Statistics Norway

Statistisk sentralbyrå

Li-Chun Zhang

Opplegg til en statistikk over familie- og husholdningsfordelingen i den norske befolkningen

– Mot et bedre grunnlag for undersøkelsesbasert personstatistikk

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder

Forord

Dette notatet dokumenterer resultatet fra et av delprosjektene i prosjektet "Definisjoner og standarder for gruppering av husholdninger og familier". Målet for hovedprosjektet er å komme fram til definisjoner av de mest sentrale familie- og husholdningsbegrepene, og definere standarder for presentasjon av familie- og husholdningsopplysninger. I tillegg skal prosjektet vurdere standardisert innsamling av opplysninger om familier og husholdninger, samt avklare ulike metodiske problemstillinger knyttet til utvalgsunderskelser for dette statistikkområdet.

Formålet med delprosjektet som rapporteres her, har vært å finne en mer enhetlig måte å beregne familie- og husholdningsfordelingen på i SSB. Slik det er i dag, gir ulike undersøkelser avvikende resultat når det gjelder disse fordelingene. Notatet skisserer en prinsipiell løsning av det metodiske problemet gjennom etablering av marginale husholdnings- og familiefordelinger (basert på utvalgsundersøkelser). Implementeringen av prosjektet er utsatt i påvente av et bedre datagrunnlag. Arbeidet med å framskaffe dette er i gang.

Resultatene fra de andre delprosjektene i "Definisjoner og standarder for gruppering av husholdninger og familier" vil publiseres i løpet av 1. kvartal neste år.

Prosjektet har en referansegruppe bestående av representanter fra de seksjoner i SSB som vil bli berørt av resultatene fra prosjektet.

1 Innledning

I dette korte notatet skisserer vi et opplegg for etablering av en statistikk over familie- og husholdningsfordeling i den norske befolkningen. Vi vil bl.a. gå inn på definisjon av, begrunnelse for, og fremgangsmåte for å etablere statistikken. Vedlegget inneholder en nærmere diskusjon av flere tekniske sider rundt de statistisk-metodiske problemstillingene. Arbeidet er resultatet fra et av delprosjektene i hovedprosjektet *"Definisjoner og standarder for gruppering av husholdninger og familier"*.

2 Statistikk over familie- og husholdningsfordelingen

2.1 Marginal familiefordeling gitt størrelsen på familien

Det er i prosjektet foreslått følgende detaljerte inndeling for familietyper:

1. Enpersonfamilie, person under 30 år
2. Enpersonfamilie, person 30-44 år
3. Enpersonfamilie, person 45-66 år
4. Enpersonfamilie, person 67 år og over
5. Ektepar uten barn, eldste person under 30 år
6. Ektepar uten barn, eldste person 30-44 år
7. Ektepar uten barn, eldste person 45-66 år
8. Ektepar uten barn, eldste person 67 år og over
9. Ektepar med små barn (yngste barn 0-5 år)
10. Ektepar med store barn (yngste barn 6-17 år)
11. Ektepar med voksne barn (yngste barn 18 år og over)
12. Samboerpar uten barn, eldste person under 30 år
13. Samboerpar uten barn, eldste person 30-44 år
14. Samboerpar uten barn, eldste person 45-66 år
15. Samboerpar uten barn, eldste person 67 år og over
16. Samboerpar med små barn (yngste barn 0-5 år)
17. Samboerpar med store barn (yngste barn 6-17 år)
18. Samboerpar med voksne barn (yngste barn 18 år og over)
19. Mor med små barn (yngste barn 0-5 år)
20. Mor med store barn (yngste barn 6-17 år)
21. Mor med voksne barn (yngste barn 18 år og over)
22. Far med små barn (yngste barn 0-5 år)
23. Far med store barn (yngste barn 6-17 år)
24. Far med voksne barn (yngste barn 18 år og over)
25. Andre familier

Enhver person i Norge skal tilhøre en av disse familietypene. Man kan derfor bruke nummeret til familietypen som et familiekjennemerke for vedkommende. Videre kan man fordele alle personer med samme familiekjennemerke etter størrelsen på familien. En slik inndeling av den norske befolkningen vil gi følgende tabell, som forteller hvor mange personer som lever i en bestemt type familie, gitt størrelsen på familien; m.a.o. den marginale familiefordelingen gitt størrelsen på familien[1]:

[1] Familiefilen i BESYS (SSBs system for produksjon av befolkningsstatistikk- og filer) kan ikke gi oss denne marginale familiefordelingen. Men ut fra husholdningsopplysninger og registre, kan personer i et utvalg klassifiseres riktig etter denne nye inndelingen for familietype.

	Familiekjennemerke						
Familiestørrelse	1	2	3	...	23	24	25
1 person							
2 personer							
3 personer							
...							
K personer							

2.2 Marginal husholdningsfordeling gitt størrelsen på husholdningen

Det er i prosjektet også foreslått følgende detaljert inndeling for husholdningstyper:

Énfamiliehusholdninger
1. Aleneboende under 30 år
2. Aleneboende 30-44 år
3. Aleneboende 45-66 år
4. Aleneboende 67 år og over
5. Par uten barn, eldste person under 30 år
6. Par uten barn, eldste person 30-44 år
7. Par uten barn, eldste person 45-66 år
8. Par uten barn, eldste person 67 år og over
9. Gifte par med små barn (yngste hjemmeboende barn 0-5 år)
10. Gifte par med store barn (yngste hjemmeboende barn 6-17 år)
11. Gifte par med voksne barn (yngste hjemmeboende barn 18 år og over)
12. Samboerpar med små barn (yngste hjemmeboende barn 0-5 år)
13. Samboerpar med store barn (yngste hjemmeboende barn 6-17 år)
14. Samboerpar med voksne barn (yngste hjemmeboende barn 18 år og over)
15. Mor med små barn (yngste hjemmeboende barn 0-5 år)
16. Mor med store barn (yngste hjemmeboende barn 6-17 år)
17. Mor med voksne barn (yngste hjemmeboende barn 18 år og over)
18. Far med små barn (yngste hjemmeboende barn 0-5 år)
19. Far med store barn (yngste hjemmeboende barn 6-17 år)
20. Far med voksne barn (yngste hjemmeboende barn 18 år og over)
Flerfamiliehusholdninger
21. Husholdninger med to eller flere enpersonfamilier
22. Andre husholdninger uten barn 0-17 år
23. Andre husholdninger med barn (yngste hjemmeboende barn 0-17 år)

