

Klargøring af data til aflevering til DDA

—

Instruks

Denne instruks gennemgår datadokumentationsformater m.v. for datasystemfiler, som skal afleveres til Dansk Data Arkiv efter indgået aftale om *ekstern oparbejdning*.

Formateringen er nødvendig for at sikre langtidsbevaring af den fulde dokumentation. Hensynet til langtidsbevaringen af fuldt dokumenterede data indebærer bl.a., at der er ganske detaljerede krav til angivelse af manglende data.

Formateringskravene omfatter alle former for kvantitative datasystemfiler. Eksemplerne tager for enkeltheds skyld udgangspunkt i spørgeskemadata.

Filformater

Datafilen afleveres i et af følgende formater:

- SPSS (*.sav*)
- STATA (*.dta*)
- SAS (*.sas7bdat* med tilknyttet formatbibliotek: *formats.sas7bcat*)

Filen skal kunne læses fejlfrit i eller konverteres til SPSS version 19.0 for Windows.

Filen kan eventuelt produceres i et konverteringsprogram.

Den institution eller forsker, der afleverer data (donor), har ansvaret for, at filen har det korrekte format.

Forklaringerne nedenfor baserer sig på SPSS-syntaks. Syntaksen svarer i udgangspunktet til syntaksen i andre statistikprogrammer.

Variabeltyper

Variable skal være af typerne *numeric* eller *string*.

Eventuelle variable i andre formater skal rekodes.



Eksempel

Variablen 'Dato' ligger i formatet date.

Variablen rekodes til tre numeriske variable: 'Dag', 'Måned' og 'År', eller den omkodes til string. Kodeværdier for 'År' angives med fire cifre.

Variable i numerisk format skal være i heltal eller med maksimalt ni decimaler.

Variable, som har kodeværdilabels, skal være heltal.

Bemærk: Parameteren Decimals i vinduet Variable View i SPSS regulerer kun fremvisningen af variabelen; den faktiske præcision af variabelen forbliver skjult.

Kodning af uoplyst

I de tilfælde, hvor respondenterne har undladt at besvare et spørgsmål, kodes værdien som uoplyst. Det gøres ved at indsætte det mindste antal 9-taller, der overstiger den største i øvrigt anvendte kodeværdi.

Eksempel

Variablen har kodeværdierne 1, 2 og 3	→	Uoplyst kodes som 9
Variablen har 273 som højeste kodeværdi	→	Uoplyst kodes som 999

Filtercheck

Der skal foretages filtercheck, hvis datamaterialet indeholder filtreringer.

En filtrering optræder, når en variabel eller en gruppe af variable ikke skal have værdier for visse respondenter.

Et spørgsmål, der anvendes som filter, således at ikke alle skal besvare et eller flere efterfølgende spørgsmål, betegnes efterfølgende en *filtrerende* variabel. Variablen for et spørgsmål, som efterfølgende ikke skal besvares af alle, kaldes en *filtreret* variabel.



Filtreringerne checkes således:

Irrelevant – Omkod kodeværdien for den filtrerede variabel til irrelevant for de respondenter, hvor filterbetingelsen er opfyldt, og hvor den filtrerede variabel er kodet som uoplyst (dvs. respondenterne har rettelig ikke besvaret spørgsmålet).

Kodeværdien for irrelevant er kodeværdien for uoplyst + 1. Dvs. kodeværdien for irrelevant bliver 100, hvis kodeværdien for uoplyst er 99.

Eksempel (eksemplet side 3 forts.):

Rekodning af besvarelse til filtreret spørgsmål pga. (korrekt) manglende værdi:

spm. 6 'Hvad er dit køn?': 1 ('Mand')

spm. 7 'Er din menstruation regelmæssig?': blank → er kodet som 9 ('Uoplyst') → kodes som 10 ('Irrelevant')

Fejl type 1 – Konstater antallet af fejl af type 1 i filtrerede spørgsmål (dvs. respondenterne har besvaret spørgsmålet, men det skulle ikke have været besvaret pga. filtret).

Eksempel (eksemplet side 3 forts.):

spm. 6 'Hvad er dit køn?': 1 ('Mand')

spm. 7 'Er din menstruation regelmæssig?': 2 ('Nej')

Undersøg om der er fejl ved data eller ved dokumentationen, hvis fejlen optræder hos mange respondenter. Korrigér så vidt muligt fejlen og dokumenter, at den er korrigeret.

Der kan eksempelvis optræde mange fejl af type 1, hvis den anvendte interviewerinstruktion adskiller sig fra den, der anvendes under klargøringen af data, eller hvis der forekommer systematiske indtastningsfejl.

Fejl type 2 – Konstater antallet af fejl af type 2 (dvs. respondenterne har besvaret det filtrerede spørgsmål, men ikke besvaret det filtrerende spørgsmål).

Eksempel (eksemplet side 3 forts.):

spm. 6 'Hvad er dit køn?': blank

spm. 7 'Er din menstruation regelmæssig?': 2 ('Nej')



Undersøg om der er fejl ved data eller ved dokumentationen, hvis fejlen optræder hos mange respondenter. Korrigér så vidt muligt fejlen og dokumenter, at den er korrigeret.

