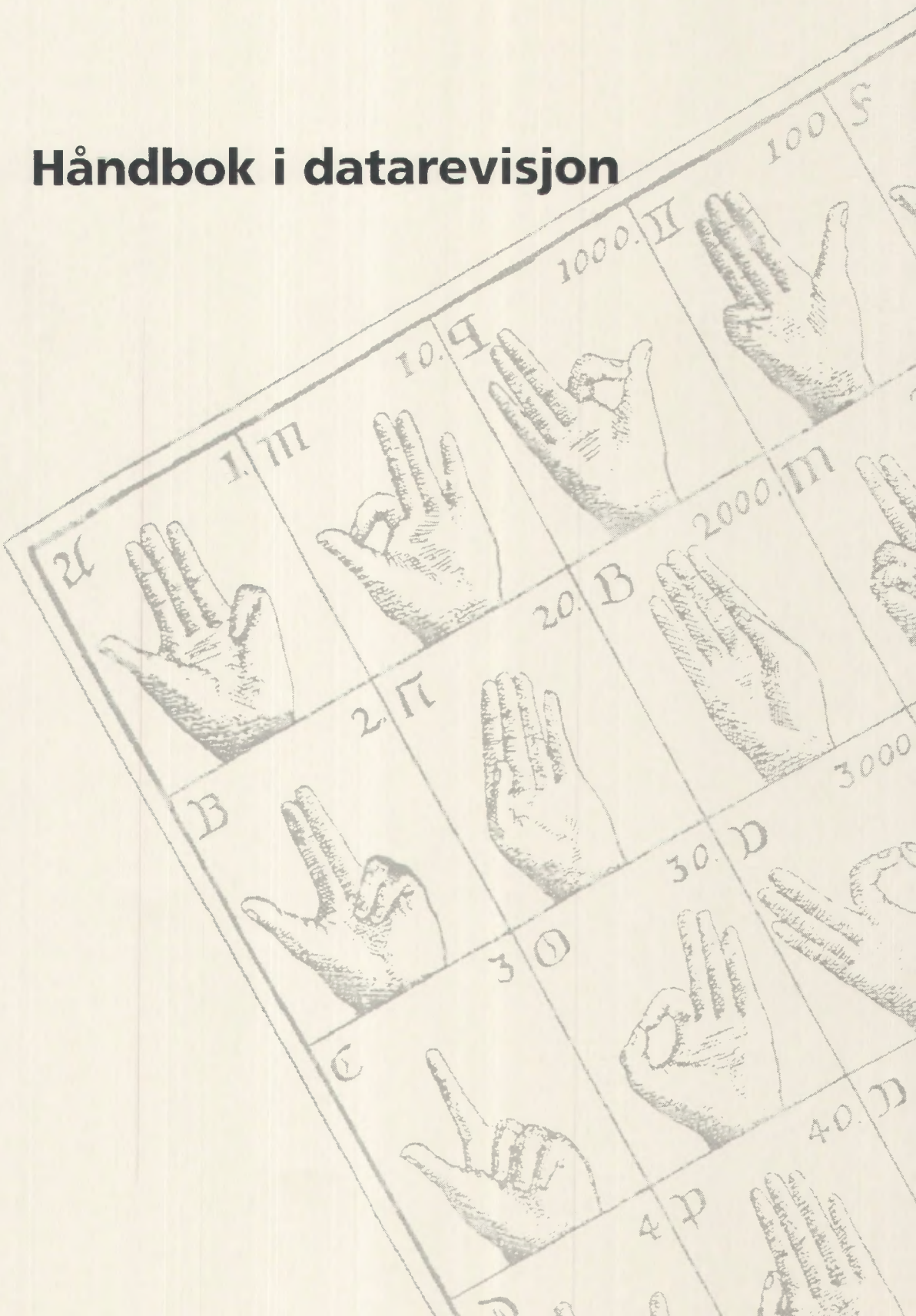
Håndbok i datarevisjon



Håndbok i datarevisjon

Statistisk sentralbyrå • Statistics Norway
Oslo — Kongsvinger, august 1998

Forord

Revisjonshåndboka gir en oversikt over rutiner for mottak og bearbeiding av data fram til publisering. Den er ment å være en oppslagsbok/huskeliste for god revisjonsskikk, men ikke en katalog hvor en kan finne konkrete løsninger til enhver revisjonsprosess. Store deler av håndboka bygger på rutiner som er i bruk rundt om i seksjonene idag og inneholder derfor mye kjent stoff. Den prøver likevel å fokusere på en del forhold som kan synes opplagte, men som vi likevel har lett for miste av synet i det daglige arbeidet.

Et viktig formål med revisjonshåndboka er å starte en prosess med å se på revisjonsarbeidet i SSB med et kritisk lys og vurdere alternative metoder. Håndboka omtaler da også en del metoder som er lite i bruk i SSB og kommer med konkrete forslag og anbefalinger av metoder og også en del krav, spesielt til lagring og dokumentasjon. Håndboka tar også opp hvordan vi bedre kan måle virkningen av revisjon. Det brukes betydelige ressurser på dette og det kan tenkes at vi bruker for mye ressurser. Effekten av innsatsen har vært lite evaluert og dermed lite kjent. Dessuten har den rivende utvikling av automatiske verktøy gitt håp om at det er betydelige gevinster å hente.

Prosjektet med revisjonshåndboka ble formelt godkjent av Direktørmøtet 16. april 1997. Som styringsgruppe for prosjektet har DM hatt saken oppe i flere møter, seinest 3. juni 1998. DM besluttet da at oppgaver knyttet til uttesting av metoder, videreutvikling av håndboka o.l. skal legges inn under det kommende metodeutvalget. Det opprettes også en ad-hoc gruppe som skal sørge for at forslagene i revisjonshåndboka tas hensyn til i VP-arbeidet for 1999.

Arbeidet har foregått i en prosjektgruppa med representanter fra de fleste fagseksjonene og metode- og IT-gruppene. Prosjektgruppa har i flere møter diskutert utkast til håndboka, sist gang 4. mars 1998. I tillegg har håndboka blitt tatt opp i IT-utvalget. Håndboka ble presentert for andre SSB-medarbeidere på et seminar 11. juni 1998.

Prosjektleder og leder av prosjektgruppa har vært rådgiver Frank Foyn. Foyn har også stått for utarbeidingen av håndboka.

Statistisk sentralbyrå
Oslo, 27. august 1998

Svein Longva

Innhold

1. Formålet med revisjonshåndboka

Revisjon er omtalt i SSBs «Strategiplan 1997 -» som ledd i en kostnadseffektiv statistikkproduksjon. I tillegg er datafangst og oppgavebyrde sentrale temaer i strategiplanen.. Denne håndboka er et ledd i oppfølgingen av strategiplanen på disse områdene, spesielt innen revisjon. Håndboka er ment å fungere som starten på en prosess med større bevisstgjøring på tema. Dette omfatter vurdering av ulike statistiske metoder og mer generelt aktivitetene i revisjonsprosessen. Målet er også at denne håndboka skal ha motiverende effekt på saksbehandlerne ved å sette deres rolle i produksjonsprosessen i perspektiv.

I løpet av de siste årene er det blitt nedlagt betydelige ressurser innen en rekke statistiske sentralbyråer på å utforme håndbøker om revisjon. I tillegg har det foregått et internasjonalt arbeid på feltet. Statistikkavdelingen i European Commission for Europe (ECE) har hatt gående et arbeid over mer enn ti år. Grunnen til dette er en erkjennelse av at det nedlegges betydelige ressurser på revisjon og at effekten av innsatsen er lite kjent. Dessuten har den rivende utvikling av automatiske verktøy gitt håp om at det er betydelige gevinster å hente gjennom en systematisk sammenligning av de forskjellige metoder for å identifisere beste praksis innen forskjellige felter. Formålet med denne håndboka er å sette igang en liknende prosess i SSB.

Som de fleste statistiske sentralbyråer bruker SSB mange forskjellige metoder i revisjonen. Målet må være at disse metoder skal være metodisk velfunderte og være en del av en samlet revisjonsstrategi. Gjennom å liste opp de forskjellige metoder skal håndboka gjøre det lettere å identifisere og bruke velfunderte metoder. I tillegg skal den foreslå metoder for evaluering og dokumentasjon av kontrollprosessen. Dette er kanskje den viktigste svakhet ved praksis i SSB idag. Det foregår lite evaluering av disse aktiviteter. På de områder det faktisk er gjort evalueringsstudier er disse sjeldent dokumentert på en måte som er til nytte for andre som skal legge opp en ny revisjonsprosess.

Håndboka er ikke en katalog hvor en kan finne konkrete løsninger til enhver revisjonsprosess. Ennå har det ikke vært mulig å identifisere en klar beste praksis innen noe felt og det har heller ikke vært mulig å utprøve alle tilgjengelige metoder og verktøy. Dette arbeid må fortsette ute i avdelingene og seksjonene. Dersom vi skal arbeide effektivt fram mot en strategi på feltet er det nødvendig at avdelingene identifiserer visse statistikker for fortsatt utprøving av programverktøy til editering. For å gjennomføre og koordinere aktiviteten bør Avdeling for samordning og utvikling få et særlig ansvar når det gjelder metoder og valg av programvare. Det bør opprettes et revisjonsforum med representanter også fra fagseksjonene.

Det er nylig gitt ut håndbøker i andre statistikkbyråer om samme tema. I mars 1997 ble rapporten «Granska effektivt» utgitt i Statistiska centralbyrån og i desember 1997 ble rapporten «Feilsøgningskatalog» utgitt i Danmarks Statistik. Den norske revisjonshåndboka bygger i stor grad på innholdet i disse rapportene. En annen referanse er Data Editing, en manual utgitt av Australian Bureau of Statistics.

2. Hva menes med revisjon og hvorfor er det nødvendig ?

2.1. Hva er revisjon og hvilken rolle har den ?

Håndboka tar for seg :
- **mottak av data**
- **koding**
- **kontroll av data**
- **feilretting**
- **imputering**
- **kobling**
- **lagring**
- **dokumentasjon**

Denne håndboka omhandler revisjon i vid forstand, dvs. klargjøring av mottatte data fram til publisering. Dette omfatter mottak av data, koding av data i henhold til ulike standarder, kontroll og feilretting, imputering, kobling, lagring og dokumentasjon av data. Med editering (eng: edit) menes revisjon i noe snevrere forstand, først og fremst knyttet til aktivitetene kontroll, retting og imputering av data.

Det er nyttig å avklare forholdet mellom statistiske metoder og revisjonsprosessen. De siste årene er det utviklet en rekke metoder for imputering av verdier for frafall, både når hele enheten mangler og ved partielt frafall. Derimot synes det fortsatt som om det å identifisere feil og mistenkelige verdier er basert på ad hoc metoder og at metoder for identifikasjon av ekstremverdier (outliers) er blitt lite systematisk anvendt på feilsøking.

Revisjonen skal :

- identifisere feil eller mangler

- rette eller korrigere for feil

- godkjenne mistenkelige verdier

Revisjonens oppgave er først og fremst å identifisere og rette eller korrigere for feil i datamaterialet for å øke datakvaliteten.
Med revisjon tenker vi tradisjonelt på den kombinerte prosessen med :

- Kontroll av data med sikte på å identifisere feil, mistenkelige og manglende verdier og
- Retting av feil, mistenkelige eller manglende verdier med korrekte eller sannsynlige verdier eller
- Godkjenning av mistenkelige verdier

Det er viktig at direkte feil i data blir rettet opp. Hvis en ikke kjenner den riktige verdien, må en være sikker på at den nye verdien er klart riktigere enn den opprinnelige.

Koding av data i henhold til ulike standarder er en del av revisjonsprosessen. Korreksjon for helt eller delvis frafall regner vi vanligvis ikke som revisjon i snever forstand, men er et viktig element i forbedring av datakvaliteten.

men også .....

Et annet viktig formål med revisjon er å identifisere problemer og feilkilder som blir avdekket under innhenting og bearbeiding av data.

- identifisere problemer og feilkilder

For alle typer undersøkelser gjelder det å få fram mulige svakheter i undersøkelsen som kan være til nytte i vurderingen av resultatene. Det er likevel spesielt viktig for regelmessige undersøkelser at erfaringer fra foregående undersøkelser blir brukt for å forbedre framtidige undersøkelser. Slike forbedringer er det forholdsvis enkelt å gjennomføre for skjemabaserte undersøkelser, endring på spørsmålsformulering mv. For administrative data er dette ofte en tyngre prosess og det vil være nødvendig med en mer langsiktig horisont.

2.2. Årsaker til feil og mangler i mottatte data

Mange årsaker til feil :

Det er flere grunner til at det er feil i innkomne data og som gjør revisjon nødvendig:

- Skjemainnhold/ - design
  - *Forståelse av spørsmålene i oppgaven*

Er skjema forståelig ?

Både ved skjemabaserte undersøkelser og intervjuundersøkelser er det viktig at oppgavegiver forstår hva det blir spurt om, hva vi er ute etter å kartlegge. Begreper og definisjoner må gå klart fram på en lett forståelig måte. Det vil være en avveining mellom en forholdsvis kort og ikke fullstendig forklaring kontra en lang og uttømmende forklaring. Faren ved lange forklaringer er at rettledningen blir for omfattende og at oppgavgiver går rett løs på spørsmålene uten å sette seg inn i definisjonene.

