

## Riktlinjer för kontrollkodning

Kodning av variabelvärden enligt en klassifikation är förenad med osäkerhet i form av risk för kodningsfel. Tillsammans med andra former av osäkerhet inverkar denna osäkerhet på tillförlitligheten i statistiken. Kodningsfel är en form av bearbetningsfel.

Kodning ska kvalitetssäkras och kvalitetskontrolleras. För automatisk kodning gäller att minst 5 procent av materialet ska kontrollkodas i en lämplig undersökning minst vart tredje år. För manuell kodning gäller att minst 5 procent av materialet ska kontrollkodas löpande, med tillräcklig kontroll även av enskilda kodares arbete.

### Begrepp

I Verksamhetsstödet förklaras flera begrepp inom kodningsområdet. Här följer några kompletteringar:

Med **första kodning** (förstakodning, grundkodning, ordinarie kodning, originalkodning) menas den kodning som alltid ingår i produktionsprocessen, oaktat ev. kontrollkodning. Förstakodningen resulterar normalt (enligt SCB-praxis) i ett unikt kodvärde på variabeln i fråga för varje objekt/post.

Med **klarkodade värden** menas kodade värden som är resultat efter det att såväl första kodning som ev. kontrollkodning och rekonciliering har genomförts. Då har operativt ”sanna” kodvärden bestämts för de kvalitetskontrollerade posterna, medan de övriga klarkodade posterna endast är förstakodade. Klarkodade poster där ”sann” kod har fastställts används för analys genom jämförelser med samma posters förstakoder.

Med (operativt) **”sant” kodvärde** (sann kod, riktigt kodvärde) menas, i detta kvalitetskontrollsammanhang, det kodvärde som efter kontroll befinns vara det troligaste för objektet. Denna kontroll grundas normalt på kontrollkodning och rekonciliering. Rekonciliering kan leda till två eller fler likvärdiga kodvärden, utifrån tillgänglig information. Ett sant kodvärde kan alternativt bestämmas genom utredning i återkontakt med uppgiftslämnaren eller genom avstämning med klassifikationsansvarig eller undersökningsledare. Observera att det i denna mening sanna kodvärdet inte nödvändigtvis stämmer överens med den okända verkliga (”korrekta”) koden, som idealt skulle erhållas om man hade fullständig kunskap om objektet. (Den verkliga koden kan f.ö. vara mer eller mindre obestämbar även idealt sett, för sådana enstaka objekt som inte helt passar in i klassificeringens förenklade verklighetsbild.)

I de fall kodaren haft kontakt med klassifikationsansvarig eller undersökningsledare kan den satta koden anses vara **expertkodad** eller ”sann”. Det är diskutabelt om denna kod då ska kontrollkodas utan direkt anses vara rätt. Om syftet med kontrollkodningen är *kvaliteten* i undersökningen är det resursmässigt tvivelaktigt av två skäl. Dels finns det risk för att klassifikationsansvarig kontaktas två gånger för samma kod, dels att rekoncilieraren sannolikt kommer att välja den första koden vilket gör arbetet med kontrollkoden redundant. I de fall klassifikationsansvarig kontaktats två gånger kan det bli svårt för rekoncilieraren att fastställa rätt kod om den föregående gjort två olika bedömningar av den sanna koden. Det kan dock vara motiverat att kontrollkoda även expertkodade poster om syftet är samla in processdata över kodningsförfarandet.

**Masskodning** innebär att koder sätts för flera undersökningsobjekt samtidigt för en eller flera variabler (klassifikationer) vanligtvis efter selektering med hjälp av särskilda regler. Reglerna är undersökningsspecifika och bestäms i förväg eller under kodningens gång. Ett exempel är kodning av bortfall som då kan göras för alla objekt som saknar tillräcklig information. Masskodning gränsar till *automatisk kodning med regler* där kodningen med sedan tidigare sparade regler sker automatiskt vid laddning (i Prisma). Ett samlingsnamn för båda typerna av kodning är *satsvis kodning* (med hjälp av regler). Satsvis kodning betraktas som expertkodning.

### Vad ska kontrollkodas och rekoncileras?

I princip gäller att koder satta *internt* på SCB ska kontrollkodas och rekoncileras. Tillämpningen skiljer sig delvis mellan å ena sidan automatisk kodning, och å andra sidan datorstödd kodning, enligt följande.

Kontrollkodning och rekonciliering tillämpas i följande fall:

- Automatisk kodning genom maskinell kodning med lexikon: kontrollkodning görs vid valda tillfällen, nämligen minst vart tredje år i en utvald enkätundersökning.
- Kodning vid intervju, intervjukodning (kan också ses som en form av datorstödd kodning): kontrollkodning görs vid valda tillfällen, nämligen minst vart tredje år i första rotationsgruppen i AKU (lämpligen för två–tre månader).
- Datorstödd manuell kodning: kontrollkodning görs löpande.

