Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

IO 75/26

27. juli 1975

NOEN BETRAKTNINGER OM DATATERMINOLOGI

Av

Jan M. Hoem

INNHOLD

1. Innledning

1A. I faget statistikk beskjeftiger en seg med prinsippene for innsamling, bearbeiding og presentasjon av data, med regler for hvordan en kan få informasjon fra data, og med hvordan en kan ta beslutninger på grunnlag av slik informasjon. Faget er etter hvert blitt meget omfattende og nokså lite enhetlig, og det har delt seg opp i underdisipliner med tildels ganske forskjellige tradisjoner. Slik utviklingen er blitt, har en kommet til å legge svært forskjellig vekt på teoridannelse i de ulike underdisipliner. På noen områder har en fått en velutviklet teori, gjerne formulert ved matematikk. Andre steder har innslaget av teori derimot vært temmelig svakt, og rent beskrivende fremstillinger av aktiviteter har vært sterkt fremtredende. Dette har f.eks. inntil nylig vært nokså karakteristisk for behandlingen av prinsippene for statistikkproduksjon. I undervisningen presenteres slike spørsmål fortsatt gjerne under navn som "deskriptiv statistikk" eller "praktisk statistikk", mer eller mindre bevisst i motsetning til "matematisk statistikk" eller "teoretisk statistikk", og dette gjenspeiler nok en ganske almen holdning til feltet. Etter hvert er det imidlertid begynt å komme levende bidrag av mer teoretisk karakter. De har riktignok foreløpig først og fremst form av brokker av en teoribygning, men forhåpentligvis kan de sees som begynnelsen til en systematisk og enhetlig fremstilling av dette området.

1B. I et fag som underkastes teoretisk behandling, vil det gjerne etter hvert utvikle seg en gjennomtenkt terminologi. Tilsvarende er det lett å forestille seg at terminologien kan bli nokså lite avklart i et fagområde der teoridannelsen er svak. Etablering av en konsekvent og enhetlig teminologi virker på sin side avklarende på mange problemer og kan derved bidra til å føre den teoretiske forståelse framover.

Det er fortsatt et stykke vei fram før i hvert fall norsk terminologi for statistikkproduksjon har fått en tilfredstillende struktur. Gjeldende språkbruk i enkelte sentrale miljøer, særlig i Statistisk Sentralbyrå og blant folk i visse kvantitativt orienterte samfunnsvitenskapelige forskningsinstitusjoner, gir viktige retningslinjer, men det er lett å påpeke unødvendige unøyaktigheter, urimeligheter og inkonsekvenser i den daglige bruk av ulike betegnelser.

På denne bakgrunn er det hyggelig å se en voksende interesse for terminologiske spørsmål knyttet til statistikkproduksjonen. Blant en rekke andre impulser har det internasjonale samarbeidet vedrørende et system for sosial og demografisk statistikk (SSDS) i regi av de statistiske sentralbyråene vært med på å aktualisere behovet for en mer velutviklet teori for statistikkproduksjon, og dette har igjen brakt nødvendigheten av et bedre begrepsapparat frem i lyset. Folk knyttet til Statistiska Centralbyrån i Stockholm har nylig laget en rekke skrifter om slike spørsmål (Forsberg, 1972, Öberg, 1972, Sundgren, 1971, Datasamordningskommittén, 1974), og i Skandinavia har dette satt i gang en interessant diskusjon som fortsatt pågår (Hoffmann, 1973, Fastbom, 1973, Hoem, 1974, Gundersen, 1974, Sundgren og Öberg, 1974). Impulsene til forslagene i diskusjonen kommer fra sosiologi/statsvitenskap (f.eks. Hellevik, 1971), og fra informasjonsvitenskap (informatologi, infologi, datalogi, eller hva en nå har funnet på å kalle dette faget på de ulike nordiske språk), to fagområder som en kan håpe vil fortsette å gi verdifulle bidrag til utviklingen av en teori for produksjon av statistikk om samfunnsforhold.

1C. Dette notatet er et innlegg i denne terminologidiskusjonen. Det tar i første rekke sikte på å bidra til en presiseringog standardisering av språkbruken i Statistisk Sentralbyrå omkring en del sentrale statistiske begreper. Jeg har valgt å presentere definisjonene verbalt i en løpende tekst. Dette har ikke gjort det lett å avveie hensynene til generalitet, abstraksjon, leselighet og presisjon. En vil derfor se at begrepene presenteres med ulikt presisjonsnivå og at det i stor utstrekning appelleres til leserens egen innsikt og intuisjon.

Begreper som det etter min oppfatning har vært særlig vanskelig å beskrive, er søkt belyst med eksempler. Fremstillingen er allikevel i første rekke egnet for en leser som er kommet nokså mye lenger enn til et innføringskurs i teorien for statistikkproduksjon.

