de data genegeerd wordt. Met de gewone kleinste kwadratenschatter (OLS) wordt een relatie tussen de variabele y en de tijd t geschat. Voorbeelden hiervan zijn:

Lineair trendmodel: $$y_t = c_1 + c_2 t$$

Kwadratisch trendmodel: $y_t = c_1 + c_2 t + c_3 t^2$

Exponentiële groei: $y_t = Ae^{rt}$ of $\log y_t = c_1 + c_2 t$

Autoregressief trendmodel: $y_t = c_1 + c_2 y_{t-1}$

Logaritmisch autoregressief model: $\log y_t = c_1 + c_2 \log y_{t-1}$

Deze manier van werken is eenvoudig, maar doorgaans weinig accuraat. Vooral schommelingen in de variabele blijken moeilijk te 'verklaren' in dergelijke modellen. De tijdreeksmodellen die hieronder besproken worden zijn daartoe veel beter in staat.

Eigenschappen van tijdreeksen

Voordat we een aantal algemene tijdreeksmodellen bespreken, kijken we eerst naar een aantal veronderstelde eigenschappen van tijdreeksen. We veronderstellen dat y_1, $y_2 ... y_T$ afkomstig zijn uit een kansverdeling. De kunst is nu een regel, model of proces te vinden dat aangeeft hoe de y's getrokken worden uit deze kansverdeling. Belangrijk is dat dit proces constant is in de tijd. Dit noemen we stationair. Een proces dat verandert in de tijd is niet-stationair. Voor stationariteit is het nodig dat het gemiddelde, de variantie en de kansverdeling van een variabele constant zijn in de tijd. Om na te gaan of dit zo is kun je bijvoorbeeld twee periodes vergelijken (1800-1899 en 1900-1999 of week 1, 26, 27 en 52).

Voor een niet-stationair proces is het moeilijk om een relatie tussen opeenvolgende waarden van y te vinden. Echter door de eerste verschillen te nemen ($\Delta y_t = y_t - y_{t-1}$) kan de niet-stationariteit opgeheven worden. Een voorbeeld illustreert dit. Investeringen It blijken niet-stationair te zijn. In de periode januari 1980 - december 1989 heeft de reeks een ander gemiddelde en een andere variantie dan in de periode januari 1990 - december 1999. Echter, ΔI_t (= I_t - I_{t-1}) blijkt wel stationair te zijn. De gemiddelde en de variantie van de verschillen zijn wel gelijk in beide periodes. Soms is het nodig om vaker verschillen te nemen, dit komt neer op de verschillen van verschillen.

Autocorrelatiefunctie

Voordat we een relatie tussen opeenvolgende waarden van y bepalen is het handig om te kijken naar autocorrelaties, zoals de correlatie tussen waarde y_t en y_{t-1} of tussen y_t en y_{t-3}. De verschillende autocorrelatiewaarden worden samengevat met behulp van de autocorrelatiefunctie:

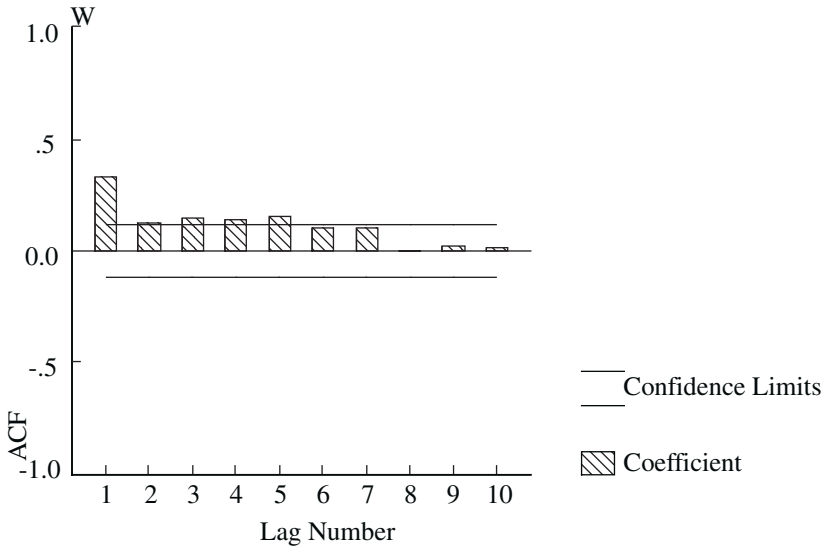
$$\rho_k = \frac{\mathrm{cov}(y_t, y_{t-k})}{\sigma_y^2}$$

ρ_k is de correlatie tussen y_t en y_{t-k}

$\mathrm{cov}(y_t, y_{t-k})$ is de covariantie tussen y_t en y_{t-k}

σ_y^2 is de variantie van y

Een grafiek met de verschillende waarden voor ρk noemen we een autocorrelatieplot. Deze ziet er in SPSS als volgt uit:



Figuur 8.1 Autocorrelatieplot

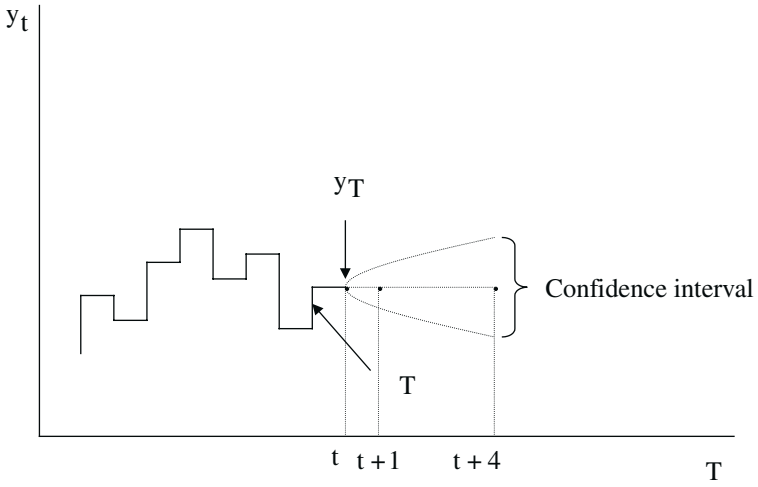
Uit bovenstaand autocorrelatieplot blijkt dat direct opeenvolgende waarden relatief sterk gecorreleerd zijn. Waarnemingen waar meer dan 7 periodes tussen zitten zijn onderling bijna niet meer gecorreleerd. Het verloop van een autocorrelatieplot is behulpzaam in het formuleren van een tijdreeksmodel voor de variabele y die we onderzoeken.

Tijdreeksmodellen
In een tijdreeksmodel proberen we y te verklaren met behulp van eigen vertraagde waarden (y_{t-j}) en huidige en vertraagde storingstermen (ε_t en ε_{t-j}). Op basis hiervan zijn verschillende modellen te formuleren, waarbij het vierde model het meest algemeen is:
- random walk;
- voortschrijdend gemiddelde (moving average, MA);
- autoregressief proces (AR);
- combinatie van 2 en 3, al dan niet na het nemen van verschillen (als proces niet-stationair is; ARIMA).

Random walk
In een random walk proces ontwikkelt y_t zich geheel willekeurig. Er is geen patroon te vinden in opeenvolgende waarden van y_t. De ontwikkeling in y_t wordt alleen bepaalde door willekeurige sprongen ε_t:

$y_t = y_{t-1} + \varepsilon_t$ oftewel $y_t = \sum_{i=1}^{t} \varepsilon_i$

De storingstermen zijn willekeurig (random) bepaald: $\varepsilon_t \sim (0, \sigma^2)$. De beste projectie naar de toekomst die je kunt doen voor y_{t+j} is y_t. Hoe verder vooruit je projecteert, hoe groter de standaardfout van je projectie wordt. Een random walk kun je herkennen aan het feit dat alle waarden ρ_k van de autocorrelatiefunctie gelijk zijn aan nul. Eventueel kun je een trend of 'drift' term toevoegen aan het random walk proces. De volgende figuur geeft een random walk proces weer:



Figuur 8.2 Random walk

Voortschrijdend gemiddelde

Een voortschrijdend gemiddelde of moving average proces wordt doorgaans aangeduid als MA(q), waarbij q het aantal termen is dat wordt opgenomen. In een dergelijk proces wordt y_t bepaald door gewogen gemiddelden van storingstermen tot q periodes terug:

$$y_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q}$$

Voor alle storingstermen geldt daarbij opnieuw: $\varepsilon_t \sim (0, \sigma^2)$. Belangrijkste eigenschap van een MA-proces is dat je het verleden tot q periodes terug kunt meenemen. De ontwikkeling in de 'fouten' in voorspellingen (storingstermen) worden als het ware in acht genomen. Wanneer uit een autocorrelatieplot blijkt dat een variabele slechts met een beperkt aantal vertraagde waarden gecorreleerd is kun je het proces het beste als een MA-proces modelleren. De autocorrelaties geven een indicatie van welke vertraagde waarden je moet opnemen.

Autoregressief proces
Een autoregressief proces wordt aangeduid met AR(*p*), waarbij p het aantal termen is dat wordt opgenomen. In een dergelijk proces wordt y_t bepaald door een gewogen gemiddelde van eigen historische waarden tot *p* periodes terug:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t$$

waarbij $\varepsilon_t \sim (0, \sigma^2)$. Bij autoregressieve processen wordt het hele verleden van een variabele in acht genomen, zelfs voor een AR(1)-proces. Door een AR(1)-proces te herschrijven wordt dit duidelijk:

$$\begin{aligned} y_t &= \alpha_1 y_{t-1} + \varepsilon_t = \alpha_1(\alpha_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \alpha_1(\alpha_1(\alpha_1 y_{t-3} + \varepsilon_{t-2}) + \varepsilon_{t-1}) + \varepsilon_t \quad \text{etc.} \\ &= \varepsilon_t + \alpha_1 \varepsilon_{t-1} + (\alpha_1)^2 \varepsilon_{t-2} + \ldots + (\alpha_1)^p \varepsilon_{t-p} \end{aligned}$$

Door het opnemen van meerdere termen kun je het verleden op verschillende manieren meenemen. De verschillen tussen MA(*q*)- en AR(*p*)-modellen zijn als volgt samen te vatten. Een MA(*q*)-model neemt informatie van slechts q periodes terug mee. Een AR(*p*)-model neemt de hele historie in beschouwing, maar met afnemend gewicht. Een MA(*q*)-proces is flexibel door verschillende parameter waarden voor β's, een AR(*p*)-proces door het opnemen van *p*-termen. Verder blijken MA(*q*)-processen meer willekeurig op en neer te schommelen, terwijl AR(*p*)-processen een duidelijker patroon hebben.

Autoregressief voortschrijdend gemiddelde
Het ligt voor de hand om een MA(q)-proces en een AR(p)-proces te combineren, om zo beide positieve eigenschappen te combineren. Zo'n model wordt aangeduid als ARMA (p,q). Met behulp van ARMA(*p,q*)-modellen kan het verloop van *y* op een betere manier beschreven worden, wat bijdraagt aan een grotere voorspelkracht. Een ARMA(p,q)-model ziet er als volgt uit:

$$y_t = \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}$$

Als de tijdreeks niet-stationair is, moet je verschillen nemen. Het aantal keren dat je verschillen moet nemen totdat de reeks stationair is, wordt aangeduid met d. Een dergelijk model duiden we aan als ARIMA(*p,d,q*). Deze kunnen ook in SPSS geschat worden.

Keuze van een model
In bovenstaande paragrafen zijn verschillende basismodellen beschreven, die afhankelijk van het aantal opgenomen termen, in staat zijn om het verloop van een variabele y op verschillende manieren te beschrijven. Een relevante vraag is welke specificatie gekozen dient te worden. Zonder een duidelijke regel zijn hierbij verschillende elementen van belang. Vooraf kan op basis van een autocorrelatieplot een keuze gemaakt worden voor de specificatie. Wanneer y_t alleen gecorreleerd is met de vertraagde waarde vier periodes terug, dan volstaat een MA(4)-model. Wanneer een duidelijke aflopende trend in de autocorrelaties is waar te nemen, volstaat een eenvoudig AR(1) of AR(2)-model. Bij meer ingewikkelde patronen in de autocorrelaties dient een combinatie van termen genomen te worden. Het aantal termen is doorgaans lastig te bepalen en kan het beste met behulp van 'trial and error' worden vastgesteld.

Een andere aanpak is het schatten van een algemeen ARIMA(p,d,q)-model met meerdere termen. Vervolgens kan gekeken worden welke geschatte parameters significant van nul verschillen. Verschillende modellen kunnen ook worden vergeleken op basis van het Akaike Informatie Criterium (AIC) of Bayesian Informatie Criterium (BIC). Beide maten geven een weging van de RSS (Residual Sum of Squares) van het geschatte model en het aantal opgenomen termen. Zie ook paragraaf 7.2.3 voor meer informatie over de twee criteria. Voor deze criteria geldt: hoe kleiner hun waarde, hoe beter het model.

Tot slot kun je kijken welk model het beste in staat is te voorspellen. Daartoe kun je voorspellingen van verschillende modellen uitzetten in een grafiek tegen werkelijke waarden.

8.4 Casestudie

In deze casestudie onderzoeken we het verloop van vleesprijzen en de aankopen van vlees door Nederlandse huishoudens. Een dataset van PVE is beschikbaar met daarin de volgende variabelen:

qv totale hoeveelheid varkensvlees gekocht door Nederlandse huishoudens (in 1.000.000 kg);
qr totale hoeveelheid rundvlees gekocht door Nederlandse huishoudens (in 1.000.000 kg);
pv prijs van varkensvlees in guldens per kilogram;
pr prijs van rundvlees in guldens per kilogram.

De gegevens zijn vierwekelijkse waarnemingen van januari 1994 tot en met december 2000. In totaal zijn er 91 waarnemingen beschikbaar. De data is beschikbaar als SPSS-file 'vleesprijs.sav' toegankelijk via de intranetsite van het Domeinteam Data en Modellen. Start het programma SPSS en open de datafile (File, Open ▸, Data... ; selecteer 'vleesprijs.sav' vanuit de juiste directory in het geopende window).

A. Structurele verandering in een regressie model
In dit onderdeel wordt de relatie tussen aankopen van rundvlees en de prijzen van rund- en varkensvlees onderzocht. De vraag is of deze relatie is veranderd sinds op 21 maart 1997 in Wilp het eerste geval van BSE in Nederland werd gevonden. Deze datum zit aan het eind van waarneming 42. Vanaf waarneming 43 zou er dus een mogelijk effect zijn. Daarom wordt de *F*-test op structurele verandering uitgevoerd.

A1. Schat eerst het volgende model voor de gehele dataset (Kies Analyze, Regression ▸, Linear... . Vervolgens selecteer je de juiste afhankelijke variabele en de verklarende variabelen. Om een constante in het model op te nemen kies je bij Options... Include constant in equation (standaard al aangevinkt). De overige instellingen kunnen blijven zoals ze zijn):

$$qr_t = \beta_1 + \beta_2 pr_t + \beta_3 pv_t + \varepsilon_i$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661[a]	.437	.424	.361478

a. Predictors: (Constant), PR, PV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.920	2	4.460	34.131	.000[a]
	Residual	11.499	88	.131		
	Total	20.418	90			

a. Predictors: (Constant), PR, PV

b. Dependent Variable: QR

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.230	.726		12.721	.000
	PV	-.145	.051	-.285	-2.842	.006
	PR	-.241	.054	-.449	-4.481	.000

a. Dependent Variable: QR

Dit is het gerestricteerde model. De aanname geldt dat in de gehele periode de relatie tussen hoeveelheid aangekocht rundvlees en prijzen van vlees gelijk is. De RSS_R (som van de gekwadrateerde residuen) is 11,499. De schattingsresultaten zijn redelijk plausibel. Een stijging in de prijs van rundvlees leidt tot een daling van de gevraagde hoeveelheid. De prijs van varkensvlees heeft ook een negatief effect op de gevraagde hoeveelheid rundvlees. Als je varkensvlees beschouwt als substituut van rundvlees is dit niet wat je verwacht. Wanneer de prijs van varkensvlees stijgt verwacht je dat mensen varkensvlees vervangen door rundvlees.

A2. Schat nu het model voor beide periodes apart (dus voor waarneming 1-42 en voor waarneming 43-91). Gebruik week als selectievariabele (<43 en >42). Tel van beide modellen de *RSS* op. Dit geeft de RSS_U van het niet-gerestricteerde model, waarbij we ervan uitgaan dat alle parameters (intercept en hellingparameters) verschillen in beide periodes.

ANOVA[b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.839	2	1.920	14.361	.000[a]
	Residual	5.213	39	.134		
	Total	9.052	41			

a. Predictors: (Constant), PR, PV

b. Dependent Variable: QR

c. Selecting only cases for which WEEK < 43

ANOVA[b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.981	2	.990	9.522	.000[a]
	Residual	4.785	46	.104		
	Total	6.765	48			

a. Predictors: (Constant), PR, PV

b. Dependent Variable: QR

c. Selecting only cases for which WEEK > 42

De $RSS_U = 5{,}213 + 4{,}785 = 9{,}998$.

A3. Vervolgens voeren we een F-test uit om te testen of alle parameters verschillen:

$$F_{test} = \frac{(RSS_R - RSS_U)/k}{RSS_U/(N-2k)} = \frac{(11{,}499 - 9{,}998)/3}{9{,}998/85} = \frac{0{,}500}{0{,}118} = 4{,}24 > F_{3,85;0,05}\ (= 2{,}72)$$

Hieruit blijkt dat we de nulhypothese van gelijke parameters over beide periodes verwerpen.

