Statistics Norway
Statistisk sentralbyrå

Johan Fosen og Leiv Solheim

Avledede variable i registerstatistikk

To metoder for klassifikasjon av sysselsettingsstatus

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Innhold

1 Innledning

FoB2001 er en registertelling, dvs. at i stedet for å innhente opplysninger fra et spørreskjema for en bestemt person, hentes dette fra ulike administrative registre. Registerverdiene (slik de forekommer i registrene som hentes inn) er imidlertid av varierende kvalitet. For sysselsettingsstatus (sysselsatt/ikke sysselsatt) er registervariabelen av så dårlig kvalitet, at en omfattende bearbeiding av registerinformasjon er nødvendig. I sluttkant av denne prosessen avleder man en ny registerbasert variabel, en klassifisert sysselsettingsstatus. I Seksjon for arbeidsmarkedsstatistikk (S260) har man implementert en metode (Bråthen & Fosen 1998) som vi vil betegne den deteministiske klassifikasjonsmetoden. Vi vil i dette notatet innføre en helt annen type klassifikasjonsmetode, en stokastisk metode, der det i klassifikasjonen er et tilfeldig element. Sistnevnte metode har en del positive sider fra metodestatistisk synsvinkel, og vi vil betrakte deterministisk metode i lys av denne metoden, samtidig som vi ser nærmere på heldige og uheldige sider ved begge metoder. Vi vil i løpet av notatet vise at deterministisk metode er en rimelig metode til tross for sine svakheter, og at den er å foretrekke fremfor den stokastiske metoden vi presenterer her.

I kapittel 2 definerer vi en del begreper, og vi identifiserer hva slags kvalitet vi ønsker å undersøke, før vi i kapittel 3 gir en skjematisk beskrivelse av registerbeidingsprosessen. I kapittel 4 definerer vi sysselsetting og ser på definisjonen bak den opprinnelige registerverdien av sysselsetting.

Kapittel 5 motiverer på en enkel måte den deterministiske metode ved å se på hvordan man kan gå fram for å forbedre den operasjonelle definisjonen av sysselsetting som er innebygget i registerfilen. Både i kapittel 4 og 5 vil beskrivelsene av registrene være forenklede for oversiktens skyld. Først i kapittel 6 beskriver vi mer i detalj konstruksjonen av den deterministiske metoden, og i samme kapittel presenterer vi den stokastiske metoden.

Ulike konsekvenser av de ulike metodene studeres i kapittel 7, og en empirisk sammenlikning utføres. Med unntak av avsnittene om varians og mikrokonsistens, og enkelte av vedleggene, har vi forsøkt å unngå matematisk notasjon mest mulig.

2 Avgrensninger og definisjoner

Vi vil benytte termen "kjennemerke" om egenskapene vi ønsker å måle i FoB, f.eks. sysselsetting, utdanning osv. Vi vil med "registerdefinisjonen av kjennemerke" mene den *operasjonelle* definisjon av kjennemerket som ligger bak den registrerte versjon av kjennemerket (registerverdi), mens "FoB-definisjonen" angir hva som skal være meningsinnholdet i begrepet i FoB, altså en normativ nominaldefinisjon[1].

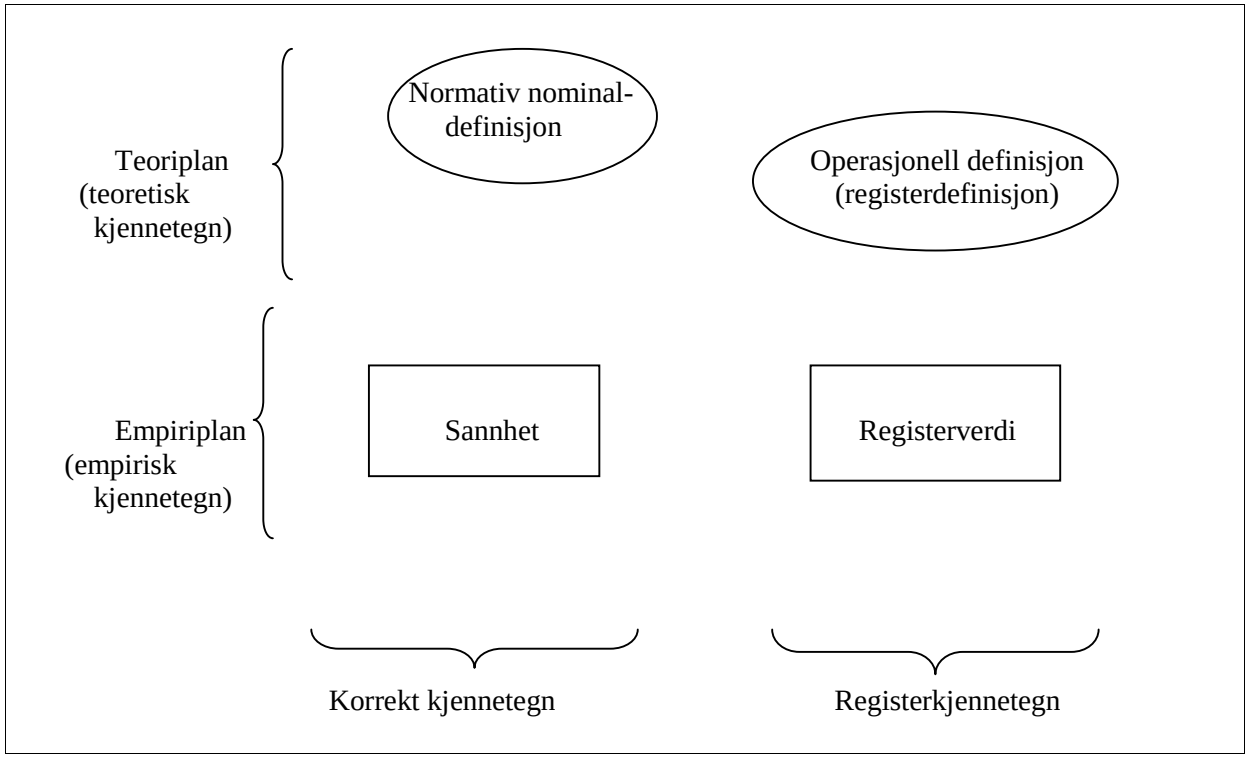
I dette notatet vil vi konsentrere oss om tilstandene sysselsatt/ikke-sysselsatt. Vi kaller det tilhørende kjennemerket for sysselsettingsstatus, og dette er en aggregering av det egentlige FoB-kjennemerket arbeidsstyrkestatus som har verdiene sysselsatt/ledig/utenfor arbeidsstyrken. Sysselsetting kan også deles videre opp i lønnstakere, selvstendige og familiearbeidere, og disse tre verdiene utgjør kjennemerket yrkesstatus.

Kjennemerket sysselsettingsstatus forekommer på flere plan: vi har begrepet, empirisk variabel og verdien som variabelen antar. Vi vil referere til "sysselsettingsstatus" i alle tre situasjoner, og det vil framgå av sammenhengen hva vi snakker om. Den opprinnelige empiriske sysselsettingsstatus slik variabelen framkommer i registeret, betegner vi med "opprinnelig registersysselsettingsstatus", mens nye forslag til variabel basert på registerinformasjon, vil bli betegnet med "klassifisert sysselsettingsstatus". Man benytter generelt termen "predikere" når man skal anslå en verdi for personer utenfor utvalget, og det å klassifisere er det spesialtilfelle av prediksjon at man predikerer hver person inn i en klasse, der klassene her er sysselsatt og ikke-sysselsatt.

[1] Ofte benyttes "teoretisk definisjon" i stedet for nominaldefinisjon, men alle definisjoner kan betraktes som teoretiske.

Dette notatet handler om to måter å gå fram på for å finne klassifisert sysselsettingsstatus. Vi vil referere til metodene som den "deterministiske metode" og den "stokastiske metode", der sistnevnte metode refererer til at det skjer en eller annen form for tilfeldig trekking. Det kan naturligvis tenkes en rekke andre både deterministiske og stokastiske metoder.

Figur 1 Oversikt over hvilke plan ulike typer definisjoner hører hjemme.



Alle filer i dette notatet er mikrodata, som i personstatistikken betyr at person (eller lavere nivå) er enhet. Med "register" vil vi mene "råregisteret" slik som SSB mottar det, mens filer som er bearbeidede i SSB betegnes som "statistikkfiler". Vi vil i dette notatet referere til både informasjonen fra registeret og fra statistikkfiler som registerinformasjon.

Klassifikasjon av sysselsettingsstatus kan betraktes som siste fase i bearbeidingsfasen av registerinformasjon. Klassifikasjonen kan betraktes som en avledning av allerede eksisterende registerinformasjon, men med det tillegg for stokastisk metode at en tilfeldig komponent inngår i formelen som utfra eksisterende registerinformasjon hos en person, gir en klassifisert sysselsettingsstatus.

Enkelte definisjoner kommer senere i notatet etter hvert som nødvendig bakgrunn er etablert, de fleste i kapittel 4 og i 6.1.1.

2.1 Kvalitet

Eurostat har laget et utkast til definisjon av kvalitet der det inngår syv komponenter (en norsk oversettelse: Stålnacke et. al 1999). Den komponenten vi vil se på her er nøyaktighet, som er definert som avviket mellom den målte verdien og sannheten. Begrepet nøyaktighet sier altså hva vi vanligvis ville betegne som hvor lite feil noe er. Vi vil dele den opp 'nøyaktighet' i to under-komponenter. Den

ene er hvor godt den operasjonelle definisjonen stemmer med den normative nominaldefinisjonen[2]. Dette kalles definisjonsmessig validitet, og er i vårt tilfelle forskjellen mellom registerdefinisjonen (operasjonell) og FoB-definisjonen.

Den andre underkomponenten av nøyaktigheten er reliabilitet, og sier hvor godt samsvar det er mellom selve dataene (empirisk sysselsettingsstatus) og registerdefinisjonen. Her vil dette si forskjellen mellom selve statistikkfilen og hva som skulle stått i filen dersom alle prosesser i registerføring og -bearbeiding var blitt gjort fullstendig korrekt. Dersom det ikke er gjort noen feil i danning og bearbeiding av registeret, er reliabiliteten perfekt.[3]

3 Registerbearbeiding

FoB-individfilen er et resultat av en lang bearbeidingsprosess. I bedriftene registreres bl.a. informasjon om arbeidstakere og dette innrapporteres til det kommunale trygdekontor. Informasjonen går videre til fylkestrygdekontor og til slutt samles det i Rikstrygdeverket (RTV) under navnet Arbeidstakerregisteret. SSB innhenter kopi av Arbeidstakerregisteret og deretter foregår en intern bearbeiding i SSB i Seksjon for arbeidsmarkedsstatistikk (S260). Etter den interne bearbeidingen har man en statistikkfil, og denne og andre statistikkfiler dannet på tilsvarende måte fra andre kilder samles, etter kopling av statistikkfiler og enda mer bearbeiding, til FoB-individfilen som Seksjon for folke- og boligtelling (S370) har ansvaret for.

Den ene typen kjennemerker er der man etter en mindre bearbeiding i form av noe konsistensbehandling m.m. i fagskesjonen, har etablert en variabel i registeret der definisjonsmessig validitet er god, og man vet at reliabiliteten til variabelen er god. For disse kjennemerkene er utfordringen å måle nøyaktigheten (som kanskje kun vil være reliabiliteten), noe vi vil komme tilbake til i et annet notat.

Den andre typen kjennemerker er der hvor registervariabelen selv etter en del bearbeiding ikke innebærer at definisjonsmessig validitet er akseptabel, og/eller at reliabiliteten ikke er god nok. Et eksempel er sysselsetting, der registerdefinisjonen er strengere enn FoB-definisjonen. Klassifikasjon av sysselsettingsstatus er derfor nødvendig, og vi vil i det følgende gå inn i siste fase av bearbeidingen i S260. Det er på dette trinnet klassifikasjonsmetoden for sysselsettingsstatus kommer inn.

3.1 Reliabiliteten til registerfiler

Det er et generelt problem ved reliabiliteten til registerfiler at registerfører har et administrativt styrt motiv for nøyaktigheten av registrering. Dette betyr f.eks. at registeret er nøyaktig ført i den grad det oppfyller den administrative bruk. Som et eksempel har vi LTO-lønn (en del av Lønns- og trekkoppgaveregisteret), der datering for start og slutt på jobbforhold er registrert, men kvaliteten på dateringen er dårlig. Dette kan skyldes at detaljert datering kanskje er av mindre interesse for den administrative bruk.

4 Nominal- og operasjonell definisjon av sysselsetting. Utfordringer

FoB-definisjonene er stort sett identisk med internasjonale anbefalinger. For arbeidsmarkeds-kjennetegn er grunnlaget "System of National Accounts 1993" (Eurostat et. al 1993) utarbeidet i samarbeid mellom FN, Eurostat, IMF[4], Verdensbanken og OECD. For sysselsetting er denne definisjonen i overensstemmelse med ILO[5]-anbefalinger. Definisjonen sier at alle som arbeider minst

[2] Definisjonen av hvordan begrepet bør forstås.

[3] Begrepene defnisjonsmessig validitet og reliabilitet er beskrevet mer utførlig i f.eks. Hellevik (1991).

[4] Det internasjonale pengefondet

[5] Den internasjonale arbeidsorganisasjonen.

en time i løpet av måleuka, skal regnes som sysselsatte, også de som er midlertidig fraværende fra slik jobb pga. sykdom, permisjon, ferie,streik eller lock-out (beskrevet detaljert i UN economic commission for Europe & Eurostat 1997)

Det opprinnelige registervariabelen for sysselsettingsstatus er sysselsatte lønnstakere, og dette kjennetegnet finnes implisitt i Arbeidstakerregisteret/Arbeidsgiverregisteret (A/A) som alle som er registrert med et aktivt jobbforhold i den aktuelle uka. Den opprinnelige registerdefinisjonen av sysselsatte lønnstakere er strengere enn FoB-definisjonen, ved at kravet for å være registrert med aktivt arbeidstakerforhold (og dermed være sysselsatt) er at gjennomsnittlig avtalt arbeidstid er minst 4 timer per uke, at arbeidsforholdet strekker seg over et visst tidsrom, og at denne jobben er aktiv i den aktuelle uka. Dette innebærer at f.eks. fast ansatte med arbeidstid en halv dag i uka (3,75 timer) ikke kommer med, og likeledes alle som er sysselsatt som selvstendige.

Egentlig kan det diskuteres om A/A definerer sysselsatte lønnstakere, eller om registeret heller definerer lønnstakere med en jobb av et visst omfang. For vårt formål er valget mellom disse to et spørsmål om hensiktsmessighet, og vi velger den førstnevnte betraktning. Vi vil også gå et skritt videre, og i fortsettelsen betrakte registerdefinisjonen av sysselsatte lønnstakere som registerdefinisjonen for sysselsatte.

Vi vil se på metoder for klassifikasjon av sysselsettingsstatus der vi tar utgangspunkt, iallfall for deterministisk metode, i et forsøk på å få en bedre definisjonsmessig validitet enn den opprinnelige registerdefinisjonen. For begge metoder vil vi se på deres evne til å takle kravet til nøyaktighet totalt, og ikke bare den definisjonsmesige validiteten.

En innfallsvinkel til en bedre definisjonsmessig validitet er å forsøke å innføre en justert registerdefinisjon som kan erstatte den opprinnelige registerdefinisjonen. Det er resultatet av denne justerte registerdefinisjonen som vi vil kalle for deterministisk klassifisert sysselsettingsstatus, og metoden for deterministisk metode. En grafisk skisse av problemstillingen finnes i figur 2.

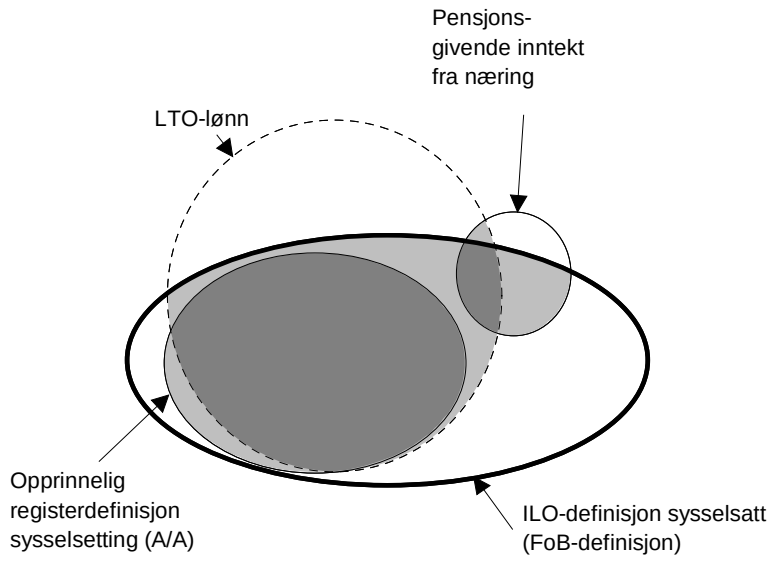
I Figur 2a) betegner hver sirkel en bestemt egenskap, og sirkelen inneholder alle personer som skal ha denne egenskapen. Innenfor den kraftig heltrukne sirkelenen er alle personer som skal være (og er)[6] sysselsatte etter FoB-definisjonen. Sirkelen for opprinnelig registerdefinisjon av sysselsetting, dvs. A/A-mengden, inneholder alle personer som iflg. denne definisjonen skal være sysselsatte. Vi ser at A/A-mengden utelukkende ligger innenfor FoB-definisjonen, men A/A-mengden er mye strengere. Altså fanger A/A opp altfor få personer.

Tilsvarende har vi for figur 2b) at hver sirkel nå indikerer hvilke personer som faktisk har de aktuelle egenskapene. Vi ser at sirkelen for registerdefinisjon av sysselsetting, ikke lenger er fullstendig inkludert i FoB-definisjonen. For de andre sirklene har vi for enkelhetsskyld tegnet dem omtrent identisk med deres posisjon i figur 2a), men dette betyr altså ikke at de gjelder eksakt de samme personene.

