Statistics Norway
Statistisk sentralbyrå

Li-Chun Zhang og Anne Vedø

Omlegging av utvalgsplan for AKU

Seksjon for statistiske metoder og standarder

Innhold

1 Sammendrag

Dagens utvalgsplanen for AKU var tatt i bruk gjennom omleggingen i løpet av 1996 og 1997. Den kan karakteriseres som følgende:

- stratifisering etter alle landets fylker;
- trekkenheter er familier ifølge det sentrale folkeregisteret (DSF);
- DSF-familiene innen hvert fylke trekkes systematisk med antatt tilfeldig sortering.

Våre vurderinger om denne plan er kort sagt som følgende:

- Stratifiseringen etter fylke er nødvendig for å kunne publisere på fylkesnivået.
- Bruket av DSF-familien som trekkenheten kan trolig begrunnes mht. kostnaden i datainnsamling. Kun i få lander i verden, der iblant Sverige, er AKU basert på et rent personutvalg. Metodisk sett øker familieutvalg usikkerheten i estimering av f.eks. sysselsetting i forholdet til personutvalg.
- Det kan vises at systematisk trekking fører til unødvendig store variasjoner i utvalgsvariansen, som ikke kan reduseres vha. etterstratifisering eller kalibrering i estimeringen.

Vi forslår derfor å bruke istedet et tilfeldig DSF-familieutvalg *stratifisert* etter

1. fylke,
2. familiestørrelse: 1-AKU-persons, 2-AKU-persons og andre familier,
3. alder til den eldste AKU-personen i familien: 16-29 år, 30-54 år, 55-67 år og 67-74 år.

Dagens antall DSF-familier fra hvert fylke videreføres. Deretter fastsettes stratumsstørreslene i utvalget proporsjonelt med tilsvarende stratumsstørrelsene i populasjonen innen hvert fylke. Hoved egenskaper ved denne utvalgsplan kan sammenfattes som følgende:

- Den er minst like effisient som et rent personutvalg tilfeldig trukket innen hvert fylke. Dette gjelder både estimatorene for nivå og endring i andelen sysselsatt og ledig.
- Den kan gi betydelig redusering i utvalgsvariansen ved dagens metode i gjennomsnitt over tid. F.eks. kan variansreduksjonen være over 50% når det gjelder estimatorene for kvartalsvis endring i andelen sysselsatt og ledig.
- Ved å bruke tilfeldig trekking fjerner vi den unødvendige variasjonen i utvalgsvariansen ved dagens metode av systematisk trekking.

Et tilleggs moment i utvalgsplanen for AKU dreier seg om det såkalte husholdningsspørsmål, som stilles i det 2. kvartal i året. Det går ut på å kartlegge forholdet hvert medlem i husholdningen har til en bestemt referansperson. Denne er for tiden den person som har likt person- og familienummer i DSF. Vi forslår at i framtiden skal referanspersonen velges tilfeldig innen DSF-familien.

2 Systematisk trekking

2.1 Generelt

Systematisk trekking gjennomføres på følgende måte:

1. Sorter populasjonen etter noen bestemte kjennemerker.
2. Velg *trekkelengden*, betegnet med L, slik at den ønskete utvalgsstørrelsen er lik det største heltall som er mindre enn
$$\frac{\text{Antall trekkenheter i populasjonen}}{\text{Trekkelengde}}.$$
3. Velg et tilfeldig heltall som er mindre eller lik trekkelengden, betegnet med r og $1 \leq r \leq L$.
4. Utvalget består da av enhet nr. r, $r + L$, $r + 2L$, osv.

I praksisen er antallet trekkenheter i populasjonen ikke alltid delebar med trekkelengden. Utvalget som er trukket på denne måten har enten akkurat den ønskete utvalgsstørrelsen eller 1 flere.

Systematisk trekking brukes i de alle fleste utvalgsundersøkelser i SSB, inkl. AKU, pga. tre ting:

- Historisk sett slo systematisk trekking gjennom i en tid da rent tilfeldig trekking, som må gjøres vha. datamaskin, ikke var så allment tilgjengelig som det er idag.
- Som regel bruker man en partiell sortering av populasjonen i person-/husholdningsundersøkelser. Sorteringen betraktes som partiell siden det siste kjennemerke, f.eks. person- eller familienummer, kan i alle praktiske tilfeller antas å være uavhengig av interessevariabelen. Dette er også grunnen til at man ofte betrakter systematisk trekking som tilnærmet stratifisert tilfeldig trekking mht. de kjennemerker i sorteringen som antas å være korrelerte med interessevariabelen.
- Systematisk trekking kan enkelt justeres slik at trekkesannsynlighetene blir proporsjonelle med en tilleggs størrelse, noe som kan være komplisert å oppnå på andre måter.

2.2 Et eksempel for nivåestimatorer

I forbindelsen med FOB2001 var det utviklet en 'ny' metode for sysselsettingsstatus basert på flere administrative registre. Man kan beregne den eksakte utvalgsvariansen ved systematisk trekking for andel sysselsatt og ledig gitt en populasjon av dette registerkjennemerke. Siden denne registerstatusen er høyt korrelert med ILO-statusen fra AKU i det siste kvartal i året, gir de beregnete variansene et godt bilde av utvalgsvariansene for nivåesitmatorene basert på AKU.