A4. Vervolgens schat je een model waarin je met behulp van dummyvariabelen rekening houdt met veranderingen in het aankoopgedrag na het eerste BSE-geval. Daarvoor maak je een aantal nieuwe variabelen aan:

$$d_1 = \begin{cases} 0 & \text{voor week } 1\text{-}42 \\ 1 & \text{voor week } 43\text{-}91 \end{cases}$$

$$pr2 = \begin{cases} 0 & \text{voor week } 1\text{-}42 \\ pr & \text{voor week } 43\text{-}91 \end{cases}$$

$$pv2 = \begin{cases} 0 & \text{voor week } 1\text{-}42 \\ pv & \text{voor week } 43\text{-}91 \end{cases}$$

en schat je vervolgens $qr_t = \beta_1 + \gamma_1 d_1 + \beta_2 pr_t + \gamma_1 pr2_t + \beta_3 pv_t + \gamma_3 pv2_t + \varepsilon_i$

De parameters γ meten op deze manier de verandering in het intercept en de verandering in de hellingsparameters voor *pr* en *pv*. Schattingsuitkomsten zijn:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.714[a]	.510	.482	.342957

a. Predictors: (Constant), PR2, PR, PV, PV2, D1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.421	5	2.084	17.719	.000[a]
	Residual	9.998	85	.118		
	Total	20.418	90			

a. Predictors: (Constant), PR2, PR, PV, PV2, D1

b. Dependent Variable: QR

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.563	1.229		8.596	.000
	PV	-.239	.100	-.469	-2.386	.019
	PR	-.255	.090	-.474	-2.836	.006
	D1	-3.382	1.558	-3.559	-2.170	.033
	PV2	.213	.118	2.706	1.806	.074
	PR2	4.493E-02	.109	.812	.411	.682

a. Dependent Variable: QR

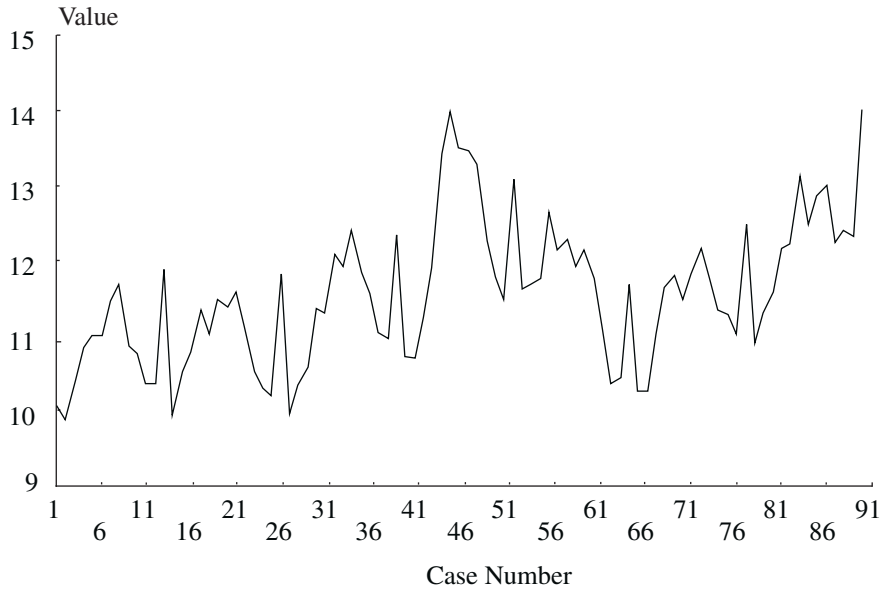
Uit de *t*-waardes en *p*-waardes blijkt dat de verandering in het intercept significant van nul verschilt bij een 5% kritisch niveau. De verandering in de hellingsparameter voor varkensvleesprijs verschilt significant van nul bij een 10% kritisch niveau. De verandering van de rundvleesprijs verschilt, opmerkelijk genoeg, niet significant van nul. Met andere woorden het effect van de rundvleesprijs op de rundvleesconsumptie is niet veranderd door de BSE-crisis. Interessant is te zien dat de verschillende parameterwaarden ook al zijn geschat bij het testen op structurele verandering toen beide modellen apart geschat werden voor beide periodes. Voordeel van het werken met dummy's is dat sommige variabelen kunnen variëren terwijl anderen constant gehouden worden voor beide periodes. Op deze manier zijn de schattingen efficiënter dan bij apart schatten voor beide periodes. Apart schatten heeft als voordeel dat de

variantie mag variëren. In het dummymodel wordt de variantie van de storingstermen constant verondersteld.

B. Tijdreeksanalyse
Dit onderdeel richt zich op de prijs van varkensvlees. De interesse gaat voornamelijk uit naar een model dat de prijs van varkensvlees goed voorspelt, zonder dat daarbij gebruikgemaakt wordt van een bepaalde relatie met andere variabelen. Met andere woorden de prijs van varkensvlees wordt niet verklaard uit het aanbod, andere prijzen en de concentratiegraad in de keten maar aan de hand van het gedrag van de prijs van varkensvlees in de tijd.

SPSS kent een aantal handige opties voor tijdreeksanalyse. Zo kun je tijdreeksmodellen schatten (Analyze, Time series ▸), maar je kunt ook variabelen transformeren (Transform, Create Time Series...). Met deze laatste optie kun je verschillen nemen, vertraagde waarden creëren, corrigeren voor seizoenseffecten enzovoorts. Ook het gebruik van grafieken is handig.

B1. Zet de varkensvleesprijs in een grafiek uit tegen de tijd (Graphs, Line...; kies Simple en values of individual cases en dan Define waar je de juiste variabelen selecteert). Kun je op basis van deze grafiek zeggen of de tijdreeks stationair is?



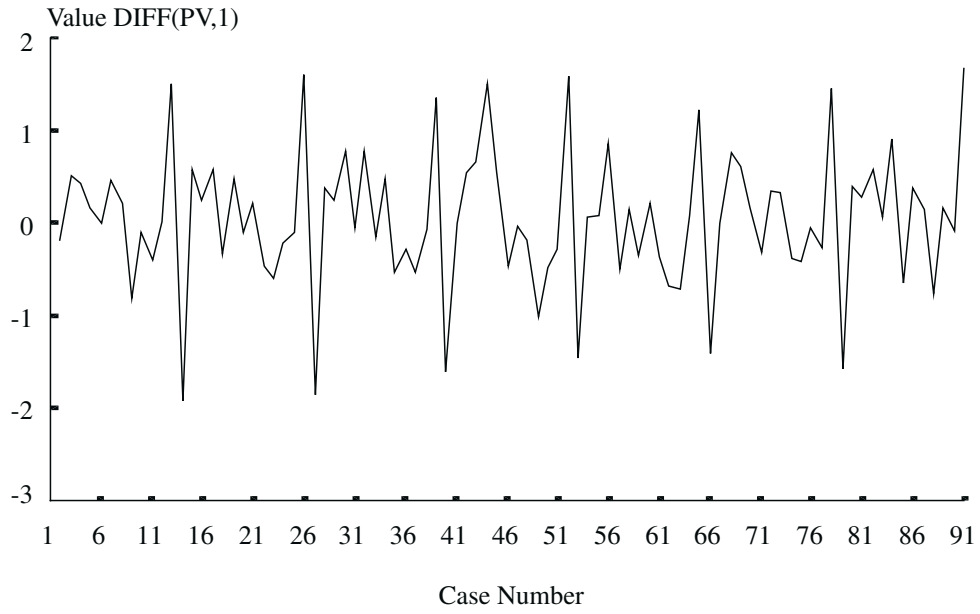
Figuur 8.3 Tijdreeks van de prijs van varkensvlees

Het gemiddelde en de variantie lijken niet constant. Je zou dus verwachten dat deze variabele niet stationair is.

B2. Maak nu een variabele met de eerste verschillen van pv en noem deze pv_1. (Transform, Create Time Series...). Zet vervolgens de eerste verschillen uit in een grafiek tegen de tijd. Zijn de eerste verschillen stationair? Kun je in één van beide grafieken een patroon onder-

scheiden en eventueel verklaren?

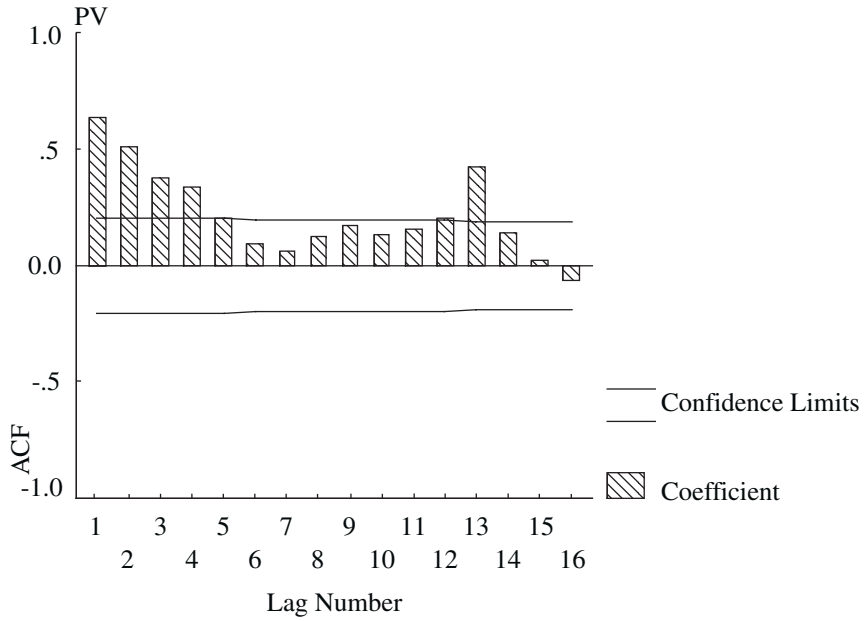
Allereerst lijkt het gemiddelde en de variantie redelijk constant voor bijvoorbeeld twee subperiodes. Er is duidelijk een patroon te onderscheiden. Rond de twaalfde vierwekelijkse periode neemt de prijs van varkensvlees duidelijk toe. Dit is rond de kerstdagen. Daarna neemt de varkensvleesprijs af. Ook mogelijke effecten van Pasen en zomervakantie (barbecuen) kun je onderscheiden. Op basis van deze eerste verschillen lijkt het mogelijk de ontwikkeling in de varkensvleesprijs te modelleren.



Figuur 8.4 Tijdreeks van de eerste verschillen van de prijs van varkensvlees

B3. Maak een autocorrelatieplot om na te gaan hoe waarnemingen in de tijd samenhangen met vertraagde waarnemingen. (Graphs, Time series ▸ Autocorrelations) Vink alleen autocorrelations aan. SPSS neemt standaard zestien vertraagde waarden (lags) mee. Dit kun je veranderen bij Options....

Het is duidelijk waar te nemen dat er eerst sprake is van een afname in de autocorrelaties, waarna deze weer toenemen. Rond de dertiende vertraging is er een duidelijke sterke autocorrelatie. Je bent dan een jaar verder en de cyclus van varkensvleesprijzen begint opnieuw.



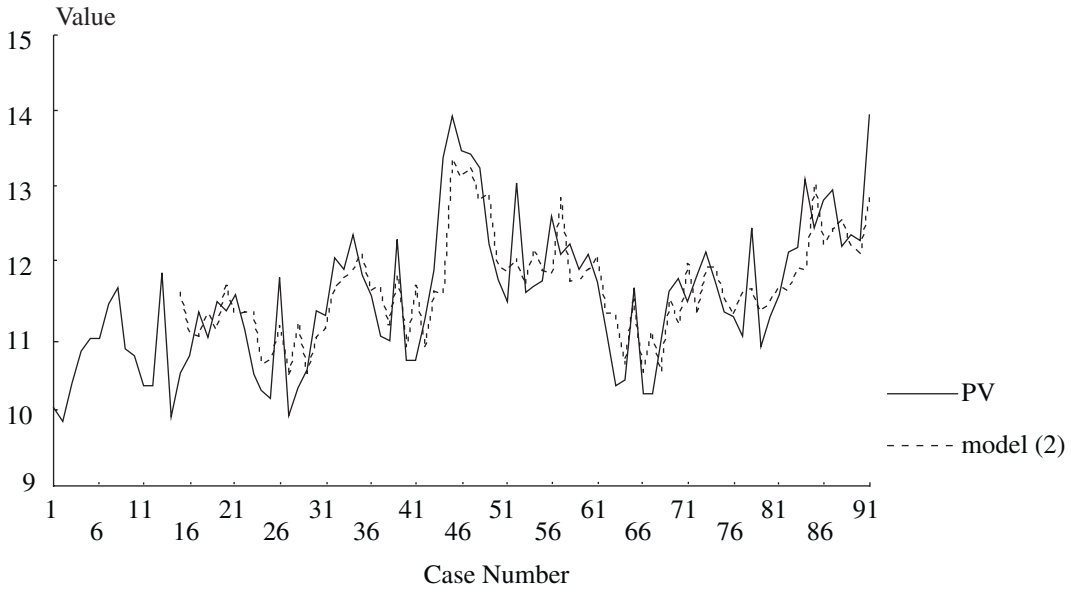
Figuur 8.5 Autocorrelatieplot

B4. Schat een aantal ARIMA-modellen in SPSS. (Analyze, Time series ▸ Arima...). Met de optie Arima... kun je namelijk beter beheersen welke termen je opneemt dan met de optie Autoregression... (deze neemt altijd een AR(1) term op). Je kunt vertraagde waarden aanmaken met (Transform, Create Time Series...).
Voorbeelden:

$$pv_t = \alpha_0 + \alpha_1 pv_{t-1} + \varepsilon_t$$

$$pv_t = \alpha_0 + \alpha_1 pv_{t-1} + \alpha_2 pv_{t-8} + \alpha_3 pv_{t-13} + \varepsilon_t$$

Creëer voor elke model voorspelde waarden en zet alle voorspelde waarden tegen elkaar uit in een grafiek.



Figuur 8.6 Model 2: $pv_t = \alpha_0 + \alpha_1 pv_{t-1} + \alpha_2 pv_{t-8} + \alpha_3 pv_{t-13} + \varepsilon_t$

Uit de gestippelde lijn kunnen we afleiden dat model 2 redelijk goed in staat is om pv te voorspellen.

8.5 Literatuur

Onderstaande literatuur gaat dieper in op de onderwerpen die in dit hoofdstuk behandeld zijn:
- Pindyck, R.S. en Rubinfeld, D.L. (1998).
- Verbeek, M. (2000).
- Hamilton, J.D. (1994).

Pindyck en Rubinfeld (1998) is zeer toegankelijk. Het besteed aandacht aan structurele verandering in regressiemodellen, maar vooral tijdreeksanalyse wordt zeer uitgebreid besproken. Verbeek (2000) is een algemeen econometrieboek dat in twee hoofdstukken dieper op tijdreeksanalyse in gaat. De beschrijving is beknopter maar ook formeler. Voor gevorderde gebruikers is het zeker een aanrader. Het boek van Hamilton is gespecialiseerd in tijdreeksanalyse en met meer dan 800 pagina's is het vooral bestemd voor diegenen die alles willen weten over tijdreeksanalyse.

9. Keuzemodellen

9.1 Inleiding

In dit hoofdstuk worden modellen besproken die gebruikt kunnen worden voor de analyse van keuzes. De modellen zijn een verbijzondering van de lineaire regressiemodellen zoals die in hoofdstuk 6 besproken zijn. De eenvoudigste modellen gaan uit van twee mogelijke keuzes. Nadat deze modellen uitgebreid besproken zijn wordt een aantal uitbreidingsmogelijkheden gegeven (meerdere keuzes, geordende keuzes, nul waarnemingen versus continue hoeveelheden).

Het hoofdstuk is als volgt opgebouwd. In paragraaf 9.2 bespreken we de theorie om keuzegedrag te analyseren. Behalve verschillende modellen wordt ook ingegaan op zaken als voorspellen en kwaliteit van het model. In paragraaf 9.3 wordt een casestudie uitgewerkt met behulp van SPSS software. Paragraaf 9.4 bespreekt ten slotte bruikbare achtergrondliteratuur.

9.2 Theorie

In hoofdstuk 6 is het lineaire regressiemodel besproken. Een afhankelijke variabele Y_i werd aan de hand van een aantal onafhankelijke variabelen X_{ij} verklaard. Deze relatie werd weergegeven met een lineaire vergelijking:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

De afhankelijke variabele Y_i werd daarbij continu verondersteld (bijvoorbeeld prijs, inkomen, productie). De onafhankelijke variabelen X_{ij} zijn doorgaans ook continu. Daarnaast kunnen er enkele onafhankelijke dummyvariabelen opgenomen zijn in het model. Dummyvariabelen worden gebruikt als indicatorvariabelen om een kwalitatief onderscheid te kunnen maken. Voorbeelden van dummyvariabelen zijn geslacht (man of vrouw), wel of geen diploma, wel of niet werkloos enzovoorts. In paragraaf 9.2.1 worden modellen beschreven waarin dummyvariabelen de te verklaren variabelen zijn. Uitbreidingen op deze modellen zijn ook mogelijk, er kunnen bijvoorbeeld meer dan twee mogelijkheden zijn: je hebt bijvoorbeeld de keuze tussen gangbare, geïntegreerde of biologische landbouw. Uitbreidingen worden in paragraaf 9.2.2 beschreven.

