

Arbeidsnotater

S T A T I S T I S K S E N T R A L B Y R Å

OSLO: Postboks 8131 Dep, Oslo 1
Tlf. (02) *41 38 20

KONGSVINGER: Postboks 510, Stasjonssida, 2201 Kongsvinger
Tlf. (066) *14 988

IO 79/5

7. februar 1979

SAMMENHENGER INNENFOR ET SYSTEM FOR SOSIODEMOGRAFISK STATISTIKK. EN OVERSIKT OVER NOEN PROBLEMSTILLINGER.

Av

Eivind Hoffmann

INNHold

	Side
Innledning	1
Ulike kilder til sammenheng	2
Merknader om enheter i sosial- og personstatistikk	3
Sammenhenger ved å knytte sammen observasjoner for de samme enheter .	5
Sammenhenger basert på logiske og formelle strukturer	7
(a) Personregnskap	7
(b) Regnskapssammenhenger ellers	8
Sammenhenger basert på tids-relasjoner	9
(a) Tidsregnskap	9
(b) Koordinering av tidsenheter, referanseperioder og defini- sjoner av aktiviteter	9
Sammenhenger avledet fra felles medlemskap i større enheter	11
Sammenhenger ved samordnede definisjoner og klassifikasjoner	12
Hvilke sammenhenger ønsker vi å etablere	13
Avsluttende merknader	14
Henvvisninger	15

Ikke for offentliggjøring. Dette notat er et arbeidsdokument og kan siteres eller refereres bare etter spesiell tillatelse i hvert enkelt tilfelle. Synspunkter og konklusjoner kan ikke uten videre tas som uttrykk for Statistisk Sentralbyrås oppfatning.

INNLEDNING

De ulike delene av statistikken over menneskelige og sosiale forhold - den sosiale og demografiske (sosiodemografiske) statistikken - inngår på mange måter i en helhet. Dette har delvis sitt utspring i de saksområdene som statistikkene tar opp, delvis i de kjennemerkene som inngår i statistikkene, og delvis i den anvendelse som gjøres av de statistiske data. Utgangspunktet for det arbeidet som er blitt betegnet som et system for sosiodemografisk statistikk, har vært behovet for å sikre at definisjoner og klassifikasjoner blir utformet slik at de samme forhold blir behandlet på en konsistent måte innenfor hele området sosial- og personstatistikk, slik at den helhet som eksisterer i den "virkeligheten" - i det samfunn - vi skal beskrive også så langt som mulig blir tatt vare på i statistikkarbeidet. Det har også vært viktig å ta vare på sammenhengene med den økonomiske statistikken - som organiseres gjennom Nasjonalregnskapssystemet - og til miljø- og ressursstatistikken, der man også har startet flere arbeider med tanke på å få en helhetlig organisering av de ulike delstatistikkene.

Nøkkelpbegreper i dette arbeidet - som "integrering", "forbindelser" og "sammenheng" - har aldri fått noen samlet beskrivelse eller definisjon. Dette har skapt adskillige vanskeligheter og usikkerhet. Bakgrunnen er de mange ulike muligheter som eksisterer for å lage sammenhenger mellom statistiske data. Hver av disse mulighetene gir grunnlag for en egen type "sammenheng". Emnet er diskutert i [1], særlig på ss. 19-22, men diskusjonen er meget knapp og det er mange sider ved emnet som ikke er tatt opp der - langt mindre besvart. I dette notatet er det forsøkt å peke på noen av de spørsmål og problemer som inngår i en diskusjon av sammenheng i et SSDS. Termen "sammenheng" vil bli brukt generelt om hele emnet; slik termen "connections" nyttes i [1], mens termer som "integrering" og "forbindelser" bør nyttes om mer presist avgrensede begreper.

ULIKE KILDER TIL SAMMENHENG

La oss først se på hvilke kilder det kan være til sammenheng mellom ulike sett av statistiske data.

De sammenhenger mellom sett av statistiske data som vi er i stand til å etablere, vil avhenge av hvordan disse statistiske data er innsamlet og organisert. Nyttien av å kunne etablere sammenhenger mellom ulike sett av statistiske data vil avhenge av hvordan disse sammenhengene er etablert og det formål som vi har for bruken av data. De ulike former for sammenhenger er ikke alternativer. Vi kan ha kombinasjoner av ulike typer sammenhenger og ulike kilder til dem, og i en konkret situasjon må vi ta hensyn til at de alle kan være tilstede. Nedenfor er gitt en oversikt over noen hovedtyper av sammenhenger.