Enhver person i Norge skal tilhøre en av disse husholdningstypene. Man kan derfor bruke nummeret til husholdningstypen som et husholdningskjennemerke for vedkommende. Videre kan man fordele alle personer som har det samme husholdningskjennemerke etter størrelsen på husholdningen. En slik inndeling av den norske befolkningen ville gi oss følgende tabell, som forteller oss hvor mange personer som lever i en bestemt type husholdning, gitt størrelsen på husholdningen; m.a.o. den marginale husholdningsfordelingen gitt størrelsen på husholdningen:

	Husholdningskjennemerke							
	Énfamiliehusholdninger					Flerfamiliehusholdninger		
Husholdningsstørrelse	1	2	3	...	20	21	22	23
1 person								
2 personer								
3 personer								
...								
K personer								

2.3 Familie- og husholdningsfordelingen

En utfylt krysstabell av de to marginale fordelingene forteller oss hvor mange personer som lever i en bestemt kombinasjon av familie- og husholdningstype gitt størrelsen på familien og husholdningen. Dette utgjør den samlede familie- og husholdningsfordelingen.

3 Begrunnelse for statistikken

En statistikk over familie- og husholdningsfordelingen som skissert ovenfor, er nyskapende i internasjonal sammenheng. Den vil først og fremst sikre konsistens mellom forskjellige familie- og husholdningsstatistikker produsert i SSB. Vi får et bedre grunnlag både for estimering og behandling av enhetsfrafall i utvalg. Siden dette gjelder alle undersøkelser med familie eller husholdning som enhet, vil den også betyr ressurssparing. Vi skal gå noe nærmere inn på disse begrunnelsene i det følgende.

3.1 Konsistens

Som produsent av befolkningsstatistikk, bør SSB til enhver tid ha en oversikt over hvor mange familier og husholdninger som finnes i Norge, og hvordan de fordeler seg i befolkningen. Man kan sikre konsistens mellom forskjellige familie- og husholdningsstatistikker ved hjelp av en statistikk over familie- og husholdningsfordelingen, fordi denne vil utgjøre et felles grunnlag for å blåse opp tall fra utvalgsundersøkelser. Slik konsistens vil klart styrke troverdigheten i de relevante statistikkene.

3.2 Mer aktuell familiestatistikk

Den nye standard inndelingen av familietype svarer bedre til brukernes behov enn den som har vært i bruk til nå. I dag baseres familiestatistikk helt og holdent på familiefilen i BESYS. Men den marginale familiefordelingen kan ikke leses ut direkte fra denne pga. problemet med samboere, da det i BESYS kun er registrert samboerpar med felles barn[2]. Det er derfor et sprik mellom registerbeskrivelsen og virkeligheten, og dermed også i forhold til det brukerne venter å få tall for. En undersøkelsesbasert familiestatistikk kan bygges på den nye og mer relevante inndelingen av familietype, og vil dermed kunne bidra til å rette opp problemer i dagens familiestatistikk.

3.3 Kvalitet

Hverken familie- eller husholdningsfordeling er i dag kjent i personstatistikken. Hver enkelt familie- eller husholdningsundersøkelse må derfor estimere familie- eller husholdningsfordeling, uansett om dette skjer på en eksplisitt måte eller ikke. Ved å etablere en egen statistikk over denne fordelingen, som kan benyttes i alle undersøkelser, kan man konsentrere ressursbruken, og dette gir oss mulighet til å gå lengre i metodevalg, datautnyttelse og, ikke minst, kvalitetsevaluering/-vedlikehold. De enkelte undersøkelser behøver ikke lenger å lage sine egne fordelinger, men kan bygge på et felles grunnlag.

3.4 Bedre grunnlag for estimering og behandling av enhetsfrafall

Når familie- og husholdningsfordelingen er gitt på forhånd, har man et bedre grunnlag for oppblåsing enn om man måtte foreta estimering ut fra hver enkelt undersøkelse. Forbedret datautnyttelse vil også bety redusert usikkerhet i estimator.

[2]Samboere uten felles barn registreres i dag som enpersonfamilier eller mor/far med barn. Dette betyr at vi ikke kan oppgi tall for denne gruppen, og at tallene for enpersonfamilier, og mor/far med barn blir for høye.

I alle utvalgsundersøkelser finnes enhetsfrafall både i registeret og utvalget. Det første vil resultere i underdekning, som kun er av betydning i enkelte undersøkelser rettet mot spesielle delpopulasjoner. Enhetsfrafall i utvalget kan derimot innføre skjevhet i en trekking som ellers er representativ. En enkel måte å behandle enhetsfrafall på, er å dele nettoutvalget og populasjonen i mindre grupper, og deretter blåse opp delutvalg mot tilsvarende delpopulasjon under antagelsen om at enhetsfrafallet er tilfeldig fordelt innen hver delpopulasjon. Familie- og husholdningsfordelingen muliggjør slik deling. Metoden vil fullt ut justere skjevhet forårsaket av enhetsfrafall, dersom enhetsfrafallet er tilfeldig fordelt innen hver kombinasjon av familie- og husholdningstype, dvs. etter standard inndeling og størrelse som tidligere definert. Faktisk vil den nesten alltid bidra til å redusere denne type skjevhet, selv når enhetsfrafallet ikke er helt tilfeldig innen hver celle i familie-/husholdningstabellen.

