NSD-stat+, Statistische Dataverwerking uit Noorwegen

Marien Lina *)

Samenvatting:

NSD-stat + is een statistisch pakket dat in het Noorse onderwijs wordt gebruikt. De stof is vergelijkbaar met dat voor inleidingen statistiek op universitair niveau in Nederland. Een Engelstalige handleiding beschrijft zowel het praktisch gebruik van het pakket als de theoretische achtergronden van de ondersteunde methoden. Het pakket is ook buiten het onderwijs bruikbaar voor het invoeren en verwerken van databestanden. Het programma is redelijk snel in het verwerken van grotere bestanden. Voor de onderzoekspraktijk lijkt het aanbod van geavanceerdere methoden en technieken een beetje mager. Het basispakket bevat eenvoudig te bedienen modulen voor het ontwerpen en invoeren van datasets. De uitwisseling van data met andere statistische paketten is mogelijk via *ASCII, SPSS, DIF* en *DBASE3* bestanden. De analyse module ondersteunt de meest gangbare statistische standaardprocedures, van het maken van tabellen tot multipele regressie. Het pakket heeft diverse grafische mogelijkheden, onder anderen histogrammen, spreidingsdiagrammen en box-plots. Ook kan men thematische kaarten maken. Het is wel jammer dat de gebruiker hiervoor geen coördinaatbestanden kan creëren. Gezien de kwaliteitprijs verhouding lijkt het pakket goed bruikbaar als men aan de ondersteunde procedures genoeg heeft. Een versie voor *Windows* is in voorbereiding.

Technische gegevens en installatie:

Programma:	NSD-Stat +, versie 93
Ontwikkeld door:	Norwegian Social Science Data Services (NSD)
Besturings-systeem:	MS-DOS 2.10 of hoger Installatie in een netwerkomgeving is mogelijk
Minimale configuratie:	XT of AT met 350 Kbytes vrij geheugen, minimaal één diskette drive (opstarten van diskette is mogelijk)
Aanbevolen extra's:	Harde schijf met 3 Mbyte vrij voor het programma, extra schijf ruimte voor data afhankelijk van hoeveelheid gegevens. Een van de volgende grafische kaarten: cga, mcga, ega, ega64, ega-mono, hercules mono, att400 (Olivetti M24), vga, pc3270, ibm8514.
Prijs voor 1 gebruiker:	Fl. 1000,-. Voor educatieve instellingen Fl. 500,-
Handleiding:	Een uitgebreide handleiding in klapper van ca. 200 pagina's met praktische gebruikerswenken en statistische theorie.
Distributeur:	iec *Pro*GAMMA Grote Rozenstraat 15 9712 TG Groningen tel. 050-636900

---

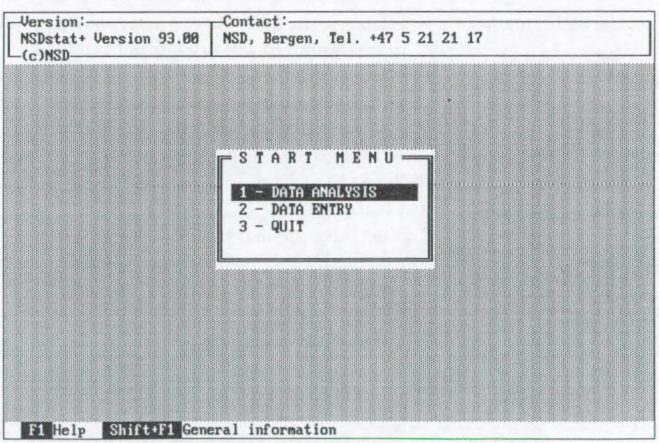
*) Centraal Bureau voor de Statistiek. Divisie Research en Ontwikkeling, Sector Statistische Informatica
tel. 070-3375139