- (1) Sammenhenger etablert ved å forbinde - integrere - observasjoner om den samme enhet fra ulike kilder. I prinsippet er dette den enkleste og mest naturlige måte å knytte sammen statistiske data så lenge enhetene er personer eller andre enkle primærenheter, men dette er langt mer uklart og vanskelig når enhetene er sammensatte eller aggregater - enheter som er avledet fra de primære enheter f.eks. ved aggregering.
- (2) Sammenhenger som følger av at de statistiske data inngår i et system der begreper, definisjoner og klassifikasjoner følger av en fast logisk struktur. Slike logiske strukturer har vanligvis én bestemt anvendelse av de statistiske data som utgangspunkt, men de kan vise seg å være nyttige også i andre anvendelser av de samme eller tilknyttede data.
- (3) Sammenhenger som bygger på at menneskelige aktiviteter og begivenheter er tidkrevende eller følger etter hverandre i tid. Til en viss grad vil dette representere spesialtilfelle av (2), men de spørsmål som knytter seg til samtidighet og rekkefølge av aktiviteter og begivenheter, gjør at det er hensiktsmessig å behandle slike sammenhenger for seg.
- (4) Sammenhenger som bygger på at statistiske enheter på et nivå er medlemmer av større enheter - enheter på et høyere nivå. De større enhetene vil ha kjennemerkeverdier som da er felles for alle sine medlemmer på det lavere nivå. Slike kjennemerkeverdier ved den høyere enheten kan være avledet av kjennemerker ved de enkelte medlemmer (f.eks. gjennomsnittsinntekten, inntektsfordelingen); eller de kan være et direkte kjennemerke ved den høyere enheten (f.eks. størrelsen).

(5) Andre sammenhenger som bygger på bruken av samordnede definisjoner av variable og samordnede klassifikasjoner.

I det som følger skal vi prøve å utvikle disse ideene noe videre, etter noen merknader om enheter innenfor sosial- og personstatistikken.

MERKNADER OM ENHETER I SOSIAL- OG PERSONSTATISTIKK

De analyser som anvender statistikk som regnes med til person- og sosialstatistikk (dvs. som vil bli organisert innenfor et SSDS) omhandler personer og grupper av personer, forbindelser mellom slike enheter og forbindelsene mellom slike enheter og samfunnets institusjoner (i vid betydning). De enhetene som er gjenstand for analyse (analyseenhetene) er vanligvis personer eller persongrupper (f.eks. familier, husholdninger, befolkninger), og i noen tilfeller de institusjoner som betjener eller på annen måte påvirker personer eller grupper. De statistiske enhetene (observasjonsenhetene eller tabelleringsenhetene) kan imidlertid også være noe annet enn personer eller persongrupper. De kan f.eks. også være begivenheter, transaksjoner eller institusjoner. Hva som er den statistiske enhet henger nær sammen med hva som er grunnlaget for statistikken - hvordan den er samlet inn - og hvordan den blir presentert. Det kan ofte vise seg at enhetene i statistikkene er forskjellige fra de enhetene som analysene dreier seg om.

Når vi diskuterer sammenhenger mellom ulike sett av statistiske data, så må vi gjøre klart hva som er enhetene i de ulike statistikkene, og hvilke analyseenheter vi er opptatt av. Kjennskap til hvilke enheter som er involvert er nødvendig for å kunne avgjøre hvilke sammenhenger som kan etableres og om de er relevante for den bruk vi ønsker å gjøre av disse dataene. Ofte er det lett å avgjøre hva som er analyseenhet, og vi skal i det som følger forutsette at vi kjenner den. Hva som er enhet i statistikken vil det ofte være mindre enkelt å avgjøre, og det vil endres med organiseringen av data.

a: La oss se på et eksempel fra helsestatistikken: De primære observasjonsenhetene kan være: "syketilfelle", "pasienter", "personer" og "institusjoner". Den samme "pasient" kan ha to (eller flere) "syketilfelle" samtidig eller i samme periode, og derfor "stå bak" mer enn et "syketilfelle"

i statistikken. Den samme "personen" kan være innlagt (behandlet) som "pasient" mer enn en gang (eller ved mer enn et sykehus) i løpet av referanseperioden, og derved "stå bak" mer enn en "pasient" i statistikken. Det er en mulig kilde både til forvirring og til å etablere sammenhenger mellom ulike datasett, at "syketilfelle" kan tilknyttes kjennemerker ved de "pasienter" og "personer" som det dreier seg om, og at "personer" kan tilknyttes kjennemerker som følger av deres "syketilfeller" og deres rolle som "pasienter" (f.eks. arten av syketilfelle, innleggelsens varighet). Deter naturlig å snakke om at enheter og kjennemerker er "kjedet sammen" i slike tilfeller der en enhet "låner" kjennemerker fra andre enheter, dvs. har et kjennemerke som er avledet av en annen enhets kjennemerke.

