Statistics Norway
Statistisk sentralbyrå

Johan Heldal og Johan Fosen

Statistisk konfidensialitet i SSB

Et diskusjonsnotat

Notater

1876 • 2001 125 år

Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Innhold

1. Bakgrunn

Gjennom hele det 20. århundre har det vært en jevnt stigende etterspørsel etter statistisk informasjon for både offentlig og privat bruk. Denne utviklingen har gått hånd i hånd med utviklingen av et stadig mer komplekst samfunn hvor både den offentlige og private oppmerksomhet har trengt stadig dypere inn i samfunnets detaljer. Den økte etterspørselen er blitt fulgt av en eksplosiv økning i både volum og detalj av statistiske tabeller som produseres for offentlig planlegging og generell informasjon.

Statistikkbrukere i tidligere perioder var henvist til å ta til takke med ferdiglagede tabeller fra statistikkprodusenten. Men fremveksten av raske datamaskiner har ført til at det ikke lenger er noen teknologiske grunner til at statistikkbrukerne skal behøve å nøye seg med dette som den eneste form for formidling av statistisk informasjon. Derfor har etterspørselen etter detaljerte mikrodata, både fra offentlige myndigheter og fra det vitenskapelige samfunn, eksplodert. I tillegg har markedet for spesialproduserte tabeller økt, og ettersom disse tabellene ønskes til et bestemt formål, er ofte detaljeringsnivået mye større enn i tabeller som skal dekke mer allmenne behov.

Denne prosessen fører til en økende fare for avsløring av sensitiv informasjon om enkeltindivider, husholdninger og bedrifter. Dette kan være til skade for dem det gjelder. Dessuten, i et demokratisk samfunn må en svært stor del av den statistiske datainnsamlingen være basert på frivillig deltakelse fra publikum, noe som fordrer tillit mellom publikum og de institusjoner og personer som står for innsamling, bearbeiding og analyse av innsamlede data. Muligheter for avsløring av oppgavegiveres identitet og følsomme karakteristika representerer en trussel mot slik tillit og kan i det verst tenkelige scenario ødelegge selve basisen for det arbeidet som gjøres i institusjoner og derved selve informasjonsbasisen til moderne demokratier.

Mange andre land har vesentlig større problemer med offentlighetens tillit til landets offisielle statistikkprodusent enn dem vi har i Norge. Dette kan være en av årsakene til svært høyt frafall i utvalgsundersøkelser som noen land erfarer. Disse landene har også implementert formelle regler for ivaretagelse av konfidensialitet ved publisering av statistiske data. Nederlands statistiske sentralbyrå (CBS) har derfor gått i bresjen i Europa ved å vie store ressurser til forskning omkring ivaretagelse av konfidensialitet i statistisk informasjon som gjøres tilgjengelig. CBS har også med støtte fra EU og i samarbeid med flere universiteter og statistiske institusjoner innen EU utviklet programpakken ARGUS som inneholder metoder for konfidensialitetssikring av statistiske data.

I USA reguleres spørsmål om statistisk konfidensialitet i spenningsfeltet mellom ”Privacy act” fra 1974 (med endringer 1993) og ”Freedom of Information Act” (FOIA, 1966 med senere tillegg). American Statistical Association, Committee on Privacy and Confidentiality, har utarbeidet definisjoner og retningslinjer for hvordan konfidensialitet og retten til informasjon skal ivaretas innen rammene av dette spenningsfeltet. Forskning omkring konfidensialitetsproblematikk har vært drevet aktivt de siste 15 årene. Programvare har vært utviklet, men ikke et så generelt program som ARGUS. En prototyp for fjerninnhenting av informasjon basert på konfidensielle data har vært under utvikling (Keller-McNulty og Unger 1998).

I Norge er det juridiske grunnlaget for å sikre statistisk konfidensialitet ivaretatt i statistikkloven. Statistikkloven § 2-4 pålegger SSBs medarbeidere og dem som mottar innhentede opplysninger taushetsplikt, § 2-5 regulerer hva opplysningene kan brukes til og § 2-6 regulerer hvordan opplysningene kan offentliggjøres.

§ 2-6. Offentliggjøring av opplysninger.

Opplysninger hentet inn etter fastsatt opplysningsplikt, eller som er gitt frivillig, skal ikke i noe fall offentliggjøres slik at de kan føres tilbake til oppgavegiver eller annen identifiserbar enkeltperson til skade for denne, eller til urimelig skade for denne dersom oppgavegiveren eller enkeltpersonen er et foretak.

Også lov om offentlighet i forvaltningen og lov om behandling av personopplysninger setter rammer for hva slags informasjon SSB kan gjøre tilgjengelig. En samlet oversikt over de lover, regler og konsesjonsbetingelser som gjelder finnes i Håndbok for datasikkerhet og fysisk sikring og etterfølgende notater og referater fra sikkerhetsutvalget.

Garantien for at de lovpålagte påbud overholdes, beror i Norge i stor grad på kontrakter og de sanksjoner mot misbruk som ligger i dem. Det arbeides også mye med intern datasikkerhet i SSB. I skrivende øyeblikk pågår prosjektet "Behandling av sensitive data i SSB" og prosjektet "Sikkerhet 2000". Det førstnevnte prosjektet er et kartleggingsarbeid av bruken av registre med sensitiv informasjon, noe som skal benyttes når SSB etter hvert skal legge all bruk av registre med sensitiv informasjon til et lukket nettverk. Det sistnevnte prosjektet har et bredere fokus og omhandler innføring av bedre teknologisk sikkerhetsarkitektur og forbedring av administrative rutiner for oppfølging av sikkerhetstiltak generelt i SSB.

Når det gjelder publisering og utlevering av mikrodatasett har det i Norge ikke utkrystallisert seg felles formelle regler for hvordan påbudet i §2.6 bør ivaretas og det har tidligere ikke tidligere foregått noen forskning på området. Likevel har det dannet seg en praksis i SSB for hva som kan tillates publisert i tabeller. Sikkerhetsutvalget gir rammekonsesjoner for hvordan søknader om utlevering av datasett direkte fra SSB skal behandles.

Utlevering av datasett til forskere skjer i stor grad via Norsk Samfunnsvitenskapelig Datatjeneste (NSD). En ny avtale om formidling av data mellom SSB og NSD ble undertegnet i 1999. NSD har også utviklet en praksis for hvor detaljerte mikrodata forskere kan få adgang til. Det finnes imidlertid et mye rikere arsenal av metoder som kan anvendes enn det som er i bruk i Norge i dag, og det er en av målsetningene med dette notatet å presentere noe av det arsenalet som foreligger.

Ulike måter å gjøre statistiske data tilgjengelig reiser ulike problemstillinger. Avsnitt 2 omtaler ulike måter å publisere statistisk informasjon og hva som primært skiller dem med hensyn på problemstillinger knyttet til konfidensialitet. Avsnitt 3 tar for seg en del grunnbegrep, blant annet hva en som skal forstår med begrepene "Statistisk avsløring" (disclosure) og identifiserende variable også kalt nøkkelvariable. Avsnitt 4 tar for seg skillet mellom statistisk informasjon basert på utvalg og informasjon basert på fullstendige tellinger eller registre. Avsnitt 5 definerer hva som betraktes som stiller spørsmål om hva som bør betraktes som sensitiv informasjon og hva som ikke behøver å betraktes som slik. Avsnitt 6 tar for seg ulike metoder for å konfidensialitetssikre statistisk informasjon, både metoder som er behandlet i ARGUS og metoder som ikke er implementert der. Det er klart at konfidensialitetssikring av statistisk informasjon alltid vil innebære tap av informasjon som må avveies mot den samfunnsmessige nytteverdien av den informasjon som ligger i data. Avsnitt 7 diskuterer dette tapet og hvordan det kan måles i ulike sammenhenger. Avsnitt 8 gir eksempler på bruk av programpakken ARGUS. Et viktig emne som ikke vil bli tatt opp i dette notatet er hvordan de enkelte metodene vil innvirke på mulighetene for å benytte ulike typer statistiske analyser.

Avsnitt 2, 5 og 8 er skrevet for et bredt publikum av interesserte i konfidensialitet, mens avsnitt 3, 4, 6 og 7 er ment for lesere med bakgrunn i matematisk statistikk, selv om deler av avsnitt 6 også vil være interessant for et bredere publikum.

2. Om ulike former for formidling av statistisk informasjon

Som nevnt innledningsvis finnes det i dag primært to former for formidling av statistisk informasjon,

- som aggregerte tabeller eller
- som filer med mikrodata fra utvalgsundersøkelser, tellinger eller registre.

Mikrodata uteleveres bare til statistisk bruk innen forskning og offentlig planlegging, jf stl § 2-5. I fremtiden vil vi nok også se nye former for publisering. Spesielt vil bruken av internett med mulighet for brukerne til interaktivt å hente ut tabeller med aggregerte data fra svært små egendefinerte subpopulasjoner, for eksempel ved bruk av kart, representere en ny form der data fra svært store datamasser kan kombineres samtidig. Det vil være en utfordring å sette de riktige grensene for hvor detaljert det skal være mulig å hente ut statistikk på denne måten.

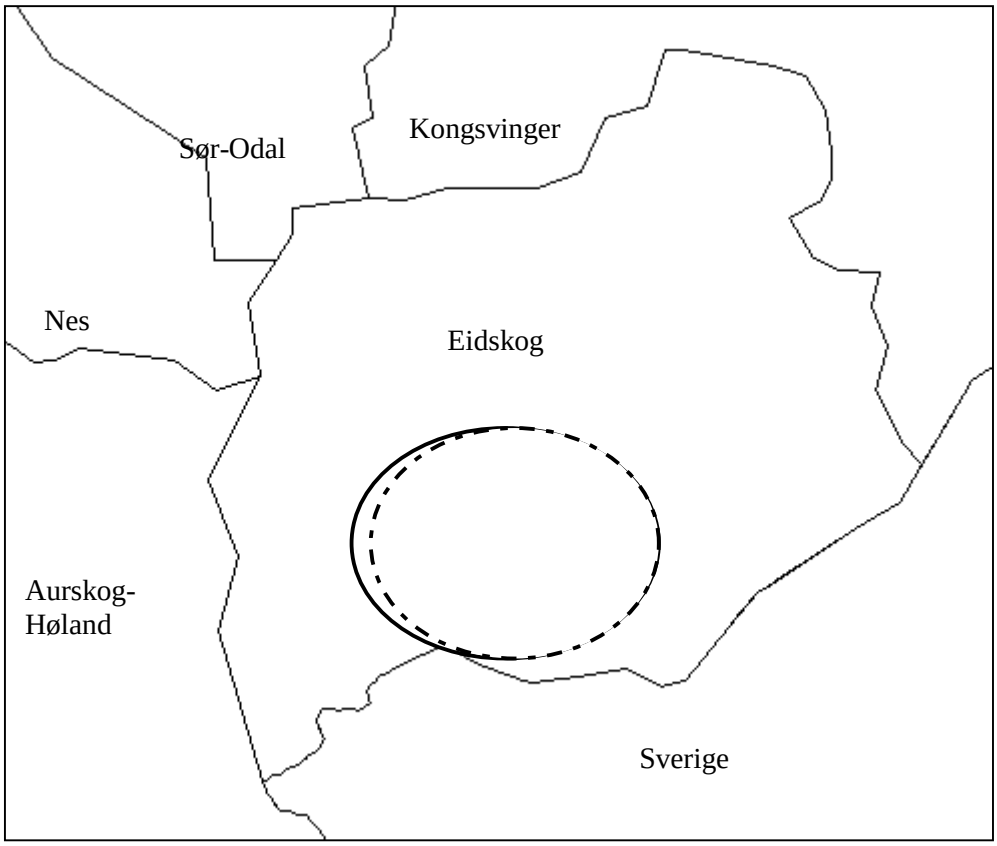
Det vil ikke være nok å begrense størrelsen på subpopulasjonen som brukeren kan velge når han velger på kartet. Figur 1 viser et kart over Eidskog kommune, og de to ellipsene illustrerer to subpopulasjoner som en bruker kan velge å få tabeller ut fra. Dersom personen etterpå sammenlikner tabellene fra de to områdene, kan han lage seg en tabell over den lille tredje subpopulasjonen som er de som kun dekkes av den ene ellipsen. Med litt jobb kan de to ellipsene plasseres slik at den tredje subpopulasjonen kun inneholder en håndfull eller kanskje til og med bare en person. En løsning på dette er å bare tillate brukeren å velge mellom på forhånd definerte områder, selv om dette gir mye mindre fleksibilitet.

Lov om offisiell statistikk og Statistisk Sentralbyrå, §1.2, definerer hva som skal forstås som offisiell statistikk.

§ 1-2. Definisjoner.

- **(1)** Statistikk er tallfestede opplysninger om en gruppe eller et fenomen, som fremkommer ved sammenstilling og bearbeiding av opplysninger om de enkelte enhetene i gruppen eller et utvalg av disse enhetene, eller ved systematisk observasjon av fenomenet.
- **(2)** Offisiell statistikk er statistikk som gjøres tilgjengelig for allmennheten av Statistisk sentralbyrå eller annet statlig organ.

Tabeller er aggregerte data og derved ”statistikk” etter denne definisjonen mens mikrodata ikke er det. Begge modi for formidling av statistisk informasjon kan imidlertid plages av problemet at individuelle oppgavegivere kan identifiseres i data på en måte som kan være til skade for denne. Men problemene kan være av noe forskjellig natur. Nedenfor vil vi diskutere de to modi separat.



Figur 1. Kart over Eidskog kommune i Hedmark, og to områder (markert ved to ellipser) som en bruker kunne tenkes å velge som subpopulasjoner i et tenkt program der man kan velge subpopulasjoner helt etter eget ønske, for deretter å få programmet til å produsere tabeller for akkurat disse subpoplasjonene.

2.1 Tabeller

Statistikk er tradisjonelt blitt formidlet i form av tabeller i publikasjoner fra SSB eller andre institusjoner som samler inn og legger til rette statistiske data. Man kan skille mellom to typer tabeller:

- i) Rene *kontingenstabeller* er ikke noe annet enn en opptelling av antall enheter i hver celle av en krysstabell basert på en eller flere kategoriske variable, hvorav noen kan være rene kategoriske (nominale eller ordinale) variable og noen kan være kategoriseringer av kontinuerlige (skala) variable. For eksempel kan en kontingenstabell vise antall individer i en befolkning eller i et utvalg etter kjønn, alder, ekteskapelig status og høyeste utdanningsnivå eller antall bedrifter i bestemte næringer etter fylke.
- ii) *Mengdetabeller* er tabeller hvor antall enheter i hver celle er erstattet av en aggregering av verdiene til en tredje variabel over de enhetene som er representert i cellen. Tabellen kan for eksempel vise den totale eller gjennomsnittlige inntekt til individene i de enkelte celler eller den samlede omsetning til bedriftene i hver celle. Bak en mengdetabell vil det alltid ligge en kontingenstabell som forteller hvor mange individer den tredje variabelen er aggregert over. Den bakenforliggende kontingenstabellen er av overordnet betydning for i hvilken grad mengdetabellen representerer et konfidensialitetsproblem.

I en kontingenstabell oppstår et konfidensialitetsproblem dersom det er mulig å identifisere individer/enheter entydig eller nesten entydig ved å kombinere *færre* enn alle de variable som krysstabuleres. Det vil gi mulighet for å trekke slutninger om den eller de øvrige variable. For eksempel, befolkningen i en kommune kan være krysstabulert etter kjønn, alder og ekteskapelig status og utdanningsnivå. Noen av kombinasjonene av de tre første variablene kan inneholde så få observasjoner at disse fordeler seg etter bare en eller kanskje to kategorier av utdanningsnivå. En leser som er i stand til å identifisere individene med de sjeldne kombinasjonene av kjønn, alder og ekteskapelig status vil også være i stand til å lære noe om disse individenes utdanningsnivå. Dette kalles en *direkte* eller *attributt avsløring* (direct/attribute disclosure). Tabell 1 viser situasjonen i et tenkt lite område, og vi ser at dersom en leser av tabellen på forhånd kjenner en ugift kvinne over 50 år, så ser han enkelt i tabellen hva slags utdanning hun har, nemlig lav utdanning, og vi har fått en attributtavsløring. Dersom de tre ugifte kvinnene over 50 år i stedet var fordelt med to på lav utdanning og en på høy utdanning, så ville en leser av tabellen som kjente de tre men manglet utdanningsinformasjon om to om dem, kunne lese dette av tabellen dersom den kvinnen han kjente utdanningen til hadde høy utdanning. Det sistnevnte tilfellet betegnes også attributtavsløring.

		Menn			Kvinner		
Sivilstatus	Utdanning	16-29 år	30-49 år	50 år +	16-29 år	30-49 år	50 år +
Ugift	Lav	6	7	5	7	5	3
	Middels	5	10	7	7	4	0
	Høy	0	6	3	0	4	0
Gift	Lav	5	10	13	4	8	11
	Middels	4	8	7	5	10	7
	Høy	3	7	4	3	7	3
Annet	Lav	0	5	13	0	6	18
	Middels	0	3	10	0	3	5
	Høy	0	4	3	0	3	3

Tabell 1: Antall innbyggere i et tenkt lite område etter sivilstatus, utdanning, kjønn og alder.

Ettersom utdanning for mange oppfattes som en følsom variabel, særlig hvis den er svært lav, vil personen som avsløres kunne føle dette plagsomt. I noen tilfeller kan det tenkes at informasjon som er tilegnet på denne måten blir brukt mot den personen det gjelder og derfor være til direkte skade.