Inleiding

Sinds 1986 is het statistisch pakket *NSD*-stat in ontwikkeling bij het Noorse instituut: *Social Science Data Services*. Het pakket is speciaal ontwikkeld voor het gebruik in het Noorse schoolsysteem. Het doel daarbij was een praktische kennismaking van de student met de rol van de sociale wetenschapper. Het programma wordt reeds enkele jaren gebruikt op Noorse scholen en universiteiten. Verschillende datasets met sociale en politieke gegevens uit Noorwegen worden meegeleverd. Een hoge verwerkingssnelheid maakt het gebruik van grote datasets zowel in het onderzoek als in de onderwijspraktijk goed mogelijk. De eenvoudige versie *NSD*-stat ondersteunt procedures van het maken van tabellen tot regressie. In de uitgebreidere versie *NSD*-stat + zijn enkele analysemethoden toegevoegd waaronder multi-variate regressie en variantie-analyse. Hierover later meer. Het pakket is ontwikkeld vanuit het statistiekonderwijs voor sociale wetenschapen en lijkt in de Nederlandse situatie goed bruikbaar voor colleges inleidende statistiek op universiteiten. De zwaardere versie is hier getest. Deze versie bevat overigens alle procedures van de eenvoudige uitvoering.

Onderwerpen in deze beschrijving

Na een *eerste indruk* komen in deze bespreking verschillende onderwerpen aan bod. In *de handleiding* wordt stapsgewijs kennis gemaakt met de procedures voor het *samenstellen van een dataset*, het importeren en exporteren van data, de analyse-modulen voor *univariate analyse* en *bivariate analyse* zoals die in de eenvoudige versie voorkomen. Daarna komt de uitgebreide versie aan de orde. Dit betreft enkele *uni- en bivariate uitbreidingen* en de *multivariate analyse* module. In de uitgebreide versie zijn tevens extra *data invoer faciliteiten* opgenomen. Het programma heeft verschillende mogelijkheden om met variabelen te *manipuleren*. Het *cartografisch* gedeelte van het pakket verdient een aparte vermelding. Na deze onderwerpen wordt deze beschrijving afgesloten met een indruk van de user-interface en de algemene indruk die het pakket nagelaten heeft.



Figuur 1
The NSD-stat start scherm

De eerste indruk

De bouwers van het systeem beogen dat men er gemakkelijk en flexibel zijn doel mee kan bereiken. Er zijn geen praktische grenzen gesteld aan de omvang van de data, en er wordt melding gemaakt dat het pakket met een minimum aan tijdverlies grote bestanden aankan. Het programma is menugestuurd. Het hoofdmenu heeft de opties *data entry module*, *data analyse module* en *quit*. De bediening die dan wordt uitgelegd spreekt voor zich voor een ieder die bekend is met programmatuur met een vergelijkbare user-interface, zoals bijvoorbeeld de Borland Pascal omgeving of Clipper applicaties.

De handleiding

Het pakket *NSD*-stat+ is beschikbaar in een goed leesbaar Engels. De handleiding maakt een frisse indruk en is overzichtelijk ingedeeld in een losbladige klapper van circa 200 pagina's. Per onderdeel van het programma wordt de theorie achter de techniek vanaf de basis toegelicht. Bijvoorbeeld, bij het beschrijven van regressie wordt het principe van de kleinste kwadratensom uitvoerig uitgelegd, met verhelderende illustraties en voorbeelden. In de volgende paragrafen zal de inhoud van de handleiding nader aan de orde komen.

Het samenstellen van een data set.

Hoofdstuk 1 van de handleiding geeft het begrippenkader voor het samenstellen van een dataset, waarbij gebruikte termen als *cases*, *variables* en *values* worden uitgelegd. Na een voorbeeld van een datamatrix worden de gebruikte variabele typen *nominal*, *ordinal* en *continuous* uitgelegd. Er is een voorziening om per invoerveld een code voor ontbrekende data op te nemen, bijvoorbeeld de waarde 999 voor een variabele waarin de leeftijd in jaren wordt vastgelegd. *Variable descriptions* en *value descriptions* van de dataset kunnen in *NSD*-stat *on-line* worden ingevoerd.

Data-invoer faciliteiten in NSD

Binnen *NSD* kan data invoer in een eenvoudige spreadsheet plaatsvinden. Er wordt naderhand een systeembestand gemaakt. Er kan een kodeboek worden vastgelegd, waarop de datainvoer wordt gecontroleerd. Bij het definiëren van variabelen worden de variabele-naam en het type (*continuous* of *discrete*) vastgelegd. Voor continue variabelen worden tevens de minimum waarde, de maximum waarde en het aantal decimalen gedefiniëerd. De waarde voor ontbrekende scores wordt vastgelegd als het eerst volgende cijfer dat geheel uit negens bestaat boven de opgegeven maximum waarde. Voor discrete variabelen wordt voor elke waarde een numerieke code en een omschrijving gevraagd. Na het vastleggen van het kodeboek kunnen de data worden geïmporteerd of ingevoerd. Eventueel kan de variabele definitie worden aangepast via manipulaties.