Ofte er det interne sammenhenger i skjema som skal stemme overens. Slike sammenhenger er det viktig at oppgavegiver blir klar over. Dette er en hyppig årsak til uoverensstemmelser i data. Design og uttesting av skjema på forhånd er viktig for å lette forståelsen hos oppgavegiver.

  - *Definisjon av variable*

Hvordan passer våre definisjoner for oppgavegiverne

Selv om spørsmål i *skjemaundersøkelser* blir forstått, hender det ofte at oppgavegiver ikke kan svare nøyaktig på spørsmålet. Dette gjelder bl.a. i økonomisk statistikk der vi spør om kjennemerker som ikke kan hentes direkte i foretakets regnskap, f.eks. vil investeringer etter nasjonalregnskapets definisjon atskille seg fra definisjoner i et finansregnskap. Dette er som regel en permanent situasjon som ikke bedrer seg over tid. Ofte kan det lønne seg å endre definisjonen etter oppgavegiverens data og heller foreta korreksjoner i det samlede materialet etterpå. Det er uheldig stadig å påpeke avvik overfor oppgavegiver. Det viser seg ofte også vanskelig å be oppgavegiver om å gi skjønnsmessige anslag. Mange oppgavegivere vegrer seg for dette og det vil også svekke statistikkens omdømme («det er ikke så nøye hva vi rapporterer til SSB»).

ved

skjema-undersøkelser

og

Ved bruk av *administrative data* må vi i enda større grad tilpasse oss de definisjoner som ligger i de administrative registre. Vi kan

ved

administrative data

likevel bruke vår innflytelse som er hjemlet i Statistikkloven om å påvirke innholdet i offentlige registre (§ 3-2) og undersøkelser som skal utføres av forvaltningsorgan (§ 3-3). SSB har også inngått en avtale med registereierne om å gjøre innholdet mest mulig til egnet til statistisk bruk. Det er likevel under oppbygging av *nye* registre vi har størst mulighet i å lykkes med dette og derfor bør være mest aktive (se mer kap. 5.2.2)

- Statistisk enhet
  - *Datatilgjengelighet på riktig nivå*

Spør på riktig enhetsnivå

I personstatistikk er enheten enten den enkelte person eller husholdning. I næringsstatistikk er de vanligste enhetene foretak eller bedrift, men også bransjeenhet og lokal enhet brukes. Det er viktig å være klar over at noen opplysninger, f.eks. finanskostnader, kan være lett tilgjengelig på foretaksnivå, men ikke på bedriftsnivå og omvendt. Det er derfor viktig å ha klarlagt på hvilket nivå vi faktisk trenger dataene og om dette er et nivå oppgavegiver faktisk kan gi opplysninger på. Avvik mellom ønsket enhetsnivå og tilgjengelig enhetsnivå er en hyppig årsak til feil og bør unngås. Der det er nødvendig med avvik vil det være en fordel å ta utgangspunkt i det nivået oppgavegiver har data og så be oppgavegiver fordele eller summere. Dette kan gjøres for kvantitative variable (f.eks. antall sysselsatte), men vanskelig for kvalitative variable (f.eks. holdningsspørsmål).

  - *Registeropplysninger*

Er det feil i registerene ?

Bedrifts- og foretaksregisteret (BOF) danner basis for næringsundersøkelsene. Målet er at BOF skal være et heldekkende og kvalitetssikret register. Det vil delvis være en målkonflikt ut fra at ressursbruken er begrenset. For mange statistikkområder er kvalitetssikring av de enhetene som finnes i BOF viktigere enn at BOF er fullstendig heldekkende. Feil ved enhetene i BOF (ikke riktig oppdatert navn, adresse mv.) fører til ekstraarbeid ved innhenting av data. «Feil» næringskode kan medføre at foretak får tilsendt skjematyper som ikke passer for virksomheten foretaket driver. I tillegg vil våre ulike undersøkelser ha behov for ulike statistiske enheter for samme konsern/foretak. Vår inndeling av foretaket i bedrifter kan også være dårlig tilpasset foretakets interne organisering og ofte går heller ikke den fullstendige inndelingen fram for foretaket. Dette medfører at vi ofte får inn oppgaver fra andre enheter enn det som var forutsatt.

Tilbakemelding viktig

En viktig kilde for oppdatering av BOF er de ulike næringsundersøkelsene. Det er derfor viktig at det er gode rutiner for tilbakemeldinger fra næringsundersøkelsene til BOF og for kvalitetssikring og oppdatering av disse opplysningene.

Data fra andre registre samsvarer ofte ikke med våre enheter og enhetsdefinisjoner. Dette er et betydelig problem, spesielt der opplysninger blir koblet mot BOF. Enheter fra andre registre vil ofte være sammenslåinger (aggregeringer) av spesielt våre bedriftsenheter (f.eks. Merverdiavgiftsregistret, A/A-registret). Dette skaper spesielt problemer ved nedbryting på næring og region. Enhetsregistret vil delvis bøte på dette, men det er langt fram før våre viktigste administrative registre er samordnet i den grad vi ideelt ønsker oss.

- Manglende motivasjon fra oppgavegiver

Manglende motivasjon gir dårlige oppgaver

Mange oppgavegivere er generelt motvillige til å fylle ut skjemaer. Denne motviljen øker klart med de forhold som allerede er nevnt. Dette sammen med bruk av oppgaveplikt gjør at mange mottatte oppgaver er av generelt dårlig kvalitet. Det er ingen enkel sak å motivere oppgavegiverne til å gi oppgaver av god kvalitet, men at skjema er tilpasset oppgavegiver og formålet med undersøkelsen kommer klart fram er forhold som bidrar til bedre motivasjon.

- Frafall

Hvordan behandle frafall ?

Ofte mangler hele eller deler av skjemaopplysningene på grunn av nekting og andre årsaker. Dette gjelder spesielt frivillige undersøkelser, men er også et problem i oppgavepliktige undersøkelser. F.eks. er frafallet i forbindelse med AKU nesten 10 prosent. Det er utviklet en rekke statistiske metoder for å redusere effektene av frafall i noen husholdningsundersøkelser, men deres egenskaper er bare delvis kjente. Innen økonomisk statistikk er det gjort lite for å vurdere effektene av frafall. Det er unntaksvis blitt foretatt frafallsundersøkelser (bl.a. Innovasjonsundersøkelsen 1992), men vi har liten erfaring med å korrigere resultatene for slike frafallsundersøkelser.

Viktigste oppgave:

Oppsummering:

Forbedre kvaliteten i data vi mottar fra oppgavegiverne

Revisjonen kan bare i en viss grad rette opp dårlig kvalitet i innkomne data. I tillegg er slik oppretting en ressurskrevende aktivitet. Den viktigste oppgaven for å forbedre datakvalitet og redusere behovet for revisjon er derfor å øke kvaliteten i de data vi mottar fra oppgavegiverne.

3. Mål for revisjonsarbeidet

Ingen datasett kan være feilfrie

Det er viktig at statistikkproduktene våre har god kvalitet. Dette betyr imidlertid ikke at alle data skal være fullstendig feilfrie. Et slikt mål vil i praksis være umulig å oppnå. Det vil heller ikke være optimalt å produsere feilfrie data. I vurderingen av hvor langt en skal gå må vi ta hensyn til både de kvalitetskrav brukerne setter til statistikken og kostnadene for å nå dette kvalitetsmålet. Det er viktig å være klar over at statistikkens aktualitet er en del av kvalitetsaspektet i tillegg til selve påliteligheten i tallene.

Vurder krav til kvalitet mot kostnadene

Ikke rett til gale verdier

Det er likevel ønskelig at alle absolutte feil og sannsynlige feil blir rettet opp i datasettet. Hvis en ikke kjenner den riktige verdien, må en være sikker på at den nye verdien er klart riktigere enn den opprinnelige. Det betyr at en bør ha et godt faktagrunnlag for å rette. Hvis ikke, bør en la være å rette. Det er viktig å unngå at riktige verdier faktisk blir rettet til gale verdier ! Dette kan skje selv i systemer med omfattende kontrolloppleggog nitidig gransking av materiale. Husk at formålet ikke er at data skal rettes slik at de ikke faller ut på nytt i en *vurderingskontroll.* Men data bør rettes slik at de ikke faller ut på nytt i *absolutte kontroller.*

Hvem er brukerne av statistikken

og

hvilke krav stiller de ?

Brukerne av en statistikk kan være en homogen gruppe med stort sett de samme krav til statistikken. Mer vanlig er at brukerne er en uensartet gruppe som stiller svært forskjellige krav. Noen kan ønske grove indikasjoner, men med svært god aktualitet (som konjunkturindikator), mens andre brukere kan sette store krav til pålitelighet ned på detaljert nivå, men med svært beskjedne krav til aktualitet (forskningsformål).

Det er derfor viktig å ha klart for seg hvem brukerne av statistikken er, både sentrale brukere og andre brukere, hvilken type informasjon de ønsker og hvordan brukerbehovene skal prioriteres.
Sentrale spørsmål i denne sammenheng er:

- Hva blir publisert ?

Er detaljer viktig ?

- Hvilke kjennemerker eller variable er sentrale; hovedposter eller også underspesifikasjoner ?

Større undersøkelser inneholder et betydelig antall spørsmål eller kjennemerker der en del kjennemerker er underspesifikasjoner av andre. Et eksempel er bl.a. regnskapsstatistikk. Det er viktig å være klar over om underspesifikasjonene primært er med som kontrollspørsmål for hovedpostene eller om også underspesifikasjonene er sentrale for brukerne. Hvis underspesifikasjonene primært er kontrollvariable, kan kvaliteten på disse reduseres sålenge hovedpostene er riktige.

- Aggregeringsnivå.

De fleste undersøkelser kan fordeles ned på region (fylke/kommune). Foretaksstatistikk kan i tillegg fordeles på bl.a. næring og størrelsesgruppe og personstatistikk på bl.a. alders

gruppe, sosioøkonomisk gruppe. Fordelinger ned på detaljert nivå krever generelt større kvalitet på enkeltopplysninger enn for undersøkelser som blir publisert på et svært aggregert nivå, f.eks. konsumprisindeksen.

Publiseres foreløpige tall ?

- Publiseres foreløpige tall ?

Krav til kvalitet på foreløpige tall vil være mindre enn til endelige tall. Det kan forsvare en mer summarisk behandling av oppgaver i første omgang for publisering av foreløpige tall og en grundigere behandling i neste omgang. Det er imidlertid viktig å unngå dobbeltarbeid i en slik prosess, også unngå at enheter blir kontaktet flere ganger for avklaring av ulike forhold.
Det er spesielt viktig at innsatsen konsentreres om enheter med betydelig innvirkning på sluttresultatet i en rutine med publisering av foreløpige tall (se kap. 4.4.1 Selektiv revisjon).

Bruk av ikke-publisert materiale

- Hvordan blir materialet brukt på annen måte enn direkte publisering ?

Selv om bare en begrenset del av datamaterialet blir publisert, kan det være betydelig etterspørsel etter mer detaljerte opplysninger. Dette kan være i form av faste abonnement/oppdrag (f.eks. utenrikshandel på detaljert varenivå, varehandelsstatistikk på kommunenivå) og mange, enkeltstående forespørsler. Slik bruk av datamateriale må tas med i vurderingen av kvalitetskrav til statistikken.

Brukes mikrodata ?

og

Hvilke konsekvenser har det ?

- Bruk av mikrodata

Med mikrodata menes informasjon for den minste statistiske enheten, som regel person/husholdning eller bedrift/foretak. Opplysningene om enheten kan foreligge med full identifikasjon (navn, adressse), avidentifisert eller anonymisert.
Mikrodata blir brukt til å generere statistikk (aggregerte tall) på primærområdet og publiseres generelt ikke. I enkelte tilfelle er det likevel sammenfall mellom den statistiske oppgaveenheten og publiseringsnivået. F.eks. publiseres økonomitall for hver kommune basert på kommuneregnskaper.

Mikrodata kan også brukes til koblinger fra et datasett til et annet datasett. Koblinger kan gjøres mot samme sett av data for foregående periode(r) (tidsserie paneldata) eller mot et helt annet datasett, f.eks. kobling av FoU-undersøkelsen mot industristatistikken. Et koblet materiale kan både brukes til publisering av statistikk for et bredere sett av data, men blir svært ofte brukt til analyse og forskning. Til forskningsformål brukes også mikrodata fra samme datasett, uten kobling mot annet materiale. Også til bruk i offentlig planlegging og forvaltning kan mikrodata bli stilt til disposisjon på spesielle vilkår.

Bruk av mikrodata stiller sterkere krav til kvalitet på detaljert nivå enn bruk av aggregerte tall. Dette må tillegges vekt ved revisjon av

data, men betyr likevel ikke at mikrodata må være feilfrie. Større krav til kvalitet på mikrodata fører til mer omfattende revisjon og kan også medføre forsinkelser i publisering av aggregerte data (hovedtall). Det må derfor foretas en avveiing av behovet for rask publisering av hovedtall og kvaliteten på mikrodata. En viktig premiss er primærformålet med undersøkelsen.

Hvis det er betydelig forskjell i kravet til revisjonsinnsats ved rask publisering av hovedtall og mer detaljert bruk, kan det være hensiktsmessig med en rask gjennomgang først for å rette opp betydelige feil i materialet og mer grundigere kontroll i neste omgang. Det kan være vanskelig å unngå en viss form for dobbeltarbeid ved en slik prosess.