Tabell 1 viser forholdet mellom utvalgsvariansen ved systematisk trekking basert på 10 forskjellige sorteringer av alle AKU-personer i Østfold i det 4. kvartal 2001, og variansen ved enkelt tilfeldig trekking. Her har vi beregnet variansen i andelen sysselsatt og ledig i det 4. kvartal 2001. I tillegg har vi beregnet variansene for det 4. kvartal 2002, der populasjonen består av de samme personer men med en oppdatert registerstatus. Vi gjør dette for å se hvordan variansene varierer over tid.

Tabell 1: Forhold mellom utvalgsvarianser ved systematisk og enkelt tilfeldig trekking.

	Sysselsatt		Ledig	
Sortering	kvt. 4, 2001	kvt. 4, 2002	kvt. 4, 2001	kvt. 4, 2002
fnr	0.968	0.817	1.218	0.955
kommnr fnr	1.016	1.272	0.943	0.965
kommnr kjonn fnr	0.968	1.578	1.094	0.945
kommnr alder fnr	0.663	1.000	1.064	0.788
kommnr kjonn alder fnr	0.759	0.876	1.000	0.930
kommnr alder kjonn fnr	0.745	1.309	1.167	1.015
famnr fnr	0.875	0.968	0.955	0.735
kommnr famstørr famnr fnr	1.033	0.746	0.867	0.970
alder fnr	0.650	0.984	1.047	0.766
alder kjonn fnr	0.860	0.937	0.955	1.078

fnr: personnummer, kommnr: kommunenr, famnr: familienummer, famstørr: størrelse til familie.

Det er store variasjoner i variansene både når det gjelder en bestemt sortering over tid, og når man sammenligner forskjellige sorteringer. Ta f.eks. sotering "alder fnr". Den gir omtrent like stor varians for andelen sysselsatt som enkelt tilfeldig utvalg i 2002, men 35% mindre varians i 2001. Den gir nesten like stor varians for andelen ledig i 2001, men 23.4% mindre varians i 2002. Eller sammenlign f.eks. sortering "alder kjonn fnr" med "kommnr alder kjonn fnr": når kommune sorteres først, er variansen økt med over 35% for andelen sysselsatt i 2002, mens den er redusert med litt over 10% i 2001.

2.3 Forklaring

Vi henviser til Appendiks A for en teoretisk utredning. Det vises at utvalgsvariansen er ustabil ved systematisk trekking basert på tilfeldig sortering. Dette kan sees på følgende måte:

1. Tilfeldig sortering av enheter med hver sin bestemt interesseverdi er det samme som (tilfeldig) permutasjon av interesseverdiene for en bestemt sortering av enhetene.

2. Variansen av utvalgsgjennomsnittet ved systematisk trekking er variansen av det tilsvarende gjennomsnittet i alle mulige systematiske utvalg.

3. Gjennomsnittet i alle mulige systematiske utvalg forandres ved permutasjon av (de samme) interesseverdiene i populasjonen, slik at variansen ved systematisk trekking forandres.

Videre er det slik at

- Gjennomsnittet til variansen ved systematisk trekking over alle mulige permutasjoner av interesseverdiene er det samme som variansen av utvalgsgjennomsnittet ved enkel tilfedig trekking.

- Variansen ved enkel tilfeldig trekking forandres ikke ved permutasjon av interesseverdiene, slik som variansen ved systematisk trekking gjør.

Derfor er variasjonen i systematisk trekking pga. tilfeldig sortering av populasjonen unødvendig.

2.4 Konklusjon

Systematisk trekking med antatt tilfeldig sortering bør erstattes med enkel tilfeldig trekking. Stratifisert systematisk trekking med antatt tilfeldig sortering innen hvert stratum bør erstattes med stratifisert tilfeldig trekking. Pinsippet gjelder alle utvalgsundersøkelser.

Dette kommer muligens overraskende for mange. Men følgende argumenter synes å være viktige:

1. Systematisk trekking med tilfeldig sortering forårsaker unødvendig variasjon i utvalgsvariansen, som ikke kan fjernes vha. etterstratifisering eller kalibrering i estimeringen. For AKU er varianskoeffisienten av utvalgsvariansen ca. 12% ved systematisk trekking med tilfeldig sortering.

2. For løpende utvalgsundersøkelser med konjunktur statistikk som viktig formål, som f.eks. AKU, er det klart uheldig hvis utvalg trukket på forskjellige tidspunkter har betydelig variert presisjon, som f.eks. opp til 50% av variansen (i gjennomsnitt) fra et kvartal til det neste.

3 Stratifisering

3.1 Nivåestimering

For estimeringen av nivået sammenligner vi følgende 6 utvalgplaner:

- enkelt tilfeldig trekking av personer innen hvert fylke, betegnet med "Pers Dir",
- proporsjonelt stratifisert tilfeldig trekking av personer etter alder (11 grupper) og kjønn innen hvert fylke, betegnet med "Pers Str", med ialt 22 strata innen hvert fylke,
- enkelt tilfeldig trekking av DSF-familier innen hvert fylke, betegnet med "Fam Dir",
- proporsjonelt stratifisert tilfeldig trekking av DSF-familier etter antallet AKU-personer i familien (3 grupper: 1-person, 2-personer, og ellers) innen hvert fylke, betegnet med "Fam Str1",
- proporsjonelt stratifisert tilfeldig trekking av DSF-familier etter den eldeste personens alder (11 grupper) innen hvert fylke, betegnet med "Fam Str2",
- proporsjonelt stratifisert tilfeldig trekking av DSF-familier etter den eldeste personens alder (11 grupper) og antallet AKU-personer i familien (3 grupper: 1-person, 2-personer, og ellers) innen hvert fylke, betegnet med "Fam Str", med ialt 33 strata innen hvert fylke.