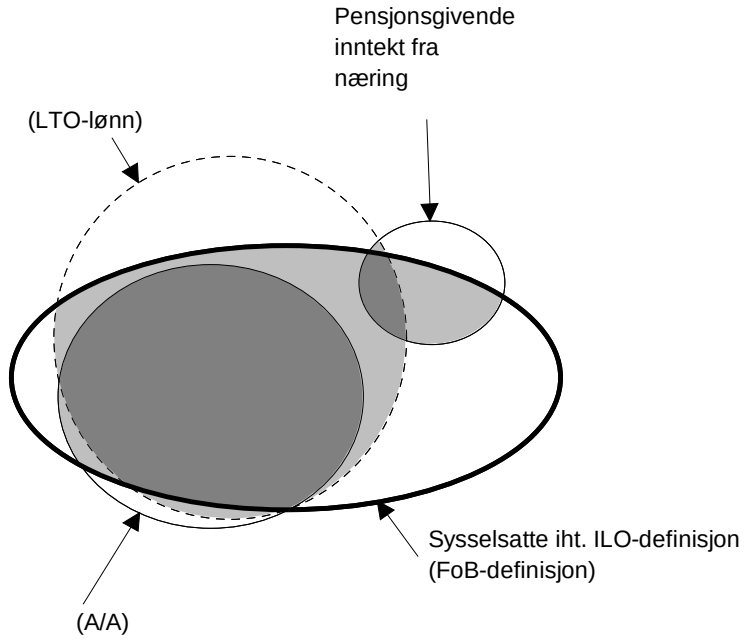
Forskjellen mellom figur 2a) og figur 2b) illustrerer reliabiliten, dvs. den delen av avviket mellom sannhet (FoB-definisjonen) og faktiske registerverdier, som ikke skyldes den manglende definisjonsmessige validiteten.

[6] For en normativ nominaldefinisjon er det per definisjon overensstemmelse mellom "skal være" og "er" .

Figur 2 En skjematisk skisse av ILO-definisjon (FoB-definisjon) av sysselsetting, opprinnelig registerdefinisjon av sysselsetting (A/A), og registerdefinisjon av pensjonsgivende inntekt og lønnsforhold (LTO-lønn). Hver mengde danner en kombinasjon av egenskaper, og hver mengde inneholder de personer som skal ha/har disse egenskapene. Overlapp mellom A/A og pensjonsgivende inntekt fra næring er ikke tatt med i figuren.



(a) skisse av situasjonen slik den skal være i henhold til registerdefinisjonen



(b) skisse av situasjonen slik den faktisk er i registrene

5 Modellering av sysselsetting

Nedenfor vil vi skissere hvordan vi vil gå fram for å få en justert registerdefinisjon som er mer i overensstemmelse med sannheten (FoB-definisjonen), før vi i kapittel 6 mer detaljert beskriver metoden. Presentasjonen i dette kapittelet vil innebære at vi beskriver registersituasjonen på en forenklet måte.

Det er nøyaktigheten (figur 2b) vi til slutt er opptatt av å få best mulig, men likevel slik at det ikke strider mot ønsket om at definisjonsmessig validitet for justert registerdefinisjon skal være best mulig (figur 2a). Registerdefinisjonen sier nemlig noe om intensjonen bak registeret, og dermed noe om hvordan registeret vil utvikle seg over tid. Ettersom figur 2b) ikke avviker mye fra figur 2a), velger vi i fortsettelsen å se bort fra figur 2a).

Med utgangspunkt i figur 2b) vil vi forsøke å finne mengder (grupper) utover A/A-mengden som vi kan klassifisere som sysselsatte, på en slik måte at avviket mellom mengden av de som er klassifiserte som sysselsatte og mengden som betegner FoB-definisjonen blir mindre enn med den opprinnelige registerdefinisjonen (A/A).[7] Vi kan se at mengden "pensjonsgivende inntekt fra næring" oppfyller dette ønsket til en viss grad, mens LTO-lønn (jobb med registrert lønnsutbetaling) er lite hensiktsmessig for å utvide fra A/A-mengden. LTO-lønn-mengden inkluderer en mengde personer som absolutt ikke er sysselsatte, enten ved at forholdet bare er feriepenger eller ved at dateringen i LTO-lønn er feil og jobben dermed egentlig gjelder for et annet tidspunkt.

Vi vil betrakte modeller som innebærer en gruppering som skissert ovenfor, dvs. at vi i relasjon til sysselsetting danner homogene ikke-overlappende grupper: en klassifikasjonsgruppe skal i størst mulig grad enten befinne seg utelukkende innenfor eller utelukkende utenfor FoB-sysselsetting-mengden. Dette er det samme som at andelen sysselsatte i hver klassifikasjonsgruppe skal presses nærmest mulig mulig 0 og 1 (homogen gruppering). Det å ta utgangspunkt i en eller annen form for gruppering av personene, vil vi kalle en *grupperingsmodell*, og en slik modell vil på litt ulike måter være utgangspunkt for både den deterministiske og den stokastiske metoden som blir beskrevet i kapittel 6. Vi skal imidlertid senere se at det kan være andre krav som gjør at det ikke er optimalt å følge en homogen gruppering slavisk.

For å kunne finne en god modell, er vi avhengig av å ha en fasit for sysselsettingsstatus (kun slik kan vi identifisere homogene grupper). Til dette formålet har vi AKU (Arbeidskraftundersøkelsen) der man bruker samme definisjon av sysselsetting som FoB (ILO-definisjonen),[8] som vi antar har god reliabilitet[9] slik at den kan fungere som en brukbar fasit. Ved å finne klassifikasjonsgrupper som vi har skissert ovenfor der vi la noen klassifikasjonsgrupper bli klassifisert som sysselsatte og noen som ikke-sysselsatte, har vi skissert en modell mellom registerinformasjon og sysselsettingsstatus, og ovenfor har vi forsøkt å få modellen best mulig i forhold til "fasiten"[10]. Med vår grupperingsmodell vil et godt første utgangspunkt være å finne mest mulig homogene grupper. Hver gruppe er representert i AKU-utvalget, og basert på andel sysselsatte i AKU-delen, kan vi klassifisere resten av gruppen i sysselsatte og ikke-sysselsatte.

Deterministisk metode vil si å bruke idéen ovenfor om å la alle i en klassifikasjonsgruppe enten bli sysselsatte eller ikke-sysselsatte: dersom andelen AKU-sysselsatte i AKU-delen av mengden "pensjonsgivende inntekt fra næring" er høy, lar vi alle i denne mengden (også de som ikke er i AKU-utvalget) bli sysselsatte. Den andre måten å klassifisere på er at vi sikrer at hver mengde får like stor

[7] Figuren skisserer omfanget av ulike operasjonelle definisjoner. ILO-definisjonen er en teoretisk definisjon, men den kan i teorien måles direkte (dersom man spurte folk og de alle var istand til å forstå definisjonen), og omfanget av den i empiri-planet som figur 1 illustrerer er dermed åpenbar.

[8] Det er AKU som er benyttet for å kunne skissere området for FoB-definisjonen i figur 2.

[9] Siden det ikke er definisjonsproblemer i AKU, er reliabilitet det samme som nøyaktighet.

[10] AKU er fasit i den grad reliabiliteten er høy (og ikke svekket pga. målefeil, kodefeil m.m).

andel som er klassifiserte som sysselsatte som gruppas andel av AKU-sysselsatte[11]. Det er den sistnevnte ideen som er grunnlaget bak den stokastiske metoden.

Vi skal i neste kapittel se nærmere på både den deterministiske og stokastiske modellen. Vi vil bl.a. se at klassifikasjonsgruppene i deterministisk metode har en annen rolle enn de tilsvarende gruppene for stokastisk metode, noe som gjør at vi vil gruppere på ulike måter. For sistnevnte metode vil vi i stedet for grupper snakke om etterstrata.

Etter at klassifikasjonen er gjennomført i hele populasjonen, kan man studere AKU-utvalget for å se hvor riktig dette ble iflg AKU. For å gjøre en slik måling, bør man benytte et annet utvalg enn det som ble benyttet for å tilpasse modellen. Årsaken til dette er at modellen på en måte er spesialsydd for det konkrete utvalget man har benyttet i tilpasningen, et utvalg man gjerne kaller treningsdata. Jo mer detaljert modellen er i forhold til utvalgets størrelse, jo større er faren for at spesialsyingen gjør at modellen fungerer mye bedre for treningsdatasettet enn for et annet datasett (valideringsdatasettet). Deterministisk metode har imidlertid så få komponenter som er avhengige av det enkelte utvalget, at faren for overtilpasning er liten.

5.1 Andre modeller

Den stokastiske klassifikasjonsmåten som vi skal se nærmere på senere, tilsvarer regresjon (f.eks. logistisk) dersom man i regresjonen kun har kategoriske forklaringsvariable, og alle samspillsledd tas med i modellen. Logistisk regresjon er presentert i vedlegg 2.

6 De to modellene

Vi har ovenfor skissert en deterministisk og en stokastisk modell som begge er basert på grupperingsmodellen. Vi vil nedenfor presentere disse modellene nærmere, og enkelte av forenklingene vi gjorde i forrige kapittel, vil her bli formulert mer presist og dermed mer korrekt. Hvordan vi skal velge grupperingen i stokastisk metode, vil vi først ta opp i neste kapittel.

6.1 Deterministisk klassifikasjon

Med utgangspunkt i figur 2, er det en naturlig tanke å prøve å endre litt på den opprinnelige registerdefinisjonen (den operasjonelle definisjonen av sysselsetting) slik at den nærmer seg FoB-definisjonen. Resultatet av denne prosessen vil være en slags justert registerdefinisjon, og det empiriske kjennemerket som dette resulterer i vil vi kalle deterministisk klassifisert sysselsettingsstatus. Metoden som skisseres nedenfor er beskrevet mer detaljert i Bråthen & Fosen (1998). Videre har man i S260 tilpasset metoden til bruk på hele populasjonen.

6.1.1 Presiseringer og grunnleggende om registrene

I deterministisk metode spiller arbeidstakerforhold (A/A), lønnsforhold (LTO-lønn) og selvstendigforhold (konstruert utfra pensjonsgivende inntekt fra næring) en sentral rolle. Som en fellesbetegnelse på slike forhold bruker vi "jobbforhold". Videre har arbeidstakerforhold og lønnsforhold fellesbetegnelsen lønnstakerforhold. Med det viktigste lønnstakerforholdet refererer vi til det forholdet som er rangert som det viktigste blant alle slike forhold som personen har og som er aktive i måleuka.[12] Viktigste selvstendigforhold har en analog betydning. I stedet for "måleuka" benyttes ofte "referanseuka". Når vi nedenfor grupperer etter type forhold, refererer vi hele tiden til personenes viktigste lønnstakerforhold eller viktigste selvstendigforhold.

Med utgangspunkt i figur 2b) kan vi identifisere de viktigste typer jobbforhold. Den opprinnelige registerdefinisjonen av sysselsetting (A/A) består av to mengder: den største er der man har overlapping med LTO-lønn. Denne overlappingsmengden er definert ved at alle personer har fått

[11] AKU-sysselsatte i en gruppe er de som har svart at de er sysselsatte i AKU-undersøkelsen.

[12] Rangeringen er gjort i tidligere faser av registerbearbeidingen i S260, og denne rangeringen blir beskrevet i Bråthen & Fosen (1998).

identifisert et viktigste lønnstakerforhold som både finnes i A/A og LTO-lønn.[13] Vi kaller denne mengden for "koplet lønnstakerforhold"-mengden. Mengden der personens viktigste forhold er A/A, men man ikke finner kopling til LTO-lønn, kalles ukoplet arbeidstakerforhold. Begge disse to mengdene kan deles i to: etter om personene også har et viktigste selvstendigforhold eller ikke. Disse to mengdene er ikke tegnet inn i figur 2, og vi vil se bort fra disse fordi de ikke viser seg å være hensiktsmessige (det viser seg at alle med koplet lønnstakerforhold bør klassifiseres som sysselsatte uansett).

Mengden LTO-lønn som ikke overlapper med A/A, består også av to deler: der personen også har et viktigste selvstendigforhold, og der personen ikke har det. Denne overlappingen er forskjellig av natur fra overlappingen vi så mellom A/A og LTO-lønn hvor det dreide seg om samme jobb. LTO-lønn/-selvstendig-overlappingen er mellom to jobber. Uansett så kaller vi den mengden der man kun har LTO-lønn, for "kun ukoplet lønnsforhold", mens overlapp-delen kalles "ukoplet lønnsforhold og selvstendigforhold"-mengden.

Det siste mengden som er indikert i figur 2b) er "kun selvstendigforhold"-mengden. Alle mengdene beskrevet i dette avsnittet er oppsummert i ramme 1.

Ramme 1 Arbeidsmarkedsgrupper definert ved figur 2b)
Alle forhold refererer til personenes viktigste lønnstakerforhold eller selvestendigforhold i referanseuka.

Klassifikasjonsgruppe
Koplet lønnstakerforhold
Ukoplet arbeidstakerforhold
Kun ukoplet lønnsforhold
Ukoplet lønnsforhold og selvstendigforhold
Kun selvstendigforhold

6.1.2 Beskrivelse

Nedenfor følger en beskrivelse av den deterministiske metoden, med vekt på motivasjonen bak hvert trinn i metoden.

Vi ser i figur 2b) at en kandidat til justert registerdefinisjon, er å snevre inn A/A-mengden til "koplet lønnstakerforhold"-mengden. Dette vil si at man krever at et A/A-forhold må ha lønn knyttet til seg for at det skal kvalifisere for sysselsetting. De vi eliminerer ved en slik innsnevring er i større grad ikke-sysselsatte enn sysselsatte.

Vi ser at etter mengden "koplet lønnstakerforhold", er det mengden "kun selvstendig-forhold" og "ukoplet lønnstakerforhold og selvstendigforhold" som i størst grad faller innenfor FoB-definisjonen av sysselsetting. Fordi vi ønsker en bestemt andel selvstendige blant de som er klassifisert som sysselsatte,[14] og fordi det viser seg at de to mengdene med selvstendigforhold er for store, må man la kun en del av disse bli klassifisert som selvstendige. Vi har ingen informasjon fra selvstendigforholdet som gjør oss i stand til å selektere personer som sysselsatte på en fornuftig måte. Ettersom vi skal klassifisere alle i samme gruppe på identisk måte, får vi en indikasjon av figur 2 på at det er fornuftig at alle blir regnet som sysselsatte i "selvstendig"-mengdene (dette ser ut til å gi mest riktig klassifikasjon). Imidlertid vil vi med en slik klassifikasjon få for mange selvstendige, så vi trenger en måte for å skille ut de som i stedet for å være selvstendige skal være lønnstakere, og det naturlige er å

[13] dersom man ikke hadde hatt kopling, ville ikke arbeidstakerforholdet og lønnsforholdet begge to kune være viktigst på samme tidspunkt, for kun en jobb kan være viktigst på et bestemt tidspunkt.

[14] Yrkesstatus skal brukes i FoB.

lete blant dem som faktisk synes å kunne være lønnstakere. Det viser seg at lønn er godt egnet til å skille de faktisk sysselsatte fra de ikke-sysselsatte i "ukoplet lønnstakerforhold"-mengdene noe som vil bli benyttet i trinn 3 nedenfor. For å skille mellom sysselsatte og ikke-sysselsatte i "ukoplet lønnstakerforhold og selvstendig"-mengden, lar man de med høyest lønn heller få mulighet til å bli bli lønnstakere på neste trinn, mens altså resten av "selvstendig"-mengdene forblir selvstendige. Man setter grensen mellom høy og lav lønn nøyaktig slik at man får det ønskede antall selvstendige (et antall som styres av det anslåtte nivået selvstendige basert på AKU).

I tillegg til at en omdøping av selvstendige i praksis ikke kan skje på noen mer fornuftig måte, er det også slik at det faktisk i "selvstendig"-mengden er de med høy lønn som i sterkest grad intuitivt synes å kunne være lønnstakere i stedet. Denne delen av regelen blir enda mer rimelig når man for produksjonsformål antakelig må benytte informasjon om selvstendig virksomhet fra ett år tidligere[15], noe som gjør at informasjon om lønnsforhold må regnes som mer pålitelig informasjon.

På trinn 3 betrakter man "kun ukoplet lønnsforhold"-mengden og den delen av "ukoplet lønnsforhold og selvstendig"-mengden som man ikke lot bli sysselsatte på trinn 2. Førstnevnte mengde er den absolutt største og vi ser i figur 2b) at denne mengden går ganske mye på tvers av FoB-definisjonen. Imidlertid, dersom man deler alle personene på trinn 3 i to etter om de har høy lønn eller lav lønn, viser det seg at de førstnevnte faller brukbart innenfor FoB-definisjonen, dvs. at de fleste faktisk er sysselsatte. Mengden med lav lønn er brukbart utenfor FoB-definisjonen. Ettersom lønn er en kontinuerlig variabel der personene er fordelt noenlunde jevnt på ulike lønnsstørrelser, vil de som nesten hadde høy nok lønn ha omtrent den samme andel faktisk sysselsatte som de som hadde så vidt høy nok lønn. I begge tilfeller vil andelen sysselsatte ligge et sted i nærheten av 50%, og dersom vi ser isolert på disse to gruppene, kan klassifikasjonen virke litt vilkårlig.

Det viser seg at det ikke finnes noen klare kandidater på trinn 4, så vi stopper ved trinn 3, men vi setter skillet mellom høy og lav lønn på trinn 3 slik at den justerte registerdefinisjonen gir like mange sysselsatte som FoB-definisjonen (målt ved AKU).

En oppsummering av den justerte registerdefinisjonen, som utgjør deterministisk metode, finnes i ramme 2.

Ramme 2 Deterministisk metode for klassifisering av sysselsetting (justert registerdefinisjon)

Alle forhold henviser til viktigste lønnstakerforhold eller viktigste selvstendigforhold i måleuke (referanseuke). Trinn 2 er litt forenklet, se pkt. 6.1.3.

Trinn1 : personer med koplet lønnstakerforhold, klassifiseres som sysselsatte og videre som lønnstakere.

Trinn2 (gjelder kun de som ikke er klassifisert som sysselsatte på trinn 1): alle med selvstendigforhold, og ikke samtidig høy samlet årslønn, klassifiseres som sysselsatte og videre som selvstendige. Grensen for høy lønn defineres slik at man får samme antall selvstendige som AKU estimerer.

Trinn 3 (gjelder kun de som ikke er klassifisert på trinn 1 eller trinn 2): de med ukoplet lønnsforhold (og som enten har eller ikke har selvstendigforhold) som i tillegg har høy samlet årslønn klassifiseres som sysselsatte og videre som lønnstakere. Grensen for høy lønn defineres slik at man etter trinn 3 får samme antall sysseslatte som estimert utfra AKU.