1D. For å gjøre notatet mer egnet til oppslag er det tatt inn et stikkordregister og et eksempelregister bakerst.

De nordiske statistiske foreninger har under trykking en flerspråklig ordliste, som tar utgangspunkt i Kendall og Bucklands velkjente statistiske ordbok. Karakteristisk nok for oppsplittinger av faget statistikk har ordlisten og dette notatet få berøringspunkter. Derimot vil interesserte kunne forfølge mange av ideene her f.eks. hos Lazarsfeld og Menzel (1961) og i skriftene nevnt i punkt 1A.

1E. Dette notatet er en revidert versjon av et tidligere diskusjonsinnlegg (Hoem, 1974), som med dette kan betraktes som foreldet. Under bearbeidingen av det nye notatet har jeg hatt stor nytte av kommentarer fra fagfellene Thormod Andreassen, Erik Botheim, Eivind Hoffmann og Gisle Skancke i Statistisk Sentralbyrå, og fra Bo Sundgren og Svante Öberg i Statistiska Centralbyrån. Jeg vil gjerne rette en særlig takk til Dag Gundersen, Norsk leksikografisk institutt, Universitetet i Oslo, for verdifulle synspunkter på språkbruken.

2. Betegnelsene "bestand", "variabel", "kjennemerke", osv.

2A. Når en statistisk undersøkelse gjennomføres av en institusjon som Statistisk Sentralbyrå, vil interessen som regel være konsentrert om en endelig samling (en populasjon, et univers) av elementer (observasjonsenheter, telleenheter, objekter), f.eks. bedrifter, husholdninger, eller personer. Denne samlingen vil vi omtale som undersøkelsesbestanden. (I eldre statistisk litteratur omtales ofte en bestand som en "masse". I nyere litteratur unngår en den siste betegnelsen, formodentlig fordi ordet gir assosiasjoner til massebegrepet i fysikken.)

Underbestanden kan være en samling av aktiviteter eller hendelser, f.eks. tyverier eller trafikkulykker, og det vi har å si, er ment å dekke også slike bestander. [Hoffmann (1973, kap. III) diskuterer slike muligheter.] For enkelhets skyld skal vi imidlertid ordlegge oss som om elementene er fysiske eller juridiske personer eller objekter.

2B. For hvert element (medlem) i bestanden tenker vi oss utført en måling, slik at en får registrert en måleverdi. (En registrerer bedriftens brannforsikringsverdi, intervjuobjektets svar på et spørsmål, husholdningens månedlige husleie, osv.). I statistisk teori representerer vi dette ved en statistisk variabel definert over bestanden. Hvis vi kaller variabelen for v, vil det da til hvert medlem av bestanden svare en verdi av v, slik at medlem nr. k har variabelverdien v(k). Hvis bestanden har N medlemmer, skal vi kalle $V = \{v(1), v(2), \ldots, v(N)\}$ for datamengden til v. V er naturligvis endelig. I mange tilfeller vil elementene i V være en samling punkter innenfor et nærmere angitt område V' på tallinjen. Det kan da være nyttig å ha betegnelsen verdimengde på V'. Som eksempel kan en tenke seg at elementene er norske skattytere i et gitt år og at v er skattbar inntekt i året. Mens V er samlingen av alle skattbare inntekter som forekommer det året, kan det være nyttig å bruke $V' = [0, \infty >$, dvs. å "tillate at alle ikkenegative skattbare inntekter i prinsippet kan forekomme".

Etter modell fra matematisk statistikk og andre fag har vi her knyttet variabelbegreper i statistikken helt til matematikkens funksjonsbegrep[1)]. For å kunne arbeide med en variabel,

1) En statistisk variabel skal være en reell avbildning av $\{1, 2, \ldots, N\}$. Denne terminologien faller ikke helt sammen med språkbruken i elementer matematikk. Hvis f er en reeell funksjon over den reelle tallinjen $\mathbb{R}$, er en vant til å skrive f.eks. $y = f(x)$ og omtale både x og y som (matematiske) variable. Dette faller naturlig i matematikk, men i statistikk ville det tilsvare at en omtalte bl.a. nummeret k på et medlem av undersøkelsesbestanden som en variabel, noe som ville føles helt fremmed. Videre er det $\bigcup_{2=1}^{N}\{v(k)\}$ som kalles verdimengde i matematikken, og dette kan være en ekte delmengde av V'.

er det således nødvendig å kjenne dens definisjonsområde (her: undersøkelsesbestanden), dens verdimengde (eller dens potensielle verdimengde), og de tilordningsprinsipper som knytter disse sammen, dvs. de prinsipper målingen utføres etter. Det er viktig å holde disse tre aspektene fra hverandre. Tilsammen definerer de variabelen. Hvordan de så symboliseres, enten det er med tall, bokstaver, eller på annen måte, er en sentral opplysning for datadokumentasjonen, men den hører ikke med i variabeldefinisjonen annet enn helt formelt.