Størrelsen til både person- og familieutvalg innen hvert fylke settes til det tilsvarende faktiske antallet i AKU, 4. kvartal 2001.

Vi beregner utvalgsvariansen til estimatoren for andelen sysselsatt og ledig for registerpopulasjonen i det 4. kvartal 2001. Tabell 2 viser fylkesvis forholdet mellom variansen ved forskjellige utvalgplanene og variansen ved Pers Dir. For andelen sysselsatt noterer vi at

Tabell 2: Fylkesvis forhold mellom varianser ved forskjellige utvalgsplaner og Pers Dir.

	Andel sysselsatt					
Fylke	Pers Dir	Per Str	Fam Dir	Fam Str1	Fam Str2	Fam Str
Østfold	1	0.771	2.188	1.053	2.003	0.893
Akershus	1	0.769	2.176	1.032	1.989	0.873
Oslo	1	0.774	2.047	1.008	1.902	0.848
Hedmark	1	0.769	2.336	1.134	2.139	0.947
Oppland	1	0.767	2.233	1.075	2.044	0.899
Buskerud	1	0.771	2.227	1.062	2.037	0.905
Vestfold	1	0.773	2.298	1.092	2.104	0.932
Telemark	1	0.766	2.386	1.145	2.193	0.974
Aust-Agder	1	0.770	2.491	1.156	2.273	0.988
Vest-Agder	1	0.770	2.309	1.061	2.110	0.915
Rogaland	1	0.772	2.362	1.076	2.158	0.934
Hordaland	1	0.769	2.312	1.069	2.112	0.915
Sogn og Fjordane	1	0.769	2.515	1.164	2.298	1.002
Møre og Romsdal	1	0.772	2.324	1.081	2.127	0.929
Sør-Trøndelag	1	0.772	2.149	1.024	1.967	0.870
Nord-Trøndelag	1	0.774	2.394	1.132	2.199	0.966
Nordland	1	0.772	2.270	1.084	2.081	0.928
Troms	1	0.770	2.144	1.019	1.969	0.863
Finnmark	1	0.771	2.369	1.114	2.177	0.950
	Andel ledig					
Fylke	Pers Dir	Per Str	Fam Dir	Fam Str1	Fam Str2	Fam Str
Østfold	1	0.995	0.946	0.938	0.945	0.936
Akershus	1	0.994	0.921	0.911	0.920	0.908
Oslo	1	0.996	0.917	0.910	0.916	0.908
Hedmark	1	0.995	1.015	1.006	1.015	0.998
Oppland	1	0.993	0.967	0.958	0.966	0.951
Buskerud	1	0.995	0.947	0.938	0.945	0.936
Vestfold	1	0.997	0.978	0.969	0.977	0.968
Telemark	1	0.991	1.025	1.015	1.022	1.007
Aust-Agder	1	0.997	0.982	0.972	0.979	0.963
Vest-Agder	1	0.998	0.917	0.905	0.916	0.905
Rogaland	1	0.996	0.921	0.910	0.921	0.908
Hordaland	1	0.995	0.925	0.916	0.924	0.912
Sogn og Fjordane	1	0.992	0.993	0.985	0.990	0.984
Møre og Romsdal	1	0.995	0.944	0.933	0.943	0.928
Sør-Trøndelag	1	0.995	0.914	0.904	0.912	0.904
Nord-Trøndelag	1	0.996	1.000	0.990	0.999	0.985
Nordland	1	0.996	0.961	0.951	0.959	0.950
Troms	1	0.994	0.902	0.894	0.900	0.885
Finnmark	1	1.000	0.946	0.940	0.945	0.934

- Stratifisert personutvalg reduserer variansen med litt over 20% i forholdet til direkte personutvalg.
- Direkte familieutvalg øker variansen med over 100% i forholdet til direkte personutvalg. Dette er effekten av trekkenhet.
- Av de to stratifiseringsvariabler for familieutvalg har familiestørrelsen den største effekt.
- Stratifisert familieutvalg etter familiestørrelsen og den eldeste personens alder er minst like effisient som direkte personutvalg.

For andelen ledig noterer vi at

- Ingen stratifiseringsvariabler brukt her gir særlige effekter i forholdet til direkte personutvalg.
- Familieutvalg er noe mer effisient enn personutvalg slik at effekten av trekkenhet går i den motsatte retningen som i tilfellet andel sysselsatt.