Kodning vid intervju sker normalt i WinDATI där intervjuaren väljer en kod från en lista efter uppgiftslämnarens svar. Kodning med lexikon går i princip ut på att aktuell kod matchas på uppgiftslämnarens svar via ett standardiserat lexikon. De svar som inte kunnat matchas skickas i regel vidare med hjälp-information till datorstödd manuell kodning. Denna kodning utförs av utbildad personal som sätter en kod med hjälp av svaret i klartext (benäm-

ning), medföljande beskrivning och hjälpinformation från beslutsstödjande system.

Ingen kontrollkodning och rekonciliering tillämpas normalt i följande fall:

- Utredande kodning: manuell kodning där återkontakt med uppgiftslämnaren skett; koderna kan då anses vara riktiga.
- Expertkodning: manuell kodning i samråd med klassifikationsansvarig eller undersökningsledare; koderna kan då anses vara riktiga.
- Masskodning: (manuell) kodning av flera undersökningsobjekt samtidigt efter utredning eller förutbestämd kodningsregel; koderna kan då anses vara riktiga. Ses som en form av expertkodning.
- Extern kodning: kodning utförd av externa uppgiftslämnare; det finns inte underlag för rekonciliering.

Objekt/poster som utreds tillsammans med uppgiftslämnare eller undersökningsledaren (expertkodade) ska tillsvidare inte gå till kontrollkodning. Det rör sig om objekt med konstiga svarscombinationer som gör att objektet blir svår att koda. En del av dem bortfallskodas men långt ifrån alla. Det anses vara ett omak att kontaktas två gånger för samma objekt om objektet väljs ut till kontrollkodning.

Den kod satt via masskodning bör normalt ses som den riktiga på grund av konsistensskäl. Alla objekt som uppfyller masskodningsregeln bör ha samma klarkod. Masskodningen är tillämplig på ett material som innehåller många objekt av samma karaktär vilka därmed ska ha samma kod. Om masskodning är aktuellt för en produkt bör den utföras innan den manuella kodningen börjar. Detta för att undvika att manuell kodning och kontrollkodning görs redundant av den av masskodningen satta koden. I Prisma sparas förstakod, kontrollkod och masskod separat vilket gör analys och användning av endera kod möjlig.

### **Kodningsverktyget Prisma**

För datorstödd manuell kodning finns det centrala kodningsverktyget Prisma som stödjer förstakodningen samt efterföljande kontrollkodning och rekonciliering. Verktyget kan även användas enbart för kontrollkodning (och rekonciliering), genom att de utvalda posterna som ska kontrollkodas laddas in separat.

### **Urval för kontrollkodning**

Vid kontrollkodning ska minst 5 procent av det kodade materialet kontrollkodas och rekoncileras. De enskilda kodarna ska följas upp.

### **Automatisk kodning med lexikon**

Vid kontrollkodning av automatiskt kodade poster (objekt) dras normalt ett enligt nedan stratifierat OSU från samtliga kodade poster, om minst 5 procent.

Vid kontrollkodning av lexikon kan det vara lämpligt att stratifiera på olika grupperingar i klassifikationen, så att urvalet någorlunda speglar alla delar av lexikonet. Fördelningen på koder kan vara mycket ojämn. Det kan även vara lämpligt att stratifiera på andra aspekter som är intressanta för processen. För kontrollkodningarna av yrkeslexikon SLEX och PLEX år 2017 användes stratifiering bl. a efter antal klassifikationer och av lexikonerna kodad yrkeskod.

### **Kodning vid intervju**

För kontrollkodning av kodning vid intervju används med fördel den metodik som togs fram i samband med Lotta-projektet *Evalvering av yrkeskodningsmetod i AKU*. I det projektet användes data från två månader för rotationsgrupp 1. Sammanlagt kontrollkodades då 2 512 poster där huvudsysslan var samma båda kvartalen; av dessa behövde ca 12 procent reconcilieras. Vid kontrollkodningarna i AKU år 2017 användes data från tre månader för rotationsgrupp 1. Frågan om yrkesbeskrivning ställs vid andra intervjutillfället för denna grupp, för de individer som har samma yrke som de hade vid första intervjutillfället.

### **Datorstödd manuell kodning**

Vid kontrollkodning av manuellt kodade poster dras ett stratifierat OSU, eller motsvarande systematiskt urval, om minst 5 procent av hela materialet. Urvalet stratifieras efter de enskilda kodarna. Urvalsfraktionen kan behöva bli större än 5 procent beroende på vilka precisionskrav man har vid skattning av felfrekvens, särskilt per enskild kodare. Närmare riktlinjer för detta finns inte f.n., utöver vad som implicit ingår i det följande. Urvalet kan dras antingen under tiden den ordinarie kodningen pågår, så att kontrollkodning sker simultant, eller efter att den ordinarie kodningen är avslutad.