Selv med en omfattende kontroll for å sikre kvaliteten på mikrodata, er det i praksis umulig å gardere seg mot at bruk av materialet til ulike analyse-, forsknings- og planleggingsformål ikke vil fange opp (mistenkelige) feil i data. I stedet for at det brukes mye tid på nitidig behandling er det derfor en forsvarlig ordning at brukerne får tilgang til mikrodata med advarsel om at datamateriale kan inneholde feil. Det er viktig at forskere og andre brukere av materiale da melder tilbake om feil de har funnet. For grove enkeltfeil eller systematiske feil i materialet bør det være en pliktig tilbakemelding.

4. Type kontroller og metoder for retting og imputering av data

4.1. Ulike typer data

Data kan grupperes i forskjellige kategorier. Hvilken kategori de tilhører har betydning for hvordan de skal kontrolleres og rettes.

Identifikasjonsdata

- Identifikasjonsdata
  Dette er variable som identifiserer enheten. Personnummer og organisasjonsnummer er de mest vanlige identifikasjonsvariable i henholdsvis personstatistikk og foretaksstatistikk, men det vil også være andre variable som identifiserer enheten. Identifikasjonsvariable bør alltid være riktig.

Kvalitative data

- Kvalitative data
  Dette er variable som har en eller flere gyldige verdier, men et begrenset sett av verdier (diskrete variable). Vanlig i holdningsundersøkelser, uttrykt som tallkoder (1 for ja, 2 for nei), men også som klassifikasjonsvariable (næringskode, utdanningskode, kjønn). Kontroll av gyldige verdier vil være absolutte kontroller, men kvalitative variable kan også inngå i vurderingskontroller sammen med andre variable.

Kvantitative data

- Kvantitative data
  Dette er numeriske verdier som vanligvis kan anta et ubegrenset sett av verdier (kontinuerlige variable), f.eks. omsetning, men det vil som regel finnes minimums- og maksimumsverdier, f.eks. lønn for ansatte i staten. Det gir mening å regne på kvantitative variable, f.eks. gjennomsnittlig omsetning i en bransje. En gjennomsnittsverdi kan imidlertid anta en verdi som er ugyldig på enkeltobservasjoner, f.eks. at gjennomsnittlig antall personer i norske husholdninger er 2,7. Tidsvariable, som datoer, hører med blant kvantitative variable.

4.2. Kontrollnivå og metoder

Kontroll av data kan foretas på:

- mikronivå
- makronivå

Mikronivå

Mikronivå.
Dette innebærer at hver enhet blir kontrollert for seg, ofte også felt for felt. Enhetene blir sjekket for absolutte feil og mulige feil (vurderingskontroller). Kontroller på mikronivå er effektive for å rette opp absolutte feil. Faren er at det brukes like mye tid på å kontrollere og rette små og store feil i datasettet siden det utføres samme type kontroller for alle enheter. Det er imidlertid mulig å differensiere mellom store og små enheter ved å la avviksgrenser variere etter feltets verdi. Det bør f.eks. godtas høyere prosentvise endringer for små beløp enn for store verdier. Det kan også legges inn kontroller på endring i absolutte størrelser. For kvalitative verdier er det vanskeligere med slik differensiering.

Makronivå

Makronivå.
Ved feilsøking på makronivå tar en utgangspunkt i dataene på et aggregert nivå, f.eks. celler i tabeller for publisering. Kontroller av samlet materiale på makronivå vil stort sett være vurderingskontroller. Kontroller på makronivå gir en god oversikt over datamaterialet, får fram mistenkelige resultater og kan også avdekke absolutte feil i materialet. For avklaring av tvilsomme forhold er det imidlertid nødvendig å identifisere enkeltenheter som forårsaker uventete resultater og rette opp på mikronivå. Ett problem med revisjon på makronivå er at feil på to enheter delvis kan oppveie hverandre og dermed ikke vises i en kontroll på makronivå.

Kombinere mikro- og makronivå

Kombinert kontrollnivå
Kontroller på mikro- og makronivå utelukker ikke hverandre, men de kan supplere hverandre. Det er vanlig at begge metoder brukes i samme undersøkelse, men vanligvis separat. Det er imidlertid mulig å kombinere mikro- og makrokontroll. Slike integrerte kontroller kan være mer effektive og treffsikre (selektiv revisjon/ grafisk revisjon). Det er likevel en klar fordel om datasettet er korrigert for absolutte feil før det kontrolles på makronivå.

(En mer systematisk oversikt over makrometoder er gitt i kap. 4.4.)

4.3. Type kontroller

Det kan skilles mellom to hovedtyper av kontroller, absolutte kontroller og vurderingskontroller. Kontrollene kan foretas både for hver enkelt enhet (mikronivå) og for materialet samlet (makronivå).

4.3.1. Absolutte kontroller:

Absolutte kontroller :

Med absolutt kontroll menes en kontroll som med 100 % sikkerhet identifiserer en direkte feil i datasettet. Det er mest vanlig å foreta slike kontroller på mikronivå.

Eksempler på absolutte kontroller:

- validitet

- *Validitetskontroll.*
  Identifikasjonsopplysninger som personnummer og foretaksnummer er enten gyldige eller ugyldige.
  Kvalitative variable kan bare ha bestemte verdier, ofte bare to alternativer som ja eller nei, mann eller kvinne. Noen felt har bare gyldige verdier i henhold til en fastsatt liste (standard), f.eks. kommunenummer, næringskode.

- dubletter

- *Dublettkontroll*
  En observasjon skal bare forekomme en gang i datasettet, f.eks. person eller person x inntektstype. Filorganisering på det enkelte området bestemmer hva som er en dublett.

- frafall

- *Enhetsfrafall og partielt frafall*
  Kontroll av identifikasjonsopplysninger mot register identifiserer og bestemmer omfanget av manglende enheter (enhetsfrafall). Manglende enkeltopplysninger for innkomne enheter er partielt frafall.

- summer, inkonsistens

- *Konsistenskontroll (inkl. sumkontroll)*
  Konsistenskontroll er en kontroll mellom to eller flere variable hvis verdier skal harmonere på en meningsfull måte. Den mest typiske konsistenskontroll er at summen av underposter skal være lik i alt-posten. Det vil også være andre logiske sammenhenger som må være oppfylt, f.eks. at eiendeler i alt og gjeld og egenkapital i alt i en regnskapsbalanse skal være like og at det i et datasett ikke finnes menn som er gravide.

4.3.2. Vurderingskontroller

Vurderings-kontroller :

Vurderingskontroller skal fange opp verdier som er lite sannsynlige eller mistenkelige, men som kan være riktige. Vurderingskontroller kan foretas både på mikronivå og makronivå.

Svært ofte vil slike vurderingskontroller se på relasjoner mellom to eller flere datafelt innenfor samme enhet/record. Det kan også sjekkes relasjoner mellom ulike datasett. Svært vanlig er kontroll på endring fra forrige periode i tilsvarende undersøkelse for samme enhet. Person-/bedriftsopplysninger fra én undersøkelse kan også kobles mot

beslektede opplysninger for samme enhet i en annen undersøkelse og opplysningene sjekkes mot hverandre.

- *Konsistenskontroll*

- konsistens

Det vil være en flytende overgang om en konsistenskontroll er absolutt eller ikke. Aldersfordeling innen en familie eller laveste alder for fullført høyere utdanning er eksempler på relasjoner der det er vanskelig å sette noen absolutte grenser. Inntektsopplysninger fra skattemyndighetene og selvangivelsen er eksempel på opplysninger fra to datasett som bør være konsistente.

- *Intervallkontroller*

- intervall

Intervallkontroll betyr at en verdi bør ligge innenfor et variasjonsområde, vanligvis høyere enn en minimumsverdi og lavere enn en maksimumsverdi. For slike kontroller på mikronivå er disse grensene som regel fastsatt på forhånd og er like for alle typer enheter. Det kan imidlertid legges inn betingelser slik at grensene kan variere etter f.eks. størrelse eller region. Slike kontroller kan enten tas for bare et felt eller for to eller flere datafelt i sammenheng. F.eks. kan det settes en minimums- og maksimumsgrense på absolutt verdi av personinntekt for å sjekke urimelige verdier. For to datafelt innenfor samme enhet/record kan f.eks. settes en nedre og øvre grense for gjennomsnittlig lønnskostnad i et foretak; totale lønnskostnader dividert på antall ansatte. Svært vanlig er også kontroll av kjennemerke for samme enhet for to ulike perioder. Grensene vil her være en nedre og øvre grense for prosentvis endring fra forrige periode, f.eks. omsetning.

Vanlige vurderingskontroller på makronivå vil være:

- forholdstall innenfor samme datasett for ulike grupper, f.eks. rentabilitet i regnskapsstatistikken etter ulike bransjer
- endringstall fra forrige periode for samme kjennemerke for ulike grupper, f.eks. månedlig produksjonsindeks etter sektor
- forholdet mellom samme eller beslektet kjennemerke fra ulike undersøkelser, f.eks. sysselsetting fra Arbeidskraftundersøkelsene og A/A-registeret.

4.4. Makrometoder for feilsøking

Identifisér ekstremverdier

Makrometoder for feilsøking er teknikker for å identifisere mistenkelige verdier i datamaterialet som må sjekkes nøyere av saksbehandler. Metodene søker å peke ut de avvikende verdiene som påvirker totaltallene mest. Dette sikrer oss at det ikke brukes for mye tid på kontroll av mindre viktige avvik.

Problem med innliggere (inliers)

Ulike makrometoder vil lett identifisere ekstremverdier eller utliggere (outliers). Langt vanskeligere er det å identifisere innliggere (inliers). Med dette menes rapporterte dataverdier som ikke avviker særlig fra gjennomsnittet, men som er feil og burde hatt en helt annet og avvikende

verdi. Slike type feil kan også ha betydelig påvirkning på totalresultatet. Eksempel er manglende rapportering av en stor investering i Investeringsundersøkelsen. God områdekunnskap ved bl.a. å følge med i nyhetsbilde eller sjekk mot andre kilder kan avdekke slike feil. Mindre betydningsfulle, men systematiske feil av denne typen kan identifiseres ved å følge enkeltenheters rapportering over en lengre periode, f.eks. butikker som ikke rapporterer prisendringer på varer til den månedlige konsumprisindeksen. Slik rapportering fra flere enheter vil kunne gi betydelig over- eller underestimering av totalindeksen selv om hver enkelt observasjon betyr lite.

Makro feilsøkingsmetoder krever at hele datamaterialet er tilgjengelig maskinelt. Metodene egner seg derfor spesielt godt for maskinelt overførte data og optisk leste data. Metoden med interaktiv registrering og kontroll av papirskjemaer passer dårlig inn i en strategi med vekt på selektiv revisjon. Det kan faktisk være hensiktsmessig for enkelte undersøkelser å gå tilbake til konvensjonell dataregistrering hvis optisk lesning ikke egner seg.

Metoder for makrorevisjon

Det finnes flere metoder for feilsøking på makronivå:

- Aggregert metode
  Dette er mer et overordnet prinsipp for feilsøking enn et konkret verktøy. Metoden går i korthet ut på å foreta feilsøking i to trinn:
  - Først feilsøking på aggregert nivå for å kartlegge mistenkelige (tabell)celler; grupper av enheter
  - Feilsøking på mikronivå i de mistenkelige cellene.
  Selektiv revisjon bygger på dette prinsippet.

- Top-Down-metode
  Dette er også et mer generelt prinsipp enn et konkret verktøy. Metoden går ut på å studere de viktigste enhetene fra toppen, f.eks. de 15 største enhetene målt for en bestemt variabel (omsetning) eller de 15 største positive og negative endringer fra forrige periode. Alle data for enhetene granskes og evt. rettes. Prosessen foregår til feilrettingene er så små at de ikke påvirker resultatene på aggregeret nivå i særlig grad. Det er utviklet metoder for poengberegning av enheter og at enhetene prioriteres for kontroll på grunnlag av de beregnede poengene (DIFF utviklet av Latouche og Berthelot).

En annen statistisk feilsøkingsprosedyre er Hidiroglou-Berthelot-metoden. I stedet for faste grenseverdier for endringer som aksepteres baseres dette på egenskapene til dataene (se kap. 4.4.3)

Det finnes også kombinerte metoder for feilsøking og imputering, f.eks. neurale nettverk (se kap. 4.5.2.2)

4.4.1. Selektiv revisjon

Selektiv revisjon innebærer at en konsentrerer revisjonsinnsatsen til feil/-feilmeldinger som har stor innvirkning på totalresultatene, dvs. på aggre-

Rett opp de største feil først

gert nivå. En må da ha et kontrollopplegg der en kan blinke ut slike mistenkelige feil og en må konsekvent prioritere nærmere undersøking og oppretting av enheter som forårsaker slike feilmeldinger. Ved kontroll og evt. oppretting av de høyest prioriterte tilfellene kan en kjøre ut kontroller og blinke ut neste runde av feilmeldinger i en iterativ prosess. Selektiv revisjon med konsekvent prioritering av viktige feilmeldinger er effektiv ressursbruk. Generelt for alle undersøkelser vil revisjonsinnsatsen være begrenset og denne metoden sikrer oss at det er de minst viktige feilene som evt. ikke blir rettet når en må sette sluttstrek for kontroll av materialet. Selektiv revisjon egner seg også svært bra for publisering av foreløpige tall.