Vi legger til følgende kommentar:

- Vi har tatt med den eldeste personens alder som en stratifiseringsvariabel i vårt forslag tidligere mest pga. månedesutvalg i AKU, som av og til mangler personer i den eldeste aldersgruppen for fylkesvis kalibrering. Vi har imidlertid forslått en grovere 4-deling av alderen enn i tilfellet her.
- Siden dagens sortering av DSF-familie antas å være tilfeldig innen hver kommune, er utvalgsvariansen i gjennomsnitt omtrent den samme som direkte familieutvalg. Stratifisert familieutvalg kan derfor redusere variansen for andelen sysselsatt med over 50% i forholdet til dagens metode.
- Stratifisert familieutvalg taper noe effisiens i forholdet til stratifisert personutvalg. Den tapete effisiensen kan likevel gjenvinnes vha. etterstratifisering og kalibrering i estimeringen.
- Dagens utvalgsplan tilsier at det er over 400 DSF-familier i hvert fylke. Antallet utvalgsstrata innen hvert fylke er 12 i vårt forslag, noe som synes å kunne forsvares. Derimot ville det ha blitt for mange små strata om man tar med kommune i stratifiseringen. Simuleringen viser imidlertid at over 430 kommuner kan forventes å være representert i kvartalsutvalget med vårt forslag.

3.2 Estimering av endring

Det kan vises at registermetoden brukt i FOB2001 er misvisende når det gjelder endringene i AKU-populasjonen. Som datagrunnlag for den empiriske evalueringen har vi derfor brukt de observerte AKU paneler, ved å anta at alle relevante andeler og fordelinger i populasjonen fra et kvartal til det neste er akkurat de samme som observert i det tilsvarende AKU-panelet. På denne måten blir f.eks. et stratumsgjennomsnitt i populasjonen satt til det tilsvarende stratumsgjennomsnitt i utvalget, og en marginal fordeling av stratumsstørrelser i populasjonen satt til den tilsvarende fordelingen i utvalget. Mao. ser vi bort fra variasjonen i trekkesannsynligheten pga. stratifisering etter fylke og frafall.

Tabell 3: Forholdet mellom varianser ved forskjellige utvalgsplaner og Pers Dir.

Panel	Endring i andel sysselsatt				Endring i andel ledig			
(kvt/år)	Pers Dir	Per Str	Fam Dir	Fam Str	Pers Dir	Per Str	Fam Dir	Fam Str
4/01 - 1/02	1	0.995	2.752	1.014	1	0.997	2.671	0.983
1/02 - 2/02	1	0.998	2.757	0.992	1	0.999	2.751	0.997
2/02 - 3/02	1	0.996	2.735	0.980	1	1.002	2.696	0.976
3/02 - 4/02	1	0.997	2.718	0.990	1	0.997	2.732	0.987
4/02 - 1/03	1	0.998	2.726	0.988	1	0.997	2.704	0.977
1/03 - 2/03	1	0.992	2.775	1.000	1	0.996	2.728	0.976
2/03 - 3/03	1	0.994	2.823	1.010	1	0.998	2.658	0.952
3/03 - 4/03	1	1.003	2.763	0.987	1	1.003	2.722	0.966

Vi konsentrerer oss om følgende 4 utvalgsplaner: Pers Dir, Pers Str, Fam Dir og Fam Str, og beregner variansen for 8 kvartalsvis endringer mellom det 4. kvartal 2001 og det 4. kvartal 2003. Utvalgsstørrelsen var satt til den i Finnmark. Tabell 3 viser forholdet mellom variansene. Vi noterer at

- Resultatene er omtrent samme for sysselsatt og ledig.
- Stratifisering av personutvalg har nesten ingen effekt.
- Effekten av trekkenhet er stor og gjelder både endringer i andelen sysselsatt og ledig.
- Stratifisert familieutvalg er minst like effisient som personutvalg. I gjennomsnitt reduseres utvalgsvariansen med over 50% i forholdet til dagens utvalgsplan.

3.3 Allokering

Med allokering sikter man til valget av stratumsstørrelser i utvalget. Den såkalte optimale allokering avhenger av utvalgsvariansen i alle strata og gjelder kun for en bestemt estimator under betraktning. Estimering av forskjellige popualsjonstotaler eller -gjennomsnitter krever hver sin optimale allokering, som ikke lar seg forene i praksis. Uten klar føring i hva som bør prioriteres høyest, er det ikke mulig å komme med noen optimale allokering.

Vi har forslått en proporsjonell allokering der antallet trekkenheter i utvalget har et konstant forhold til det tilsvarende antallet i populasjonen i alle strata inne hvert fylke. For AKU gir dette et selvveied utvalg av både DSF-familier og personer innen hvert fylke. Faktisk er utvalget mer kontrollert enn et selvveied utvalg, siden vi stratifiserer etter familiestørrelsen. Dvs, antallet personer i utvalget har et konstant forhold til det tilsvarende antallet i populasjonen i alle strata for 1- eller 2-persons DSF-familier, mens det samme konstant forholdet gjelder forventningsmessig i alle strata for DSF-familier med flere enn 2 personer, som et vanlig selvveied utvalg.

4 Referansperson

I det 2. kvartal i året stilles det et husholdningsspørsmål i AKU, som går ut på å kartlegge den faktiske husholdningen til en bestemt referansperson i hver DSF-familie. Denne er fortiden den person som har likt person- og familienummer i DSF.

Det kan oppstå problem med dagens valg av referansperson når medlemmene i en DSF-familie faktisk bor i flere husholdninger. Ta f.eks. en borteboende student som ikke har meldt flytting og, dermed, formelt sett fortsatt bor hjemme hos sine foreldre. Med dagens praksis blir som regel en av foreldrene til studenten valgt som referanspersonen, siden det er denne person som 'bærer' familienummer i DSF. Dette fører til at den faktiske husholdningen til studenten ikke blir kartlagt. Videre er det slik at denne student har ingen mulighet for å bli kartlagt, dersom den bor alene eller sammen med en annen person som heller ikke 'bærer' sitt eget familienummer, fordi husholdningen til denne student i slike tilfeller har ingen referansperson under dagens opplegg.