9.2.1 Binaire keuzemodellen

Bij (binaire) keuzemodellen heeft de afhankelijke variabele Y_i slechts twee mogelijke waarden: 0 of 1. De verklarende variabelen zijn opnieuw continu of binair (dummy). Dergelijke modellen duidt men doorgaans aan als binaire keuzemodellen, kwalitatieve keuzemodellen of

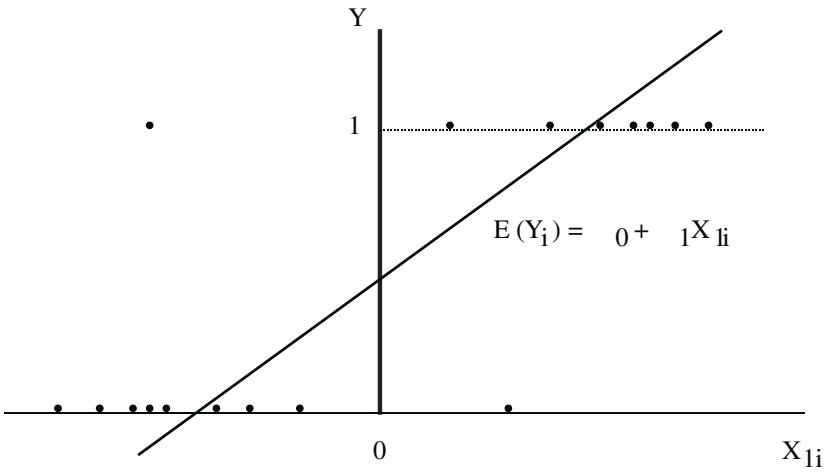
beperkt afhankelijke variabele modellen. De toepassingsmogelijkheden van deze modellen zijn groot. Enkele voorbeelden zijn het onderzoeken van het al dan niet aankopen van een bepaald product, buiten de landbouw werken door boeren, de switch naar biologische landbouw, investeringsgedrag of stemgedrag van boeren en werkloosheid onder boeren. De afhankelijke variabele is:

$$Y_i = \begin{cases} 0 & \text{gangbaar, niet geïnvesteerd, werkloos, etc.} \\ 1 & \text{biologisch, geïnvesteerd, werkzaam, etc.} \end{cases}$$

Lineair Kans Model
In een lineair kansmodel wordt niets anders gedaan dan het schatten van de vergelijking zoals die hierboven weergegeven is met OLS, waarbij Y_i dus slechts twee mogelijke waarden kent. Daarbij kun je aantonen dat de uitkomst van het model (bijvoorbeeld bij voorspellen) een kans is:

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$$
$$E(Y_i) = 1 \cdot P_i + 0 \cdot (1 - P_i) = P_i$$
$$\Rightarrow \; P_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$$

De eerste vergelijking geeft de verwachtingswaarde voor Y_i. De tweede vergelijking geeft ook de verwachtingswaarde maar kijkt naar de mogelijke uitkomsten. Die zijn 1 met een kans P_i en 0 met een kans (1-P_i), zodat de verwachtingswaarde gelijk is aan P_i. Met andere woorden de verwachtingswaarde van het model is gelijk aan de kans op $Y_i = 1$. De volgende figuur geeft een aantal waarnemingen (0 of 1) voor een keuzemodel en het daarbij behorende geschatte lineaire kansmodel weer:

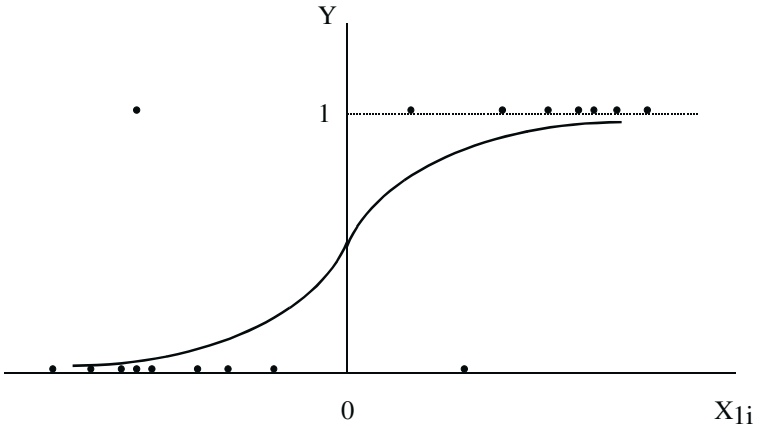


Figuur 9.1 Lineair kansmodel

Het geschatte lineaire kansmodel wordt weergegeven met de diagonale lijn. Uit deze lijn zijn direct een aantal (negatieve) eigenschappen van het lineair kansmodel af te lezen. Zo kunnen voorspelde kansen groter dan 1 zijn of kleiner dan 0. Wat verder opvalt is dat het effect van een variabele X_j overal gelijk is omdat het om een lineair model gaat. Dit geldt zowel dichtbij 0, 0,5 als bij 1. Deze eigenschap is ook niet altijd wenselijk. Als laatste valt op dat het model niet goed in staat is om te voorspellen. Lineaire kansmodellen hebben doorgaans een lage R^2. Andere problemen met een lineair kansmodel zijn heteroscedasticiteit en niet normaal verdeelde storingstermen. De consequenties daarvan zijn dat de schattingen niet efficiënt zijn en dat toetsen voor parameters formeel niet uitgevoerd kunnen worden. Voordelen van het lineair kansmodel zijn uiteraard dat het eenvoudig (te schatten) is en dat de parameterschattingen zuiver zijn.

Logit en Probit

De hierboven genoemde nadelen van het lineair kansmodel hebben er toe geleid dat er alternatieve modellen ontwikkeld zijn voor het analyseren van keuzes door economische agenten. Vooral het genoemde nadeel, van kansen die groter dan 1 en kleiner dan 0 kunnen zijn, heeft hieraan bijgedragen. Een alternatief model zou er als volgt uit moeten zien:



Figuur 9.2 Logit- en probitmodel

Het S-vormige verloop garandeert dat (voorspelde) kansen niet groter zijn dan 1 en niet kleiner dan 0 en heeft ook tot gevolg dat het effect van de verklarende variabele het sterkst is in het midden. Een toename in X zorgt rond de waarde 0 voor een relatief veel grotere stijging in Y dan bij andere waarden van X. De feitelijke oplossing voor de problemen van het lineair kansmodel ligt in het transformeren van het verklarende deel van het model met een functie die een dergelijk S-vormig verloop heeft. Met andere woorden zoek een functie *h* die $\beta_0+\beta_1X_{1i}+\ldots+\beta_kX_{ki}$ transformeert naar het interval [0,1]. Het model komt er als volgt uit te zien:

$$P_i = h\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k\right)$$

Een functie die een S-vormig verloop heeft is een cumulatieve verdelingsfunctie. Wanneer het verklarende deel van het model (verklarende variabelen en parameters) getransformeerd wordt met een normale verdelingsfunctie wordt het ontstane model een probit model genoemd. Wanneer de meer gebruiksvriendelijke verdelingsfunctie van de logistische verdeling gebruikt wordt noemen we dit een logitmodel.

De verdelingsfunctie van een normale verdeling die leidt tot een probit model is gedefinieerd als:

$$P_i = F(Z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-t^2/2} dt$$

De verdelingsfunctie van de logistische verdeling (logitmodel) is:

$$P_i = F(Z_i) = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki})}}$$

Beide verdelingen lijken sterk op elkaar alleen zijn de 'staarten'van de logistische verdeling wat dikker. Dat betekent dat probit- en logitmodellen doorgaans ook tot vergelijkbare uitkomsten leiden met betrekking tot voorspellingen. Het is duidelijk dat de verdelingsfunctie van de logistische verdeling een meer gebruiksvriendelijke vorm heeft. Daarom wordt bij handberekeningen doorgaans het logitmodel gebruikt. Wanneer men gebruik maakt van een software pakket als SPSS maakt het uiteraard niet uit welk model men gebruikt.

Het gebruik van een verdelingsfunctie om het model te transformeren heeft als gevolg dat de directe relatie tussen de uitkomst van het model (kans) en het verklarende deel verandert. Stel het verklarende deel van het model gelijk aan een niet-observeerbare (latente) en continue variabele Y_i* :

$$Y_i^* = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i$$

Deze latente variabele is te interpreteren als *voorkeur voor*, *neiging tot* of *nut van* de keuze die wordt aangeduid met 1. De relatie tussen deze latente variabele en de waargenomen keuzes is als volgt. Je neemt een aankoop of investering waar als de latente variabele (voorkeur) een drempelwaarde overschrijdt. En je neemt deze aankoop of investering niet waar als de latente variabele een waarde heeft die kleiner is dan de drempelwaarde. Met andere woorden:

$$Y_i = \begin{cases} 1 & \text{als } Y_i^* > 0 \\ 0 & \text{als } Y_i^* \leq 0 \end{cases}$$

Doordat de verdelingsfunctie toeneemt in Y_i* bestaat er ook een positieve relatie tussen Y_i en Y_i*. De bovenstaande twee vergelijkingen beschrijven samen een probit- en logitmodel. Ook voor dergelijke modellen kan je laten zien dat de uitkomst van het model (bij voorspellen) een kans is:

$$
\begin{aligned}
P_i &= P(Y_i = 1) = P\left[\varepsilon_i > -(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)\right] \\
&= 1 - F\left[-(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)\right] \\
&= F(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)
\end{aligned}
$$

Schatten van een probit- en logitmodel

Het transformeren van de latente variabele met behulp van een verdelingsfunctie heeft tot gevolg dat de OLS-methode niet gebruikt kan worden om een probit- of logitmodel te schatten. OLS kan immers alleen gebruikt worden bij het schatten van parameters van lineaire vergelijkingen. De transformatie heeft de vergelijking echter niet-lineair gemaakt. Hoe een dergelijk model dan wel te schatten? Dat gebeurt met de methode van de grootste aannemelijkheid, beter bekend als Maximum Likelihood (ML). De essentie van Maximum Likelihood is dat een doelfunctie met parameters en variabelen gemaximaliseerd wordt zodanig dat de kansverdeling van waarnemingen zo goed mogelijk benaderd wordt. Hier wordt niet verder op deze methode ingegaan. Het enige dat hier van belang is, is de eindwaarde van deze functie. Deze wordt aangeduid als de Log-Likelihoodwaarde (of LogL of L). Deze waarde zegt iets over de kwaliteit van de schatting en is ook bruikbaar bij het testen.

Voorspellen met probit- en logitmodellen

Met behulp van geschatte parameters $\hat{\beta}$ en waarnemingen kun je kansen P_i voorspellen, bijvoorbeeld in een logitmodel:

$$
P_i = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki})}}
$$

Let er echter op dat het model de kans op $Y_i = 1$ (P_i) voorspelt en niet direct $Y_i = 1$ of $Y_i = 0$ als uitkomst geeft! Gebruik daarom een beslisregel om voorspellingen te doen. Bijvoorbeeld:

$P_i > 0{,}5$ voorspel dan $Y_i = 1$
$P_i \leq 0{,}5$ voorspel dan $Y_i = 0$

Kwaliteit van het model

Er zijn twee eenheden die vaak gebruikt worden om de kwaliteit van een (geschat) logit of probit model te beoordelen. De eerste is de proportie goede voorspellingen oftewel Count R^2. Deze bereken je als volgt:

- voorspel de kans P_i voor elke waarneming;
- gebruik een beslisregel
  bijvoorbeeld $P_i > 0{,}5$, voorspel dan $Y_i = 1$
  $P_i \leq 0{,}5$, voorspel dan $Y_i = 0$;
- tel het aantal correcte voorspellingen. Dan geldt:

$$\text{Count } R^2 = \frac{\text{aantal goede voorspellingen}}{\text{totaal aantal observaties}}$$

Deze maatstaf is vaak hoog. Dit komt doordat je ook zonder model altijd een aantal correcte voorspellingen doet (voorpel dat alles 1 is of alles 0 is; in beide gevallen doe je al een aantal goede voorspellingen). Je moet dus eigenlijk kijken hoeveel beter je kunt voorspellen met het model.

De tweede maatstaf kijkt meer naar de bijdrage van de verklarende variabelen in het model. Hiervoor wordt de Log-Likelihoodwaarde van het geschatte model vergeleken met de Log-Likelihoodwaarde van een model met alleen een constante (een dergelijk model voorspelt voor elke waarneming dezelfde kans). Gebruikmakend van beide waarden wordt McFaddens's R^2 gedefinieerd:

$$\text{McFadden's } R^2 = 1 - \frac{L(\beta)}{L(0)}$$

$L(\beta)$ is de waarde van de log-likelihood van het geschatte model is
$L(0)$ is de log-likelihoodwaarde van een model met alleen een constante

McFaddens's R^2 ligt tussen 0 en 1. In het geval dat het model perfect voorspelt, is $L(\beta)=0$ (maximale waarde), $R^2 = 1$. Wanneer de variabelen niets toevoegen aan het model (alle βs =0), dan geldt $L(\beta) = L(0)$ en $R^2 = 0$.

Effect van een variabele op P_i
Een verandering van een verklarende variabele heeft tot gevolg dat de kans op de gebeurtenis die we onderzoeken ook verandert. We zagen al dat voor het lineair kansmodel deze verandering overal hetzelfde is. Voor het lineaire kansmodel geldt:

$$\frac{\partial P_i}{\partial X_{ij}} = \beta_j \qquad \text{oftewel} \quad \partial P_i = \beta_j \cdot \partial X_{ij}$$

Zoals uit de grafiek voor het probit- en logitmodel blijkt is het effect van een variabele op de kans (het marginale effect) niet constant voor deze modellen. Met andere woorden het niveau van de waarde van een verklarende variabele heeft invloed op het marginale effect. Voor probit- en logitmodellen gelden de volgende formules om de marginale effecten te berekenen:

Logit: $$\frac{\partial P_i}{\partial X_{ij}} = \beta_j \cdot P_i\left(1 - P_i\right)$$

Probit: $$\frac{\partial P_i}{\partial X_{ij}} = \beta_j \cdot f\left(Z_i\right)$$

$f(Z_i)$ is de kansdichtheidsfunctie van de standaard normale verdeling.

Vergelijken van modellen
Met een gegevensset kun je een lineair kansmodel, probit- en een logitmodel schatten en vergelijken. De vergelijking kun je maken op basis van *Count* R^2, *McFadden's* R^2 of op basis van de marginale effecten. Parameters van deze drie modellen kun je echter niet direct vergelijken doordat er verschillende transformatiefuncties gebruikt zijn om de modellen te schatten. Om parameters te kunnen vergelijken moet je ze omrekenen. Hiervoor gelden de volgende vuistregels:

$$\hat{\beta}_{\text{Probit}} \cong 0{,}625 \cdot \hat{\beta}_{\text{Logit}}$$
$$\hat{\beta}_{\text{LKM}} \cong 0{,}25 \cdot \hat{\beta}_{\text{Logit}}$$
$$\hat{\beta}_{\text{LKM}} \cong 0{,}25 \cdot \hat{\beta}_{\text{Logit}} + 0{,}5 \quad \text{voor constante}$$
$$\hat{\beta}_{\text{Probit}} \cong 2{,}5 \cdot \hat{\beta}_{LKM}$$
$$\hat{\beta}_{\text{Probit}} \cong 2{,}5 \cdot \hat{\beta}_{\text{LKM}} - 1{,}25 \quad \text{voor constante}$$

9.2.2 Uitbreiding keuzemodellen

Tot nog toe zijn we uitgegaan van slechts twee mogelijke keuzes. Alhoewel de meeste keuzeprocessen teruggebracht kunnen worden tot twee mogelijkheden (definieer een groep waarin je geïnteresseerd bent en voeg alle andere categorieën samen tot een restgroep) kan dit toch beperkend zijn. Hieronder wordt een aantal uitbreidingen besproken op het basismodel met twee keuzes.

Meerdere keuzen met ordening
De eerste uitbreiding bouwt direct voort op het binaire probit- en logitmodel in die zin dat de structuur van het model hetzelfde is behalve dat er meerdere (geordende) keuzen zijn. Voorbeelden van dergelijke geordende keuzes zijn: niet werken, parttime werken of voltijds werken; gangbare, geïntegreerde of biologische landbouw; aankoop gangbare, volière of scharreleieren. Opnieuw is er een latente variabele aanwezig (voorkeur voor, neiging tot of nut van), maar nu met meerdere klassen. Deze latente variabele is net als in de binaire keuzemodellen essentieel voor een geordend model.

$$Y_i^* = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i$$
$$Y_i = j \quad \text{als } \gamma_{j-1} < Y_i^* \leq \gamma_j$$

γ_{j-1} en γ_j zijn klassengrenzen

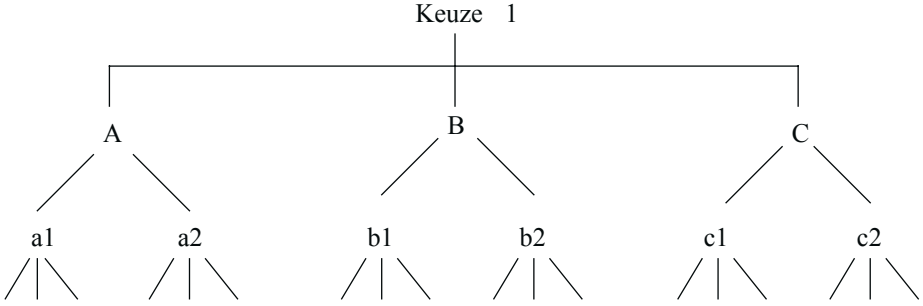
Meestal worden de klassengrenzen ook in het model geschat.

Meerdere keuzen zonder ordening
Er zijn ook keuzes waaraan geen ordening ten grondslag ligt. In modellen waarin dit het geval is, wordt verondersteld dat aan elke keuze een bepaald nut wordt ontleend. De waargenomen keuze hangt dan samen met het maximale nut. Het nut voor individu *i* van keuze *j* kan met de volgende functie worden weergegeven: $U_{ij} = X_{ij}\beta + \varepsilon_{ij}$. Op basis hiervan kan een multinomial logitmodel worden opgesteld:

$$P(Y_i = j) = \frac{e^{X_{ij}\beta}}{1 + e^{X_{i2}\beta} + \ldots + e^{X_{iM}\beta}}$$

Het idee van een multinomial logitmodel is dat de kenmerken X_{ij} van in totaal M alternatieven worden gewogen. De waargenomen keuze heeft de hoogste gewogen som.

Meerdere keuzemomenten
Het idee van het multinomial logitmodel op basis van vergelijking van nutsniveaus kan verder worden uitgebreid naar een model met meerdere keuzes op meerdere momenten. Eerst wordt een hoofdkeuze gemaakt waarna een vervolgkeuze gemaakt moet worden en daarna eventueel nog meerdere vervolgkeuzes. Nested logitmodellen kunnen met dit soort geneste keuzestructuren omgaan. Het keuzeproces dat hieraan ten grondslag ligt heeft een boomstructuur, zoals aangegeven in figuur 9.3



Figuur 9.3 Nested logitmodel

Voorbeelden van dergelijke geneste keuzen zijn:
- vleesaankopen: rund- of varkensvlees; vervolgens riblappen of bieflappen (rund) óf speklappen of karbonade(varken);
- stoppen of doorgaan met een bedrijf. Stoppen: verkopen of hobbybedrijf. Doorgaan: maatschap of overdracht aan opvolger;
- multifunctionele landbouw: Ja of nee. Zo ja, wat (camping, zorgboerderij, natuurbeheer enzovoorts).

Wel of niet; indien wel: hoeveel?
In het binaire keuzemodel gingen we na of een bepaalde gebeurtenis wel of niet plaatsvindt. Voor een aantal keuzeprocessen is dat echter een beperking. Voor bijvoorbeeld aankopen, investeringen of arbeidsdeelname is het interessanter om de hoeveelheid te verklaren in het geval deze gebeurtenis plaatsvindt. We onderscheiden bijvoorbeeld het geval dat er niet geïnvesteerd wordt van het geval dat er wel geïnvesteerd wordt. En als er wel geïnvesteerd wordt ben je geïnteresseerd in het bijbehorende bedrag. Een dergelijk model heet een Tobit model. De analogie met een probit model is duidelijk:

$$Y_i^* = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i$$
$$Y_i = \begin{cases} Y_i^* & \text{als } Y_i^* > 0 \\ 0 & \text{als } Y_i^* \leq 0 \end{cases}$$

De latente variabele wordt nu deels waargenomen (als er wel wordt geïnvesteerd) en deels niet (als er niet wordt geïnvesteerd). In het geval een hoeveelheid wordt waargenomen is deze gelijk aan de latente variabele. Een dergelijk model wordt ook wel een censored regression genoemd (de latente variabele is voor nulwaarnemingen gecensureerd). Dit model kent vele uitbreidingen (zie Amemiya, 1984 voor een overzicht). Zo is het mogelijk om gescheiden beslisregels en hoeveelheidvergelijkingen te modelleren. Dergelijke modellen worden veelal gebruikt voor het modelleren van drempels en selectieprocedures. In SPSS bestaan er geen standaard routines voor dergelijke modellen. Stata en Limdep zijn twee alternatieve gebruiksvriendelijke software pakketten die voor dergelijke modellen wel bruikbaar zijn.

