Variabelroller för mikro- och makrodata

1 Syften och inledning

1.1 Syften

Detta dokument, som är skrivet för SCB, har två huvudsyften. Det ena huvudsyftet är att beskriva en struktur för variabler på likartat sätt som tidigare har gjorts för objekt, se Objekt i statistikproduktionen, roller och relationer. Situationen är dock mer komplex för variabler än för objekt. Rollerna är flera och relationerna är mindre tydliga.

Det andra huvudsyftet är terminologi. Det underlättar om alla på SCB använder samma terminologi. Kvalitetshandboken Kvalitet för den officiella statistiken – en handbok är utgångspunkten här, men flera benämningar behövs. I många fall finns sedan gammalt synonymer. Det förekommer även termer som bör undvikas, t.ex. för att de har flera betydelser. På senare tid har internationella beskrivningsmodeller tillkommit.

Att synliggöra kopplingar till metadata är ytterligare ett syfte, kopplingar som också påverkar strukturen för rollerna.

1.2 Inledning

Mikrodata används ofta som benämning för statistiska data avseende variabler knutna till objekt, t.ex. kan en generell variabel som ålder knytas till en objekttyp som person. En datamängd på mikronivå – ett observationsregister på mikronivå – kan, schematiskt, ses som en eller flera tabeller där rader och kolumner innehåller objekt respektive variabler.

Ett *statistikvärde* är en skattning av en statistisk storhet. En *statistisk storhet* bildas genom att ett statistiskt mått sammanfattar enskilda värden för variabler hos objekten i en redovisningsgrupp för en eller flera referenstider. Både statistiska storheter och statistikvärden är makrodata, aggregerade mikrodata. Fokus för makrodata ligger på statistikvärden.

De statistiska storheter som den statistikansvariga myndigheten (SAM) – här i första hand SCB – bestämmer ska skattas kallas *målstorheter*. Dessa byggs upp av målvariabler, målobjekt, målpopulationer etc. Prefixen *intresse* respektive *mål* används, då det behövs, för att särskilja önskemål och val. När valet är givet eller inga missförstånd uppstår, kan prefixet mål utelämnas. Ofta sägs population i stället för målpopulation, som exempel.

Det är vanligt att tabellceller och marginaler i en *statistisk tabell* motsvarar statistiska storheter och att nedbrytningarna visar indelningar av populationen i redovisningsgrupper eller anger referenstider (ofta båda). I en korstabell med två indelningar, säg, visar marginalerna var och en av indelningarna samt hela populationen. En statistisk tabell kan också visa osäkerhetsintervall. En tabellplan sätts upp i designen, och i statistikproduktionen fylls varje tabell, så långt möjligt, med statistikvärden.

Vissa undersökningar samlar in och bearbetar makrodata, för att därefter redovisa makrodata som är delvis andra än de insamlade. Även mikrodata kan samlas in och ingå i bearbetningarna.

Variabler kan indelas i olika typer efter hur de mäts. Kvalitativa variabler mäts på antingen nominal- eller ordinalskala, medan kvantitativa variabler mäts på antingen intervall- eller kvotskala. Se vidare Frågor och svar – om frågekonstruktion i enkät och intervjuundersökningar (scb.se), t.ex. s. 81.

En variabels värdemängd består av de värden som variabeln kan anta. Värdemängdens karaktär bestäms av variabeltypen. Vissa värdemängder utgår från klassifikationer, som ofta är hierarkiskt ordnade med internationell samordning åtminstone på de översta nivåerna. Värdemängder ingår i innehållsmetadata (avsnitt 2).

En och samma variabel kan, i likhet med objekt, ha flera roller. Det är vanligt förekommande.

2 Fyra kategorier metadata – omfattar variabelroller

2.1 Metadata i fyra kategorier

SCB har en indelning av metadata i fyra kategorier: innehållsmetadata, styrande metadata, processdata och referensmetadata.

- Innehållsmetadata omfattar bland annat information om objekt, populationer och variabler.
- Styrande metadata omfattar bland annat regler i statistikproduktionsprocessen, t.ex. för påminnelser och för granskning på mikro- och makronivå.
- Processdata är utfallsdata, t.ex. information om inflöde och om genomförda imputeringar. Processdata används i styrande metadata och vid utvärdering, i det senare fallet ofta på aggregerad nivå.
- Referensmetadata ingår i dokumentation internt och externt, t.ex. en sammanställning av uppgiftslämnarkontakter eller en kvalitetsdeklaration.