På den annen side vil en tabell som kun gir mulighet til identifisering dersom verdiene til alle variablene i tabellen er kjente, ikke gi grunnlag for avsløring. For eksempel viser tabell 2 antall fødsler i et tenkt område etter morens alder. I cellen med *alder=14 år* ser vi at det er kun én fødsel. Dette er ikke en avsløring. Dersom noen skal kunne identifisere den 14 år gamle moren må denne allerede kjenne til moren og vite at hun er 14 år og at hun har fått barn. Det har i så fall funnet sted en *identitetsavsløring* (identity disclosure). Vedkommende vil derfor ikke ha lært noe nytt om moren, bortsett fra at hun kanskje er den eneste i sitt slag et bestemt år. Om det i seg selv er en sensitiv informasjon vil vi ikke ta stilling til her. Det har ikke funnet sted noen attributt avsløring.

Morens alder	Antall fødsler
14	1
15	25
16-19	75
20-24	150
25-29	221
30-34	134
35-39	101
40-44	43
45-49	12
50-54	3
55-59	0

Tabell 2. Antall fødsler etter morens alder

I en mengdetabell for omsetning etter næring kan en bedrift alene eller sammen med kanskje bare en eller to andre bedrifter representere en bestemt næring. Det vil da ofte kunne være rimelig lett å avsløre omsetningen til den aktuelle bedriften med tilstrekkelig presisjon til at det er interessant selv om den inngår i tabellen sammen med andre. Spesielt, hvis det er to bedrifter i samme celle etter næring vil den ene kunne regne ut omsetningen til den andre ved å subtrahere seg selv fra totalen. Dette kalles *residual avsløring* (Residual disclosure, Fellegi 1972). Fellegi betraktet residual avsløring i større generalitet og gav følgende definisjon:

Hvis en lineærkombinasjon av publiserte data resulterer i et tall som selv representerer en direkte avsløring, da er den publiserte statistikken en indirekte avsløring.

For å unngå residual avsløring er det vanlig i SSB å forlange at det skal ligge minst tre enheter bak hvert celleatotal som kan publiseres. I mange land er dette ikke betraktet som tilstrekkelig, og kravet kan f.eks. være at de tre største bidragsyterne til cellens total ikke skal utgjøre mer enn 80% av totalsummen i cellen. Dette innebærer at man ikke oppgir tall for f.eks. samlet omsetning selv om tallet er basert på fem bedrifter dersom én av bedriftene er så mye større enn de andre at celletotalen er tilnærmet lik omsetningen til den største bedriften. . Dette prinsippet kalles i litteraturen *(n,k) dominans regelen* (*the (n,k) dominance rule*): Undertrykk cellen hvis de n største enhetene i cellen utgjør mer enn k% av cellens total.

Vi kan f.eks. tenke oss at en bestemt celle skal inneholde omsetning for en bestemt næringskode i en bestemt kommune, og at dette er summen av omsetning for seks bedrifter som har omsetning på hhv 2,3,3,7,8 mens den siste bedriften har en ukjent omsetning Y. I denne situasjonen viser tabell 3 hvor stor omsetningen Y kan være før vi må undertrykke denne cellen, gitt ulike verdier av n og k. Vi ser f.eks. at dersom vi ikke vil at de to største bedriftene (vi velger n=2) skal utgjøre mer enn 75% av total omsetning (vi velger k=75%), så må den ukjente omsetningen Y være maksimalt 37. Dette kan vi sjekke ved at da er samlet omsetning for alle bedriftene 60[1], mens de to største til sammen er 37+8=45, noe som utgjør 75% av alle de seks bedriftenes omsetning.

Programpakken ARGUS anvender (n,k) dominans regelen. Det er mulig eksplisitt å formulere hvilket krav av denne typen man vil stille til tabellen for at en celletotal skal publiseres.

[1] De seks bedriftene har nå omsetning 2,3,3,7,8 og 37 og summen blir da 60.

	n		
k (%)	3	2	1
60	-	14,5	34,5
65	-	19,9	42,7
70	-	27,0	53,7
75	9,0	37,0	69,0
80	17,0	52,0	92,0
85	30,3	77,0	130,3
90	57,0	127,0	207,0
95	137,0	277,0	437,0

Tabell 3: Den maksimale størrelsen til omsetningen til en bestemt bedrift dersom cellen denne bedriften tilhører ikke skal måtte undertrykkes, når regelen for undertrykking er (n,k)-dominansregel (se beskrivelse i teksten), og flg situasjon antas: cellen skal inneholde samlet omsetning for den nevnte bedrift og fem andre bedrifter, der sistnevnte bedrifter har omsetningstallene 2,3,3,7 og 8.

Både kontingenstabeller og mengdetabeller kan opptre "koblet" til andre tabeller. At to tabeller er koblet vil si at de har en eller flere marginaler felles. Dette kan være den sensitive variabelen og/eller noen av dem som kan brukes til å identifisere individer. Et eksempel er dersom vi i tillegg til tabell 1 har en tabell som er identisk bortsett fra at utdanning er erstattet med arbeidsstyrkestatus (sysselsatt,ledig eller utenfor arbeidsstyrken). Vi kan for enkelhetsskyld anta at cellene får samme antall personer som i tabell 1. Fra tabell 1 kan vi se at det blant de yngste ugifte menn er seks personer med lav utdanning, fem med middels utdanning og ingen med høy utdanning. Fra tabellen med arbeidsstyrkestatus vil vi da få at det blant de yngste ugifte menn er seks personer utenfor arbeidsstyrken, fem arbeidsledige og ingen sysselsatte. Kjenner man noen av de seks mennene som er arbeidsledige vil man kunne se av tabellene at disse også har lav utdannining (og vise versa).

Som vi ser ovenfor vil to koblede tabeller holdt sammen gi større mulighet for å identifisere og foreta avsløring enn hver av de to tabellene alene. Mulighetene vil imidlertid være mindre enn om alle variablene i de to tabellene var krysset. Når vi utnytter koplingen av tabell 1 og den tenkte tabellen med arbeidsstyrkestatus i stedet for utdanning, vet vi fra tabell 1 at det er seks unge ugifte menn med lav utdanning, og fra den andre tabellen vet vi at ingen av de seks er sysselsatte. Imidlertid vet vi ikke hvor mange av de seks som er ledige og hvor mange som er utenfor arbeidsstyrken, men dette ville vi visst dersom vi hadde en tabell der alle variablene var krysset slik at vi hadde hatt tabellen *inbyggere etter sivilstatus, utdanning, kjønn, alder og arbeidsstyrkestatus*.

ARGUS har bare begrensede muligheter til å håndtere koblede tabeller. En mulighet er lagt inn, men betraktes i den nåværende utgaven av programmet som eksperimentell og garanteres ikke.

Mye av den tidlige litteraturen omkring konfidensialitet dreide seg naturlig nok mye om tabeller. Fellegi (1972) var den første som gav emnet en skikkelig formell behandling og utledet teoremer for identifiserbarhet. Dalenius (1977) presenterte en formell definisjon av begrepet "statistical disclosure" og gav en rekke eksempler på hvordan det kunne oppstå i tabeller.

2.2 Mikrodata

Mikrodata inneholder data på vesentlig mer detaljert nivå enn tabeller. I utgangspunktet inneholder mikrodata gjerne formelle identifikatorer som navn, adresse og personnummer i tillegg til de variable som har statistisk interesse. Formelle identifikatorer blir alltid strippet av datasettene før de stilles til disposisjon for brukere. De øvrige kjennemerker deles gjerne inn i "*sterkt identifiserende*", "*identifiserende*", "*ikke-identifiserende*" og "*sensitive*" (ARGUS). Grenseoppgangen mellom disse er skjønnsmessig. En variabel som er sensitiv kan i prinsippet også være identifiserende. En variabel som *bostedskommune* regnes gjerne som sterkt identifiserende. Slike variable strippes gjerne også fra datasettet eller gis en grovere inndeling (eks: bostedsfylke) før det gjøres tilgjengelig.

Likevel vil det i et mikrodatasett være store muligheter for å identifisere individer dersom en person med tilgang til datasettet har opplysninger om enkeltindivider som kan gjenfinnes i datasettet. Et sentralt begrep er *entydighet i populasjonen*. En statistisk enhet er entydig i en populasjon med hensyn til en gitt kombinasjon av variable dersom det bare finnes en enhet med en gitte kombinasjonen i populasjonen. En av de vanligste metodene for å redusere muligheten til å produsere entydige, er å slå sammen verdier eller kategorier i svært grove inndelinger. I litteraturen kalles dette *global omkoding*. Både SSB og NSD benytter i stor grad global omkoding før de frigir datasett til forskningsformål. Begrensning av antall tilgjengelige variable er en annen metode som benyttes.

2.3 Tabeller kontra mikrodata

Tabeller er aggregerte data mens mikrodata ikke er det. En annen forskjell mellom tabeller og mikrodata er at tabeller kan være utstyrt med marginaler mens mikrodata ikke er det. Ved siden av global omkoding, som også kan brukes direkte på tabellnivå og da kalles *tabellredesign* i litteraturen, er *prikking* (cell suppression) den vanligste måten å forhindre at sensitive opplysninger skal kunne avsløres gjennom tabeller. Når dette gjøres i tabeller som er utstyrt med marginaler, ønsker en ofte å publisere marginalene dersom disse ikke i seg selv er "farlige" fra et konfidensialitetssynspunkt. Dette har konsekvenser når enkeltceller i tabeller prikkes. For at innholdet av den prikkete cellen ikke skal kunne utledes av marginalene og de øvrige celletallene, må også andre celler, fortrinnsvis i samme rad og kolonne også prikkes. Dette kalles *sekundærprikking* (secondary suppression). Metoder for valg av celler til sekundærprikking vil bli omtalt i avsnitt 6.

Den metoden som motsvarer prikking i mikrodatasett kalles i litteraturen *lokal sensurering* (local suppression). Den går ut på å erstatte sensitive verdier i enkeltposter med uoppgitt (missing). Men mikrodata har ikke marginaler. Ved bruk av lokal sensurering vil en aggregert tabell med fullstendige celletall svarende til de undertrykte verdiene ikke kunne reproduseres korrekt fra mikrodata.

Dersom den aggregerte tabellen ikke er "sikker" vil det heller ikke være ønskelig å kunne gjennomføre en slik aggregering. På den annen side, heller ikke marginaler eller globale totaler som fra et konfidensialitetssynspunkt er "sikre" vil kunne reproduseres korrekt. En mulig løsning på dette er å publisere sikre korrekte marginaler som tillegg. En annen vil være å erstatte de riktige mikrodata for de enheter som tilhører celler som må undertrykkes (primært eller sekundært) med andre tall enn de korrekte, men på en slik måte at marginaler som kan publiseres blir ivaretatt.

Et problem med lokal sensurering vil også være hvilken variabels verdi bør undertrykkes. Hvis et datasett inneholder kombinasjonen

Bostedsfylke:	*Oslo*
Yrke:	*Bonde*
Kjønn:	*Kvinne*
Sivilstatus:	*Partner*

vil variabelverdien *Partner* kunne oppfattes som sensitiv. Ettersom kombinasjonen *Oslo, Bonde, Kvinne* er sjelden (hvis den i det hele tatt forekommer), vil denne kunne identifisere individet bak. Spørsmålet blir så hvilken variabels verdi bør undertrykkes? Bør det være sivilstatusverdien siden dette er den sensitive variabelen eller bør det være verdien til en av de tre variablene som kan være identifiserende? Hvis det er verdien *Partner* som skal sensureres må det også finnes andre individer i datasettet som mangler verdi på variabelen sivilstatus av andre grunner enn at de er sensitive. Hvis ikke vil en manglende verdi være synonym for den sensitive og man er like langt. Det er derfor mer rimelig å sensurere verdien til en av de identifiserende variablene. Det er da rimelig å søke å undertrykke verdien til den variabelen hvis verdi må antas å ha minst informasjonsmessig verdi. Hvilken det er må sees i sammenheng med spørsmålet om det også finnes andre kombinasjoner av variable som kan være identifiserende. Global omkoding kan her også være et alternativ. Hvis yrkeskoden som svarer til Bonde kan for eksempel reduseres til ettsiffernivå.

2.4 Eksempel

Tabell 4 viser et eksempel på en mikrofil, og tabell 5 viser hvordan en mulig tabell utfra denne mikrofilen kan se ut. For oversiktens skyld har vi kun fylt inn for de to cellene som gjelder de 12 personene vi ser i tabell 4. Ettersom det viser seg at det kun er to personer i den ene cellen, ønsker vi å undertrykke dette tallet.

Dersom vi gjennomfører tabellprikking, vil vi basert på tabell 5 prikke cellen med kun to personer, dvs. cellen med antall kvinnelige bønder som er registrert partner. Samtidig kan vi beholde marginalene 42 og 1 256. Riktignok måtte vi ha endret minst en til av cellene i kolonnen for kvinnelige partnere og minst en av cellene i raden for bønder, for ellers kan man regne ut hvor mange som er i cellen vi forsøker å skjule (residualavsløring).

Personident	Yrke	Sivilstatus	kjønn	…
⋮				
n	bonde	partner	kvinne	
n+1	bonde	partner	kvinne	
m	bonde	partner	mann	
⋮	⋮	⋮	⋮	
m+9	bonde	partner	mann	
⋮				
M				

Tabell 4: Utdrag fra en mikrofil

Yrke	I alt	Menn				Kvinner			
		…	partner	…	Missing	…	partner	…	Missing
Yrke	I alt	…	partner	…	Missing	…	partner	…	Missing
I alt	45 492		59		0		42		0
⋮	…	…	…	…	0	…	…	…	0
bonde	1 256	…	10	…	0	…	2	…	0
⋮	…	…	…	…	0	…	…	…	0

Tabell 5: Ikke-sikker tabell over innbyggere etter yrke, kjønn og sivilstatus i et tenkt område. Symbolet "…" indikerer flere celler.

I stedet for tabellprikking basert på originalt mikrodatasett, kunne vi tenke oss at det var gjennomført lokal undertrykking i mikrodatasettet først. Dersom vi undertrykket verdien "partner" for person *n* og *n+1*, ville vi ikke produsert tabell 5, men derimot tabell 6. Som vi ser blir marginalen feil for totalt antall kvinner som er registrert partnere (ved at vi får 40 i stedet for 42). I dette tilfellet er det i tillegg ingen som har manglende oppgitt sivilstand i utgangspunktet, og dersom en inntrenger vet både dette og at det er "partner" som er sensitiv verdi, har ikke den lokale undertrykkingen skjult noe som helst.

Yrke	I alt	Menn				Kvinner			
		…	partner	…	Missing	…	partner	…	Missing
I alt	45 492		59		0		40		2
⋮	…	…	…	…	0	…	…	…	0
bonde	1 256	…	10	…	0	…	0	…	2
⋮	…	…	…	…	0	…	…	…	0

Tabell 6 Innbyggere etter yrke, kjønn og sivilstatus i et tenkt område, men basert på mikrofil som er lokalt undertrykket. Symbolet "…" indikerer flere celler.

3. Grunnbegrep og scenarier

3.1 Grunnbegrep

I litteraturen har det forekommet flere definisjoner av begrepet "disclosure" som vi i de foregående avsnittene har oversatt med "avsløring". Den mest generelle definisjonen, som også kalles *inferentiell avsløring* (inferential disclosure), stammer fra Dalenius (1977):

> *If the release of the statistic S makes it possible to determine a (microdata) value D more accurately than is possible without access to S, then a disclosure has taken place.*

S kan her være en tabell med aggregerte data, en annen form for estimat som publiseres på makronivå, eller *S* kan være mikrodata fra et mikrodatasett. I prinsippet betyr definisjonen at enhver form for publisering av statistiske data vil representere en avsløring. Deler av littera-

turen tar utgangspunkt i denne definisjonen og formaliserer en bayesiansk tilnærming der den publiserte observatoren S brukes til å oppdatere en aprioriforfordeling $f_D(d)$ om verdien av et kjennemerke D til en sikrere aposteriorifordeling $f_D(d \mid S)$. Bruk av en slik metode krever at man har en mer eller mindre formalisert modell for populasjonen og at denne kan anvendes sammen med aprioriforfordelingen. Den krever ikke at individuelle statistiske enheter identifiseres for at man skal kunne snakke om en avsløring. Den krever heller ikke at en statistisk enhet skal ha vært med i det statistiske grunnlaget for S (f.eks. et utvalg) for at avsløring skal kunne finne sted. Basert på desisjonsteori utviklet Duncan & Lambert (1986, 1989) usikkerhetsmål $U(\cdot)$ for graden av avsløring for hver enkelt enhet i populasjonen. Dette avhenger av f_D og det "tap" $L(\delta, d)$ en "inntrenger" vil lide ved å tro at verdien av den interessante variabelen hos en interessant enhet er δ mens den riktige er d. Formelt kan U defineres som

$$U(f_D) = \min_\delta EL(\delta, d)$$

hvor forventningen er tatt over f_D før eller etter at S er frigitt. Observatoren S skulle bare kunne publiseres dersom $U(f_D(d \mid S)) > \tau$ for alle enheter i populasjonen der τ er en nedre grense for usikkerheten om de enkelte enhetens kjennetegn. Et typisk valg av L er $L(\delta, d) = (\delta - d)^2$ som gir at $U = Var(d)$, men andre tapsfunksjoner er også aktuelle.