Uansett hvor stort omfang denne type problem har, har man ingen grunn til å holde fast ved en utvalgsplan som ikke dekker hele populasjonen, siden en del personer har ingen sannsynlighet til å bli inkludert i utvalg, riktig nok kun mht. husholdningspørmålet. Vi forslår derfor at i framtiden skal referanspersonen velges tilfeldig innen hver DSF-familie i AKU. Med denne endring kommer utvalgsplanen til å dekke alle personer som bor i husholdningene med minst en person mellom 16 og 74 år. Såvidt vi kan se er det ikke nødvendig å endre spørresekvensen ellers ved interjvuet.

Et alternativ er å først trekke personer og deretter ta med DSF-familien til en trukket person i AKU. Disse personer kan da brukes som referanspersonene mht. husholdningsspørsmålet. Det er imidlertid et par ting som gjør at man kan se bort fra dette alternativ:

- I strata for DSF-familier med flere enn 1 person finnes det nå en liten sannsynlighet for at man unødig trekker flere personer fra den samme DSF-familien i en tilfeldig trekking.
- I strata for DSF-familier med flere enn 2 personer avhenger nå trekkesannsynligheten til en person av dens DSF-familiestørrelse, noe som gjør at utvalget ikke lenger er selvveied for personer fra DSF-familier av forskjellige størrelser. Dette er antageligvis grunnen til at husholdninger trukket via personer hadde større utvalgsvarianser enn husholdninger trukket direkte i simuleringsstudiet av Vedø og Rafat (Notat 2003/56).

A Systematisk trekking

A.1 Standard systematisk trekking

Anta populasjon $U = \{1, ..., N\}$. La y_i være interessevariablen, der

$$E_1(y_i) = \mu, \quad E_1\{(y_i - \mu)^r\} = \mu_r \quad (i \in U), \quad E_1(y_i y_j) = \mu^2 \quad (i \neq j)$$

hvor E_1 betegner forventning mht. modellen for y_i. Betrakt systematisk trekking basert på tilfeldig sortering av populasjonen. La $\bar{y}_m$ være gjennomsnittet i det $m$te utvalget, for $m = 1, ..., k$. Anta at $N = nk$. Utvalgsvariansen av $\bar{y}_m$ er da gitt som

$$V_{sys} = \frac{1}{k}\sum_{m=1}^{k}(\bar{y}_m - \bar{Y})^2 = \frac{1}{k}\sum_{m=1}^{k}(\bar{y}_m - \mu)^2 - (\bar{Y} - \mu)^2$$

slik at

$$E_1(V_{sys}) = \frac{1}{k}k(\frac{\mu_2}{n}) - \frac{\mu_2}{N} = (1 - \frac{1}{k})\frac{\mu_2}{n}.$$

Videre,

$$V_{sys}^2 = \frac{1}{k^2}\{\sum_{m=1}^{k}(\bar{y}_m - \mu)^4 + \sum_{p\neq m}(\bar{y}_m - \mu)^2(\bar{y}_p - \mu)^2\} - \frac{2}{k}\sum_{m=1}^{k}(\bar{y}_m - \mu)^2(\bar{Y} - \mu)^2 + (\bar{Y} - \mu)^4$$

der $(\bar{y}_m - \mu)^2(\bar{Y} - \mu)^2$ kan skrives som

$$(\bar{y}_m - \mu)^2\{\frac{1}{k}\sum_{p=1}^{k}(\bar{y}_p - \mu)\}^2 = \frac{1}{k^2}\{(\bar{y}_m - \mu)^4 + \sum_{p\neq m}(\bar{y}_m - \mu)^2(\bar{y}_p - \mu)^2 + (\bar{y}_m - \mu)^2\sum_{p\neq q\neq m}(\bar{y}_p - \mu)(\bar{y}_q - \mu) + (\bar{y}_m - \mu)^3\sum_{p\neq m}(\bar{y}_p - \mu)\}$$

slik at

$$E_1\{(\bar{y}_m - \mu)^2(\bar{Y} - \mu)^2\} = \frac{\mu_{4,n}}{k^2} + \frac{k-1}{k^2}\frac{\mu_2^2}{n^2} \quad \text{for} \quad \mu_{4,n} = E_1\{(\bar{y}_m - \mu)^4\} = \frac{\mu_4}{n^3} + \frac{3\mu_2^2}{n^2}$$

der $\mu_{4,n}$ er den fjerde sentrale moment av $\bar{y}_m$. Vi har

$$\begin{aligned}
E_1(V_{sys}^2) &= \frac{1}{k^2}\{k\mu_{4,n} + k(k-1)\frac{\mu_2^2}{n^2}\} - \frac{2}{k}k\{\frac{\mu_{4,n}}{k^2} + \frac{k-1}{k^2}\frac{\mu_2^2}{n^2}\} + \mu_{4,N} \\
&= \frac{k-2}{k^2}\mu_{4,n} + \mu_{4,N} + \frac{(k-1)(k-2)}{k^2}\frac{\mu_2^2}{n^2} \\
V_1(V_{sys}) &= \frac{k-2}{k^2}\mu_{4,n} + \mu_{4,N} - \frac{k-1}{k^2}\frac{\mu_2^2}{n^2} > \frac{2k-5}{k^2}\frac{\mu_2^2}{n^2} \\
CV_1(V_{sys}) &= \frac{SD_1(V_{sys})}{E_1(V_{sys})} > \sqrt{\frac{2k-5}{(k-1)^2}} \doteq \sqrt{\frac{2}{k}}.
\end{aligned}$$