9.3 Casestudie

Biologische landbouw staat volop in de belangstelling, ook al is het aandeel in de totale landbouw nog niet groot. In deze casestudie onderzoeken we of we de keuze voor biologische landbouw kunnen relateren aan een aantal kenmerken van boeren. Op deze manier krijgen we meer inzicht in de karakteristieken van beide groepen en kunnen we nagaan hoe we potentiële bio-boeren kunnen onderscheiden. In de SPSS-file bioboer.sav (File, Open ▸, Data... ; selecteer 'bioboer.sav' vanuit de juiste directory in het geopende window) zijn de volgende gegevens opgenomen voor biologische melkveehouders:

- *biodum* indicator voor biologische bedrijfsvoering (0 = gangbaar, 1 = biologisch)
- *biotype* indicator soort biologische landbouw (0 = gangbaar, 1 = EKO, 2 = biodynamisch)
- *age* leeftijd van de boer in jaren
- *educ* dummy voor opleidingsniveau (0 = geen hoger of middelbaar onderwijs, 1 = wel)
- *sizeha* bedrijfsgrootte in hectares
- *sizequo* grootte melkquotum (100.000 kg)
- *tenure* aandeel pacht in totale hoeveelheid land

- *clay* dummy voor grondsoort (0 = zand of veen, 1 = klei)
- *crratio* verhouding eigen vermogen/totaal vermogen
- *animalha* aantal dieren per hectare
- *prof* korte termijn winst (totaal saldo) (€ 100.000 van 1990)

Deze file is te benaderen via de intranetsite van het Domeinteam Data en Modellen. Er zijn in totaal 473 waarnemingen over de periode 1994-1999, waarvan 120 van biologische bedrijven.

1. Schat een Lineair Kans Model waaraan je de keus voor biologische landbouw (biodum) relateert aan de verklarende variabelen 3 tot en met 11 en een constante. Kies ervoor om voorspelde waardes te bewaren. (Kies Analyze, Regression ▸, Linear...). Vervolgens selecteer je de juiste afhankelijke variabele en de verklarende variabelen. Om een constante in het model op te nemen kies je bij Options... Include constant in equation (standaard aangevinkt). Bij Save... kies je voor (Unstandardized) Predicted Values). Dit geeft de volgende output:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.537[a]	.289	.275	.37

a. Predictors: (Constant), PROF, AGE, TENURE, CLAY, EDUC, ANIMALHA, CRRATIO, SIZEHA, SIZEQUO

b. Dependent Variable: BIODUM

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25.873	9	2.875	20.901	.000[a]
	Residual	63.683	463	.138		
	Total	89.556	472			

a. Predictors: (Constant), PROF, AGE, TENURE, CLAY, EDUC, ANIMALHA, CRRATIO, SIZEHA, SIZEQUO

b. Dependent Variable: BIODUM

Coefficients[a]

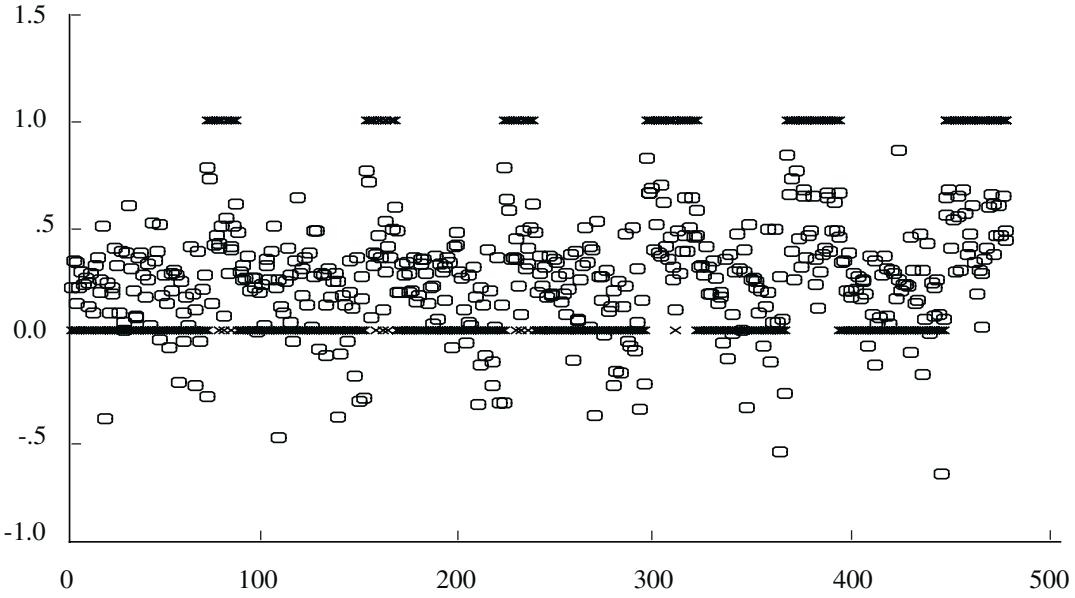
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.879	.152		5.780	.000
	AGE	-5.11E-03	.002	-.136	-3.256	.001
	EDUC	.200	.038	.229	5.235	.000
	SIZEHA	4.927E-03	.002	.244	3.085	.002
	SIZEQUO	-6.84E-02	.017	-.355	-4.062	.000
	TENURE	-2.86E-02	.056	-.022	-.514	.607
	CLAY	4.205E-02	.037	.048	1.132	.258
	CRRATIO	-.303	.188	-.081	-1.612	.108
	ANIMALHA	-.116	.031	-.227	-3.763	.000
	PROF	1.108E-02	.029	.032	.383	.702

a. Dependent Variable: BIODUM

Naar aanleiding van de bovenstaande output kunnen we de volgende vragen stellen:

- Welke variabelen hebben parameters die significant van nul verschillen (Met andere woorden welke variabelen doen ertoe?)
  Alle variabelen met een p-waarde (Sig.) kleiner dan bijvoorbeeld 0,05 (als we met 95% betrouwbaarheid uitspraken willen doen). Dat zijn dus: constante, age, educ, sizeha, sizequo en animalha.
- Hoeveel verandert de kans op biologische landbouw als de leeftijd van een boer 10 jaar toeneemt?
  $\partial P_i = \beta_j \cdot \partial X_{ij} = -0{,}005 \cdot 10 = -0{,}05$ *. Met andere woorden de kans neemt met 0,05 af*.
- Hoeveel verandert de kans als de boer van lager opgeleid naar hoger opgeleid verandert?
  $\partial P_i = \beta_j \cdot \partial X_{ij} = 0{,}200 \cdot 1 = 0{,}200$ *. Met andere woorden de kans neemt dan met 0,2 toe*.
- Geef een interpretatie van de gevonden R^2.
  De R^2 geeft aan dat slecht 28,9% van de variatie van het al dan niet biologisch boeren wordt verklaard door het model. Dit is tamelijk laag, maar we weten dat R^2 doorgaans laag is voor een lineair kansmodel.

2. Bekijk de voorspelde kansen op biologische landbouw. Zijn dit vreemde uitkomsten? (Je kunt de voorspelde waarden direct bekijken, of de samenvattende statistieken (minimum, maximum, gemiddelde enzovoorts) of een over-lay scatter plot maken voor een grafisch overzicht (Graphs, Scatter...; kies Overlay en koppel *biodum* aan *index* en vervolgens de *voorspelde kans* aan *index*)



Figuur 9.4 *Voorspellingen aan de hand van een lineair kansmodel*
De kruisjes op de lijnen biodum=0 en biodum=1 geven de waargenomen waarden weer. De rondjes zijn de voorspelde waardes. De voorspelde waarden (kansen) kunnen allerlei waardes aannemen. Duidelijk is te zien dat veel voorspelde kansen kleiner dan 0 zijn.

3. Schat vervolgens een probit model met dezelfde variabelen als bij 1 en laat SPSS ook hier de voorspelde kansen uitrekenen. In SPSS kun je dit op twee manieren doen. De parameterschattingen zijn bij beide methodes hetzelfde, maar de output van de tweede manier is veel bruikbaarder:

- als je kiest voor Analyze, Regression ▸ , Probit, selecteer bij Response Frequency variabele *biodum*. Bij Total Observed moet je een variabele selecteren met voor elke waarneming waarde. Deze moet je dus vooraf aanmaken. De reden voor deze omslachtige manier van werken is dat SPSS probit met name gebruikt voor analyses van groepseffecten;
- een beter manier is te kiezen voor Analyze, Regression ▸ , Ordinal... Bij Dependent geef je de afhankelijk variabele *biodum* op en bij Covariates de verklarende variabelen 3 tot en met 11. Vervolgens kies je bij Option... bij het knopje Link: voor Probit (of Logit als je dat later wilt doen). Om voorspelde kansen uit te rekenen kies je bij Output... Saved variables, Estimated respons probabilities. Het is ook handig om deze kans direct te vertalen in een voorspelde waarde 0 of 1 (Predicted Category);

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	535.780			
Final	317.652	218.127	9	.000

Link function: Probit.

Pseudo R-Square

Cox and Snell	.369
Nagelkerke	.545
McFadden	.407

Link function: Probit.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[BIODUM = 0]	-4.769	.875	29.674	1	.000	-6.484	-3.053
Location	AGE	-2.08E-02	.007	8.877	1	.003	-3.454E-02	-7.128E-03
	EDUC	1.356	.210	41.826	1	.000	.945	1.767
	SIZEHA	-4.43E-03	.008	.273	1	.601	-2.103E-02	1.218E-02
	SIZEQUO	-.200	.088	5.196	1	.023	-.371	-2.799E-02
	TENURE	1.532E-02	.278	.003	1	.956	-.530	.560
	CLAY	1.722E-02	.175	.010	1	.921	-.325	.359
	CRRATIO	-1.966	.953	4.255	1	.039	-3.835	-9.790E-02
	ANIMALHA	-1.785	.254	49.470	1	.000	-2.282	-1.288
	PROF	.167	.171	.960	1	.327	-.167	.502

Link function: Probit.

Naar aanleiding van de bovenstaande output kan de volgende vraag worden gesteld: Bereken met de gegeven output de proportie goede voorspellingen (Count R^2) en McFadden's R^2. Vergelijk je berekende waarde met de gegeven waarde. Om Count R^2 te berekenen is het handig om ook de voorspelde categorie (Predicted Category) uit te laten rekenen. Je kunt dan een kruistabel maken (Analyse... Descriptive Statistics... Cross-tabs...) waarin je de werkelijke waardes van *biodum* uitzet tegen de *voorspelde biodum*. Zo weet je direct het aantal correcte voorspellingen. Ook kun je zo nagaan waar de meeste foute voorspellingen zijn. Met andere woorden voorspelt het model vaak biologische bedrijven als gangbaar, of juist vaak gangbare bedrijven als biologisch.

BIODUM * Predicted Response Category Crosstabulation

Count

		Predicted Response Category		
		0	1	Total
BIODUM	0	329	24	353
	1	48	72	120
Total		377	96	473

$$\text{Count } R^2 = \frac{\text{aantal goede voorspellingen}}{\text{totaal aantal observaties}} = \frac{329 + 72}{473} = 0{,}848$$

Voor het berekenen van de McFadden's R^2 kun je de waarden van de log-likelihoods aflezen uit de tabel Model Fitting Information.

$$\text{McFadden's } R^2 = 1 - \frac{L(\beta)}{L(0)} = \frac{-158{,}83}{-267{,}89} = 0{,}407$$

Dit is hetzelfde als de door SPSS gegeven waarde.

4. Schat op dezelfde manier ook een Logitmodel.

Parameter Estimates

							95% Confidence Interval	
		Estimate	Std. Error	Wald	df	Sig.	Lower Bound	Upper Bound
Threshold	[BIODUM = 0]	-8.801	1.604	30.110	1	.000	-11.945	-5.657
Location	AGE	-3.97E-02	.012	10.394	1	.001	-6.386E-02	-1.557E-02
	EDUC	2.588	.406	40.691	1	.000	1.793	3.383
	SIZEHA	-1.06E-02	.015	.494	1	.482	-4.022E-02	1.899E-02
	SIZEQUO	-.347	.157	4.908	1	.027	-.654	-4.004E-02
	TENURE	-2.21E-02	.518	.002	1	.966	-1.037	.993
	CLAY	.121	.315	.149	1	.700	-.496	.739
	CRRATIO	-3.840	1.719	4.992	1	.025	-7.209	-.472
	ANIMALHA	-3.232	.473	46.591	1	.000	-4.160	-2.304
	PROF	.308	.318	.938	1	.333	-.315	.932

Link function: Logit.

De volgende vragen hebben betrekking op de geschatte logit / probit en lineaire kansmodellen:

- vergelijk de voorspelde kansen van een Lineair kansmodel, een Probit en een Logitmodel;
  g*eringe verschillen tussen Probit en Logit, meer verschillen tussen Lineair Kansmodel en Probit/Logit;*
- vergelijk McFadden's R^2 voor Logit en Probit;
  voor Logit iets hogere waarde, 0,414; echter gering verschil;
- vergelijk enkele parameters van de drie modellen door gebruik te maken van de vuistregels (zie paragraaf 9.2.1);
  de onderstaande tabel vergelijkt een aantal parameters:

Tabel 9.1 Vergelijking parameters

	Lin. Kans M.	Logit	Probit	β*LKM* → β*Probit*	β*Logit* → β*Probit*
Constante	0,879	-8,801	-4,769	0,947	-5,501
Educ	0,200	2,588	1,356	0,500	1,618
Sizequo	-0,068	-0,347	-0,200	-0,170	-0,217
Animalha	-0,116	-3,232	-1,785	-0,290	-2,020

Wat opvalt is dat (na transformatie) de parameters van het Lineaire Kans Model behoorlijk afwijken van die van het Probit model. De Probit- en Logitmodellen liggen dicht bij elkaar.

5. Je kunt de keus voor biologisch versus gangbaar ook zien als een geordende keuze. Van gangbaar naar EKO naar biodynamisch zou dan steeds een stap verder zijn. Schat daarom een Ordered Probit model voor deze 3 mogelijkheden (gebruik variabele *biotype*). Je kunt hier opnieuw kijken hoe goed het model voorspelt enzovoorts. Wat valt je op in de voorspelde kansen?

Case Processing Summary

		N	Marginal Percentage
BIOTYPE	0	389	82.2%
	1	53	11.2%
	2	31	6.6%
Valid		473	100.0%
Missing		0	
Total		473	

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	553.081			
Final	418.098	134.983	9	.000

Link function: Probit.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[BIOTYPE = 0]	-2.933	.773	14.403	1	.000	-4.447	-1.418
	[BIOTYPE = 1]	-2.140	.766	7.818	1	.005	-3.641	-.640
Location	AGE	-1.00E-02	.007	2.123	1	.145	-2.347E-02	3.454E-03
	EDUC	1.120	.205	29.875	1	.000	.719	1.522
	SIZEHA	-1.39E-02	.007	3.420	1	.064	-2.856E-02	8.290E-04
	SIZEQUO	-.141	.088	2.598	1	.107	-.313	3.054E-02
	TENURE	.162	.260	.388	1	.534	-.347	.671
	CLAY	9.685E-02	.166	.341	1	.559	-.228	.422
	CRRATIO	-1.806	.897	4.052	1	.044	-3.565	-4.753E-02
	ANIMALHA	-1.331	.229	33.645	1	.000	-1.780	-.881
	PROF	.261	.161	2.634	1	.105	-5.418E-02	.576

Link function: Probit.

BIOTYPE * Predicted Response Category Crosstabulation

Count

		Predicted Response Category		
		0	2	Total
BIOTYPE	0	389		389
	1	43	10	53
	2	26	5	31
Total		458	15	473

Allereerst valt op dat dit model een lagere McFadden's R^2 heeft (0,244) dan in het geval van slechts twee keuzes. Ook de Count R^2 is lager (0,833). Blijkbaar is dit model niet slechter. Kijkend naar de voorspellingen valt op dat het moeilijk is om het onderscheid tussen EKO en biodynamisch te voorspellen. Geen enkel bedrijf wordt als EKO-bedrijf voorspeld. In totaal worden maar 5 van de 69 biologische bedrijven in de juiste categorie voorspeld. Dit model is dus zeker geen verbetering.

6. Schat een multinomial logitmodel voor deze 3 keuzes (Analyze, Regression ▸ , Multinomial Logistic; Vraag bij Statistics... om een classification table. Deze geeft meteen een handig overzicht van de goede en foute voorspellingen).

- Vergelijk de uitkomsten met het ordered probit model. Welke heeft je voorkeur?

Classification

	Predicted			
Observed	0	1	2	Percent Correct
0	380	9	0	97.7%
1	38	11	4	20.8%
2	18	11	2	6.5%
Overall Percentage	92.2%	6.6%	1.3%	83.1%

Count R^2 is iets lager voor het multinomial logitmodel, maar McFadden's R^2 is iets hoger (0,289). Het multinomial logitmodel is echter wel beter in staat onderscheid te maken tussen EKO en biodynamisch. Het laatste model heeft dan ook onze voorkeur en je kunt concluderen dat de keuze tussen niet-biologisch, EKO en biodynamisch geen ordinale keuze is. Er spelen blijkbaar andere overwegingen een rol om te kiezen tussen EKO en biodynamisch.

9.4 Literatuur

De meeste algemene econometrieboeken zoals die in paragraaf 6.4 zijn genoemd, besteden ruim aandacht aan het modelleren van keuzes. Daarnaast is er een veelheid aan gespecialiseerde literatuur:

- Amemiya, T. (1981);
- Amemiya, T. (1984);
- Maddala, G.S. (1983).

De bovenstaande literatuur is redelijk toegankelijk. Maddala wordt beschouwd als een standaardwerk dat, alhoewel al wat gedateerd, de meeste uitbreidingen op het basismodel bespreekt.

10. Multivariate technieken

10.1 Inleiding

Dit hoofdstuk bespreekt een drietal multivariate data-analyse technieken. De term multivariate technieken is een verzamelnaam voor verschillende data-analyse technieken. Men kan deze term op de volgende manieren omschrijven: 'Gelijktijdige analyse van meer dan twee variabelen' of 'Meten, verklaren en voorspellen van de mate van samenhang in gewogen combinaties van variabelen' (Hair, 1998). Onder deze algemene definitie valt een aantal technieken waarvan er één al uitgebreid is besproken in de hoofdstukken 7 en 9: regressieanalyse. In dit hoofdstuk worden hoofdcomponentenanalyse, factoranalyse en clusteranalyse besproken. Een andere multivariate techniek, discriminantanalyse, heeft veel overeenkomsten met de keuzemodellen die in hoofdstuk 9 besproken zijn. In hoofdstuk 11 worden nog twee andere multivariate technieken besproken: conjunctanalyse en multidimensional scaling.