Denne desisjonsteoretiske tankegangen har vært anvendt på en lang rekke scenarier, ikke minst for å studere hvordan ulike "inntrengere" med ulike typer motiver og dertil knyttede tapsfunksjoner og med ulik grad av kunnskap om enhetene i populasjonen vil agere forutsatt at de opptrer rasjonelt. Dalenius' definisjon ble i 1978 anbefalt som definisjon på "statistical disclosure" av "Subcomittee on Disclosure Avoidance Techniques" i USA's kongress.

En enklere form for "avsløring" er det som kalles *identitetsavsløring* (identity disclosure) og som ble nevnt i avsnitt 2. Som det fremgår av navnet er dette synonymt med at identiteten til en statistisk enhet i et datasett avsløres. I motsetning til inferensiell avsløring krever identitetsavsløring at den enhet hvis identitet avsløres må være med i det datasettet som publiseres eller er grunnlag for en tabell. Flere forfattere (Spruill 1983, Paas 1985, Strudler et. al 1986) betrakter identitetsavsløring som den formen for avsløring som det er viktig å sikre seg mot. Årsaken til at denne betraktes om viktig er at den ofte er en forutsetning for *attributtavsløring*, det å lære verdien av et kjennetegn som ikke var kjent fra før.

En person som prøver å avsløre noe ukjent om en statistisk enhet (person, husholdning, bedrift) på grunnlag av statistisk informasjon som han eller hun har tilgang til kalles i litteraturen for en *inntrenger* (intruder). I det scenariet som er oftest beskrevet i litteraturen har inntrengeren, i tillegg til de rent statistiske data, også informasjon om verdier av noen aktuelle variablene for en eller flere kjente enheter i populasjonen. Disse aktuelle variable finnes igjen i de statistiske dataene og betegnes gjerne med *identifiserende variable* eller *nøkkelvariable* (identifying variables/key variables). De enheter som inntrengeren kjenner slike verdier for kalles vekselvis *identifikasjonsfil* eller *bekjentskapskrets* (identification file/circle of aquaintancies). Det er ikke avgjørende at inntrengeren kjenner verdiene til nøkkelvariablene helt presist. Typiske identifiserende variable i Norge i dag er demografiske variable som kjønn, alder, ekteskapelig status og bostedskommune. Inntekt kan også regnes som en identifiserende variabel ettersom ligningen her i landet er offentlig tilgjengelig. I prinsippet kan ethvert kjennemerke som en mulig inntrenger kan kjenne verdien av (mer eller mindre sikkert) og legge inn i sin identifikasjonsfil, betraktes som en identifiserende variabel, selv om denne verdien ikke er direkte offentlig tilgjengelig.

Kjennemerker hvis verdi det statistiske byrået eller en statistisk enhet (person/bedrift) ikke ønsker at skal avsløres for utenverden, kalles *sensitive variable.* Det er verdien av sensitive variable som er objektet for forsøk på avsløring. Det er mer korrekt å si at det er visse verdier av variabelen heller enn variabelen selv som er sensitiv. Slike verdier vil vi kalle *sensitive verdier*. Eksempler på sensitive variable kan være seksuell legning og kriminell fortid. For variabelen "seksuell legning" betraktes "Heterofil" ikke som en sensitiv verdi, men andre verdier av variabelen vil være det. Inntekt oppfattes gjerne som sensitiv selv om den kan være offentlig tilgjengelig. Utdanningsnivå oppfattes noen ganger som sensitivt, særlig hvis det er lavt.

Det er ikke noe skarpt skille mellom identifiserende variable og sensitive variable. For eksempel kan utdanning og inntekt være både identifiserende og sensitiv. Videre kan de nye kategoriene av ekteskapelig status som går på partnerskap være sensitive idet de definerer en seksuell legning. Men i analysen av statistisk konfidensialitet vil det være den rolle variabelen spiller i en aktuell situasjon som er avgjørende. En inntrenger som benytter verdien til en sensitiv variabel som identifiserende har ikke lært den fra statistiske data. Derfor er det likevel legitimt å opprettholde et skarpt skille mellom identifiserende og sensitive variable i modeller som beskriver muligheter for avsløring og inntrengeres adferd.

3.2 Scenarier

I litteraturen om statistisk konfidensialitet tas det ofte utgangspunkt i eksempler som representerer ulike typer data med ulike statistiske egenskaper og inntrengere med ulik motivasjon og med ulike mengder kunnskap som utgangspunkt for å identifisere personer/-bedrifter. Disse eksemplene representerer forskjellige scenarier. Selv om situasjonen kan være ulike, kan den beskrives med en felles struktur:

- En inntrenger **I** sitter med en identifikasjonsfil/bekjentskapskrets og har tilgang til statistiske data.
- **I** forsøker å matche nøkkelvariablene i bekjentskapskretsen til de statistiske dataene for å identifisere dem.
- Hvis **I** lykkes med identifiseringen vil han/hun kunne lære noe mer om dem.

Eksempel: Hvis Per Hansen er i bekjentskapskretsen til **I** og **I** har påvist at en celle med en observasjon må være Per Hansen.

Strukturen i de tre punktene kan gi inntrykk av at **I** må være en målrettet person som i utgangspunktet har til hensikt å bruke statistiske data til å avsløre personer eller bedrifter. I deler av litteraturen fremstilles også **I** på denne måten. En (etter vår mening) mer realistisk situasjon er den hvor:

- Personen har en legitim bruk for de statistiske data.
- Personen oppdager noen særlig interessante enheter i data.
- Personen prøver å identifisere dem og blir derved en **I**.

Personen må da skaffe seg en "identifikasjonsfil". "Spontan" gjenkjenning kan også finne sted. De siste situasjonene unndrar seg imidlertid i noen grad en formalisert behandling, og det er derfor den beste tilnærming også å behandle disse situasjonene innenfor rammeverket ovenfor.

4. Populasjon og utvalg

Det er en ofte avgjørende forskjell mellom statistikk/mikrodata fra utvalg og statistikk/mikrodata fra altomfattende registre eller tellingssituasjoner. En enhet som er unik med hensyn på en verdi av en kryssklassifisering av flere kategoriske nøkkelvariable i et register eller et tellingsdatasett er *unik i populasjonen* med hensyn på en kombinasjon av disse variablene. Denne enheten vil kunne identifiseres og avsløring av verdien på andre variable vil kunne foretas med sikkerhet. På den annen side, en enhet som er unik med hensyn på en slik verdi i utvalget er ikke nødvendigvis unik i populasjonen. Det vil imidlertid da være et spørsmål om det er mulig å påvise at personen er unik også i populasjonen eller hvor mange av dem som er *unike i et utvalg* som også er det i populasjonen. I tilfeller der nøkkelvariablene er variable som også finnes i registre vil det være lett å kontrollere hvilke og hvor mange personer som er entydige i populasjonen og i et gitt utvalg. La U_p være antall entydige i populasjonen og la U_s være antall personer som er entydige både i populasjon på N individer og i et selvveiende utvalg på n individer. Da er $E(U_s) = nU_p / N$. Bethlehem & al. (1990) foreslår å bruke antall eller andel unike i utvalget som et kriterium for hvor detaljert informasjon som bør gjøres tilgjengelig ved å definere en absolutt eller relativ øvre grense C_a eller C_r og kreve at $U_s \le C_a$ eller $U_s / n \le C_r$.

I alminnelighet vil entydighet i et utvalg være mindre farlig enn entydighet i populasjonen. På den annen side, hvis en inntrenger vet at en enhet i hans/hennes identifikasjonsfil har svart i et utvalg som han/hun har tilgang til, vil entydighet i utvalget være like avslørende som entydighet i populasjonen. I norsk statistikk er entydige med hensyn på demografiske variable lette å identifisere i befolkningsstatistikken. Hvis vi betrakter variablene kjønn (K), alder (A), ekteskapelig status (S) og fylke (F) kan vi utlede tabell 7 fra befolkningsstatistikken pr. 1/1-2000 .

Med flere nøkkelvariable (f.eks. yrke utdanning, nasjonalitet) blir antall entydige i populasjonen fort stort. Antall entydige kombinasjoner i populasjonen er imidlertid ikke alltid like tilgjengelig. I forhold til et gitt utvalg vil det interessante være hvor mange kombinasjoner som er entydige i det utvalget og hvilke og hvor mange av disse som er entydige i populasjonen. Hvis svaret på det spørsmålet ikke kan gis ved en tabulering fra registre som ovenfor, må man søke å estimere sannsynlighetene for at en kombinasjon som er entydig i utvalget også skal være det i populasjonen. Det er etter hvert nedlagt mye forskning og publisert atskillige artikler for å utvikle gode metoder for å estimere dette. Nedenfor vil vi skissere de linjer som de fleste av disse arbeidene går langs og noen referanser.

Kombinasjoner	K	U_p
K×A(ett-årsgr. 16 år +)	2×95 = 190	6
K×A(ett-årsgr. 16-74 år)	2×59 = 118	0
K×A(ett-årsgr. 16 år +)×S[1]	2×95×5 = 940	32
K×A(ett-årsgr. 16-74 år)×S[1]	2×59×5 = 590	6
K×A(5-årsgr. 15 år +)×S[2]×F	2×16×6×19 = 3684	140

Tabell 7. Antall entydige kombinasjoner i den norske befolkning

[1] Ikke medregnet registrerte, gjenlevende, separerte eller skilte partnere.

[2] Registrerte, gjenlevende, separerte eller skilte partnere slått sammen til en kategori. Disse utgjør 81 av de 140.

Anta at vi har M kategoriske nøkkelvariable $X_1,\ldots,X_M$. Disse har $K_1,\ldots,K_M$ kategorier og kan kryssklassifiseres i $K \le K_1K_2\cdots K_M$ gyldige kategorier som vi nummererer med $k=1,\ldots,K$.

La

$$F_k = \#\text{ganger kombinasjon } k \text{ forekommer i populasjonen,}$$
$$f_k = \#\text{ganger kombinasjon } k \text{ forekommer i et utvalg, og}$$
$$I_k = \begin{cases} 1 \text{ hvis } F_k = 1 \\ 0 \text{ ellers} \end{cases}$$

$F_k = I_k = 1$ betyr da at kategori k representerer en unik i populasjonen og kan representere et problem. $F_k = 2,3$ eller 4 kan også representere et problem dersom en ønsker at inntrengeren skal ha stor usikkerhet med hensyn på om enheten i identifikasjonsfilen er den samme som i utvalget, eventuelt også dersom enhetene i kategori k er svært like med hensyn på verdien av en sensitiv variabel. Generelt kan man definere en kombinasjon k som usikker dersom $\boldsymbol{F_k \le c}$ der c er en på forhånd valgt grense. Noen ganger brukes begrepet *MINimum Unsafe Combination* (MINUC) når det er behov for å formalisere problemstillingen. En verdi k av en kombinasjon av m variable er MINUC (av orden m) dersom den er usikker (har $\boldsymbol{F_k \le c}$), men ingen kombinasjon av noen undergruppe på *m - 1* av variablene har en usikker verdi (dvs. alle har $\boldsymbol{F_k > c}$).

Spørsmålet om evidens for entydighet i populasjonen for en kombinasjon som er entydig i utvalget kan nå formuleres som problemet å estimere

$$P(F_k = 1 \mid f_k = 1) = \frac{P(F_k = 1 \cap f_k = 1)}{P(f_k = 1)} = \frac{P(f_k = 1 \mid F_k = 1)P(F_k = 1)}{\sum_x P(f_k = 1 \mid F_k = x)P(F_k = x)}$$

Her vil $P(f_k = 1 \mid F_k = x)$ avhenge av utvalgsdesignen og vil derfor ofte være rimelig lett å beregne. Estimater for sannsynlighetsfordelingen $P(F_k = x)$ vil imidlertid kreve modellering av populasjonen. Den vanligste tilnærmingsmåten tar utgangspunkt i å betrakte F_k - ene som uavhengige poissonfordelte:

$$P(F_k = x) = \frac{\lambda_k{}^x}{x!}\exp(-\lambda_k)$$

Antallet entydige i populasjonen, $U_p = \sum_{k=1}^{N} I_k$ vil med en slik modell ha forventning

$$EU_p = \sum_{k=1}^{K} EI_k = \sum_{k=1}^{K} P(F_k = 1) = \sum_{k=1}^{K} \lambda_k \exp(-\lambda_k)$$

De fleste forsøk på å estimere disse sannsynlighetene baserer seg på å legge en aprioriforfordeling på λ_k. Bethlehem & al. (1990), Skinner & al. (1994) og Fienberg Makov (1998) baserer seg på en såkalt Poisson-gamma modell med

$$\lambda_k \sim Gamma(\alpha, \beta).$$

Denne gir en negativ binomisk fordeling for F_k som er den samme for alle verdier av *k* og spesielt

$$P(I_k = 1) = P(F_k = 1) = (1 + N\beta)^{-(1+\alpha)}.$$

α og β kan estimeres på grunnlag av utvalget. Dette gir også at $P(F_k = 1 | f_k = 1)$ vil være den samme for alle kombinasjoner av nøkkelvariablene, noe som apriori sjelden vil være troverdig. Tilpasningen til empirisk kjente fordelinger for F_k -ene har imidlertid vært så som så og andre aprioriforfordelinger enn gammafordelingen har også vært prøvd (Skinner & Holmes (1993), Chen & Keller McNulty (1998)).
Det ville være mer tilfredsstillende å kunne estimere ulike sannsynligheter for entydighet i populasjonen for ulike kombinasjoner som er entydige i utvalget, dvs. slik at "risikoen" $P(F_k = 1 | f_k = 1)$ kan variere over ulike verdier av *k* (såkalt "per record risk"). For å minimere antall unike i utvalget bør en først foreta seg noe med de utvalgsenhetene som har størst risiko målt på denne måten. Skinner og Holmes (1998) foreslår en modell der λ_k modelleres som en log-lineær modell med overdispersjon:

$$\lambda_k = \mu + u_{x_1}{}^{X_1} + \cdots + u_{x_H}{}^{X_H} + \varepsilon_k$$

der $x_1, \ldots, x_H$ er verdiene av de kategoriske variablene $X_1, \ldots, X_H$ som danner celle *k* i krysstabuleringen.. $\varepsilon_k \sim Normal(0, \sigma^2)$. Eksperimenter hvor estimerte sannsynligheter for entydighet ble sammenlignet med virkelige andeler gav gode resultater. Den log-lineære komponenten av modellen må holdes så enkel som mulig for å unngå ustabilitet i estimatene. Modeller med to-faktoreffekter kan imidlertid også gi godt resultat.

En ennå annen metode som har vært forsøkt for å estimere antall i populasjonen når det kun er en forekomst i utvalget er syntetisk estimering. Vi vil ikke gå i detalj om den metoden her, men henviser til Willenborg og de Waal (1996) avsnitt 5.2.

Kontinuerlige nøkkelvariable slik som for eksempel inntekt må behandles på andre måter. I prinsippet er enhver verdi av en kontinuerlig variabel unik i populasjonen så vel som i utvalget. Bare det forhold at de alltid måles med et begrenset nøyaktighet gjør at to like verdier likevel forekommer i data. Kall den kontinuerlige variabelen *Y*. Hvis en forutsetter at en inntrenger kanskje bare kjenner verdien av *Y omtrentlig* for de enhetene han hun ønsker å identifisere, vil man kunne godta en presis verdi av en kontinuerlig nøkkelvariabel i datasettet. Willenborg og de Waal (1996, Kap. 5.3) bruker begrepet *sammenlignbare* (*comparable*) verdier. To enheter i populasjonen med samme kombinasjon *k* av andre (diskrete) variable og forskjellige verdier $\boldsymbol{Y}_1$ ***og*** $\boldsymbol{Y}_2$ på en kontinuerlig variabel sies å være *p*% sammenlignbare dersom forskjellen mellom dem ikke er større enn *p%* av den største:

$$100 \frac{|\boldsymbol{Y}_2 - \boldsymbol{Y}_1|}{\boldsymbol{max}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)} \leq \boldsymbol{p}$$

Hvis tilstrekkelig mange enheter kan sies å være sammenlignbare i denne forstand vil man kunne si at den gitte kontinuerlige variabelen er sikker under de gitte forutsetninger om nøyaktig kunnskap. Hvis dette ikke holder må man gjøre noe med data. Verdien av *p* som eventuelt benyttes vil måtte avhenge av hvor nøyaktig en vurderer at en potensiell inntrenger vil kunne kjenne verdien av den kontinuerlige variabelen. Dette vil være avhengig av hvilke eksterne datakilder som kan være tilgjengelige om *Y.*

Spesielt viktige er de ekstreme verdiene av den kontinuerlige variabelen. Hvis $\boldsymbol{Y}_1$ er den største verdien som forekommer i gruppe *k* i populasjonen vil verdien være *p%* sammenlignbar med den hvis

$$100\frac{\boldsymbol{Y}_1 - \boldsymbol{Y}_2}{\boldsymbol{Y}_1} \leq \boldsymbol{p}$$

Hvis *d* er en nedre skranke for antall enheter man kan akseptere at er sammenlignbare med Y, kan man formulere følgende sikkerhetsregel:

Regel 1: For enheter med samme kombinasjon *k* av et gitt sett kategoriske nøkkelvariable skal det være minst *d* populasjonselementer i intervallet $[(1-p)Y_1, Y_1]$ hvor $\boldsymbol{Y}_1$ er den største verdien av *Y* som forekommer i kombinasjon *k*.