De fyra kategorierna utesluter inte varandra. Referensmetadata består av utvalda metadata från de andra kategorierna. Processdata kan bli styrande metadata i ett senare steg, t.ex. i en adaptiv design.

2.2 Variabelroller och metadata

Variabler har roller som det är praktiskt att strukturera och gruppera. Ett sätt är att knyta variabelroller till metadatakategorierna, det vill säga till innehållsmetadata, styrande metadata och processdata. Denna struktur presenteras i avsnitt 3-5 nedan. Referensmetadata som hämtar metadata från de övriga kategorierna har inte egna variabelroller.

I många fall har en och samma variabel flera roller, t.ex. både en innehållsrelaterad och en styrande roll.

3 Variabelroller knutna till innehållsmetadata

I den här gruppen avser variablerna i första hand egenskaper hos, eller kopplade till, de objekt som statistiken är avsedd att ge information om.

Fem framträdande variabelroller beskrivs nedan samlat för mikrodata och makrodata i avsnitt 3.1-3.5. De flesta rollerna finns på båda nivåerna, och det finns samband däremellan.

3.1 Intressevariabel

Intresse betecknar det som ingår i användarnas behov och önskemål. Det gäller inte bara variabler utan även statistiska storheter med t.ex. populationer. Rollen intressevariabel finns främst i dialoger med statistikanvändare inom ramen för den statistiska undersökningen och dess statistik samt i avvägningar under undersökningens utformning. När SCB är producent till en annan SAM är även den myndigheten involverad i dialoger. Intressevariabeln är en variabel på mikronivå, och på makronivå ingår den i statistiska intressestorheter.

Intressevariabel är en roll här, medan intresseobjekt inte är en roll bland de få objektrollerna. Intressen ingår i referensmetadata men knappast i IT-system.

3.2 Målvariabel

Mål betecknar det som statistikproducenten siktar på att åstadkomma. Dialoger förs med statistikanvändare om statistikens ändamål, kvalitet i övrigt, kostnader och uppgiftslämnarbörda. I avvägningarna ingår bland annat relevans och tillförlitlighet.

En målvariabel kan vara densamma som en intressevariabel, vilket är ett exempel på en variabel med flera roller. En målvariabel kan vara densamma som en observationsvariabel (se avsnitt 3.3 nedan). Målvariabeln kan alternativt vara modellbaserad eller härledd, i allmänhet med osäkerhet, men härledningar utan osäkerhet förekommer.

Rollen målvariabel förekommer främst under den statistiska undersökningens utformning, bearbetningar och redovisning. Målvariabeln är en variabel på mikronivå, och på makronivå ingår den i statistiska målstorheter och därmed i statistikvärden.

3.3 Observationsvariabel

Observation betecknar det som samlas in på mikro- och makronivå, t.ex. via register, nya datakällor, direktinsamling eller från andra undersökningar. Vad som samlas in beror bland annat på vilken statistik som ska tas fram, vad som finns tillgängligt i administrativa data och andra datakällor samt vad som lämpligen kan ingå i en eventuell direktinsamling (särskilt för mikrodata). Rollen observationsvariabel finns främst under den statistiska undersökningens utformning, insamling och bearbetningar.

En statistisk tabell kan ses som ett observationsregister på makronivå. En observationsvariabel är en variabel på mikronivå. På makronivån ingår observationsvariabler i statistikvärden som samlas in. Ett observationsregister på mikro- eller makronivå innehåller observationsvariabler eller målvariabler eller båda.

3.4 Indelningsvariabel

För var och en av de tre rollerna intressevariabel, målvariabel och observationsvariabel finns stora likheter mellan mikro- och makronivåerna. Däremot gäller det för rollen *indelningsvariabel* (variabel för gruppindelning) att den utspelar sig i gränslandet mellan mikro- och makronivåerna. Variabelvärdena som används för gruppindelningarna finns på mikronivå, men behövs för att kunna framställa och på makronivå redovisa statistik för olika redovisningsgrupper. Ett enskilt statistikvärde avser en viss redovisningsgrupp.