2C. De symbolene som brukes til å representere verdiene til en variabel, vil vi kalle for dens koder. Hvert symbol omtales altså som én kode, og en oppstilling over dem kalles en "kodeliste". For variabeldefinisjonen er det likegyldig hvilken kodeliste som benyttes, så lenge den er hensiktsmessig og kjent. Fra definisjonens synspunkt vil enhver éntydig transformasjon av kodelisten være like god.

Dersom en endrer på definisjonen av bestanden, eller på den potensielle verdimengde (f.eks. ved å redusere antall mulige koder), går en imidlertid i prinsippet over fra én variabel til en annen, la oss si fra v_1 til v_2. Imidlertid vil en ofte beholde samme betegnelse på de to variablene, slik at v_1 og v_2 får samme navn. For eksempel snakker vi om "ekteskapelig status" både når kodelisten er {ugift, gift, separert, skilt, enke/enkemann} og når den er {ugift, gift, før gift}. Selv om navnet er det samme, er det allikevel viktig at en holder klart for seg at det fortsatt dreier seg om to variable, ikke bare én. En må ikke blande sammen betegnelse og begrep. Tenk bare på hviken forskjell det er mellom variable som betegnes "arbeidsløshet" eller "flytting" når de er definert henholdsvis på individnivået (v_1) og i statistikk for kommuner (v_2).

Andre ganger er forskjellen mellom to variable (etter denne definisjonen) imidlertid lite betydningsfull, f.eks. hvis det bare er undersøkelsesbestandene som ikke faller sammen. Bortsett fra dateringen er det bare en uinteressant og formell forskjell mellom på den ene side variabelen "ekteskapelig status" definert med en gitt kodeliste for norske menn pr. 31/12 1974, og på den andre side variabelen med samme navn, kodeliste og bestand pr. 31/12 1975, fordi registreringspraksisen (bruken av kodelisten) neppe har endret seg i mellomtiden. Vi kommer tilbake til momenter som dette i avsnitt 4 nedenfor.

Betegnelser som "kjennemerke" og "kjennetegn" er synonyme med "statistisk variabel". Tilsvarende brukes "kjennemerkeverdi" synonymt med "variabelverdi", osv.

I avsnittene nedenfor skal vi utdype vår omtale av disse begrepene.

3. Betegnelsene "klassifisering", gruppering", osv.

3A. La oss tenke oss datamengden V delt opp i et antall disjunkte delmengder $V_1,\ldots,V_j$. (Disse kan eksempelvis være avledet av en oppdeling av V' i delintervaller.) Dette gir opphav til en klassifisering av elementene i bestanden ved at element nr. k sies å tilhøre klasse j dersom $v(k)\varepsilon\checkmark j$. Vi har her nummerert klassene fra 1 til J. I mange sammenhenger, f.eks. i offisielle standarder, vil en utvikle en systematisk klassebetegnelse ved hjelp av sifre og eventuelt andre symboler. En slik betegnelse kalles klassekoden. Hvis hver klasse får et treffende navn, genereres det en tilsvarende klassebetegnelse for elementene i klassen. (Eksempel: betegnelsene på kommunetypene; Rideng 1974.)

En klassifisering av elementene i en bestand gir umiddelbart opphav til en ny variabel, avledet av den som ligger til grunn for klassifiseringen. Hvis vi kaller den opprinnelige variabelen for v, får vi avledet en klassifiseringsvariabel v^* ved å sette $v^*(k) = j$ hvis element nr. k tilhører klasse nr. j, slik at

$$v^*(k) = j \text{ hvis og bare hvis } v(k)\varepsilon V_j.$$

Igjen vil v og v^* ofte få samme navn. Ikke desto mindre dreier det seg om to ulike variable (bortsett fra i det trivielle tilfelle da hver V_j bare har ett element). (Eksempel: ett- og femårige aldersklasser.)

3B. La oss tenke oss at det er definert et avstandsmål d over bestanden, slik at avstanden mellom elementene k og ℓ er $d(k,\ell)$. Elementene i bestanden kan da grupperes slik at elementene kommer i samme gruppe dersom $d(k,\ell)$ er liten. Hvis en krever $d(k,\ell) = 0$ for at elementene k og ℓ skal komme i samme gruppe, får en i prinsippet ingen problemer med å angi grensene mellom ulike grupper. Dersom en tillater d- avstanden mellom elementene i en gruppe å bli positiv, må det fastlegges kriterier for hvor stor d- avstanden skal kunne bli innen en gruppe, og for hvor grensene mellom gruppene skal trekkes.