#### *Kontrollkodning samtidigt som den ordinarie kodningen*

Urvalet av kontrollkodningsposter dras systematiskt under den ordinarie kodningens gång. Ett minimiantal  $n$  poster per kodare måste kontrollkodas för att man ska kunna dra slutsatser om kodarens felfrekvens i hela materialet. Tills vidare sätts  $n$  till 20. De  $n$  första posterna väljs ut till kontrollkodning. Därefter selekteras med en frekvens på 5 procent, dvs. var 20:e post (talet 20 här beror inte av  $n$ ), helst med en slumpmässigt vald startpunkt (vilket dock för tillfället tyvärr inte går att göra automatiskt i Prisma). För att minska andelen kontrollkodningsposter ned mot 5 procent, kan frekvensen (efter att  $n$  poster har valts ut) väljas lägre än 5 procent.

För kontrollkodning väljs således först post 1–20, och därefter post  $k + 20$ ,  $k + 40$ ,  $k + 60$ ,  $k + 80$  osv., där  $k$  är ett tal mellan 1 och 20. För tillfället används  $k = 20$ . Men så snart verktyget Prisma har kompletterats och ger möjlighet till det med automatik, ska istället  $k$  sättas slumpmässigt (likformigt fördelat) mellan 1 och 20. Detta innebär att minst 5 procent av det kodade materialet kontrollkodas, och vi har en viss grund för att bedöma de enskilda kodarnas felfrekvenser.

I de fall där man anser det vara viktigt att kodarna inte vet att de utför en kontrollkodning används med fördel denna metodik. Urvalsmetoden stöds av kodningsverktyget.

Ett problem med urvalsmetoden är att urvalet av poster utgör ett OSU endast om posterna laddas i slumpmässig ordning (dvs. inte beroende av region, intervjuare eller annan faktor) avseende svårighetsgrad för kodning och att kodarna inte betar sig annorlunda vid de  $n$  första kodningarna (ingen träningseffekt).

#### *Kontrollkodning efter den ordinarie kodningen*

Urvalet av kontrollkodningsposter dras som ett OSU stratifierat per kodare. Till skillnad från då urvalet av poster dras samtidigt som den ordinarie kodningen känner vi här till antalet kodade poster totalt och per kodare. Fördelen är att urvalet kan designas effektivare. Ett minsta antal kontrollposter per kodare ska gälla. Urvalsfraktionen totalt kan ibland sättas till exakt 5 procent (medan fraktionen däremot alltid är större än 5 procent i det samtidiga urvalet). Urvalsmetoden stöds av kodningsverktyget.

#### *Hantering av kodkombinationer med flera klassifikationer*

Varje benämning motsvarar ofta en kodkombination för flera klassifikationer (t ex SEI - ISCO-08 - SSYK). Beroende på vilken benämning förstakodaren har utgått ifrån, kan en eller flera koder i kombinationen skilja sig åt. De koder som är olika går till rekonciliering och rättas upp. I en del fall kan då kombinationen av koder bli felaktig, eftersom rekoncilieraren bara ska ta ställning till de koder som fallit ut som omaka. Vid framtida kontrollkodning med mer än en klassifikation, bör därför rekoncilieraren även kontrollera kombinationen av koder. Exempel: förstakodaren kodar administrativ assistent med SSYK, ISCO-08 och SEI varpå enbart ISCO-08-klassifikationen väljs ut till kontrollkodning. Om kontrollkodaren då gör bedömningen att yrket borde vara bokföringskontorist istället, stämmer inte den ISCO-koden mot den SSYK och SEI som är satt i förstakodningen.

Åtgärden är att dra urvalet till kontrollkodning per individ (post) så att alla klassifikationer kontrollkodas samtidigt. Detta urval görs med fördel manuellt utanför Prisma.

## Bestämning av oacceptabel felfrekvens

Enligt GD-beslutet om kvalitetskontroll av kodning kommer det an på respektive undersökning/produkt att i samråd med processansvarig skatta felfrekvenser och bedöma vad som är "oacceptabel felfrekvens" eller "frekventa fel". Arbetet med bestämning av oacceptabel felfrekvens sker inom ramen för produktens förvaltningsarbete, dock i samråd med objektspecialisten för Samla in och mikrogranska (SIM). Finansiering ska ske inom ramen för berörd undersökning.

Uppföljningen görs både för kodargruppen som helhet och för enskilda kodare. En ökad felfrekvens för gruppen som helhet kan tyda på generella problem som dykt upp, t.ex. ökad tidspress eller ökad personalomsättning. En hög felfrekvens för vissa koder inom en klassifikation ger anledning till att förbättra instruktionen inom det problemområdet.

