

Vejledning i opbygning og check af dokumentation for kvantitative datamaterialer

(vers. 2)

For at kunne arbejde effektivt med analyse af data fra spørgeskemaer er det væsentligt, at dokumentationen af data er opbygget systematisk og omhyggeligt.

Dette er Dansk Data Arkivs forslag til opbygning og check af dokumentation. Eksemplerne tager udgangspunkt i interviewbaserede data, men anbefalingerne er universelle for kvantitative data.

Der skelnes i det følgende mellem tre dokumentationsniveauer: studieniveau, variabelniveau og kategoriniveau.

Dokumentation på studieniveau

Dokumentation på studieniveau er informationer om et datamateriales tilblivelse: navnet på den ansvarlige forsker/organisation, tidspunkt for dataindsamling, antal respondenter etc.

Vores lokaliseringsskema afspejler de direkte krav til studiedokumentationen ved aflevering.

Dokumentation på variabelniveau

Dokumentation på variabelniveau er systematisk beskrivelse af de enkelte variable i systemfilen. Denne foretages direkte i statistik-programpakken. De fleste variable afspejler spørgsmål i det anvendte spørge-/registreringsskema, men der kan også være tale om konstruerede variable.

Ikke alle programpakker har plads til al den nødvendige dokumentation. I så fald må den resterende dokumentation opbygges separat.

Variabelnavn

Der er ikke særlige krav til variabelnavne.

Variabelbredde

Den største bredde, variabelen må have, angives.

Decimaler

Har en variabel decimaler, skal værdien for decimaler være sat til det største antal decimaler, som anvendes. Den nødvendige bredde kan konstateres ved at fremstille frekvenstabeller for variable med decimaler.

Variabellabels

Variabellabels skal indeholde spørgsmålsnummer og spørgsmålstekst fra spørgeskemaet (fx "Spm. 22: Hvad er din senest opnåede skoleuddannelse?").

Dokumentation på kategoriniveau

Dokumentation på kategoriniveau er beskrivelse af kodeværdier, dvs. svarkategorier for de enkelte spørgsmål i spørge-/registreringsskemaet og for rekodede værdier.

Missing

Der bør skelnes mellem tre forskellige former for manglende dataværdier: uoplyst, irrelevant og deltager ikke.

Uoplyst er tilfælde, hvor en variabel burde have en gyldig værdi, men ikke har det, fx fordi spørgsmålet ved en fejl ikke er blevet besvaret. Koden 9, 99, 999 etc. anbefales. For koderne 1, 2 og 3 bliver uoplyst 9, for koderne 1-XX bliver uoplyst 99, for koderne 1-XXX bliver uoplyst 999.

Irrelevant anvendes, når der retmæssigt ikke foreligger gyldig værdi for en variabel, fx hvis respondenterne pga. en given besvarelse er blevet bedt om at springe til et spørgsmål længere fremme i spørgeskemaet (filtrering). Koden uoplyst + 1 anbefales, dvs. koden bliver 10, 100, 1000 etc. Kodeværdien 10 anvendes, hvis kodeværdien for uoplyst er 9.

Deltager ikke anvendes, når en respondent som følge af undersøgelsens design har været udelukket fra en del af dataindsamlingen, fx hvis kun en bestemt del af respondenterne indgår i en supplerende helbredstest. Den anbefalede kodeværdi er uoplyst + 2. Dvs. kodeværdien bliver 101, hvis kodeværdien for uoplyst er 99.

Kodeværdilabels

Der skal være labels til alle kodeværdier for kategoriske variable. Der skal altid tilføjes labels til manglende og andre ikke-valide dataværdier (missing) i en variabel.

Kodeværdier for svarkategorier skal indeholde teksten fra spørge-/registreringsskemaet til den pågældende kodeværdi (fx 1 "Folkeskole", 2 "Studentereksamen, HF eller lignende", 3 "Ved ikke". Desuden tilføjes kodeværdilabels for missing (fx 9 "Uoplyst", 10 "Irrelevant", 11 "Deltager ikke").

Datachecks inden analyse og aflevering

Inden analyse og aflevering af data til arkivering i Dansk Data Arkiv er det en god ide at gå materialet igennem for eventuelle fejl. En stor fejlkilde er filtreringer.

Filtercheck

Mange gange vil det ikke være alle respondenter, der skal besvare et spørgsmål, men kun den del af respondentgruppen, der har besvaret et forudgående spørgsmål på en bestemt måde, jf. ovenfor.

Man taler i den forbindelse om den filtrerende variabel som den variabel, der afgør, hvilke respondenter, der skal besvare det følgende spørgsmål (fx "Har du erhvervsarbejde?"). Den filtrerede variabel er den variabel, der kun besvares af en del af respondentgruppen (fx "Hvis ja, hvor mange timer arbejder du om ugen i gennemsnit?").

Eksempel – fejltyper og valide/ikke-valide udfald for filtreret variabel:

		<i>Har du erhvervsarbejde (ja/nej)?</i> (Filtrerende variabel)		
		Ja, er i arbejde	Nej, er ikke i arbejde	Ubesvaret
<i>Hvis ja, hvor mange timer arbejder du om ugen i gennemsnit?</i> (Filtreret variabel)	Besvaret	Svar Spm. er besvaret som anvist.	Fejl type 1 Spm. skulle ikke have været besvaret.	Fejl type 2 Uvist om besvarelse skulle finde sted eller ej.
	Ikke besvaret	Uoplyst Spm. skulle have været besvaret.	Irrelevant¹ Spm er på korrekt vis ikke besvaret.	Uoplyst Uvis

¹ Er disse svar kodet som uoplyst, bør de omkodes til irrelevant, så der kan skelnes i dataanalysen.

Det er en god ide at optælle svar af type 1 og fejl af type 2. Fejl af type 2 optælles for den filtrerende variabel, mens fejl af type 1 optælles for den filtrerede variabel.

Filterfejl er almindelige, men er der mange, bør omstændighederne undersøges nærmere.

Check af kodeværdier

Vha. frekvenstabeller for hver variabel kan det checkes, at alle kodeværdier har kodeværdilabels.