Regel 1 danner grunnlaget for såkalt *toppkoding*, dvs. at den definerer når det er behov for tiltak mot ekstremt høye verdier. En tilsvarende regel kan defineres for minste verdi av *Y* og definerer behov for *bunnkoding*. Hvis mikrodatasettet er et utvalg på *n* enheter vil vi ikke nødvendigvis kjenne $\boldsymbol{Y}_1$. Sortert i synkende rekkefølge vil vi observere $\boldsymbol{y}_1 \geq \boldsymbol{y}_2 \geq \cdots \geq \boldsymbol{y}_n$. La $\gamma = d/N$. La $\boldsymbol{g}_s = \#[\boldsymbol{Y}_i \geq \boldsymbol{y}_s]/\boldsymbol{N}$ være andelen av enheter i populasjonen med $\boldsymbol{Y}_i \geq \boldsymbol{y}_s$. Vi kan nå ikke anvende regel 1 direkte, men vi kan søke å bestemme *s* slik at $\boldsymbol{P}[\boldsymbol{g}_s \geq \gamma] \geq 1-\alpha$ og velge den minste verdi *s* som oppfyller dette som erstatning for $(1-\boldsymbol{p})\boldsymbol{Y}_1$. $\boldsymbol{P}[\boldsymbol{g}_s \geq \gamma]$ blir uavhengig av fordelingen til Y. I et enkelt tilfeldig utvalg er $\boldsymbol{P}[\boldsymbol{g}_s \geq \gamma]$ en sum av hypergeometriske sannsynligheter og kan approksimeres ved beta- integralet

$$\boldsymbol{P}[\boldsymbol{g}_s \geq \gamma] = \frac{\Gamma(\boldsymbol{n}+1)}{\Gamma(\boldsymbol{s}-1)\Gamma(\boldsymbol{n}-\boldsymbol{s}+1)} \int_{\gamma}^{1} \boldsymbol{x}^{s-2}(1-\boldsymbol{x})^{n-s+1}\,\boldsymbol{dx}\,.$$

Eksisterende metoder som kan benyttes for å sikre kategoriske så vel som diskrete variable vil bli omtalt i avsnitt 6.

5. Sensitiv og beskyttelsesverdig informasjon.

5.1 Data om individer, familier og husholdninger

Når man diskuterer konfidensialitet er begrepet *sensitiv informasjon* sentralt. For personer har personopplysningslovens §2 nr. 8 en definisjon av hva som er en *personsensitiv opplysning.* Dette er rasemessig/etnisk bakgrunn, politisk/filosofisk/religiøs oppfatning, mistenkt/siktet/-tiltalt/dømt for straffbar handling, helseforhold, seksuelle forhold og medlemskap i fagforening. Men også andre opplysninger kan oppleves som følsomme og private for den enkelte.

F.eks. vil opplysning om utdanning, inntekt og formue være variable som en del mennesker ikke ønsker skal kunne leses av alle. ”Verdien” til noen av disse er lett synlige og er kanskje vel så mye identifiserende som følsomme. Inntekt og formue er offentlig tilgjengelig fra ligningen og derfor sterkt identifiserende. Den oppfattes imidlertid også som noe privat som ikke hvem som helst bør kunne lese den nøyaktige verdien av. På bakgrunn av at skatteligningen er offentlig tilgjengelig, nå også på CD, må en oppfatning av inntekt og formuesforhold som noe følsomt i statistikk sees på som mer en emosjonell og derved subjektiv holdning enn noe som er direkte farlig for den enkelte. Mange av disse variablene dekkes ikke av det juridiske begrepet "sensitiv" i henhold til personopplysningsloven. En bredere betegnelse er *beskyttelsesverdige* opplysninger.

Avsløring av individers egenskaper på visse områdene kan i noen tilfeller være alvorlig til skade for individet i forhold til dets omgivelser. En opplysning som kan være til skade for ett individ vil kunne være fullstendig harmløs for en annen. Dette har også et rent subjektivt element. Graden av "skade", er ikke bare et spørsmål om objektive kriterier, men like mye om hvordan individet oppfatter den informasjonen som er avslørt. Et individ som er åpen omkring et tema som gjelder det selv vil ikke kunne presses av en inntrenger som har fått tak i informasjon. Det er derfor ofte individenes egne subjektive holdninger til sine egenskaper som gjør dem beskyttelsesverdige. Størst mulig åpenhet i samfunnet vil derfor i alle sammenhenger måtte sees som en fordel for den som skal samle inn statistisk informasjon. Dette kan imidlertid statistikkprodusenten ikke selv gjøre noe med.

Hensikten med å hindre at enkeltindivider eller husholdninger blir identifisert i statistiske data er å svekke muligheten for at opplysninger som fra en hvilken som helst oppgavegivers synspunkt er beskyttelsesverdige, blir avslørt.

5.2 Data for bedrifter og foretak

For bedrifter er det gjerne mulighetene for at konkurrerende bedrifter kan utnytte informasjon til å skaffe seg en markedsmessig fordel som er den sterkeste drivkraften bak et ønske om å holde informasjon beskyttet. Det er også mest av hensyn til bedrifter den såkalte (n, k) dominansregelen er innført, selv om den benyttes også i forbindelse med personer. (Se side 7). Regelen i SSB er $n = 3$ og $k = 90\%$ (Sikkerhetshåndboken, kapittel 6). Praktiseringen av 90% regelen ser imidlertid ut til å variere. Blant de variablene som den enkelte bedrift kan finne ønskelig å hemmeligholde finner man produksjon, produksjonsverdi, salg, lønnskostnader, import og eksport.

Vi vil også reise spørsmål om hensynet til at informasjoner om enkeltbedrifter ikke skal kunne leses av statistikken er noe overdrevet i SSB og vil anføre to grunner til det. For det første er det slik at svært mange av de opplysningene som søkes skjult i statistikken er offentlig tilgjengelige fra andre kilder. Det gjelder ikke minst bedriftsregnskapene som foruten å danne grunnlag for ligningen, også må overleveres foretaksregisteret Brønnøysund. Offentlig statistikk er derfor ikke det første sted en bedrift som vil ha informasjon om konkurrenten, vil søke i. For det andre vil neppe bedriftene føle at all den informasjon som SSB samler inn til rent statistisk bruk er beskyttelsesverdig.

Bedrifter og foretak skal føle seg sikre på at informasjon som ikke er offentlig tilgjengelig, og som de ikke ønsker skal bli kjent for offentligheten eller konkurrentene, heller ikke blir det via offentlig statistikk. Det vil imidlertid være en fordel om omfanget av de tiltak som må gjøres for å unngå uønskede avsløringer, ikke er større enn nødvendig. En reduksjon av tiltakene vil kanskje kunne oppnås ved såkalt informert samtykke (informed consent) hvor

hver oppgavegiver tillates, men ikke oppfordres til, å markere den informasjon som de ikke ønsker skal kunne bli kjent av andre. Samtidig som enkeltbedrifter vil kunne ønske å holde noe informasjon skjult for konkurrentene ved ikke å gi riktige opplysninger der SSB kan be om eller kreve dem, er de samme bedriftene gjerne brukere av den samme statistikken og vil ha interesser av at den er mest mulig informativ. Det er derfor grunn til å stille spørsmål om ikke bedriftene selv bør ta noe av ansvaret for å vurdere hvor mye som kan være synlig og hvor mye som må beskyttes av den informasjon de gir.

5.3 SSBs holdning

Som nevnt i avsnitt 1 er statistikkloven premissleverandør for SSBs konfidensialitetskrav i statistikken. Detaljerte tabeller øker muligheten for identifisering av personer og bedrifter (oppgavegiver) i statistikk fra SSB. Ved publisering av data for mindre geografiske områder vil lokalkunnskap hos brukerne lettere bidra til identifisering. Det samme vil fleksible muligheter for brukerne til å selv å definere geografiske områder for statistikken, f.eks. ved ulike kartbaserte løsninger. Dette aktualiserer behovet for konkrete retningslinjer for anonymisering av statistikker.

SSB er i ferd med å avklare strategi og retningslinjer for detaljeringsgraden ved publisering av statistikk (egen DM-sak i løpet av 2001). Retningslinjene er basert på følgende hovedprinsipp:

- **Hovedregel:** Opplysninger må ikke offentliggjøres slik at de kan føres tilbake til oppgavegiver eller annen identifiserbar enkeltperson eller foretak.
- **Unntaksregel:** Opplysninger kan unntaksvis av hensyn til en hensiktsmessig oppbygging av statistikken offentliggjøres slik at de kan tilbakeføres til oppgavegiver eller annen identifiserbar enkeltperson dersom dette ikke er til skade for vedkommende, eller til urimelig skade for juridisk person.

Hva bør forstås med uttrykket ”til skade for, eller urimelig til skade for denne” i statistikklovens §2.6? Et slikt uttrykk kan være gjenstand for ulik tolkning innen Statistisk sentralbyrå så vel som i den offentlige opinion. Etter vår oppfatning er det ikke bare et spørsmål om objektive kriterier, men vel så mye om hvordan en oppgavegiver subjektivt oppfatter den informasjonen som gis. En type informasjon som en oppgavegiver vil oppfatte som harmløs vil av en annen kunne oppfattes som svært sensitiv. En oppgavegivers villighet til å avgi korrekt informasjon vil imidlertid uvegerlig avhenge av dennes egen forståelse av hva som er ”til skade”, ikke av hvordan SSB eller noen annen vil forstå det. Det vil her også være et skille mellom personlige oppgavegivere på den ene side og bedrifter og foretak på den andre. For bedrifter og foretak vil objektive interesser, i første rekke spørsmålet om konkurrenter vil kunne utnytte informasjonen til egen konkurransefordel, være det styrende hensynet som kan begrunne et ønske om konfidensialitet. For fysiske personer, familier og husholdninger vil visse typer informasjon kunne føre til sosial stigmatisering om de ble kjent. En subjektiv frykt for slik stigmatisering vil også kunne spille en vesentlig rolle selv om den ikke behøver å være rasjonelt begrunnet.

I tillegg til respondentenes egne interesser kan det være samfunnsmessige hensyn som tilsier at visse typer informasjon om enkelte respondenter ikke blir offentlig kjent. Dette gjelder kanskje spesielt personer i militære yrker og informasjon knyttet til militær produksjon.

Det er derved ikke mulig å gi en klar definisjon av til skade-begrepet. Dette er en rettslig standard som vil kunne variere fra tilfelle til tilfelle, statistikkområde, opplysningstype og skiftende samfunnsholdninger. Det er derfor viktig å holde fast på hovedregelen om at statistikk ikke skal være identifiserende, og derved unngå til dels umulige avveininger av til-skade begrepet knyttet til unntaksregelen.

6. Metoder for beskyttelse av konfidensialitet

I dette avsnittet vil vi presentere en oversikt over de metodene for å beskytte konfidensialitet som i dag er i bruk eller som er omtalt i litteraturen. I likhet med tidligere avsnitt vil metoder for tabeller og metoder for mikrodatasett behandles i egne underavsnitt. Metoder for mikro-datasett har også implikasjoner for tabeller som produseres på grunnlag av et mikrodatasett som er behandlet med de samme metodene. Det ville være ønskelig at konfidensielt sikre tabeller fra et mikrodatasett og som også er publisert offentlig skal kunne reproduseres med korrekte marginaler fra et mikrodatasett som er behandlet med metoder for å sikre konfidensi-aliteten. Denne form for konsistens mellom metoder for tabeller og metoder for mikrodatasett har imidlertid blitt lite fokusert i litteraturen. Noen av de metodene som omtales nedenfor er i vanlig bruk i SSB, men ofte i en litt primitiv form som kan utføres "for hånd" med en viss bruk av intuisjon. Andre krever avansert bruk av programvare for å kunne anvendes. Noen metoder er tilgjengelig i programpakken ARGUS som egentlig er to pakker, τ-ARGUS for tabeller og μ-ARGUS for mikrodatasett. SSB har nå denne programpakken. Manualer er til-gjengelig fra internett. (www.cbs.nl/sdc/argus.htm).

De ulike metodene anvendes i tilknytning til ulike *kostnadskriterier*, dvs. kriterier for hvordan man skal måle den mengden informasjon som går tapt når tabellene eller mikrodatasettene behandles for å sikre konfidensialiteten. Her vil kun metodene beskrives. Kostnadskriteriene vil bli diskutert i avsnitt 7.

Det må også presiseres at bruk av en metode ikke nødvendigvis automatisk produserer en sikker tabell eller mikrodatasett. Hvorvidt en gitt anvendelse av en metode gir den nødven-dige sikkerhet eller om det må gås mer drastisk til verks, må avgjøres i hvert enkelt tilfelle.

6.1 Metoder for tabeller

Prikking (Cell suppression) går ut på å erstatte sensitive celler i en tabell med kolon. Dette kalles en primærprikking. Når tabellene er utstyrt med marginaler må man også gjennomføre sekundærprikking for å unngå residual avsløring, dvs. at tallet i den prikkede cellen skal kunne beregnes ved hjelp av de fullstendige marginalene. Hvilke og hvor mange celler som må sekundærprikkes vil avhenge av hvordan de sensitive cellene konfigurerer seg i tabellen og av hvilke kostnadskriterier vi ønsker å minimere. Problemet å prikke en stor tabell slik at kostnadene blir minimert er ikke trivielt. Det er et lineært, eventuelt også heltalls program-meringsproblem som krever stor datakraft med effektiv programvare for å gjennomføre. Et svært stort antall mulige løsninger vil måtte gjennomgås. τ-ARGUS har innebygget metoder for optimalisert prikking, men låner en modul fra den numeriske optimeringspakken XPRESS (Dash software ltd.). For å aksessere XPRESS via τ-ARGUS kreves en såkalt hardware aksess nøkkel eller "dongle". Mens ARGUS for øvrig er gratis programvare koster denne donglen 600GBP.

Tabell-redesign er bare navnet på prosessen å slå sammen to eller flere "tilstøtende" kategorier (kolonner, rader etc.) i noen av variablene. I store tabeller som inneholder mange sensitive celler anbefales det å gjennomføre tabell-redesign før en gjennomfører prikking, ellers blir oppgaven (for τ-ARGUS) med å utføre optimal prikking for stor. Tabell-redesign er nært knyttet til det som kalles Global omkoding i mikrodatasett. Global omkoding av et mikrodatasett bevirker tabell-redesign av de tabellene som lages av det.

Kontrollert avrunding betyr å avrunde tallene i alle cellene til et multiplum av en gitt base, f.eks. 5, 10 50 eller 100. Hvor mye beskyttelse kontrollert avrunding gir avhenger av størrelsen på basen. Problemet med avrunding er å avrunde alle celler og marginaler på en slik måte at additivitet mellom celler og marginaler blir bevart samtidig som den avrundede tabellen blir så lik den opprinnelige som mulig. I likhet med optimal prikking er optimal avrunding et lineært og heltalls programmeringsproblem som krever XPRESS. Kontrollert avrunding for to-veis tabeller finnes i τ-ARGUS. I tabeller av høyere dimensjon er problemet med å få avrundede celler og marginaler til å stemme ikke alltid løsbart. Kontrollert avrunding er uegnet for tabeller der celletallene eller verdien av aggregeringsvariabelen over enhetene varierer ekstremt. Basen må i slike situasjoner gjøres så stor at det meste av informasjonsinnholdet i cellene forsvinner.

Gruppe-prikking (Group suppression) er en klasse metoder for simultan prikking av flere tabeller med en eller flere felles marginaler. Det er implementert i τ-ARGUS som et eksperiment, men har ikke full funksjonalitet.

Log-lineær tilpassing har betydelige muligheter med tanke på å konstruere tabeller som endrer cellenes innhold i større eller mindre grad uten å endre på marginalene. Denne teknikken er i første rekke utviklet med tanke på log-lineær analyse av kontingenstabeller, men det er teknisk ikke noe problem å anvende den på mengdetabeller. Log-lineær tilpassing bevarer de marginalene vi ønsker uforandret mens innholdet i tabellen endres ("glattes") mer eller mindre. Dersom en hel tabell tilpasses med et log-lineært program vil imidlertid hele tabellen påvirkes, også de cellene som er sikre og ikke behøver noen bearbeiding. Det vil imidlertid være mulig å velge ut bare deler av en tabell til en slik tilpassing. For eksempel kan man først anvende τ-ARGUS til å finne en optimal prikking og deretter tilpasse en subtabell som inneholder alle primært og sekundært prikkede celler med et log-lineært program. Man vil da få en tabell hvor også de prikkede cellene inneholder "glattede" tall i stedet for de opprinnelige. Vi har ikke sett at denne typen bruk av log-lineære tabeller har vært presentert i litteraturen.

Permutering (Data Swapping) er en klasse metoder som tar sikte på å bytte om enheter mellom celler. Dette påvirker antall enheter bak hver celle og tallene i dem, men kan gjennomføres på sofistikerte måter som bevarer visse strukturer i tabellen. For eksempel kreves at det forventede antall enheter i hver celle i den ombyttede tabellen må være lik det opprinnelige antallet (se PRAM metoder under avsnitt 6.2). I stedet for å anvendes på de bakenforliggende enhetene i tabellen kan permutering anvendes på enheter av den aggregerte variabelen. Disse enhetene kan defineres i den størrelse man ønsker.

Det er i ferd med å utvikles metoder som er i stand til å foreta ombyttinger på en slik måte at tabellens marginaler forblir uforandret. Disse metodene er svært sofistikerte og benytter teori for såkalte Gröbner baser (Fienberg et al 1998, Diaconis and Sturmfels 1998). Anvendelse av slike metoder krever simuleringer med markovkjedemodeller. Disse metodene er nært forbundet med log-lineære modeller. Resultatene blir svært like dem man får fra log-lineær tilpassing, men i stedet for å erstatte de opprinnelige celletallene med forventede verdier når tabellens marginaler er gitt, simuleres det innholdet stokastisk under forutsetning av de gitte marginalene.