Hele utledningen er betydelig forenklet basert på følgende tilnærming

$$V_{sys} = \frac{1}{k}\sum_{m=1}^{k}(\bar{y}_m - \mu)^2 - (\bar{Y} - \mu)^2 \doteq \frac{1}{k}\sum_{m=1}^{k}(\bar{y}_m - \mu)^2.$$

Det følger at

$$\begin{aligned} E_1(V_{sys}) &\doteq \frac{\mu_2}{n} \\ E_1(V_{sys}^2) &\doteq \frac{\mu_{4,n}}{k} + \frac{k-1}{k}\frac{\mu_2^2}{n^2} \\ V_1(V_{sys}) &\doteq \frac{\mu_{4,n}}{k} + (\frac{k-1}{k} - 1)\frac{\mu_2^2}{n^2} > \frac{3\mu_2^2}{kn^2} - \frac{\mu_2^2}{kn^2} = \frac{2}{k}\frac{\mu_2^2}{n^2} \\ \text{og} \quad CV_1(V_{sys}) &> \sqrt{\frac{2}{k}}. \end{aligned}$$

A.2 Systematisk πps trekking

Systematisk πps trekking er standard systematisk trekking mht. tilleggs *kumulativ total* x_i for $i \in U$. For enkelhetsskyld, og uten praktisk betydning, anta at x_i er et heltall. La $X = \sum_{i\in U} x_i$. Trekkelengden er nå gitt som $k = X/n$, der vi antar at k er et heltall slik som X og n. På denne måten kan standard systematisk trekking betraktes som et spesielt tilfelle av systematisk πps trekking der $x_i \equiv 1$ for alle i. Enheten i med kumulativ total x_i er med i x_i forskjellige systematiske utvalg. Trekkesannsynligheten er slik at

$$\pi_i = n\frac{x_i}{X} < 1.$$

Basert på et systematisk πps utvalg, betegnet med s_m for $m = 1, ..., k$, har vi

$$\hat{Y}_m = \sum_{i\in s_m}\frac{y_i}{\pi_i} = \frac{X}{n}\sum_{s_m} b_i = X\bar{b}_m \quad \text{for} \quad b_i = \frac{y_i}{x_i} \quad \text{og} \quad \bar{b}_m = \sum_{i\in s_m} b_i/n.$$

Det følger at

$$\begin{aligned} E_{sys}(\bar{b}_m) &= \frac{1}{k}\sum_{m=1}^{k}(\sum_{i\in s_m}\frac{b_i}{n}) = \frac{1}{k}\sum_{i\in U}\frac{x_i b_i}{n} = \frac{1}{kn}\sum_{i\in U} y_i = \frac{Y}{X} = B \\ E_{sys}(\hat{Y}_m) &= XE_{sys}(\bar{b}_m) = XB = Y \\ V_{sys}(\hat{Y}_m) &= X^2 V_{sys}(\bar{b}_m) = X^2\{\frac{1}{k}\sum_{m=1}^{k}(\bar{b}_m - B)^2\}. \end{aligned}$$

Vi skal nå betrakte variansen ved systematisk πps trekking under følgende modell

$$b_i = \frac{y_i}{x_i} = \beta + \epsilon_i \quad \text{der} \quad E_1(\epsilon_i) = 0 \quad \text{og} \quad V_1(\epsilon_i) = \mu_2 = \sigma^2$$

og ϵ_i er uavhengig av ϵ_j for $i \neq j$. Legg merke til at, siden ϵ_i er uavhengig av x_i under modellen,

enhver sortering av populasjonen basert på x_i kan betraktes som tilfeldig mht. ϵ_i. Som i tilfellet standard systematisk trekking ovenfor, er utledningen betydelig forenklet ved følgende tilnærming

$$V_{sys}(\bar{b}_m) = \frac{1}{k}\sum_{m=1}^{k}(\bar{b}_m - \beta)^2 - (B-\beta)^2 \doteq \frac{1}{k}\sum_{m=1}^{k}(\bar{b}_m - \beta)^2 = \frac{1}{k}\sum_{m=1}^{k}\bar{\epsilon}_m^2.$$

Det følger at

$$E_1(V_{sys}) \doteq \frac{\mu_2}{n}.$$

Videre har vi

$$V_{sys}^2 \doteq \frac{1}{k^2}\{\sum_{m=1}^{k}\bar{\epsilon}_m^4 + \sum_{p\neq m}\bar{\epsilon}_m^2\bar{\epsilon}_p^2\}.$$