3.5 Hjälpvariabel – innehållsrelaterad

Utöver de fyra uttalade rollerna i avsnitt 3.1-3.4 finns en mängd innehållsrelaterade variabelroller. Benämningen *hjälpinformation* är vanligt förekommande både för enskilda variabler och som samlingsbenämning. När variabelroller ska beskrivas är det naturligt att säga *hjälpvariabel*.

Många hjälpvariabler är variabler på mikronivå, men även aggregat på makronivå förekommer.

Hjälpvariabler kan också vara styrande variabler, se avsnitt 4.

4 Styrande variabel

Rollen *styrande variabel* är knuten till styrande metadata. Det finns två huvudtyper, där den första är mer renodlat styrande än den andra som involverar flera roller.

1. Variabler som är baserade på undersökningsdesignen och som bär information om statistikproduktionsprocessens genomförande, men som inte avser innehåll eller är processdata. Det kan t.ex. vara uppräkningstal, regler för vägval i produktionsflöden eller information som avser datalagring.
2. Variabler som också är innehållsrelaterade hjälpvariabler (avsnitt 3.5 ovan) eller som också är processvariabler (avsnitt 5 nedan). Det kan t.ex. vara variabler som ingår i populationsavgränsningar eller som är statuskoder.

Några exempel på variabler som är styrande är:

- *Identifierande variabel*, eller synonymt *identitetsvariabel*, på mikronivå används för att tillföra information till en datamängd och för att integrera flera datamängder.
- Information om objekt används i ramförfarande på mikronivå med avgränsningar av ramelement och eventuellt med ett urvalsförfarande, speciellt stratifiering och allokering, i många fall med ett storleksmått. Även vid insamling på makronivå behövs ramförfaranden med hjälpinformation, där populationer avgränsas.
- Variabel som används i bearbetningar på mikro- eller makronivå (benämningar som *transformation* förekommer för bearbetningar, även *process* med input och output). Det kan vara härledningar och modellbaserade beräkningar av målvariabler från observationsvariabler. Det kan vara ett skattningsförfarande från mikro- till makronivå med hjälpinformation, t.ex. vid kalibrering där hjälpinformation kan vara mikrodata eller på aggregerad nivå. Vid säsongrensning används en *kalendervariabel* för transformation på makronivå.
- Variabler som styr flaggningar, grupperingar och hanteringar i granskningsprocessen.

5 Processvariabel

Rollen *processvariabel* är knuten till processdata. Den är mer renodlad än rollen styrande variabel. Processvariabler bär information om resultat av processens genomförande, t.ex. statuskoder som speglar resultatet av kontaktförsök eller uppgifter som beskriver inflöde (exempelvis andel inkomna objekt eller vägda och ovägda bortfallsandelar).

Processvariabler kan användas för styrning av statistikproduktionsprocessen (t.ex. vägval beroende på inflöde), och de används vid utvärdering (t.ex. information från kontrollkodning, om hur kontaktförsök har lyckats och om träffsäkerhet i granskningsförfarandet). Användningen kan gälla den pågående omgången eller kommande omgångar (eller båda).

6 Benämningar i beskrivningsmodeller

De beskrivningsmodeller som tas upp nedan används i flera olika sammanhang vid SCB.

6.1 Benämningar i GSIM

GSIM (*Generic Statistical Information Model*) är en informationsmodell för statistikproduktion som består av fyra delar innehållande informationsobjekt: Verksamhet (*Business*), Innehåll (*Concept*), Datautbyte (*Exchange*) och Datastrukturer (*Structures*).

Variabler och det som relaterar till det finns i innehållsdelen och kan hierarkiskt beskrivas som att allt utgår från begrepp (”*concept*” vilket trots samma namn är en delmängd av *Concept* ovan). Genom att kombinera ett begrepp med en objekttyp får man en variabel, och genom att kombinera en variabel med hur den representeras får man en representerad variabel. Representationen kan bestå av kontinuerliga eller uppräknade värdemängder där de sistnämnda består av kodlistor som i en del fall är klassifikationer. När en representerad variabel används i ett specifikt dataset kallas den för en instansvariabel. Alla informationsobjekt har egenskaper vilka kan jämföras med attribut. Läs mer på Clickable GSIM - Clickable GSIM - UNECE Statswiki

SCB använder GSIM som informationsmodell i utvecklingen av processer och IT-stöd för statistikproduktionen.