6.1.3 Utfyllende kommentarer og presiseringer

Man er ikke garantert at alle med selvstendigforhold blir sysselsatte. Dette kan skje på to måter: den ene er at lønnsgrensen på trinn 2 blir lavere enn på trinn 3. En del av de som faller ut på trinn 2 fordi de heller bør være lønnstakere, vil da ikke komme med på trinn 3. Personer med selvstendigforhold

[15] pga. produksjonstiden for likningsregisteret. Tallene for år *t* er først ferdig i mars/april år *t+1*.

som også har lønnstakerforhold av et middels omfang aktivt i referanseuka, vil dermed risikere å ikke bli sysselsatt, mens en som bare har et selvstendigforhold blir sysselsatt. Dette er urimelig, men i praksis er lønnsgrensen på trinn 2 høyere enn på trinn 3 med klar margin. De to lønnsgrensene bør sammenliknes ved implementasjon av den deterministiske metoden, men helst burde kanskje metoden modifiseres slik at alle med selvstendigforhold regnes som sysselsatte.

Den andre måten som kan forårsake at ikke alle selvstendige blir sysselsatte, er for personer der lønn er erhvervet utelukkende på lønnstakerforhold som ikke er aktive på tidspunktet vi måler på. En slik person bør ikke bli klassifisert som lønnstaker, men han blir altså heller ikke klassifisert som selvstendig. Sammenliknet med en person som bare har et selvstendigforhold, står vel personen med like sterk informasjon. Dette betyr at blant alle i mengden "kun selvstendigforhold", er det noen få som ikke blir klassifisert som sysselsatte (selvstendige) imotsetning til de fleste i denne mengden, og det er fordi de hadde et lønnstakerforhold et annet tidspunkt på året.

I deterministisk metode benyttes samlet årslønn over alle personens arbeidstakerjobber dette året. Det viste seg i bakgrunnsmaterialet til Bråthen & Fosen (1998) at samlet årslønn ga mindre feilklassifikasjon (iflg. AKU) enn om man benytter årslønn for kun det viktigste forholdet[16]. Årsaken til at ikke timelønn eller daglønn beregnes, er at dateringen i LTO-lønn ikke er pålitelig. En mulig løsning som kunne vært prøvet ut kunne imidlertid vært at man eliminerte lønn av den typen som man med stor sikkerhet vet ikke er relevant. Slik lønn er lønn for koplet lønnstakerforhold som ikke er aktivt på referanstidspunktet. Et eksempel: dersom referansetidspunktet er første uke i november, og personen har en heltidsjobb i A/A som ble avsluttet 1. august, er det rimelig å se bort fra denne lønnen når man vurderer om personen er sysselsatt 1. uke i november.

6.1.4 Om motivasjon

Den deterministiske klassifikasjonsrutinen er laget utfra et ønske om at det skal være mikrokonsistens, men det er også en rekke andre motiver. Man har prioritert konsistens i den forstand at med to individer med ulik registrert arbeidsmarkedsinformasjon, så skal en ev. forskjell mellom dem i klassifisert status aldri favorisere personen med svakest arbeidsmarkedsinformasjon. I tillegg er den laget enkel (få og ukontroversielle parametre og grupperingsvariable) slik at den ikke skal være sårbar overfor regelendringer i skatteregler, endringer i standarder m.m. Metoden er enkel i den forstand at det er lite bruk av statistisk metode for å implementere metoden, men metoden er ikke enkel når det gjelder registerbearbeiding. Det siste er et mindre problem ettersom det i S260 (som skal bruke metoden) er høy kompetanse på registerbearbeiding.

Det har vært et bevisst valg å ikke bruke AKUs struktur for aktivt. Ved danningen av deterministisk regel er AKU brukt en del gjennom sammenlikning på individnivå mellom AKU-svar og registerinformasjon, for å finne klassifikasjonsgruppene som hhv skal klassifiseres som sysselsatte og ikke-sysselsatte. Når først regelen er funnet, er det imidlertid kun de to lønnsgrensene som vil bli justert fra et år til et annet, og dette betyr at det kun er AKU-nivået som kommer til å bli benyttet aktivt i regelen. I kap. 7 vil vi gå nærmere inn på fordeler og ulemper ved metoden.

6.2 Stokastisk klassifikasjon

Med stokastisk klassifikasjon er utgangspunktet et annet enn ved deterministisk klassifikasjon. Etter at klassifikasjonen er gjennomført, ønsker vi at AKU-utvalget når det gjelder oppgitt sysselsettingsstatus skal være mest mulig likt populasjonen når det gjelder klassifisert sysselsettingsstatus. Dette kan vi bl.a. få til ved hjelp av stokastisk klassifikasjon i etterstrata. Vi vil her se nærmere på en slik metode, og refererer til den som stokastisk metode.

Vi inndeler populasjonen og utvalget i grupper, slik at vi har grupperingsmodell, men nå kaller vi gruppene for etterstrata, og etterstrataene har en annen funksjon enn klassifikasjonsgruppene. Analogt

[16] Imidlertid ville bruk av lønn fra viktigste forhold unngått at personer med selsvtendigforhold ikke blir klassifisert som sysselsatte selvstendige pga. lønnsforhold et annet tidspunkt på året.

med deterministisk metode er vi interessert i andel faktisk sysselsatte i hvert etterstratum. For deterministisk metode nøyde vi oss med å sikre at andelen var ganske stor, og gikk da til det skrittet å klassifisere alle i gruppa som sysselsatte (eller ingen dersom andelen veldig liten). Når prosedyren skulle gjentas et annet år var det ikke nødvendig å finne klassifikasjonsgruppeandelen sysselsatte. Nå er vi interessert i at hvert etterstratum i populasjonen skal ha lik andel sysselsatte som det vi måler i AKU-delen av etterstratumet.

6.2.1 Datering

Datering av registerinformasjon dukker opp som et viktig tema med så aktiv bruk av AKU som stokastisk metode legger opp til. For deterministisk metode er kun nivåtallet på måletidspunktet nødvendig, men dette holder ikke for stokastisk metode. Vi bruker AKU-svar som mål på sysselsettingsstatus, og dette svaret gjelder for en periode som er personens intervju-uke AKU-uka, (noe som gir en individuell uke for hver person). Vi skal imidlertid klassifisere for 1. uke i november,[17] dvs. at AKU-svar og klassifisert verdi ikke er direkte sammenliknbart på individnivå.

Ettersom AKU-svaret gjelder AKU-uka, lager vi etterstrata etter registerinformasjonen for AKU-uka. Deretter finner vi andel AKU-sysselsatte. Når vi går fra AKU-uka til 1. uke i november for å klassifisere resten av populasjonen, vil etterstrataene få en annen sammensetning av personer, men vi antar at andelene sysselsatte i hvert etterstratum forblir den samme, og vi bruker dermed den målte andelen i den versjonen av etterstratumet som er målt i AKU-uka. Dette forutsetter at sammenhengen mellom faktisk sysselsetting (målt ved AKU) og register-informasjon for 1. uke i november ikke skiller seg ut fra gjennomsnittet over året av denne sammenhengen. Dette er en svakere antakelse enn at sammenhengen mellom faktisk sysselsetting og registersituasjon er konstant over hele året; det holder at 1. uke i november er en vanlig uke, og at det ikke finnes så mange spesielle uker at gjennomsnittet over hele året avviker fra det normale.

Dersom siste uke i desember hadde vært valgt i stedet, kunne antakelsen derimot bli feil: det er kanskje et annet omfang av småjobber i slutten av desember, og da ville sammensetningen av de etterstrataene som er en del av mengden "ukoplet lønnsforhold" bli annerledes enn ellers. Dette ville kunne føre til at disse etterstrataene egentlig hadde en annen sysselsettingsandel enn tilsvarende etterstrata et annet tidspunkt på året.

Gitt at vi velger en uke som 1. uke i november, er spørsmålet om det finnes så mange spesielle uker at gjennomsnittet av de andre ukene avviker fra november. Jul og påske er så få uker at de ikke kan slå ut, men hva med sommerjobb-effekten? Dersom vi antar at sommerjobbene stort sett er ukoplede lønnsforhold, kan etterstrata som er en del av mengden "ukoplet lønnsforhold" endre sin sammensetning så mye at fordelingen av syssselsettingsstatus i etterstratumet i sommermånedene blir helt annerledes. I så fall burde personer med AKU-uker om sommeren vært fjernet fra analysen før andel sysselsatte ble beregnet.

Dersom alle etterstratifiseringsvariable er konstante over hele året, vil etterstrataene være de samme uansett datering. Når vi samtidig vet at vi benytter samme andel sysselsatte i et etterstratum uansett datering, vil dette forutsette at vi må anta likt sysselsettingsnivå gjennom hele året slik at det å kun bruke slike etterstratifiseringsvariable er uheldig. Slike etterstratifiseringsvariable som ikke endrer seg i løpet av året er alder[18], kjønn, utdanning og delvis selvstendigforhold og ukoplet lønnsforhold (de sistnevnte refererer altså til hhv. en persons viktigste selvstendigforhold og viktigste lønnstakerforhold, som ofte er registrert fra 1.1-31.12).

6.2.2 Klassifikasjon ved trekking

Når vi skal klassifisere for 1. uke i november, er utgangspunktet etterstrataene for denne uka, samt sysselsettingsandelene for tilsvarende etterstrata for AKU-uka. Klassifkasjonen skal skje ved trekking, og det er her stokastisk klassifikasjon skiller seg mest fra deterministisk klassifikasjon.

[17] AKU-uka er heller ikke definert for andre enn AKU-utvalget og dermed helt uegnet som uke for klassifikasjon.

[18] definert som alder ved slutten av året.

Etter at vi har klassifisert i hele populasjonen, ønsker vi at AKU-utvalget med AKU-sysselsettingsstatus skal være representativt for populasjonen med klassifisert verdi. Følgelig vil vi at andelene sysselsatte (klassifisert) i etterstrataene i populasjonen skal være lik de tilsvarende andelene AKU-sysselsatte i utvalget. Å sikre like andeler betyr imidlertid at vi ikke har noen informasjon om hva slags folk i etterstratumet som faktisk er sysselsatte, vi ser kun på andelen i hele etterstratumet. Når vi skal klassifisere må vi imidlertid velge hvem i etterstratumet i populasjonen som skal bli sysselsatte, og vi ønsker å unngå systematisk overrepresentasjon av enkelte undergrupper (som vil gi skjevhet som vi ikke har kontroll over). Derfor trekker vi tilfeldig hvilke individer som skal bli sysselsatte, men slik at andelen vi klassifiserer som sysselsatte i etterstratum h blir lik andelen AKU-sysselsatte $\hat{p}_h$ for etterstratum h i utvalget. Dette er "trekke fra hatt"-type trekking eller enkel tilfeldig trekking, og betyr bl.a. at alle personer i et etterstratum har samme sannsynlighet for å bli klassifisert som sysselsatt (derav unngår vi systematisk overrepresentasjon av undergrupper). Vi vil komme i kapittel 7 komme inn på hva slik trekking innebærer matematisk.

Vi har eliminert muligheten for systematisk overrepresentasjon ved metoden ovenfor, men vi har muligheten for tilfeldig overrepresentasjon. Vi vil se nærmere på dette forholdet senere, og også se på at når vi kommer til å skulle estimere på små områder, så er det å velge ut personer fra et lite område gitt klassifikasjonen, en ikke-tilfeldig prosess som kan medføre at vi overrepresenterer undergrupper som sysselsatte likevel: dersom vi ser på et lite område der andelen faktisk sysselsatte er den samme som i befolkningen, kan det være lokale særegenheter som gjør at veldig mange personer er i et etterstratum der estimert sysselsettingsandel er høy, men der altså de lokale personene fra dette etterstratumet i mindre grad er sysselsatte. Dermed vil vi systematisk overestimere sysselsatte her pga. de lokale særegenhetene.

7 Deterministisk eller stokastisk

Vi har ovenfor sett at deterministisk og stokastisk metode både har likheter og ulikheter i sin oppbygging. Vi vil se nærmere på dette nedenfor, både de åpenbare forskjellene og hvordan metodene fungerer i forhold til skjevhet og varians. I forbindelse med dette ser vi også på hvordan man skal lage etterstrataene i stokastisk metode, og i hvilken utstrekning liknende betraktninger gjelder for deterministisk metode.

7.1 Parvis registerkonsistens

Vi har tidligere sett på forskjellen som ligger i det at den ene metoden er deterministisk og den andre stokastisk. Den deterministiske er laget for bl.a. å være "parvis registerkonsistent", som vil si at to personers ulikhet i registertilknytning skal reflekteres i ulikheten i klassifisert sysselsettingsstatus.

Eksempel 1. "Parvis registerkonsistens"

Vi antar at person A og B har samme registerinformasjon bortsett fra at person B har høyere lønn. Høyere lønn er sterkere indikasjon på sysselsetting enn lav lønn, så dersom det da viser seg at A er klassifisert som sysselsatt mens B ikke er det, så har vi ikke parvis registerkonsistens.

Stokastisk metode vil ikke gi parvis konsistens fordi man trekker sysselsettingsstatus og dermed ikke kan sikre at sterkere registerindikasjon på sysselsetting faktisk slår ut i samme retning for klassifisert sysselsettingsstatus.

7.2 Tabuleringsbegrensninger

Deterministisk metode er laget slik at man gir alle personer i samme klassifikasjonsgruppe lik sysselsettingsstatus, det er slik man kan få parvis registerkonsistens (se ovenfor). Som en følge av dette er det farlig å benytte klassifikasjonsgrupperingsvariablene som tabuleringsvariable. Ettersom de

med høy lønn overrepresenteres[19], vil en tabell av sysselsetting etter lønn gi for store sysselsettingstall i gruppen høy lønn og altfor lav andel i gruppen lav lønn. Disse avvikene skal likevel ikke overdramatiseres: de fleste med høy lønn er i klassifikasjonsgruppen "koplet lønnstakerforhold" og de er som oftest faktisk sysselsatte. For enkelte grupper kan det imidlertid slå ut noe, f.eks. gruppen med lav lønn, som i ganske stor grad utgjøres av klassifikasjonsgruppen "ukoplet lønnsforhold med lav lønn" (som alle klassifisert ikke-sysselsatte mens en del faktisk er sysselsatte) og i mindre grad av "koplet lønnstakerforhold" (der alle er klassifisert som og de fleste faktisk er sysselsatte).

7.3 Klassifikasjonsgruppenes rolle i forhold til etterstataenes rolle

Mens man ved stokastisk metode for hver person i et etterstratum trekker verdien sysselsatt med sannsynlighet lik AKU-sysselsettingsandelen i etterstratumet, innebærer deterministisk metode at man først avrunder andelen i en klassifikasjonsgruppe til nærmeste heltall, og så lar enten alle eller ingen bli klassifisert som sysselsatte. Det å la alle bli sysselsatte, kan sees på som at vi trekker sysselsettingsstatus med sannsynlighet 1 (dvs. 100%), mens det å la ingen bli sysselsatte betyr å trekke med sannsynlighet 0. Dette betyr at man i deterministisk metode trekker statusen sysselsatt med sannsynlighet lik den avrundede andelen sysselsatte i en klassifikasjonsgruppe mens man i stokastisk metode trekker statusen sysselsatt med sannsynlighet lik den eksakte andelen sysselsatte i et etterstratum. De to metodene kan således formuleres veldig likt.

Ovenfor synes deterministisk metode å være åpenbart urimelig, men grunnen til at det ikke er tilfellet er at vi stiller bestemte krav til danningen av klassifikasjonsgrupper, mens etterstrataene kan dannes ganske vilkårlig uten at metoden faller sammen.

Slik som den deterministiske regel er presentert, ser vi at vi har valgt klassifikasjonsgrupper der AKU-sysselsettingsandelen er nærmest mulig 0 eller 1, og i den grad de avviker en del fra det, er det rimelige intuitive grunner til at alle skal klassifiseres som f.eks. sysselsatte (hensynet til parvis registerkonsistens som definert i 7.1). Variablene som danner klassfikasjonsgruppene, kan ikke benyttes som tabuleringsvariable (f.eks. lønn, se 7.2).

7.3.1 Etterstratifisering i deterministisk metode

Vi så ovenfor at klassifikasjonsgruppene er dannet etter strengere krav enn etterstrataene. Nedenfor vil vi se at det ikke er naturlig å se på klassifikasjonsgruppene bare som et alternativ til etterstrataene.

Deterministisk metode kunne ha vært utvidet til å inneholde etterstratifisering. Dette ville gå ut på at man først etterstratifiserte, og deretter gjennomførte tre-trinnsregelen separat for hvert etterstratum. I praksis ville trinn 1 bli likt for alle etterstrata, mens lønnsgrensene på trinn 2 og trinn 3 ville blitt annerledes ved at de da sikrer riktig nivå innenfor hvert etterstratum.[20]

Vi kan nå tenke oss en felles etterstratifisering etter f.eks. kjønn og alder. Da kan det tilsynelatende virke som om de tre trinnene i deterministisk metode kan betraktes som et alternativ til trekking i stokastisk metode, men hva dersom vi tenker oss at vi utvider etterstratifiseringen til også å inkludere klassifikasjonsgrupperingen. Med denne etterstratifiseringen vil stokastisk metode fremdeles innebære trekking innenfor hvert etterstratum, men deterministisk metode vil degenerere ved at den deterministiske metoden fanges opp av etterstratifiseringen. Metoden reduseres til å enten la alle eller ingen i et etterstratum bli klassifisert som sysselsatte, og ettersom metoden dermed ikke forsøker å sikre et representativt nivå av personer klassifiserte som sysselsatte i hvert etterstratum, er det naturlig å si at deterministisk metode ikke er definert med slik etterstratifisering.

Selv om klassifikasjonsgrupperingen i deterministisk metode likner litt på en helt bestemt etterstratifisering, er det antakelig mest naturlig å betrakte grupperingen som et alternativ til den rollen trekkeprosessen har i stokastisk metode.

[19] ved at en bestemt undergruppe av dem i sin helhet klassifiseres som sysselsatte.

[20] Dette ville krevd visse avveininger med tanke på hvordan man skulle finne totalene i hvert etterstratum.

Dersom man skal sammenlikne stokastisk og deterministisk metode empirisk, ville det ideelle kanskje være å etterstratifisere så likt som mulig i deterministisk metode og stokastisk metode. Samtidig er det en versjon av deterministisk metode uten etterstratifisering som er implementert og som er vårt analysegrunnlag. I dette notatet vil vi imidlertid uansett begrense oss til mer summarisk sammenlikning empirisk. Fokus ligger på beskrivelse av à priori egenskaper ved metodene, samt konsekvenser for anvendelse av resultatene.