Små enheter har større vekt i utvalgsundersøkelse enn i totaltelling

Det er viktig å være klar over at betydningen av mindre enheter vil være forskjellig i en totaltelling og i en utvalgsundersøkelse. I en totaltelling vil en liten enhet bare representere seg selv og bety lite i totalbildet. I en undersøkelse med f.eks. 10 prosent utvalg vil den tilsvarende enheten telle 10 ganger så mye. Feil i enheten vil derfor ha langt større konsekvenser enn i en totaltelling.

Mer aktiv bruk av selektiv revisjon vil kunne erstatte mye av den systematiske og grundige kontrollen på mikronivå av mistenkelige feil som idag i stor grad utføres på mange undersøkelser. Selektiv revisjon egner seg spesielt godt for periodiske undersøkelser der en har god bakgrunn for vurdering av mistenkelige feil.

Makrokontroll krever at hele datamaterialet er tilgjengelig. Et mulig hinder for en effektiv makrokontroll, spesielt ved skjemabasert årsstatistikk, er at ikke alle oppgaver er kommet inn innen fristen. En del oppgaver, også for viktige enheter, kan komme inn forholdsvis lenge etter fristen. Et tiltak for delvis å bøte på dette er å være langt strengere ved innvilgning av utsettelser.

4.4.2. Grafisk revisjon

Sjekk ekstremverdier grafisk

Grafiske presentasjoner er en effektiv måte å få oversikt over datamaterialet på. Det er likevel riktigere å kalle dette et nyttig *verktøy* i revisjonsarbeidet enn *metode*. Chart-diagram viser hvordan enhetene (observasjonene) fordeler seg og blinker spesielt ut enheter som avviker betydelig fra de øvrige enhetene (outliers). Det fins program som gjør det mulig å hente fram enkeltenheter for å sjekke om enheten inneholder feil og evt. rette opp direkte. I slike program (f.eks. SAS Insight) får en lett opp informasjon både på mikro- og ønsket makronivå. Slike program kan derfor brukes for kontroll både på mikro- og makronivå. Programmene kan også identifisere absolutte feil for oppretting.

4.4.3. Automatisk fastsatte grenseverdier.

Ved intervallkontroller er fastlegging av grenseverdier (min/max) viktig. Grenseverdiene må ikke være for snevre slik at det blir for mange

kontrollutfall. Dette gjør det vanskeligere å sortere ut de viktige utfallene fra de mindre viktige og vil kunne føre til at en mister oppmerksomheten og også respekten for kontrollen. For vide grenser kan føre til at reelle feil ikke blir oppdaget og slipper igjennom. Manuelt fastsatte grenseverdier bør være basert på emnekunnskap og/eller empiriske erfaringer, f.eks. hvordan kontrollen har fungert i tidligere undersøkelser.

La datamaterialet fastsette grenseverdier

Et alternativ til manuelt fastsatte grenseverdier er bruk av automatiserte metoder. Fordelen med slike metoder er at de baserer seg på fordelingen av hele datamaterialet og kan på mer objektivt grunnlag fastslå hva som er avvikende og hva som ligger innenfor et akseptabelt nivå. Bruk av *median* og *kvartiler* som parametre kan være bedre enn gjennomsnitt og standardavvik som påvirkes kraftig av ekstremverdier (outliers).

Hidiroglou-Berthelot-metoden (HB-metoden)

HB-metoden er en statistisk feilsøkningsprosedyre basert på egenskapene til dataene. Metoden er utviklet ved Statistics Canada og er fullstendig maskinell når den først er programmert og testet. I Statistisk sentralbyrå er metoden blant annet brukt innen Produksjonsindeksen og Ordrestatistikken for industrien, samt husleieundersøkelsen i Konsumprisindeksen. Metoden er spesielt egnet til periodiske undersøkelser, men kan også brukes på tverrsnittsdata og på tvers av undersøkelser. I periodiske undersøkelser, kan man f. eks. benytte relativen i forhold til foregående eller tilsvarende periode foregående år. I tverrsnittsdata, f.eks. priser i ulike områder av landet, kan man beregne relativen i forhold til ett område.

HB-metoden

Metoden har flere fordeler. Kontrollen av dataene er maskinell. Det settes ikke konstante grenser for hva som aksepteres eller ikke, men utnytter egenskapene ved fordelingen til dataene. Ved hjelp av parametre kan man begrense antall observasjoner som må undersøkes og styre hvor stor vekt som skal legges på de store enhetene i forhold til de små.

Et maskinelt feilidentifiseringssystem bør være en del av et større IT-system for dataklargjøring, feilidentifisering og imputering, der stor vekt legges på systemets funksjon som et støttesystem for revisjonen. Systemet identifiserer feil i data utifra et sett med rutiner og regler. Data som ikke er flagget for feil blir heller ikke behandlet i revisjonen. Etter manuell kontroll eller rekontakt med oppgavegiver rettes feil i mikrodata. Systemet kjøres videre for imputering fram til produksjonsklar fil.

Etter beregninger og aggregeringer vil det foretas makrokontroller på seriene, noe som kan utløse ytterligere kontroller og feilretting på mikronivå.

4.5. Retting og korrigering av datamaterialet

To typer feil :

- frafall

og

- gale verdier

Det er i hovedsak to typer av feil eller mangler i et datamateriale:

- frafall i undersøkelsen
  - enhetsfrafall (enheter i undersøkelsesbestanden mangler fullstendig, f.eks. pga. opphør, nekting e.l.)
  - partielt frafall (enheten i undersøkelsesbestanden er med i datamaterialet, men en del opplysninger mangler; ufullstendig oppgave)
- feil eller mistenkelige verdier for enheter i undersøkelsesbestanden

Det er ulike metoder for å korrigere datamaterialet avhengig av type feil:

- Veiing for frafall av enheter
- Imputering for partielt frafall
- Manuell retting av feil
- Automatisk retting av feil

4.5.1. Korrigering for frafall

Veiing for frafall

4.5.1.1. Veiing for frafall av enheter

Det utvalg en står igjen med når frafallet er identifisert og tatt bort er nettoutvalget. Ved husholdningsundersøkelser er det vanlig å fordele nettoutvalget etter visse registervariable for å se om det er representativt. Hvis ikke gir en vekter til hver observasjon slik at nettoutvalget blir representativt for registervariablene. Dette viser seg å være utilstrekkelig i noen tilfeller og da må en imputere i tillegg til eller istedet for veiing. Metodeseksjonen har betydelig erfaring med slike teknikker.

4.5.1.2. Imputering for partielt frafall

Dersom bare deler av opplysningene i oppgaven mangler, er det en vanlig metode å imputere verdier for disse manglende opplysningene. Det er ulike metoder for imputering (se mer under 4.5.2.2).

Hvis omfanget av manglende data er begrenset, er det også mulig å sette inn manglende verdier manuelt (se 4.5.2.1).

4.5.2. Retting og korrigering for feil

Manuell oppretting har fordeler

4.5.2.1. Manuell oppretting

Manuell oppretting av saksbehandler på grunnlag av kontrollutfall (lister i batch eller interaktivt kontroll) er den vanligste metoden i bruk. Denne metoden har også flere fordeler i forhold til automatisert retting. Med manuell retting har saksbehandler full kontroll med datasettet og bestemmer selv hvilke verdier som tilordnes enheten.

Manuell retting foretas i mange tilfelle uten å ta kontakt med oppgavegiver. Dette er fullt forsvarlig i de tilfelle det er mer eller mindre

opplagt hva feilen er og hvilken verdi som er den riktige. Grunnlag for å rette slike feil kan f.eks. være opplysninger om enheten fra annen statistikk, foretakets regnskap eller andre pålitelige eksterne kilder. I flere typer undersøkelser vil det i praksis heller ikke være mulig å kontakte oppgavegiver på nytt. Dette gjelder for de fleste personbaserte undersøkelser og administrative datasett.

Kontakt oppgavegiver helst bare en gang

Selv om kontakt med oppgavegiver både er ressurskrevende og kan være plagsom for respondenten, er dette i mange tilfelle både ønskelig og nødvendig. Antall kontakter bør imidlertid reduseres til et minimum, helst bare en gang. Kontakt kan enten tas pr. telefon, brev, faks eller e-mail. Hva som er best av muntlig eller skriftlig kontakt avhenger av kompleksiteten i forespørselen. Enkle spørsmål for avklaring kan greit tas over telefon. Ved mer omfattende forespørsel kan det lønne seg med faks eller e-mail først slik at respondenten får tid til å avklare spørsmålene i forespørselen på forhånd.

Lær opp og lær av oppgavegiver

Behov for kontakt gjelder spesielt i tilfelle der det er uklarheter i oppgaven og tvil om hva riktig verdi skal være. I periodiske undersøkelser vil kontakt med oppgavegiver også være nyttig med sikte på å «oppdra» oppgavegiveren til bedre forståelse av definisjoner, sammenhenger i skjema m.v. Direkte kontakt med en del oppgavegivere er også nyttig for våre saksbehandlere, både for å få bedre innsikt på fagområdet og få negative/positive reaksjoner fra oppgavegivere på skjemaet mv. for forbedring av senere undersøkelser. Denne kontakten er også SSBs ansikt utad til «grasrota» av respondenter.

4.5.2.2. Automatisk retting og imputering

Retter et dataprogram like bra ?

I undersøkelser med mye manuell oppretting og der opprettinger i stor grad skjer på grunnlag av normtall/gjennomsnittstall, oppgaver fra forrige periode mv. bør det absolutt vurderes om opprettinger kan gjøres maskinelt i en eller annen form. Det er spesielt aktuelt i periodiske undersøkelser. I engangsundersøkelser med betydelige opprettinger og beregninger kan det også være aktuelt hvis det finnes tilrettelagte dataverktøy slik at investeringskostnadene er overkommerlige.

Metoder for automatisk retting :

Det finnes fra enkle og meget oversiktlige maskinelle beregninger til mer kompliserte intelligente systemer. Metodene kan brukes både til beregning av manglende data og til beregning av nye verdier for feil rapporterte data. Ved bruk og valg av maskinelle metoder bør en ha rimelig god oversikt over *årsakene* til feil og mangler ved dataene.

Ulike metoder for beregning av data (imputering):

Rateestimering

- Rateestimering
  Baserer seg på at det er et (forholdsvis) stabilt forhold mellom to variable i samme undersøkelse (tverrsnitt) eller samme variable i to

påfølgende undersøkelser (tidsserie). For eksempel kan beregnet gjennomsnittspris på varenivå for enheter som har oppgitt både mengde og verdi brukes til å beregne mengder for enheter som bare har oppgitt verdi.

Regresjon

- Regresjonsestimering
  Baserer seg på at det er et avhengighetsforhold mellom to eller flere variable (tid- og/eller tverrsnittsdata). I forhold til enkel rateestimering kan det trekkes inn flere forklaringsvariable og metoden tar også bedre hensyn til variasjoner mellom enhetene i datamaterialet.

Imputering :

- Deduktiv eller deterministisk imputering

- Deduktiv

Manglende verdi(er) for en enhet fastsettes fra andre data for samme enhet på grunnlag av logiske regler. F.eks. kan en sumpost beregnes som summen av underpostene (hvis disse er oppgitt). Hvis det går fram av oppgaven at en person er mor til barn, kan manglende avkryssing for kjønn settes til kvinne.

- Cold-deck imputering

- Cold-deck

Beregner manglende verdier på grunnlag av data for samme enhet fra foregående undersøkelse. Den enkleste form for cold-deck imputering er ren kopiering av f.eks. priser fra forrige undersøkelse, men det kan legges inn mer avanserte beregninger.

- Hot-deck eller donorimputering

- Hot-deck

Beregner manglende verdier på grunnlag av data for en annen enhet i samme undersøkelse; en enhet som i en eller forstand likner mest på «mottakende» enhet, f.eks. foretak i samme bransje eller personer i samme aldersgruppe.

- Neurale nettverk

Neurale nettverk

Kunstige neurale nettverk kan avsløre sammenhenger i data som ellers er vanskelige å beskrive og som må betegnes som atypiske når en tar hensyn til mange variable under ett. Nettverket kan brukes til imputering ved å gi et kvalifisert estimat på dataverdier som normalt vil forekomme, på manglende og feil dataverdier.

Kunstige neurale nettverk er et modellapparat som legger til grunn at det finnes sammenhenger mellom variablene man legger inn i modellen. Det kan f.eks være at endring i husleie antas å være korrelert med geografisk beliggenhet. Disse sammenhengene forsøker man så å tallfeste ved hjelp av algoritmer for numerisk optimering. Dette omtales ofte som *læring*, mens tradisjonell økonometrisk terminologi kaller dette *estimering*.

Hvordan nettverket utformes, med antall sammenhenger og antall lag i modellen, er av stor betydning for modellens egenskaper. Også valg av algoritme for læringen, og kriterier for stopp av læringen er viktig. Desverre finnes det pr. i dag få generelle retningslinjer å støtte seg til ved disse valgene. Man er i stor grad avhengig av å prøve seg fram for å finne en kombinasjon av valg som gir en modell med akseptable egenskaper.

Når nettverket er lært opp, kan det så f.eks brukes til å imputere manglende verdier. Dette gjøres ved å gi nettverket partielle observasjoner som input. Nettverket bruker så sin «kunnskap» om sammenhengene mellom variablene til å beregne de manglende verdiene.

Denne metoden finner vi blant de nyere framstøtene innen rutiner for automatisk imputering. Det er gjort forsøk i flere statistikkbyråer, Danmarks statistik, Central Office of Statistics (UK), Statistics Canada og her i Statistisk sentralbyrå (husleieundersøkelsen, landbruksstatistikk). I disse landene er metoden utprøvet i tester hvor man har ønsket å imputere partielt frafall. Befolkningsstatistikk og økonomisk statistikk er områder hvor metoden er testet.