Multivariate data-analyse technieken kunnen worden onderverdeeld in twee hoofdgroepen: afhankelijkheids (dependence)-technieken waar afhankelijke en onafhankelijke variabelen onderscheiden worden en onderlinge afhankelijkheids (interdependence)-technieken waarbij geen onderscheid tussen afhankelijke en onafhankelijke variabelen wordt gemaakt. Bij deze laatste groep worden variabelen gezamelijk onderzocht. Regressieanalyse is een voorbeeld van een afhankelijkheidstechniek. Hoofdcomponentenanalyse, factoranalyse en clusteranalyse zijn onderlinge afhankelijkheidstechnieken.

Dit hoofdstuk is verder als volgt opgebouwd. Paragraaf 10.2 bespreekt hoofdcomponenten- en factoranalyse. Paragraaf 10.3 bespreekt clusteranalyse. In paragraaf 10.4 wordt een casestudie besproken waarin beide technieken gebruikt worden in SPSS. Paragraaf 10.5 geeft immers een kort overzicht van bruikbare literatuur voor een verdere verdieping.

10.2 Hoofdcomponenten- en factoranalyse

Hoofdcomponentenanalyse en factoranalyse zijn twee aparte technieken. Toch worden ze doorgaans niet apart onderscheiden (zie bijvoorbeeld SPSS). Dit komt mede doordat de eerste stap in een factoranalyse vaak een hoofdcomponentenanalyse is. En doordat er naast verschillen ook een aantal overeenkomsten in beide technieken zijn. Zo analyseren beide technieken relaties tussen een (groot) aantal variabelen en definiëren zij gemeenschappelijke onderliggende dimensies (hoofdcomponenten of factoren). Het doel bij beide technieken is het samenvatten van data of datareductie. Toepassingsmogelijkheden zijn bijvoorbeeld: het bepalen van hoofdkenmerken van een bedrijfsstructuur, het bepalen van het imago van een bepaald product of het karakteriseren van management. In alle drie de gevallen gebeurt dit aan de hand van een aantal hoofdkenmerken. Bij deze technieken wordt gezocht naar breed geformuleerde kenmerken die meerdere variabelen omvatten. Dit wordt in het volgende voorbeeld weergegeven:

X2 A-merk X8 Prijs X9 Verpakking	Product
X3 Bereidingstijd X4 Extra benodigdheden X5 Kant-en-klaar	Gemak
X1 Aantal KJ X6 Kleurstoffen X7 Onverzadigde vetzuren X10 Vitamines	Gezondheid

De 10 variabelen zijn samen te vatten in drie hoofdeigenschappen van een product. De producten kunnen vervolgens aan de hand van deze drie hoofdeigenschappen gekarakteriseerd worden.

10.2.1 Hoofdcomponentenanalyse

In een hoofdcomponentenanalyse nemen we de p variabelen in een dataset ($X_1, X_2...X_p$) en zoeken daarmee p lineaire combinaties ($Z_1, Z_2...Z_p$) die niet gecorreleerd zijn:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

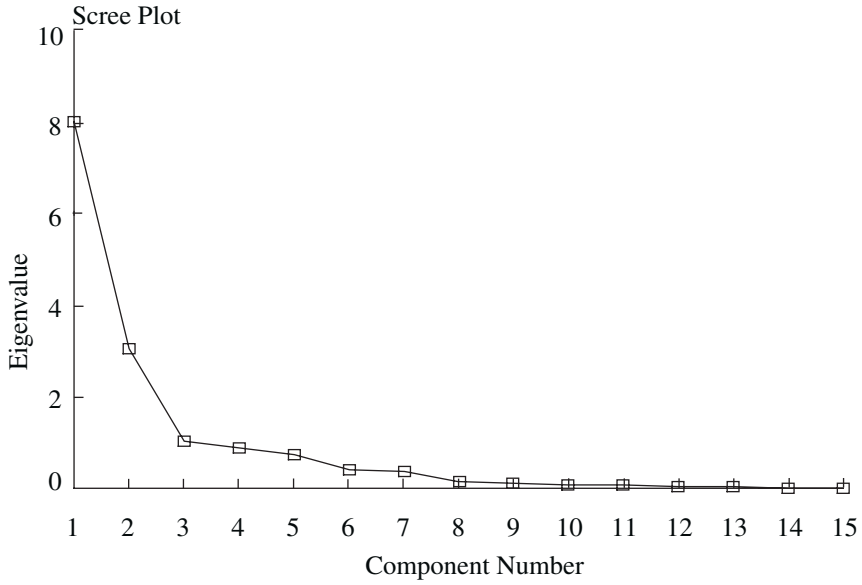
We noemen Z_i hoofdcomponenten. Omdat deze hoofdcomponenten niet gecorreleerd zijn beschrijven ze verschillende dimensies van data. Bij het afleiden van de hoofdcomponenten geldt:

$\text{var}(Z_1) \geq \text{var}(Z_2) \geq ... \geq \text{var}(Z_p)$

Wanneer bijvoorbeeld Z_1 en Z_2 samen 90% van de variantie in de totale dataset omvatten, kunnen p variabelen samengevat worden met twee hoofdcomponenten. Om een hoofdcomponentenanalyse uit te voeren moet je de volgende stappen doorlopen:

- standaardiseer de p variabelen zodat ze allen een gemiddelde van 0 en variantie van 1 hebben. Dit voorkomt dominantie van variabelen met grote variantie;
- bereken de covariantie matrix C van alle variabelen (nadat stap 1 is uitgevoerd is deze eigenlijk een correlatie matrix);

- bereken de eigenwaarden λ_1, λ_2,...,λ_p en de bijbehorende eigenvectoren a_1, a_2,...,a_p. De eigenwaarde λ_i geeft de variantie van component *Zi* en de eigenvector a_i de coëfficiënten van deze component;
- bepaal hoeveel hoofdcomponenten je kiest. Dit kun je doen op basis van een aantal criteria:
  - totale variantie die minimaal in hoofdcomponenten moet zitten (90 of 99%);
  - hoofdcomponenten met variantie (eigenwaarde) > 1;
  - statistische significantie;
  - scree plot (grafiek van eigenwaarden; zoek naar de knik/elleboog);
  - interpreteerbaarheid van hoofdcomponenten;
- interpreteer de hoofdcomponenten! Dit kan op basis van coëfficiënten a_{ij}. Gebruik daarbij de hoofdcomponent scores van waarnemingen. Wanneer de interpretatie moeilijk is kun je overwegen om een factoranalyse uit te voeren.



Figuur 10.1 Scree plot

Bovenstaande scree plot geeft de eigenwaarden van de verschillende componenten weer. Hierbij zoek je de eerste duidelijke knik of elleboog in de grafiek. Vanaf die hoofdcomponent wordt er relatief weinig variantie weergegeven in de hoofdcomponenten. In bovenstaande grafiek is dat duidelijk na de derde hoofdcomponent. Neem in dit geval niet meer dan drie hoofdcomponenten mee.

10.2.2 Factoranalyse

Bij factoranalyse volgen we een iets afwijkende procedure. Druk nu *p* gestandaardiseerde variabelen X_i uit in *m* ongecorreleerde factoren F_i met allen een gemiddelde van 0 en een variantie van 1:

$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m + e_i$

a_{ij} zijn de factorladingen
e_i is een variabele specifieke factor

Bij factoranalyse worden de variabelen uitgedrukt in een beperkt aantal factoren, terwijl hoofdcomponentenanalyse het omgekeerde doet, namelijk de hoofdcomponenten uitdrukken in variabelen. Bij factoranalyse zoeken we in principe ook slechts een beperkt aantal factoren. De variantie in een variabele kunnen we nu als volgt uitdrukken:

$$\begin{aligned}\text{var}(X_i) &= \text{var}(a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m) + \text{var}(e_i) \\ &= a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2 + \text{var}(e_i)\end{aligned}$$

De variantie is uit te splitsen in twee delen. Een gedeelte van de variantie is gerelateerd aan factoren ($a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2$). Dit noemen we de communaliteit. Daarnaast is er een deel variabele-specifieke variantie, weergegeven als var(e_i).

Het volgende voorbeeld dient om factoranalyse te verduidelijken. Hierbij gaan we uit van de examencijfers voor een aantal verschillende vakken: engels (E), frans (F), wiskunde (W), natuurkunde (N), scheikunde (S) en geschiedenis (G). Als we een factoranalyse uitvoeren blijkt dat we deze variabelen uit kunnen drukken in twee factoren:

$E = 0{,}20I + 0{,}80IJ + e_E$ $\quad N = 0{,}70I + 0{,}30IJ + e_N$
$F = 0{,}25I + 0{,}85IJ + e_F$ $\quad S = 0{,}60I + 0{,}30IJ + e_S$
$W = 0{,}80I + 0{,}20IJ + e_W$ $\quad G = 0{,}15I + 0{,}82IJ + e_G$

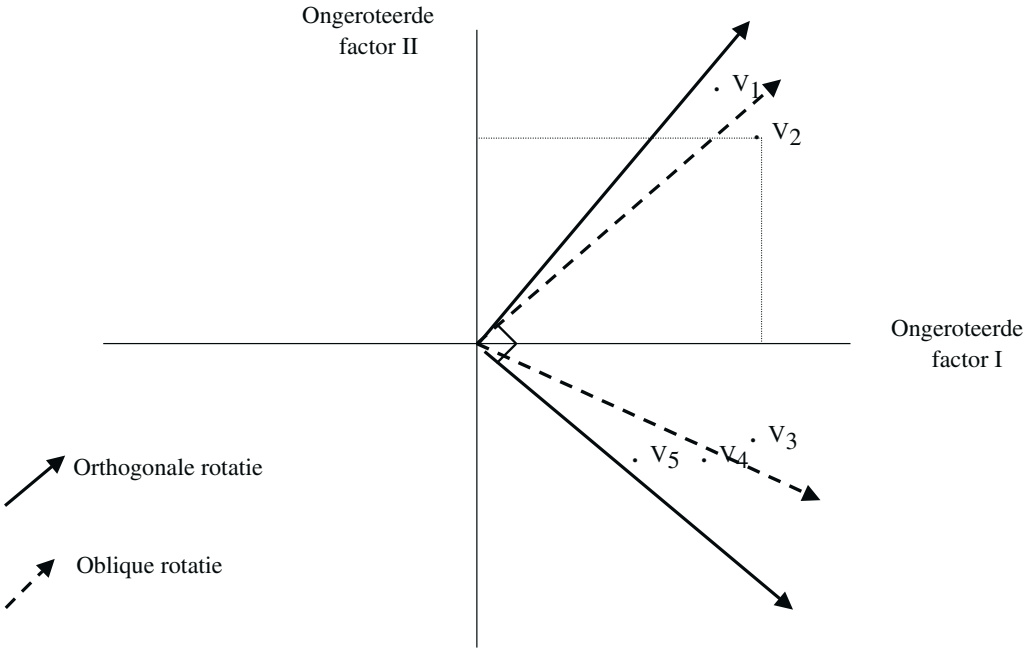
Deze factoren zouden we tijdens de interpretatiefase kunnen benoemen als intellect (I) en ijver (IJ). Een andere benoeming is ook mogelijk, deze hangt af van de interpretatie. Alle zes de vakken hebben bepaalde scores op deze factoren. Er is duidelijk een analogie met regressieanalyse. Het verschil is echter dat we vooraf geen waarnemingen hebben voor intellect en ijver. Factoren en factorladingen (coëfficiënten) worden gezamenlijk bepaald. Met deze factorladingen kun je ook een aantal dingen berekenen. Zo is de variantie van de communaliteit van bijvoorbeeld I met W 0,64. Verder is de correlatie tussen cijfers voor W en G: 0,80*0,15+0,20*0,82=0,284. Om een factoranalyse uit te voeren moet je een aantal stappen doorlopen:

- bereken de correlatie of covariantiematrix;
- bepaal de factorladingen a_{ij}. Dit kan bijvoorbeeld met de eerste *m* hoofdcomponenten (niet helemaal correct want e_i hangen dan af van overige *p-m* hoofdcomponenten en zijn daarom gecorreleerd). Principal axis factoring itereert ladingen verder op basis van be-

rekende communaliteiten. Deze methode erkent expliciet het verschil tussen de communale en de specifieke variantie;
- roteer de factoren om ze beter te kunnen interpreteren. Dit houdt in dat factorscores kleiner of groter worden;
- bereken de factorscores voor elke waarneming.

Roteren
Factoren en factorladingen zijn niet uniek. Je kunt variabelen op veel verschillende manieren uitdrukken in factoren. Het veranderen van de uitdrukking leidt tot veranderingen in de factorladingen. Dit noemen we het rotatieprobleem. Door de factoren te roteren kunnen ze doorgaans beter geïnterpreteerd worden. Let erop dat bij een gelijk aantal factoren de communaliteiten wel gelijk zijn. Er zijn twee hoofdsoorten van rotatie: orthogonale (loodrechte) en vrije of oblique rotatie. Voor elke hoofdgroep is er een aantal verschillende rotatiemethodes:
- voor orthogonale rotatie zijn de verschillende rotatiemethodes:
  - Varimax (variabele: hoge lading op één factor en lage op andere factoren);
  - Quartimax (alle variabelen hoge lading op één factor en verder elke variabele een hoge op andere factor en lage op andere);
  - Equimax.;
- Voor de oblique rotatie zijn dat:
  - Direct Oblimin;
  - Promax.

De volgende figuur geeft schematisch het roteren van 2 factoren weer.



Figuur 10.2 Rotatie van factoren

Met de aangegeven rotatie scoren de variabelen V_1 en V_2 hoger op factor II en lager op factor I dan zonder rotatie. De variabelen V_3, V_4 en V_5 scoren hoger op factor I en lager op factor II na rotatie. Door een oblique rotatie wordt dit effect nog wat versterkt ten opzichte van het effect bij een orthogonale rotatie.

10.2.3 Vergelijking hoofdcomponenten- en factoranalyse

Figuur 10.3 geeft een vergelijking van hoofdcomponenten- en factoranalyse.

Er is echter ook een aantal gemeenschappelijke elementen. Beide methoden zijn alleen bruikbaar als er voldoende correlatie tussen variabelen is. Je hebt als het ware groepen van variabelen nodig. De interpretatie van de hoofdcomponenten of de factoren kan voor beide methoden moeilijk zijn en is subjectief. Dit vraagt inzicht en creativiteit. Tot slot, zoals al eerder gesteld, hoofdcomponentenanalyse wordt vaak gezien als een speciaal geval van factoranalyse (zie SPSS). Uit de bespreking van beide methoden blijkt dat dit niet geheel terecht is.

	Hoofdcomponentenanalyse	Factoranalyse
1.	Zoek nieuwe variabelen (hoofdcomponenten)	Zoek onderliggende factoren in data
2.	Verklaar zoveel mogelijk van variantie in data.	Verklaar correlatie tussen variabelen in data
3.	Gemeenschappelijke en specifiek variantie	Op basis van gemeenschappelijke variantie
4.	Variabelen vormen index (hoofdcomponent)	Variabelen opgebouwd uit factoren
5.	Hoofdcomponenten loodrecht op assen	Factoren kunnen roteren voor interpretatie

Figuur 10.3 Vergelijking hoofdcomponenten- en factoranalyse

10.3 Clusteranalyse

Bij clusteranalyse vormen we homogene groepen van observaties uit een dataset van *n* waarnemingen met *p* variabelen. Waar we dus bij hoofdcomponenten- en factoranalyse variabelen probeerden samen te vatten, richten we ons bij clusteranalyse op (groeperen/clusteren van) waarnemingen. Elke groep/cluster bevat waarnemingen met dezelfde scores op bepaalde kenmerken. Er worden dus homogene groepen/clusters gevormd. De clusters verschillen op basis van die kenmerken. Het aantal clusters is echter op voorhand niet bekend. Daarin verschilt clusteranalyse van bijvoorbeeld de keuzemodellen zoals die in hoofdstuk 9 besproken worden. De clustering geschiedt op basis van een numerieke procedure. Voorbeelden van toepassingen van clusteranalyse zijn:

- bedrijfsstijlen onderzoek;
- onderscheiden van groepen consumenten;
- clusteren van multifunctionele landbouwbedrijven.

De clusters zijn gebaseerd op de 'afstand' tussen waarnemingen. Er zijn verschillende definities van afstand mogelijk. Een algemene afstandsdefinitie is:

$$D_{ij} = \sqrt[m]{\left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^n \right)}$$

waarbij *m* en *n* kunnen variëren. Aan de hand van de bovenstaande algemene formule kunnen verschillende afstandsmaten worden geformuleerd:

- meetkundige afstand (n=m=2);
- gekwadrateerde meetkundige afstand (n=2, m=1);
- Minkowski afstand (algemene definitie, n=m=2 maar met absolute afstanden);
- Block afstand (Minkowski, n=m=1).

Daarnaast zijn er methoden die gebruikmaken van correlatiecoëfficiënten.
Hieronder bespreken we twee typen van clusteranalyse: hiërarchische clustering en niet-hiërarchische clustering.

Hiërarchische clusteranalyse
Hiërarchische clustering gaat stapsgewijs. Eerst wordt een cluster gevormd van waarnemingen met de kleinste onderlinge afstand. Vervolgens wordt gekeken naar de volgende kleinste afstand voor de vorming van een nieuw cluster of uitbreiding van een al bestaand cluster. Dit gaat zo door totdat alle waarnemingen ingedeeld zijn in een aantal clusters. Dit aantal moet vooraf worden opgegeven, anders wordt slechts één cluster gevormd. Nadat je een afstandsdefinitie hebt gekozen is het wel de vraag hoe je met afstanden omgaat. Hier zijn verschillende methodes voor:

- centroid clustering (afstand tot gemiddelde waarneming in cluster);
- median clustering (afstand tot mediaan waarneming in cluster);
- nearest-neighbour (afstand tot dichtstbijzijnde waarneming in cluster);
- farthest-neighbour (afstand tot meest verre waarneming in cluster);
- average linkage methoden (gemiddelde afstand tot alle waarnemingen in cluster);
- Ward's-methode (op basis van minimale RSS in cluster; homogeniteit).