6.2 Benämningar i XBRL

XBRL (*eXtensible business reporting language*) är en standard för att beskriva finansiell information. Taxonomier i XBRL utgör dels en detaljerad beskrivning av så kallade basbegrepp inom ett sakområde, dels en beskrivning av hur begrepp relaterar till varandra inom olika rapporter. En taxonomi har många olika skikt, men en del har en tydlig koppling till vad som skulle benämnas variabler. Dessa kallas begrepp och har utöver namn en datatyp, en balanstyp (debet eller kredit) och en periodtyp (flöde eller stock). Begrepp kan representeras av en vallista (*choice list*), genom ett värde i en angiven valuta eller genom ett värde i form av en fritext. Innehållet i taxonomierna finns på flera nivåer – myndighetsgemensamma basbegrepp utgör grunden, men kan kompletteras med sektorspecifika begrepp inom specifika tillämpningar. Till begreppen finns normalt beskrivande dokumentationstexter eller referenser till lagstiftning och andra standarder (eller båda). I en specifik tillämpning kan ytterligare referenser (till specifik normgivning eller lagstiftning) tillföras och avgränsningar av gemensamma begrepp kan göras. I den specifika tillämpningen kan även summeringsregler och andra kontroller tillföras.

SCB använder XBRL för att möjliggöra maskin-till-maskin-inhämtning av finansiell information från företag och organisationer, genom att

systemleverantörer bygger rapporteringsmöjligheter enligt specifikationer i taxonomierna.

6.3 Benämningar med anknytning till SSD

Redan på 1970-talet introducerades på SCB en struktur med dimensioner kallade *alfa-beta-gamma-tau* som ett angreppssätt för att beskriva innehållet i statistiska tabeller. Enkelt uttryckt motsvarar alfa en målpopulation, beta ett statistiskt mått och en målvariabel, gamma indelningsvariabler (typiskt korsklassificeringar) och tau referenstider. Statistikvärden i en statistisk tabell kan arrangeras med avseende på dessa dimensioner. För mer information, se t.ex.
The AlfaBetaGammaTau-model: A theory of multidimensional structures of statistics.
Denna struktur har påverkat SSD och dess beskrivningar.

I SSD förekommer benämningen *tabellinnehåll*. Den avser typiskt statistiskt mått och målvariabel, t.ex. andel sysselsatta. Den kan även användas för exempelvis osäkerhetsintervall, t.ex. konfidensintervall för andel sysselsatta.

7 Benämningar: synonymer – eller inte

Det här avsnittet listar och kommenterar ytterligare några variabelbenämningar som förekommer i SCB:s statistikproduktion.

7.1 Benämningar som kan användas

Benämningen *bakgrundsvariabel* förekommer i flera sammanhang. Det handlar typiskt om information som finns tillgänglig på mikronivå och med potential att användas som hjälpinformation på mikro- eller makronivå.

Benämningen *registervariabel* förekommer om variabler som hämtas från register och tillförs ett observationsregister, för att användas som observationsvariabel, målvariabel eller hjälpvariabel.

I anslutning till identitetsvariabel förekommer benämningarna kopplingsvariabel och länk.

- En *kopplingsvariabel* är en variabel som används för entydig identifiering av enskilda objekt eller grupper av objekt.
- En *länk* mellan två register utgörs av en eller flera kopplingsvariabler.

Insamlingsvariabel används ofta om variabler som samlas in, oftast på mikronivå. Även *mätvariabel* förekommer. De är båda ett något smalare begrepp än observationsvariabel. Även variabler som härleds ur de direkt observerade variablerna kan ingå bland observationsvariablerna. Om de härledda variablerna är målvariabler förs de dit.

Fördelningsvariabel används ibland om en indelningsvariabel (variabel för gruppindelning). Denna term förekommer t.ex. i beskrivningar av strukturen i SSD.

7.2 Benämningar som bör undvikas

Undersökningsvariabel används ibland om en variabel i en undersökning. Termen är oprecis och kan avse flera roller. Den bör undvikas.

Spridningsvariabel används ibland om en indelningsvariabel (variabel för gruppindelning). Eftersom ordet spridning används om statistikredovisning finns en risk för sammanblandning med målvariabel. Termen är således mindre lämplig och bör undvikas.

Benämningen *statistikdata* förekommer som synonym till statistikvärde, t.ex. i Verksamhetsstödet och där främst i anvisningar till SSD. Denna term bör undvikas.