7.4 Behov for mikrodata

Et av produktene i FoB er mikrodata (individdata) til forskere. Stokastisk metode er konstruert slik at man ikke kan feste tiltro til de enkelte observasjonene i mikrodataene, og dette skyldes naturligvis trekkingen. Samtidig er det først på aggregert nivå at tallene vil bli brukt av forskerne, og ved store talls lov[21] vil tilfeldighetene i trekkingen jevne seg ut. Deterministisk metode er heller ikke pålitelig for hver enkelt person, men for den metoden er det registrerte opplysninger og ikke terningkast som er metoden. Dersom man virkelig trenger mikrodata, er deterministisk metode mye bedre.

I FoB er man interessert i bruttostrøm, dvs. hvor mange personer som endrer verdi over tid. Dette er også interessant i årsstatistikken til fagseksjonen som eier tallene, og for å kunne beregne dette trengs mikrodata. Stokastisk metode kunne vært modifisert slik at dataene reflekterte den virkelige bruttostrømmen, og således eliminerte behovet for mikrodata som kilde til bruttostrømtall. Dette ville imidlertid kreve avanserte løsninger som vi ikke vil forsøke å gå inn på her.

7.5 Hvor god er AKU?

Vi har allerede sett at begge metodene benytter AKU i litt varierende grad, og det ligger litt i kortene at det er AKU som er fasiten vi forsøker å etterlikne. Dette er imidlertid ikke et helt opplagt valg.

Den ultimate sannheten ville vi oppnådd dersom vi fikk til et dybdeintervju av hver person der vi kunne lese ut den virkelige sannheten av svarene og avsløre bløff og eliminere feil svar pga. uvitenhet. Tre former for datafangst er relevante i FoB-sammenheng (som et forsøk på å måle sannheten):
1) Klassisk folke- og boligtelling: hver person svarer på et skjema mottatt per post.
2) Utvalg. Her har vi AKU som er en telefonundersøkelse.
3) En ren registerfangst: informasjon hentes fra registre alene.

Deterministisk metode likner mye på en ren registerfangst fordi metoden i stor grad går ut på at registeret bearbeides best mulig og at man lar det være med det. Gitt registerregelen er AKU kun benyttet for å finne totalnivået, men i utformingen av regelen ble AKU brukt atskillig mer. Likevel er det antakelig riktig å betrakte metoden som en registerfangst-liknende rutine. I stokastisk metode er derimot AKU mye mer sentral. Deterministisk metode er altså å lene seg på registerinformasjon, mens stokastisk metode er å lene seg på AKU. Begge metoder gjør imidlertid bruk av den andre informasjonskilden som støtte, deterministisk ved at AKU styrer sysselsettingsnivået, og stokastisk ved at registerinformasjon brukes for å etterstratifisere.

AKU likner i større grad enn registerfangst på klassisk FoB. Den eneste forskjellen er innsamlingsmetoden og at det er utvalg i stedet for fulltelling. Spørsmålet er om man skal forsøke å nærme seg sannheten via klassisk FoB/AKU eller om man skal gå via registerfangst. Det ene spørsmålet er hva som faktisk er nærmest sannheten, men et annet spørsmål er hvilken av veiene som synes mest passende dersom man likevel kun i begrenset grad kan nærme seg sannheten.

Feilen med AKU/klassisk FoB er bl.a. frafall, noe som skyldes at individer enten glemmer å svare eller motsetter seg å svare. FoB har tradisjonelt hatt høy svarprosent, etter 1. purring bør man være oppe i 95%, men selv om frafrallet ikke er stort, er det absolutt slik at det kan gå utover kvaliteten av

[21] Store talls lov sier grovt sett at når man gjør tilstrekkelig antall uavhengige observasjoner av en variabel så vil gjennomsnittsverdien av variabelen stabilisere seg.

produserte tabeller. I tillegg til feilene som oppstår pga. frafallet dersom populasjonens profil avviker fra den som svardelen har, kommer målefeil: noen vil av uvitenhet besvare spørsmål feil, andre vil bevisst pynte på svarene sine. Det kan for noen føles sosialt stigmatiserende (selv med kun en telefonstemme som intervjuer) å innrømme at man ikke har arbeid. Til og med når det er et skjema som leveres, vil enkelte ikke føle seg nok anonymisert til å fortelle den hele sannheten. Noen få pynter kanskje til og med på svarene fordi de faktiske forholdene er noe man ikke helt vil innrømme overfor seg selv.

Som vi har sett tidligere, er det åpenbare problemer med å lage en deterministisk klassifisert sysselsettingsstatus som et alternativ til den opprinnelige registerverdien, men den deterministiske metoden oppnår likevel samme klassifikasjon som AKU-svar for i overkant av 89% av utvalget. Dette betyr at deterministisk metode likner på AKU, men sier ikke hva som er nærmest sannheten. Det er grunn til å tro at AKU er en del nærmere sannheten, men det er en viss feil knyttet til begge metodene, og det er et strategisk valg om man vil velge å publisere tall basert på hva nøye bearbeidet registerinformasjon sier (noe deterministisk metode til syvende og sist innebærer), eller om man vil publisere tall basert på hva et utvalg har svart over telefon. I sistnevnte tilfelle må man godta enkelte uheldige sider som skyldes det nødvendige tilfeldige elementet i klassifikasjonen.

7.6 Nøyaktighet utfra perspektivet i statistisk metode

Vi vil nå betrakte nøyaktighet under ett og ikke skille mellom komponentene definisjonsmessig validitet og reliabilitet definert i 2.1. I tråd med vanlig språkbruk, vil vil benytte termen "feil" der hvor det er mangel på nøyaktighet. Hensikten med dette avsnittet er å forsøke å beskrive feil i lys av en annen inndeling av nøyaktighet og som utgjør to sentrale begreper i statistisk metode, nemlig skjevhet (systematisk feil) og varians (tilfeldig variasjon). I tillegg vil vi se på mikrokonsistens.

For å få et bedre grep om disse to størrelsene, kan det være nyttig å skille feilen etter når den forekommer. Vi har to faser: fase 1 er prosessen fram til klassifikasjonen starter (dvs. i produksjonen av datagrunnlaget, både av registerinformasjonen og utvalget), mens fase 2 er klassifikasjons-prosessen.

7.6.1 Estimatets totalfeil

Målet for totalfeil til sysselsettingsestimatet er avstanden i absoluttverdi (tallverdi) mellom estimatet basert på klassifisert verdi og sannheten, $|\hat{\theta}_{klass} - \theta|$. Her lar vi θ betegne sann sysselsettingsandel, mens $\hat{\theta}$ betegner estimatet. Vi har sjelden mulighet til å måle totalfeil i det enkelte tilfellet, og i stedet ser vi da på forventet (gjennomsnittlig) kvadrat[22]-totalfeil. Dette kalles gjerne bruttovarians (engelsk: mean squared error):

$$brutto\,\text{var}\,ians(\hat{\theta}) = MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

Videre kan bruttovarians deles videre opp i skjevhet og varians:

$$brutto\,\text{var}\,ians(\hat{\theta}) = Var(\hat{\theta}) + \left[skjevhet(\hat{\theta})\right]^2$$

Skjevhet er en systematisk feil.

Vi har feil på to trinn: feil i prosessen der utvalget er generert og feil i klassifikasjonen. Førstnevnte feil består i tillegg av to komponenter, nemlig feil ved trekkingen og feil pga. frafall.

7.6.2 Skjevhet

Skjevhet pga. trekkingen av utvalget antas å ikke forekomme (utvalgsplanen skal være laget slik at dette unngås).

[22] av regnetekniske årsaker benyttes kvadratet av feilen i stedet for absoluttverdien.

Dersom frafallet ikke er et tilfeldig underutvalg av bruttoutvalget, er frafallet skjevt etter en eller flere variable. Dersom de variable som frafallet er skjevt i forhold til, er sysselsettingsstatus selv eller er korrelert med sysselsettingsstatus, vil frafallet være skjevt i forhold til sysselsettingsstatus -- man sier også at frafallet er ikke-ignorerbart for sysselsettingsstatus. Et eksempel er dersom det spesielt er de yngste som har frafall, for de unge er i mindre grad sysselsatt enn andre. Ettersom man ikke måler frafallet, vil man måtte estimere for populasjonen som om frafallet var likt nettoutvalget. Med et skjevt frafall, blir anslaget på sysselsatte skjevt dersom man ikke f.eks. etterstratifiserer.

Etterstratifisering innebærer at vi vekter hvert etterstratum etter etterstratumets størrelse i populasjonen. Dersom vi etterstratifiserer etter aldersgruppe, vil vi vekte aldersgruppene slik at aldersfordelingen er slik den ville vært uten frafall. Som vi har sett tidligere, kan både stokastisk og deterministisk metode inneholde en etterstratifisering.

Når vi skal se på skjevhet som følge av klassifikasjonsprosessen, må vi se det i relasjon til på hvilket nivå vi skal anslå sysselsetting. Vi kan anslå sysselsatte totalt eller vi bryte ned sysselsatte etter tabuleringsvariable.

For å anslå sysselsatte totalt, vil det med stokastisk metode ikke være noen skjevhet forårsaket av klassifikasjonsprosessen, dvs. at gitt utvalget har vi ingen skjevhet. Stokastisk metode gjenspeiler kun AKU-utvalget og der AKU ikke gir informasjon (dvs. om akkurat hvem som skal klassifiseres til hva innenfor et etterstratum), så trekker man tilfeldig (se 6.2.2). Dersom vi i tillegg antar at frafallet er ignorerbart (dvs. kan sees bort fra) , er stokastisk metode (ubetinget utvalget) forventningsrett (har ingen skjevhet) for å anslå sysselsatte totalt dersom trekkingen av utvalget ikke innfører skjevhet. Dvs. at stokastisk metode gjennomsnittlig gir riktig svar (i form av samme andel sysselsatte i hvert etterstratum som det er i populasjonen dersom man hadde stilt AKU-spørsmålene til alle) dersom man går gjennom utvalgstrekkeprosessen mange ganger.

Deterministisk metode innfører skjevhet ettersom man etter en arbeidsmarkedsgruppering lar alle i en klassifikasjonsgruppe enten bli sysselsatte eller ikke sysselsatte. Blant personene i gruppa "koplet lønnstakerforhold" er det ca. 4% som ikke skulle vært sysselsatte iflg. AKU, i klassifikasjonsgruppen "selvstendige" er det i størrelsesorden 15% som ikke skulle vært sysselsatte. Deterministisk metode lar imidlertid alle i disse to klassifikasjonsgruppene bli sysselsatte. Den største overrepresentasjonen har man likevel i overgangen mellom klassifikasjonsgruppene "ukoplet lønnsforhold med lav lønn" og "ukoplet lønnsforhold med høy lønn". Lønnsgrensen viser seg å bli i størrelsesorden 25 000,- , og dersom man for hver av klassifikasjonsgruppene ser på de som hhv. er i lønnsområdet 20 000-25 000 og 25 000-30 000, viser det seg at nesten halvparten av disse blir klassifisert feil iflg. AKU.

Som vi var inne på i 7.2 er konsekvensene av skjevhetsinnføringen at man vil få feilaktige svar dersom man tabulerer sysselsetting etter lønn eller en variabel som er korrelert med lønn. Det blir for mange sysselsatte i tabuleringsgruppen høy lønn og for få med lav lønn. Dersom alder er korrelert med lønn slik at unge har mindre lønn enn eldre, betyr det at man i gruppen ”unge personer” klassifiserer for få som sysselsatte, mens man gjør motsatt i gruppen ”godt voksne personer”. For totalt antall sysselsatte i landet veier de to feilene opp hverandre slik at antallet blir riktig. Antall sysselsatte etter fylke vil også bli riktig dersom hvert fylke har lik fordeling av lønn. Enkelte mulige skjevheter ble sett på i Bråthen & Fosen (1998), men man fant ingen resultater som skulle tyde på at lønn bidrar til stor skjevhet. Blant annet ble det funnet at det ikke gir vesentlig bedre sysselsettingstall blant unge dersom man korrigerer for skjevhet som skyldes ulik lønnsfordeling i ulike aldersgrupper. På lavere geografisk nivå er det langt vanskeligere å måle kvalitet, og dermed vanskelig å si noe om skjevhet.

Dersom vi skal bryte ned tall etter tabuleringsvariable, f.eks. fylke, er stokastisk metode kun forventningsrett dersom vi i tillegg til ignorerbart frafall har modellantakelsen at hvert fylke oppfører seg på samme måte som hele landet, dvs. at andelen sysselsatte i hvert etterstratum også gjelder når man først bryter ned på fylke og så ser på andel sysselsatt i hvert etterstratum i hvert fylke. Som

eksempel kan vi tenke oss at vi etterstratifiserer etter kjønn, og at det i populasjonen er 70% av mennene som er sysselsatte og 62% av kvinnene. Så skal vi tabulere for Finnmark fylke, der vi tenker oss at det i virkeligheten er 65% sysselsatte både blant menn og kvinner. Vi har imidlertid i klassifikasjonen trukket statuser slik at andelene er hhv 70% og 62%, og dermed vil disse tallene være omtrent de vi får når vi tabulerer for Finnmark, og ikke 65% for begge kjønn. Denne kilden til skjevhet er beskrevet mer matematisk i 7.6.3.1.

Deterministisk metode vil på tilsvarende måte kunne bli skjevere (eller mindre skjevt) når man bryter ned, med mindre andelen sysselsatte i de ulike klassifikasjonsgruppene skulle vise seg å være de samme innenfor hvert fylke som i hele landet. For ett fylke kunne det f.eks. i teorien vise seg at lønn egner seg dårlig for å skille sysselsatte og ikke-sysselsatte blant de med LTO-lønn som viktigste forhold.

Det er klart fra det som står ovenfor at stokastisk metode sørger for å fordele de sysselsatte riktig etter etterstratifiseringsvariablene, f.eks. etter alder. Spørsmålet nå er om etterstratifisering for deterministisk metode i tillegg til å rette opp for en etterstratifiseringsvariabel som alder, også kan rette opp skjevhet mhp. lønn mer direkte?

Dersom alder og lønn er korrelerte kan dette bety at når man i deterministisk metode behandler aldersgrupper hver for seg, så er lønn bedre til å skille mellom sysselsatte og ikke-sysselsatte. Dersom man slår sammen aldersgruppene, vil lønn ikke fungere så bra. Eksempelet nedenfor illustrerer dette poenget. Det mest ekstreme er dersom vi innenfor hver aldergruppe har at lønn skiller perfekt mellom sysselsatte og ikke-sysselsatte, men der lønnsgrensene er ulike i hver aldersgruppe, for når man da ser på alle aldre sammen, vil ingen lønnsgrense kunne skille mellom sysselsatte og ikke-sysselsatte i stor grad.

Eksempel 2

Deterministisk metode overrepresenterer de med høy lønn som sysselsatte. Vi ser her på trinn 3, dvs. de personene som er vanskeligst å klassifisere korrekt. Vi etterstratifiserer etter aldersgruppe, og tenker oss at det viser seg at når man ser på aldersgruppene hver for seg, så overrepresenterer vi i mye mindre grad etter lønn. Dette er høyst mulig, og dermed har man minsket skjevheten etter lønn for deterministisk metode. Vi vil nedenfor se hvordan dette konkret kan gå til.

Vi antar for enkelhetsskyld at vi deler inn utvalget i tre like store etterstrata etter alder. Vi antar videre at lønnsgrensene viser seg å bli 10000,- , 25000,- og 40000,- for hhv. de yngste, de mellomste og de eldste. Fra før vet vi at lønnsgrensen uten etterstratifisering er ca. 25000,-. Tabell 1 viser et eksempel på andel sysselsatte i hver gruppe for hvert etterstratum, og siste rad er den summen dette gir når man ser bort fra etterstratum. For de uskraverte feltene, er andel feilklassifiserte i prosent lik tallet i cellen, mens for de skraverte er feilklassifikasjonsandelen i prosent lik 100 minus tallet i cellen. Summen nederst viser at man feilklassifiserer en del i lønnsgruppene 10000-40000,- uten etterstratifisering. Derimot ser vi at det er lite feilklassifisering med etterstratifisering: for de yngste feilklassifiserer man aldri mer enn 30%, og slik er det også for de med middels alder og for de eldste.

Tabell 1 Andel AKU-sysselsatte i en tenkt situasjon i ulike lønnsgrupper, etter aldersgruppe blant personer der viktigste forhold er ukoplet lønnsforhold. Det antas at innenfor en lønnsgruppe eralle aldersgruppene like store for oversiktens skyld. Mørke celler markerer grupper der alle personene er klassifisert som sysselsatte iht. deterministisk metode. Lyse celler markerer at ingen er klassifisert som sysselsatte.

Andel AKU-sysselsatte	Samlet lønn per år			
	< 10 000,-	10 000,- - 25 000,-	25 000,- -- 40 000,-	> 40 000,-
yngste aldersgruppe	20 %	70 %	80 %	90 %
middels aldersgruppe	5 %	25 %	70 %	80 %
eldste aldersgruppe	5 %	20 %	30 %	70 %
Totalt	10 %	38 %	60 %	80 %

7.6.3 Varians

På samme måte som med skjevhet, er den ene komponenten av variasjon den som skyldes genereringen av utvalget og frafallet. Den andre variasjonen er den som skyldes klassifikasjonen. Sistnevnte kalles gjerne Monte Carlo-usikkerhet (MC-usikkerhet) dersom det inngår trekking av status, slik som i stokastisk metode.

Vi er mest interessert i hva som er feil ved bruk av nettopp det utvalget vi har for hånden. For skjevhet er en slik problemstilling irrelevant, for skjevheten pga. genereringen av utvalget er den samme uansett hvilket utvalg som trekkes og sier hvor systematisk feil dette utvalget er. Variasjonen er derimot per definisjon hvordan resultater varierer. Ettersom vi har ett bestemt utvalg tilgjengelig når vi skal klassifisere, er vi mest interessert i variasjonen gitt dette utvalget. Gitt utvalget så klassifiserer deterministisk metode deterministisk slik at det ikke er noen variasjon, mens vi for stokastisk metode får MC-usikkerheten. Vi ser på den nedenfor, før vi ser på variasjonen pga. utvalgsgenereringen.

7.6.3.1 Variasjon pga. trekking i klassifikasjonsprosessen

Ved trekking av statuser på nytt, ville det blitt et annet resultat. Det er denne variasjonen som er MC-usikkerheten. Dersom ikke annet er eksplisitt skrevet, er all varians i dette delkapittelet betinget det utvalget vi har til rådighet.