3.5 Ressurssparing og mulighet til videre utvikling

Etablering av statistikk over familie- og husholdningsfordelingen vil lette analysebyrden i alle[3] undersøkelsesbaserte familie- og husholdningsstatistikker. De enkelte undersøkelsene behøver ikke lenger å lage egne familie-/husholdningsfordelinger, men kan bygges på et felles grunnlag. Dermed sparer man ressurser både på estimering og harmonisering mellom de ulike undersøkelsene. Naturligvis kan innføringen av en slik statistikk medføre produksjonsendringer i utvalgsbasert personstatistikk, akkurat som ethvert annet standardiserende tiltak. Men slike omlegginger vil også være gunstige med tanke på framtidens muligheter, som f.eks. et eget boligregister, fordi de mulighetene som innføres ved et slikt boligregister vil være dekket i den etablerte statistikken over familie- og husholdningsfordeling.

4 Fremgangsmåte og videre arbeid

Statistikken skal baseres på løpende utvalgsdata. Estimeringsmetoden skal inneholde både imputering og oppblåsing. Den første justerer for skjevhet forårsaket av enhetsfrafall, mens den andre tar seg av skjevhet pga. tilfeldighet i trekking. I vedlegget er det forklart flere tekniske detaljer ved metoden. Her skal vi skissere hovedtrekk i det videre arbeidet.

4.1 Datagrunnlag

Inntekts- og formuesundersøkelsen (IF) er per dato det mest omfattende datamaterialet med husholdningsopplysninger i SSB. Den benytter et sammensatt utvalg[4] av Panelutvalget til IF (IFP), Forbruksundersøkelsen (FU), Levekårsundersøkelsen (LKU), Boforholdsundersøkelsen (BFU), Helseundersøkelsen (HU) og et tilleggsutvalg av personlig næringsdrivende (IFN). Utvalgsstørrelsen varierer noe avhengig av hvilke undersøkelser som er tilgjengelige i det aktuelle året. Men i de senere årene har bruttoutvalget holdt seg stabilt over ti tusen husholdninger. Også frafall varierer i deler av IF. LKU er best med ca. 20% frafall, IFN verst med ca. 40%, og hele IF sett under ett $25 - 30\%$.

Siden omleggingen i 1996 har AKU begynt å samle inn opplysninger om husholdningene i utvalget en gang i året (2. kvartal). Bruttoutvalget inneholder rundt tolv tusen husholdninger og familier. Halvparten av dette overlapper fra et år til det neste. Etter omleggingen benytter AKU systematisk trekking innen hvert fylket. Men materialet er til nå ikke kodet, så det har ikke vært mulig å undersøke kvaliteten på dataene. AKU dekker heller ikke personer over 74 år, slik at en del av IF uansett må brukes som tilleggsutvalg.

[3] Det fantes 12 faste utvalgsundersøkelser på personstatistikk ifølge Aarberge utvalget (Samordning av levekårsrelatert statistikk, 1994, Intern Notat). I tillegg kommer enkeltstående prosjekter på oppdragsbasis.

[4] Noen av de her nevnte undersøkelsene kan falle bort i framtiden.

Videre arbeid Man bør undersøke de relevante dataene i AKU før man avgjør sammensetning av datagrunnlaget. Det endelige materialet skal inneholde koder for familie- og husholdningstype og familie/husholdningsstørrelse for hver nøkkelperson som har svart, eller årsaken til frafall.

4.2 Registergrunnlag

Flere registre i SSB, som f.eks. BESYS, gir oss verdifull tilleggsinformasjon både for behandling av frafall og estimering av familie- og husholdningsfordelingen. Forbedring i registergrunnlaget har derfor stor betydning i den nåværende sammenhengen. Spesielt finnes det flere muligheter når det gjelder husholdninger. Et boligregister, dersom det blir etablert, vil gi oss en oversikt over bohusholdninger, noe som uten tvil vil være nyttig i estimeringen av husholdningsfordelingen. Uansett skal det i FoB2000 gjennomføres en totaltelling av boliger. Utfordringen ligger i å legge forholdet til rette slik at man kan dra nytte av denne tellingen også etter år 2000[5].

Videre arbeid Bedre utnyttelse av registergrunnlaget krever samordnet innsats. Nyttig tilleggsinformasjon skal også kodes inn i det endelige datamaterialet (det som skal brukes for å etablere den marginale familie/husholdningsfordelingen).

4.3 Modellering av enhetsfrafall

Justering for skjevhet forårsaket av enhetsfrafall forutsetter fornuftige antagelser om mekanismene i frafallet, dvs. en god modell for enhetsfrafall.

I dag utgjør enhetsfrafall i AKU ca. 10 prosent av bruttoutvalget, dvs. rundt 2 400 personer hvert kvartal, uten at man regner med husholdningsdelen. En enkel gjennomgang viser at frafallet er lite avhengig både av kjønn og område (fylke). Variasjoner mht. alder er også små, bortsett fra en merkbar lav svarandel blant de eldste (67-74). Det viser seg at frafallet[6] inneholder ca. 700 "nekting", 600 "ikke å treffe", 500 "ikke å lokalisere", 300 "antageligvis ikke private husholdninger" (f.eks. langtidssyke som likevel hører til AKU populasjonen), og resten fordeler seg på forskjellige andre grunner. Av erfaring er det f.eks. ikke vanskelig å forestille seg at familie- eller husholdningsstørrelsen er forholdsvis liten blant de ca. 600 "ikke å treffe", kanskje til dels også de 500 "ikke å lokalisere", samtidig som andre årsaker kanskje er mer avgjørende for "nekting". Det kan derfor være hensiktsmessig å bruke forskjellige modeller på forskjellige deler av frafallet. Det samme gjelder også data fra IF. Dagens rutine tilsier at man erstatter husholdningsopplysninger med lignende familieopplysninger fra BESYS i tilfellet frafall. Metoden er utilstrekkelig[7], spesielt mht. vårt formål da den lett kan innføre skjevhet i husholdningsfordelingen.

Videre arbeid Modeller for forskjellige typer frafall må undersøkes i samarbeid med fageksperter med erfaringer fra familie- og husholdningsdata. Spesielt bør det legges vekt på samspillet mellom frafall og familie- og husholdningstype/størrelse.