Importeren en exporteren van data.

In de data-invoer module kunnen *files* worden geconverteerd naar andere formaten. Deelsets kunnen worden geëxporteerd naar een standaard *ASCII -file*. De uitgebreide versie maakt export van *DBASE-III*, *DIF* en *SPSS-system files* mogelijk, en aggregatie per klasse van de te benoemen aggregatievariabele. Als aggregatiewaarde kan de som, de frequentie, de minimum of maximum waarde, het gemiddelde of de standaard deviatie worden gebruikt. Data *files* van het zelfde type (*ASCII* , *SPSS-portable*, *DBASE-III* en *DIF files*) kunnen tevens worden geïmporteerd en worden dan geconverteerd naar *NSD system files*. Het importeren van *ASCII-files* vereist enig

datadefinitie werk. Hoe dat in zijn werk gaat staat beschreven in de handleiding. Meerdere matrixen kunnen worden ingevoerd en samengevoegd.

De data analyse module

Via het hoofdmenu komen we in de *data analyse module*. Hier kunnen we een bestaande dataset opvragen door een bestandsnaam en een zoekpad te specificeren in een invoervenster. In de analyse module kunnen we kiezen uit *univariate analyse* en *bivariate analysetechnieken*. Daarnaast zijn er opties voor het bekijken en afdrukken van een datamatrix, het selecteren van een sub-set, het creëren van nieuwe variabelen en een cartografische presentatie. In de analysemodule is telkens informatie over de variabelen opvraagbaar met functietoetsen. De variabelen worden dan geselekteerd door de naam in de lijst op te zoeken en op <Enter> te drukken.

Een handige optie voor het berekenen van foutmarges bij steekproeven.

Een aparte optie van de analysemodule zit verborgen onder de <F6> toets. Hierin kan een zogeheten *Error Margin Table* worden opgeroepen. Deze procedure kan worden gebruikt om de foutmarges te berekenen voor binomiale verdelingen in steekproeven. In een tabel kun je per regel een steekproefomvang invullen, een gewenst betrouwbaarheidsinterval (80, 90, 95 of 98%) kiezen en een verwachte proportionele verdeling aangeven (p=0,02, p=0,05, p=0,10, p=0,15, p=0,20, p=0,40 en p=0,50). *NSD* geeft dan het gebied aan waarbinnen afwijkingen tussen de gevonden en verwachte proporties niet significant zijn. Deze procedure kan tevens hulp bieden bij het bepalen een steekproefomvang als men een bepaalde nauwkeurigheid van het meetresultaat van een proportie wenst. Men kan dan (in verschillende rijen van de tabel) meerdere steekproefomvangen invullen en kijken bij welke steekproefomvang de gewenste precisie is bereikt.

Univariate analyse

De univariate procedures betreffen de sub-procedures *frequency* en *descriptive*. Een frequentietabel wordt overzichtelijk op het scherm weergegeven.

Figuur 2
Een frekwentietabel in NSD-stat

Return Scroll Measure Graphics

v4 EDUCATION: Age when full-time education completed

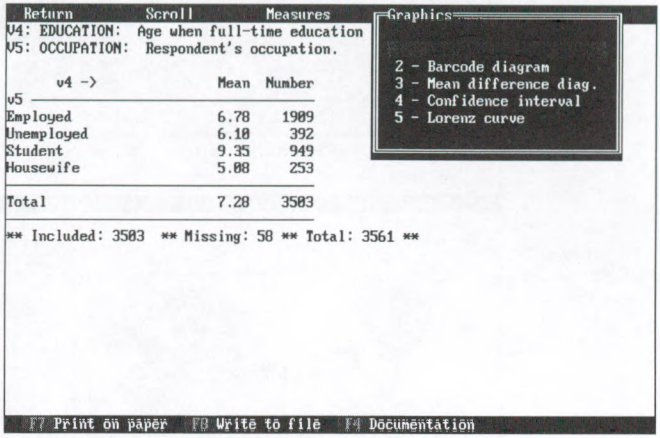
CATEGORY NAME	CODE	NUMBER	% OF ALL	% OF VALID
12 years or less	1	41	1.15	1.16
13 years	2	50	1.40	1.42
14 years	3	166	4.66	4.70
15 years	4	200	5.62	5.66
16 years	5	551	15.47	15.60
17 years	6	324	9.10	9.17
18 years	7	516	14.49	14.61
19 years	8	321	9.01	9.09
20 years	9	230	6.46	6.51
21 years or more	10	1134	31.84	32.10
Not answered	99	28	0.79	-
SUM =		3561	100.00	100.00