Trekkingen skjer som tidligere nevnt ved enkelt tilfeldig utvalg, dvs. at den andelen som skal klassifiseres som sysselsatte, velges slik at alle kombinasjoner av utvalg av ønsket størrelse er like sannsynlige. Dette innebærer at vi i hvert etterstratum trekker sysselsettingsstatuser utfra en hypergeometrisk fordeling. En vilkårlig person blir klassifisert som sysselsatt med sannsynlighet

$$\hat{p}_h = \textit{AKU-sysselsettingsandel i etterstratum } h = \textit{anslått nivå på andelen sysselsatte i hele etterstratumet, } p_h \ ,$$

og forutsatt at ingen etterstrata er mindre enn 50,[23] betyr dette at hver person tilnærmet klassifiseres uavhengig av alle andre personer, og at vi kan betrakte situasjonen som om vi trekker ved "myntkast" der sannsynligheten ikke er 50%, men $\hat{p}_h$ for å bli sysselsatt. Slik trekking kalles enkel tilfeldig trekking med tilbakelegging, og under forutsetning om slik trekking har vi da at

$$Y_h = \textit{Antall klassifisert som sysselsatte i etterstratum nr. } h \\ \sim \textit{binomisk}(N_h, \hat{p}_h)$$

[23] Det holder med 50 dersom $\hat{p}_h$ er mellom 5% og 95%

der N_h er antall personer i etterstratum nr. *h.*

Den største forskjellen på binomisk og hypergeometrisk fordeling i vår situasjon, er at sistnevnte innebærer at andelen klassifisert som sysselsatte i hvert etterstratum er eksakt $\hat{p}_h$, og at de to første variansformlene nedenfor, som gjelder binomisk fordeling, dermed ikke gjelder. Derimot gjelder den binomiske fordelingen for nivået i andre grupper, slik at variansformlene for slike grupper g nedenfor er mer interesante for oss.

Vi lar

$$\hat{p}_h^{klass} = \frac{Y_h}{N_h} = \textit{ andel klassifisert som sysselsatte i etterstratum } h,$$

og det følger da at

$$\text{var}(\hat{p}_h^{klass} \mid \hat{p}_h) = \frac{N_h \hat{p}_h (1-\hat{p}_h)}{N_h^2}.$$

Dersom vi har *r* etterstrata, har vi

$$\hat{p}^{klass} = \sum_{h=1}^{r} \frac{Y_h}{N} = \textit{ andel klassifisert som sysselsatte i hele populasjonen},$$

der hver Y_h er uavhengig av alle andre Y_h ettersom hver eneste person klassifiseres uavhengig av alle andre personer. Dermed får vi med $\underset{\sim}{p} = (p_1, ..., p_r)$ at

$$\text{var}(\hat{p}^{pred} \mid \underset{\sim}{\hat{p}}) = Var\left(\sum_{h=1}^{r} \frac{Y_h}{N} \mid \underset{\sim}{\hat{p}} \right) = \sum_{h=1}^{r} \frac{N_h \hat{p}_h (1-\hat{p}_h)}{N^2},$$

Dersom vi skal se på variansen for f.eks. Strand kommune, tar vi fremdeles utgangspunkt i de *r* etterstrataene. For hvert etterstratum ser vi imidlertid bort fra alle observasjoner som ikke gjelder denne kommunen. Vi lar $Y_{h,g}$ betegne antallet klassifisert som sysselsatte i område *g* i etterstratum *h*, og har da at

$$Y_{h,g} \sim \textit{binomisk}(N_{h,g}, \hat{p}_h).$$

Vi lar

$$\hat{p}_g^{klass} = \sum_{h=1}^{r} \frac{Y_{h,g}}{N_g},$$

der hver komponent er uavhengig som tidligere, og vi får variansuttrykket

$$\text{var}(\hat{p}_g^{klass} \mid \underset{\sim}{\hat{p}}) = \text{var}\left(\sum_{h=1}^{r} \frac{Y_{h,g}}{N_g} \mid \underset{\sim}{\hat{p}} \right) = \sum_{h=1}^{r} \frac{N_{h,g} \hat{p}_h (1-\hat{p}_h)}{N_g^2},$$

ettersom trekkesannsynligheten fortsatt er $\hat{p}_h$ for alle i etterstratum *h*. Derimot er det nå ikke N_h personer som klassifiseres med denne sannsynligheten, men kun $N_{h,g}$ som er antall i Strand kommune som klassifiseres med denne sannsynligheten.

Når $\hat{p}_g^{klass}$ varierer, varierer den rundt en forventet verdi:

$$E(\hat{p}_g^{klass} \mid \underset{\sim}{\hat{p}}) = E\left(\sum_{h=1}^{r} \frac{Y_{h,g}}{N} \mid \underset{\sim}{\hat{p}} \right) = \sum_{h=1}^{r} \frac{N_{h,g} \hat{p}_h}{N_g},$$

mens den sanne verdien er

$$p_g = \sum_{h=1}^{r} \frac{N_{h,g} p_{h,g}}{N_g} .$$

Ved at man benytter p_h (estimert ved $\hat{p}_h$) i stedet for $p_{h,g}$ (som ikke kan estimeres direkte fordi utvalget i område g i etterstratum h er altfor lite), er en kilde til mulig skjevhet, som vi var inne på i 7.6.2 men da formulert mer i ord og ikke i formler.

Et mål for MC-usikkerheten er den utstrekningen som et 95% prediksjonsintervall[24] får i begge retninger fra den forventede verdien ovenfor. Dette kaller vi usikkerhetsmarginen, og for etterstratum h (og område g dersom alle klasifikasjonssannsynlighetene $\hat{p}_h$ er like store),[25] har vi at

$$Usikkerhetsmargin = 1,96 \cdot SE(\hat{p}_h^{klass}) = 1,96 \cdot \sqrt{\text{var}(\hat{p}_h^{klass})} . \qquad (7.1)$$

Tabell 2 viser hvor stor usikkerhetsmarginen er i et etterstratum etter størrelsen på populasjonen og andel sysselsatte. Tabellen kan også direkte brukes på et område *g* der alle klassifikasjonssannsynligheter $\hat{p}_h$ er like, for da kan man bruke tabellen som om hele området var ett etterstratum.[26] Vi ser at dersom usikkerhetsmarginen skal være mindre enn ett prosentpoeng (de cellene som ikke er grå), må etterstratumet (i populasjonen) være større enn 1000 personer.

Vi ser at vi får den største variansen dersom vi har etterstratum der andel sysselsatte er 50% (da er det klassifikasjon ved myntkast). Jo nærmere 0 eller 100% andelen blir i et etterstratum, jo mer homogene blir etterstrataene. I et homogent etterstratum har nesten alle samme sysselsettingsstatus slik at variasjonen er liten, noe vi også ser fra tabellen. I det ekstreme tilfellet at andelen er 100%, er det ingen usikkerhet betinget utvalget i det hele tatt.

Tabell 2 Usikkerhetsmargin i etterstratum ved tilfeldig trekking med tilbakelegging av sysselsettingsstatus innenfor et etterstratum. Etter etterstratumstørrelsen i utvalget og andel sysselsatte i etterstratumet.

Populasjons-størrelsen av	Sannsynlighet for å klassifisere som sysselsatt										
etterstratum	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95
10	0,135	0,186	0,248	0,284	0,304	0,310	0,304	0,284	0,248	0,186	0,135
50	0,060	0,083	0,111	0,127	0,136	0,139	0,136	0,127	0,111	0,083	0,060
100	0,043	0,059	0,078	0,090	0,096	0,098	0,096	0,090	0,078	0,059	0,043
200	0,030	0,042	0,055	0,064	0,068	0,069	0,068	0,064	0,055	0,042	0,030
500	0,019	0,026	0,035	0,040	0,043	0,044	0,043	0,040	0,035	0,026	0,019
1000	0,014	0,019	0,025	0,028	0,030	0,031	0,030	0,028	0,025	0,019	0,014
5000	0,006	0,008	0,011	0,013	0,014	0,014	0,014	0,013	0,011	0,008	0,006
10000	0,004	0,006	0,008	0,009	0,010	0,010	0,010	0,009	0,008	0,006	0,004

[24] Et prediksjonsintervall er slik at dersom vi gjentar klassifikasjonsprosessen uendelig mange ganger, så vil intervallet i 95 av 100 tilfeller dekke den predikerte verdien $\hat{p}_h^{klass}$.

[25] Dersom ikke alle klassifikasjonssannsynlighetene er like store, er ikke antall klassifiserte i område g binomisk fordelt, og da er det mer komplisert å finne prediksjonsintervallet.

[26] For hver person trekkes status tilnærmet uavhengig av alle andre. Dersom sannsynligheten for statusen sysselsatt er p i alle etterstrataene, og der etterstratumstørrelsene er $N_{1,g}, N_{2,g}, ..., N_{k,g}$, er antall klassifiserte som sysselsatte i område *g,* gitt utvalget, binomisk med parametre p og $(N_{1,g}+N_{2,g}+...+N_{k,g})$. Førstnevnte parameter kan da leses av i tabellhodet til tabell 2, mens sistnevnte parameter leses av i forspalten.

Dersom vi antar at vi har fått til en etterstratifisering der AKU-sysselsettingsandelene i etterstrataene alle er 10 prosentpoeng fra 0 eller 100%, vil et 95% prediksjonsintervall for andelen sysselsatte i et område med 10000 innbyggere utstrekke seg seks promille i hver retning fra den observerte andelen. For et område med 1000 innbyggere vil intervallet utstrekke seg 1,9 prosent. Når man da i tillegg skal splitte opp etter alder, næring osv., kan usikkerheten bli stor (mer om små områder senere).

Som vi ser ovenfor har valg av etterstrata betydning for størrelsen til variansen. Dersom vi kan forbedre etterstratifisering slik at andelen sysselsatte øker fra 0.7 og 0.3 til 0.8 og 0.2, vil man ved å bruke tabell 2 se at usikkerhetsmarginen avtar med 11 prosent for et etterstratum med 10 000 personer, og relativt mer når etterstratumet er mindre. I det ekstreme tilfellet at andel sysselsatte er 0 eller 100%, vil det ikke bli noen MC-usikkerhet i det hele tatt.

Når det gjelder MC-usikkerheten, gjelder altså det utelukkende trekking av statuser, og det eneste viktige for denne usikkerheten er hvor homogene eterstrataene er mht. sysselsettingsstatus. Dersom alle i hvert etterstratum har samme status, vil man ikke få noen MC-usikkerhet, mens det verste er dersom det er halvparten av hver status.

7.6.3.2 Variasjon totalt

Variasjonen som skyldes genereringen av utvalget, er den variasjonen som skyldes at andelene sysselsatte i hvert etterstrata er ukjente slik at klassifikasjon ved trekking av status skjer utfra de observerte andelene i utvalget, noe som ville vært annerledes dersom det var et annet utvalg som var blitt trukket. Vi vil nedenfor tallfeste den ubetingede variansen (som tar hensyn til både utvalgsgenereringen og MC-usikkerheten).

Størrelsen på variansen til $\hat{p}_h$ er enkel å tallfeste dersom vi antar at trekkingen av utvalget er tilnærmet enkelt tilfeldig med tilbakelegging. I så fall blir antall sysselsatte i hvert etterstratum tilnærmet binomisk fordelt, og andel sysselsatte får en varians som likner den vi har i avnsitt 7.6.3.1:

$$\text{var}(\hat{p}_h) = \frac{p_h(1-p_h)}{n_h},$$

der

$$p_h = \textit{sann andel sysselsatte i etterstratum h (i populasjonen)}.$$

Hva dette innebærer for variasjonen på den predikerte verdien $\hat{p}_{h,g}^{pred}$ er mer komplisert, bl.a. fordi det ikke er et fast antall personer som trekkes fra hvert område g. Man kan imidlertid benytte at

$$\text{var}(\hat{p}_g^{klass}) = E\left[Var(\hat{p}_g^{klass} \mid \{\underset{\sim}{\hat{p}}\})\right] + Var\left[E(\hat{p}_g^{klass} \mid \{\underset{\sim}{\hat{p}}\})\right] . \qquad (7.2)$$

Ettersom

$$E\left(\hat{p}_h(1-\hat{p}_h)\right) = E(\hat{p}_h - \hat{p}_h^2) = E\left[\hat{p}_h - \text{var}\,\hat{p}_h - (E\hat{p}_h)^2\right] = p_h(1-p_h) - \frac{p_h(1-p_h)}{n_h},$$

har vi at

$$E\left[Var(\hat{p}_g^{klass} \mid \{\underset{\sim}{\hat{p}}\})\right] = E\left[\sum_{h=1}^{r} \frac{N_{h,g}\hat{p}_h(1-\hat{p}_h)}{N_g^2}\right] = \sum_{h=1}^{r} \frac{N_{h,g}p_h(1-p_h)}{N_g^2}\left[\frac{n_h - 1}{n_h}\right]$$

og

$$Var\left[E(\hat{p}_g^{klass} \mid \{\underset{\sim}{\hat{p}}\})\right] = Var\left[\sum_{h=1}^{r} \frac{N_{h,g}\hat{p}_h}{N_g}\right] = \sum_{h=1}^{r} \frac{N_{h,g}^2 p_h(1-p_h)}{N_g^2 n_h}.$$

Dersom vi setter dette inn i (7.2), får vi

$$\text{var}(\hat{p}_g^{klass}) = \sum_{h=1}^{r} \frac{N_{h,g} p_h (1-p_h)}{N_g^2} \left[\frac{n_h - 1}{n_h} \right] + \sum_{h=1}^{r} \frac{N_{h,g}^2 p_h (1-p_h)}{N_g^2 n_h}$$

Dette betyr at jo mindre område g man skal finne tall for, jo mer betydning har MC-usikkerheten for det endelige estimatet (det er første ledd i formelen ovenfor som øker mest når N_g avtar. Dette er jo også rimelig: på små områder vil tilfeldig tilordning av statuser kunne gi svært varierende resultater.

For deterministisk metode er det også andelen sysselsatte i hvert etterstratum som benyttes. Med lik etterstratifisering vil utvalgsgenereringen bidra likt med variasjon til stokastisk og deterministisk metode. Imidlertid er deterministisk metode ikke avhengig av etterstratifisering for å fungere, og uten etterstratifisering, er det kun andelen sysselsatte i hele utvalget som benyttes, og variansen til denne er meget liten. Deterministisk metode slik den er foreslått uten etterstratifisering har veldig liten varians pga. genereringen av utvalget og dermed veldig liten varians totalt, mens stokastisk metode altså har større variasjon.

7.6.4 Oppsummering

Det endelige valget av etterstratifiseringsvariable er en blanding av hensynet til å redusere skjevhet og varians for sysselsettingsestimatet for hele populasjonen, men også av hensynet til å redusere skjevhet (og varians) for estimater for publiseringsgrupper. Skjevhet etter alder betyr f.eks. lite for estimering av totalt sysselsettingsnivå, men er viktig når vi krysstabulerer sysselsetting mot alder.

7.6.5 Valg av etterstrata i praksis og sammenlikning av metodene

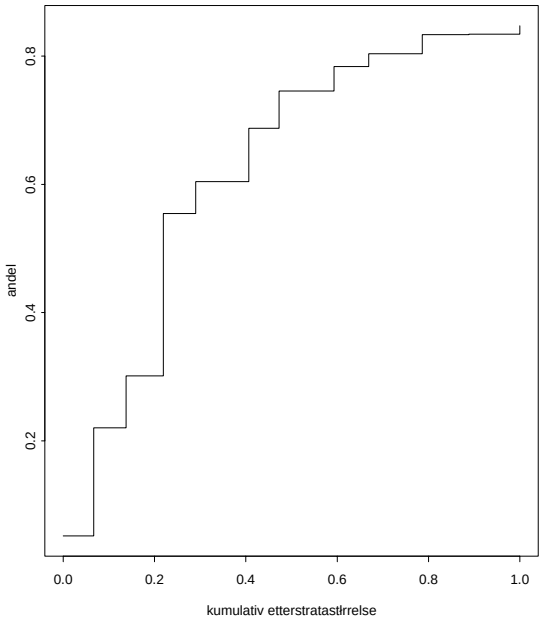
Vi vil her se på de to metodene anvendt i praksis på data fra 1992. Vi vil se mest grundig på stokastisk metode, ettersom deterministisk metode er behandlet i Bråthen & Fosen (1998). For tabeller og figurer over andel AKU-sysselsatte etter arbeidsmarkedsvariabelen (vedlegg 4), vil den etterstratumstørrelsen som er målt ikke være størrelsen til etterstratumet gjeldende for AKU-uka, men størrelsen til det tilsvarende etterstratum den 1. uka i november (som er tidspunktet man skal klassifisere for). Årsaken til et slikt valg er at det kun er for tidspunktet 1. uke i november at vi kan fordele hele populasjonen i etterstrata, og det er for dette tidspunktet vil skal klassifisere (se 6.2.1). Ideellt sett skulle vi i analysen kun benyttet det utvalget som har AKU-uke lik 1. uke i november, men det utvalget ville vært for lite.

Tabell 3 viser andel AKU-sysselsatte etter aldersgruppe, og fordelingen av alder i hhv. populasjon og utvalg. Vi ser at andelen sysselsatte stort sett ligger over 80%, men 19,5% av populasjonen ligger i etterstrata der andelen AKU-sysselsatte er mellom 25% og 75%. Figur 3 illustrerer dette grafisk. Her illustrerer lengden på x-aksen størrelsen på etterstrataene, og y-aksen viser andel AKU-sysselsatte. Etterstrataene er sortert etter økende AKU-sysselsettingsandel. Jo brattere figuren er i området der andelen AKU-sysselsatte er mellom f.eks. 0,3 0g 0,7, jo bedre er det.

Tabell 3: Fordelingen av alder i populasjonen og AKU-utvalget. Siste kolonne er AKU-sysselsettingsnivå etter alder.

Alders-gruppe	Andel av populasjonen	Andel av utvalget	AKU-sysselsettings-nivå
16-19	7,6	7,7	30,8
20-24	10,8	10,4	61,4
25-29	10,8	10,5	76,2
30-34	10,2	10,4	82,7
35-39	10,1	10,1	85,0
40-44	9,8	10,0	85,9
45-49	9,6	9,2	84,5
50-54	6,8	6,9	80,2
55-59	5,8	6,0	70,4
64-64	6,1	6,4	55,5
65-69	6,3	6,4	22,0
70+	6,1	6,0	5,6

Figur 3 Sysselsettingsandeler iht. AKU etter kumulativ etterstratumstørrelse. Etterstrataene er sortert etter økende AKU-sysselsettingsandel. Etterstratifisering etter alder (12 verdier).