---

[5] Metodisk sett er det ikke nødvendig med en totaltelling for å få god nok husholdningsstatistikk i FoB2000. Men når man første har valgt å gjøre det, bør man forsøke å få mest mulig ut av det.

[6] Med forbehold om klassifisering av frafallsgrunn som kan endre seg med hensyn til forskrifter fra datatilsynet.

[7] Belsby og Bjørnstad (1997). Modeling and estimation methods for household size in the presence of nonresponse — Applied to the Norwegian Consumer Expenditure Survey. Discussion papers No. 206, Statistics Norway.

Epland (1999). Longitudinal non-response: Evidence from the Norwegian Income Panel. Documents 99/6. Statistics Norway.

4.4 Oppblåsing

Dagens oppblåsingsmetode i IF heter kalibrering, og denne metoden tar hensyn både til tilleggsinformasjon om inntekt/formue og befolkningsstatistikk (med person som enhet). Av flere grunner er den lite hensiktsmessig for den nye statistikken som omtales. For det første har man problemstillingen knyttet til frafallet. Videre er det f.eks. vanskelig å forstå at husholdningsfordelingen skal ha mye å gjøre med inntektsfordelingen i samfunnet, samtidig som det er lett å forestille seg at registerinformasjon om familier har en bedre sammenheng med familie- og husholdningsfordelingen enn registerinformasjon om personer. I stedet skal man utarbeide et opplegg der oppblåsing (mot den norske befolkningen) er basert på imputert bruttoutvalg. Det skal justere for skjevhet pga. tilfeldighet i trekking. Her må man kunne benytte tilleggsinformasjon fra forskjellige registre. Teknikken er etterstratifisering, raking og kalibrering.

Videre arbeid Den meste aktuelle problemstillingen her handler om å finne ut hvor omfattende den nye statistikken skal bli. Det vil vise seg om en krysstabell på det mest detaljerte inndelingsnivået er for stor ut fra det tilgjengelige datagrunnlaget. Det antas imidlertid at etablering av de to marginale fordelingene er fullt innen rekkevidde. Alternativt kan man derfor ta sikte på de to marginale fordelingene, i tillegg til en mindre krysstabell (med noen sammenslåtte marginaler).

5 Oppsummering

Ovenfor har vi skissert et opplegg til en statistikk over familie- og husholdningsfordelingen i den norske befolkningen. Statistikken er nyskapende i internasjonal sammenheng, og vil kunne gi oss et bedre felles grunnlag for utvalgsundersøkelsesbasert personstatistikk. Som begrunnelse for en slik fordeling kan man bl.a. nevne sikring av konsistens, aktualisering av familiestatistikken, forbedring av kvalitet, fleksibilitet i oppblåsing og behandling av enhetsfrafall i de enkelte undersøkelser, og sist, men ikke minst, ressurssparing både i analyse og harmonisering mellom forskjellige statistikker. Statistikken skal baseres på løpende utvalgsdata først og fremst fra Inntekts- og formuesundersøkelsen og Arbeidskraftundersøkelsen. Det endelige produksjonsopplegget skal inneholde to deler, nemlig imputering og oppblåsing. Den første behandler enhetsfrafallet i dataene, og skal resultere i et utfylt bruttoutvalg. Den andre blåser så dette bruttoutvalget opp mot populasjonen ved å justere for skjevhet i trekking. Det planlegges å gjennomføre det videre arbeidet i et kommende prosjekt.

A Familie- og husholdningsstatistikk basert på undersøkelse med frafall

A.1 Estimering av (familie- og) husholdningsfordeling med frafall i utvalg

1 Tre typer enheter er i bruk for SSBs personstatistikk, nemlig (i) person, (ii) familie, og (iii) husholdning. Opplysninger om de to første finner man i SSBs person- og familieregistre. Det mangler registerinformasjon om husholdning i den norske befolkning.

2. Med *(familie- og) husholdningsfordeling* mener man den frekvenstabellen, der alle personer i den aktuelle populasjonen er kryssklassifisert etter familie- og husholdningsgruppering. For å lette notasjonen, bruker vi her kun størrelsen til familier og husholdninger. Anta at alle personer hører til i alt H familiegrupper etter størrelsen, nemlig 1-personsfamilie, 2-personsfamilie, ..., og famlie med H eller flere personer, og lignende for husholdning. Da er husholdningsfordeling i populasjonen gitt som

Husholdningsfordeling i den aktuelle populasjonen (Antall personer N)

Antall	*Husholdning*			
Familie	1	2	$\cdots$	$\geq H$
1	N_{11}	N_{12}	$\cdots$	N_{1H}
2	N_{21}	N_{22}	$\cdots$	N_{2H}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\geq H$	N_{H1}	N_{H2}	$\cdots$	N_{HH}

$\overset{N_{ij}=p_{ij}N}{\Longleftrightarrow}$

Andel	*Husholdning*			
Familie	1	2	$\cdots$	$\geq H$
1	p_{11}	p_{12}	$\cdots$	p_{1H}
2	p_{21}	p_{22}	$\cdots$	p_{2H}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\geq H$	p_{H1}	p_{H2}	$\cdots$	p_{HH}

3 Anta at et utvalg består av n personer med underliggende kryssklassifisering $\{s_{ij}\}$, dvs. uten frafall, s.a. $\sum_{i,j=1}^{H} s_{ij} = n$. Anta i tillegg en frafallsfordeling $\{r_{ij}\}$, der r_{ij} er svarsannsynlighet gitt at en person er trukket fra familiegruppe i og husholdningsgruppe j.