Perturbasjon er en klasse av metoder som legger stokastiske feilledd til celletallene, men slik at disse feilleddene summerer seg til 0 over hver rad og over hver kolonne. Størrelsen på feilleddene vil kunne variere fra celle til celle. Slike metoder er ikke inneholdt i τ-ARGUS.

6.2 Metoder for mikrodatasett

Langt flere metoder er etter hvert blitt utviklet for å behandle mikrodatasett mot avsløringer enn det finnes metoder for tabeller. For å systematisere metodene er en del metoder forsøkt delt inn i klasser. En slik klasse er *den lineære transformasjonsklassen*. La $\boldsymbol{X}$ være den originale datamatrisen med n rader (enheter) og p kolonner (variable). La $\boldsymbol{Z}$ $(n\times q)$ være resultatmatrisen etter at $\boldsymbol{X}$ er behandlet. La $\boldsymbol{A}$ være en $n\times n$ matrise som opererer på radene til $\boldsymbol{X}$ og $\boldsymbol{B}$ en $p\times q$ matrise som opererer på kolonnene til $\boldsymbol{X}$. Hvis $\boldsymbol{X}$ og $\boldsymbol{Z}$ kan relateres med en bilineær transformasjon av typen

$$\boldsymbol{Z = AXB + C}$$

sies metoden å tilhøre den lineære transformasjonsklassen. Matrisen $\boldsymbol{C}$ $(n\times q)$ er en perturbasjonsmatrise som kan brukes til å tilsløre produktet $\boldsymbol{AXB}$.

Blant de metodene som blir omtalt nedenfor tilhører noen den lineære transformasjonsklassen mens andre ikke gjør det. Programpakken μ-ARGUS kan utføre noen av metodene.

Støylegging (perturbation methods) er en annen klasse av metoder som går ut på å legge støy på variabelverdier, globalt eller lokalt, slik at identifisering gjøres mer usikker.

Global omkoding går i korthet ut på å slå sammen kategorier i kategoriske variable for å unngå at noen kategorier virker identifiserende, eventuelt i kombinasjon med andre variable. Å redusere antall sifre i yrkes, nærings eller utdanningskoder er en form for global omkoding. Også gruppering av kontinuerlige variable, f.eks. inntekt eller alder, går inn under denne betegnelsen. Metoden er global i den forstand at den virker på samme måte på alle rader i datamatrisen $\boldsymbol{X}$. Global omkoding er den mest brukte metoden for å redusere mulighetene for identitetsavsløring. I forhold til en del andre metoder som vil bli omtalt her har global omkoding den fordel at den aldri ødelegger det som kalles *dataintegriteten*, dvs. at den aldri fører til at det oppstår logiske inkonsistenser i data og som ikke var der fra før. Dersom det er mange unike kombinasjoner i data anbefales det ofte at man gjør global omkoding først og så benyttes andre metoder på det som da blir igjen. Å utføre global omkoding på et mikrodatasett svarer til å gjøre tabell-redesign på en tabell. μ-ARGUS kan utføre global omkoding. Det kan enten gjøres fullstendig styrt av brukeren eller automatisk. I siste tilfelle søker μ-ARGUS å minimere informasjonstapet i data (målt ved økning i dataenes entropi). Hvis verdiene til variablene som skal rekodes globalt, er kodet som dummyvariable, er metoden en lineær transformasjonsmetode med $\boldsymbol{A = I}$ (identitetsmatisen), $\boldsymbol{B}$ er en matrise av nuller og enere og $\boldsymbol{C = 0}$.

Lokal sensurering (local suppression) består i å sensurere verdiene til noen nøkkelvariable i noen linjer hvor det forekommer kombinasjoner som er tilstrekkelig sjeldne. Lokal sensurering har den fordel i forhold til global omkoding at informasjonstapet blir mindre. På den annen side vil lokal sensurering introdusere skjevhet i aggregerte data. Hvor alvorlig dette er avhenger av omfanget. Også lokal sensurering må søkes gjennomført slik at informasjonstapet blir så lite som mulig. μ-ARGUS kan gjøre lokal sensurering. Den minimerer det totale antall sensureringer som det er nødvendig å gjennomføre. Hvis for eksempel kombinasjonene

”Bostedskommune = Oslo”, ”Yrke = bonde” og ”Kjønn = kvinne” og samtidig

”Bostedskommune = Oslo”, ”Yrke = bonde” og ”Status = ugift”

begge forekommer så sjelden at de representerer risiko for identifisering, vil μ-ARGUS foretrekke å sensurere enten bostedskommune eller yrke heller enn både kjønn og status. Sensurering av en kombinasjon, som ovenfor, kan få konsekvenser for hvorvidt kombinasjoner i andre linjer hvor den sensurerte variabelen inngår oppfattes som sjeldne og må sensureres. Lokal sensurering kan derfor ikke utføres uavhengig linje for linje i datasettet, men må baseres på et optimaliseringskriterium som tar hensyn til hele datasettet under ett.

Lokal sensurering kan gjennomføres ut fra ulike optimeringskriteria. Ett slikt kriterium kan være at man ønsker færrest mulig sensureringer. En forsker som skal ha et bestemt datasett kan av hensyn til de analysene han/hun ønsker å gjøre, heller ønske at få kategorier av variablene berøres av sensureringen og spesielt at noen ikke blir det. de Waal og Willenborg (1998) beskriver metoder for å gjøre optimal sensurering under ulike kriteria. Problemet formuleres som et heltalls (0-1) programmeringsproblem. Disse optimeringsproblemene kan imidlertid ta lang tid på datamaskin og mer heuristiske fremgangsmåter er foreslått anvendt. Hurkins og Tiorine (1998) studerte hvordan global omkoding og lokal sensurering kan brukes i optimal kombinasjon.

Støylegging, også kalt maskering av data, er en klasse av metoder som går ut på å legge støy på verdiene til nøkkelvariable for å gjøre det vanskeligere for en inntrenger å kjenne igjen en person, husholdning eller bedrift i sin identifikasjonsfil. Hvis $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')$ er matrisen av observerte datavektorer fra et mikrodatasett, kan man for hver enhet legge til en stokastisk støy- eller perturbasjonsvektor $\boldsymbol{u}_j$ slik at $\boldsymbol{x}_j$ erstattes av $\boldsymbol{y}_j = \boldsymbol{x}_j + \boldsymbol{u}_j$. Fordelingen til $\boldsymbol{u}_j$ kan avhenge av fordelingen til $\boldsymbol{x}_j$. Man kan tenke seg at den som frigir perturberte data også offentliggjør den metoden som er brukt til å støylegge data med. Slik offentliggjøring kan gjøre det lettere for en inntrenger som ønsker å trenge gjennom støyen, men vil også være av stor betydning for at støylagte data skal kunne brukes til sine legitime formål.

Delvis basert på Sullivan og Fuller (1989,1990) og Fuller (1993) kan man sette opp følgende krav til perturberte mikrodata som skal frigis til forskningsformål:

a. Gjennomsnitt og kovariansmatrise for de perturberte variablene må være så like dem for de tilsvarende opprinnelige variablene som mulig.
b. De marginale kumulative fordelingsfunksjonene til de støylagte variablene må være så lik de opprinnelige som mulig.
c. Konsistente estimater for høyere ordens størrelser, som f.eks. tre-veistabeller, kurtose- og skjevhetsledd må kunne konstrueres med minst mulig økning i variansen.
d. Sannsynligheten for at en inntrenger skal kunne identifisere et individ må være tilstrekkelig liten.
e. Det bør være mulig å kvantifisere på en graden av konfidensiell sikkerhet på en måte som er forståelig for publikum.
f. Støyleggingen av variablene må ikke bryte dataintegriteten.

Dette er ideelle fordringer som ikke alle trekker i samme retning. Det siste punktet går på det forhold at perturbering kan føre til logiske inkonsistenser mellom verdiene til forskjellige variable. Det er ikke ønskelig i et datasett som skal kunne brukes.

Fuller (1993) foreslår en metode hvor hver enkelt variabel transformeres først til en uniformt fordelt variabel ved hjelp av den kumulative fordelingsfunksjonen og deretter til en standard

normalfordelt variabel som pålegges støy som er $u_j \sim N(0, \delta)$, uavhengig for ulike observasjoner *j*. Den perturberte variabelen transformeres så tilbake til en uniformt fordelt variabel som deretter transformeres tilbake til opprinnelig skala ved den inverse av den opprinnelige kumulative fordelingsfunksjonen. Metoden kan anvendes både på kontinuerlige og diskrete data. Den produserer en perturbert kumulativ fordelingsfunksjon som er ganske lik den opprinnelige. Graden av perturbasjon avhenger av δ. Metoden fungerer best for variable som ikke er for langt fra å være normalfordelte. Ved svært skjevfordelte variable, slik som inntektsfordelinger og andre økonomiske variable, vil bivariate sammenhenger kunne påvirkes betydelig. Metoden kan modifiseres til å ta hensyn til at ikke alle variable maskeres og at det kan være logiske sammenhenger mellom de mulige verdiene til ulike variable.

PRAM (Post RAndomisation Method) (Gouweleeuw et al. 1998) er en perturberingsmetode for kategoriske data. Den går ut på å forandre verdiene til kategoriske variable i et mikrodatasett ved hjelp av en sannsynlighetsmekanisme, bestemt ved en gitt Markov overgangsmatrise. Hvis denne overgangsmatrisen er kjent for den som skal analysere data, vil det fortsatt være mulig å utføre en rekke forskjellige statistiske analyser på data.

La $\boldsymbol{x}$ være en kategorisk variabel med *K* kategorier, $\boldsymbol{k} = 1, \ldots, \boldsymbol{K}$. La $\boldsymbol{y}$ være den tilsvarende variable etter at $\boldsymbol{x}$ er blitt perturbert. La $\boldsymbol{P} = \{\boldsymbol{p}_{kl}\}$ være en $\boldsymbol{K} \times \boldsymbol{K}$ matrise hvor radene er sannsynligheter:

$$\boldsymbol{p}_{kl} = \boldsymbol{P}(\boldsymbol{y} = \boldsymbol{l} \mid \boldsymbol{x} = \boldsymbol{k})$$

$\sum_{l} \boldsymbol{p}_{kl} = 1$. $\boldsymbol{P}$ er derfor en markovmatrise. Siden det ikke er ønskelig med sterkere perturbering enn det som er nødvendig for å sikre datasettet konfidensielt, bør sannsynligheten for at en variabelverdi ikke blir perturbert, $\boldsymbol{p}_{kk}$, være stor mens sannsynlighetene for alle mulige perturbasjoner, $\boldsymbol{p}_{kl}, \boldsymbol{l} \neq \boldsymbol{k}$, bør være små. På den annen side bør sannsynligheten for at en sjelden og derved sterkt identifiserende verdi er korrekt i mikrodatasettet være liten. Altså bør $\boldsymbol{p}_{kk}$ ikke være stor hvis *k* er en sjelden kategori.

Hvis $\boldsymbol{x}$ er representert som en vektor av dummyvariable, er $\boldsymbol{E}(\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{P}'\boldsymbol{x}$. For vektorer $\boldsymbol{T}_x = \sum \boldsymbol{x}_i$ og $\boldsymbol{T}_y = \sum \boldsymbol{y}_i$ av totaler i en populasjon eller et utvalg, $\boldsymbol{T}_y$ vil $\boldsymbol{E}(\boldsymbol{T}_y \mid \boldsymbol{T}_x) = \boldsymbol{P}'\boldsymbol{T}_x$. Hvis $\boldsymbol{P}$ er kjent for brukeren av datasettet vil han/hun kunne estimere $\boldsymbol{T}_x$ forventingsrett ved $\hat{\boldsymbol{T}}_x = (\boldsymbol{P}^{-1})'\boldsymbol{T}_y$ forutsatt at $\boldsymbol{P}$ er en ikke-singulær matrise. Hvis $\boldsymbol{x}$ er en kombinasjon av alle kategoriske variable vil $\boldsymbol{T}_x$ representere en fullstendig krysstabellering av dem.

PRAM kan anvendes på forskjellige kategoriske variable uavhengig av hverandre. Alternativt kan $\boldsymbol{x}$ være en kombinasjon av flere kategoriske variable som i avsnitt 4. Dette siste gjør det mulig å spesifisere $\boldsymbol{P}$ slik at ulovlige kombinasjoner av kategoriske variable kan unngås i det perturberte datasettet.

Det et mulig å velge $\boldsymbol{P}$ slik at $\boldsymbol{E}(\boldsymbol{T}_y \mid \boldsymbol{T}_x) = \boldsymbol{P}'\boldsymbol{T}_x = \boldsymbol{T}_x$, dvs. at den perturberte totalen $\boldsymbol{T}_y$ blir en forventningsrett estimator for $\boldsymbol{T}_x$. Dette vil være oppfylt hvis $\boldsymbol{T}_x$ er en høyre-egenvektor for matrisen $\boldsymbol{P}'$ svarende til egenverdien 1. Det blir da ikke nødvendig å premultiplisere $\boldsymbol{T}_y$ med matrisen $(\boldsymbol{P}')^{-1}$ for å få et forventningsrett estimat. Gouweleeuw et al. spesifiserer hvordan en slik $\boldsymbol{P}$ matrise kan konstrueres. Det vil imidlertid være nødvendig med adgang til $\boldsymbol{P}$ dersom brukeren ønsker å estimere varianser og konfidensintervall.

Men hvis $\boldsymbol{T}_x$ er en egenverdi for $\boldsymbol{P}$ med multiplisitet 1, vil en bruker av det perturberte datasettet som også har adgang til $\boldsymbol{P}$-matrisen selv, kunne utlede hele $\boldsymbol{T}_x$ ved å regne ut egenvektorene til $\boldsymbol{P}$. Hvis $\boldsymbol{x}$ representerer alle variable i datatsettet, vil hele det uperturberte datasettet være avslørt. Dette kan unngås ved å la $\boldsymbol{T}_x$ være en egenverdi med multiplisitet større enn 1. Dette innebærer imidlertid så sterke restiksjoner på $\boldsymbol{P}$ at det sjelden vil være mulig i praksis.

PRAM har vært anvendt i praksis i Nederland på "Dutch National Travel Survey" hvor brukere utenfor CBS skulle ha tilgang til mikrodata. Anvendelsen av metoden måtte tillempes slik at de tilgjengelige data ble akseptable for brukerne Her ble blant annet postnummerkode perturbert. For flere detaljer om den anvendelsen henvises til Gouweleeuw et al.

Det viser seg at PRAM er ekvivalent med visse såkalte Randomised Response metoder som er metoder som kan brukes ved feltintervju for å få respondenter til å svare på sensitive spørsmål uten å avsløre seg for intervjueren. Forskjellen er at i PRAM anvendes metoden kunstig etter at data er samlet inn, og ikke bare på sensitive spørsmål. Dette gjør det også mulig å tilpasse anvendelsen av (parametrene i) metoden til de marginale fordelingene i utvalget som ikke er kjent på intervjutidspunktet.

Datapermutering (Data swapping) er en klasse av metoder som går ut på å bytte om de observerte verdier av noen variable mellom enhetene i datasettet for på den måten å gjøre det umulig å sammenstille dem med verdier av flere andre til en identifiserende nøkkel. I et tellingsdatasett eller et selvveiende utvalg vil denne metoden fullstendig ivareta alle endimensjonale fordelinger i data og deres tilhørende gjennomsnitt, varianser, kvantiler etc. Anvendt i "rå" form vil metoden imidlertid gjøre alle analyser som både involverer permuterte og ikke-permuterte variable meningsløse.

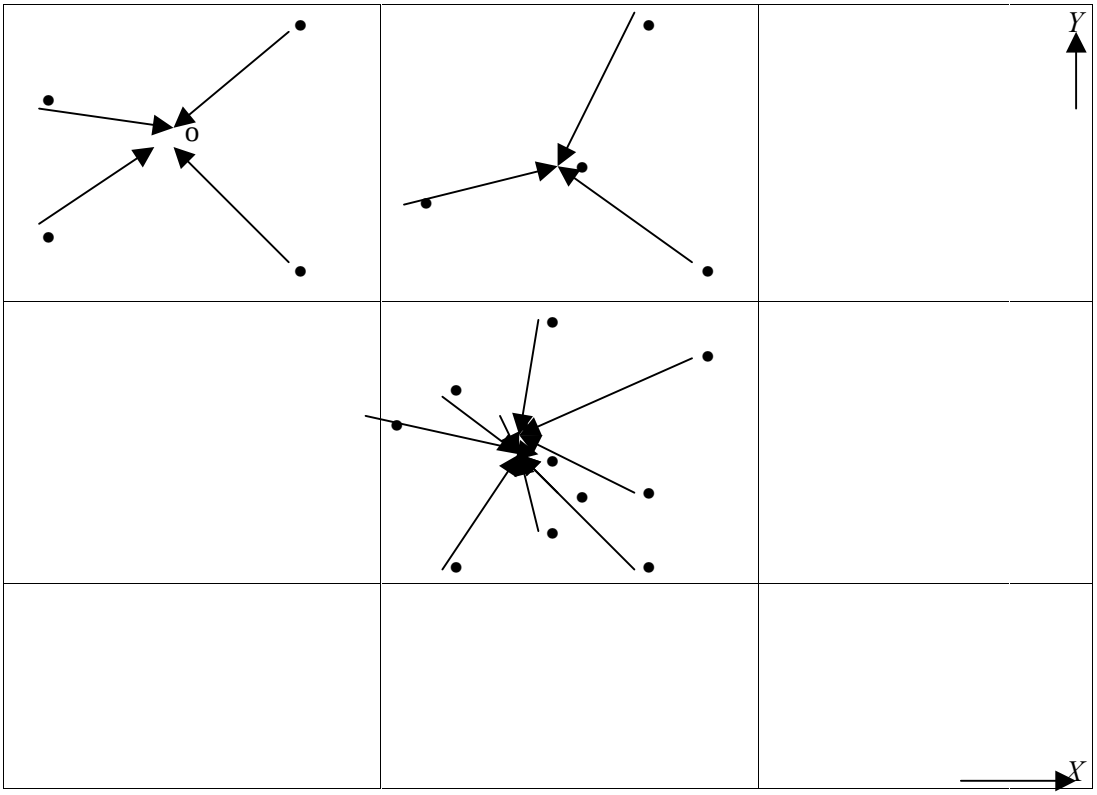
Det finnes mer sofistikerte permuteringsmetoder. For kontinuerlige data kan dette innebære å først dele variabelens variasjonsområde inn i intervaller og deretter permutere innen disse intervallene. Det er også mulig å gjennomføre permutering bare innen grupper bestemt av andre variable. Fullers metode for perturbering av data på side 23 reduserer seg til en permuteringsmetode dersom den enkle empiriske kumulative fordelingsfunksjonen benyttes ved transformasjon begge veier. Kravene a, b og c ovenfor vil da alltid være eksakt oppfylt univariat for hver variabel som behandles med denne metoden. Anvendt uten modifikasjoner som beskrevet ovenfor vil imidlertid de mest ekstreme observasjonene få minst tilbøyelighet til å bli permutert. Dette vil være uheldig siden det gjerne er disse som vil være mest identifiserene. Metoden kan imidlertid modifiseres slik at denne ulempen unngås. Graden av permutering vil avhenge av parameteren δ.