b: Når vi grupperer sammen de primære observasjonsenheter som har bestemte verdier på et kjennemerke, og så beregner hvor mange som faller i hver gruppe, eller en parameter som beskriver fordelingen av en annen variabel innenfor hver gruppe, da har vi samtidig etablert et nytt sett statistiske enheter og fått "observasjoner" av kjennemerker ved disse enhetene. Hvis vi f.eks. beregner gjennomsnittsinntektene for alle inntektstakere i hvert fylke, så har vi samtidig etablert enhetene "inntektstakere i fylke x" og observert kjennemerket "gjennomsnittsinntekt" for disse enhetene. Vi har også etablert et kjennemerke ved enheten "fylke x" (nemlig "gjennomsnittsinntekt" for inntektstakerne i fylket") og andre kjennemerker kan opprettes på tilsvarende måte - ved å avlede dem fra observasjoner av enheter som ikke er fylket, men som er knyttet til det på en eller annen måte. Dette er et eksempel på en litt annen form for sammenkjedete kjennemerker enn den vi tok opp foran.

c: Det er viktige forskjeller når det gjelder hvor permanente de ulike enheter er. Noen enheter kan for alle praktiske formål regnes for å være helt stabile, jfr. enheten fylke. For en del enheter er det nokså enkelt å avgjøre når de oppstår og når de opphører å eksistere, f.eks. personer. En del enheter er høyst ustabile eller graden av stabilitet er sterkt avhengig av hvordan de er definert. Vi kan skille mellom tre ulike måter å definere persongruppe-enheter - og de gir opphav til enheter med særdeles ulik grad av stabilitet:

(1) Hvis, for eksempel, en familieenhet defineres som å bestå av et gitt sett personer NN, PP, RR og SS, da vil vi ha en ny familieenhet når SS forlater enheten, eller hvis KK slutter seg til gruppen. Dette er altså eksempel på en gruppe-enhet der definisjonen krever at den skal bestå av et

bestemt sett av personer som har bestemte relasjoner til hverandre, eller som har et bestemt sett kjennemerkeverdier. Enheten blir oppløst så snart et av medlemene forlater gruppen, eller så snart et nytt medlem slutter seg til den fordi personen(e) forandrer en kjennemerkeverdi eller sin relasjon til de andre i gruppen. Gruppe-enheter definert på denne måten kan da være høyst ustabile.

(2) En familieenhet kan defineres som å bestå av det settet av personer som har en bestemt form for relasjoner til en referanseperson. En enhet definert på denne måten vil da eksistere så lenge referansepersonen eksisterer (som referanseperson), uavhengig av hvilke andre personer som er med i gruppen. En slik gruppe-enhet vil normalt være mer stabil - mer varig - enn tilsvarende enheter som bygger på definisjoner av type (1).

(3) Mange gruppe-enheter defineres som det settet av personer (innenfor en gitt befolkning) som har et bestemt sett kjennemerkeverdier. Gruppe-enheten eksisterer så lenge det er minst en person som har de definerende kjennemerkeverdier, og det kan skje en fullstendig utskifting av de personer som inngår i gruppen uten at gruppe-enheten opphøres. Gruppe-enheter definert på denne måten vil gjerne være svært varige. Ofte vil den eksplisitt formulerte definisjon av gruppe-enheter innenfor sosial og demografisk statistikk være av denne typen, men det er ofte en underliggende antakelse at gruppe-enheten i det alt vesentlige består av de samme personer over tid (og innenfor ulike sett av data). I praksis holder ofte ikke en slik underliggende antakelse.

SAMMENHENGER VED Å KNYTTE SAMMEN OBSERVASJONER FOR DE SAMME ENHETER

Diskusjonen i dette avsnittet vil forutsette - med mindre noe annet er uttrykkelig sagt - at vi har to sett av statistiske data som er blitt samlet inn på samme tidspunkt (for samme referanseperiode) fra to ulike kilder/undersøkelser. Vi vil diskutere mulighetene for å knytte sammen disse datasett med tanke på å analysere forhold ved personer og deres adferd.

a: Enklest er det i de tilfeller der vi har individ-data, enkeltobservasjoner om personer, der hver person har et identifiserende kjennemerke (f.eks. personnummer) som er det samme i de to datasett. For alle de personer som

er observasjonsenheter i begge data-sett vil det være mulig å nytte personnummeret for å få til et fullstendig integrert datasett, og det vil være mulig å nytte alle de tilgjengelige opplysningene simultant. For personer utenfor dette integrerte datasett vil det da mangle noen opplysninger, nemlig de kjennemerker som inngår i "det andre" datasettet. I noen tilfeller av slike delvis integrerte datasett kan det være mulig å sette inn "estimerte" kjennemerkeverdier der hvor de faktiske mangler ved hjelp av de teknikker og modeller som er utviklet for å håndtere manglende observasjoner. I andre tilfeller vil det være nødvendig å la de manglende observasjoner mangle, eller å bare utnytte settet av helt integrerte data.