** Included: 3533 ** Missing: 28 ** Total: 3561 **

F7 Print on paper F8 Write to file F4 Documentation

Voor continue variabelen kan het aantal gewenste klassen worden opgegeven. De grenzen van de klassen worden dan door *NSD*-stat bepaald. De tabellen geven absolute aantallen, ongecorrigeerde percentages en percentages die gecorrigeerd zijn voor ontbrekende waarden. Je kunt door de tabel bladeren als die te groot is om in eens op het scherm te plaatsen. De tabel kan worden gesorteerd op de celfrequenties. Verder kan er een staafdiagram en een taartdiagram worden weergegeven. Ook enkele centrummaten (mediaan, modus en gemiddelde) kunnen via *Frequency* worden opgevraagd. In *Descriptive* kan een frequentiepolygoon worden opge-

Figuur 3
Descriptives



vraagd. Door deze polygoon kan de normale verdeling worden afgebeeld, en er kan een box-whisker-plot worden bijgeplaatst. Verder kan een Lorenz curve worden opgevraagd voor discrete variabelen. De scheefheid wordt aangegeven met de *GINI*-index. In de standaartabel van *Descriptive* worden nog enkele kengetallen weergegeven: het rekenkundig gemiddelde, de som, het minimum, het maximum, de standaarddeviatie en het randtotaal. De mediaan wordt niet berekend voor continue variabelen.

Bivariate analyse

Deze procedure bevat de onderdelen *Crosstable*, *Descriptive* en *Scattergram*. Via de optie *Crosstable* kan voor twee diskrete variabelen kan een kruistabel worden opgevraagd met absolute aantallen, rij- en kolom percentages en totale percentages (zie figuur 4).

Continue variabelen kunnen ook hier automatisch worden ingedeeld in een nader op te geven aantal klassen. Staaf- en taartdiagrammen kunnen voor alle categorieen van één van de kolomvariabele worden opgevraagd (zie figuur 5). De staafdiagrammen kunnen worden gestapeld. Via *Descriptives* wordt voor een continue variabele, per klasse van een diskrete variabele, dezelfde resultaten verkregen als met frequencies voor de totale steekproef. De grafische uitvoer kent hier nog een paar extra's, bijvoorbeeld een staafdiagram toont de afwijkingen van het totaalgemiddelde per klasse van de diskrete variabele.

Figuur 4
Een kruistabel in NSD-stat

Return Scroll Measure Table

1 - Column percentage
3 - Percent of total
4 - Raw numbers
5 - Expected cell freqency
6 - Deviation from expected
7 - Standardized deviation
8 - Column header over
9 - Column header under
A - No text
B - Turn the table

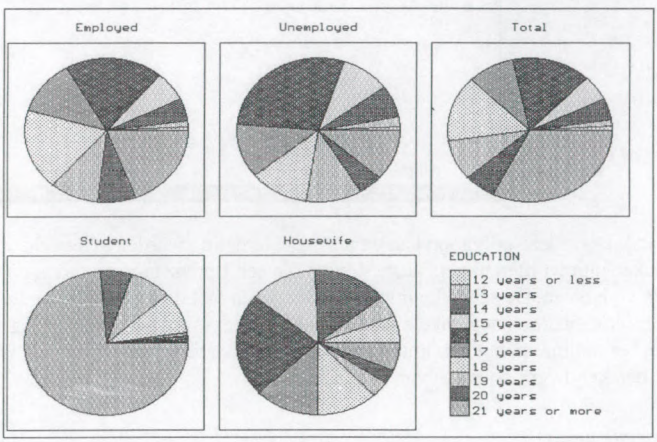
→ v5 OCCUPATION Respondent's occupation.
↓ v4 EDUCATION Age when full-time education com