Bij hiërarchische clustering is er elke stap één cluster minder. De vraag is echter wanneer te stoppen met clusteren? Je kunt vooraf een aantal clusters opgeven, maar hoe bepaal je nu het aantal gewenste clusters? Daar is een aantal maatstaven voor, die het beste gezamenlijk kunnen worden beoordeeld.

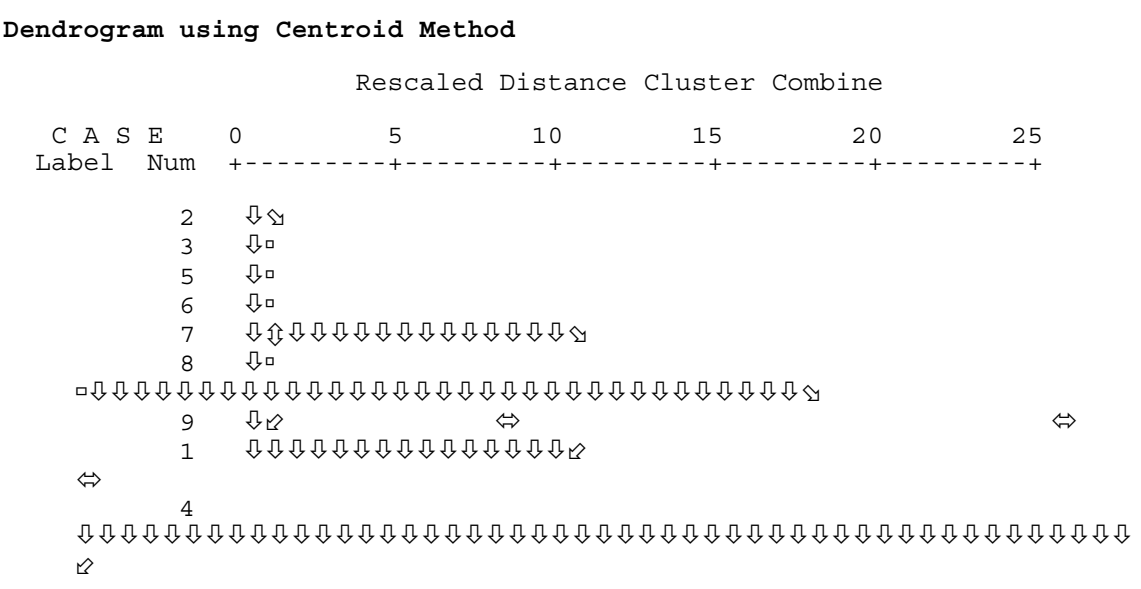
- root-mean-square standard deviation van een cluster:

$$RMSSTD = \sqrt{\sum_{j=1}^{p} \hat{s}_j^2 \Big/ p}$$

Deze maatstaf kan gebruikt worden om te kijken hoeveel de standaardafwijkingen van variabelen veranderen in een cluster als je één waarneming toevoegt;

- $R^2 = \frac{SS_{Between}}{SS_{Totaal}}$

  Deze maatstaf is gebaseerd op het feit dat de som van de kwadraten van variabelen binnen een cluster zo klein mogelijk moet zijn, maar tussen clusters zo groot mogelijk. Bij een maximale som van de kwadraten tussen clusters heeft deze R^2 de waarde 1 (maximale verschillen tussen clusters) en als er geen verschillen tussen de clusters zijn (geen onderscheid) dan heeft R^2 een waarde 0;
- semi-partiële R^2. Vergelijk de SS_{nieuw} en de som van de SS_{oud} van beide samengevoegde clusters. Deze maatstaf vergelijkt de homogeniteit van het nieuw gevormde cluster met de homogeniteit van de twee 'oude' samengevoegde clusters. Als de SS_{nieuw} veel groter is dan de som van de 'oude' *SS* dan wordt de homogeniteit erg aangetast en is het wellicht beter niet te clusteren;
- afstand tussen clusters (afhankelijk van methode). Om het aantal clusters op basis van afstand te bepalen is het handig gebruik te maken van een dendrogram. Deze geeft schematisch weer welke (clusters van) waarnemingen zijn samengevoegd bij welke afstand (zie figuur 10.4).



Figuur 10.4 Dendrogram using centroïd method

Niet-hiërarchische clusteranalyse

Niet-hiërarchische clusteranalyse is een andere methode om een clusteranalyse uit te voeren. De stappen in een dergelijke analyse zijn als volgt:

- verdeel de data in een van tevoren bepaald aantal (*k*) groepen (in SPSS heet dit: K- Means clustering);

- observaties worden in de niet-hiërarchische clusteranalyse ingedeeld in clusters op basis van afstand tot het clustergemiddelde (centroid);
- bereken het nieuwe gemiddelde na het toevoegen van een observatie of cluster aan een cluster;
- herhaal stappen 2 en 3 tot er geen of nog maar zeer kleine veranderingen optreden in de clustergemiddeldes.

Deze manier van clusteren verschilt met name van hiërarchische clustering doordat het clusterproces iteratief is. Clusters kunnen worden samengevoegd maar ook weer gesplitst worden. Dit laatste is niet mogelijk bij hiërarchische clustering dit is meer éénrichtingsverkeer. In het iteratieve proces is een aantal variaties mogelijk doordat gebruikgemaakt wordt van verschillende startwaarden en verschillende updatemogelijkheden.

Welke methode te gebruiken?
Het verschil tussen hiërarchische en niet-hiërarchische clusteranalyse roept de vraag op welke methode het beste gebruikt kan worden? De verschillen tussen beide methodes zijn weergegeven in figuur 10.5 (met + voor positieve eigenschap, en - voor negatieve eigenschap):

	Hiërarchische clustering	Niet-hiërarchische clustering
1.	Aantal clusters niet van tevoren bepaald (+)	Aantal clusters van tevoren bepaald (-)
2.	Waarnemingen kunnen niet opnieuw ingedeeld worden (-)	Waarnemingen kunnen van cluster veranderen (+)
3.	Verschillende methoden geven verschillende resultaten, afhankelijk van datastructuur (+/-)	
4.		Mogelijkheid tot testen van tevoren bepaalde groepen (+)

Figuur 10.5 Vergelijking hiërarchische en niet-hiërarchische clusteranalyse

In algemene zin is het aan te bevelen om eerst een hiërarchische clustering te doen, gevolgd door niet-hiërarchische clustering, om te zien of het iteratieve proces nog veel verandert in de gevormde clusters. Tot slot nog een aantal algemene opmerkingen:
- clusteranalyse is alleen zinvol als er variabelen zijn die verschillen tussen (groepen) van waarnemingen veroorzaken. Clustervorming moet ergens op gebaseerd zijn;
- de interpretatie van de gevormde clusters vergt creativiteit. Interpretatie is vaak handig op basis van clustergemiddelden;
- om de betrouwbaarheid van clustering na te gaan is het (bij een voldoende grote dataset) aanbevolen om de clustering na te gaan op basis van subsamples;
- maak een vergelijking met externe indicatoren indien deze voorhanden zijn (bedrijfsscores, aandelenkoersen enzovoorts).

10.4 Casestudie

In deze casestudie gaan we informatie over 104 verschillende landen, vastgelegd in 15 kernvariabelen analyseren met behulp van factoranalyse en clusteranalyse. De SPSS-file landen.sav bevat de benodigde gegevens. De file is te vinden via de intranetsite van het Domeinteam Data en Modellen. Open deze file in SPSS. De variabelenamen staan boven aan elke kolom. Als je met de cursor op een variabelenaam gaat staan krijg je een variabelelabel te zien met een omschrijving van elke variabele. Omdat SPSS bij deze technieken veel output genereert, wordt alleen de kernoutput gegeven.

A. Factoranalyse
A1. Voordat er een factoranalyse wordt uitgevoerd, worden de variabelen eerst gestandaardiseerd (zodat voor elke variabele geldt dat het gemiddelde 0 is en de standaard afwijking 1). SPPS noemt dit Z-waarden. Dit kun je doen bij Analyze, Descriptive Statistics ▸, Descriptives... . Vink hier Save standardized values as variables aan. Doe dit voor alle variabelen. Let erop dat *life expectancy males* wordt hernoemd om verwarring met *life expectancy females* te voorkomen.

A2. Voer een factoranalyse uit op basis van alle beschikbare gestandaardiseerde variabelen. Kies Analyze, Data Reduction ▸, Factor... Vervolgens selecteer je variabelen waarmee je de factoranalyse wilt uitvoeren. Vervolgens zijn er vijf knoppen waarmee je van alles kunt (de-)selecteren:

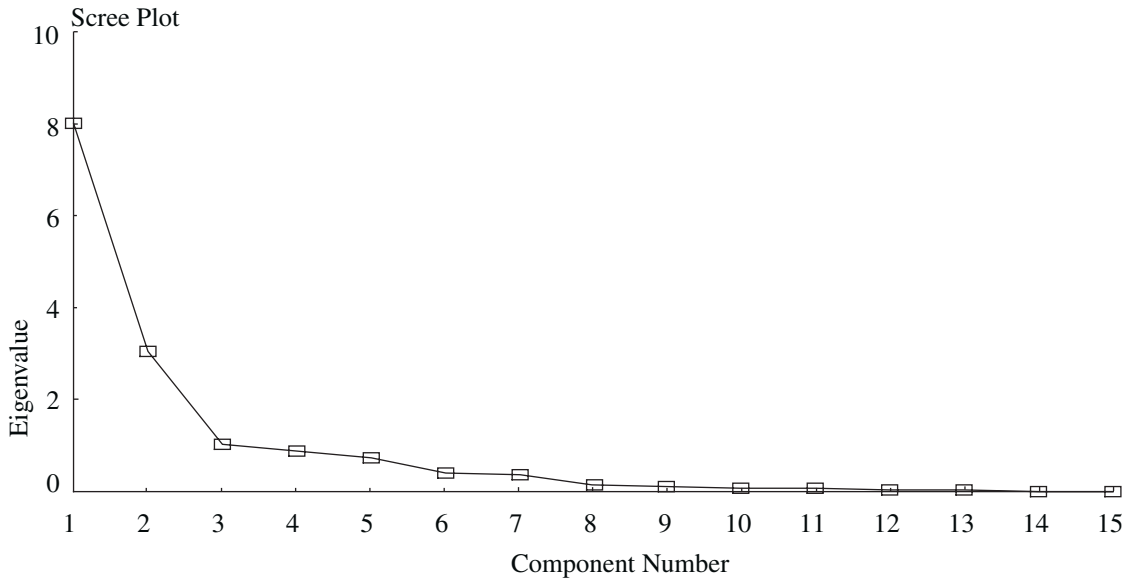
- Descriptives... geeft de mogelijkheid een aantal statistieken op te vragen. Hiermee kun je onder andere vaststellen of er voldoende correlatie is tussen de variabelen. Is er weinig onderlinge correlatie dan is factoranalyse geen bruikbare techniek. Selecteer Univariate descriptives, Initial solution en Coefficients.
- Bij Extraction... kun je onder andere de methode kiezen om de factoranalyse uit te voeren. Kies hier voor een Scree plot en verander verder niets.

Aan de overige knoppen verander je niets. Voer de factoranalyse uit door op Ok te drukken.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.023	53.488	53.488	8.023	53.488	53.488
2	3.042	20.277	73.766	3.042	20.277	73.766
3	1.047	6.981	80.747	1.047	6.981	80.747
4	.873	5.822	86.569			
5	.723	4.819	91.389			
6	.411	2.738	94.127			
7	.361	2.406	96.533			
8	.163	1.088	97.621			
9	.112	.748	98.369			
10	8.349E-02	.557	98.926			
11	7.451E-02	.497	99.422			
12	3.464E-02	.231	99.653			
13	3.000E-02	.200	99.853			
14	1.424E-02	9.496E-02	99.948			
15	7.748E-03	5.166E-02	100.000			

Extraction Method: Principal Component Analysis.



Figuur 10.6 Factoranalyse scree plot

Aan de hand van de bovenstaande output kunnen we de volgende vragen stellen:

- Heeft het op basis van de correlatiecoëfficiënten zin om een factoranalyse uit te voeren?
  Ja, er is een behoorlijk aantal variabelen dat onderling sterk correleert.
- Bekijk de communaliteiten. Welke variabelen zijn sterk gerelateerd aan de factoren?
  Dat zijn er een behoorlijk aantal. Bijvoorbeeld levensverwachting van vrouwen/mannen, kindersterfte, geboorteratio enzovoorts
- Hoeveel factoren selecteer je als je kijkt naar eigenwaarden?
  Drie. Er zijn drie hoofdcomponenten met een eigenwaarde > 1, dus kun je ook met drie factoren gaan werken.
- Hoeveel factoren selecteer je als je kijkt naar de scree plot?
  Ook drie. De knik of elleboog vindt plaats na drie hoofdcomponenten. Dat betekent dat er daarna nog maar weinig variantie verklaard wordt door volgende componenten.

A3. Voer opnieuw dezelfde factoranalyse uit, maar kies ervoor om slechts twee factoren te genereren (Extraction... Number of factors) én om te roteren. Dit doe je bij Rotation... Kies eerst voor de orthogonale Varimax rotatie. Vraag om de Rotated solution en een Loading plot. Voer vervolgens een oblique rotatie uit (Direct Oblimin) en vraag opnieuw om de Rotated solution en een Loading plot. Omdat de loading plot moeilijk te lezen is met de lange namen voor variabelen, kun je ervoor kiezen ze korte namen te geven.

Total Variance Explained

	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.023	53.488	53.488	6.478	43.189	43.189
2	3.042	20.277	73.766	4.586	30.576	73.766

Extraction Method: Principal Component Analysis.

Rotated Component Matrix [a]

	Component	
	1	2
Zscore: Number of people / sq. kilomete	.183	-5.28E-02
Zscore: People living in cities (%)	.798	-.144
Zscore: Average female life expectancy	.950	-.270
Zscore(LIFEEXPM) Average male life expe	.955	-.196
Zscore: People who read (%)	.779	-.456
Zscore: Population increase (% per year	-.368	.864
Zscore: Infant mortality (deaths per 10	-.903	.334
Zscore: Gross domestic product / capita	.519	-.565
Zscore: Region or economic group	-.105	.796
Zscore: Birth rate per 1000 people	-.738	.634
Zscore: Death rate per 1000 people	-.858	-.408
Zscore: Number of aids cases / 100000 p	-.527	-.205
Zscore: Birth to death ratio	.172	.915
Zscore: Fertility: average number of ki	-.742	.584
Zscore: Predominant climate	.122	-.803

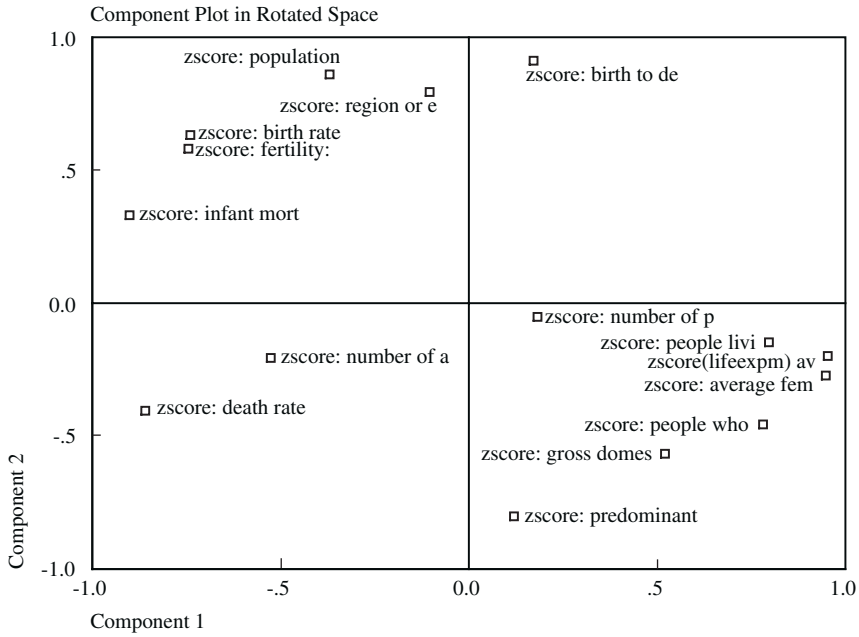
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Transformation Matrix

Component	1	2
1	-.831	.557
2	.557	.831

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.



A4. Probeer aan de hand van de factorscores en de loading plots de beide factoren te benoemen.
Dat blijkt tamelijk moeilijk te zijn. De eerste factor is gemakkelijk aan te duiden als welvaartsniveau van een land. De tweede factor scoort hoog op alles wat met bevolkingsgroei te maken heeft en kan op deze wijze geduid worden.

B. Clusteranalyse
B1. Voer een hiërarchische clusteranalyse uit op basis van dezelfde variabelen als bij A. Kies Analyze, Classify ▸, Hierarchical Cluster... . Vervolgens selecteer je variabelen waarmee je de clusteranalyse wilt uitvoeren. Ook hier is het belangrijk om met gestandaardiseerde waarden te werken. Je kunt daarom de bij A1 aangemaakte variabelen gebruiken. SPSS geeft je bij clusteranalyse ook direct de mogelijkheid om variabelen te transformeren. Ga naar Method... en kies onder Transform Values bij Standardize voor Z scores en vink By variable aan. Kies hier ook bij Cluster Method voor Centroid clustering en onder Measure voor Euclidian distance. Kies Continue. Terug in het hoofdwindow voor hiërarchische clusteranalyse kies je bij Plots... voor een Dendrogram, en kies je bij Icicle voor None. Kies Continue en voer de clusteranalyse uit. Dit kan enige tijd duren. Probeer het dendrogram te interpreteren (dubbel-klik in het plaatje om het geheel te kunnen zien).

B2. Een probleem bij het vorige onderdeel is dat het moeilijk is om na te gaan hoeveel clusters je het beste kunt onderscheiden. Dit is vaak een probleem bij grote databestanden. Daarom kun je er bij hiërarchische clustering ook voor kiezen om een opgegeven aantal clusters te creëren.

Dit doe je bij Statistics... Bij Single solution kun je een exact aantal clusters (bijvoorbeeld vijf) opgeven. Je kunt ook kiezen voor een Range of solutions. Doe dit laatste en kies voor vier tot zes clusters. In de output kun je de clusterindeling vergelijken op basis van het

aantal clusters. Clusters met één of twee waarnemingen zijn blijkbaar moeilijk in te delen. Ga na welke landen moeilijk in te delen zijn. Aan welke variabelen ligt dat? Hoe los je dit op?