b: Noe mer komplisert er det i de tilfeller der vi enten har enkeltobservasjoner om de samme personer i to ulike data-sett, men intet felles identifiserende kjennemerke; eller vi har observasjoner om de samme personer, men bare tilgang til deres fordeling etter noen kjennemerker - og altså ikke til de enkelte individopplysninger. Vi kjenner m.a.o. fordelingene n_{ijk}^A og n_{ijl}^B fra datasettene A og B, der kjennemerket \bar{I} (med verdiene $i=1, \dots, I$) og kjennemerket \bar{J} ($j=1, \dots, J$) er observert i begge datasett, mens kjennemerkene \bar{K} ($k=1, \dots, K$) og \bar{L} ($l=1, \dots, L$) bare er observert for enheten i henholdsvis datasettene A og B. Dette betyr at $\sum_k n_{ijk}^A = \sum_l n_{ijl}^B$ ($i=1, \dots, I; j=1, \dots, J$). For disse enhetene, dvs. de individgrupper som blir definert av ulike kombinasjoner av verdier for kjennemerkene \bar{I} og \bar{J} , har vi da et integrert datasett med henblikk på \bar{L} og \bar{K} - vi kjenner fordelingen av \bar{L} og \bar{K} for hver av de persongruppe-enheter som er definert ved verdier av \bar{I} og \bar{J} . Disse persongruppe-enheter er altså definert slik at de består av de samme personer innenfor begge datasett - jfr. det som er sagt ovenfor og i avsnittet om enheter. Alt dette trenger imidlertid ikke ha særlig verdi dersom vår analyse gjelder personer og ikke slike grupper som man får definert ved kjennemerkeverdier for \bar{I} og \bar{J} . Hvis vi f.eks. vet at 33 prosent av 19-årige norske menn i 1971 deltok i utdanningssystemet, og at det samme år var 2 prosent av 19-årige norske menn siktet for forseelser eller forbrytelser, så er det naturligvis ikke mulig på grunnlag av dette å si noe om størrelsen på andelen av 19-årige menn i utdanningssystemet som ble siktet, men det er mulig å si noe om hvilket mulighetsområde det er for denne andelen.

I praksis har vi i de fleste tilfeller observert N personer i datasett A og M personer i B, og har bare utilstrekkelige kunnskaper om i hvilken grad de to datasettene har med de samme personer. Vi vil da kjenne fordelingene av de N og M personene over kjennemerkene \bar{I} og \bar{J} (og finner f.eks. at $n_{ij} < m_{ij}$), uten å vite i hvilken grad noen av de n_{ij} personene også er med blant de m_{ij} . (Hvis A og B er uavhengige tilfeldige utvalg, kan vi - med stor sikkerhet - forutsette at A og B ikke har noen personer felles.) For hver kombinasjon av verdier av \bar{I} og \bar{J} kan vi også i dette tilfellet definere persongrupper som statistiske enheter, men nå i en svakere forstand enn ovenfor ved at disse gruppene er definert som enheter uavhengig av hvilke personer som er med i gruppene - det er tilstrekkelig at de har de definerende kjennemerker. Dette betyr igjen at den analytiske verdi av disse enhetene kan være begrenset, og at mulighetene for å integrere statistiske data for slike enheter vil være av begrenset verdi.

Metodiske spørsmål av denne typen har i stor grad vært diskutert i samfunnsvitenskaplig metodelitteratur - delvis under betegnelsen "økologiske feilslutninger". Det er noe underlig at det nesten ikke er spor av denne diskusjonen i [1].

SAMMENHENGER BASERT PÅ LOGISKE OG FORMELLE STRUKTURER

Nasjonalregnskapssystemet (NR) er en samling av formelle definisjonssammenhenger og standard klassifikasjoner. Sammenhengene mellom de ulike sett av statistiske data som blir organisert av NR er nesten utelukkende basert på eller avledet fra disse logiske definisjonssammenhenger og klassifikasjonsstandarder:

(a) Personregnskap

Anta at en persongruppe blir definert på tidspunkt T ved å angi en kombinasjon av kjennemerkeverdier som skal gjelde for alle personer i gruppen på dette tidspunkt. Det vil da være en definisjonsmessig sammenheng mellom denne gruppen og en gruppe personer definert på tidspunkt $T+t$ ved den samme kombinasjon av kjennemerkeverdier: Gruppen på tidspunkt $T+t$ må bestå av de personer som var i gruppen på tidspunkt T pluss de som er kommet til gruppen i perioden fra T til $T+t$ (ved fødsel eller

ved å oppnå de nødvendige "kvalifikasjoner" for å høre til gruppen) minus de som har forlatt gruppen (ved død eller ved å miste de nødvendige "kvalifikasjoner" for å høre til gruppen). For eksempel, så er verdens befolkning på tidspunkt T og T+t knyttet sammen ved de som har overlevet, fødsler og dødsfall i perioden.

