Hvordan jobbe med data du ikke ser

Temakurs

19.10.2022

Johan Sjøberg (SSB), Hans Christian Halvorsen (SSB), Trond Pedersen (Sikt)

Hvordan ser datasettene ut?

- Wide-datasett (import)

Eksempel: Datamatrise ved bruk av import (4 variabler)

ID	Variabel 1	Variabel 2	Variabel 3	Variabel 4
123456	1	200000	0301	1
135791	1	410000	0301	1
147036	2	515000	1201	sysmiss
159371	2	309011	1101	sysmiss
160505	2	357000	1101	1
173951	2	399000	0301	3

import

Variabel 1: Status per 2005-10-01

Idnr	Verdi	Start	Stopp
A	1	2000-01-01	2002-12-10
A	2	2002-12-11	2007-10-10
B	1	2010-04-02	2011-05-05
B	2	2011-05-06	2014-01-01
B	3	2014-01-01	2016-12-31
C	1	2005-06-20	2016-12-31
D	2	1999-04-05	2001-08-16
D	3	2001-08-17	2005-04-14
D	4	2005-04-15	2011-03-05

Variabel 2: Status per 2005-10-01

Idnr	Verdi	Start	Stopp
A	10 000	2003-01-01	2003-12-31
A	150 000	2004-01-01	2004-12-31
A	160 000	2005-01-01	2005-12-31
B	220 000	2003-01-01	2003-12-31
B	250 000	2004-01-01	2004-12-31
B	300 000	2005-01-01	2005-12-31
C	90 000	2003-01-01	2003-12-31
C	120 000	2004-01-01	2004-12-31
C	170 000	2005-01-01	2005-12-31
D	320 000	2003-01-01	2003-12-31
D	400 000	2004-01-01	2004-12-31
D	440 000	2005-01-01	2005-12-31

Variabel 1

Idnr	Verdi V1	Start	Stopp
A	2	2002-12-11	2007-10-10
C	1	2005-06-20	2016-12-31
D	4	2005-04-15	2011-03-05

Variabel 2

Idnr	Verdi V2	Start	Stopp
A	160 000	2005-01-01	2005-12-31
C	170 000	2005-01-01	2005-12-31
D	440 000	2005-01-01	2005-12-31

Variabel 2

Idnr	Verdi V2	Start	Stopp
A	160 000	2005-01-01	2005-12-31
B	300 000	2005-01-01	2005-12-31
C	170 000	2005-01-01	2005-12-31
D	440 000	2005-01-01	2005-12-31

Hvordan ser datasettene ut?

- Long-datasett (import-panel og reshape-to-panel)

Eksempel: Datamatrise ved bruk av import-panel (3 variabler, 3 måletidspunkt)

ID	Tid	Variabel 1	Variabel 2	Variabel 3
123456	2000-01-01	1	200000	0301
123456	2001-01-01	1	210000	0301
123456	2002-01-01	2	215000	1201
135791	2000-01-01	2	305011	1101
135791	2001-01-01	2	301000	1101
135791	2002-01-01	3	299000	0301
147036	2000-01-01	1	150000	2030
147036	2001-01-01	1	159000	2030
147036	2002-01-01	3	199000	0301

import-panel

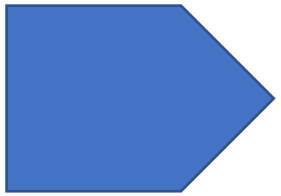
Variabel 1 og 2: Status per 2005-01-01 og 2006-01-01

Variabel 1

Idnr	Verdi	Start	Stopp
A	1	2000-01-01	2002-12-10
A	2	2002-12-11	2007-10-10
B	1	2010-04-02	2011-05-05
B	2	2011-05-06	2014-01-01
B	3	2014-01-01	2016-12-31
C	1	2005-06-20	2016-12-31
D	2	1999-04-05	2001-08-16
D	3	2001-08-17	2005-04-14
D	4	2005-04-15	2011-03-05

Variabel 2

Idnr	Verdi	Start	Stopp
A	10 000	2003-01-01	2003-12-31
A	150 000	2004-01-01	2004-12-31
A	160 000	2005-01-01	2005-12-31
B	220 000	2003-01-01	2003-12-31
B	250 000	2004-01-01	2004-12-31
B	300 000	2005-01-01	2005-12-31
C	90 000	2003-01-01	2003-12-31
C	120 000	2004-01-01	2004-12-31
C	170 000	2005-01-01	2005-12-31
D	320 000	2003-01-01	2003-12-31
D	400 000	2004-01-01	2004-12-31
D	440 000	2005-01-01	2005-12-31



Paneldatasett: Variabel 1 og 2, alle måletidspunkt

Idnr	date@panel	Verdi V1	Verdi V2
A	2005-01-01	2	160 000
A	2006-01-01	2	sysmiss
B	2005-01-01	sysmiss	300 000
B	2006-01-01	sysmiss	sysmiss
C	2005-01-01	sysmiss	170 000
C	2006-01-01	1	sysmiss
D	2005-01-01	3	440 000
D	2006-01-01	4	sysmiss

Hvordan ser datasettene ut?