3C. Formelt sett er en gruppering og en klassifisering akkurat det samme, i den forstand at enhver klassifisering er en gruppering og enhver gruppering er en klassifisering. La nemlig bestanden være klassifisert etter variabelen v, og la

$$d(k,\ell) = |j_k - j_\ell|,$$

der j_k og j_ℓ er numrene på de klassene elementene k og ℓ tilhører. Da er klassifiseringen en gruppering av elementene generert av kravet $d = 0$.

La det omvendt foreligge en gruppering av bestanden, og la en variabel v være definert ved at $v(k) = g_k$, der g_k er nummerert på den gruppen elementet k tilhører. Da er grupperingen samtidig en klassifisering etter v.

Det er allikevel nyttig å ha begge betegnelsene, både gruppering og klassifisering. Betegnelsen gruppering leder nemlig tanken i retning av at det skal foreligge en inndeling av bestanden i grupper som i en meningsfylt forstand er homogene. En har bruk for en terminologi som kan ta med dette substansielle aspektet, selv om det formelt sett er overflødig.

3D. Ovenfor har vi definert en klassifisering med utgangspunkt i en enkelt variabel. Formelt sett er dette tilstrekkelig, siden en med utgangspunkt i flere variable definert over den endelige bestanden alltid en-tydig kan definere en enkelt avledet variabel som alene representerer dem alle. Det er allikevel nyttig å kunne omtale en fler-dimensjonal klassifisering.

La variablene $v_1, v_2, \ldots, v_n$ være definert over bestanden, og la $\underset{\sim}{y} = (v_1, \ldots, v_n)$. Datamengden til $\underset{\sim}{y}$ er $\underset{\sim}{Y} = \{\underset{\sim}{y}(k): k = 1,2, \ldots, N\}$. Det foreligger en fler-dimensjonal klassifisering av bestanden tilsvarende hver oppdeling av $\underset{\sim}{Y}$ i disjunkte delmengder $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_J$. Den fler-dimensjonale klassifiseringen kan kalles en kryss-klassifisering etter $v_1, v_2, \ldots, v_n$ hvis den er fremkommet ved at datamengdene $V_1, \ldots, V_n$ til de enkelte variable er delt opp i disjunkte undermengder V_{ij} ($i = 1,2, \ldots, n$; $j = 1,2, \ldots, J_i$) og hver $\underset{\sim}{Y}_j$ er fremkommet som et kryssprodukt av typen

$$\underset{\sim}{Y}_j = V_{1j_1} \times V_{2j_2} \times \ldots \times V_{nj_n}$$

$$= \{\underset{\sim}{y}(k): v_i(k) \in V_{ij_i} \text{ for } i = 1,2, \ldots, n\}.$$

3E. I praksis kan en gruppering fremkomme f.eks. ved at klasser i en opprinnelig, mer finmasket klassifisering (eventuelt en kryssklassifisering) slås sammen til større grupper.

En kan også få en gruppering av personer med utgangspunkt i en opprinnelig klassifisering av sosiale systemer som disse personene tilhører. For eksempel kan det foreligge en klassifisering av bedrifter, la oss si i næringer, og dette gir opphav til en tilsvarende gruppering av arbeidstakere etter næring.

Stadig grovere grupperinger av elementene i en gitt bestand kan gi opphav til nye bestander av elementer på stadig høyere aggregeringsnivåer når det som representerer en gruppe elementer på ett nivå, oppfattes som et element på et høyere nivå. Eksempelvis kan personer grupperes i husholdninger, husholdninger i kommunale befolkninger, disse i fylkesbefolkninger, osv. Variable kan være definert over bestander på hvert av disse nivåene. Noen ganger kan en variabel på et mer aggregert nivå fremkomme ved summering over variabelverdier fra et lavere nivå. Slik fremkommer f.eks. husholdningsinntekt, samlet personlig inntekt i en kommune, osv. Andre slike transformasjoner er gjennomsnitt, ulike spredningsmål, empiriske regresjonskoeffisienter, osv. Variable fremkommet på dette viset, kan omtales som aggregatvariable. Det er igjen viktig å være oppmerksom på at de variable en arbeider med "før" og "etter" en slik aggregering, kan ha samme betegnelse til tross for at det dreier seg om ulike variable.

Andre ganger defineres en variabel direkte for elementene i aggregater uten at den selv fremkommer ved summering over variabelverdier. Antall folkeskoler i en kommune er et eksempel. Slike variable omtales ofte som globale variable (i relasjon til elementene i aggregatene).

En kan finne en nærmere diskusjon av momenter som disse hos Lazarsfeld og Menzel (1961).