Utvalg uten frafall

Antall	*Husholdning*			
Familie	1	2	$\cdots$	$\geq H$
1	s_{11}	s_{12}	$\cdots$	s_{1H}
2	s_{21}	s_{22}	$\cdots$	s_{2H}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\geq H$	s_{H1}	s_{H2}	$\cdots$	s_{HH}

Frafallsfordeling

Sannsynlighet	*Husholdning*			
Familie	1	2	$\cdots$	$\geq H$
1	r_{11}	r_{12}	$\cdots$	r_{1H}
2	r_{21}	r_{22}	$\cdots$	r_{2H}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\geq H$	r_{H1}	r_{H2}	$\cdots$	r_{HH}

Kommentar 1 *Utvalget her består kun av personer som direkte er trukket. Det inneholder f.eks. ikke andre personer i den samme husholdningen som er tatt med i en husholdningsundersøkelse.*

4 Anta at det utvalget med n personer inneholder n_r som svarte og m frafall, der $n_r + m = n$, $n_r = \sum_{i,j=1}^{H} n_{ij}$, $m = \sum_{i=1}^{H} m_i$, og $\sum_j n_{ij} + m_i = \sum_j s_{ij}$, dvs.

Utvalget med frafall

Antall	*Husholdning*				
Familie	1	2	$\cdots$	$\geq H$	*Frafall*
1	n_{11}	n_{12}	$\cdots$	n_{1H}	m_1
2	n_{21}	n_{22}	$\cdots$	n_{2H}	m_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\geq H$	n_{H1}	n_{H2}	$\cdots$	n_{HH}	m_H

som har følgende betinget fordeling gitt $\{s_{ij}\}$:

Betinget fordeling til utvalget med frafall

Antall/Frekvens	*Husholdning*				
Familie	1	2	$\cdots$	$\geq H$	*Frafall*
1	$s_{11}r_{11}$	$s_{12}r_{12}$	$\cdots$	$s_{1H}r_{1H}$	$\sum_{j=1}^{H} s_{1j}(1-r_{1j})$
2	$s_{21}r_{21}$	$s_{22}r_{22}$	$\cdots$	$s_{2H}r_{2H}$	$\sum_{j=1}^{H} s_{2j}(1-r_{2j})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\geq H$	$s_{H1}r_{H1}$	$s_{H2}r_{H2}$	$\cdots$	$s_{HH}r_{HH}$	$\sum_{j=1}^{H} s_{Hj}(1-r_{Hj})$

5 Estimering av husholdningsfordelingen i populasjonen består av to trinn, nemlig (1) komplementering av utvalget $\{s_{ij}\}$, og (2) estimering av husholdningsfordelingen i populasjonen basert på det komplementerte utvalget. Trinn (1) heter også *imputering*, og handler om rekonstruering av utvalget pga. frafall. I trinn (2) tar man ikke lenger hensyn til frafallsproblemstilling, da det handler om justering av den imputerte husholdningsfordelingen i utvalget pga. tilfeldighet i trekking.

6 En frafallsmodell sies å være *ignorerbar* dersom svarsannsynligheten kun avhenger av familiestørrelsen, som er kjent selv blant frafall.

Eksempel 1 *Anta* $r_{ij} = r_i$, *s.a.* $\hat{r}_{ij} = \hat{r}_i = (\sum_j n_{ij})/(m_i + \sum_j n_{ij})$ *og* $\hat{s}_{ij}/\hat{s}_{ik} = n_{ij}/n_{ik}$.

7 En frafallsmodell er *ikke-ignorerbar* dersom r_{ij} avhenger av husholdningsstørrelsen, som er ukjent blant frafall, til tross for kjennskap til familiemerke. Til imputering under en slik ikke-ignorerbar frafallsmodell, kan man bruke følgende iterativt algoritme, gitt at modellen er identifiserbar:

Frafall		*Frafall*					*Imputert (komplett) tabell*			
m_1		m_{11}^*	m_{12}^*	$\cdots$	m_{1H}^*		$n_{11}+m_{11}^*$	$n_{12}+m_{12}^*$	$\cdots$	$n_{1H}+m_{1H}^*$
• m_2	$\Rightarrow$	m_{21}^*	m_{22}^*	$\cdots$	m_{2H}^*	$\Rightarrow$	$n_{21}+m_{21}^*$	$n_{22}+m_{22}^*$	$\cdots$	$n_{2H}+m_{2H}^*$
$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$	$\ddots$	$\vdots$
m_H		m_{H1}^*	m_{H2}^*	$\cdots$	m_{HH}^*		$n_{H1}+m_{H1}^*$	$n_{H2}+m_{H2}^*$	$\cdots$	$n_{HH}+m_{HH}^*$
		der $m_{ij}^* = m_i \frac{\hat{s}_{ij}(1-\hat{r}_{ij})}{\sum_{j=1}^{H}\hat{s}_{ij}(1-\hat{r}_{ij})}$					s.a. $\hat{s}_{ij} = n_{ij} + m_{ij}^*$.			

• oppdater $\hat{r}_{ij}$ basert på $\{n_{ij}\}$ og $\{m_{ij}^*\}$ under frafallsmodellen som om man hadde observert m_{ij}^*; og iterer hvis ikke endringene i parametrene er små nok.

Eksempel 2 *Anta* $r_{ij} = r_j$, *dvs. frafall kun avhenger av husholdningsstørrelsen, s.a. man oppdaterer* $\hat{r}_j$ *med* $(\sum_i n_i)/\sum_i (n_{ij} + m_{ij}^*)$. *Belsby og Bjørnstad (1997)*[8] *brukte imidlertig en annen metode under akkurat denne frafallsmodellen som gav de samme resultatene: dersom man erstatter* m_{ij}^* *med* $n_{ij}(1-\hat{r}_j)/\hat{r}_j$ *i ligning* $\sum_j m_{ij}^* = m_i$, *får man*

$$1/\hat{\mathbf{r}} = \mathbf{n}^{-1}\mathbf{s}, \tag{1}$$

der $\mathbf{r} = (r_1, ..., r_H)^T$, *og* $\mathbf{s} = (\sum_j s_{1j}, ..., \sum_j s_{Hj})^T$, *og* $\mathbf{n}$ *er* $k \times k$*-matrise med element* n_{ij}.