Mikroaggregering kan enklest beskrives ved å tenke seg to kontinuerlige variable som begge kan brukes som nøkkelvariable. De to variablene deles begge inn i intervaller av passende lengde. I et spredningsplott av de to variablene kan disse inndelingene tegnes som i figur 2.

Spredningsplottet deles inn i celler med typisk 3-10 punkter i hver. For hver variabel beregnes gjennomsnitt, eventuelt median av punktene i hver celle og disse erstatter de opprinnelige data. Dersom det er gjennomsnittet av hver celle som benyttes vil totalt gjennomsnitt og sum av variabelen bli upåvirket. Variansene i det nye datasettet vil bli noe deflatert, men vanligvis ikke mye. Korrelasjoner blir lite påvirket. Se for øvrig Defays og Anwar (1998). En enkel versjon av mikroaggregering i en dimensjon finnes i ARGUS.

Et alternativ til å erstatte alle verdier i en celle med en enkelt verdi er å permutere variabelverdiene til enhetene i samme celle (mikropermutering). Dette vil gjøre det umulig for en inntrenger å bruke variabelverdiene til å identifisere en enhet i filen. En slik metode vil bevare

både gjennomsnitt, varianser og alle andre univariate statistikker for hele datasettet uforandret. Hvis cellene ikke blir for få og store i forhold til hele datasettet er det grunn til å tro at også korrelasjoner og multivariate statistikker vil bli lite påvirket. Dette må imidlertid studeres nærmere. En ulempe med mikropermutering er at det vil kunne være mulig å identifisere



Figur 2. Noen punkter i et spreningsplott med illustrasjon av mikroaggregering i celler i to dimensjoner

at et individ er med i et datasett, og også i en gitt permuteringscelle. Har man flere mulige identifiserende variable vil det fort være mulig også å identifisere et individ innen cellen.

6.3 Topp og bunnkoding

På side 17 er det forklart hva topp- og bunnkoding er. Mikroaggregering og permutering er metoder som kan benyttes for å behandle ekstreme verdier. metodene som er beskrevet på side 17 kan benyttes til å bestemme størrelsen på aggregerings eller permuteringscellen for de mest ekstreme verdiene. Topp og bunnkoding er med i ARGUS.

7. Tap av informasjon

Det informasjonstap som en gitt metode for konfidensialitetssikring nødvendigvis må føre med seg, vil avhenge av hvilke analyser en bruker ønsker å gjennomføre.

7.1 Tabeller

For tabeller er det hovedsakelig fire ulike mål på informasjonstap som anvendes og som søkes minimert ved anvendelse av metoder for konfidensialitetssikring. Tre av dem knytter seg utelukkende til anvendelse av prikkemetoder:

- Antall celler som må prikkes
- Antall enheter i de cellene som prikkes
- Verdiene i cellene som prikkes (mengdevariabel)

I τ-ARGUS kan man velge mellom disse kriteriene. Hvilke og hvor mange celler som prikkes vil avhenge av hvilket kriterium som er valgt. Disse målene er også uavhengig av bruk av statistisk modell for data.

Det fjerde målet, entropi (Shannon og Weaver 1949), er et mer generelt informasjonsteoretisk mål som kan anvendes også ved andre metoder enn prikking og også for mikrodata. Den knytter seg imidlertid til hvilken modell en analytiker velger for data. Informasjonstapet må sees i relasjon til modellen. I en ren frekvenstabell generert på grunnlag av en standard modell (multinomisk, poisson eller produktmultinomisk), med *I* rader og *J* kolonner og cellesannsynligheter π_{ij} kan tabellens entropi defineres som

$$\boldsymbol{U}(\Pi) = -\sum_i \sum_j \pi_{ij} \log \pi_{ij} \ .$$

Entropien er minst (0) i en tabell hvor all sannsynlighetsmassen er konsentrert i en celle og størst hvis den er spredd likt over alle cellene. En tabell er derfor mest informativ hvis entropien er liten. Vi kan estimere entropien ved å sette inn $\hat{\pi}_{ij} = \boldsymbol{n}_{ij} / \boldsymbol{n}$. I en tabell der fire celler er prikket, fordelt på to rader og to kolonner, si cellene (2,4), (2,6), (5,4) og (5,6), vil vi kunne definere marginalene til de prikkede cellene som $\pi_{2+} = \pi_{24} + \pi_{26}$, $\pi_{5+} = \pi_{54} + \pi_{56}$, $\pi_{+4} = \pi_{24} + \pi_{54}$ og $\pi_{+6} = \pi_{26} + \pi_{56}$. Basert på en uavhengighetsmodell for sammenhengen mellom rad- og kolonnevariabelen vil det være mulig å estimere sannsynlighetene for de prikkede cellene som

$$\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j}, \ i = 2,5; \ j = 4,6.$$

Settes disse estimatene inn i $\boldsymbol{U}(\cdot)$ kan entropien i den prikkede tabellen estimeres. Denne entropien vil alltid være større enn eller lik entropien i den fullstendige tabellen. Økningen i entropi vil være et mål for informasjonstapet ved prikkingen.

Entropibegrepet er nært knyttet til likelihoodfunksjonen som for en multinomisk modell for den originale (og fullstendige) tabellen ser ut som

$$L(P \mid X) = \sum_i \sum_j x_{ij} \log p_{ij}$$

og som har maksimum for $p_{ij} = x_{ij} / n \equiv \hat{\pi}_{ij}$, sannsynlighetsmaksimeringsestimatet. Estimering av π_{ij} for prikkede celler er et missing data problem som krever en modell for hvordan missing mekanismen virker. Estimater som fremkommer, f.eks. $p_{ij} = x_{i+}x_{+j} / n$ for

uavhengighetsmodellen vil gi lavere likelihood innsatt i $\boldsymbol{L}(\boldsymbol{P} \mid \boldsymbol{X})$ enn estimater basert på fullstendige data. Differansen er ekvivalent med økningen i tabellens entropi. Ved tabell-redesign, sammenslåing av rader og kolonner, vil økning i entropi kunne måles på tilsvarende måte ved å sette inn estimerte cellefrekvenser for hver av de sammenslåtte radene/kolonnene.

For mengdetabeller er det i prinsippet også mulig å estimere økning etter de samme prinsipper. Likelihoodfunksjonen. I tillegg til en likelihood for antall observasjoner per celle krever det en likelihood knyttet til mengdevariabelen gitt celle. Denne gir seg ikke selv, og vil ikke så lett kunne implementeres i et standardprogram.

7.2 Mikrodatasett

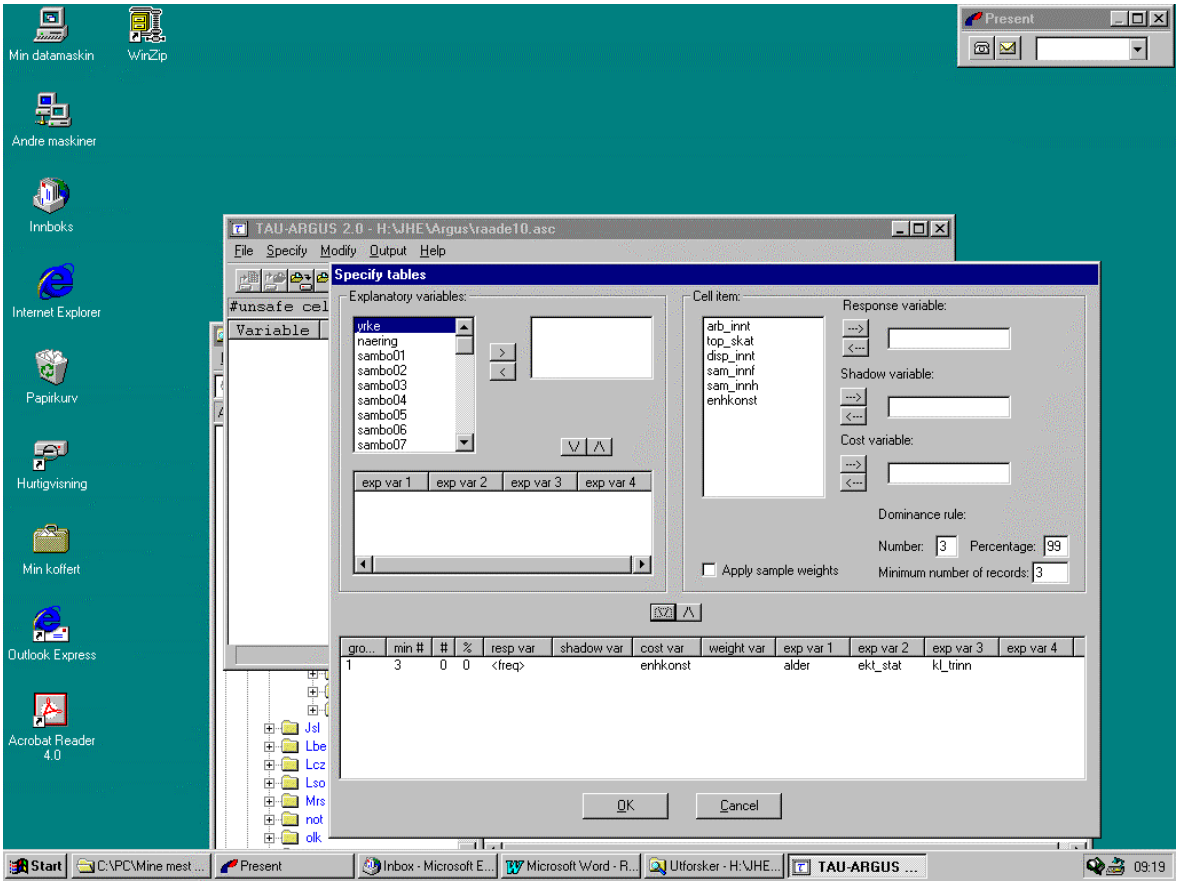
For en forsker som arbeider ut fra en modell som han/hun vil estimere på grunnlag av mikrodata, er endring i entropi, beregnet med utgangspunkt i likelihoodfunksjonen for modellen, det teoretisk ideelle målet for informasjonstap. Ved global omkoding (det å slå sammen kategorier for diskrete data) er det i prinsippet mulig å anvende det også, men i en mer begrenset forstand. For en dataprotektor som skal lage et sikkert datasett av et som ikke er sikkert, vil entropi ikke være et praktisk mål. Det vil være ganske umulig for dataprotektoren å forholde seg til spesielle modeller. Av samme grunn er ikke entropi i en slik spesifikk forstand noe egnet kriterium som kan implementeres i programpakker som ARGUS. Ved bruk av lokal sensurering søker ARGUS å minimere det nødvendige antall sensureringer for å oppnå en sikker fil. For global rekoding bruker ARGUS dataprotektorens klassifisering av variabelens viktighet i kombinasjon med en vurdering av hvor identifiserende de ulike variabelverdier er. For diskrete variable er det viktig å finne den rette balanse mellom global rekoding og lokal sensurering.

8. Eksempler på anvendelser med ARGUS

Vi har forsøksvis anvendt τ-ARGUS på noen problemstillinger med data fra SSB.

8.1 Tre-veis frekvenstabell

Panel 1

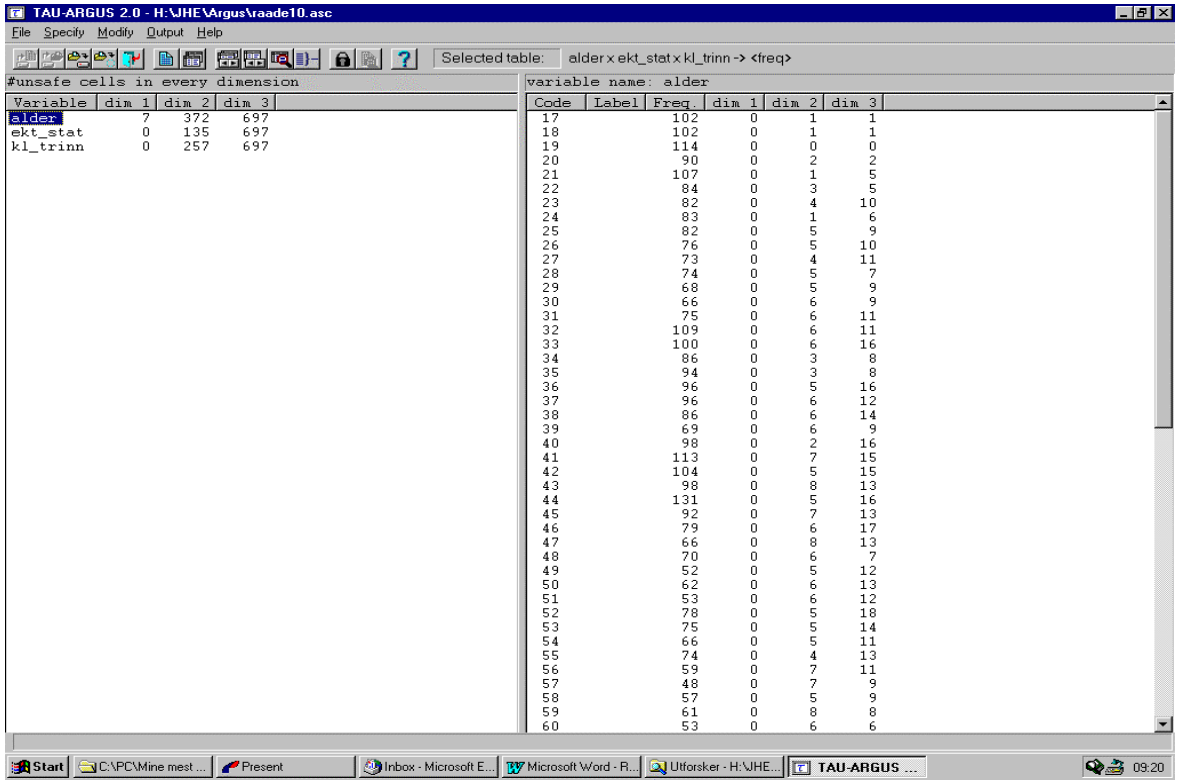


Dette eksempelet tar utgangspunkt i folketellingsdata fra en fulltellingskommune i folketellingen 1990. å grunnlag av en fil skal vi lage en tabell for alder x sivilstatus x lengden av høyeste fullførte utdanning. Både *alder*, og utdanningens lengde (variabelen *kl_trinn*) er målt i ettårige kategorier. Det er i utgangspunktet 83 alderskategorier, fem kategorier av sivilstatus og 13 +missing=14 kategorier for utdanningens lengde. I alt inneholder tabellen 5910 celler. Tabeller spesifiseres i skjermbildet i panel 1. Teksten "*Minimum number of records: 3*" betyr at vi krever minst fire personer i en celle for at den skal regnes som "sikker". Dette innebærer at alle celler med tre eller færre individer blir klassifisert som "usikre".

I det lille vinduet er variabelen *enhkonst* valgt som "cost var", mens ingen variable velges som *response*- eller *shadow*-variabel (dette gjøres først i avsnitt 8.2). Variabelen *enhkonst* er lik 1 for alle personene i datasettet. Aggregert opp per celle blir denne lik cellefrekvensen. Ved å angi denne kostvariabelen sier vi at vi ønsker en sekundærprikking som minimerer antall individer i de prikkede cellene.

Ved å trykke *OK* produseres nå panel 2.