(Hvis gruppen er definert som en enhet uavhengig av dens personmessige sammensetning, så blir de formelle sammenhenger som er beskrevet ovenfor sammenhenger mellom sammensetningen av gruppen på de to tidspunkter.)

Mellom undergrupper er det videre slike sammenhenger som følger av at de tilsammen skal utgjøre totalpopulasjonen (hele gruppen). Dette kravet medfører logisk nødvendige sammenhenger på bevegelsene mellom gruppene, f.eks. verdens befolkning oppdelt i ulike nasjonaliteter.

Sammenhenger av dette slaget forutsetter naturligvis begreper, definisjoner og klassifikasjoner som er fullstendige og konsistente; og det må også forutsettes at der hvor det er ulike kilder til data (ulike sett av data) om den samme persongruppen må de begrepene, definisjonene og klassifikasjoner som er nyttet i de ulike datasett være samordnet og konsistente. For eksempel, hvis "tilstedeværende befolkning" ikke er konsistent definert i alle land, så er det ikke mulig å komme fram til verdens befolkning ved å legge sammen befolkningstallene fra de enkelte land.

Det er i [1] gitt flere eksempler på modeller der mulighetene for slike sammenhenger er beskrevet i detalj. Det er imidlertid bare et begrenset antall problemstillinger eller analytiske modeller der denne typen av sammenhenger er sentrale for analysen og organiseringen av data.

(b) Regnskapssammenhenger ellers

Statistikk om økonomiske, finansielle og - til en viss grad - andre ressurser, kan organiseres og bindes sammen ved regnskapssystemer. Personer og grupper av personer eier (disponerer) slike ressurser og anvender dem for ulike formål. Statistikk som nyttes for å beskrive og/eller analysere personer eller persongrupper og deres herredømme over ressurser og bruken av disse for ulike formål, kan da bindes sammen gjennom utnytting av de formelle regnskapssammenhenger. De vil også bidra til å knytte data om personer o.s.v. med andre data organisert gjennom nasjonalregnskapssystemet.

SAMMENHENGER BASERT PÅ TIDS-RELASJONER

(a) Tidsregnskap

All menneskelig aktivitet krever tid. Innenfor en gitt referanseperiode kan hver tidsenhet anvendes til ulike aktiviteter, og tallet på tidsenheter innenfor perioden er begrenset. Derved kan man lage et regnskap for disse tidsenheter og aktiviteter ved hjelp av et passende utformet regnskapssystem. Dette systemet vil da bidra til å organisere sammenhenger mellom ulike sett av statistiske data som gjelder aktiviteter definert for tidsenheter som inngår i samme referanseperiode.

Til nå er det hovedsaklig utviklet regnskapssystemer for de helt korte referanseperioder som blir anvendt i tidsbruksundersøkelser. Det eksisterer imidlertid eksempler på regnskapsopplegg også for lengre referanseperioder, jfr. Yrkeshistorieundersøkelsen.

(b) Koordinering av tidsenheter, referanseperioder og definisjoner av aktiviteter

Innenfor de ulike deler av den sosiale og demografiske statistikken nyttes et stort antall tidsenheter, referanseperioder og definisjoner av aktiviteter. Innenfor noen av disse statistikkområdene er disse tingene koordinert på en systematisk måte (f.eks. innenfor befolkningsstatistikken), men de er ikke koordinert på tvers av statistikkområder, selv ikke for de "samme" kjennemerker. Definisjonen av "sysselsatt" gir for eksempel ulike kjennemerker avhengig av hvilken referanseperiode som blir valgt. ILO-anbefalingene inneholder to ulike referanseperioder, og i anbefalingene om 1980-folketellingene kan man finne andre referanseperioder. Det er åpenbart behov for å få til en koordinering av tidsenheter, referanseperioder og aktivitetsdefinisjoner mellom de ulike statistikkområder innenfor SSDS, slik at man kan få utnyttet de muligheter som ligger i tidsregnskapssystemer og gjennom de noe mindre slagkraftige muligheter som følger av samordnede definisjoner og klassifikasjoner.