Er is hier een aantal landen dat moeilijk in te delen is. Het gaat hier om de cases 85, 44, 104 en 98 oftewel Singapore, Hongkong, Zambia en Uganda. De eerste twee zijn zeer dichtbevolkte, westers georiënteerde en welvarende Aziatische landen. Zij scoren extreem hoog op bevolkingsdichtheid. Zambia en Uganda scoren extreem hoog op de aids ratio. Deze twee kleine groepen hebben een behoorlijk verstorend effect op de clustering. Je zou opnieuw een clusteranalyse uit kunnen voeren waarbij deze landen niet worden meegenomen. Een andere optie is om deze variabelen niet mee te nemen in de analyse.

B3. Voer opnieuw een clusteranalyse uit, maar nu zonder de moeilijk in te delen landen. Vraag om exact zes clusters. Bij Save... kun je nieuwe variabelen aanmaken die de clusterindeling aangeven. Kies ook hier voor zes clusters. Vergelijk de clusterindeling met clusterindelingen gebaseerd op andere clustermethoden: Nearest neighbour, Furthest Neighbour en Ward's method. Gebruik Euclidian distance. Zijn er veel verschillen?

Er zijn grote verschillen in de gebruikte methodes. Bij de centroi- methode worden veel landen in een cluster (cluster twee) ingedeeld. Cluster vier, vijf en zes bestaan allen uit één land (Bangladesh, Barbados en USA respectievelijk). Bij de single linkage-methode een vergelijkbaar beeld, maar nu bijna alle landen in cluster één ingedeeld. De complete linkage-methode en Ward's-methode leiden tot een betere differentiatie in groepen. De conclusie is dat het nogal wat uitmaakt welke methode men gebruikt. En dan hebben we de afstandsmaat nog constant gehouden. Van belang is te letten op uitschieters in de waarnemingen en te experimenteren met verschillende methoden en afstandsmaten.

B4. Vervolgens doen we een niet-hiërarchische clustering met zes clusters. Let erop dat je hier niet kunt kiezen voor gestandaardiseerde waarden (Z scores) van variabelen. Deze moet je dus als bij A1 aanmaken.

Kies Analyze, Classify ▸, K-means Cluster... Selecteer dezelfde variabelen (gestandaardiseerde waarden / Z scores). Kies voor zes clusters. Bij Save... vraag je om de clusterindeling op te slaan. Eventueel kun je het aantal iteraties verhogen.

Number of Cases in each Cluster

Cluster	1	18.000
	2	40.000
	3	2.000
	4	37.000
	5	1.000
	6	2.000
Valid		100.000
Missing		.000

Vergelijk de clusterindeling met die van de hiërarchische clusteranalyse.
De niet-hiërarchische clusteranalyse zorgt, net als sommige hiërarchische methodes, voor een redelijke differentiatie in de groepen. Tussen Ward's-methode en de niet-hiërarchische clusteranalyse is duidelijk een patroon te ontdekken.

10.5 Literatuur

Tot slot een drietal bruikbare boeken voor een diepgaandere bespreking:
- Manly, B.F.J. (1994);
- Sharma, S. (1996);
- Hair, J.F., Anderson, R.E., Tatham, R.L. en Black, W.C. (1998).

Het boek van Manly is net als dit hoofstuk bedoeld als een eerste kennismaking met deze technieken. Naast hoofdcomponenten-, factor- en clusteranalyse bespreekt het ook een aantal andere gangbare multivariate technieken op een bondige manier. Sharma is een gedegen en duidelijk boek dat meer technisch ingaat op de besproken methodes. Dit boek is een aanrader. Hair ten slotte is een veel gebruikt standaard werk dat echter te veel tekst en te weinig details bevat. Voor praktische vragen is dit boek minder geschikt.

11. Multidimensional scaling en conjunctanalyse

11.1 Inleiding

Multidimensional scaling (MDS) en conjunctanalyse (CA) zijn beide methoden waarmee percepties en voorkeuren van consumenten/respondenten kunnen worden gemeten en weergegeven. In tegenstelling tot de andere methoden die in deze cursus aan de orde zijn geweest, kunnen beide methoden worden toegepast op een individuele respondent. MDS en CA zijn recente toevoegingen aan SPSS. In SPSS kan slechts een beperkte verzameling MDS-analyses worden uitgevoerd. CA kan uitsluitend via de syntaxvensters worden uitgevoerd.

In paragraaf 11.2 wordt MDS beschreven. Paragraaf 11.3 behandelt vervolgens CA en in paragraaf 11.4 wordt een overzicht gegeven van de relevante literatuur.

11.2 Multidimensional scaling

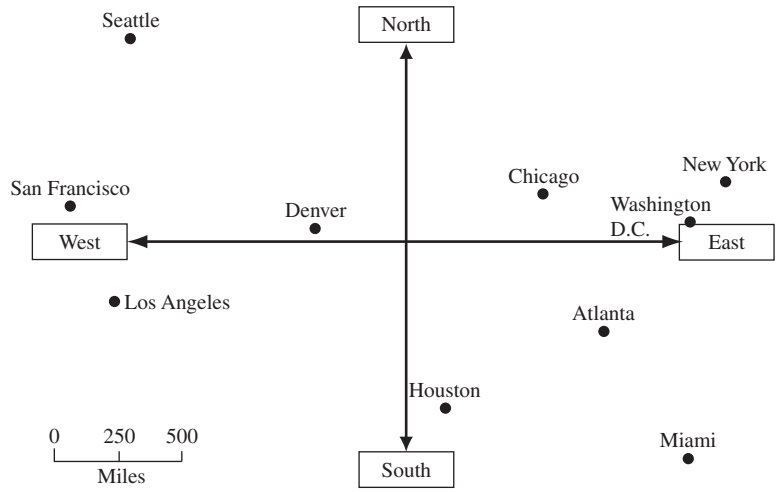
Het doel van MDS is te komen tot een ruimtelijke weergave van percepties of preferenties. Een afgeleide doelstelling is het bepalen van de dimensies waarop de vergelijkingen en of voorkeuren zijn gebaseerd. In paragraaf 10.2.1 wordt MDS aan de hand van een voorbeeld geïllustreerd. In paragraaf 10.2.2 wordt een stappenplan gegeven voor het uitvoeren van MDS.

11.2.1 Voorbeeld MDS

Het principe van MDS kan geïllustreerd worden aan de hand van onderstaand voorbeeld (Green et al., 1988). In figuur 11.1 is een vereenvoudigde kaart van de VS afgebeeld. Op basis van deze kaart kan een afstandentabel zoals in figuur 11.2 worden gemaakt. Hierbij wordt dus op basis van een ruimtelijke voorstelling een afstandentabel gemaakt. Het principe van MDS gaat de andere kant op. Op basis van een afstanden tabel wordt een ruimtelijke voorstelling gemaakt. Je kunt je voorstellen dat, als je heel veel geduld zou hebben, je op basis van de afstanden in figuur 11.2, een plaatje zoals in figuur 11.1 zou kunnen construeren. Er is maar één mogelijke samenstelling waarbij het plaatje alle afstanden goed representeert. Wel kan het kaartje worden gedraaid (andere noord zuid richting).

Het principe van MDS is dus op basis van een afstandentabel een ruimtelijke weergave te construeren die de afstanden zo goed mogelijk weergeeft. Bij MDS gaat het meestal niet over fysieke afstanden maar om psychologische afstanden. De respondent moet dan aangeven hoe ongelijk of gelijk bepaalde stimuli (bijvoorbeeld producten zijn). Indien deze beoordelingen paarsgewijs worden gemaakt kan een soort afstandtabel worden gemaakt die de afstanden tussen de producten weergeeft. Deze tabel is de invoer voor MDS die een n-dimensionale ruimte probeert te construeren waarin de afstanden in deze ruimte zo veel mogelijk overeenkomen met de afstandstabel. Het doel van de MDS is dan ook: het plaatsen van merken in een meer-dimensionale ruimte op basis van gelijkheids- of voorkeurs- uitspraken en het achterha-

len van de dimensies die aan deze uitspraken ten grondslag liggen.



Figuur 11.1 Kaart van de Verenigde Staten

Cities	Atla.	Chic.	Denv.	Hous.	L.A.	Miam.	N.Y.	S.F.	Seat.	Wash. D.C.
Atlanta		587	1212	701	1936	604	748	2139	2182	543
Chicago			920	940	1745	1188	713	1858	1737	597
Denver				879	831	1726	1631	949	1021	1494
Houston					1374	968	1420	1645	1891	1220
Los Angeles						2339	2451	347	959	2300
Miami							1092	2594	2734	923
New York								2571	2408	205
San Francisco									678	2442
Seattle										2339
Washington D.C.										

Figuur 11.2 Afstandstabel

MDS kan bijvoorbeeld worden toegepast bij imagometing, marktsegmentatie, productontwikkeling en bij de evaluatie van een reclamecampagne. De benodigde invoerdata bestaat uit:

- gelijkheidsbeoordelingen (ratio, interval, ordinaal of nominaal geschaald) of,
- voorkeursuitspraken (ratio, interval of ordinaal geschaald)

Er moet een duidelijk onderscheid worden gemaakt tussen beide. Voorkeursbeoordelingen kunnen op basis van andere dimensies worden gemaakt dan gelijkheidsbeoordelingen. Indien vakantiebestemmingen worden vergeleken dan kan het zijn dat in de gelijkheidsbeoordelingen twee bestemmingen als totaal anders worden gezien, maar indien de respondent beide bestemmingen graag zou bezoeken dan kunnen deze bij preferentiebeoordelingen dicht bij elkaar in de buurt komen te liggen.

11.2.2 Stappen in het uitvoeren van een MDS-analyse

Probleemformulering
In deze fase moet het doel van het onderzoek worden geformuleerd. In het verlengde hiervan moeten de stimuli (de te vergelijken objecten) worden bepaald. De stimuli zijn medebepalend voor de dimensies die gevonden zullen worden. Ten aanzien van het aantal stimuli verdient het aanbeveling minstens acht en maximaal vijfentwintig stimuli op te nemen.

Verkrijgen van invoerdata
Hierbij moeten paarsgewijze vergelijkingen worden gemaakt tussen alle mogelijke combinaties van stimuli. De vraagstelling zou bijvoorbeeld de volgende vorm kunnen hebben. In hoeverre vindt u een Opel Kadett vergelijkbaar met een ford escort. Dit kan worden beantwoord op een negenpuntsschaal waarbij de uiteinden zijn gelabeld van zeer onvergelijkbaar tot zeer vergelijkbaar. Het aantal paarsgewijze vergelijkingen bedraagt: n * (n - 1). Deze paarsgewijze vergelijkingen resulteren in een afstandstabel.

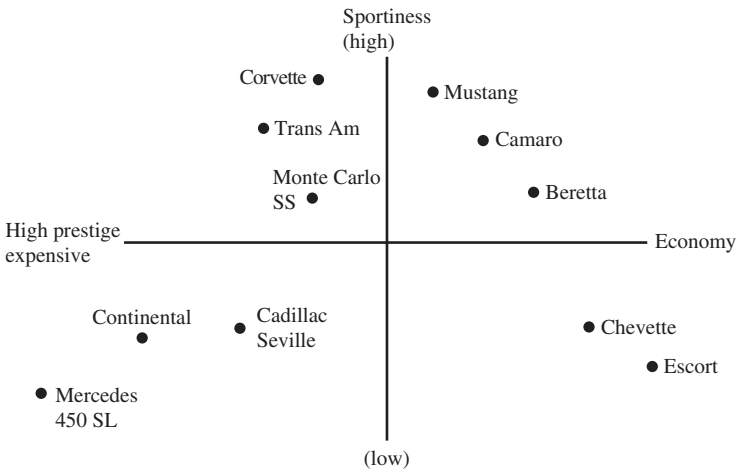
Selecteren MDS-procedure
Afhankelijk van de verzamelde data (preferenties of gelijkenis; nominaal, ordinaal, interval- of ratiogeschaalde gegevens) en de vraag of je analyses wilt uitvoeren op groeps- of individueel niveau, dient de juiste procedure geselecteerd en toegepast te worden (zie Green et al., 1989 voor een beschrijving van alle mogelijkheden), in dit voorbeeld is sprake van interval geschaalde gelijkheidsbeoordelingen.

Keuze van aantal dimensies
Vervolgens moet het aantal dimensies worden vastgesteld. Hierbij geldt een afweging tussen een betere fit met meer dimensies en een eenvoudigere interpretatie met een kleiner aantal dimensies. De fit neemt toe met het toenemen van het aantal dimensies, net zoals de fit van een regressiemodel verbetert indien extra variabelen worden toegevoegd. Voor het vaststellen van het aantal dimensies is een aantal technieken / vuistregels beschikbaar:

- ten eerste kan men gebruikmaken van de stress van de oplossing. De stress is een maatstaf voor de badness of fit. Des te groter de stresswaarde, des te slechter de oplossing. De stresswaarde kan worden uitgezet ten opzichte van het aantal dimensies. Vervolgens kan men op zoek gaan naar het punt waarbij een uitbreiding van het aantal dimensies slechts een geringe vermindering van de stress laat zien (een knik in de figuur);
- interpreteerbaarheid van de oplossing is bepalend voor de keuze van het aantal dimensies;
- vuistregel: het aantal dimensies is kleiner of gelijk aan het aantal stimuli gedeeld door vier. Bij acht stimuli kan men dan ten hoogste twee dimensies gebruiken;

- in de pratijk komt men meestal uit op twee of drie dimensies.

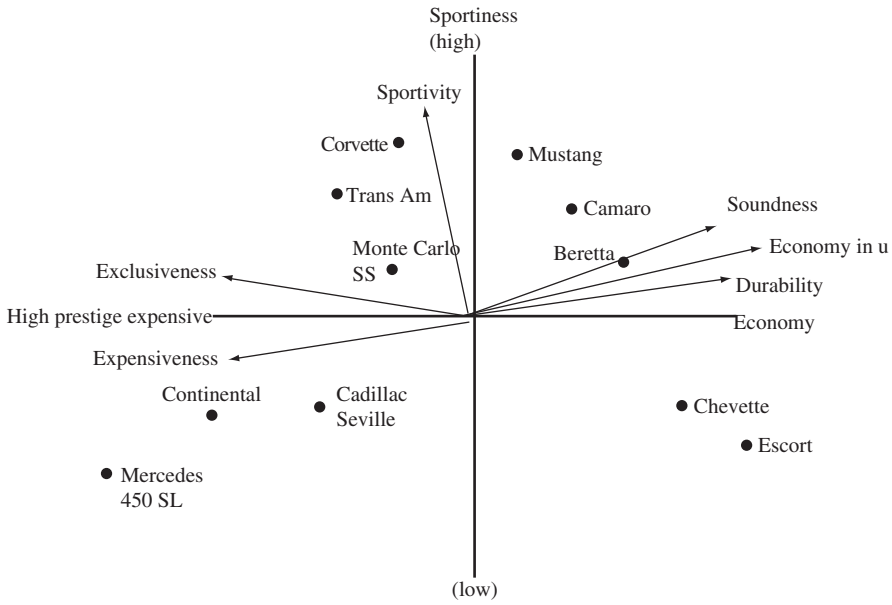
Interpreteer oplossing
Indien het aantal dimensies is vastgesteld moet de oplossing worden geïnterpreteerd. Hierbij probeert men vast te stellen welke dimensies in de beoordelingen van de respondent hebben meegespeeld. Bij de interpretatie kan een aantal strategieën worden gevolgd. Men kan de respondenten vragen welke dimensies een rol hebben gespeeld bij het maken van de beoordelingen. Ook kan men de respondent de oplossing voorleggen en vervolgens om een interpretatie vragen. Een andere mogelijkheid is dat de onderzoeker op basis van de kennis van de producten tot een interpretatie komt. Op basis van de kennis van de automerken kan de onderzoeker bijvoorbeeld concluderen dat de verticale as de mate van sportiviteit weergeeft. De horizontale as geeft iets weer van de zuinigheid van de auto en het prestige.



Figuur 11.3 Oplossingsruimte plus interpretatie

De meest geavanceerde oplossing omhelst het laten beoordelen van de producten op een aantal kenmerken. Dit kan door middel van een vraagstelling zoals: 'In hoeverre vindt u de escort sportief?' Deze vraag kan bijvoorbeeld worden beoordeeld op een zevenpuntsschaal met als uiteinden zeer onsportief en zeer sportief.

Deze kenmerken kunnen vervolgens in de oplossingsruimte worden afgebeeld als vectoren. Deze vectoren kunnen bij de interpretatie van de uitkomstenruimte worden gebruikt. Naarmate een as dichter bij een vector ligt zal deze vector een belangrijkere rol spelen bij de interpretatie van die as (zie figuur 11.4). Sportiviteit is een vector die dicht bij de verticale as ligt. Deze as hangt dus sterk met sportiviteit samen. Deze methode wordt property fitting genoemd. Helaas wordt deze niet ondersteund door SPSS.



Figuur 11.4 Interpretatie door middel van property fitting

Bepaal validiteit en betrouwbaarheid
Bij de beoordeling van de kwaliteit van de geconstrueerde oplossingsruimte (zie figuur 11.3) speelt de stress een belangrijke rol. Zoals gezegd geeft de stress de slechtheid van de oplossing weer. Het brengt tot uitdrukking in hoeverre de af te leiden afstanden uit de oplossingsruimte overeenstemmen met de afstanden zoals die door de respondent zijn beoordeeld. Des te hoger de stresswaarde des te groter het verschil in afstanden in de afstandentabel en de afstanden in de oplossingsruimte. Een veel gebruikte maatstaf is Kruskal's stress:

$$\text{Stress} = \sqrt{\frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j (d_{ij} - \bar{d})^2}}$$

$\bar{d}$ gemiddelde afstand ($\sum d_{ij} / n$) op de kaart

$\hat{d}_{ij}$ afgeleide afstand

d_{ij} afstanden zoals beoordeeld door respondenten

Kruskal geeft de volgende vuistregels voor verschillende stresswaarden (stresswaarde ligt tussen 0 en 1).

Tabel 11.1 Stresswaarden

Stresswaarde	Mate van fit
0,2	Poor
0,1	Fair
0,05	Good
0,025	Excellent
0	Perfect

11.3 Conjunctanalyse

Conjunctanalyse beschouwt een product als een pakketje attributen. Een televisie heeft bijvoorbeeld een beeldbuis met een bepaalde omvang, al dan niet een afstandbediening en al dan niet teletekst. Conjunctanalyse probeert vervolgens het nut van attribuutniveaus (part worth oftewel deelnut) en het belang van attributen te achterhalen.

Conjunctanalyse heeft als doel het achterhalen van het nut van attribuutniveaus en het belang van attributen van producten en eventueel de verschillen tussen respondenten.