Så langt har ingen tatt metoden i bruk i regulær produksjon av statistikk. Men forsøkene ansees av flere land å være lovende, og man forventer at metoden vil tas i bruk på mange områder.

5. Feil og feilsøking i de ulike faser i produksjonsprosessen

Hvordan data hentes inn og registreres har stor betydning for hvordan kontroll- og rettearbeidet kan og bør legges opp. I stor grad kan SSB selv velge innhentingsmetode og øvrig produksjonsopplegg; i andre tilfelle må vi tilpasse oss eksterne opplegg (spesielt administrative data).

Vi kan skille mellom to hovedmetoder for innhenting og registrering av data:

Metoder for datainnhenting :

- skjemabasert

- Skjemabasert innhenting enten med
  - Konvensjonell dataregistrering, med enkle kontroller for gyldige koder mv.
  - Interaktiv registering og omfattende kontroll av data foretatt av saksbehandler
  - Optisk lesing av skjema med kontroll av gyldige koder mv.

- elektronisk

- Elektronisk innhenting
  - Rådata på maskinelle filer (i hovedsak administrative data)
  - Elektroniske skjemaer utfylt av oppgavegiver med og uten innebygde kontroller (overføring på diskett eller linje, telefoninntasting)
  - Interaktiv registrering og kontroll av saksbehandler, CATI/CAPI - teknikk (direkte inntasting ved datainnhenting fra oppgavegiver pr. telefon eller ved besøksintervju)

Feilsøking kan foretas i følgende faser i produksjonsprosessen:

- ved mottak, før dataregistrering (mikronivå)
- under dataregistrering (mikronivå)
- etter dataregistrering
  - mikronivå
  - makronivå

5.1. Feilsøking ved skjemabasert innhenting

Ved skjemabasert innhenting vil feilsøkingen vanligvis foretas i flere trinn; både før, under og etter dataregistrering.

Feilsøking kan foretas i flere faser

Det kan variere etter type undersøkelse hvor omfattende feilsøkingen på mikronivå skal være før og under dataregistreringen. I vurderingen av dette må en ta hensyn til:
- den generelle kvaliteten på de innkomne oppgavene
- faren for dobbeltarbeid.

Under alle omstendigheter må det foretas kontroll og feilsøking etter dataregistrering (mikro og/eller makrokontroll). Det er i denne fasen den tyngste delen av revisjonsinnsatsen bør settes inn.

Selv om revisjonsprosessen for en aktuell undersøkelse er sterkt mikropreget, er det uansett nødvendig med en samlet makrokontroll før publisering. Et egnet nivå for denne kontrollen vil være de tabeller som skal publiseres eller fortrinnsvis på et noe mer detaljert tabellnivå.

Mottakskontroll av skjemaer

Hvor grundig skal data sjekkes ved mottak ?

Det vil for alle undersøkelser være nødvendig med visse rutiner for mottak av skjema fra oppgavegivere. Dette gjelder bl.a.:

- behandling av returer fra posten pga. feil adresse, opphør av virksomhet mv.
- behandling av ikke utfylte skjemaer og henvendelser fra personer/ virksomheter som ikke mener de er oppgavepliktige
- registrering av godkjente innkomne oppgaver
- purring av manglende oppgaver

Ved mottak må det være visse regler og stilles visse minimumskrav i en primærkontroll av oppgaven før den blir godkjent for videre behandling i produksjonsprosessen. Dette gjelder først og fremst kontroll på om oppgaven er tilstrekkelig utfylt eller ikke inneholder åpenbare feil. Om en skal gå et skritt videre med kvalitetssjekk av de innrapporterte oppgavene i denne primærkontrollen må vurderes og ses i sammenheng med den resterende produksjonsprosessen. Generelt må det likevel advares mot utvidet mottakskontroll.

Fordeler

- Fordeler ved utvidet mottakskontroll

Fordelen med en forholdsvis omfattende primærkontroll er at *feil og mangler oppdages tidlig* og kan tilbakemeldes oppgavegiver umiddelbart for oppretting mens rapporteringen er fersk. Det er langt vanskeligere å innhente tilleggsopplysninger fra oppgavegiver lenge etter at vi har mottatt oppgaven. Dette kan også gå utover vårt renomme på lengre sikt hvis vi stresser tidsfrister og tvangsmulkt, men bearbeider oppgavene langt utover disse tidsfristene.
En effektiv primærkontroll vil også kunne *avdekke systematiske feil* i oppgavene eller misforståelser blant oppgavegiverne på et tidlig stadium med bedre muligheter for å iverksette tiltak for å bøte på dette.

Med for dårlig mottakskontroll *mister* vi også *sanksjonsmidler* som tvangsmulkt overfor oppgavegivere med mangelfulle oppgaver hvis de først er blitt godkjent som innkommet.

Ulemper

- Ulemper ved utvidet mottakskontroll

Det er imidlertid problemer med å gjennomføre en omfattende mottakskontroll. For det første er det *tidspress* i denne fasen av arbeidet. Rundt tidsfristen mottas en omfattende mengde oppgaver og en må ofte konsentrere seg om returer og direkte henvendelser for å sikre seg at vi får oppgaver fra de enheter som skal være med i undersøkelsen. Det er også viktig å sikre seg at innkomne enheter ikke blir feilaktig purret eller ilagt tvangsmulkt.

En mottakskontroll må nødvendigvis bli mikroorientert. Dette betyr at alle enheter stort sett vil gjennomgå samme type behandling. Det blir derfor *vanskelig* å få til *prioritert behandling* av viktige enheter og feil og dermed en mest mulig effektiv samlet ressursinnsats ved revisjon. For stor vekt kan bli lagt på mindre vesentlige feil og mangler i denne fasen.

En omfattende mottakskontroll vil heller aldri kunne erstatte en grundig feilsøking seinere i prosessen. Det er derfor *fare for dobbeltarbeid* ved at delvis samme typer kontroller foretas flere ganger i produksjonsprosessen. Det er også uheldig overfor oppgavegiverne om de blir kontaktet flere ganger for avklaring i oppgavene og spesielt uheldig med *flere kontaktpersoner*. Antall kontakter og kontaktpersoner bør begrenses til en pr. oppgavegiver selv om det for store og viktige enheter kan være nødvendig med flere forespørsler.

Sentral datafangstseksjon

Unngå dobbeltarbeid ved sentral datafangst

Med en sentral datafangstseksjon kan en effektivisere rutiner for mottak, registrering og purringer mv. generelt. En mer omfattende primærkontroll vil være vanskeligere å gjennomføre i en sentral datafangstseksjon og videre bearbeiding og ferdiggjøring av oppgavene i fagseksjon. Et problem er at personer i en sentral datafangstseksjon ikke vil ha den samme spesialkompetanse innenfor hvert fagfelt som fagseksjonene har for å oppdage feil og mangler i de mottatte oppgavene. Et annet forhold er problemet, som lett vil oppstå, med flere kontaktpersoner overfor oppgavegiveren. Med en sentral datafangstseksjon er det derfor spesielt viktig med avklaring av arbeidsfordeling mellom denne enheten og fagenheten og at fagenheten bidrar til den kompetanseoppbygging som er nødvendig i datafangstenheten for å gjennomføre en mest mulig effektiv primærkontroll.

Med optisk lesing eller konvensjonell dataregistrering kan det være hensiktsmessig å utføre enkel revisjon som *koding* av informasjon i henhold til standarder før dataoverføring skjer. Slike oppgaver kan i større utstrekning enn annen revisjon og kontroll også gjennomføres av en sentral datafangstseksjon. Kompetanse på standarder kan være lettere å etablere. Det er viktig at slik koding før dataregistrering er av en slik kvalitet at det er ingen eller lite behov for etterkontroll på mikronivå av kodingen.

5.2. Feilsøking ved elektronisk innhenting

Deler av prosessen med mottak, kontroll og feilsøking av data blir svært forskjellig i forhold til skjemabasert innhenting. Dette skyldes to forhold:

- dataene er allerede tilgjengelig maskinelt og en manuell forkontroll og oppretting for å redusere maskinelle feilmeldinger er derfor ikke aktuelt
- ved elektronisk innhenting er det langt større muligheter for å legge maskinelle kontroller hos oppgavegiver før data sendes SSB og de bør derfor inneholde langt færre feil enn et papirskjema.

Det vil likevel være forskjeller i feilsøkingsmetoder mellom elektronisk innhenting ved hjelp av våre egenutviklede datafangstprogram og mottak av administrative data fra andre etater. Bruk av egenutviklede rutiner gir langt bedre muligheter for å legge inn optimale kontrollsystemer som en del av innhentingen.

En samlet makrokontroll av materiale før publisering av resultater vil være nødvendig også ved elektronisk innhenting på samme måte som for skjemabasert innhenting.

5.2.1. Elektroniske skjemaer

Elektroniske skjemaer :

For elektroniske skjemaer har vi store muligheter for å bygge inn kontroller ved innregistrering av data. Dette gjelder først og fremst kontroller internt i det aktuelle datasettet. Data for samme enhet for tidligere perioder kan også legges ved og det gir muligheter også for kontroll av endringer mot forrige periode. Derimot er det vanskelig med kontroll mot annet datamateriale.

mindre behov for mikrokontroll

Elektroniske skjemaer med innlagte kontroller er et effektivt tiltak for å sikre oss data uten alvorlige feil. Det er meget velegnet til å luke ut absolutte feil som f.eks. at summen av gjeld og egenkapital er lik summen av eiendeler i en regnskapsbalanse. Vi må imidlertid være varsomme med å bygge inn for mange vurderingskontroller i jakten på å sikre oss «feilfrie» data. Blant oppgavegiverne vil det være store variasjoner og kontrollene må tillate et forholdsvis vidt spenn i mulige sett av verdier. For omfattende og snevre kontrollgrenser med mange feilmeldinger vil for det første kunne føre til irritasjon blant oppgavegiverne. En vel så viktig konsekvens er at oppgavegiverne presses til å rapportere data som godtas av systemet, men som ikke er representative for enheten. Oppgavegiverne lærer etterhvert kontrollene og knepene for å unngå feilmeldinger.

For elektroniske skjemaer i en eller annen form fylt ut av oppgavegiver kan det være hensiktsmessig med mikrokontroll ved mottak, selv om behovet bør være klart mindre enn ved papirskjemaer. Med et fleksibilt kontrollopplegg kan det være betydelige feil eller mangler også ved slike oppgaver. En utlisting av visse sentrale størrelser (på papir eller skjerm) kan være hensiktsmessig for å identifisere enheter med åpenbare feil eller mangler og rapportere tilbake til enheten umiddelbart. Dette vil likevel avhenge av typen statistikk (korttidsstatistikk kontra årsstatistikk) og den videre behandlingen.

For elektroniske skjemaer vil neste trinn naturlig være en kombinert mikro- og makrokontroll.

For interaktiv innhenting og registrering av SSB (CATI-/CAPI-teknikk med innlagte kontroller) bør rutinene stort sett være de samme som for elektroniske skjemaer. Det kan imidlertid legges inn flere

vurderingskontroller, men det krever at intervjuer utøver skjønn om intervjuobjektet skal konfronteres med feilmeldinger eller ikke i hvert tilfelle.

For denne typen er det naturlig å gå direkte til makrokontroll etter dataregistrering, men med mulighet for å sjekke enkeltdata. I intervjuundersøkelser, spesielt overfor personer/ husholdninger, er det bare unntaksvis aktuelt å kontakte oppgavegiver på nytt.

5.2.2. Administrative (registerbaserte) data

Administrative data:

makrokontroll

Sentrale registre som brukes av SSB i statistikkproduksjon er bl.a. flere av Skattedirektoratets registre: Momsregister, LTO-registeret, Likningsregister og Selvangivelsesregister, Toll- og avgiftsdirektoratets TVINN-register (utførsel og innførsel av varer), Rikstrygdeverkets Arbeidsgiver- og Arbeidstakerregister, Regnskapsregisteret i Brønnøysund, Kommuneregnskapsdata og Statens kartverks GAB-register. Felles for disse registre er at vi mottar dataene for alle enhetene samlet og på en maskinell form. Det er derfor lite hensiktsmessig med omfattende mikrokontroll for å blinke ut spesielle enheter ved mottak. Det er heller ikke aktuelt å ta direkte kontakt med den enkelte enhet. Kommuneregnskapsdata avviker fra dette mønsteret ved at oppgaveenheten og publiseringsnivå er sammenfallende og det derfor er nødvendig med mikrokontroll.

Administrative data brukes ofte sammen med andre kilder og matches også mot andre registre og databaser på mikrodata. Kontroll mot andre kilder kan derfor være en viktig sekvens i registerbasert statistikk.

Administrativ kontroll ofte mangelfull for statistisk kontroll

Ved mottatte registerdata er vi prisgitt den kvalitetskontroll som er foretatt av registereier. Denne kvalitetssikringen kan være god eller dårlig, men den vil uansett primært være rettet mot kontroll for administrative formål. Slike kontroller *kan* være til begrenset nytte for bruk av materialet til statistiske formål. F.eks. vil Skattedirektoratet være opptatt av om utliknet skatt og nettogrunnlaget for dette er riktig, men mindre opptatt av om bruttooppgavene for inntekter og kostnader og underspesifikasjoner er korrekte. I tillegg godtas ulike føringsmåter.