Etterstratifisering etter alder bidrar til å redusere frafallsskjevhet. Vi ser av tabell 3 at utvalget er ikke-representativt etter alder: unge personer er underrepresentert og dette skyldes antakelig frafall. Samtidig ser vi av tabellen at ulike aldersgrupper har ulik sysselsettingsandel i AKU. Dette betyr at skjevheten i aldersfordelingen påvirker de anslåtte sysselsettingstallene -- frafallet er ikke-ignorerbart. Ved å etterstratifisere etter alder eliminerer man denne feilkilden (fordi man lar hver person telle mer der hvor utvalget har blitt unaturlig lavt pga. frafall). Uten etterstratifisering ville vi fått samme andel AKU-sysselsatte i populasjonen som i AKU-utvalget, dvs. 65,3%, men etter etterstratifiseringen viser det seg at dette tallet ikke nærmer seg det offisielle tallet på sysselsatte i populasjonen (de publiserte

AKU-tallene). I dette tilfellet slo skjevhetene i ulike retninger slik at andel sysselsatte i hele populasjonen ikke ble annerledes enn om man ikke hadde etterstratifisert etter alder.

En arbeidsmarkedsvariabel med 12 verdier er konstruert som et forsøk på å redusere arbeidsmarkedsinformasjonen ned til en variabel (se vedlegg 4). Vi ser av tabell 4 at utvalget også er systematisk skjevt etter denne arbeidsmarkedsvariabelen, antakelig pga. frafall. Samtidig ser vi her i enda større grad enn da vi brukte alder, at arbeidsmarkedsvariabelen har innvirkning på sysselsettingsstatus i AKU. Dersom vi retter opp skjevheten i fordelingen av arbeidsmarkedsvariabelen ved å benytte den som etterstratifiseringsvariabel i tillegg til alder, viser det seg at vi i populasjonen får klassifisert 64,4% som stysselsatte, noe som er lik det offisielle estimatet basert på AKU, dvs. at vi har fått korrigert for skjevheter skapt via frafall. Skjevheten i fordelingen av arbeidsmarkedsvariabelen førte altså til en feilestimering av sysselsettingsandelen i populasjonen på nesten ett prosentpoeng. Selv om man har funnet en etterstratifisering som gir riktige tall for totalt andel sysselsatte, har vi imidlertid ingen garanti for at det ikke er frafall etter andre variable og som gir skjevheter i våre anslag på sysselsetting etter f.eks. fylke.

Tabell 4 viser også hvor god etterstratifiseringen er når det gjelder å få liten varians[27] i sysselsettingstallet basert på stokastisk klassifikasjon når vi etterstratifiserer utelukkende etter arbeidsmarkedsvariabelen. Vi ser at det kun er 8,6% av populasjonen som er i etterstrata med AKU-sysselsettingsandel mellom 30% og 70%. Da vi kun benyttet alder som etterstratifiseringsvariabel, var det tilsvarende tallet ca. dobbelt så stort, som vi har sett tidligere.

Tabell 4 Fordelingen av arbeidsmarkedsvariabelen (vedlegg 4) i populasjonen og AKU-utvalget. Siste kolonne er AKU-sysselsettingsnivå etter arbeidsmarkedsvariabelen.

Arbeidsmarkeds-variabel	Andel av populasjonen (1. uke i nov)	Andel av utvalget (1. uke i nov)	AKU-sysselsettings-nivå (AKU-uke)[28]
00	6,1	5,5	13,5
06	18,5	17,6	5,9
07	0,3	0,3	38,6
15	1,3	1,2	22,0
16	2,9	2,9	45,8
18	0,4	0,4	38,7
23	7,1	7,0	25,8
24	4,7	4,5	67,8
25	3,5	3,4	88,4
26	2,1	2,0	82,7
27	0,3	0,3	34,8
28	52,8	54,9	95,7

Vi etterstratifiserer nå etter både alder og arbeidsmarkedsvariabelen. Slik som figur 3 illustrerte andelen sysselsatte i ulike etterstrata ved etterstratifisering etter alder, viser figur 4 situasjonen ved etterstratifisering etter både alder og arbeidsmarkedsvariabelen. Vi ser umiddelbart at grafen i figur 4 er mer bratt enn grafen i figur 3, men vil ikke være særlig brattere enn ved etterstratifisering kun etter arbeidsmarkedsvariabelen. Grunnen til å inkludere alder også, er å korrigere for skjevheten i aldersfordelingen i utvalget; selv om dette ikke hadde betydning for sysselsettingen totalt, kan

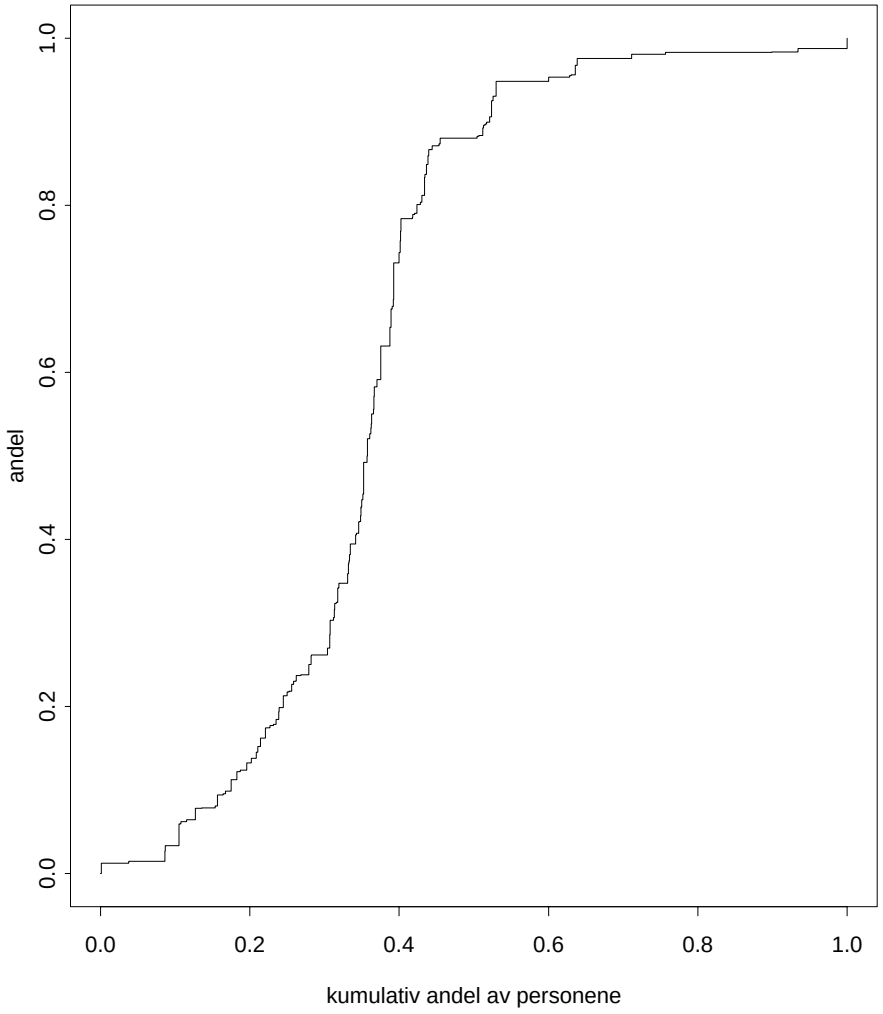
[27] Vi betrakter kun varians betinget utvalget, dvs. Monte-Carlo-usikkerheten.

[28] For denne siste kolonnen er personene gruppert etter hva slags arbeidsmarkedsvariabel de hadde i sin AKU-uke. Det er disse andelene som ville være grunnlaget for trekking av statuser i stokastisk metode dersom man kun etterstratifiserte etter arbeidsmarkedsvariabelen.

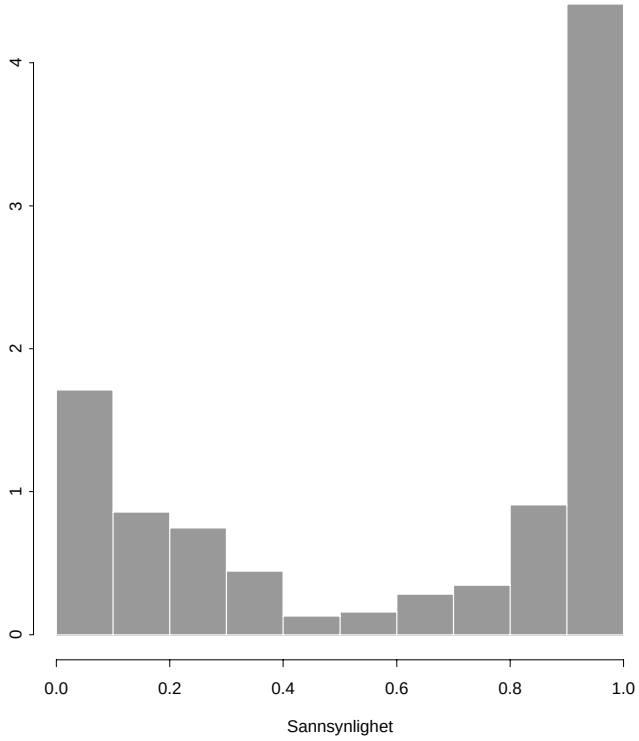
skjevheten i aldersfordelingen bidra til skjevhet når sysselsetting brytes ned etter f.eks. næring, dersom vi ikke retter opp aldersskjevheten ved å etterstratifisere etter alder.

Figur 5 handler om samme situasjon som figur 4, men viser mer direkte hvor mange personer som blir klassifisert med de ulike sannsynlighetene. Vi kan se direkte hvor stor andel personer som f.eks. blir klassifisert som sysselsatte med sannsynlighet 40-60%, nemlig 2,8% av populasjonen (andelen er arealet av søyle 5 og 6, og dersom arealet multipliseres med 100 så får vi tallet i prosent). "Myntkast" som metode for å klassifisere sysselsettingsstatus forekommer altså for relativt få personer.

Figur 4 Sysselsettingsandeler iht. AKU etter kumulativ etterstratumstørrelse. Etterstrataene er sortert etter økende AKU-sysselsettingsandel. Etterstratifisering etter arbeidsmarkedsvariabel (12 verdier) og alder (12 verdier).



Figur 5 Histogram over de beregnede sannsynlighetene for at hvert enkelt individ skal klassifiseres som sysselsatt ved stokastisk klassifikasjon. Sannsynligheten framkommer ved at vi innenfor hvert av de ca. 100 etterstrata finner andelen AKU-sysselsatte, og denne andelen tilordnes hvert individ i etterstratumet som individets sannsynlighet for å klassifiseres som sysselsatt. Arealet til hver søyle er andelen individer som får trukket sin sysselsettingsstatus med den sannsynlighet som indikeres på x-aksen. Etterstratifisering etter arbeidsmarkedsvariabelen og alder.



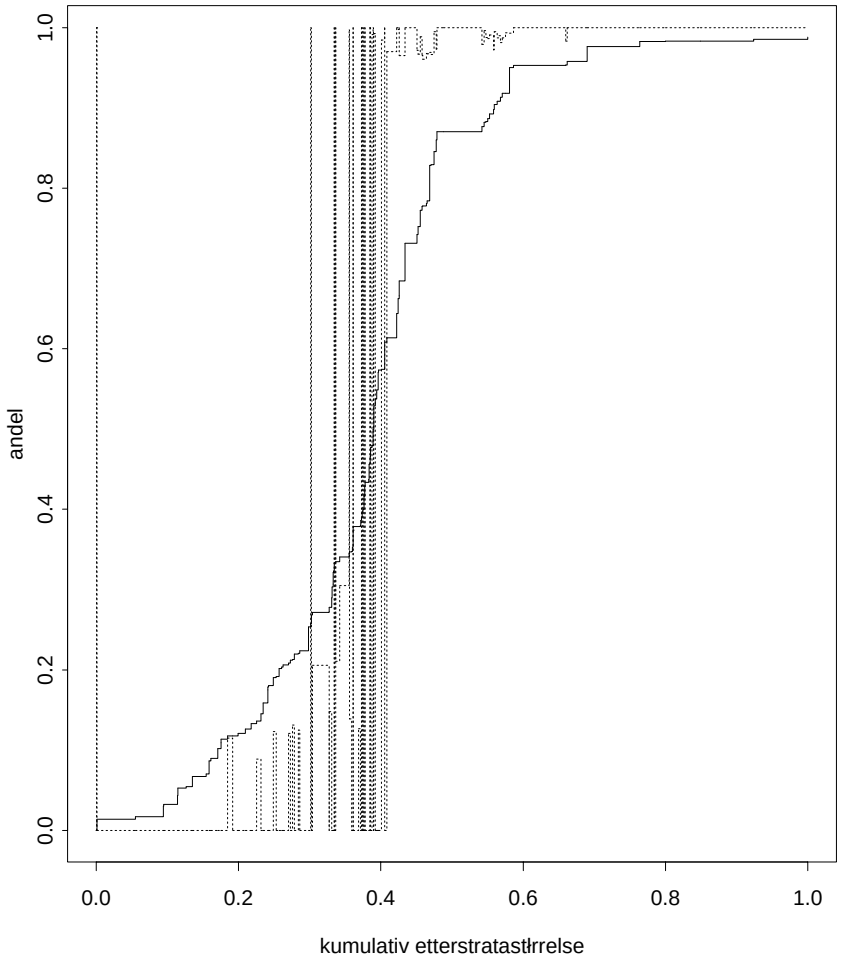
7.6.6 Stokastisk vs. deterministisk

Figur 6 viser avviket mellom nivået på andel sysselsatte innenfor hvert etterstratum i utvalget for hhv. stokastisk (etterstratifisert etter arbeidsmarkedsvariabelen og alder) og deterministisk metode. Akkurat når man ser på sysselsetting i hvert etterstratum er stokastisk metode helt overlegen ved at den iht. AKU gir riktig andel klassifiserte som sysselsatte i utvalgsdelen av hvert etterstratum (ettersom det trekkes enkelt tilfeldig). Dette innebærer imidlertid at vi kan si ganske mye om kvaliteten til den deterministiske metoden for å måle andel sysselsatte etter etterstratifiseringsvariablene. Vi ser at deterministisk metode bommer noe. Dette skyldes den skjevheten man har innført ved at alle med lik arbeidsmarkedsinformasjon får samme status.

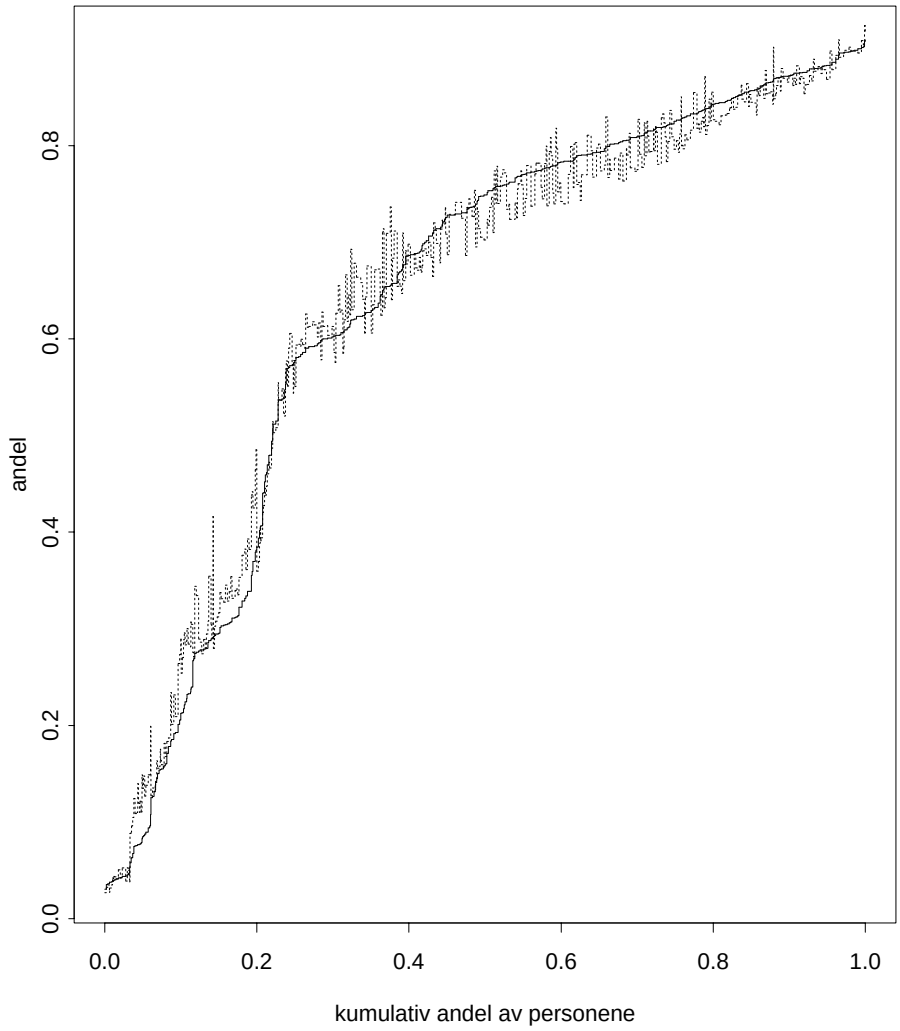
I figur 7 ser vi hvilket sysselsettingsnivå som hhv. stokastisk og deterministisk klassifikasjon gir i populasjonen for en mulig publiseringsgruppering, nemlig krysssklassifisering av kjønn, alder og fylke. Vi ser at avviket stort sett ligger innenfor ± 5 prosentpoeng, men at det i ett tilfelle kommer opp i 10 prosentpoeng. Stokastisk metode leder til varians[29] i andel sysselsatte ettersom vi har en tilfeldig mekanisme som tilordner sysselsettingsstatus. Når grupper blir små kan denne variansen bli stor (se 7.6.3), slik at det ikke er mulig å si hvilken av metodene som er mest feil.

[29] Dette er ikke utvalgsvarians, men en simuleringsvarians som uttrykker variasjon gitt det utvalget vi faktisk har.

Figur 6) Sysselsettingsandeler i ulike etterstrata i populasjonen for 1. uke i november med stokastisk metode (heltrukket linje) og deterministisk metode (stiplet linje) etter kumulativ etterstratumstørrelse. Etterstrataene er sortert etter økende andel klassifisert som sysselsatte (stokastisk metode). Etterstratifisering etter arbeidsmarkedsvariabel og alder (begge 12 verdier).



Figur 7 Som figur 6, men nå ser man ikke på andeler sysselsatte etter etterstrata, men etter krysssklassifisering av kjønn, alder (12 verdier) og fylke. Gruppene er som etterstrataene i figur 6 sortert etter økende andel sysselsatte beregnet ved stokastisk klassifikasjon.