4. Tidsdimensjonen.

4A. I det ovenstående har vi sett bort fra ett aspekt ved statistiske undersøkelser, nemlig det at de alltid vil referere seg til et tidspunkt eller en periode. Vi skal nå ta dette med i betraktning.

Hvis en følger en gruppe av statistiske undersøkelsesenheter over tid, vil medlemsstokken kunne endre seg ved at noen forlater gruppen (bedrifter nedlegges, personer skifter egenskaper eller dør) mens andre kan tre inn i den. En slik gruppe av undersøkelsesenheter der det kan være tilgang og avgang av elementer, kan vi kalle en dynamisk bestand.

4B. Siden elementene i en slik bestand kan skifte egenskaper, er det ofte problematisk å avgjøre hva som skal få betraktes som "samme element" av bestanden på to ulike tidspunkter. Hvis elementene er personer, er vel dette nokså likeframt, men det er ikke alltid så lett å bestemme seg for hva som skal få anses som samme husholdning over tid, for eksempel. Øberg (1972, avsnitt 3.3) har gitt en grei oversikt over slike spørsmål, og han innfører et interessant skille mellom identiske og gjenidentiske enheter. Slik disse begrepene beskrives av Hoffmann (1973, side 8), er medlemmer av en dynamisk bestand identiske bare hvis de på alle tenkelige variable har de samme verdier, også på tidsvariabelen. Et element er derfor bare identisk med seg selv. For å få sammenheng ved bevegelse langs tidsvariabelen innføres så begrepet gjenidentitet, som svarer til det vi i dagligtale mener når vi sier at ting vi har observert på ett tidspunkt er den samme som den vi observerer på et annet tidspunkt. En gitt person i skattemanntallet pr. 31/12 1972 er identisk med samme person i Det sentrale personregister pr. samme dato, og det er slike sammenknytninger mellom ulike datakilder som gjør identitetsbegrepet interessant. Den ovennevnte personen er "bare" gjenidentisk med samme person pr. 26/2 1973. Mer relevant er dette skillet naturligvis ved grupper av personer eller ved institusjoner, der identifiserende variable kan skifte verdi over tid. (Eksempel: Familier med samme familienummer i Byråets familieregister på to ulike tidspunkter kan ha helt ulik sammensetning. Det eneste som trenger å være felles, er den referansepersonen som familienummeret refererer seg til.) En må ha et sett av regler for å bestemme hvor store endringer som skal tillates i hvilke variable før en slutter å snakke om gjenidentiske elementer.

4C. Spørsmål vedrørende gjenidentitet er betydningsfulle for en diskusjon av hva identifikasjonsnumre for elementer i en dynamisk bestand eventuelt skal inneholde av informasjon. I Skandinavia kan vi jo lese en persons kjønn og fødselsdato rett ut av personnummeret og fylket

ut av kommunenummeret. Men hvordan skal en bygge opp identifikasjonsnumre for bedrifter (i bedrifts- og foretaksregisteret) eller helseinstitusjoner (i Det økonomisk-medisinske informasjonssystem)? Statistisk Sentralbyrås holdning er nå vanligvis at det er uheldig å ta inn informasjon om enheten i identifikasjonsnummeret, og en begrunnelse er at enheten vil skifte registrert identitet når en innebygget kjennetegn skifter verdi. Hvis kjennetegnene i identifikasjonsnummeret var sentrale nok, kunne det imidlertid være fordelaktig med et slikt automatisk identitetsskifte. Om en legger opp et system uten slik automatikk, kan gjenidentitet lett bli knyttet til mindre vesentlige (kanskje vilkårlige) kriterier, som f.eks. eksakt lokalisering. [Poengene i dette avsnittet skyldes Eivind Hoffmann.]

4D. La oss imidlertid nå forutsette at en har løst de praktiske problemene knyttet til definisjonen av gjenidentitet, og la elementene i en dynamisk bestand entydig være tildelt permanente identifikasjonsnumre. Enkelte variable vil da være definert over de elementene som er medlemmer av bestanden på et gitt tidspunkt, nemlig tilstandsvariable og beholdningsvariable. Hvis samme formelle definisjon brukes til å definere slike variable over en dynamisk bestand for flere tidspunkter, slik at en får frem en serie variable med samsvarende definisjoner, kan en si at variabeldefinisjonene er samordnet over tid og at de tilsvarende datamengdene er sammenliknbare over tid.