Statistikkloven gir oss visse muligheter for å påvirke innholdet i offentlige registre og vi har inngått avtaler med de fleste registereierne. Det kan likevel ha begrenset verdi å utvide et register med nye kjennemerker hvis de bare skal brukes til statistiske formål. Faren er stor for at slik informasjon ikke blir tilstrekkelig kvalitetssikret av registeransvarlig. Det er derfor viktig at vi strengt prioriterer våre ønsker om nye kjennemerker og at vi i dialog med registeransvarlig sikrer oss at disse kjennemerkene får en tilstrekkelig god kvalitet. Dette betyr først og fremst at statistikkdata er kvalitetssikret innholdsmessig, men også at datamaterialet er kontrollert og rettet for trivielle sumfeil mv.

6. Evaluering og dokumentasjon av produksjons- og revisjonsprosessen

6.1. Indikatorer for evaluering og kvalitetssikring av revisjonsprosessen

Det er viktig med evaluering av revisjonsprosessen på hvert statistikkprosjekt, både for å vurdere om ressurser til revisjon står i rimelig forhold til effekten av revisjon og for å vurdere om nye metoder eller teknikker kan tas i bruk som vil kunne gi bedre avkastning. Endringer i selve undersøkelsen, inkl. tilgang til administrative data, utvikling av statistiske metoder og nye programverktøy er de viktigste grunnene for å vurdere nye rutiner.

For å kunne foreta en best mulig evaluering er det nødvendig med noen kvantitative mål for revisjonsprosessen. Sentrale mål er kostnadene til revisjon, effekten av revisjon på sluttresultatet og treffsikkerheten i kontrollopplegg.

Noen sentrale indikatorer og metoder:

6.1.1. Kostnader

Hvor mye ressurser bruker vi på revisjon ?

- Timeverk brukt på revisjon
  - utførte timeverk til revisjon ialt
  - utførte timeverk som andel av totalt utførte timeverk på statistikkproduktet
  - utførte timeverk pr. oppgaveenhet

Innsamling av data og Bearbeiding av data er allerede idag separate aktivitetskoder i Produktregisteret. Problemet er at disse aktivitetskodene ikke brukes fullt ut av alle som utfører disse aktivitetene. Dette må innskjerpes. Grensedragningen mellom de ulike aktivitetene kan være uklar, men ikke vanskeligere enn at saksbehandlernes egne anslag bør være tilstrekkelig for dette formålet.

6.1.2. Omfang og treffsikkerhet i kontrollrutiner

Er kontrollene gode nok ?

- Opprettingsfrekvens: $\frac{\text{Antall opprettinger}}{\text{Antall objekter (enheter/ datafelt)}}$

Opprettingsfrekvensen kan splittes opp i :

- Feilmeldingsfrekvens: $\frac{\text{Antall feilmeldinger}}{\text{Antall objekter (enheter/ datafelt)}}$

og

- Treffsikkerhet: (hitrate)

$$\frac{\text{Antall opprettinger}}{\text{Antall feilmeldinger}}$$

Feilmeldingsfrekvensen forteller hvor mange enheter i undersøkelsen en bestemt kontroll slår ut for, mens treffsikkerheten forteller hvor mange opprettinger som blir foretatt på grunnlag av feilmeldinger fra denne kontrollen. Poenget er å ha en høy treffsikkerhet, kombinert med relativt lav feilmeldingsfrekvens. Dette kjennemerker en effektiv kontroll.
Høy treffsikkerhet kombinert med høy feilmeldingsfrekvens, tyder på at det er noe fundamentalt galt i datamateriale.
Lav treffsikkerhet indikerer at kontrollen ikke er særlig effektiv, spesielt hvis feilmeldingsfrekvensen er relativt høy. Det tyder på at mulighetsområdet (min/max-verdier) er definert for snevert slik at for mange enheter faller ut på kontrollen.
En feil kan fanges opp av ulike kontroller og settet av kontroller kjøres simultant på datamaterialet. Det vil derfor være noe vanskelig å skille de mest effektive kontrollene fra de mindre effektive.

6.1.3. Effekten av revisjon

Sammenlikne før og etter revisjon

- Sammenlikne statistikkdata og rådata.

Totaleffekten av revisjon kan måles som :

$$\frac{|\text{Statistikkdata} - \text{Rådata}|}{\text{Statistikkdata}}$$

For å studere effekten av revisjonen bør en se på tallverdien av avviket mellom statistikkdata og rådata og se på effekten av både positive og negative avvik. Positiv verdi betyr at de endelige statistikkdata er blitt endret til en høyere verdi enn de opprinnelige og omvendt. En klar overvekt av enten positive eller negative avvik tyder på skjevheter enten i det innkomne datamaterialet (f.eks. avvik i forhold til definisjon) eller i kontroll- og opprettingsrutinene.

Det er viktig å være klar over at statistikkdata ikke nødvendigvis er de korrekte verdier (fasit).

En effektiv metode for å studere enkeltobservasjoners betydning for totalresultatet er å kjøre ut kumulativ fordeling av avvikene for hver enhet sortert etter avtakende størrelse. En bratt stigende kurve med avflating betyr at noen få enheter betyr mye for totalresultatet. Det er viktig at disseenhetene fanges opp tidlig i kontrollrutinene og rettes.
Kurve med 45° helling betyr at alle opprettede enheter betyr like mye. Dette er sjelden tilfelle i praksis.

For en mest mulig omfattende vurdering bør slike indikatorer beregnes for de ulike variable og kontroller.

- Parallell revisjon og koding

Parallell koding og revisjon

En teknikk som kan brukes for kvalitetssikring er å la to grupper revidere og kode det samme materiale enten ved å bruke de samme metoder eller to forskjellige metoder. Slike gjentakelser av prosessen gjøres meget sjeldent. Det skyldes at kostnadene er høye ved slikt dobbeltarbeid, både i form av tid og ressurser. Dette kan derfor ikke anbefales som en permanent ordning, men kan være aktuelt som enkeltstående tiltak for å teste ut revisjonsarbeidet.
I forbindelse med store tellinger har denne teknikken blitt brukt under navnet "acceptance sampling". En kontrollerer her arbeidet løpende ved å trekke et tilfeldig utvalg av et kodet materiale og foretar kodingen på nytt. Dersom de feilkodinger en finner utgjør en andel som er større enn en på forhånd bestemt prosentandel, legges alt tilbake til omkoding. Størrelsen på utvalget samt hvor mye feil en skal akseptere, bestemmer den endelige kvaliteten på kodingen. Ofte starter en opp med høye ambisjoner for deretter å redusere disse for å få arbeidet gjort innen den fastsatte tidsramme.
I tillegg til å kontrollere kvaliteten til en bestemt undersøkelse gir slike opplegg betydelig innsikt i kodeprosessen og kvaliteten til kodeinstruksen og opplæring av koderne. Slik innsikt er også nyttig for opplegget av andre statistiske undersøkelser.

6.2. Lagring og dokumentasjon av data og revisjonsprosessen

For evaluering og kvalitetssikring av hele produksjonsprosessen er det nødvendig å dokumentere og lagre informasjon om data og revisjonsprosessen. Sentrale elementer er dokumenterte datafiler og revisjonsinstrukser.

6.2.1. Data og metadata

For å beregne effekter av revisjon ved å sammenlikne innkomne data og ferdigbearbeidete data er det nødvendig å lagre både:

- rådata (opprinnelig primærdata) og bruttoutvalget av enheter
- statistikkdata (data for publisering) for nettoutvalget

For enheter som er med i bruttoutvalget, men ikke i nettoutvalget, bør årsaken framgå, som opphør, feil ved uttrekk (skulle ikke vært med), nekting, kommet inn for seint mv. Tilsvarende må det framgå årsaken til at enheter i nettoutvalget ikke er med i det opprinnelige bruttoutvalget, tilganger av ulike slag.

Rådata må også lagres,

men

Det vil være betydelige forskjeller mellom de ulike datasett hvor stort avvik det er mellom rådata og statistikkdata, både avvik i totaltallene og antall rettede dataelementer. Det må derfor vurderes i hvert tilfelle om både rådata og statistikkdata skal lagres fullt ut (stor grad av dobbeltlagring) eller om rådata og statistikkdata bare skal lagres når de er forskjellig; dvs. bare korrigerte felt.

hva er rådata ?

Hva er rådata ?
I en del tilfelle kan det være noe uklart hva som skal defineres som «rådata».

- Skjemabaserte undersøkelser

ved papirbasert innhenting

I skjemabaserte undersøkelser er en del innkomne oppgaver av svært variabel kvalitet. Noen er helt eller delvis ubesvart og andre inneholder store feil som bl.a. skyldes at oppgavegiverne har misforstått oppgaven. Slike feil bør primært fanges opp i mottakskontrollen og returneres til oppgavegiver for oppretting. Hvis omfanget av mangelfulle oppgaver er betydelig, bør dette registreres på en eller annen måte med sikte på tiltak for å forbedre kvaliteten på innkomne oppgaver. Slike mangelfulle oppgaver bør likevel ikke betraktes som rådata og lagres i denne sammenheng. Med rådata menes de data som er blitt «godkjent» som innkomne og klar for videre bearbeiding. En slapp mottakskontroll vil kunne slippe gjennom for dårlige oppgaver.

og

For skjemabaserte undersøkelser er det nødvendig med spesielle registreringsrutiner for å lagre rådata. Metoden med interaktiv registrering og kontroll er ikke spesielt egnet for lagring av rådata, mens tradisjonell dataregistrering og optisk lesning tilfredstiller dette kravet.

- Elektronisk innhenting

ved elektronisk innhenting

  - For elektroniske skjemaer ligger allerede de mottatte dataene på maskinell form og det er forholdsvis uproblematisk med lagring av rådata. Også elektroniske skjemaer kan imidlertid være mangelfullt utfylt. Problemet vil likevel være klart mindre enn for papirskjemaer siden det bør være innebygde kontroller i utfyllingen av de elektroniske før de oversendes oss.

  - Ved direkte registrering og kontroll under datainnhenting (CAPI-/CATI-teknikk) blir feil svar stort sett rettet opp umiddelbart. Data som blir lagret etter at intervjuet er avsluttet må karakteriseres som rådata. Selv om en del etterkontroller kan resultere i korreksjoner, vil det likevel være mindre avvik mellom rådata og statistikkdata ved denne innhentingsmetoden.

  - For administrative data er de mottatte datafiler rådata og finnes allerede på maskinell form. Også for administrative data

- vil gjennomgående antall korrigerte felt være lavt i forhold til totalt antall datafelter.

Dokumentasjon av hvert datafelt
For lagring og dokumentasjon av statistikkdata finnes det allerede etablerte rutiner (se Datadok 98). For å tilfredsstille behovet for evaluering og dokumentasjon av hele produksjonsprosessen må det likevel stilles enkelte nye krav til dokumentasjon.

Dokumentasjon av rettede datafelt

For hvert datafelt må det framgå om feltet er :

- ukorrigert verdi (opprinnelig primærdata)
- manuelt korrigert verdi for feil oppgitt verdi
- manuelt fastsatt verdi for uoppgitt verdi
- maskinelt beregnet eller imputert verdi

Koding av type retting kan gjøres i ettertid på følgende måte: Det legges inn egen kode for datafelt som blir maskinelt beregnet eller imputert (under beregningen). Ved å matche statistikkdatafilen mot rådatafilen kan ukorrigerte data kodes for de felter der statistikkdata og rådata er identiske. Manuelle korrigeringer vil være felt som har avvik mellom rådatafilen og statistikkfilen samtidig som de ikke er blitt maskinelt beregnet. Manuelt fastsatt for uoppgitt verdi kodes for blanke felt i rådatafilen (forutsatt at blanke felt ikke godtas i statistikkfilen), mens manuelt korrigert for oppgitt verdi vil være felt som ikke er blanke i rådatafilen.

Det vil også være ønskelig med opplysninger om:
- hvilke feilkontroller som (hyppigst) er falt ut
- årsak til at opprinnelig primærdata er feil (årsakskode)

Det er likevel ikke krav om at slik informasjon blir lagret regelmessig, men det vil være nyttig med slik informasjon i en evaluering og revidering av produksjonsprosessen. Ved evaluering av de ulike feilkontroller kan antall feilmeldinger mv. simuleres på datamaterialat i ettertid. Det er ikke nødvendig at dette registreres simultant som en del av den ordinære produksjonsprosessen.

6.2.2. Dokumentasjon av revisjonprosessen

Dokumentér undersøkelsen for

- eget bruk (revisjonsintruks)

og

- eksterne brukere

I tillegg til lagring av data og dokumentasjon av datafiler må det for hver statistikk foreligge en fullstendig og samlet dokumentasjon av statistikken. Dette vil omfatte en dokumentasjon av revisjons- og kodingsprosessen og er det vi tradisjonelt mener med en revisjonsinstruks. Den tradisjonelle revisjonsinstruksen er først og fremst beregnet på internt bruk innen seksjonen for saksbehandlerne. og er forholdsvis teknisk med detaljert spesifikasjon av alle typer kontroller. Det er i tillegg behov for annen informasjon om statistikken; begreper, definisjoner mv. Slik informasjon finnes idag stort sett i tekstavsnittet i publikasjoner. I tillegg bør dokumentasjonen inneholde en mer kritisk vurdering av datamaterialet, både gode og dårlige egenskaper ved statistikken. For brukerne av statistikken er f.eks. svakheter eller problemer i populasjon/ utvalg,

kjennemerker, definisjoner mv. viktig og slik informasjon bør ikke gjemmes bort.