7.6.7 Mikrokonsistens

Under visse antakelser vil mikrokonsistensen bli større ved deterministisk metode enn ved stokastisk metode.

Mikrokonsistens vil si i hvor stor grad hver enkelt person er klassifisert riktig. Dette er viktig fordi at vi ikke har mulighet til å måle feilen i klassifisert sysselsettingsnivå innenfor små områder på en sikker måte. Ved å få en høy mikrokonsistens, garderer vi oss litt mot nivåavvik for små områder. Likevel er denne garderingen av begrenset betydning så lenge ikke mikrokonsistensen er svært høy.

En innvending mot stokastisk metode er at det for mange etterstratifiseringer faktisk viser seg at andel som er riktig klassifiserte blir lavere enn ved deterministisk klassifikasjon. Dette gjelder når klassifikasjonsgrupperingen som ligger bak den deterministiske metoden inngår som etterstratifiseringsvariabel, og så lenge etterstrataene ikke er for små, kan man alltid tilnærmet legge til

klassifikasjonsgrupperingsvariabelen som etterstratifiseringsvariabel uten at den stokastiske metoden blir dårligere (se vedlegg 3).

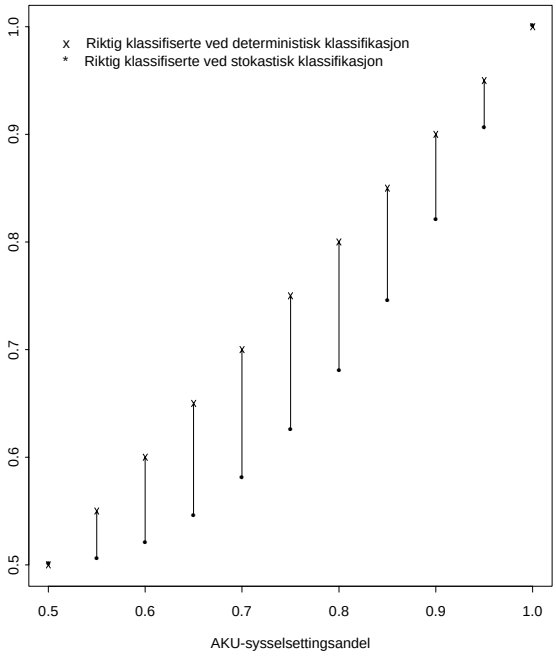
Med situasjonen nevnt ovenfor vil alle etterstrata enten tilhøre en gruppe der alle iht. deterministisk klassifikasjon er sysselsatte eller alle er ikke-sysselsatte. For et etterstratum der andel AKU-sysselsatte er $\hat{p}$ (og vi antar nå at AKU er riktig), vil den deterministiske metoden få en andel riktig klassifiserte på eksakt $\hat{p}$ dersom $\hat{p} > 0.5$, og $1 - \hat{p}$ imotsatt fall , noe som kommer av at med $\hat{p} > 0.5$ vil alle bli klassifisert som sysselsatte. Da vil $\hat{p} \cdot 100\%$ av disse faktisk være sysselsatte og ergo riktig klassifiserte. Med stokastisk klassifikasjon viser det seg imidlertid (vedlegg 1) at med D som betegnelse på antall riktig klassifiserte gitt utvalget, er $E(D_h / N_h) = \hat{p}_h^{\,2} + (1 - \hat{p}_h)^2$ og $Var(D_h / N_h) = \hat{p}_h(1 - \hat{p}_h)/N_h$. Det stokastiske i D_h skyldes som den tilfeldige tilordningen av sysselsettingsstatuser. For en delpopulasjon på størrelse 100 vil standardavviket til D_h / N_h være 4-5 prosentpoeng når $\hat{p}_h$ er mellom 20% og 80%. For mer ekstreme $\hat{p}_h$ blir standardavviket lavere.

Figur 8 viser hva $E(D_h/N_h)$ er for ulike AKU-sysselsettingsandeler $\hat{p}_h$.[30] Vi ser at for $\hat{p}_h = 1$ og $\hat{p}_h = 0.5$, blir det like riktig som deterministisk, men for alle andre $\hat{p}_h$ blir det lavere andel riktig klassifiserte med stokastisk prediksjon. Et tilnærmet 95% konfidensintervall er $\hat{E}(D_h / N_h) \pm 1{,}96 \cdot SE(D_h / N_h)$, der *SE* betyr estimert standardavvik. Dette vil vil si et avvik på mellom ± 8 og ± 10 prosentpoeng dersom delpopulasjonen er av størrelse 100 og $\hat{p}$ er mellom 0,2 og 0,8. Dersom vi i figur 8 betrakter punktet som indikerer andel riktig klassifiserte for stokastisk metode for AKU-sysselsettingsandel på 70%, ser vi at dersom vi går åtte prosentpoeng lavere enn forventet feilklassifikasjon, havner vi på et nivå under 50%. Dersom vi tenker oss mange etterstrata med AKU-sysselsettingsandel på 70% betyr dette at stokastisk klassifikasjon for 5% av disse etterstrataene gir korrekt klassifikasjon for mindre enn 50% av personene. Alternativt kan vi tenke oss at vi gjennomfører den stokastiske metoden mange ganger. For et etterstratum med AKU-sysselsettingsandel på 70% vil man 5% av gangene få at færre enn halvparten av personene er korrekt klassifisert.

Dersom man tar hensyn til at AKU kun tilnærmet er sannheten, vil stokastisk klassifikasjon være mest sårbar for dette, ettersom det er denne metoden som benytter AKU mest aktivt. Deterministisk klassifikasjon slik den er definert her er mer robust overfor svakheter ved AKU ettersom registeret benyttes mer aktivt.

[30] For $\hat{p}_h < 0.5$ blir det en helt symmetrisk situasjon (symmetri om den vertikale linjen dannet ved AKU-andel lik 0.5).

Figur 8 Forventet andel feilklassifiserte ved den stokastiske og ved den deterministiske klassifikasjonsmåten når klassifikasjonsgrupperingen bak sistnevnte metode inngår som etterstratifiseringsvariabel (se forklaring i teksten), etter faktisk andel sysselsatte.



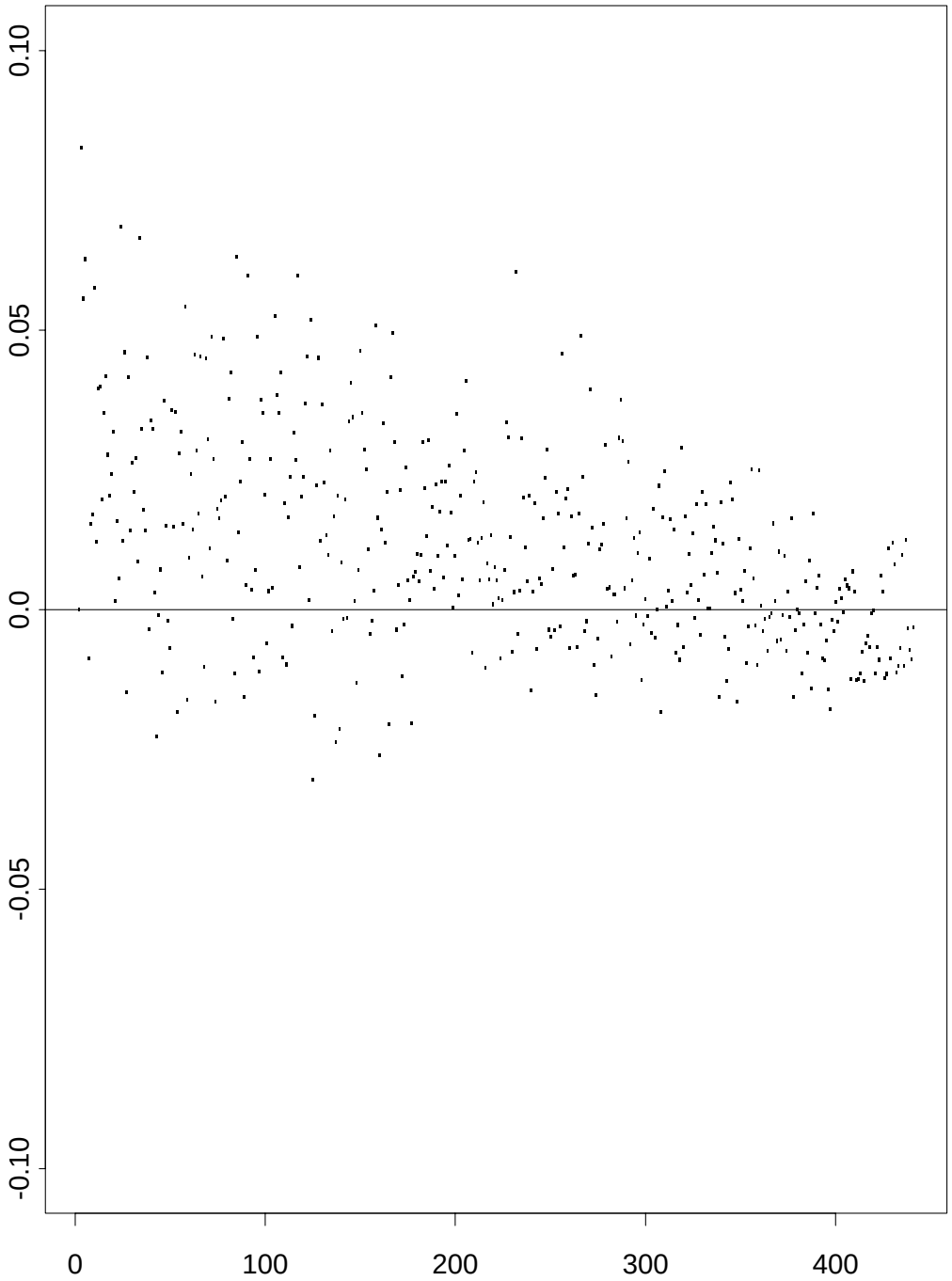
7.6.8 Små områder

Figur 9 viser avviket i prosentpoeng mellom andelen sysselsatte ved stokastisk og deterministisk klassifikasjon i populasjonen, for hver av kommunene i landet. Der hvor figuren markerer en figur over null, er det deterministisk metode som gir høyest andel. Tilsynelatende gir deterministisk klassifikasjon flere sysselsatte enn stokastisk metode, men dette skyldes at punktene lengst til høyre på figuren alle har færre sysselsatte ved deterministisk metode, og disse punktene (Oslo, Bergen, Trondheim) representerer ganske store deler av befolkningen.

Innenfor en gitt gruppe vil hver person ha fått tilordnet sysselsettingsstatus etter ulik sannsynlighet. For enkelhetsskyld antar vi at alle tilordninger skjer med sannsynlighet 0,8 eller 0,2 (da har vi funnet en ganske god etterstratifisering). Ved å bruke tabell 2 i avsnitt 7.6.3 ser vi at for et område med 200 innbyggere, er usikkerhetsmarginen på 5,5 prosentpoeng. Altså vil et 95% prediksjonsintervall i dette tilfellet (med verdi 80%) ligge fra 79% til 91%. For 5% av slike tabeller vil prediksjonsintervallet, med bredde på ca. 11 prosentpoeng, likevel ikke treffe andelen sysselsatte i AKU. Dette viser at usikkerheten er svært stor på små områder. Ikke minst er det et pedagogisk problem dersom man for en gitt kommune og den samme tilgjengelige informasjon tilgjengelig, kanskje ville få 60% dersom man gjør klassifikasjonen en dag, mens man får 62% dersom man gjør klassifikasjonen på nytt en time senere. Når man skal bryte ned tallene enda mer, f.eks. for alder x kjønn innenfor en liten kommune vil dette få enda mer dramatiske utslag.

På små områder kan man tenke seg mer avanserte metoder som ville kunne rette opp noen av problemene vi får når vi bryter ned tallene på mindre områder, og det kanskje viser seg at modellantakelsene om at globale sammenhenger (f.eks. andel sysselsatte i etterstrata) ikke holder lenger. Vi vil ikke gå nærmere inn på denne problemstillingen her.

Figur 9 Differansen populasjonen mellom andel sysselsatte iht. deterministisk og iht. stokastisk klassifikasjon, etter kommune. Kommunene er sortert etter økende størrelse og er representert med løpende indeks på x-aksen. Avviket er i prosentpoeng (dvs.at dersom andelene er 50% og 56% for en kommune, så er avviket 6 prosentpoeng). Etterstratifisering etter arbeidsmarkedsvariabel (12 verdier) og alder (12 verdier).



7.6.9 Problemet med multiple forsøk

I alt skal mange tusen tabeller lages i FoB, og dersom man skulle velge stokastisk metode og dermed i hver av disse tabellene generere sysselsettingsstatus stokastisk, ville andelen feilklassifisering og avviket i nivå variere mellom de tusenvis av tabellene. Vi så i forrige delkapittel at vi for fem prosent av alle sysselsettingstall for områder på størrelse 200, vil bomme med mer enn ca. 5 prosentpoeng. I FoB kan det være snakk om mange tusen slike tall som skal produseres, og 5% av 5000 er 100 tall som til dels blir grovt feil.

For deterministisk metode (uten etterstratifisering) vil man på tilsvarende måte ha feil som skyldes skjevheten man har innført og at feilene også her varierer stokastisk fra gruppe til gruppe (Det kan jo umulig være samme forekomster av feil i alle grupper). Deterministisk metode har imidlertid den fordelen at den individuelle sysselsettingsstatus har en stabil forankring i registerkjennetegn.

8 Prediksjon eller klassifikasjon

Vi ha ovenfor skissert to måter å klassifisere på. La oss nå anta at vi hadde valgt stokastisk metode. For denne metoden er sysselsettingsandel i etterstratumet utgangspunktet for klassifikasjon ved at man trekker hvilke individer som skal bli sysselsatte. Dersom man kunne unngå denne trekkingen ville man unngå mye av usikkerheten. Dersom hvert individ kun er predikert med en sannsynlighet for sysselsetting, kan man likevel benytte dette til å lage tabeller over sysselsatte. For et tabuleringsområde, f.eks. unge kvinner i Etnedal, vil summen av sannsynlighetene for dette området, være et godt anslag på andelen sysselsatte i området. Dette anslaget vil faktisk være bedre enn med forutgående trukket status. Dette skyldes at trekkingen egentlig innebærer at hvert individs sannsynlighet forskyves fra den estimerte sannsynlighet og til 0 eller 1. Når man skal summere opp disse 0-er og 1-ere vil disse gjennomsnittlig bli lik den estimerte sannsynlighet dersom man summerer over en hel gruppe. Dersom man derimot summerer over få individer, f.eks. en liten del av en kommune, vil de individuelle forskyvingene lede til feil. Denne feilen skyldes utelukkende den tilfeldige trekkingen og er således en usikkerhetsfeil.

Beregningsmessig vil ren prediksjon lede til problemer, og det er når man skal benytte sysselsetting ikke som tabuleringsvariabel, men derimot f.eks skal se på utdanningsnivået etter sysselsettingsstatus. For å løse denne oppgaven uten å klassifisere sysselsettingsstatuser, trengs en veldig stor deltakelse av metodestatistikere. Årsaken er at tabellverket i FoB blir meget omfattende med en rekke fleksible kombinasjoner, og at arbeidsmarkedsdelen av FoB-individfilen skal være utgangspunkt for arbeidsmarkedsstatistikk i S260 (som også er den seksjonen som har ansvaret for produksjonen for denne delen av FoB-individfilen).

Med tanke på at resten av FoB-fila er en individfil der hvert individ har en verdi på alle variable, så bryter det litt dersom hvert individ ikke skal betraktes som sysselsatt eller ikke, men som sysselsatt med en viss sannsynlighet. Sett med statistikerens briller er ikke dette noe problem, men for alle andre virker dette litt uheldig.

I alle celler i alle tabeller skal man ha antall sysselsatte. Dette betyr at man uten klassifikasjon må runde av anslått celleantall til nærmeste hele person. For en tabell ville avrunding i en celle kreve justering av andre celler for å få marginaler til å stemme, og når man aggregerer fra kommune til fylkesnivå, vil avrundingene som er gjort på deltaljerte tabeller på kommunenivå kunne føre til at når man summerer alle tabeller fra alle kommuner i et fylke, så vil summen ikke stemme med de fylkestallene man har publisert. Denne inkonsistensen vil være relativt liten (kanskje noen titalls personer i en celle i fylkestabellen), men det kan være et pedagogisk problem å forklare denne inkonsistensen på en overbevisende måte for observante brukere.

Vedlegg 1 Fordelingen til feilklassifiserte ved stokastisk klassifikasjon

Vi vil nedenfor se på fordelingen av feilklassifiserte gitt utvalget ved stokastisk klassifikasjon, og lar

$$Y_i = 1, \textit{ dersom person } i \textit{ er klassifisert sysselsatt}$$
$$p_0 = \textit{andelen faktisk sysselsatte i populasjonen}$$
$$= \textit{ sannsynligheten for at en tilfeldig trukket person er sysselsatt.}$$

Først antar vi for enkelhetsskyld at vi kun har ett etterstratum (hele populasjonen). Sannsynligheten for at person *i* blir klassifisert som sysselsatt er

$$P(Y_i = 1) = 1 - P(Y_i = 0) = \hat{p} = \sum_{i=1}^{n} \frac{Y_i}{n} = \textit{ andel sysselsatte i utvalget}\ .$$

Vi ser her på situasjonen betinget utvalget, og ønsker å predikere antall feilklassifiserte personer,

$$t = \sum_{i=1}^{N} D_i\ ,$$

der

$$D_i = \begin{cases} 1, \text{ dersom riktig klassifisert} \\ 0, \text{ ellers} \end{cases}\ .$$

Vi lar

$$I_i = \begin{cases} 1, \text{ hvis faktisk sysselsatt} \\ 0, \text{ ellers} \end{cases}\ ,$$

og vi ser dermed at

$$\{D_i = 1 \mid I_i = 1\} \Leftrightarrow \{Y_i = 1 \mid I_i = 1\} \quad \textit{som inntreffer med sannsynlighet } \hat{p}$$
$$\{D_i = 1 \mid I_i = 0\} \Leftrightarrow \{Y_i = 0 \mid I_i = 0\} \quad \textit{som inntreffer med sannsynlighet } (1 - \hat{p})\ .$$

Antall riktig klassifiserte i hele populasjonen er

$$D = D^{(0)} + D^{(1)}\ ,$$

der

$$D^{(0)} = \sum_{i=1}^{N} D_i (1 - I_i) = \textit{antall riktig klassifiserte gitt at } \{I_i = 0\}\ .$$

$$D^{(1)} = \sum_{i=1}^{N} D_i I_i = \textit{antall riktig klassifiserte gitt at } \{I_i = 1\}$$

Alle personer klassifiseres uavhengig av hverandre, og ettersom denne for alle personer med $I_i = 1$ er $\hat{p}$, er $D^{(1)}$ binomisk fordelt $binom(N\hat{p}, \hat{p})$. Tilsvarende er $D^{(0)} \sim binom(N(1-\hat{p}), 1-\hat{p})$. Ettersom $D^{(0)}$ og $D^{(1)}$ er basert på to disjunkte mengder av personer (hhv. de faktisk ikke-sysselsatte og de faktisk sysselsatte) vil disse to være uavhengige ettersom alle personer klassifiseres uavhengig av hverandre. Dette betyr at *D* har flg. egenskaper

$$E(D) = N\hat{p}^2 + N(1-\hat{p})^2$$
$$Var(D) = N\hat{p}\hat{p}(1-\hat{p}) + N(1-\hat{p})(1-\hat{p})\hat{p} + 2\,\mathrm{cov}(D^{(0)}, D^{(1)})$$
$$= N\hat{p}(1-\hat{p})$$

Man kan i stedet for å ha kun ett etterstratum, ha flere etterstrata, og da i stedet for D beregne en D_h for hvert etterstratum. Ettersom D_i er uavhengig for hvert eneste individ, kan vi betrakte hvert etterstratum for seg, og til slutt finne $D = \sum_{\forall h} D_h$.