	Employ	Unempl	Studen	Housew		
12 years or less	48.8	9.8	4.9	36.6		
13 years	54.0	10.0	2.0	26.0		
14 years	57.8	16.9	3.0	22.3		
15 years	62.6	19.2	1.0	17.2		
16 years	67.2	20.6	1.5	10.8		
17 years	72.9	15.0	2.2	10.0		
18 years	68.0	9.2	16.8	6.1		
19 years	58.3	13.5	25.4	2.8	100.0	319
20 years	63.3	8.3	24.9	3.5	100.0	229
21 years or more	32.4	3.9	62.4	1.3	100.0	1122
SUM	54.5	11.2	27.1	7.2	100.0	3503

** Included: 3503 ** Missing: 58 ** Total: 3561 **

F7 Print on paper F8 Write to file F4 Documentation

Figuur 5
Taartdiagram onder bivariate descriptions



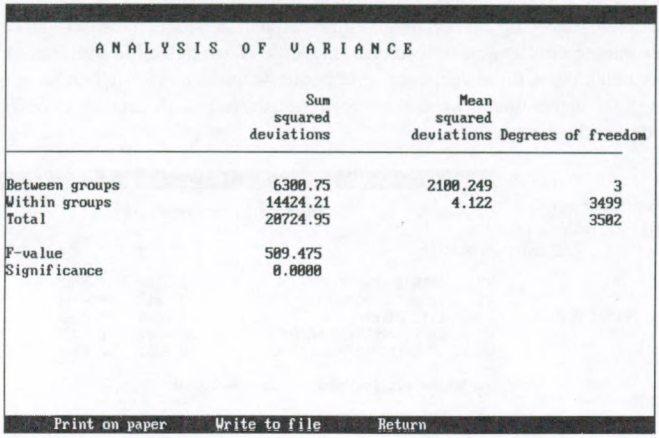
Scattergram levert een spreidingsdiagram voor twee continue variabelen. Een regressielijn kan er in worden afgebeeld. Er kan op worden ingezoomd in het diagram. De beschrijving van de mogelijkheden wordt ondersteund met een uitleg van de beginselen van lineaire regressie. In de uitgebreide versie, waarin regressie een apart onderdeel is, wordt hier meer werk van gemaakt. Behalve een regressie kan voor de twee geselekteerde variabelen een variantie-analyse worden uitgevoerd (zie figuur 6).

Uni-,en bivariate toevoegingen in NSD-stat+

Het reeds behandelde valt onder de standaard versie van *NSD*-stat. De multivariate analysemodule behoort bijde zwaardere versie *NSD*-stat+. Onder de univariate procedure *descriptives* zijn in *NSD*-stat+ de scheefheid , de kurtosis en een betrouwbaarheidsinterval toegevoegd. Onder de bivariate procedure *descriptives* zijn dezelfde maten beschikbaar voor een continue variabele,

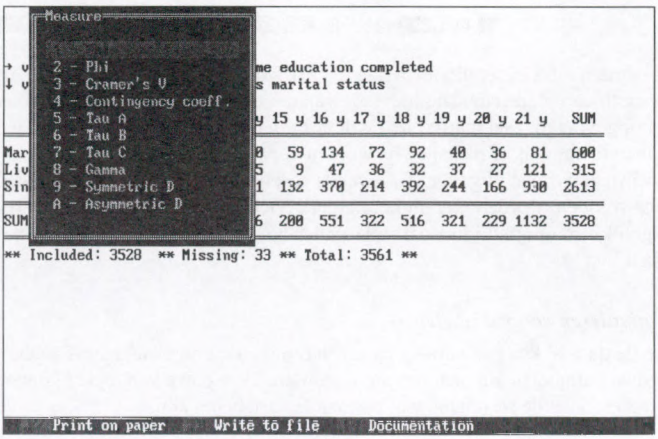
per klasse van de diskrete variabele. In kruistabellen kan men de verwachte waarden, de afwijking van verwachte waarden en de standaardafwijking per cel opvragen.

Figuur 6
De uitvoer van een variantie-analyse in NSD-stat



Verder zijn de volgende associatiematen toegevoegd in de uitgebreide versie: Chi^2, Phi, Cramer's V, de Contingency Coefficient, Kendall's Tau A, B en C, Yules Gamma en de symmetrische en de a-symetrische versie van Somer's D.