Hva skal dokumenteres ?

Dokumentasjonen bør ha en slik form at den er oversiktlig for brukerne av statistikken, både interne og eksterne. Dette kan f.eks. gjøres ved at de mest detaljerte og tekniske beskrivelser av kontroller gjøres i vedlegg. Dokumentasjonen ("revisjonsinstruksen") må inneholde en beskrivelse av de elementer som er nevnt under. Denne inndelingen er i samsvar med retningslinjene som foreligger for tekstdelen i NOS-publikasjoner og annen publisering på papir og web. En god del av denne informasjonen ligger i dag i Produktregisteret.
Det finnes også et skjema for dokumentasjon av utvalgsplan (populasjon, utvalg, frafall mv.) som ligger tilgjengelig som SSB-mal i Word.

1. Bakgrunn og formål
Hva kartlegger undersøkelsen, formål, hyppighet og viktige brukere. Det bør gis en emnekode for undersøkelsen (etter SSBs standard for emneinndeling).

2. Opplegg og gjennomføring

2.1 Omfang
- Populasjon
Det skal gis en beskrivelse av populasjonen (massen) som undersøkes, hvilke enheter undersøkelsen er ment å dekke, og hvilke avgrensninger som eventuelt er foretatt (næring, institusjonell sektor, størrelse, aldersgrupper mv.).
Det skal oppgis hvilket register eller annet grunnlag som er brukt for identifikasjon av populasjonen. Sentrale registre er Bedrifts- og foretaksregisteret og Personregisteret. Omtal svakheter eller mangler i registergrunnlag i forhold til ønsket populasjon.

- Statistisk enhet
Det skal gis en beskrivelse av de statistiske enheter som er basis, analyseenhet, observasjonsenhet, oppgaveenhet, rapporteringsenhet. Eksempel på enheter er bedrifter, foretak, kommune, person, husholdning.

2.2 Datakilder
Dersom det gjelder en registerbasert undersøkelse, identifiseres de registre som brukes, og det gis en kort beskrivelse av den type informasjon de inneholder. Svakheter eller mangler i registeret omtales også.

2.3 Utvalg av enheter
Det gis en beskrivelse av utvalgsplan. Den skal omtale hvilke typer enheter som omfattes, om det er totaltelling eller representativt utvalg, trekkemetode, antall enheter og dekningsgrad for brutto- og nettoutvalg.

2.4 Datainnsamling og -organisering.
Det gis en orientering om datainnsamlingen og innhentingsmetoder. Hvis det brukes andre datakilder enn registre, tas også med under dette avsnittet eventuelle kommentarer knyttet til spørreskjemaene. For undersøkelser der det er aktuelt, bør det også gis opplysninger om innsamlingsperiode,

(erstatning av uttrukne oppgavegivere), hvem som skulle svare innen en uttrukket husholdning eller bedrift o.l.

2.5 Kontroll- og revisjonsprosessen
Det redegjøres for mottak og godkjenning av oppgaver, revisjonsrutiner, kontakt med oppgavegivere mv. Ulike kontroller foretatt på materiale og metoder brukt skal spesifiseres.

2.6 Analyse-/estimeringsmetoder
Det redegjøres for opprettinger og imputeringer som er foretatt på rådataene og hvilke metoder som er brukt. Det redegjøres også for de analyse-/estimeringsmetoder som er brukt for å frambringe tallene som publiseres og de forutsetninger disse metodene bygger på (oppblåsninger til totaltall, indeksberegninger mv.).

3. Begreper, kjennemerker og grupperinger
Beskrivelse av hvilke kjennemerker som kartlegges og hvordan de er definert. Det er særlig viktig å presisere begreper som brukes med noe annet innhold i dagligtale, og fagtermer som fagmiljøet selv tar som en selvfølge. I visse tilfeller kan en henvisning til spørreskjema i vedlegg gi opplysninger om hvordan enkelte begreper og kjennemerker er definert. Spesielle problemer med å innhente data i samsvar med definisjonene omtales.
Det bør også gis en henvisning til hvilke standarder som brukes.

4. Feilkilder og usikkerhet
Eurostat er nå i gang med å utarbeide forslag til kvalitetsmål og kvalitetsrapportering for økonomisk statistikk. Disse rapporteringskravene ventes vedtatt i nær framtid og det vil være naturlig å tilpasse feilkilde-/usikkerhetsvurderingen til disse kravene. En del feil vil være vanskelige å estimere, men en kan i slike tilfeller foreta en kvalitativ vurdering.

4.1 Innsamlings- og bearbeidingsfeil
Både feilkilder som skyldes oppgavegiver, intervjuer, dataregistrering, koder eller andre forhold i bearbeidingen omtales. Også de feil som kan oppstå dersom register ikke helt dekker den populasjon som skal undersøkes, eller inneholder feilklassifiseringer eller andre direkte feil omtales. Videre kan det være grunn til å nevne faren for misvisende resultater som følge av måten oppgaver registreres eller grupperes på. Feil som skyldes forsinket oppdatering av registre, skal tas med under pkt. 4.3

4.2 Utvalgsvarians
For utvalgsundersøkelser bør i den grad det er mulig beregnes utvalgsvariansen for noen av de viktigste resultatene i undersøkelsen. Det må klart framgå hvilken metode som er brukt i beregningen av utvalgsvarians.

4.3 Utvalgsskjevhet/frafall
I utvalgsundersøkelser bør begrepet utvalgsskjevhet forklares, og dersom det er mulig, konkretiseres. Frafall er den viktigste kilden til skjevhet i utvalgsundersøkelser. Omfanget av frafall, metoder for å redusere frafall (purringer mv.) og metoder for å korrigere for frafall beskrives så presist som mulig. I registerundersøkelser kan forsinkelser i oppdatering anses som en form for frafall, og bør derfor vurderes i denne sammenhengen.

4.4 Kontroll og revisjon
En bør her vurdere påliteligheten av de revisjonsrutiner som benyttes og angi hvordan eventuelle feil i rutinene kan påvirke statistikken. Se kap. 6.1 for metoder for evaluering av revisjonsrutiner.

4.5 Annet
Andre feilkilder, svakheter eller mangler ved undersøkelsen

5. Lagring og formidling
Det redegjøres for hvilke data/-serier som lagres og hvordan data blir publisert og gjort tilgjengelig

Litteraturliste

Oversiktspublikasjoner:

- Australian Bureau of Statistics (1993), "Data Editing - An ABS-Manual"
Omfattende manual om hele revisjonsprosessen

- Danmarks Statistik (1997), "Feilsøgningskatalog"
Systematisk oversikt over kontrollmetoder og verktøy for editering og imputering

- Eurostat (1998), "Handbook on the design and implementation of Business surveys"
Omfattende manual om planlegging, gjennomføring og publisering av statistiske undersøkelser for foretakssektoren. Bygger på manual utarbeidet av Statistics Netherlands.

- Statistiska centralbyrån (1997), "Granska effektivt!"
Grundig om formålet med revisjon, datainnhenting/-registrering, kontrollmetoder og evaluering av revisjonsprosessen

- UNSO/ECE (1994), "Statistical Data Editing, Volume No.1: Methods and Techniques". Conference of European Statisticians, Statistical Standards and Studies - No.44
Publikasjonen inneholder en rekke grunnleggende artikler og fyldig litteraturliste. Temaer:
Review of statistical data editing methods and techniques
Macro-editing procedures
Implementation of data editing procedures
Bibliography

- UNSO/ECE (1997), "Statistical Data Editing, Volume No.2: Methods and Techniques". Conference of European Statisticians, Statistical Standards and Studies - No.48
Publikasjonen inneholder artikler under følgende temaer:
What to do when an edit fails
Designing sets of edits
Graphical editing
Evaluation of the data editing process
Impact of new technology on data editing
Automated coding
Glossary

Vedlegg A

Konsekvenser for IT-systemer og -verktøy

Datarevisjon i vid forstand omfatter flere faser i statistikkproduksjonen:
- mottak av data
- koding
- inspeksjon, kontroll, feilretting
- imputering
- kopling og lagring av data

Metodene i de ulike fasene skal være en del av en samlet revisjonsstrategi. Dette må også gjenspeiles i de tekniske løsningene. Informasjonsteknologien (IT) skal støtte opp under revisjonsarbeidet og de metodene som benyttes i de enkelte fasene. IT omfatter standard programvare, egenutviklede system, filer og databaser.

Målsettinger
Datarevisjonsmetodene skal bidra til å
- effektivisere datarevisjonen
- måle og høyne kvaliteten; som aktualitet og pålitelighet
- måle effekten av revisjonen
- synliggjøre problemer og feilkilder
- samordne statistikkproduksjonen generelt
- dokumentere prosessen, de enkelte fasene og resultatene
- standardisere bruken av metoder, løsninger og brukergrensesnitt

Konsekvenser for IT-systemene
Metodene skaper derfor føringer for,og stiller krav til, hvordan systemene utformes og brukes. IT-systemene skal bidra til å muliggjøre disse målsettingene
- gjennom økt bruk av selektiv revisjon
    - grafisk revisjon for å prioritere ressursbruken om de viktigste enhetene/ feilene (makrorevisjon)
    - mer bruk automatiserte rutiner for grenseverdier, oppretting/imputering
- ved å utnytte de muligheter for revisjon som nye rapporteringsløsninger kan innebære (f.eks. EDI og elektroniske skjemaer)
    - mulighetene for å legge inn kontroller hos oppgavegiver ved innlasting av data (summasjonskontroll, kryssreferanse, etc.)
- ved å utnytte de maskinelle revisjonsmulighetene som finnes i produktene for optisk lesing
- ved at systemutvikling/løsninger og verktøybruk bidrar til å sikre gjenbruk og samordning på tvers
- ved at lagringsstrukturer og datamodeller utformes slik at målsettingene kan realiseres
- ved at rådata og statistikkdata og endringer/opprettinger/resultat i de enkelte fasene må lagres og dokumenteres
- gjennom en standardisert og mest mulig kostnadseffektiv lagring og dokumentasjon. Det må utarbeides generelle retningslinjer for lagring og mellomlagring av data
    - hvor
    - hva
    - på hvilken måte (strukturer)

- ved å lagre nye kjennemerker (f.eks. e-post-adresser, faks-nr, kontakt-person, etc.)
- ved å strukturere arbeids- og oppgaveflyten
- ved at det skal være enkelt å kjøre ut tabeller, grafiske presentasjoner mv. fra produksjonsbasen for å skaffe seg god oversikt over og innsikt i datamaterialet
- ved å benytte alternativ til papirbaserte kontrollister (filer, baser,..)
- ved å sørge for at nye resultater etter opprettinger skal være enkelt å få fram.
- ved at feil som oppdages og rettes i en type program i en prosess ikke i tillegg må rettes av saksbehandler et annet sted (produksjonsbase)
- ved å foreta kontroll mot alternative eller andre kilder
- gjennom standardisering i ulike former for innenfor elektronisk datafangst
    - standard formater
    - standard program
    - standard kontrollrutiner

Konsekvenser for verktøyporteføljen
Oracle utviklingsproduktene (Designer/2000 og Developer/2000), Oracle DBHS og produktene fra SAS Institute utgjør de sentrale strategiske IT-produktene i SSB. Disse forventes å tilfredsstille de fleste krav som vil bli stilt i forbindelse med utvikling og anvendelse av nye revisjonsrutiner. På bekostning av optimale løsninger vil vi ofte måtte foreta en avveining mellom de krav som metodene stiller og de muligheter som IT-verktøy, IT-kompetanse eller IT-kapasitet kan tilby.

Samordnet bruk og videretilpasning av disse verktøyene innenfor produksjons- og revisjonsrutiner og dokumentasjon, er prioriterte oppgaver.

Både ved bruk av SAS-program og Oracle-verktøyene er det uansett behov for standardisering av opplegg og rutiner for bruk i ulike undersøkelser (moduler). Ut fra forskjellige forutsetninger må en forvente forskjellige måter å utforme produksjonen, inklusive revisjon, på.

Blaise er et viktig verktøy som fortsatt vil bli benyttet på spesielle områder. Fame er et annet sentralt produkt i SSB, men perifert i datarevisjonssammenheng.

SSB har allerede høstet en del erfaring med bruk av nye revisjonsmetoder og systemer utviklet for disse. Erfaring fra utvikling og bruk av disse må tas med i det videre arbeidet. Likevel må en i neste omgang utvide erfaringsgrunnlaget ved bevisst valg og gjennomføring av pilot-prosjekt ("case") for å kunne avklare mulige nye krav som IT-verktøy og -løsninger må tilfredsstille.

Støtteverktøy
En må også arbeide videre med andre støtteverktøy og -teknologier. I flere statistikkprodukter i SSB benyttes allerede i dag "billedhåndtering" ("imaging") ved optisk lesing og/eller i revisjonsarbeidet. Bruk av innscannede skjemaer i revisjonsarbeidet vil i mange tilfelle ha gunstige/positive konsekvenser:
- raskere gjenfinning av det enkelte skjema
- lettere å sortere og dermed prioritere behandlingsrekkefølgen
- enklere å fordele arbeidsoppgaver til spesialt-grupper (forskjellige næringer etc.)
- letter samhandlingen på tvers
- samtidig bruk av skjemaene

Ved utprøving av nye revisjons-/produksjonsprosesser bør en se på mulige anvendelsesområder for arbeidsflytsystemer ("workflow").
I en senere fase må det vurderes om en skal arbeide videre med neurale nettverk og "kunstig intelligens".