Det kan her være tale om atskilte tidspunkter t_1, t_2, ..., eller om observasjon som i prinsippet er kontinuerlig over et gitt tidsrom (slik som ved et løpende personregister), eller en annen tidsmengde. Vi vil i alle fall kalle samlingen av tidspunkter som kommer i betraktning, for T. For hvert tidspunkt $t\varepsilon T$ tenker vi oss da angitt en variabel v_t definert over elementene i bestanden. Vi antar at $\{v_t : t\varepsilon T\}$ er en samling av ulike variable med samsvarende definisjoner. I dagligtalen vil en gjerne omtale disse som "samme variabel v målt på ulike tidspunkter t i T", og det er ingen grunn til at man skulle forsøke å "forby" en slik språkbruk i fagspråket heller. Den er enkel, presis, og fullt tillatt etter den matematiske definisjonen av begrepet variabel. [Fra matematisk synspunkt vil det her dreie seg om en funksjon v med verdimengde $\{v_t(k) : t\varepsilon T, k\varepsilon B_t\}$, der B_t er samlingen av elementer i bestanden på tidspunkt t.]

4E. Tilsvarende betraktninger gjelder for strømningsvariable, men det er vel nødvendig med en viss utdypning. [Vi bygger her delvis på Haavelmo, 1951.] Vi skal sondre mellom to typer av strømningsvariable, nemlig

(i) slike som refererer seg til et gitt tidsrom, f.eks. en persons skattbare inntekt i et gitt år og en bedrifts investeringer i et gitt kvartal, og

(ii) slike som refererer seg til et gitt tidspunkt, som f.eks. en persons inntekt på årsbasis regnet etter inntekten pr. 1/7 1973.

Den første typen vil vi kalle periodiserte strømningsvariable. Den andre typen vil vi kalle intensitetsvariable. La oss betegne en strømningsvariabel definert for perioden $[t, t+s\rangle$ méd $v(t,s)$. For å konkretisere, kan vi tenke på en persons inntekt i dette tidsrommet. Hans inntekt pr. tidsenhet i $[t,t+s)$ er da $\bar{v}(t,s) = v(t,s)/s$, og hans inntekt pr. tidsenhet på tidspunkt t er

$$w(t) = \lim_{s \to 0} v(t,s)/s$$

hvis denne grensen eksisterer. Mens v og $\bar{v}$ er periodiserte strømningsvariable, er w en intensitetsvariabel.

Språkbruken vedrørende samordning og sammenliknbarhet over tid osv. kan nå overføres direkte til intensitetsvariable. For periodiserte strømningsvariable går det like greit når en bare skifter ut ord som "tidspunkt" med ord som "periode" overalt. Eksempelvis blir det tale om "samme periodiserte strømningsvariabel v målt for ulike perioder t i T", der T nå blir en samling

tidsintervaller. Hvis disse utgjør ikke-overlappende delintervaller av en lengere periode, er det naturlig å nummerere dem fortløpende og la t betegne periodenummeret eller noe liknende.

For en strømningsvariabel definert for en gitt periode blir definisjonsområdet vanligvis samlingen av de elementer som er medlemmer av den dynamiske bestanden en eller annen gang i løpet av perioden.

4F. Det er velkjent at en kan ha målingsproblemer knyttet til "vanlige" statistiske variable, altså problemer med påliteligheten av de innsamlede oppgavene. La oss nevne at slike problemer også kan oppstå i tilknytning til måling av tid, særlig til fiksering av nøyaktige tidspunkter. Ved analyse av individuelle flyttedata fra Det sentrale personregister, må en eksempelvis ha klart for seg forskjellen mellom faktisk flyttedato, oppgitt flyttedato (som angis av flytteren på flyttemeldingen), og registrert flyttedato (som skal være den dagen folkeregisteret mottar flyttemeldingen). En regner med at personregisterets opplysning om den siste er meget pålitelig (folkeregistere antas stort sett å oppgi den dato de faktisk mottar flyttemeldingen), men forskjellen mellom registrert og faktisk flyttedato gjør relevansen av registrert flyttedato for analyser av de virkelige flyttingene noe problematisk. Oppgitt flyttedato er mer relevant som indikator på faktisk flyttedato, men vi vet at den dato som oppgis på mange flyttemeldinger i praksis ikke er den dato flyttingen faktisk fant sted. Det foreligger forøvrig ikke noen presis definisjon av faktisk flyttedato - dette begrepet er vanskelig å fange.

[I sosiologisk literatur omtales "pålitelighet" og "relevans" ofte som "reliabilitet" og "validitet" etter engelskspråklig mønster. Se f.eks. Hellevik, 1971.]