Figuur 7
Selectie van toetsingsgrootheden bij kruistabellen



Men kan een grafiek opvragen van het betrouwbaarheidsinterval van de gevonden waarden per klasse. Verder wordt een t-test ondersteund, en als aanvullling daarop Scheffe's test, waarmee gecorrigeerd wordt voor het aantal waarnemingen in sub-groepen. Verder wordt een correlatie-matrix ondersteund, waarin de coëfficiënten paarsgewijs worden berekend.

Multivariate analyse

NSD-stat+ ondersteunt multipele regressie. Men kan voor één afhankelijke variabele een regressie uitvoeren voor een praktisch oneindig aantal onafhankelijke variabelen. In de praktijk bleek dat een regressie op acht onafhankelijke variabelen voor een dataset van ca. 3500 cases binnen vier minuten het gewenste resultaat opleverde. Toen hetzelfde geprobeerd werd voor twintig variabelen bleek dit te veel voor de beperkte vermogens van de hardware (geheugen beperking). Bij de resultaten worden per toegelaten variabele de regressie-coëfficienten aangegeven en de verklaarde variantie.

Figuur 8
De uitvoer van een regressie-analyse in NSD-stat

```
 Return    Model    Scroll    Save    Graphics   Table

Dependent:  V6      EXPECTATION OF LIFE:  1983

Predictor                                  B      Beta

V7    INFANT DEATHS                     -0.1100  -0.5280
V9    INTAKE OF CALORIES                 0.0018   0.0934
V10   LITTERATES %                       0.0860   0.2246
V5    GROSS DOMESTIC PRODUCT          3.071e-05   0.1289
V8    INHABITANTS/PHYSICIANS            -0.0001  -0.1098

Variables not included in the equation

V3    RESIDENT POPULATION            1.046e-06   0.0124

Constant                    57.58

Multiple R     0.95463              F-value     187.02
Multiple R²    0.91131              F-sign.       0.0000
Adjusted R²    0.90644

Valid cases          97.00

 F1 Help  Shift+F1 General information
```

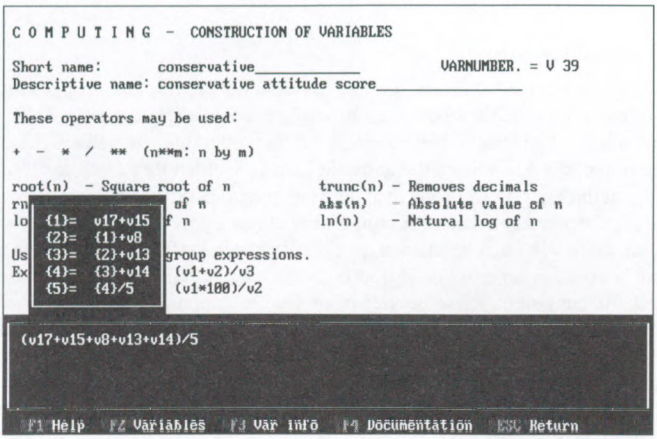
B-waarden die kleiner uitvallen dan 0.0001 worden in wetenschappelijke notatie weergegeven. De significantie, de standaarddeviatie van de regressiecoëfficiënt en de tolerantie kunnen worden opgevraagd. Een matrixplot en de residuen kunnen grafisch worden weergegeven. Cases kunnen lijstgewijs of paarsgewijs worden verwijderd. De regressie kan stapsgewijs (*forward* of *backward*) worden uitgevoerd. De reeds geselecteerde variabelen die in een volgende stap niet meer voldoen, worden *niet* uit het model verwijderd. Naast de multipele regressie bestaat de mogelijkheid om multidimensionele kruistabellen te produceren met enkele kengetallen en grafieken.

Manipuleren van variabelen.

Van de dataset kan een sub-set geselecteerd worden met statements zoals: v1<30 and v2=2-3. Nieuwe variabelen kunnen gecreëerd worden. De nieuwe variabelen kunnen via *Recode* discrete, gehercodeerde resultaten van bestaande variabelen zijn.

De variabelen kunnen ook via *Compute* berekend worden uit de waarde van reeds bestaande variabelen. Nadien kan men overtollige variabelen uit de set verwijderen. Uiteraard kan de gewijzigde dataset worden opgeslagen (zie figuur 9).