Slik koordinering vil kreve arbeid med en rekke spørsmål:

- (1) Datering - det må være samordnede regler for valg av den dato (de aktuelle datoer, når det gjelder perioder) som observasjonene skal referere seg til.
- (2) Varighet - det må være samordnede regler for å angi hvor lenge en tilstand eller aktivitet har vært. Det mest nærliggende eksempel er definisjonen av alder som har vært definert på ulike måter innenfor de

forskjellige statistikker: antall fødselsdager og fødeår er de to vanligste. I Norge er det jo bestemt at fødeår skal danne grunnlaget for aldersdefinisjonen overalt der dette er mulig.

(3) Sekvenser av aktiviteter og begivenheter - på mange områder er det institusjonelt bestemte sekvenser av aktiviteter (eller tilstander) og begivenheter. Noen av disse institusjonelle sekvenser gjelder bare som en hovedregel, mens andre følger av lovregler og bestemmelser. For å oppnå en bestemt utdanning (en tilstand) må man gjennomgå en bestemt utdanningsaktivitet (kurs, skole) og avlegge bestemte eksamener (begivenheter) - for bare å nevne et eksempel. Slike sekvenser må systematisk kartlegges og de muligheter for å binde sammen ulike sett av statistikk på dette grunnlag må utnyttes. Noen av disse mulighetene vil inngå i de personregnskapsopplegg som er nevnt tidligere.

(4) Tidsenheter og referanseperioder må bli koordinert ved at man danner et sammenhengende sett av enheter og perioder. Det synes rimelig å foreslå at det etableres et standard sett av referanseperioder av ulike lengder slik at den kortere periode alltid er tidsenhet for de lengre perioder, for eksempel slik som det er skissert her:

	<u>Tidsenhet</u>		<u>Referanseperiode</u>
1.	Kvarter (15 min.)	-	Time
2.	Time	-	Dag
3a.	Dag	-	Uke
3b.	Time	-	Uke
4.	Uke	-	Måned
5.	Måned	-	År
6.	År	-	10-års periode (hovedstadium i livssyklus)
7.	10-års periode (hovedstadium i livssyklus)	-	Hele livet.

Kombinasjonene 1-3a er vanlige i tidsbruksundersøkelser. 3b er den vanligste brukt i Arbeidskraftsundersøkelser.

Med utgangspunkt i listen ovenfor må man si at det nå er en blanding av ulike tidsenheter og referanseperioder som er i bruk - ofte innenfor samme datasett (f.eks. for ulike kjennemerker). Vanligvis kan de(n) løsning(er) som er valgt gis en god begrunnelse ut fra hva som er den primære anvendelsen av data, men de(n) skaper problemer når disse data

skal utnyttas i sammenheng med andre data. Det synes rimelig å anta at en hensiktsmessig liste over tidsenheter/referanseperiode bør være litt kortere enn den som er satt opp foran, begrenset f.eks. til 5 trinn. Det er rimelig å tenke seg at alle standard tidsenheter kortere enn referanseperioden skal kunne være tidsenheter i en gitt referanseperiode. Dette vil f.eks. eliminere det "problemet" som har ført til at både inndelingen 3a og 3b er ført opp i listen ovenfor, men det vil nok gjøre at selve regnskapssystemet og aktivitetsdefinisjonene vil få et noe mer komplisert opplegg.

(5) Når man har et koordinert sett av regler for tidsenheter og referanseperioder vil man ha et godt grunnlag for å definere begivenheter, aktiviteter og tilstander som simultane, sekvensielle eller adskilte, og så utnytte disse forhold til å knytte sammen ulike sett av data når det er hensiktsmessig.

(6) For mange formål er det nyttig og nødvendig å definere både primære og sekundære aktiviteter med referanse til en bestemt tidsenhet. Det vil i tilfelle være de primære aktiviteter som skal inngå i et tidsregnskap for en gitt referanseperiode. Registreringen av sekundære aktiviteter vil imidlertid gi verdifull og nødvendig tilleggsinformasjon. Definisjonene av primære og sekundære aktiviteter må være koordinert for de forskjellige tidsenheter og referanseperioder - for eksempel definisjonene av "sysselsetting" og "være under utdanning (skolegang)". Noen aktiviteter (eller tilstander) vil det nok bare være mulig å gi meningsfulle og nyttige definisjoner for korte tidsenheter - for eksempel aktiviteter knyttet til egenarbeid eller vanlige, kortvarige syketilfelle - mens andre bare er meningsfulle for lengre tidsenheter.