Conjunctanalyse kan bijvoorbeeld worden toegepast om het belang van attributen in de aankoopbeslissing vast te stellen, marktaandelen te voorspellen, productontwerp te ondersteunen of de markt te segmenteren. De data neemt de vorm aan van voorkeuren van respondenten voor (hypothetische) producten gemeten op een interval- of rangordeschaal. Op basis van deze voorkeuren wordt getracht het nut van niveaus van attributen vast te stellen zodanig dat die de invoerbeoordelingen zo goed mogelijk weerspiegelen.

11.3.1 Stappen in conjunctanalyse

Formuleer het probleem

In de beginfase moet de doelstelling van de conjunctanalyse worden geformuleerd. Een belangrijk element is de keuze van de attributen die in het onderzoek worden meegenomen. De attributen kunnen vastgesteld worden aan de hand van een literatuuronderzoek of een (groeps)discussie. Daarnaast moet worden aangegeven welk niveau van de attributen wordt meegenomen in het onderzoek. Hierbij geldt een pragmatische overweging dat men niet al te veel niveaus moet meenemen omdat het aantal te maken beoordelingen sterk toeneemt. Daarnaast geldt dat er enigszins een evenwicht moet zijn tussen het aantal niveaus van de verschillende attributen.

In het vervolg van deze beschrijving wordt een voorbeeld van sportschoenen uitgewerkt (Malhotra, 1993). Er worden drie attributen onderscheiden: de zool, het bovenwerk en de prijs. De attribuut niveaus zijn in onderstaande tabel weergegeven.

Tabel 11.2 Attributen

Attribuut	Omschrijving	Nr
Zool	Rubber	1
	Polyurethane	2
	Plastic	3
Bovenwerk	Leer	1
	Canvas	2
	Nylon	3
Prijs	$15.00	1
	$30.00	2
	$45.00	3

Construeren van de profielen

Op basis van de onderscheiden attributen en attribuutniveaus kunnen (hypothethische) producten worden samengesteld door alle mogelijk combinaties van attribuutniveaus te bepalen. Een dergelijke (hypothetische) product wordt een profiel genoemd. Zo kan men bijvoorbeeld een schoen samenstellen met een rubberen zool, een leren bovenwerk en een prijs van dertig dollar. Het definiëren van alle mogelijke combinaties wordt het factoriele design genoemd. Bij drie attributen met drie niveau levert dit zevenentwintig mogelijke combinaties op. Al deze combinaties zouden aan de respondent kunnen worden voorgelegd.

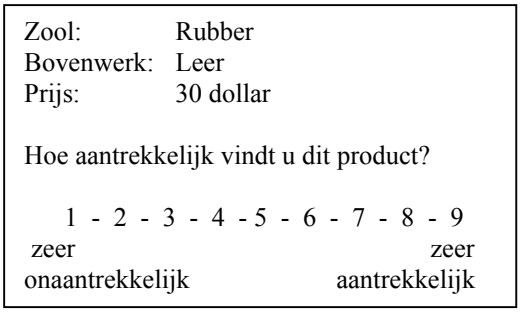
Bij een groter aantal attributen of niveaus resulteert dit al gauw in een onhandelbaar groot aantal combinaties. Om het aantal combinaties te beperken kan gebruik worden gemaakt van zogenaamde fractionele ontwerpen. Hierbij wordt slechts een deel van de mogelijke combinaties aan de respondent voorgelegd. De combinaties kunnen niet willekeurig worden gekozen. De combinaties moeten zodanig worden samengesteld dat de hoofdeffecten onafhankelijk kunnen worden geschat. Voor het samenstellen van deze profielen kan gebruik worden gemaakt van gepubliceerde tabellen (zie bijvoorbeeld Addelman, 1962) of van de SPSS functie Orthoplan. Het nadeel van de fractionele opzet is dat niet alle interactie-effecten kunnen

Tabel 11.3 Profielen

Profiel	Zool	Bovenwerk	Prijs
1	1	1	1
2	1	2	2
3	1	3	3
4	2	1	2
5	2	2	3
6	2	3	1
7	3	1	3
8	3	2	1
9	3	3	2

In het onderstaande fractionele ontwerp worden negen profielen geconstrueerd (ten opzichte van de zevenentwintig in een compleet ontwerp). Het nummer geeft het nummer van het attribuutniveau weer (zie tabel 11.3).

Bepaal de vorm van de invoerdata
Net als bij MDS bestaan er veel varianten ten aanzien van de invoerdata in conjunctanalyse. Varianten zijn paarsgewijze vergelijken, rangordenen of het toekennen van een score aan de profielen (zie Green et al. voor een beschrijving van de diverse alternatieven). Hier zal gebruik worden gemaakt van de laatste. De invoergegevens worden bijvoorbeeld op de volgende wijze verzameld. De (hypothetische) producten worden op een kaartje beschreven. Deze kaartjes worden vervolgens aan een respondent voorgelegd, zodat deze kan aangeven hoe aantrekkelijk hij of zij dit product vindt.



Figuur 11.5 Verzameling invoerdata

In de onderstaande tabel is een kolom toegevoegd met de scores voor de producten (voor één respondent).

Tabel 11.4 Scores op de profielen

Profiel	Zool	Bovenwerk	Prijs	Beoordeling
1	1	1	1	9
2	1	2	2	7
3	1	3	3	5
4	2	1	2	6
5	2	2	3	5
6	2	3	1	6
7	3	1	3	5
8	3	2	1	7
9	3	3	2	6

worden geschat. Praktisch is dit probleem niet zo groot omdat deze effecten vaak slechts 5 tot

10% van de variantie voor hun rekening nemen.

Door de scores te bekijken kan je een indruk krijgen welke attributen een respondent belangrijk vindt. Als de respondent systematisch een hoge score toekent aan producten met een leren bovenwerk, dan kan je de conclusie trekken dat de respondent dit attribuutniveau sterk waardeert. Een conjunctanalyseprocedure doet hetzelfde maar dan op een wat consistentere manier.

Selecteer een conjunctanalyseprocedure

De te hanteren procedure is afhankelijk van de invoerdata. Bij scores kan gebruik worden gemaakt van regressieanalyse (in geval van rangordedata moet een andere methode worden gebruikt).

In geval van het toekennen van scores aan profielen ziet het onderliggende model er als volgt uit. Het totale nut U van een alternatief is gelijk aan de som van het nut dat wordt toegekend aan de afzonderlijke attribuutniveaus waarover dat product beschikt. α_{ij} is het deelnut oftewel het nut dat toegekend wordt aan een specifiek attribuutniveau. x_{ij} is een dummyvariabele met een waarde één als dat product dat niveau heeft en een waarde nul als dat niet het geval is. k_i geeft het aantal attribuutniveaus van attribuut i en m is het aantal attributen.

$$U = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \alpha_{ij} x_{ij}$$

De range voor attribuut i is het nut van het hoogst gewaarde attribuut minus het nut van het laagst gewaardeerde attribuut.

$$R_i = \{\text{Max}\ (\alpha_{ij}) - \text{Min}\ (\alpha_{ij})\ \}$$

Het belang van attribuut i is vervolgens gelijk aan:

$$W_i = R_i \,/ \sum_{i=1}^{m} R_i$$

Het belang W_i van attribuut i wordt bepaald door de range te delen door de som van de ranges. Zie het voorbeeld voor een uitwerking hiervan.

De deelnutten kunnen door middel van regressieanalyse worden bepaald. De regressiefunctie kan als volgt worden beschreven.

$$U = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + b_6 X_6$$

Hierbij wordt elk attribuut beschreven door twee dummyvariabelen. Het aantal dummyvariabelen is altijd gelijk aan het aantal niveaus minus één. De zool wordt weergegeven door de eerste twee dummies X1 en X2, het bovenwerk door X3 en X4 en de prijs door X5 en X6. Twee dummies is voldoende omdat één van de deelnutten als basisniveau gekozen wordt en de twee dummies vervolgens de veranderingen ten opzichte van dit basisniveau beschrijven. In het onderstaande voorbeeld is niveau drie de basis en dummy X1 beschrijft het verschil tussen niveau één en niveau drie.

niveau één en niveau drie.

Tabel 11.5 Dummy variabelen

	X1	X2
Level 1	1	0
Level 2	0	1
Level 3	0	0

Op basis van deze dummies kunnen de negen profielen uit tabel 11.4 als volgt gecodeerd worden. Het eerste profiel heeft voor alle attributen het eerste attribuutniveau. Volgens tabel 11.5 wordt dit niveau aan de hand van de dummies gecodeerd als één en nul. Deze codering is dan ook in tabel 11.6 terug te vinden.

Tabel 11.6 Codering profielen voor regressie

	Attribuut					
Beoordelings Score	Zool X1	Bovenwerk X2	X3	Prijs X4	X5	X6
9	1	0	1	0	1	0
7	1	0	0	1	0	1
5	1	0	0	0	0	0
6	0	1	1	0	0	1
5	0	1	0	1	0	0
6	0	1	0	0	1	0
5	0	0	1	0	0	0
7	0	0	0	1	1	0
6	0	0	0	0	0	1

Deze negen profielen vormen de input voor een regressieanalyse. Uitrekenen van deze regressiefunctie leidt tot de volgende parameterschattingen:

b1 = 1,000
b2 = -0,333

Omdat een attribuut gecodeerd is met twee dummies, moeten de twee dummyvariabelen vervolgens worden omgerekend tot een deelnut van elk van de drie afzonderlijke attribuutniveaus. Zoals gezegd geeft de dummyvariabele het verschil weer ten opzichte van een basisniveau. Daarom geldt:

a11 - a13 = b1

a12 - a13 = b2

Om deze waarden te kunnen uitrekenen moet een extra voorwaarde worden gedefinieerd (twee waarden en drie onbekenden). De absolute waarde van de deelnutten doet er niet toe. Hierdoor kan de aanvullende eis gesteld worden dat de som van de deelnutten gelijk is aan nul.

a11 + a12 + a13 = 0

Dit leidt tot de volgende deelnutten van de attribuutniveaus.

a11 = 0,778 ; a12 = -0,556 ; a13 = -0,222

Voor de andere attributen wordt een soortgelijke berekening gemaakt. Het nut van de attribuutniveaus is afgebeeld in tabel 11.7.

De som van de ranges is gelijk aan:
[0,778-(-0,556)] + [0,445 – (-0,556)] + [1,111-(-1,222)] = 4,668.

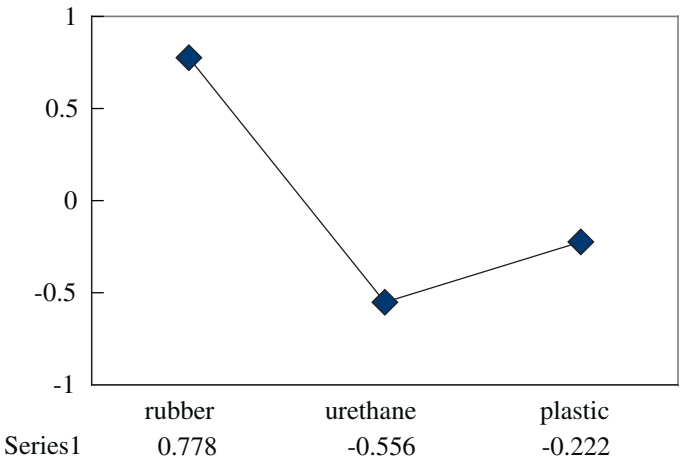
Het belang van het attribuut zool is vervolgens 1,334 / 4,668 = 0,286. Het belang van het bovenwerk en van de prijs is respectievelijk 0,214 en 0,5.

Tabel 11.7 Uitkomsten van conjunctanalyse

Attribuut	Niveau	Deelnut	Belang
Zool	Rubber	0,778	
	Polyurethane	-0,556	0,286
	Plastic	-0,222	
Bovenwerk	Leer	0,445	
	Canvas	0,111	0,214
	Nylon	-0,556	
Prijs	$15,00	1,111	
	$30,00	0,111	0,5
	$45,00	-1,222	

Interpreteer de resultaten

Na het uitrekenen van het nut van attribuutniveaus en het belang van attributen kan de uitkomst worden geïnterpreteerd. In figuur 11.6 is het nut van de attribuutniveaus voor de zool weergegeven. Uit dit plaatje is duidelijk af te leiden dat deze respondent een voorkeur heeft voor een rubberzool. Uit het belang in tabel 11.7 is af te leiden dat deze respondent het attribuut prijs het belangrijkste vindt bij het kiezen van een sportschoen.



Figuur 11.6 Nut van attribuutniveaus voor attribuut zool

Het voorgaande is een analyse van één respondent. Op soortgelijke wijze kunnen de analyses voor alle respondenten worden uitgevoerd. Vervolgens kunnen de uitkomsten voor verschillende respondenten vergeleken worden door bijvoorbeeld te kijken of er groepen te onderscheiden zijn in het belang dat ze toekennen aan attributen.

Bepaal validiteit en betrouwbaarheid
Door middel van een R^2 kan worden gekeken hoe goed het model de scores die de respondent geeft beschrijft. Daarnaast kan gebruik worden gemaakt van zogenaamde hold-out profielen. De respondent beoordeelt deze profielen op soortgelijke wijze, maar deze scores worden niet gebruikt bij het berekenen van het nut dat wordt ontleend aan attribuutniveaus. Vervolgens kan gekeken worden in hoeverre de scores op de hold-out alternatieven zijn te voorspellen uit de berekende scores in tabel 11.7.

11.4 Literatuur

Literatuur voor multidimensional scaling (oplopende moeilijkheidsgraad):
- Zwart, P.S. (1993);
- Malhotra, N.K. (1993);
- Churchill, G.A. (1991);
- Green, P.E., D.S. Tull en G. Albaum (1988);
- Hair, J.F., R.E. Anderson, R.L. Tatham en W.C. Black (1998);
- Green, P.E., F.J. Carmone en S.M. Smith (1989).

Literatuur voor Conjunctanalyse (oplopende moeilijkheidsgraad):

- Zwart, P.S. (1993);
- Malhotra, N.K. (1993);
- Churchill, G.A. (1991);
- Green, P.E., D.S. Tull en G. Albaum (1988);
- Hair, J.F., R.E. Anderson, R.L. Tatham en W.C. Black (1998).

Literatuur

Addelman, S., 'Orthogonal main-effect plans for asymmetrical factorial experiments'. In: *Technometrics* 4 (1962), pp. 21-46.

Amemiya, T., 'Qualitative Response Models: A Survey'. In: *Journal of Economic Literature* 19 (1981), pp. 1.483-1.536.

Amemiya, T., 'Tobit models: A Survey'. In: *Journal of Econometrics* 24 (1984) pp. 3-61.

Brink, W.P. van den & Koele, P., *Statistiek deel 1*. Boom Meppel, Amsterdam, 1985.

Brink, W.P. van den & Koele, P., *Statistiek deel 2*. Boom Meppel, Amsterdam, 1986.

Brink, W.P. van den & Koele, P., *Statistiek deel 3*. Boom Meppel, Amsterdam, 1987.

Cattin, P.C. en D.R. Wittink, 'Commercial Use of Conjoint Analysis: A Survey'. In: *Journal of Marketing* vol. 46 (1982).

Churchill, G.A., *Marketing Research, methodological foundations*. Dryden Press, Fort Worth, 1991.

Cochran, W.G., *Sampling Techniques*. Wiley series, 1977.

Dijk, J.P.M van, K. Lodder en H.C.J. Vrolijk, *De steekproef voor het Bedrijven-Informatienet van het LEI, Bedrijfskeuze 2000 en selectieplan 2001*. LEI-Rapportage 1.02.01. LEI, Den Haag, 2002.

Dol, W., *Small area estimation, A synthesis between sampling theory and econometrics*. Wolters-Noordhoff, 1991.

Green, P.E., D.S. Tull and G. Albaum, *Research for marketing decisions*, Prentice-Hall International Editions, 1988.

Green, P.E., V. Srinivasan, 'Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice'. In: *Journal of Marketing* 54 (1990), pp. 3-19.

Greene, W.H., *Econometric Analysis. 4th edition*, 2000.

Hair, J.F., Anderson, R.E., Tatham, R.L. en Black, W.C., *Multivariate Data Analysis, 5th edition*. Prentice Hall, pp. 730, 1998.

Hamilton, J.D., *Time-series analysis*. Princeton University Press, 1994.
Harnett, D.L., *Statistical methods*. Addison Wesley, 1982.

Huizing, K.R.E., *Inleiding SPSS 11.0 voor Windows en Data Entry*. Academic Service, Schoonhoven, pp. 366, 2002.

Maddala, G.S., *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, 1983.

Maddala, G.S., *Introduction to Econometrics. 3rd edition*. Wiley, 2000.

Malhotra, N.K., *Marketing Research an applied Orientation*. Prentice-Hall International Edition, 1983.

Manly, B.F.J., *Multivariate Statistical Methods. A Primer. 2nd edition*. Chapman & Hall, pp. 215, 1994.

Polman, N., J. Luijt, M. Mulder en G. Thijssen, *Going concern waarde en marktprijs van landbouw-bedrijven*. Wageningen Universiteit en LEI, Wageningen en Den Haag, 1999.

Pindyck, R.S. en Rubinfeld, D.L., *Econometric models and economic forecasts. 4th edition*. McGraw-Hill, 1998.

Reinhard, S., L. van Staalduinen en M. Spijkerman, *Handleiding voor de mogelijkheden en het gebruik van paneldata op het LEI, Het Informatienet en de Landbouwtelling*. LEI-notitie 01.03. LEI, Den Haag, 2001.

Sarndal, C.E., B. Swensson en J. Wretman (1992) Model Assisted Survey Sampling, Springer verslag, New York.

Sharma, S., *Applied Multivariate Techniques*. Wiley & Sons, pp. 493, 1996.

Slotboom, A., *Statistiek in woorden*. Wolters-Noordhoff, Groningen, 1996.

SPSS, *SPSS Base 11.0 for Windows User's Guide by Spss Inc*. Inc SPSS, 2001.

Thomas H. en Ronald J. Wonnacott, *Introductory Statistics for Business and Economics, Fourth Edition.* Wiley, New York, 1990.

Verbeek, M., *A Guide to Modern Econometrics*. Wiley, 2000.

Vrolijk, H.C.J., W. Dol en G. Cotteleer, *Schatten van kenmerken van kleine deelgebieden*. LEI-Rapportage 8.02.05. LEI, Den Haag, 2002.

Vrolijk, H.C.J. en K. Lodder, *Voorstel tot vernieuwing van het steekproefplan voor het Bedrijven-Informatienet*. LEI-Rapportage 1.02.02. LEI, Den Haag, 2002.

Zwart, P.S., *Methoden van marktonderzoek*. Educatieve Partners Nederland BV, 1993.