Andre revisjonsverktøy
I andre statistikkbyråer er det tatt i bruk spesialutviklet, generell programvare for revisjon. Mest kjent er kanskje GEIS (General Edit and Imputation System fra Statistics Canada). GEIS er både et feilsøkings- og imputeringsprogram.
Slike produkter som har bred internasjonal anerkjennelse og utbredelse må vurderes seriøst av SSB. Gode produkter er i seg selv et incitament for å ta dem i bruk. Videre er det viktig som et ledd i internasjonal samordning

Vedlegg B

Oversikt over datafangstmetoder og kontrollrutiner

Tabellen gir en oversikt over de viktigste datafangsmetoder og de ulike faser i kontrollrutinene. Dette er likevel bare en summarisk oversikt og den konkrete revisjonsprosessen må utformes spesielt for hver enkelt statistikk. En avgjørende faktor er kvaliteten på de mottatte dataene, men også brukernes behov (aktualitet, pålitelighet, detaljeringsnivå).

Ulike metoder for oppretting av datamaterialet (manuelle/maskinelle rutiner) vil stort sett være uavhengig av selve datafangsten. Metoder for oppretting vil til en viss grad være avhengig av type undersøkelse (personstatistikk, bedriftsstatistikk, korttidsstatistikk, strukturstatistikk, engangs-undersøkelser), men de fleste metoder for oppretting vil likevel være aktuelle for de forskjellige undersøkelsene.

	Skjemabasert innhenting			**Elektronisk innhenting**		
Rutiner/Metoder	**Konvensjonell dataregistrering**	**Interaktiv dataregistrering**	**Optisk lesing**	**Administrative data**	**Elektroniske skjemaer**	**CATI/CAPI**
Enkel mottakskontroll på mikrodata ?	Ønskelig	Ønskelig	Ønskelig	Nei	Bør være unødvendig	Bør være unødvendig
Absolutte kontroller på mikronivå	Etter registrering	Under registrering	Etter lesing/ reg., evt. under lesing	Etter mottak	Bør være unødvendig (avhengig av data-kvalitet og kontroll-opplegg)	Bør være unødvendig (avhengig av data-kvalitet og kontroll-opplegg)
Vurderingskontroller på mikronivå - Konsistenskontroller - Intervallkontroller	Etter absolutt kontroll, evt. sammen med makrokontroll	Etter absolutt kontroll, evt. sammen med makrokontroll	Etter absolutt kontroll, evt. sammen med makrokontroll	Etter absolutt kontroll, kombinert med makrokontroll	Etter mottak, kombinert med makrokontroll	Etter registrering, kombinert med makrokontroll
Feilsøking på makronivå - Selektiv revisjon - Top-down revisjon - Grafisk revisjon	Etter, evt. sammen med, mikrokontroll	Etter, evt. sammen med, mikrokontroll	Etter, evt. sammen med, mikrokontroll	Etter absolutt kontroll, kombinert med mikrokontroll	Etter mottak, kombinert med mikrokontroll	Etter registrering, kombinert med mikrokontroll

Vedlegg C

Ordliste, begreper

Absolutt kontroll Kontroll som identifiserer verdier i datasettet som med sikkerhet er feil, f.eks. ugyldig verdi eller sumfeil

Administrative data Data innhentet av andre etater for administrative formål

Aggregert metode Generell feilsøkingsmetode i to trinn: Først makrokontroll for å identifisere mistenkelige verdier, dernest mikrokontroll av de mistenkelige verdiene

Automatisk oppretting Maskinelle korreksjoner, uten innblanding av saksbehandler, foretatt på grunnlag av på forhånd fastlagte regler. Dette kan enten være logiske sammenhenger i skjema, basert på data fra samme enhet i foregående undersøkelse eller fra andre enheter i samme undersøkelse; f.eks. gjennomsnittverdier

Cold-deck imputering Imputeringsmetode som beregner manglende verdier på grunnlag av tilgjengelige data kjent på forhånd, f.eks. for samme enhet fra foregående undersøkelse

Deduktiv imputering Imputeringsmetode som beregner manglende verdier ut fra logiske regler fastlagt på forhånd, også kalt deterministisk imputering

Diskret variabel Variabel som har et begrenset sett av gyldige verdier, f.eks. kjønn, næringskode (kvalitativ variabel)

Dublettkontroll Kontrollerer om enheten/observasjonen bare finnes en gang i datasettet

Editering Kontroll av data for å oppdage feil og korrigere feil samt imputere manglende verdier

Ekstremverdi (outlier) Dataverdi som avviker betydelig fra andre verdier i datasettet og som kan mistenkes for å være feil (utligger). Det må kontrolleres om verdien er riktig eller feil og evt. må rettes. Ved imputering av manglende data for andre enheter vil det alltid være en vurdering om en riktig ekstremverdi skal inngå i korreksjonsgrunnlaget eller ikke.

Estimering Beregne verdier for manglende data. Gjøres vanligvis maskinelt på grunnlag av en statistisk modell.

Frafall *Enhetsfrafall:* Enheter i undersøkelsesbestanden som mangler fullstendig av en eller annen grunn (nekting, opphør mv.)

Partielt frafall: Enheter med i undersøkelsesbestanden, men enkelte opplysninger mangler (ufullstendig oppgave)

Grafisk revisjon Bruk av grafiske presentasjoner for å få oversikt over datamaterialet, bl.a. avvikende verdier. Slike diagrammer kan være interaktive slik at en både kan identifisere og rette enheter og automatisk få fram ny presentasjon av datamaterialet etter oppretting.

Grenseverdier	Øvre og nedre grense en verdi må ligge innenfor for ikke å indikere mulig feil i en intervallkontroll. Grenseverdier kan fastsettes manuelt eller maskinelt.
Hidiroglou-Berthelot-metoden	Maskinell statistisk feilsøkingsprosedyre basert på egenskapene til dataene. Tar hensyn til både relativ og absolutt endring i dataverdi fra foregående periode. Utviklet av Statistics Canada
Hot-deck imputering	Imputeringsmetode som beregner manglende verdier på grunnlag av data som blir oppdatert under undersøkelsen, f.eks. data for andre enheter i samme undersøkelse (donorimputering)
Identifikasjonsdata	Variable som identifiserer enheten, f.eks. personnummer eller organsisasjonsnummer
Imputering	Maskinell prosedyre for å erstatte en manglende verdi med en akseptabel verdi etter fastlagte regler, uten å ta kontakt med oppgavegiver.
Innligger	Dataverdi som ikke avviker særlig fra gjennomsnittet, men som er feil. Kan ha betydelig påvirkning på totalresultatet, men kan være vanskelig å oppdage
Interaktiv revisjon	Data blir kontrollert og samtidig rettet opp av saksbehandler direkte på dataskjerm. Dette er mest vanlig ved innregistrering av data og for kontroll av enkeltenheter
Intervallkontroll	Kontrollerer om en verdi ligger innenfor visse fastsatte grenseverdier
Iterativ prosess	Prosess som gjentas og som brukes ved kontroll av samlet materiale: Oppretting av de viktigste feilene først, dernest nye kontroll- og retterunder inntil datamaterialet er akseptabelt
Kjennemerke	Egenskap ved den statistiske enheten, f.eks. en persons alder eller et foretaks omsetning.
Koding	Gruppering av data etter en kodeliste (standard) og påføring av gyldig verdi
Konsistenskontroll	Kontroll av logiske sammenhenger mellom to eller flere variable
Kontinuerlig variabel	Variabel som kan ha et ubegrenset sett av verdier, f.eks. omsetning (kvantitativ variabel)
Korreksjon	Endring av verdi som er feil eller antatt å være feil med en riktig verdi eller antatt riktigere verdi.
Kvalitativ variabel	Variabel som beskriver egenskaper ved en enhet. Egenskapene er kategorisert i klasser. Klassene kan ha numeriske verdier, men det har ingen mening å regne på disse verdiene, f.eks. fylkeskode.
Kvantitativ variabel	Variabel som uttrykker en viss mengde, målt i en fysisk enhet, eller beløp for en enhet. Det har mening å regne på kvantitative variable, f.eks. gjennomsnittlig høyde på soldater.
Makrodata	Mikrodata aggregrert (summert, slått sammen), f.eks. til celler i tabeller for publisering

Makrokontroll	Kontroll på samlet materiale for å identifisere individuelle feil
Manuell oppretting	Saksbehandler setter selv inn verdi for feil eller manglende data på grunnlag av regler og egen vurdering, evt. i samråd med oppgavegiver.
Metadata	Informasjon om dataene
Mikrodata	Informasjon for den minste statistiske enheten, som regel person/husholdning eller bedrift/foretak. Opplysningene om enheten kan foreligge med full identifikasjon (navn, adresse), avidentifisert eller anonymisert
Mikrokontroll	Kontroll av individuelle enheter enkeltvis, ofte kjennemerke for kjennemerke
Neurale nettverk	Statistisk modell som kan avsløre sammenhenger i data som ellers er vanskelig å beskrive og som må betegnes som atypisk når en tar hensyn til mange variable under ett. Nettverket kan brukes til imputering på manglende og feil dataverdier
Observasjon	Det bestemte kjennemerket, verdien av en variabel eller egenskap, for en enhet i undersøkelsen
Optisk lesing	Skjema blir maskinelt "fotografert" og dataene lagt direkte på fil
Paneldata	Tidsseriedata for samme enhet for flere perioder (måned, kvartal, år)
Rateestimering	Estimeringsmetode som baserer seg på at forholdet mellom to variable i utvalget er det samme som for resten av populasjonen
Regresjonsestimering	Estimeringsmetode som baserer seg på at avhengighetsforholdet mellom to eller flere variable i utvalget også kan overføres til resten av populasjonen
Respondent	Oppgavegiver. Den som svarer på undersøkelsene, fyller ut skjemaene
Revisjon	I vid forstand: Klargjøring av mottatte data fram til de er klare for publisering Dette omfatter mottak av data, koding av data i henhold til ulike standarder, kontroll og feilretting, imputering, lagring og dokumentasjon av data
Rådata	Opprinnelige data mottatt fra oppgavegiver, før kontroll og oppretting starter
Selektiv revisjon	Makrokontrollrutine som identifiserer verdier/feil som har stor innvirkning på totalresultatet. Revisjonen blir konsentrert om disse feilmeldingene
Statistikkdata	Bearbeidete data klare for publisering etter at kontroll og oppretting er avsluttet
Statistisk enhet	Observasjonsenheten er den enheten vi skal kartlegge egenskaper ved, svært ofte person eller bedrift. Rapporteringsenheten kan være forskjellig fra observasjonsenheten, f.eks. kan et foretak rapportere for alle bedriftene i foretaket og person kan rapportere for husholdningen samlet
Systematisk feil	Gjennomgående feil for enhetene i undersøkelsen. Dette kan skyldes misforståelse f.eks. blant oppgavegiverne under utfylling eller ved koding/registrering av data.

Top-down metode	Makrokontrollrutine der en kontrollerer de viktigste enhetene først; f.eks. de største enhetene i absolutt verdi eller de enhetene med størst endring
Validitetskontroll	Kontrollerer om verdien er gyldig i henhold til en fastsatt liste (standard)
Variabel	Egenskap vi kartlegger ved de statistiske enhetene i en undersøkelse, f.eks. alder eller omsetning.
Veiing	Hver observasjon/enhet blir tillagt vekt slik at nettoutvalget blir representativt for hele populasjonen
Vurderingskontroll	Kontroll av verdier eller sammenhenger mellom verdier som virker mistenkelige, men som kan være riktige

De sist utgitte publikasjonene i serien Statistisk sentralbyrås håndbøker

45 Håndbok i datasikkerhet og fysisk sikring. 1994. 53s.

46 Telefonkatalog. 1998. 89s.

47 EØS-avtalen. Det statistiske samarbeid og konsekvenser for Statistisk sentralbyrås statistikkproduksjon. 1994. 55s.

48 Håndbok i tilsettingssaker. 1994. 32s.

49 Oppgaveplikt og tvangsmulkt. 1995. 55s.

50 Emneinndeling 1995. 1995. 43s.

51 Intervju: EDB-arbeidsbok. 1995.

52 Intervju: EDB-oppslagsbok. 1995.

53 Intervju: Opplæring og administrasjon. 1995.

54 Internkontroll: Revidert utgave 1997. 25s.

55 Nordisk statistikk på CD-ROM: Veiledning. 20s.

56 PC-Axis versjon 2.2: Brukerhåndbok. 69s.

57 Produktregister versjon 4.0: Brukerveiledning. 49s.

58 Håndbok i prosjektstyring. 20s.

59 Personalreglement for Statistisk sentralbyrå. 22s.

60 Produktnummerkatalog pr. 28.02.1996. 55s.

61 Innkjøpshåndbok. 1996.

62 Timeplan versjon 3.0: Brukerveiledning. 16s.

63 Håndbok i EDB-metode. 52s.

64 Publiseringshåndbok: Regler og retningslinjer for publisering i Statistisk sentralbyrå. 93s.

65 Håndbok i utvikling av statistikksystemer: Med vekt på IT-metode. 52s.

66 Håndbok i datarevisjon. 48s.

B *Returadresse:*
Statistisk sentralbyrå
N-2225 Kongsvinger

Statistisk sentralbyrå

Oslo:
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger:
2225 Kongsvinger

Telefon: 62 88 50 00
Telefaks: 62 88 50 30

Statistisk sentralbyrå
Statistics Norway