Panel 2



Panel 2 forteller at for variabelen alder alene med ett-årige aldersklasser er det syv usikre kategorier. I de andre to variablene alene er det ingen. La x være antall usikre kombinasjoner i 2-veisgrupperingen *alder* x *sivilstatus*. La y være antall usikre i kombinasjonen *alder* x *klassetrinn* og z være antall usikre i *sivilstatus* x *klassetrinn*. Da kan man lese av panelet at

$$x + y = 372$$
$$x + z = 135$$
$$y + z = 257$$

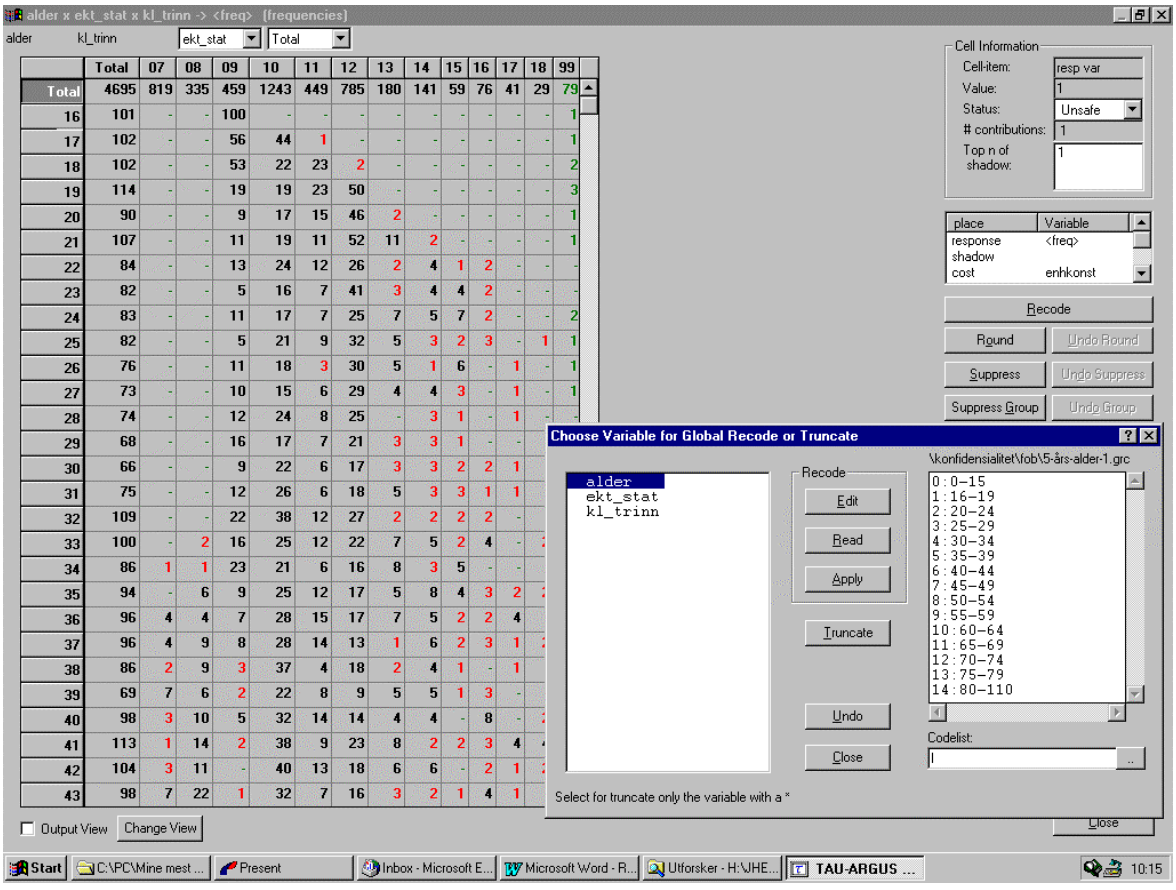
som gir $x = 125, \quad y = 247, \quad z = 10$. I 3-veiskombinasjonen av alle tre variablene er det 697 usikre kombinasjoner av i alt 5395 (når man ekskluderer celler som representerer missingkoder på noen av variablene).[1] Siden alder er markert i venstre side av panelet viser høyre side av panelet hvordan de usikre cellene der alder inngår i en-, to- eller

[1] Ved å krysse 83 alderskategorier, fem sivilstatuskategorier og 13 utdanningskategorier, får man 5395 kombinasjoner totalt.

tredimensjonale tabeller fordeler seg på alder. For eksempel: under "Code" står alderskategorien. For alderskategorien 23 år i alt med 82 personer, vil det ikke være noen usikre kombinasjoner for denne alderskategorien betraktet alene. I kombinasjon med enten sivilstatus eller klassetrinn vil det finnes fire usikre kombinasjoner og i kombinasjon med begge vil det finnes 10. Ved å markere en annen variabel i venstre panel fås en tilsvarende tabell for denne variabelen.

Panel 3 nedenfor viser hvordan de usikre cellene i to-veismarginalen *alder* x klassetrinn[1] fordeler seg og hvor mange det er i hver celle i det vinduet av tabellen som plassen tillater at vises. På skjermen fremtrer de usikre cellene i rødt, de sikre i svart og missingkolonnen (99) i grønt. Det er for mange usikre celler til at (optimal) prikking kan benyttes. Det matematiske optimeringsproblemet har for stor dimensjon. Vi omkoder derfor alder i gruppene 16-19 år, 20-24 år osv. i 5-årsgrupperinger tom 75-79 år og med 80 år + som siste (Alle usikre ett-årskategorier for alder var for aldersgrupper over 90 år.) Vi trykker på Recode i panel 3 og får opp det lille panelet nederst til høyre i bildet. Aldersomkodingene skrives inn eller hentes fra en fil ved hjelp av Read. Deretter trykkes det på Apply. Den omkodede to-veismarginalen kommer til syne i panel 4.

Panel 3



[1] Toveismarginalen er her den toveistabellen av *alder x klassetrinn* man får når man summerer over sivilstatus.

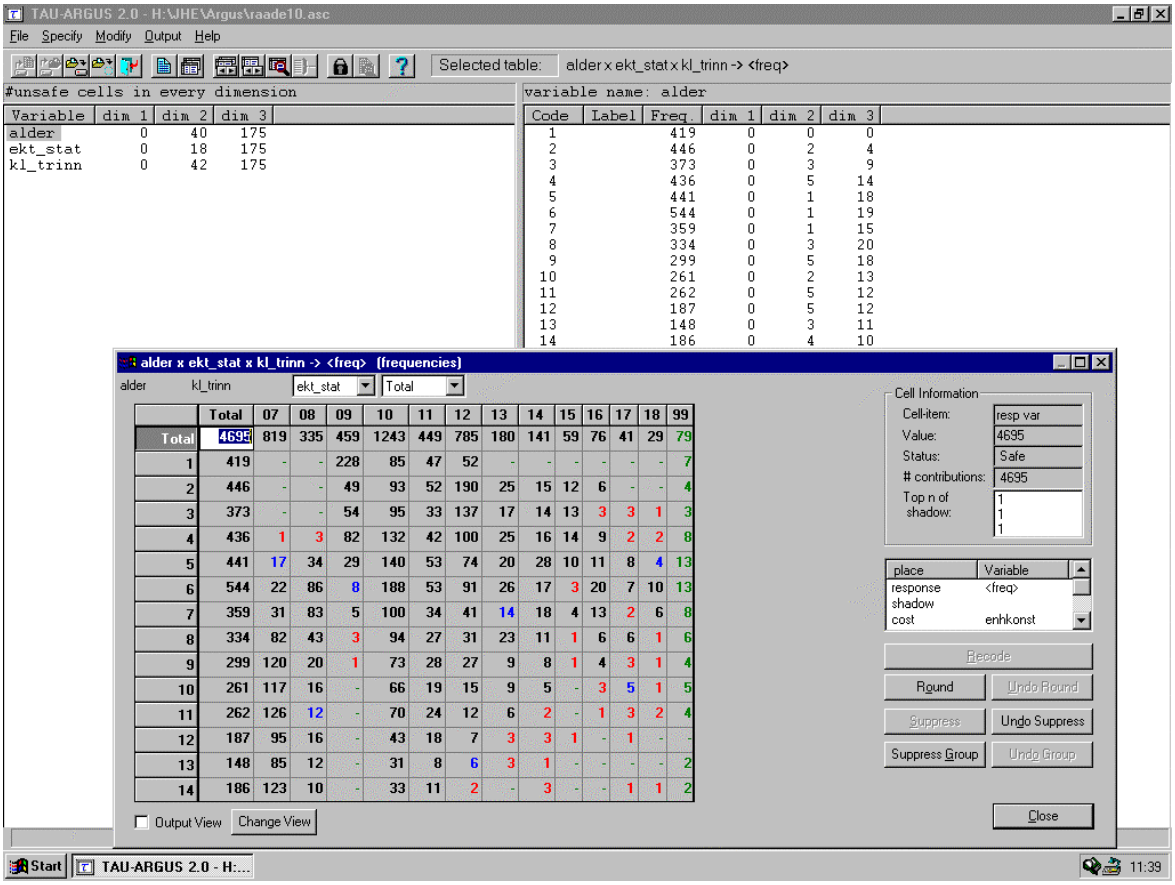
Vi ser av øvre del av panel 4 at nå er

$$x + y = 40$$
$$x + z = 18$$
$$y + z = 42$$

som gir $x = 8,\ y = 32,\ z = 10$. Det var 175 usikre celler i alle tre dimensjonene i en tabell med 910 celler. Dette er fortsatt mye.

Vi kan likevel prøve å foreta prikking og trykker på ”Suppress”. Vi får da spørsmål om ”minimum” og ”maximum” range som i litteraturen kalles ”Feasibility interval”. Det er et mulighetsintervall som et undertrykket tall tillates å ligge i gitt marginalene og synlig innmat i tabellen. Det angis i prosent i forhold til det opprinnelige tallet i cellen. I τ-ARGUS mulighetsintervallet ”by default” fra 70 til 130 % av den opprinnelige verdien. Siden vi her har en ren frekvenstabell bør disse grensene settes større og vi setter 0 og 200%. Det vil gi flere sekundære prikkinger. SUPPRESS med denne opsjonen tok vel 11 minutter når ingen andre programmer kjørte og resulterte i to-veismarginalen *alder* x *klassetrinn* i tabellen i panel 4. Det lille panelet i panel 4 viser marginalen *alder* x *klassetrinn* i 3-veistabellen. De leserne som leser dette direkte fra skjerm vil se de usikre cellene i rødt, de som skal sekundærprikkes i blått og kolonnen med missing i grønt.

Panel 4



Den ferdig prikkede tabellen kan nå lagres.

Hvis man i tillegg foretar RECODE på klassetrinn med omkodingene 1:7-9 år, 2:10 år, 3: 11-12 år, 4:13-16 år og 5:17-18 år, resulterer det i $x = 8,\ y = 3,\ z = 2$. I hele 3-veistabellen var det fortsatt 80 usikre celler av i alt 350 (ikke-missing). "Suppress" produserte 50 sekundærprikkede celler i treveistabellen og 18 i toveismarginalene etter ca. 17 minutters kjøring.

Hvis vi er litt mer liberale og setter "Minimum number of records" lik 2, vil det innebære at vi godtar alle celler med tre eller flere personer som sikre. Vi vil da få $x = 97,\ y = 282,\ z = 6$. I alle tre dimensjoner bir det 607 usikre celler. Aldersomkoding til 14 alderskategorier gir $x = 7,\ y = 20,\ z = 6$ og 149 usikre celler i tre dimensjoner. "Suppress" produserte 61 sekundærprikkinger i tre dimensjoner og 29 i to dimensjoner. Det ble også fortatt omkoding av klassetrinn til fem kategorier. En samlet oversikt over resultatene finnes i tabell 4.

	N=4			N=3		
	Uten omkoding	Alder omkodet	Alder+kl-trinn omkodet.	Uten omkoding	Alder omkodet	Alder+kl-trinn omkodet.
# celler eksl. missing	5395	910	350	5395	910	350
#3D usikre (primære)	697	175	80	607	149	67
# 3D sekundære		57	50		61	54
x / p(x)	125/226	8/14	8/14	97/	7/11	7/11
y/ p(y)	247/444	32/56	3/6	282/	20/26	2/3
z/ p(z)	10/20	10/20	2/3	6/8	6/8	2/3
s(x)/ p(s(x))		16/	14/		8/171	8/171
s(y)/ p(s(y))					13/119	4/67
s(z)/ p(s(z))		5/87	4/115		8/43	4/115

Tabell 8: Oversikt over resultater av prikking vha ARGUS. *x* er antall primærprikkede celler i marginalen alder x sivilstatus, *y* i marginalen alder x klassetrinn og *z* i marginalen sivilstatus x klassetrinn. *p(x), p(y)* og *p(z)* er antall personer i de tilsvarende cellene. *s(x), s(y)* og *s(z)* er antall sekundærprikkede celler. *p(s(x)), p(s(y))* og *p(s(z))* er antall personer i disse cellene.

Tabell 8 inneholder noe mer informasjon enn det som rapportfilene (*.rep) forteller. blant annet sier ikke rapportfilene noe om kostnaden (the cost variable), verdien av den størrelsen som søkes minimert i prikkingen. I disse eksemplene er kostnaden antall personer i de primær eller sekundærprikkede cellene. For de to-dimensjonale marginalene kan disse størrelsene finnes ved å studere skjermbildene av dem mens man er inne, men går så tapt. I noen grad kan de også finnes fra de bevarte tabellene, men de inneholder ikke alle marginaler. Av den grunn er ikke tabell 4 helt fullstendig.

8.2 Prikking av en mengdetabell

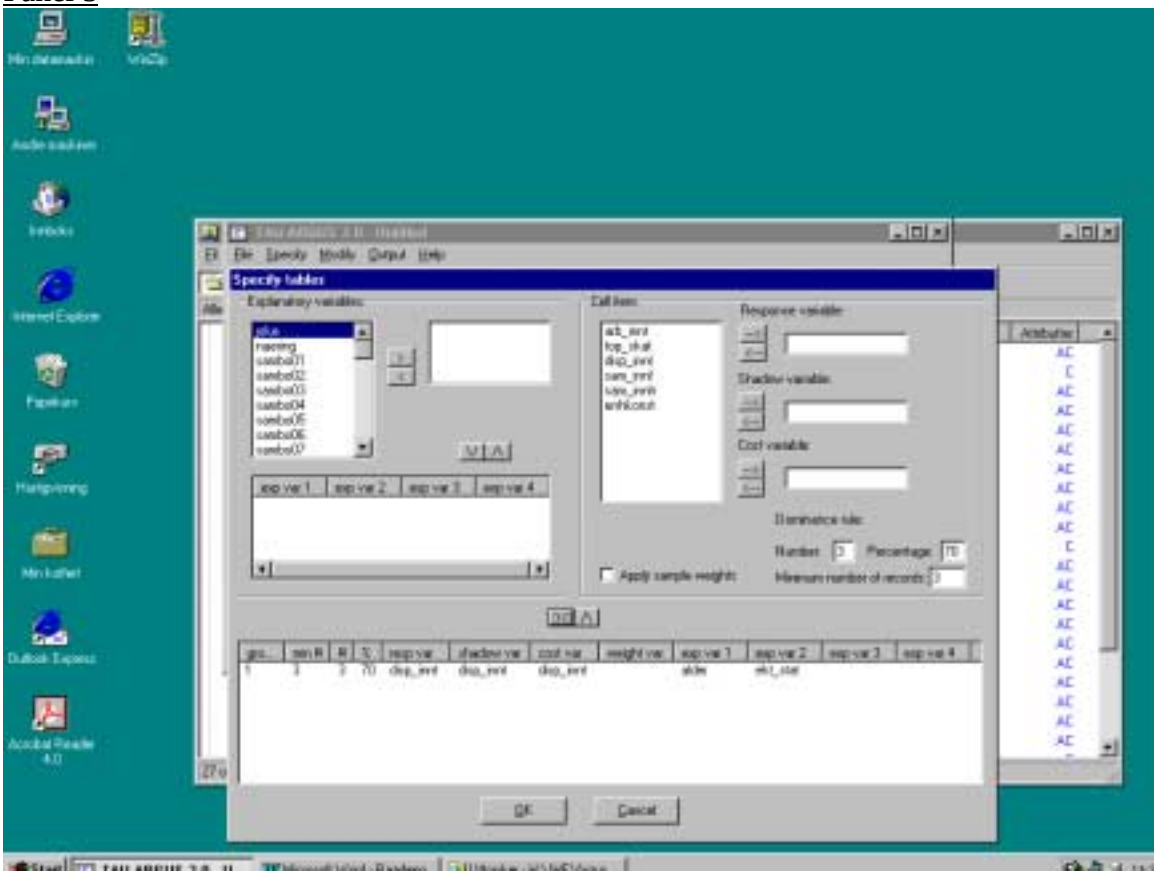
Vi går nå over til å studere en tabell med en mengdevariable, *disponibel inntekt*, i cellene for en krysstabulering av *alder* x *sivilstatus*. Disponibel inntekt brukes både som respons, skygge og kostvariabel. "Respons" variabelen er det samme som mengdevariabelen og er den som skal aggregeres i cellene. Skyggevariablene er den som benyttes for å beregne kriterier for hvorvidt en celle er sikker eller ikke. Vanligvis er dette responsvariabelen, men det kan også

være en annen. Vi tillater at de tre største disponible inntektene i hver celle utgjør høyst 70% av cellens innhold. Kostvariabelen er den hvis sum skal minimeres i de prikkede cellene. Se panel 5.