I motsetning til tilfellet standard systematisk trekking, er ikke $\bar{\epsilon}_m$ and $\bar{\epsilon}_p$ alltid uavhengige av hverandre siden det kan finnes enheter som er med både i s_m og s_p. Likevel,

$$E_1(\bar{\epsilon}_m^2\bar{\epsilon}_p^2) = \frac{1}{n^4}\sum_{(i_1,i_2)\in s_m,(j_1,j_2)\in s_p} E_1(\epsilon_{i_1}\epsilon_{i_2}\epsilon_{j_1}\epsilon_{j_2}),$$

der et ledd på den høyre siden ikke er lik null, dvs. positiv, bare hvis det har form $E_1(\epsilon_i^4)$ eller $E_1(\epsilon_i^2\epsilon_j^2)$. Nærmere sagt, la s_{mp} være snittet til s_m og s_p. La s_m^p være enhetene i s_m som ikke er med i s_p, og la s_p^m være enhetene i s_p som ikke er med i s_m. Vi har

$$\begin{aligned}
E_1(\bar{\epsilon}_m^2\bar{\epsilon}_p^2) &= \frac{1}{n^4}E_1\{\sum_{i\in s_{mp}}\epsilon_i^4 + \sum_{i\in s_{mp},g\in s_p^m}\epsilon_i^2\epsilon_g^2 + \sum_{i\in s_{mp},h\in s_m^p}\epsilon_i^2\epsilon_h^2 + \sum_{g\in s_p^m,h\in s_m^p}\epsilon_g^2\epsilon_h^2\} \\
&\geq \frac{1}{n^4}\{\sum_{i\in s_{mp}}E_1^2(\epsilon_i^2) + \sum_{i\in s_{mp},g\in s_p^m}\mu_2\mu_2 + \sum_{i\in s_{mp},h\in s_m^p}\mu_2\mu_2 + \sum_{g\in s_p^m,h\in s_m^p}\mu_2\mu_2\} \\
&= \frac{1}{n^4}\{(\sum_{i\in s_m}\mu_2)(\sum_{i\in s_p}\mu_2)\} = \frac{\mu_2^2}{n^2}.
\end{aligned}$$

Likheten inntreffer hvis s_{mp} er tomt. La $\mu_{4,n}$ være den fjerde sentrale moment av $\bar{\epsilon}_m$. Vi har

$$\begin{aligned}
V_1(V_{sys}) &> \frac{\mu_{4,n}}{k} + (1 - \frac{1}{k} - 1)\frac{\mu_2^2}{n^2} > \frac{2}{k}\frac{\mu_2^2}{n^2} \\
CV_1(V_{sys}) &> \sqrt{\frac{2}{k}}.
\end{aligned}$$

A.3 Etterstratifisering

Den etterstratifiserte estimatoren for populasjonsgjennomsnittet $\bar{Y} = \sum_{i=1}^{N} y_i/N$ er gitt som

$$\hat{\bar{Y}}_{pst} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h$$

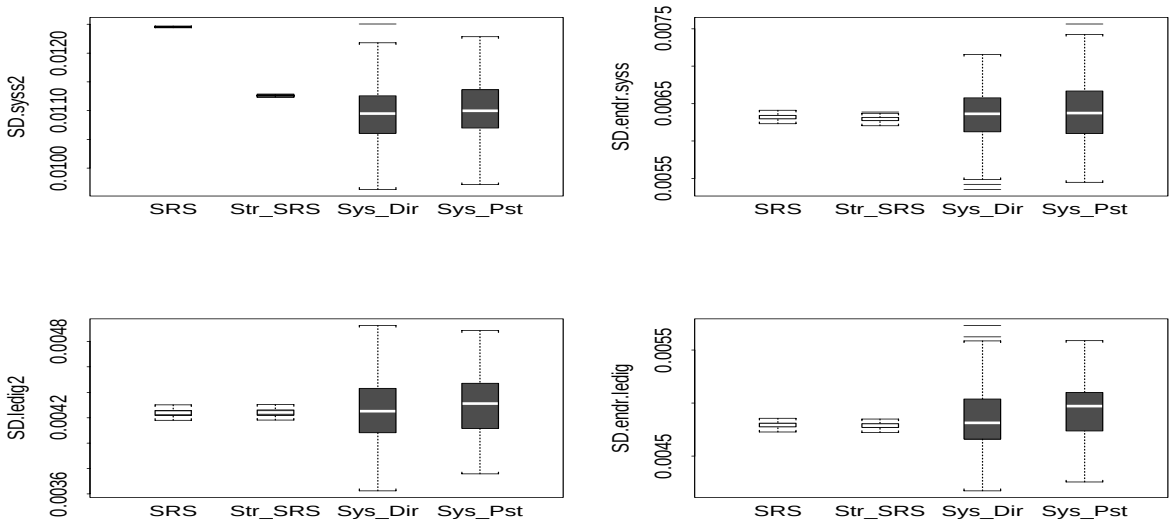
der h betegner etterstratum med N_h enheter, for $h = 1, ..., H$, og $N = \sum_h N_h$, og $\bar{y}_h$ betegner det tilsvarende etterstratumsgjennomsnitt i utvalget. Variansen til $\hat{\bar{Y}}_{pst}$ bestemmes derfor av variasjonen i $\mathbf{y} = (\bar{y}_1, ..., \bar{y}_h)$ under en gitt utvalgsplan. Siden variasjonen i $\mathbf{y}$ ved systematisk trekking forandres med tilfeldig sortering av populasjonen, forandres også variansen av $\hat{\bar{Y}}_{pst}$ ved systematisk trekking.