Vedlegg 2 Prediksjon basert på likelihood

2.1. Notasjon

For å komme videre må det innføres litt notasjon som gjelder langt mer generelt enn den situasjonen som er beskrevet overfor:

Y	-Responsvariabelen slik denne ville vært dersom vi intervjuet personene f. eks. i AKU
X_k	-Registervariabler som måles direkte i registrene(men der oppretting av feil er gjennomført), k=1,2,...K
$\hat{Y}$	-Den predikerte responsvariablen fra registrene
β_l	-Ukjente parametre som skal estimeres, l=1,2,...,L
G	-En klassifisering av personene, f. eks. etter region, utdanning, næring osv. G vil altså bestå av flere dimensjoner. Vi vil bruke betegnelsen g på en verdi av G.
i	-Person i i populasjonen
n	-Antall personer i hele utvalget
m	-Antall personer i treningsutvalget
$\hat{s}_G$	-Treningsutvalget som tilhører G
$\tilde{s}_G$	-Evalueringsutvalget som tilhører G
m_G	-Antall personer i treningsutvalget som tilhører G
$(n-m)_G$	-Antallet i evalueringsutvalget som tilhørere G

2.2. Finne prediksjonsfunksjonen

Problemstillingen er å finne en metode for å predikere en registerbasert variabel $\hat{Y}$ som vi kan lage FoB-statistikk på grunnlag av og samtidig også måle kvaliteten til denne i forhold til den "sanne" variabelen Y. Legg merke til at denne prediksjonsfunksjonen skal kunne ta de samme verdiene som Y. Er Y en binær variabel skal også $\hat{Y}$ være en binær variabel.

Helt generelt kan vi formulere problemstillingen som at vi ønsker å finne en predikert registervariabel

$$(1) \qquad \hat{Y} = \hat{Y}(X_1, X_2, \ldots, X_K, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_L | G)$$

der $\hat{\beta}_l$,l=1,2,...,L betegner estimerte parameterverdier og | G betyr at dette er gjennomført betinget med hensyn på hver verdi av G. For å kunne evaluere metoden riktig må vi i første omgang bestemme (1) ved hjelp av treningsutvalget som består av m personer.

Den metodestatistiske oppgaven er å finne den regelen i (1) som gir best mulig kvalitet sammenliknet med den sanne verdien. Vi kan tenke oss to typer kriterier her. For hver gruppe G ønsker vi å finne en prediksjon som har minst mulig kvadratavik fra den sanne verdien:

$$(2) \qquad \min_{(\beta_1,\beta_2,\ldots,\beta_L)} E[(\hat{Y}-Y)^2 | G] = \min_{(\beta_1,\beta_2,\ldots,\beta_L)} E\{[\hat{Y}(X_1, X_2, \ldots, X_K, \beta_1, \beta_2, \ldots, \beta_L) - Y]^2 | G\}$$

Et slikt kriterium garanterer ikke at den predikerte variabelen er forventningsrett. Dersom vi ønsker å begrense oss til prediksjoner som er forventningsrette innen hver gruppe har vi følgende kriterier

(3)

$$\min_{(\beta_1,\beta_2,\ldots,\beta_L)\,E[(\hat{Y}-Y)|G)]=0} Var(\hat{Y}|G) = \min_{(\beta_1,\beta_2,\ldots,\beta_L)\,E[(\hat{Y}-Y)|G)]=0} Var[\hat{Y}(X_1, X_2,\ldots,.X_K, \beta_1, \beta_2,\ldots,\beta_L)|G]$$

For å bestemme formen på prediksjonsfunksjonen og estimere parametrene bruker vi likelihood-funksjonen til de m observasjonene i treningssettet. I neste avsnitt beskrives dette for to tilfeller.

En tredje mulighet er å lage en sammensatt prediksjon der begge er funnet gjennom kriteriet (3), men der den ene bygger på en grov regional inndeling (f. eks. fylke) den andre en finere regional fordeling (f. eks. kommune). Dersom vi skal lage tall for kommuner vil den første regelen ikke lenger oppfylle (3) betinget på kommune slik at heller ikke den sammensatte oppfyller kriteriet. Vi kan da formulere dette mer presist ved at den sammensatte prediksjonen oppfyller (3) for en grov regional inndeling og (2) for en fin regional inndeling.

2.3 Likelihoodfunksjonen

Dersom Y er en binær variabel, f. eks. sysselsetting, eller en kontinuerlig variabel, f. eks. arbeidstid i uka, og at Y i dette tilfellet er normalfordelt. Dersom vi så antar at Y_i, i=1,2,..., er identisk fordelt og uavhengige innenfor hver gruppe G kan vi beskrive likelihoodfunksjonen og hvordan parameterverdiene finnes.

Først må imidlertid den betingede punktsannsynligheten, når Y er binær, og sannsynlighetstettheten når Y er en målevariabel innføres. betinget betyr betinget mhp :

(4) $f(y; \beta_1, \beta_2,..., \beta_L \mid x_1,x_2,...,x_K)$

For observasjonene i treningssettet som tilhører G,Y_i, $i \in \hat{s}_G$, kan vi nå definere likelihoodfunksjonen, L, for dette utvalget. Vi gjør dette gjennom å definere -2lnL

$$(5) \qquad -2\ln L(\beta_1,\beta_2,\ldots,\beta_L) = -2\sum_{i\in\hat{s}_G} \ln[f(Y_i;\beta_1,\beta_2,\ldots,\beta_L|X_{i1},X_{i2},\ldots,X_{iK})]$$

Ved å maksimere (5) med hensyn på parametrene finner vi estimater som i de to tilfellene vi har beskrevet tilfredsstiller (2) eller (3). Dersom vi bruker en logitmodell for Y binær og en multippel linære regresjonsmodell for Y normalfordelt kan vi beskrive (5) direkte gjennom formlene i (6) og (8).

For en logitregresjon finner vi estimatene gjennom å maksimere følgende uttrykk med hensyn på β_k-ene.

$$(6) \qquad -2\ln L(\beta_1,\beta_2,\ldots,\beta_L) = -2\sum_{i\in\hat{s}_G} \{Y_i \ln[p(\underline{X_i};\underline{\beta})] - (1-Y_i)\ln[1-p(\underline{X_i};\underline{\beta})]\}$$

der

$$(7) \qquad \ln\frac{p(\underline{X_i};\underline{\beta})}{1-p(\underline{X_i};\underline{\beta})} = \beta_0 + \sum_k \beta_k X_{ik}$$

Dersom Y er normalfordelt og vi bruker en lineær regresjonsmodell får vi følgende uttrykk som skal maksimeres med hensyn på β_k-ene.

(8) $$-2\ln L(\beta_1,\beta_2,...,\beta_L) = \frac{1}{\sigma^2}\sum_{i\in\hat{s}_G}(Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_K X_{iK}) + m_G \ln(2\pi\sigma^2)$$

2.4 Evaluere prediksjonsfunksjonen

Dersom vi nå betegner den predikerte variabelen $\hat{Y}_i$ for person i og skal evaluere hvor god metoden er må vi evaluere metoden på grunnlag av resten av utvalget, i=m+1,m+2,...,n, dvs treningssettet. La oss betegne disse verdiene for $\tilde{\tilde{Y}}_i$ som betyr

(9) $$\tilde{\tilde{Y}}_i = \hat{Y}_i(X_{i1}, X_{i2},..., X_{iK}, \hat{\beta}_1, \hat{\beta}_2,..., \hat{\beta}_L \mid G)$$

der vi altså bruker regelen slik den er utformet på treningssettet i=1,2,...m for evalueringssettet i=m+1,m+2,...,n

For å evaluere kvaliteten på (9) kan vi estimere skjevhet og varians gitt ved

(10) $$\hat{E}(\tilde{\tilde{Y}} - Y \mid G) = \frac{1}{(n-m)_G}\sum_{i\in\tilde{\tilde{s}}_G}(\tilde{\tilde{Y}}_i - Y_i) = \overline{\tilde{\tilde{Y}}}_{\tilde{\tilde{s}}_G} - \overline{Y}_{\tilde{\tilde{s}}_G}$$

(11) $$\hat{V}(\tilde{\tilde{Y}} \mid G) = \frac{1}{(n-m)_G - 1}\sum_{i\in\tilde{\tilde{s}}_G}(\tilde{\tilde{Y}}_i - \overline{\tilde{\tilde{Y}}}_{\tilde{\tilde{s}}_G})^2$$

Vedlegg 3 Hvorfor mer detaljert etterstratifisering

Vi antar enkelt tilfeldig utvalg (som AKU ikke er, men likevel ikke er så veldig langt unna.Variasjonen i hele populasjonen og i hvert etterstratum defineres på flg måte:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$\sigma_h^2 = \frac{1}{N_h}\sum_{i\in U_h}(y_i - \bar{y}_h)^2$$

der

$$U_h = etterstratum\ nr.\ h\ .$$

Fra grunnleggende litteratur om utvalgsteori (f.eks. Bjørnstad 1996, s. 38+s.46) har vi under visse betingelser om bl.a. etterstrataenes størrelser, at med $\hat{t}_{est}$ og $\hat{t}_e$ lik totalestimatorer basert på enkel oppblåsing hhv. innenfor hvert etterstratum og uten noen bruk av etterstrata,

$$Var(\hat{t}_{est}) \approx \frac{N^2}{n}(1-\frac{n}{N})\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2$$

$$Var(\hat{t}_e) \approx \frac{N^2}{n}(1-\frac{n}{N})\sigma^2$$

Imidlertid har vi at

$$\sigma^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in U_h}([y_i - \bar{y}_h] + [\bar{y}_h - \bar{y}])^2 = \sum_{h=1}^{H}\frac{N_h}{N}\sum_{i\in U_h}\frac{1}{N_h}\left\{(y_i - \bar{y}_h)^2 + (\bar{y}_h - \bar{y})^2\right\} > \sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2\ ,$$

som betyr at variansen til en etterstratifiseringsestimator er lavere enn om man ikke etterstratifiserer.

Vi ser av uttrykkene for variansene at variansen for etterstratifiseringsestimator framkomnmer ved at σ^2 erstattes med en vektet sum av σ_h^2-er. Dersom vi vurderer å dele et etterstratum opp i flere nye etterstrata, vil dette bety at den nye variansformelen kun innebærer at hver aktuell σ_h^2-er erstattes med en vektet sum av variasjonen innenfor hver delgruppe av dette etterstratumet, slik at variansen til totalestimatet reduseres enda mer. Denne prosessen må imidlertid ikke føre til at etterstrataene blir mindre enn størrelsesorden 50-100, for da gjelder ikke lenger formelen for etterstratifiseringssituasjonen ovenfor.

Vedlegg 4 Arbeidsmarkedsvariabelen

Arbeidsmarkedsinformasjonen i registerfilene er omgjort til en arbeidsmarkedsvariabel på nominalnivå[31] med 12 verdier på flg. måte der egenskapene nedenfor refererer seg til personens viktigste jobbforhold nærmest den aktuelle uka:

28: koplet lønnstakerforhold aktivt den aktuelle uka og samlet kontantlønn i året > 5000,-
27: koplet lønnstakerforhold aktivt den aktuelle uka og samlet kontantlønn i året < 5000,-
26: Tilfredsstiller ikke 27-28, men næringsinntekt fra jordbruk, skogbruk, fiske
25: Tilfredsstiller ikke 27-28, men næringsinntekt fra annen næring.
24: Tilfredsstiller ikke noe ovenfor, men ukoplet lønnsforhold eller ukoplet arbeidstakerforhold innenfor +- 2 uker fra den aktuelle uka, og og samlet kontantlønn i året > 25 000,-
23: Som 24, men der samlet kontantlønn i året < 25 000,-
18: Tilfredsstiller ikke noe ovenfor, men koplet lønnstakerforhold eller ukoplet arbeidstakerforhold innenfor +- 2 uker fra den aktuelle uka.
16: Tilfredsstiller ikke noe ovenfor, men koplet lønnstakerforhold lenger unna enn +- 2 uker fra den aktuelle uka, og og samlet kontantlønn i året > 20 000,-
15: Som 16, men der samlet kontantlønn i året < 20 000,-
07: Tilfredsstiller ikke noe ovenfor, men ukoplet lønnsforhold eller ukoplet arbeidstakerforhold lenger unna enn +- 2 uker fra den aktuelle uka, og og samlet kontantlønn i året > 60 000,-
06: Som 07, men samlet kontantlønn i året < 60 000,-
00: Resten av personene, dvs. de som ikke har noe jobbforhold aktivt på tidspunktet.

[31] Verdiene på variabelen har ingen naturlig rangordning.

Litteratur

Bjørnstad, J. F. (1995), *Utvalgsundersøkelser og prediksjon*, kompendium ved Universitetet i Trondheim-AVH.

Bjørnstad, J. F. (1996), *Generell utvalgsteori*, Internt notat.

Bråthen, M. & Fosen, J. (1998), *Definisjon av sysselsetting basert på registerinformasjon. Utarbeidelse av klassifikasjonsrutine*, Notat 98/64

Eurostat, IMF, OECD, UN & World Bank (1993), *System of national accounts 1993.*

Hellevik, O. (1991), *Forskningsmetode i sosiologi og statsvitenskap*, Universitetsforlaget.

Stålnacke, M., Hustoft, A. G. & Solheim, L. (1999), *Vurdering av kvalitet i statistikk,* Notat 99/42.

United Nations Economic Commission for Europe & Eurostat (1997?), *Recommendations for the 2000 censuses of population and housing in the ECE region*, Statistical standards and studies - No. 49.

De sist utgitte publikasjonene i serien Notater

1999/79 P.M. Holt og T. Vevle: Skattestatistikk for rederier 1996 og 1997: Dokumentasjon. 26s.

1999/80 T. Bye, Ø. Døhl og J. Larsson: Klimagasskvoter i kraftintensive næringer. Konsekvenser for utslipp av klimagasser, produksjon og sysselsetting. Regionale konsekvenser. 11s.

1999/81 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 1998. 39s.

1999/82 Ø. Kleven, E. Dalheim og D. Roll-Hansen: Innvandreres utdanning: - en pilotundersøkelse. 61s.

1999/83 E. Fidjestøl og I. Håland: Yrkeskatalog: Pr. desember 1999. 136s.

1999/84 T. Solberg: Virkning av revisjon på Avlingsstatistikk for jordbruksvekster i 1998. 24s.

1999/85 R. Choudhury, T. Eika og L. Haakonsen: KVARTS i praksis II: Systemer og rutiner i den daglige driften. 66s.

1999/86 G. Frøiland: Økonometrisk modellering av husholdningenes konsum i Norge: Demografi og formueseffekter. 55s.

1999/87 Y. Li: Beregning av elementæraggregater i konsumprisindeksen ved hjelp av generalisert gjennomsnitt. 41s.

1999/88 L. Rogstad og S.T. Vikan: Kobling av adresseregistrene i DSF og GAB 1999: Dokumentasjon av samsvar og avvik. 31s.

1999/89 E. Dalheim, J-A. S. Lie og D. Roll-Hansen: En skjemabasert komplettering av registeret over befolkningens høyeste utdanning - forprosjekt med fokus på innvandrere. 60s.

1999/90 K-A. Hovland og Å. Nossum: Flyreiser i konsumprisindeksen. 39s.

2000/1 E. Rønning: Utenlandske statsborgere og kommunestyrevalget 1999: Dokumentasjonsrapport. 34s.

2000/2 M. Bråthen: Personer registrert som yrkeshemmet i SOFA-søkerregisteret. 25s.

2000/3 A.K. Johnsen og Ø. Hokstad: FoB2001: Kvalitativ testing av boligskjema - prøveundersøkelse 1999: Dokumentasjonsnotat. 32s.

2000/4 C. Hendriks, Ø. Hokstad og R. Sønsterudbråten: FoB2001: Boligtelling - prøveundersøkelse 1999: Dokumentasjonsnotat. 60s.

2000/5 K. Bjønnes, G. Dahl og B.R. Joneid: FD - Trygd: Dokumentasjonsrapport: Økonomisk sosialhjelp 1992-1997. 31s.

2000/6 B.R. Joneid og J. Lajord: FD - Trygd: Dokumentasjonsrapport: Demografi 1992-1997. 117s.

2000/7 J. Heldal: Kalibrering av AKU: Dokumentasjon av metode og program. 28s.

2000/8 H. Hågård og L. Rogstad: FoB2001: Adresser i folkeregisteret og GAB: Rapport fra en arbeidsgruppe for adressesamordning og utredning av elektronisk datautveksling mellom DSF og GAB. 51s.

2000/9 B. Sundby: Rutiner for produksjon av statistikk over pleie- og omsorgstjenestene i kommunene 1997. 84s.

2000/10 E. Aas: På leting etter målefeil - en studie av pleie- og omsorgssektoren. 31s.

2000/11 I. Øyangen: Lokalvalgsundersøkelsen 1999: Dokumentasjonsrapport. 36s.

2000/12 E. Engelien: Arealbruksstatistikk for tettsteder: Dokumentasjon av arbeid med metodeutvikling 1999. 50s.

2000/13 F. Gundersen og A.E. Hustad: Statistikk over anmeldte lovbrudd og registrerte ofre: Dokumentasjon. 51s.

2000/14 T. Martinsen: Prosjekt over industriens energibruk. 58s.