Eksempel 3 *Man kan sette opp modeller der* r_{ij} *varierer også blant personer fra den samme husholdningsgruppen, f.eks. pga. at mange en-personsfamilie personer som lever i en fire-persons husholdning bor kollektivt, i motsetning til fire-personsfamilie personer som faktisk lever i en fire-persons husholdning.*

[8] *Belsby og Bjørnstad (1997).* Modeling and estimation methods for household size in the presence of nonresponse — Applied to the Norwegian Consumer Expenditure Survey. *Discussion Papers No. 206, Statistisk sentralbyrå.*

A.2 Frafallseffekten på familie- og husholdningsstatistikk

1 Forskjellige estimatorer for familie- og husholdningsfordelingen i populasjonen, dvs. $\{N_{ij}\}$, svarer til forskjellige frafallsmodeller, og kan dermed vurderes på denne bakgrunnen.

Imputert familie- og husholdningsfordeling i utvalget

Antall	*Husholdning*								
	Observert					*Frafall (imputert)*			
Familie	1	2	$\cdots$	$\geq H$	(*Frafall*)	1	2	$\cdots$	$\geq H$
1	n_{11}	n_{12}	$\cdots$	n_{1H}	(m_1)	m^*_{11}	m^*_{12}	$\cdots$	m^*_{1H}
2	n_{21}	n_{22}	$\cdots$	n_{2H}	(m_2)	m^*_{21}	m^*_{22}	$\cdots$	m^*_{2H}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\geq H$	n_{H1}	n_{H2}	$\cdots$	n_{HH}	(m_H)	m^*_{H1}	m^*_{H2}	$\cdots$	m^*_{HH}

Eksempel 4 *Enkel oppblåsing, $\tilde{N}_{ij} = N(n_{ij}/n_r)$, svarer til frafallsmodell $r_{ij} \equiv r$, der $m^*_{ij} = m(n_{ij}/n_r)$. Siden $\sum_j m^*_{ij} = m_i$ bare hvis $m_i/m = (\sum_j n_{ij})/n_r$, har denne modellen vanligvis dårlig tilpassning til et gitt utvalg.*

2 Generelt er den *imputerte etterstratifiserte estimatoren* for $\{N_{ij}\}$ gitt som

$$\hat{N}_{ij} = N_i(\hat{s}_{ij}/s_i) = N_i \cdot (n_{ij} + m^*_{ij})/(\sum_j n_{ij} + m_i). \tag{2}$$

Kommentar 2 *Estimator* (2) *er annerledes under forskjellige frafallsmodeller, siden de gir forskjellige m^*_{ij}. God tilpassning til data kan ofte virke som nødvendig, ikke suffisient, kriterium til modellvalg.*

Eksempel 5 *Den vanlige etterstratifiserte estimatoren svarer til ignorerbar frafallsmodell $r_{ij} = r_i$, der $m^*_{ij} = m_i(n_{ij}/\sum_j n_{ij})$, slik at, $\hat{N}_{ij} = N_i(n_{ij}/\sum_j n_{ij})$, der N_i er kjent antall personer i familiegruppe i i populasjonen. Denne modellen har alltid perfekt tilpassning til data gitt hvilket som helst utvalg.*

Eksempel 6 *Under ikke-ignorerbar $r_{ij} = r_j$ har man $m^*_{ij} = n_{ij}/\hat{r}_j - n_{ij}$, der $1/\hat{r}_j$ er den $j$te komponent i $\mathbf{n}^{-1}\mathbf{s}$ som tidligere forklart* (1)*, som også gir perfekt tilpassning til data (Belsby og Bjørnstad, 1997).*

3 Delpopulasjon av de N_{ij} personer består av N_{ij}/i familier og N_{ij}/j husholdninger for $1 \leq i, j < H$. For familier med H eller flere personer kjenner man til både antall personer, nemlig N_H og antall familier, betegnet med F_H, der $N_H/F_H > H$. Belsby og Bjørnstad (1997) brukte derfor $\hat{N}_{iH}(F_H/N_H)$ til å estimere antall husholdninger blant de N_{iH} personer. På samme måte estimeres antall familier blant de N_{Hj} personer med $\hat{N}_{Hj}(F_H/N_H)$.

Kommentar 3 *Man kan eventuelt vurdere å slå sammen alle F_H familier for familiestatistikk.*

4 Betegn med y interessevariabelen for undersøkelsen. Frafallsfordeling $\{r_{ij}\}$ forutsetter at frafall innen hver kombinasjon av familie- og husholdningsgrupper er ignorerbar, dvs. interessevariabelen der har den samme fordelingen blant de som svarer som blant frafall. Det er derfor mulig å lage forventningsrett estimator for gjennomsnitt i denne delpopulasjonen basert på det tilsvarende netto-delutvalget, uansett om det er person-, familie-, eller husholdningsstatistikk. La $\hat{\bar{y}}_{ij}$ betegne en slik estimator, da er den imputerte etterstratifiserte estimatoren for $Y_{ij} = \sum_{k=1}^{N_{ij}} y_k$ gitt som

$$\hat{Y}_{ij} = \hat{U}_{ij}\hat{\bar{y}}_{ij}, \tag{3}$$

der $\hat{U}_{ij}$ er estimator for antall enheter, basert på $\hat{N}_{ij}$ ved (2), til den aktuelle statistikken.

Eksempel 7 *I SSBs familie- eller husholdningsundersøkelser tar man som regel hele familien, eller husholdningen, som er knyttet til den person direkte trukket. Da er $\hat{\bar{y}}_{ij}$ basert på n_{ij} familier, eller husholdninger.*

5 Anta at alle personer fra den samme familiegruppen er trukket med lik sannsynlighet. Da er $\hat{N}_{ij}$ gitt ved (2) konsistent for N_{ij}, dermed også $\hat{Y}_{ij}$ gitt ved (3) for Y_{ij}, under den tilsvarende frafallsmodellen.