Vad gäller **produktkvaliteten** borde man idealt fråga sig: *hur påverkar kodningsfelet osäkerheten i statistiken?* Vad som ska ses som acceptabel felfrekvens varierar förstås mellan olika produkter.

En nyttig övning kan vara att formulera ett tänkt mål av typen: *På en viss redovisningsnivå får inte nettofelet vara större än  $x$  procent, eftersom en större felfrekvens påverkar statistiken. Basen för nettofelsberäkningen ska vara minst  $y$  observationer.*

Men detta ska inte hårdras, för kodningsosäkerheten kan inte ses isolerat. Frågan är även dels vad som är möjligt att uppnå, dels och inte minst hur kodningsosäkerheten samverkar med andra former av osäkerhet, och då kanske främst mätosäkerheten. Osäkra kodvärden kan delvis ha sin grund i diffusa svar, och dessa åter delvis i brister i mätinstrumentet; det måste man också tänka på.

Vilken typ av fel man ska beakta har att göra med typen av slutprodukt. Vi använder oss av två typer av fel: *Nettofelet* är antal enligt förstakodningen minus "sant" antal, dvs. antal felaktigt medtagna minus antal felaktigt uteslutna. *Bruttofelet* är det totala felet på mikronivå, dvs. antal felaktigt medtagna plus antal felaktigt uteslutna. Felandelar erhålls genom att felen divideras med antal enligt förstakodningen. Ett stort bruttofel behöver inte betyda ett stort nettofel, då det skulle kunna vara så att antalet felkodade poster från en grupp till en annan är lika stor som felkodningen åt motsatt håll.

För vanlig statistikredovisning är nettofelet mera direkt relevant än bruttofelet, som dock även där indikerar en potentiell osäkerhet. Direkt och genomgripande relevant är bruttofelet för analyser av t.ex. samband eller flöden. Exempel på sådana bruttofelskänsliga analyser är epidemiologisk forskning

på ULF-data och arbetsmarknadsrelaterade rörlighetsstudier på AKU-paneler.

Skattning av felfrekvens för bruttofel och nettofel görs antingen ovägt (för faktiskt kontrollkodade) eller vägt (till populationsnivå med angivande av medelfel).

**Kvaliteten i enskilda kodares arbete ska också kontrolleras.** Avvikande höga felfrekvenser för enskilda kodare kan nämligen ses som brister i processen som går att åtgärda, genom bättre instruktioner m.m. För respektive kodare beräknas ovägd felfrekvens. Denna behöver inte delas upp på kod eller i bruttofel och nettofel, eftersom materialet vanligen är för litet för det.

Det kontrolleras att ovägda felfrekvenser inte är oacceptabelt höga. Vad som är oacceptabelt har inte specificerats ännu. I ett mer sofistikerat system kan det vara aktuellt att använda hypotesprövning eller (simultana) konfidensintervall för att räkna fram vilka kodare som har signifikant högre felfrekvenser än övriga kodare eller än en fastställd gräns, t.ex. 20 procent, eller än i tidigare produktionsomgångar.

Om felfrekvenserna överstiger den acceptabla felfrekvensen vidtas planerade åtgärder, t.ex. extra utbildning via den övningsapplikation som tagits fram inom förvaltningen för Samla in och mikrogranska (SIM).

Det återstår att slutligt utforma standardiserade processdatautdrag för återkoppling till kodarna och underlag för dialogen mellan kodarna och deras arbetsledning.

### **Analys av resultat**

Analys av processdata från kodningsverktyget och resultat från kontrollkodningen ska tas om hand. Målet är att ständigt förbättra kodningsprocessen och kodningsresultaten på SCB.

- Resultaten av analysen används för att sätta in åtgärder såsom extra utbildning och förbättrade instruktioner.
- Kontrollkodningen dokumenteras så att erfarenheter kan tas tillvara både inom produkten och generellt på SCB.
- Resultat av analysen tas med i alla relevanta produkters kvalitetsdeklarationer.

Det ankommer på förvaltningsteamet för objektet Samla in och mikrogranska (SIM) att samordna och följa upp dessa insatser och sammanfatta dem i de årliga förvaltningsrapporterna. Förvaltningsteamerna för respektive produkt ansvarar för därtill hörande analys och kvalitetskommentarer.

I nuläget siktar vi främst på att utnyttja kvalitetskontrollen för att förbättra kodningen efter aktuell kodningsomgång. Det finns ingen tid i exempelvis AKU att justera skattningarna direkt utifrån nettofelen. För vissa andra produkter, lämpligen årsundersökningar, kommer dock troligen klarkodade värden att användas när statistiken tas fram. Att räkna upp från underurvalet av kontrollposter skulle införa ett extra urvalsfel, men minska det systematiska bearbetningsfelet från kodning. Om uppräkningsfel blir aktuellt är ännu inte fastlagt.