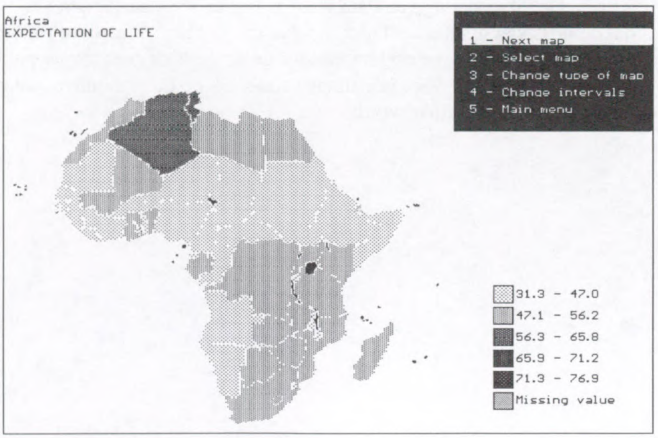
Figuur 9
Het berekenen van nieuwe variabelen



Thematische kaarten.

Wanneer een gedigitaliseerde kaart aanwezig is en met de data-set is verbonden, dan kan een chloropletenkaart (zie figuur 10) aangemaakt worden, of symbolenkaart (een staafdiagram waarin elke staaf in het centrum van het betreffende gebied wordt geplaatst).

Figuur 10
Cartografische weergave van statistische data



De module is eenvoudig te bedienen, en de kaartjes verschijnen met een legenda. Binnen de onderscheiden regio's kan een staafdiagram of een cirkeldiagram worden weergegeven (de omvang variëert mee met de somfrequenties). De kaart kan geschaald worden en deelgebieden kunnen apart worden weergegeven. Sommige kaarten verschijnen nogal traag, niet ongebruikelijk, want het heeft te maken met de omvang van het coördinatenbestand. De gebiedsindeling kan helaas niet door de gebruiker worden ontworpen. Het is jammer dat er slechts een beperkt aantal kaarten beschikbaar is, waaronder de landenkaart van West-Europa en die van Afrika. Dit gedeelte van het pakket is dus eigenlijk alleen voor onderwijsdoeleinden geschikt.

Het gebruiksgemak

Zoals gezegd is *NSD*-stat menu-gestuurd. Nadat het programma is opgestart verschijnt het hoofdmenu vanwaaruit telkens sub-procedures en vensters kunnen worden geactiveerd. De interface is tekst georiëerd. Na enige tijd met het programma te hebben gestoeid, blijkt de bediening eenvoudig. De keuzen die gemaakt kunnen worden liggen telkens intuïtief voor de hand en dat ontbreekt nog wel eens in andere pakketten. Bijvoorbeeld, als er een kruistabelletje op je scherm is verschenen dan verschijnt er tevens een menu waaruit je de toetsgrootheden kunt uitkiezen die je wilt laten berekenen, je kunt direct een grafiekje laten maken, of als de tabel te groot is voor het scherm kun je doorbladeren naar het volgende deel van de tabel. De interface is redelijk consistent, maar niet helemaal. De ontsnappingstoets <Esc> doet in sommige gevallen niet wat er in de handleiding wordt beloofd. In sommige gevallen kan een gedefiniëerde variabele niet van een continue in een diskrete veranderd worden, zonder hem te verwijderen en er een nieuwe variabele voor te creëren. In een enkel extreem geval werd mijn minder elegante intentie om het programma vast te laten lopen gehonoreerd met een run-time error. Over het algemeen gedroeg het geheel zich overigens redelijk stabiel.

Conclusie

NSD-stat + is een bruikbaar pakket voor het uitvoeren van survey research en voor educatieve doeleinden, vooral als men geen andere multi-variate technieken gebruikt dan regressie-analyse. De verwerkingssnelheid van grote bestanden is redelijk hoog. Het is een basis-pakket zonder poespas. De uitwisseling van data via o.a. *SPSS* en *ASCII* bestanden verloopt probleemloos. Doordat de prijs van het pakket lager ligt dan dat van de "zwaardere pakketten" kan het programma interessant zijn voor onderzoeksinstanties die zich geen zwaar pakket kunnen veroorloven of nodig hebben, of voor inleidingen statistiek op universiteiten, waar voor het in Noorwegen gemaakt is en gebruikt wordt.