5. Mer om betegnelsene "sammenliknbarhet", "samordning", osv.

5A. I avsnitt 4D ovenfor omtalte vi bl.a. variabeldefinisjoner som er samordnet over tid og datamengder som er sammenliknbare over tid. Tilsvarende skal vi si at det foreligger samordning innen et statistikkområde dersom en har samsvar mellom begreper, enheter og definisjoner for all statistikk innen området. (Analogt for deler av et statistikkområde.) Dette innebærer f.eks. at en har samsvar mellom periodiserte strømningsvariable målt på kvartalbasis og korresponderende variable målt på årsbasis. Kvartals- og årsstatistikkene blir da sammenliknbare. Samordnet statistikkproduksjon gir altså opphav til sammenliknbare data (sammenliknbar statistikk). Der hvor samsvaret er ivaretatt når det gjennomføres aggregering (over tid eller over elementene i en bestand eller begge deler), kan en snakke om vertikal samordning.

Analogt sier en at en har sammenliknbarhet mellom statistikken fra to statistikkområder (eller deler av dem) hvis det foreligger tilsvarende samsvar områdene imellom, og det sies å foreligge horisontal samordning mellom områdene. Skal en eksempelvis ha sammenliknbarhet mellom bedriftsopplysninger i næringsstatistikk og lønnsstatistikk, må bl.a. definisjonen av"bedrift" være den samme i begge områder.

5B. Sammenlikning mellom datamengder (tabeller, statistikk) kan naturligvis foregå på ulike aggregeringsnivåer. Noen ganger kan en sammenlikne individuelle variabelverdier for medlemmene i to bestander. Andre ganger foretar en sammenlikninger mellom summer eller gjennomsnitt av variabelverdier (aggregattall) for grupper av elementer, eller mellom totalsummer for to eller flere bestander. Ordet "sammenliknbarhet" brukes til dels upresist om alle slags muligheter til å holde ulike tallmaterialer opp mot hverandre. Hvilke datamengder som godtas som sammenliknbare, avhenger av de kvalitetskrav en har. Ofte kan en akseptere mindre forskjeller i variabeldefinisjonen.

I neste kapittel skal vi ta for oss to begreper som er relevante når det forligger sammenliknbarhet.

6. Sammenkobling og sammenbinding

6A. Verdier for to ulike variable, la oss si v_1 og v_2, definert over én og samme bestand, vil kunne foreligge i to forskjellige datakilder. Når disse variabelverdiene hentes ut fra de to datakildene og føres sammen på ett datamedium, sier vi at de to materialene sammenkobles. Skal sammenkobling kunne finne sted, må en altså ha observasjoner på det som i den gitte situasjon betraktes som elementnivået (individnivået). Det er også en forutsetning at definisjonen av bestandens enheter og omfang er felles for de to datakildene. (Eksempel: Levekårsundersøkelsen og Forbruksundersøkelsen 1973.) Hvis den bestand B_1 som er variabel v_1 er definert over, utgjør en ekte del av den bestand B_2 som en annen variabel v_2 er definert over, kan en naturligvis sammenkoble datamengdene for v_1 og restriksjonen av v_2 til B_2.

[I Statistisk Sentralbyrå har en gjerne omtalt sammenkobling som samkobling. Gundersen (1974) fraråder den siste betegnelsen og gir gode argumenter for dette.]

6B. La oss nå tenke oss at det foreligger en klassifisering K av elementene i en bestand, og at en kjenner fordelingen av elementene i hver klasse i K etter to variable, w_1 og w_2, uten at en har tilgjengelig individobservasjonene av verdiene for begge disse variable.

En kan da lage aggregattabeller kryssgruppert etter K og w_1, og separat etter K og w_2, men ikke aggregattabeller kryssgruppert etter w_1 og w_2 (eller K,w_1 og w_2 simultant). Vi vil foreslå at en i et slikt tilfelle omtaler K som et bindeledd mellom w_1 og w_2. Dette er et forsøk på oversettelse av det svenske ordet "länk" for samme begrep. (Se Fastbom, 1973, punkt 2.) Vi foreslår at w_1 og w_2 sies å være bundet sammen ved hjelp av K.

Sammenbundne materialet inviterer til visse typer av feilslutninger som det er grunn til å advare mot. Selv om en får vite at 33 prosent av norske 19-åringer var under utdanning i 1971, og at 2 prosent av norske 19-åringer var siktet for forbrytelser det året, kan en ikke slutte noe eksakt om hvor stor del av 19-åringene under utdanning var siktet. Hellevik (1971, side 276 ff) gir en interessant diskusjon av feilslutninger av denne typen.

7. Etterord

7A. Fremstillingen ovenfor tar ikke på noen måte sikte på å være fullstendig. Nødvendigheten av en ytterligere utdypning av mange av de momentene som vi tar opp, er åpenbare. Det finnes også mange begreper og betegnelser vi i det hele tatt ikke er kommet inn på her. Vi skal avslutte notatet med noen refleksjoner over visse betegnelser som er sentrale nok for notatets egen problemstilling, men der jeg ikke har følt det aktuelle behovet for (eller ønskeligheten av) en fastlegging av språkbruken like klart som andre steder.