Eksempel 8 *Anta $r_{ij} = r_i$. Da er $\hat{N}_{ij}$ med tilsvarende m^*_{ij} forventingsrett for N_{ij}, og dermed også $\hat{Y}_{ij}$ for Y_{ij} dersom $\hat{\bar{y}}_{ij}$ er uavhengig av $\hat{s}_{ij}$.*

Eksempel 9 *Anta $r_{ij} = r_j$. Da er $\hat{s}_{ij} = n_{ij}/\hat{r}_j$ (beskrevet tidligere) den maksimum likelihood estimatoren (m.l.e.) for s_{ij} (Belsby og Bjørnstad, 1997), s.a. $\hat{N}_{ij}$ er konsist for N_{ij} som følge av generelle egenskaper til m.l.e., og dermed også $\hat{Y}_{ij}$ for Y_{ij}.*

Eksempel 10 *Anta $r_{ij} = r_j$. Uten å ta hensyn til ikke-ignorerbarhet i frafall, medfører den vanlige etterstratifiserte estimatoren skjevhet i estimator for Y_{ij}, som kan estimeres med forskjellen mellom $\hat{Y}_{ij}$ basert på imputering under henholdsvis $r_{ij} = r_i$ og $r_{ij} = r_j$.*

6 Følgende Bootstrap simulering[9] kan brukes til å metodeevaluering:

Observert utvalg

n_{11}	n_{12}	$\cdots$	n_{1H}	m_1
n_{21}	n_{22}	$\cdots$	n_{2H}	m_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n_{H1}	n_{H2}	$\cdots$	n_{HH}	m_H

$\Rightarrow$ multinormisk $(n^\star_{i1}, ..., n^\star_{iH}, m^\star_i)$ innen hver rad med sannsynlighet $(n_{i1}, ..., n_{iH}, m_i)/s_i$ der $\sum_j n_{ij} + m_i = \sum_j n^\star_{ij} + m^\star_i = s_i$ $\Rightarrow$

Re-sampel

$n^\star_{11}$	$n^\star_{12}$	$\cdots$	$n^\star_{1H}$	$m^\star_1$
$n^\star_{21}$	$n^\star_{22}$	$\cdots$	$n^\star_{2H}$	$m^\star_2$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$n^\star_{H1}$	$n^\star_{H2}$	$\cdots$	$n^\star_{HH}$	$m^\star_H$

$\Rightarrow$ imputering basert på re-sampel som om det var observert $\Rightarrow$

Imputert utvalg

$n^\star_{11} + m^*_{11}$	$n^\star_{12} + m^*_{12}$	$\cdots$	$n^\star_{1H} + m^*_{1H}$
$n^\star_{21} + m^*_{21}$	$n^\star_{22} + m^*_{22}$	$\cdots$	$n^\star_{2H} + m^*_{2H}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$n^\star_{H1} + m^*_{H1}$	$n^\star_{H2} + m^*_{H2}$	$\cdots$	$n^\star_{HH} + m^*_{HH}$

$\Rightarrow$ $\hat{N}^\star_{ij}$ ved etterstratifisert estimator (2)

• gjenta prosedyren, si, B ganger, og estimer henholdsvis forventningen og variansen til estimator (2) med gjennomsnitt og utvalgsvarians for disse B re-samplete $N^\star_{ij}$.

Kommentar 4 *Ikke-parametrisk re-sampling her er gjort for gitte marginaler, dvs. $\{s_i\}$, siden etterstratifisering* (2) *benytter den kjente marginale fordelingen til familiestørrelse i populasjonen,.*

7 Dersom $\hat{\bar{y}}_{ij}$ er uavhengig fordelt med $\hat{N}^\star_{ij}$ betinget på $\{n_{ij}, m_i\}$, har man, for estimator (3),

$$Var(\hat{Y}^*_{ij}) = Var(\hat{U}_{ij})Var(\hat{\bar{y}}_{ij}) + Var(\hat{U}_{ij})E^2[\hat{\bar{y}}_{ij}] + E^2[\hat{U}_{ij}]Var(\hat{\bar{y}}_{ij}). \quad (4)$$

Ellers må man utvide simuleringsopplegg slik at det også omfatter $\hat{\bar{y}}_{ij}$.

[9] Efron og Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

A.3 Tilfelle: Husholdningsstørrelse basert på Forbruksundersøkelsen 1992 (FoU92)

1 Følgende observert utvalg er hentet fra FoU92:

Familie- og husholdningsfordeling i utvalget FoU92

Antall	*Husholdning (By)*						*Husholdning (Land)*					
Familie	1	2	3	4	≥ 5	*Frafall (By)*	1	2	3	4	≥ 5	*Frafall (Land)*
1	28	24	7	2	0	78	55	24	13	7	2	75
2	6	70	12	3	0	84	3	107	25	1	3	76
3	4	8	57	11	3	40	6	17	74	29	3	51
4	0	3	15	80	5	43	2	10	22	151	12	80
≥ 5	0	1	0	6	66	28	1	3	4	11	115	32

2 Inspeksjon av data viser at det er fornuftig å behandle frafall på by og land hver for seg (Belsby og Bjørnstad, 1997). Under den enkle ikke-ignorbare frafallsmodell $r_{ij} = r_j$ gitt bosted, har man funnet følgende estimerte svarsannsynligheter:

Estimerte svarsannsynligheter i FoU92 (Belsby og Bjørnstad, 1997)

Sannsynlighet	*Husholdning*				
Bosted	1	2	3	4	≥ 5
By	0.352	0.509	0.748	0.701	0.729
Land	0.508	0.624	0.769	0.706	0.831

Kommentar 5 *Største forskjeller skjer i de to halene på det husholdningsspektrum.*

3 Belsby og Bjørnstad brukte den imputerte etterstratifiserte estimatoren (2), etter å ha slått sammen imputert utvalg på by og land, og sammenlignet estimerte marginale husholdningsstørrelser med tilsvarende estimater henh. under frafallsmodell $r_{ij} = r_i$ og $r_{ij} = r$:

Estimerte marginale husholdningsstørrelser for FoU92 (i 1000)

	Husholdning				
Metode	1	2	3	4	≥ 5
Imputert etterstratifisert estimator ($r_{ij} = r_j$)	596.6	523.6	250.0	268.9	126.2
Vanlig etterstratifisert estimator ($r_{ij} = r_i$)	486.0	507.8	286.2	270.6	131.3
Enkel oppblåsing ($r_{ij} = r$)	390.5	496.5	283.9	279.9	148.0