- Forløpsdatasett (import-event)

Eksempel: Datamatrise ved bruk av import-event (tidsintervall: 2000-01-01 - 2003-01-01)[1]

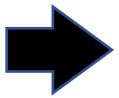
ID	Start	Stopp	Variabel
123456	2000-01-01	2000-05-30	1
123456	2000-05-31	2001-12-31	4
123456	2002-01-01	2003-08-15	2
135791	2000-04-10	2002-03-03	2
135791	2002-03-04	2002-11-11	3
147036	2002-02-28	2004-07-16	1

[1] Merk at alle hendelser som overlapper med perioden 2000-01-01 - 2003-01-01 tas med ved import

import-event

Variabel: Alle hendelser fra 2000-01-01 til 2005-12-31

Idnr	Verdi	Start	Stopp
A	1	2000-01-01	2002-12-10
A	2	2002-12-11	2007-10-10
B	1	2010-04-02	2011-05-05
B	2	2011-05-06	2014-01-01
B	3	2014-01-01	2016-12-31
C	1	2005-06-20	2016-12-31
D	2	1999-04-05	2001-08-16
D	3	2001-08-17	2005-04-14
D	4	2005-04-15	2011-03-05



Idnr	Verdi	Start	Stopp
A	1	2000-01-01	2002-12-10
A	2	2002-12-11	2007-10-10
C	1	2005-06-20	2016-12-31
D	2	1999-04-05	2001-08-16
D	3	2001-08-17	2005-04-14
D	4	2005-04-15	2011-03-05

collapse(count) var1, by(idnr)

Idnr	Verdi
A	2
C	1
D	3

NB! Etter collapse(count) erstattes variabelverdien med antall observasjoner/ records per enhet

Hvorfor er ikke tall fra microdata.no identiske med offisiell statistikk?

- Datakilden har mye å si for avvikene
- Forskjellige datakilder gir også forskjellige tall
- Ulike måletidspunkt
- Ulike populasjoner
- Forskjeller i produksjon/tilrettelegging (på tross av samme datakilde)
- Personvernfilter i microdata.no støylegger frekvenser og sensurerer ekstremverdier

Eksempel 1: Befolkning - kjønn og alder

	Personer	
	2020	
	Menn	Kvinner
0-5 år	180 788	170 371
6-15 år	329 313	313 063
16-19 år	131 197	124 017
20-66 år	1 683 981	1 607 430
67 år eller eldre	381 283	446 137

		kjønn		
		1 - Mann	2 - Kvinne	*Total*
alder	1 - 0-5 år	180770	170357	*351117*
	2 - 6-15 år	329248	312947	*642189*
	3 - 16-19 år	131084	123899	*254977*
	4 - 20-66 år	1683527	1607202	*3290728*
	5 - 67 år eller eldre	381271	446127	*827410*
	Total	*2705908*	*2660533*	*5366433*

Avvik			
	Menn	Kvinner	
0-5 år	-0,01 %	-0,01 %	-0,01 %
6-15 år	-0,02 %	-0,04 %	-0,03 %
16-19 år	-0,09 %	-0,10 %	-0,09 %
20-66 år	-0,03 %	-0,01 %	-0,02 %
67 år eller eldre	0,00 %	0,00 %	0,00 %
	-0,02 %	-0,02 %	-0,02 %

Kilde: SSBs befolkningsstatistikk (Statistikkbanken)
vs microdata.no (BEFOLKNING_KJOENN og BEFOLKNING_FOEDSELS_AAR_MND, seleksjon: BEFOLKNING_STATUSKODE per 2020-01-01 = 1 (kun permanent bosatte individer med fnr)

Eksempel 2: Regsys

	Sysselsatte personer etter bosted
	2020
0 Hele landet	2 681 541

jobbstatus	
0 - Utenfor arbeidsstyrken (kun årlig)	1280919
1 - Lønnstaker	2521502
2 - Selvstendig (kun årlig)	160036
3 - Helt ledig (annen definisjon enn i statistikken over arbeidsledige, se dokumentasjonen)	78912
Total	*4041374*

2 681 538 (≈ likt)

Kilde: SSBs registerbaserte sysselsettingsstatistikk (Statistikkbanken: 13470: Sysselsatte per 4. kvartal, etter region, statistikkvariabel og år)
vs microdata.no (REGSYS_ARB_ARBMARK_STATUS 2020-11-16)

Eksempel 3: Arblonn

	2020K4
Antall lønnstakere	2 625 347

Variabel	Gj.snitt	Std.avvik	Antall	1%	25%	50%	75%	99%
stillingspst	87.481	27.9165	2625623	5	90	100	100	140

+ 276 (≈ likt)

		dnr		
		0 - Ikke D-nummer	1 - D-nummer	*Total*
idnrtype	1 - Fødselsnummer	2556547	1118	*2557670*
	2 - D-nummer	-	59615	*59617*
	SYSMISS	264	8065	*8332*
Total		*2556821*	*68801*	*2625623*