SAMMENHENGER AVLEDET FRA FELLES MEDLEMSKAP I STØRRE ENHETER

De enkelte enheter i to forskjellige sett av statistiske data kan ha de samme kjennemerkeverdier fordi de alle er med i en større enhet, som f.eks. kan være definert geografisk. Disse kjennemerkeverdiene for den større enhet kan være bestemt ut fra kjennemerker ved dens enkelte medlemmer (f.eks. gjennomsnittsinntekt, inntektsfordeling), de kan være for

kjennemerker som er knyttet direkte til den større enhet (f.eks. størrelsen) eller avledet fra andre enheter knyttet til den større enheten (f.eks. tallet på skoler i et fylke, tallet på forbrytelser). Det kan være at slike "økologiske" kjennemerker vil være viktigere i en analytisk sammenheng (når analyseenhetene er personer) enn som hjelpemiddel for å organisere sammenhenger mellom ulike sett av data. Det kan derfor diskuteres om det er riktig å ta dette med på en liste over kilder til sammenhenger mellom ulike sett av statistiske data. Imidlertid må det vel være nyttig å ha også dette momentet i minne når vi organiserer statistiske data for ulike enheter og skal koordinere definisjoner og klassifikasjoner.

Det er noen økologiske kjennemerker som har en noe annen karakter enn de vi har nevnt ovenfor. Enheter kan ha samme kjennemerkeverdi på et sentralt kjennemerke (har et sentralt forhold felles) fordi de er observert på samme tidspunkt - eller for samme periode. Det kan f.eks. være begivenheter eller forhold utenfor enhetene selv som har stor betydning - krigstilstand, naturkatastrofer o.l. Selv for de personer som ikke kan sies å være direkte berørt av slike begivenheter, vil det være av betydning at de har funnet sted. Dette er nok også forhold som hovedsaklig er av betydning i en analysesammenheng, men også slike forhold bør kunne være av interesse når man skal koordinere definisjoner og klassifikasjoner for ulike enheter.

SAMMENHENGER VED SAMORDNEDE DEFINISJONER OG KLASSIFIKASJONER

De fleste av de kilder til sammenhenger som er nevnt foran kan bare utnyttes hvis det eksisterer koordinerte definisjoner og klassifikasjoner av enheter, variable og tidsreferanser. Arbeidet med å koordinere de forskjellige definisjoner og klassifikasjoner vil derfor være fundamentalt som grunnlag for et SSDS som skal sikre at så mange som mulig av de sammenhenger som eksisterer i den virkelighet som vi skal beskrive faktisk overlever den forenkling som det er nødvendig å foreta når vi skal nytte statistikk til beskrivelsene. Slik koordinering er viktig både når det er de samme observasjonsenhetene i de ulike datasettene, og når observasjonsenhetene er forskjellige - kanskje særlig da. Som vi har vært inne på ovenfor er mange kjennemerker for personer avledet av kjennemerker ved andre enheter, akkurat som kjennemerker for andre enheter er avledet fra kjennemerker ved personer - kanskje særlig ved aggregering. Ved slik

"sammenkjeding" av kjennemerker er det nødvendig å være klar over alle ledd i kjeden for å kunne finne fram til fornuftig koordinering. I en del tilfeller vil det være slik at en enhet kan ha flere kjennemerker av samme slag som er avledet fra kjennemerker ved andre enheter (f.eks. mer enn en diagnose, arbeid i mer enn en næring, deltakelse i mer enn en aktivitet). Det vil da være nødvendig å samordne definisjonene av hva som er det "viktigste" eller primære kjennemerke når det er nødvendig å innskrenke seg til et. I slike tilfeller vil det ofte være nødvendig også å kunne definere sekundære kjennemerker etter et fast mønster (sekundær diagnose, næring, aktivitet).

HVILKE SAMMENHENGER ØNSKER VI Å ETABLERE

Dette spørsmålet vil ikke bli besvart i dette notatet. Det må i siste instans være administrative behov og analytiske krav som må avgjøre hvilke sammenhenger som må og bør etableres, og hvor sterke de skal være. Dette vil derfor variere mellom land og over tid. De erfaringene vi har, tyder imidlertid på at det er en stor kjerne av sammenhenger som det vil være nyttig å ha i de fleste tilfeller. Innholdet i denne kjernen kan bare klarlegges ved en gjennomgang av de eksisterende analytiske hovedangrepsmåter som brukes ved analyse av sosiale og demografiske data og de krav som disse angrepsmåter stiller til dataorganisering og sammenheng i data. Tilsvarende kan man undersøke hvilke krav til sammenheng i data og dataorganisering som forsøkene på å analysere den totale situasjon for enkelte grupper har ført med seg (de eldre, de funksjonshemmede, mennesker som bor i spesielle strøk av landet). Det vil være nødvendig å undersøke konkret om eksisterende eller planlagte statistikksett tilfredsstillende kravene til sammenheng i data som kommer fra slike analytiske angrepsmåter eller totale analyser, og hvor svakhetene i tilfelle ligger. Slike undersøkelser bør kunne gi grunnlag for videre utvikling av den sosiale og demografiske statistikken som et sammenhengende hele.