Kommentar 6 *Den estimerte marginale husholdningsfordelingen er skjøvet lengre og lengre mot høyre fra under frafallsmodell $r_{ij} = r_j$ til $r_{ij} = r_i$, og så til $r_{ij} = r$.*

4 Boostrap simulering (med 500 re-sampel) gav oss følgende resultater:

Bootstrap for estimering av marginale husholdningsstørrelser i FoU92 (i 1000)

		Husholdning				
Metode		1	2	3	4	≥ 5
Imputert etterstratifisert estimator	Forventning	596	523	250	270	127
	Standardavvik	40	12	5	2	1
Vanlig etterstratifisert estimator	Forventning	484	507	286	271	131
	Standardavvik	36	11	5	2	1

Kommentar 7 *Simulering her følger estimeringsopplegget brukt i Belsby og Bjørnstad, s.a. estimator* (2) *er anvendt etter å ha slått sammen de to imputerte delutvalgene på by og land. Antall re-sampel er blitt satt til 500 etter at man har observert at resultatene endrer seg lite fra 250 til 500 re-sampel.*

5 Hoved konklusjon: (i) utvalget i FoU92 er så pass stort at konsistens i begge estimatorer er gyldig under hver sin frafallsmodell, og (ii) det er liten forskjell mellom de to estimatorene mht. effisiens.

Kommentar 8 *Man har i dette tilfellet anvendt teknikken (beskrevet i dette notatet) på by og land, dvs. etter kjennemerke Bosted. Man skal ikke se bort fra at andre kjennemerker kan trekkes inn dersom de påvirker frafallet. Eksempel på slike kjennemerker er bl.a. alder (høyre frafallsandel bl. eldre), med/uten barn, og barnas alder, osv..*

De sist utgitte publikasjonene i serien Notater

99/34 E. Birkeland (red.): Forskjeller i levekår: Hefte 3: Bruk av velferdsordninger. 126s.

99/35 E. Birkeland (red.): Forskjeller i levekår: Hefte 4: Regionale forskjeller. 118s.

99/36 M. Stålnacke, J-A. Sigstad Lie og L. Solheim: En analyse av SSBs generelle utvalgsplan fra 1995 basert på næringsvise sysselsettingstall. 83s.

99/37 B.O. Lagerstrøm: Trivsels- og arbeidsmiljøundersøkelse blant intervjuere i Statistisk sentralbyrå. 155s.

99/38 K.J. Einarsen: Evalueringsrapport for pilotforsøket for FylkesKOSTRA-utdanning. 55s.

99/39 L. Rogstad: FoB2000: Adressesamsvar mellom folkeregister og adresseregister i GAB: – rapport fra Lysebu-seminar 8. og 9. desember 1998, – tiltaksplan for bedre adressesamsvar. 39s.

99/40 D. Roll-Hansen: Samordnet levekårsunder-søkelse 1998 – tverrsnittsundersøkelsen: Dokumentasjonsrapport. 102s.

99/41 R. Johannessen: Kommunale gebyrer knyttet til bolig. Januar 1999. 30s.

99/42 M. Stålnacke, A.G. Hustoft og L. Solheim: Vurdering av kvalitet i statistikk: En oversettelse av notater fra Eurostat om kvalitetsrapportering. 77s.

99/43 E. Engelien, K. Myklebust, J.A. Paulsen og L. Rogstad: FoB2000: Stedfesting av bedrifter – forprosjekt. 40s.

99/44 I. Hauge, C. Hendriks, Ø. Hokstad og A.G. Hustoft: Standard for begreper og kjennemerker knyttet til familie- og husholdningsstatistikken. 37s.

99/45 E. Rønning: Omnibusundersøkelsene 1998: Dokumentasjonsrapport. 123s.

99/46 C. Torp: Situasjonsuttak fra Bedrifts- og foretaksregisteret. 33s.

99/47 T.N. Evensen: Utlendingers konsum i Norge: En vurdering av eksporttallene for reisetrafikk i nasjonalregnskapet på bakgrunn av statistikk som belyser forbruket til utenlandske turister i Norge. 28s.

99/48 H. Hartvedt (red.): Definisjonskatalog for grunnskoleopplæring for barn og voksne. 14s.

1999/49 K. Bjønnes og J. Johansen: FD - Trygd: Dokumentasjonsrapport. Attføringspenger, 1992-1997. 126s.

1999/50 E. Høydahl: FoB2000: Rapport fra seminar 4. juni 1999 om kommuneprodukter fra Folke- og boligtellingen 2000. 32s.

1999/51 P.E. Tønjum: Teknisk dokumentasjon av beregningsopplegget for kvartalsvis nasjonalregnskap (KNR). 91s.

1999/52 F. Gundersen: Statistikk over etterforskede lovbrudd: Dokumentasjon. 46s.

1999/53 N. Arnesen og Ø. Skullerud: Statistikk over emballasjeavfall: Beregningsresultater for 1997. 36s.

1999/54 Ø. Kleven: Bruk av kreftundersøkelsen PSA blant menn i alderen 50 til 65 år. 19s.

1999/55 P.M. Holt og L. Wiker: Inntekts- og formuesundersøkelsen for aksjeselskaper 1996: Dokumentasjon. 30s.

1999/56 B.O. Lagerstrøm: Små og mellomstore bedrifters vurdering av kostnader ved lover og regelverk: Hovedresultater. 129s.

1999/57 L.H. Thingstad: Regnskapsstatistikk for varehandel 1996: Dokumentasjon av produksjonsrutiner. 36s.

1999/58 P.E. Tønjum: Teknisk dokumentasjon av det årlige realregnskapets FAME-databaser og rutiner. 53s.

1999/59 E.J. Fløttum: Konsumgrupperinger i offisiell statistikk. 103s.

1999/60 R. Johannessen: Kvalitetssikring av korttidsstatistikk. 26s.