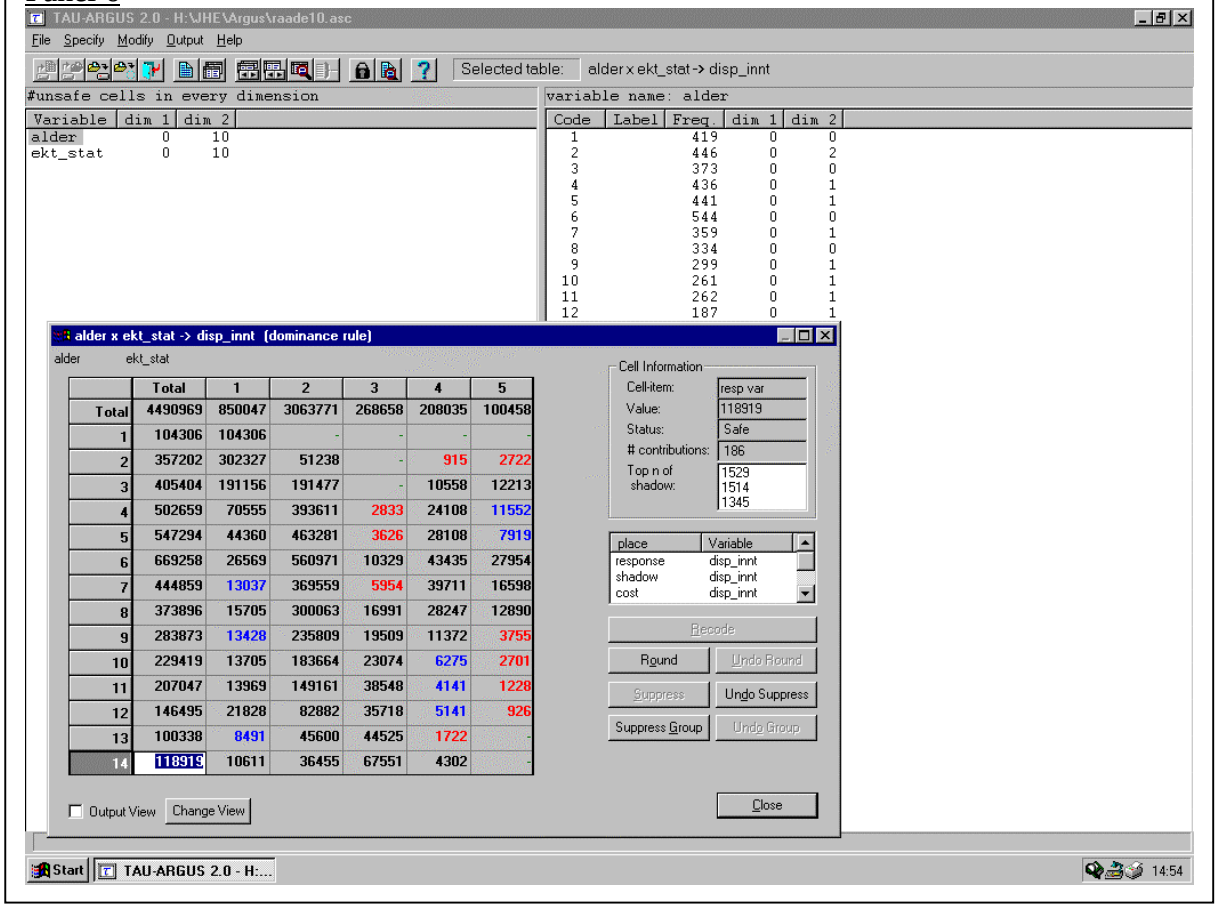
Med ett-årige aldersgrupper får tabellen 425 celler i to dimensjoner. Ut fra de gitte kriteriene definerer ARGUS 158 av disse som usikre og fortsatt syv usikre i kategorier i aldersmarginalen. Det er for mye. Vi gjør samme aldersomkoding som før. Det resulterer i ti usikre celler av

70 som vist under. åtte celler måtte sekundærundertrykkes. Panel 6 viser resultetet. Igjen fremtrer usikre, og sekundærprikkede celler i ulike farger på skjermen.

Panel 5



Panel 6



Den prikkede tabellen lagres på fil i formatet i panel 7. ”s” angir ”Suppressed” celle.

Panel 7

ekt_stat	total	1	2	3	4	5
alder						
total	4490969	850047	3063771	268658	208035	100458
1	104306	104306	0	0	0	0
2	357202	302327	51238	0	s	s
3	405404	191156	191477	0	10558	12213
4	502659	70555	393611	s	24108	s
5	547294	44360	463281	s	28108	s
6	669258	26569	560971	10329	43435	27954
7	444859	s	369559	s	39711	16598
8	373896	15705	300063	16991	28247	12890
9	283873	s	235809	19509	11372	s
10	229419	13705	183664	23074	s	s
11	207047	13969	149161	38548	s	s
12	146495	21828	82882	35718	s	s
13	100338	s	45600	44525	s	0
14	118919	10611	36455	67551	4302	0

8.3 En vanskelig to-veistabell

Tabellen som skal betraktes her er tar utgangspunkt i data for terminvis omsetningsstatistikk for Sogn- og Fjordane fylke. Tabellen lages i utgangspunktet for opptil fem siffer næringskode med 56 kategorier for fylket totalt og for 26 kommuner i fylket. Tabellen inneholder da 1456 celler hvorav mange er tomme. Svært mange celler prikkes. Dersom det legges til grunn at en celle må inneholde minst tre bedrifter for å være "sikker", vil ARGUS definere 360 celler som usikre, 85 som sikre mens 1011 er tomme. Tabellen under viser hvordan dette endrer seg når en reduserer antall sifre i næringskoden. Først med to sifre (50 og 52) vil tabellen bli håndterlig for automatisk prikking. Men da forsvinner enhver fordeling av omsetningstall i syv kommuner.

	Kjøring 1	Kjøring 2	Kjøring 3	Kjøring 4
Celler i alt	1456	832	312	52
Nullceller	1011	520		1
Sikre	85	61		
Primære prikkinger i 2.dim	424	251	104	7
Primære p i 1.dim	15	9	1	0
Siffer i næring	5	4	3	2

Tabell 9: Resultater av kjøringer av en vanskelig tabell.

En alternativ omkoding av næring i fem kategorier etter følgende nøkkel

```
500:50101-50500
521:52110-52120
52A:52200-52399
524:52400-52499
52B:52500-52799
```

produserte en tabell med 130 celler, hvorav 31 var usikre, 18 var 0 og resten, 81 celler var sikre. Ved forsøk på å anvende SUPPRESS på denne tabellen svarte ARGUS at det ikke var mulig å sikre tabellen på grunn av en eller flere celler. ARGUS opplyste ikke hvilke celler det var. Forsøk på å slå sammen kommuner hjalp ikke.

8.4 ARGUS i et produksjonsopplegg

Programmet τ-ARGUS må ha mikrodata som input, dvs. at ved bruk av programmet i et produksjonssystem, må τ-ARGUS kjøres *før* produksjon av layout-ferdige tabeller i SAS. I sin nåværende form kan vi med τ-ARGUS kun produsere tabeller interaktivt, dvs. at det ikke er mulig å masseprodusere tabeller automatisk. I tillegg er programmet på PC-plattform slik at mikrodata må overføres fra unix-plattform til PC-plattform. Likevel kunne det tenkes bruk av τ-ARGUS for produksjon av spesielle tabeller. Output fra τ-ARGUS er en ascii-tabell med forspalte og tabellhode, men uten tabellhodetekst. Se panel 8. Den må greit kunne hentes inn i et videre bearbeidingssystem, som f.eks. SAS, for å få et standardisert utseende med riktige forspalter og tabellhoder.

8.5 Mikrodatasett

Det vil i dette notatet føre for langt også å gi eksempler på bruk av µ-ARGUS for sikring av mikrodatasett. Det er bedre at den interesserte leser prøver programmet på egen hånd. Metodene som finnes omfatter automatisk og interaktiv global omkoding og lokal sensurering, topp- og bunnkoding. Det er også mulig å støylegge variable og å foreta avrundinger. µ-ARGUS analyserer i hvilken grad dette er nødvendig basert på tabeller som kan spesifiseres manuelt eller ved automatiske rutiner. Manualer er tilgjengelige på internett.

Introduksjonen er nå gitt. Det er nå opp til leseren å fordype seg videre:

Referanseliste

Bethlehem, J.G., Keller, W.J. og Pannekoek, J. (1990): *Disclosure Control of Microdata*. J. of the Am. Stat. Assoc., vol 85 no. 409 s.38-45.

Chen, G. og Keller McNulty, S. (1998). *Estimation of Identification Disclosure Risk in Microdata.* Journal of Official Statistics, 14, s79-95.

Cox, L. H. og Sande, G. (1979): *Techniques for Preserving Statistical Confidentiality*. I Proceedings of the 42nd Meetings of the International Statistical Institute (Vol. 3), s. 499-512.

Dalenius, T. (1974): The invasion of Privacy Problem and Statistics Production - An Overview. Statistisk tidsskrift , 3 s. 377-385

Dalenius, T. (1977): *Toward a methodology for statistical disclosure control*. Statistisk tidsskrift 1977:5 s428-444.

Defays, D. og Anwar, M.N. (1998): *Masking Microdata using Micro-Aggregation*. Journal of Official Statistics,vol 14 no. 4 s.449-461

Diaconis, P. og Sturmfels, B (1998): *Algebraic algorithms for sampling from conditional distributions.* Annals of Statistics, 26 s.363-397

Duncan, G. T. og Lambert, D. (1986): *Disclosure-Limited Data Dissemination.* J. of the Am. Stat. Assoc. vol 81 no. 393 s 10-18

_____ (1989): *Risk of Disclosure for Microdata.* J. of Business & Economic Statistics, Vol 7., no. 2 s. 207-217

Fellegi, I. P. (1972): *On the Question of Statistical Confidentiality*. J. of the Am. Stat. Assoc., vol 67 no. 337 s.7-18.

Fienberg, S. E., Makov, U. E. og Steel, R. J. (1998): *Disclosure Limitation Using Perturbation and Related Methods for Categorical Data Analysis.* Journal of Official Statistics,vol 14 no. 4 s.485-512 (med diskusjon)

Fuller, W. A. (1993): *Masking Procedures for Microdata Disclosure Limitation.* Journal of Official Statistics, vol 9 no. 2 s. 383-406.

Gates, G.W. (1999): *Making Data more Accessible in a Climate where perception matters.* In Statistical Data Confidentiality, Proceedings of the Joint Eurostat/UN-ECE work session on Statistical data Confidentiality, Tessaloniki Mach 1999.

Gouweleeuw, J. M., Kooman, P., Willenborg, L.C.R.J. og de Wolf, P.-P*.: Post Randomisation for Statistical Disclosure Control: Theory and Implementation*. Journal of Official Statistics, vol 14 no. 4 s. 463-478

Horm, J. (1999): *National Center of Health Statistics approaches to protection of microdata*. In Statistical Data Confidentiality, Proceedings of the Joint Eurostat/UN-ECE work session on Statistical data Confidentiality, Tessaloniki Mach 1999.

Hurkens, C.A.J. og Tiourine, S.R. (1998*): Models and Methods for the Microdata Protection Problem.* Journal of Official Statistics, 14, s.437-447.

Keller-McNulty, S. og Unger, E. A. (1998): *A Database System Prototype for Remote Access to Information Based on Confidential Data.* Journal of Official Statistics, vol. 14 no. 4 s.347-360.

Little, R.J.A.: *Statistical Analysis of Masked Data*. Journal of Official Statistics, vol 9 no. 2 s. 407-426.

Lov om offisiell statistikk og Statistisk sentralbyrå, 16. juni 1989 nr. 54. http://www.ssb.no/omssb/statlov/

Paas, G. (1988): *Disclosure Risk and Disclosure Avoidance for Microdata*. J. of Business & Economic Statistics, Vol. 6., no. 4 s. 487-500.

Paas, G. og Waushkuhn, U. (1985): Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von Anonymisierten Individualdaten. München: Oldenburg Verlag

Skinner, C.J. and Holmes, D.J. (1993): *Modelling Population Uniqueness*. Proceedings of the International Seminar on Confidentiality, Dublin, s. 175-199.

_____ (1998): *Estimating the Re-identification Risk Per Record in Microdata.* Journal of Official Statistics, 14, s361-372.

Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). *Disclosure Control for Census Microdata.* Journal of Official Statistics, 10, s31-51.

Spruill, N. L. (1982): *Measures of Confidentiality*. i Statistics of Income and Related Administrative Record Research: 1982. Washington DC: US Dept. of Treasury, Internal Revenue Service, Statistics of Income Div. s 131-136.

_____(1983): *The Confidentiality and Analytic Usefulness of Masked Business Microdata.* Proceedings of the Section on Survey Research Methods, American Statistical Assoc. s 602-607

Strudler, M., Oh, H. L. og Scheuren, F. (1986): *Protection of Taxpayers Confidentiality With Respect to the Tax Model*. I Proceedings of the Section on Survey Research Methods, American Statistical Assoc. s 602-607

Subcommittee on Disclosure Avoidance Techniques (Federal Committee on Statistical Methodology) (1978), Statistical Working Paper 2 (Federal Policy and Standards), Washington DC: U.S. Dept. of Commerce.

Sullivan, G. og Fuller, W.A. (1989): The Use of Measurement Error to avoid Disclosure. Proceedings of the Section on Survey Research Methods. American Statistical Association, s. 802-807.

_____(1990): *Construction of Masking Error for Categorical Variables.* Proceedings of the Section on Survey Research Methods, American Statistical Association, s. 435-439.

de Waal, T. og Willenborg, L. (1998): *Optimal Local Suppression of Microdata*. Journal of Official Statistics, 14, s.421-435.

Willenborg, L. og de Waal, T. (1996): *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics no. 111. Springer Verlag.

De sist utgitte publikasjonene i serien Notater

2000/77 P.O. Lande og J. Kittelsen: Forbruksundersøkinga 2000. Innlasting/Innsjekking: Brukardokumentasjon. 17s.

2000/78 J. Fosen, A.K. Johnsen og G. Røyne: Frafall blant innvandrere. En undersøkelse av frafall i Utdanningsundersøkelsen 1999 og i valgundersøkelser blant innvandrere. 53s.

2000/79 J. Kittelsen og P.O. Lande: OPPSLAG - Forbruksundersøkelsen. Brukerdokumentasjon. 39s.

2000/80 J. Kittelsen og P. O. Lande: Forbruksundersøkinga 2000. Systemdokumentasjon. 156s.

2000/81 J.T. Lind: Testing av stokastiske individuelle effekter i paneldatamodeller. 17s.

2001/2 D.Q. Pham: Innføring i tidsserier - sesongjustering og X-12-AMIRA. 110s.

2001/3 O. Rognstad: Eiendomsomsetning. Dokumentasjon av datagrunnlag og bearbeidingsrutine. 72s.

2001/4 T. Nøtnæs: Innføring i kognitiv kartlegging. 20s.

2001/5 T. Bye, M. Hansen og B. Strøm: Hvordan framskrive utslipp av klimagasser? 16s.

2001/6 A. Langørgen og R. Aaberge: KOMMODE II estimert på data for 1998. 16s.

2001/7 B.R. Joneid og J. Lajord: FD - Trygd: Dokumentasjonsrapport. Stønader til enslig forsørger. 1992-1999. 39s.

2001/8 T. Karlsen, E. Karstensen og E. Evensen: Beregningsrutiner og teknisk programstruktur for fylkesfordelt nasjonalregnskap. 27s.

2001/9 L. Rognstad, N.M. Stølen, T. Jakobsen og P. Schøning: Regional statistikk og analyse - strategi og prioriteringer. 45s.

2001/10 A. Akselsen og B.R. Joneid: FD - Trygd: Dokumentasjonsrapport. Pensjoner. Grunn- og hjelpestønader. 1992-1998. 94s.

2001/11 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 1999. 34s.

2001/12 A. Rognan og N. Barrabés: NUS2000. Dokumentasjonsrapport. 36s.

2001/13 K.I. Bøe, J. Johansen og Ø. Sivertstøl: FD - Trygd: Dokumentasjonsrapport. Attføringspenger, 1992-1998. 88s.

2001/14 O. Klungsøyr: Ekstremverdimodell for industrinæringenes investeringer i 90-årene. 30s.

2001/15 O. Klungsøyr: Markovkjede Monte Carlo i varianstkomponentmodell for sysselsettingsdata. 30s.

2001/16 M. Bråthen og T. Pedersen: Tilpasning på arbeidsmarkedet for personer som går ut av status som yrkeshemmet i SOFA- søkerregisteret - 1998. 27s.

2001/17 T. Martinsen: Statistikk over energibruk i Statistisk sentralbyrå - evaluering, brukerbehov og forutsetninger. 87s.

2001/18 L. Vågane: Undersøkelse om holdninger til frukt- og grøntabonnement blant foreldre med barn i grunnskolen. Dokumentasjonsrapport. 26s.

2001/19 H. Madsen og A. Langørgen: Anslag over antall etterspørrere av grunnskoleopplæring for voksne. 23s.

2001/20 B. Indahl, D.E. Sommervoll og J. Aasness: Virkninger på forbruksmønster, levestandard og klimagassutslipp av endringer i konsumentpriser. 27s.

2001/21 A. Barstad: På vei mot det gode samfunn? Utredning til Finansdepartementet i forbindelse med arbeidet med nytt Langtidsprogram, 2002-2005. 363s.

2001/23 L. Østby: Beskrivelse av nyankomne flykningers vei inn i det norske samfunnet. Notat til Lovutvalget som skal utrede og lage forslag til lovgivning om stønad for nyankomne innvandrere. 32s.

2001/24 T. Nøtnæs: Innføring i bruk av fokusgrupper. 22s.