Som ved den anden fejltype kan der f.eks. optræde mange fejl af type 2, hvis den anvendte interviewerinstruktion adskiller sig fra den, der anvendes under klargøringen af data, eller hvis der forekommer systematiske indtastningsfejl.

Kodning af designbetingede manglende data

I dataindsamlinger med komplekse undersøgelsesdesigns er det ofte givet på forhånd, at ikke alle respondenter skal deltage i alle dataindsamlingsrunder.

Hvis en respondent på forhånd har været udelukket fra en del af dataindsamlingen, kodes værdien som 'deltog ikke'. Kodeværdien for 'deltog ikke' er 'uoplyst' + 2. Dvs. kodeværdien bliver 101, hvis kodeværdien for uoplyst er 99.

Eksempel

Spørgeskema 1

-spørgeskema til
patienter i medicinsk
behandling

Spørgeskema 2

- spørgeskema til
patienter i
kontrolgruppen

Ved efterfølgende fletning af data for de to grupper af respondenter kodes variablene på basis af spørgeskema 1 som 'deltog ikke' for alle, der ikke var udtaget til denne del af undersøgelsen (de var i kontrolgruppen og modtog spørgeskema 2). Tilsvarende kodes variablene på basis af spørgeskema 2 som 'deltog ikke' for de respondenter, der ikke udtages til denne del (de var i medicinsk behandling og modtog spørgeskema 1).

Bemærk: Kodekategorien 'deltog ikke' ligner i princippet kodekategorien 'irrelevant'. Forskellen er, at 'deltog ikke' anvendes, når besvarelserne mangler som følge af undersøgelsens design, mens 'irrelevant' anvendes, når besvarelsen mangler pga. et filter i det aktuelle spørgeskema.

Check af kodeværdier

Kør frekvenstabeller for alle numeriske variable. Check at dokumentationen til datafilen indeholder information om alle faktisk forekommende kodeværdier.

Tilkobling af datadokumentationen

Den tilhørende datadokumentation skal kobles til systemfilen (og for SAS's vedkommende: formatbiblioteket). Følgende skal angives:

1) Variable Name

Der er intet særligt krav til variabelens navn. Indholdet vil blive bevaret af DDA, men vil ikke blive anvendt til noget specifikt formål.

2) Variable Type

Check at alle variable i filen er enten *numeric* eller *string*.

3) Variable Width

Angiv herigennem det største antal karakterer, som kodeværdierne for variabelen kan antage.

4) Decimals

Angiv herigennem det største antal decimaler, som anvendes i kodeværdierne for den pågældende variabel.

Det er nødvendigt at gennemgå dette for alle variable, da statistikprogrammerne sjældent anvender korrekt decimalangivelse. Brug frekvenstabeller til at konstatere det faktiske antal anvendte decimaler.

Der kan maksimalt angives ni decimaler.

5) Variable Label

Sæt label på variabelen. Label'en skal indeholde en henvisning til spørgsmålsnummeret i spørgeskemaet efterfulgt af den samlede relevante spørgsmålstekst fra spørgeskemaet.

Alle variable skal have en label.

Eksempel

"Spm. 3: Hvor ofte tager du smertestillende medicin (hovedpinepiller o.lign.), som ikke er udstedt på recept?"

6) Value Labels

Sæt kodeværdilabels på numeriske, heltallige kodeværdier, herunder også på kodeværdierne for uoplyst ("Uoplyst"), irrelevant ("Irrelevant") og deltog ikke ("Deltog ikke"), men kun hvis der faktisk forefindes uoplyste, irrelevante eller deltager ikke. Hvis kodeværdierne repræsenterer angivne svarmuligheder i spørgeskemaet, skal hver kodeværdilabel gengive den komplette og ordrette tekst fra spørgeskemaet til den pågældende kategori.

Alle heltallige numeriske variable, med op til 25 betydende kodeværdier, skal have value labels.

7) Missing

Der kan forekomme tre kategorier af ikke-valide kodeværdier (missing): uoplyst, irrelevant og deltog ikke.

For datasæt, der afleveres i SPSS-format: alle ikke-valide kodeværdier sættes som *discrete values*.

For datasæt, der afleveres i STATA-format: undlad at bruge missing-funktionen i STATA. Den er inkompatibel med det endelige arkiveringsformat. Kod i stedet alle ikke-valide kodeværdier som betydende kodeværdier

Personhenførbare data

Hvis materialet indeholder personhenførbare data, skal disse afleveres i en separat nøglefil.



STATENS ARKIVER

DANSK DATA ARKIV

Anden relevant tekst

- Medsend undersøgelsens spørgeskema (registreringsskema) i pdf-format. Hvis spørgeskemaet oprindeligt er udarbejdet i et tekstbehandlingsprogram (f.eks. Word), kan det alternativt medsendes i tekstbehandlingsformatet (.doc).
- Udfyld og medsend det metodiske spørgeskema på DDA Sundheds hjemmeside (sundhed.dda.dk). Skemaet findes under Bevaring af data / Ekstern oparbejdning.
- Eventuelle kommentarer til særlige forhold for enkeltvariable, f.eks. uforholdsmæssigt mange fejl af type 1 eller 2, medsendes som yderligere dokumentation i en tekstbehandlingsfil (.doc) eller som tekstfil (.txt).

Vers. 5 (070312)