En av de analytiske tilnæringsmåter som har konsekvenser som det er rimelig å studere nærmere, er den livs-syklus modell som er diskutert i detalj i [1]. I [3] er det løst skissert en annen tilnæringsmåte som for en periode tar utgangspunkt i samspillet mellom personer, de ressurser

han/hun har i utgangspunktet, omgivelsene og dens ressurser, de aktiviteter han/hun engasjerer seg i og resultatene fra disse aktivitetene. Modellstrukturer med et slikt utgangspunkt har vært nyttet bl.a. som grunnlag for de levekårsundersøkelser som har vært gjennomført f.eks. i Norge og Sverige, og som også har dannet et grunnlag for mye av diskusjonen om sosiale indikatorer.

AVSLUTTENDE MERKNADER

I [2] og i dokumenter om SSDS som er utarbeidet senere, er det avgrenset et sett av felles kjennemerker og klassifikasjoner der behovet for å samordne definisjonene er særlig viktig for å få utviklet sammenhengene innenfor et SSDS. Det er også anbefalt at man skal analysere hvilke behov for sammenhenger i statistikken man får når man skal analysere situasjonen for spesielle grupper. Dette synes å være et fornuftig opplegg for det videre arbeid, og det er helt nødvendig for å komme videre med å utvikle et SSDS. Det er imidlertid mye arbeid som må gjøres med en grundig diskusjon om definisjoner av enheter og sammenhenger mellom ulike slag enheter og mellom enheter på ulike aggregeringsnivåer og deres kjennemerker. På samme måte er det mye arbeid som må gjøres når det gjelder tidsenheter, referanseperioder og samordning av regler for datering og bestemmelse av varighet. Arbeidet med tidsenheter og andre tidsrelaterte begreper er nødvendig både fordi en diskusjon av felles kjennemerker og klassifikasjoner ellers ikke kan være fullstendig, og fordi det gir et grunnlag for å knytte de mer sektor-spesifikke kjennemerker og klassifikasjoner til den generelle SSDS-strukturen. Man må imidlertid huske at den forenkling av virkeligheten som er en nødvendig del av statistikkproduksjon gjør at det ikke er mulig å ta vare på alle de sammenhenger som eksisterer i virkeligheten - eller som vi ønsker å beskrive og analysere - når vi lager statistikk; selv om vi har individ-data med felles identifiserende kjennemerker gjennom hele statistikk-systemet. Det er mitt håp at dette notatet kan gi et lite bidrag til videre arbeid som vil føre til at den sosiale og demografiske statistikk i mindre grad enn i dag beskriver virkeligheten ved isolerte fragmenter.

Dette notatet er en oversettelse til norsk av notatet "Connections within the framework of social and demographic statistics. A survey of some issues." (EiH/ÅMJ, 13/11-78) som er et grunnlag for diskusjon i en gruppe under Working Party on a Framework of Demographic and Social Statistics som arbeider under Conference of European Statisticians. Det notatet er igjen en revidert og utvidet utgave av [3] og [4]. I arbeidet med notatet har jeg hatt nytte av merknader fra de statistiske sentralbyråer i Danmark, Norge, Sverige, Finland, USSR, Ungarn, Canada og Nederland. Diskusjoner med og merknader fra Jan M. Hoem, Lennart Fastbom, Svante Øberg, Bo Sundgren, Tapio Leppo, Hans Adler og Leroy O. Stone har vært meget nyttige. Moralsk press og støtte fra Per Sevaldson gjorde det mulig å få ferdig denne versjonen av notatet. Det er imidlertid bare forfatteren som er ansvarlig for de synspunktene som det er gitt uttrykk for her. De er ikke uten videre uttrykk for Statistisk Sentralbyrås, ECEs eller UNSOs standpunkter. I denne oversettelsen er nyttet den nordiske betegnelsen SSDS, ikke FSDS som nå er FNs betegnelse på arbeidet.

HENVISNINGER

- [1] Towards a System of Social and Demographic Statistics
United Nations publications, Sales No, E.74.XVII.8. New York 1975.
- [2] Strategy for further Work on a System of Social and Demographic Statistics
CES/WP.34/21 18 March 1976 and E/CN.3/489 26/2-76.
- [3] E. Hoffmann: Linkages within the System for Social and Demographic Statistics: A Framework for Discussion. Oslo 1972.
(Mimeographed note.)
- [4] E. Hoffmann: Connections within the System of Social and Demographic Statistics. An Introductory Discussion of their Thematic and Methodological Basis.
Central Bureau of Statistics of Norway, Oslo, 6/5-76. (Mimeographed note.)