Inkluderer alle med dnr, som utgjør 68 801 arbeidstakere

Kilde: SSBs sysselsettingsstatistikk fra A-ordningen (Statistikkbanken: 11652: Lønnstakere, jobber og lønn, etter statistikkvariabel og kvartal)
vs microdata.no (ARBLONN_PERS_SUM_STILLINGSPST 2020-11-16, BEFOLKNING_MRK_FNR, ARBLONN_PERS_DNR 2020-11-30)

Eksempel 4: Trygd

	Januar	Februar	Mars	April	Mai	Juni	Juli	August	September	Oktober	November	Desember
I alt	**340687**	**341954**	**343438**	**345015**	**346074**	**346825**	**347606**	**347859**	**348778**	**350296**	**351314**	**352197**
Tom. 49	2119	2147	2196	2245	2282	2328	2374	2392	2432	2491	2536	2558
50-69	39456	39485	39592	39690	39789	39827	39798	39681	39643	39704	39629	39415
70-99	15282	15406	15547	15722	15805	15872	15947	15956	16022	16129	16211	16240
100	283830	284916	286103	287358	288198	288798	289487	289830	290681	291972	292938	293984

Variabel	Gj.snitt	Std.avvik	Antall	1%	25%	50%	75%	99%
uførgrad	93.5137	15.5367	352201	50	100	100	100	100

+ 4 (≈ 0%)

Kilde: NAV-statistikk. PST306 Mottakere av uføretrygd. Uføregrad. Kjønn. År. Måned. (https://www.nav.no/no/nav-og-samfunn/statistikk/aap-nedsatt-arbeidsevne-og-uforetrygd-statistikk/uforetrygd/uforetrygd-ma%CC%8Anedsstatistikk/arkiv-uforetrygd-manedsstatistikk-januar-desember-2019)
vs microdata.no (UFOERP2011FDT_GRAD 2019-12-31)

Eksempel 5: Inntekt

	2019		2020	
	Talet på personar med beløp	Millionar kroner	Talet på personar med beløp	Millionar kroner
YRKESINNTEKTER	3 205 647	1 448 852,4	3 163 370	1 460 959,8
Lønnsinntekter	3 096 908	1 366 035,9	3 057 342	1 378 913,0
Netto næringsinntekter	315 830	82 816,5	310 727	82 046,7

Snitt = 451 016,93

Variabel	Gj.snitt	Std.avvik	Antall	1%	25%	50%	75%	99%
lønn	442746.2326	334952.6703	3057336	2050	162491	431563	620633	1641021

-8 271 (-1,8%)

Tall uten sdc:

Variabel	Gj.snitt	Std.avvik	Antall	1%	25%	50%	75%	99%
lønn	451017.403	427432.3262	3057339	1	162491	431563	620633	154698773

+0,5 (≈ 0%)

Faktisk gjennomsnitt: + 8 271 (+ 1,9%)

Kilde: SSBs inntekts- og formuesstatistikk for husholdninger (https://www.ssb.no/inntekt-og-forbruk/inntekt-og-formue/statistikk/inntekts-og-formuesstatistikk-for-husholdninger) vs microdata.no (INNTEKT_LONN 2020-12-31)

Teknikker for å kontrollere sine data

- Kjør ut frekvenstabeller for å kontrollere alle operasjoner du gjør. Virker frekvenstallene riktige ut i fra det du forventer? Har populasjonen forventet størrelse? Er kjønnsfordelingen som forventet?

- Starte med en liten populasjon og få variabler, sjekke at alt virker logisk/konsistent vha. tabeller, og bygge på mer og mer til du har datasettet du trenger

<u>Eksempel:</u>

Starte med et enkelt årskull, sjekke at dette stemmer i antall. Importere kjønn - sjekk kjønnsbalansen (ca 50/50). Importere utdanning – etter 20 år skal de fleste ha fullført VG

Teknikker for å kontrollere sine data II

- Kontrollere opp mot det vi vet om yrkesdeltakelse:
    - Sjekk lønnsfordelingen: Kvinner tjener generelt mindre enn menn
    - Deltid: Typisk yngre og kvinner
    - Sjekk yrkesdeltakelse i ulike aldersgrupper
    - Tilhører dine individer forventet næringsgruppe/yrkesgruppe?

Teknikker for å kontrollere sine data III

- Visualisere/skissere datamatriser -> gjør det lettere å skjønne hva som skjer

- Har du tilgang til rådata? Lag et tilsvarende datasett i for eksempel Stata og se hvordan operasjoner påvirker datasettet. Bør kunne forvente liknende effekt på datasett i microdata.no

Feilkilder mot offisiell statistikk

- Populasjonen er ikke den samme
- Datakilden er ikke den samme
- Statistikken benytter imputerte verdier
- Støyleggingen i microdata.no og i statistikken er ikke den samme