7B. Det finnes en rekke betegnelser på de talloppgaver (tallmaterialer) som fremkommer på de ulike stadier under bearbeidelsen av en statistisk undersøkelse. En snakker om data, rådata, grunndata, primærmateriale, grunnmateriale, listing, statistikk, osv., uten at det ennå synes å foreligge noen rimelig grad av enighet om hva en skal legge i de forskjellige ordene. Naturligvis har en visse opplagte presiseringsretninger: rådata forekommer tidlig i produksjonsprosessen; statistikk er trolig vel organisert numerisk informasjon presentert som tabeller, diagrammer, o.l. Bildet kompliseres imidlertid av at det som for noen er et sluttprodukt (f.eks. en tabellpublikasjon fra Byrået), for andre utgjør et utgangspunkt (et råmateriale) for videre bearbeiding. Det er mulig at det mest fruktbare er å dele opp produksjonsprosessen i delprosesser og knytte terminologi og teoridannelse til hver enkelt av disse. En kan da kanskje ta i bruk betegnelsen fra svensk datamaskinterminologi, som inndata og utdata. En slik oppdeling i delprosessen fordrer imidlertid en forutgående avklaring av begreper som datakilde/statistikkilde og statistikkområde, noe vi heller ikke har gitt oss ut på i dette notatet. Selv bruker jeg "data" som fellesbetegnelse på et sett sammenhørende talloppgaver uansett hvor de opptrer i statistikkproduksjonen. Forøvrig overlates det til den senere utvikling å bringe fram en ytterligere avklaring på dette punkt.

Stikkordregister med sidehenvisning

Eksempelregister med sidehenvisning

Referanser

[1] Datasamordningskommittén (1974): "Integreral informationssystem för samhällsplaneringen". DASK-rapport utarbetet av en arbetsgrupp vid SCK under medverkan av personal vid statskontoret. To bind.

[2] Fastbom, L. (1973): "Några kommentarer avseende Eivind Hoffmann: Linkages within the System of Social and Demographic Statistics: A Framework for Discussion." Nordiska utskottet för socio-demografisk statistik, Kommentarer 1973-02-23.

[3] Forsberg, Robert (1972): "Standard avseende vissa statistiska grunnbegrepp." Statistiska Centralbyrån, Stockholm, SCB metodinformation Nr. 72-4.

[4] Gundersen, Dag (1974): Brev til JMH datert 13. juni 1974.

[5] Hellevik, Ottar (1971): "Forskningsmetode i sosiologi og statsvitenskap." Universitetsforlaget, Oslo etc.

[6] Hoem, Jan M. (1974): "Noen begreper vedrørende statistikk og primærdata. En diskusjon av terminologi." Statistisk Sentralbyrå, ANO IO 74/6, side 3-15.

[7] Hoffmann, Eivind (1973): "Noen informasjonsteoretiske grunnbegreper og deres relevans for arbeidet med et system for sosiodemografisk statistikk." Notat EH/EH, 24/12-73. Gjengitt på side 16 til 24 i Metodehefte 10, Statistisk Sentralbyrå, ANO IO 74/6.

[8] Haavelmo, Trygve (1951): "Noen alminnelige merknader om tidsrekker og tidsrekkeanalyse." Avsnitt III.7 i "Innføring i teoretisk statistikk. Del II" av T. Haavelmo, redigert av Arne Amundsen; Memorandum av 15/2 1951 fra Sosialøkonomisk institutt, Universitetet i Oslo. Også gjengitt på side 54-59 i "Utdrag av forelesninger i teoretisk statistikk fra kompendier av Ragnar Frisch og Trygve Haavelmo", redigert av Herdis Thorén Amundsen; Universitetsforlaget, 1974.

[9] Kendall, Maurice G. and William R. Buckland: "A Dictionary of Statistical Terms". Oliver and Boyd, for the International Statistical Institute. Flere utgaver.

[10] Lazarsfeld, Paul F. og Herbert Menzel (1961): "On the relation between individual and collective properties." Side 422-440 i A. Etzioni (red.): "Complex Organizations." Holt, Renehart & Winston, New York.

[11] Rideng, Arne (1974): "Klassifisering av kommunene i Norge 1974." Statistisk Sentralbyrå, Artikkel nr. 67.

[12] Sundgren, Bo (1974): "The infological approach to data bases." Statistiska Centralbyrån, ADB-information M74:4, 1974-06-07.

[13] Sundgren, Bo og Svante Öberg (1974): Kommentarer til JMH datert 1974-06-20.

[14] Öberg, S. (1972): "Ett system för socio-demografisk statistik (SSDS), grunnläggande begrepp." Statistiska Centralbyrån, Stockholm, Rapport 1970.09.18.