Vi har undersøkt variasjonen i $V(\hat{\bar{Y}}_{pst})$ vha. følgende simulering:

- Fiks registerpopulasjonen i det 4. kvartal 2001.
- For hver person i populasjonen, generer en ny sysselsettingsstatus etter overgangs sannsynligheter observert i AKU 4. kvartal 2001 til 1. kvartal 2002.

For hver simulert populasjon beregner vi variansen av andelen sysselsatt og ledig i 1. kvartal 2002, og endringen i de to andeler i forholdet til det 4. kvartal 2001, ved

- systematisk trekking med sortering "kommnr alder kjonn fnr", betegnet med "Sys Dir",
- enkelt tilfeldig trekking uten tilbakelegging, betegnet med "SRS",
- proporsjonelt stratifisert tilfeldig trekking etter alder og kjønn, betegnet med "Str SRS",
- etterstratifisering (etter alder og kjønn) av systematisk utvalg, betegnet med "Sys Pst".



Figur 1: Boxplot av standard avvik ved forskjellig utvalgsplan og estimator fra 150 simuleringer.

Figur 1 viser 150 setter av slike simulerte standard avvik for Østfold. Bildet er nesten det samme når simuleringen gjentas i andre fylker. Man legger merke til at (i) i gjennomsnitt gir stratifisert tilfeldig trekking omtrent den samme utvalgsvarians som systematisk trekking, og (ii) mens variansen til SRS og Str SRS varierer veldig lite fra populasjon til populasjon, er variasjonen veldig stor ved systematisk trekking, noe som ikke reduseres nevneverdig vha. etterstratifisering.

De sist utgitte publikasjonene i serien Notater

2004/54 T.M. Normann: Samordnet levekårsundersøkelse 2001 - panelundersøkelsen. Dokumentasjonsrapport. 54s.

2004/55 T.M. Normann: Samordnet levekårsundersøkelse 2002 - panelundersøkelsen. Dokumentasjonsrapport. 89s.

2004/56 T. Guldbrandsen og A. Holmøy: Omnibusundersøkelsen april/mai 2004. Dokumentasjonsrapport. 54s.

2004/57 Ø. Brekke: Praktisk guide for teknisk utstyr og dataprogrammer i brukertester. 33s.

2004/58 K. Henriksen: Ny metode for prismåling av personbiler i konsumprisindeksen. 24s.

2004/59 A.S. Abrahamsen, J. Heldal, og D. Rafat: UT- Undersøkelsene i 2004 for ikke-finansielle foretak. Utvalgsplaner og utvalg til kvartals og årsundersøkelsene. 48s.

2004/60 Ø. Bolsgård og L-C. Zhang: Prisindeks for engoshandel . 35s.

2004/61 T. Guldbrandsen og B.O. Lagerstrøm: Undersøkelse om arbeids- og boligforhold. Dokumentasjonsrapport. 27s.

2004/62 G. Dahl: Trygd blant innvandrere 1992-2000. 79s.

2004/63 A. H. Sætre og N .Buskoven: Lokalvalgundersøkelsen 2003. Dokumentasjonsrapport. 79s.

2004/64 Kravspesifikasjon for elektronisk innberetning, kjennemerke og filbeskrivelse for lønnsstatistikken. Oppdatert 2004. 16s.

2004/65 L. Østby: Innvandrere i Norge - Hvem er de, hvordan går det med dem? Del I Demografi. 156s.

2004/66 L. Østby: Innvandrere i Norge - Hvem er de, hvordan går det med dem? Del I Levekår 154s.

2004/67 L. Lerskau, K.M. Heide, E. Holmøy og I.F. Solli: Virkningsberegninger på MSG6. Appendiks til Rapporter 2004/18 "Macroeconomic Properties of the Norwegian Applied General Equilibrium Model MSG6". 140 s.

2004/68 A. Holmøy, R. Johannessen og L. Solheim: Etablering av ny husleiestatistikk (indeks) - en forstudie. 19s.

2004/69 E.E. Eibak og F. Haraldsen: Undersøking om foreldrebetaling i barnehagar, august 2004. 45s.

2004/70: A. Raknerud, D. Rønningen og T. Skjerpen: Dokumentasjon av kapitaldatabasen. En database med data for varige driftsmidler og andre økonomiske data på foretaksnivå. 12s.

2004/71 M. T. Dzamarija: Norske barn i utlandet. Utvalgte land: Pakistan, Marokko, Tyrkia og Spania. 32s.

2004/72 A. S. Abrahamsen og A. Seierstad: Analyse av revisjon. KOSTRA kommunehelse. 49s.

2004/73 E. Mørk og E. Willand-Evensen: Husholdningers forbruk. En sammen-likning av forbruksundersøkelsen og nasjonalregnskapet. 36s

2004/74 M. Aamodt: Kvalitetsprosjektet for videregående opplæring. Utført på oppdrag fra Utdannings- og forskningsdepartementet i perioden mars 2003-september 2004. 187s.

2004/75 S. Blom: Holdninger til innvandrere og innvandring 2004. 53s.

2004/76 A. Rolland: En inspeksjon av Elevinspektørene. 50s.

2004/77 A. Rolland: KOSTRA og kvaliteten på de kommunale tjenester. 31s.

2004/78 J. A. Osnes: Beregningsutvalget. Dokumentasjon av SAS-systemet. 97s.

2004/79 T. Eika og T. Skjerpen: Hvitevarer 2005. Modell og prognose. 17s.