Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

IO 74/50

21. november 1974

VARIANSESTIMERING FOR NIVÅTALLESTIMATER OG ENDRINGSTALLESTIMATER VED BYRÅETS ARBEIDSKRAFTUNDERSØKELSER

av

John Dagsvik*)

*) Jeg vil takke Ib Thomsen, Petter Laake, Stein Østerlund Petersen og medarbeiderne i Sosiodemografisk forskningsgruppe for nyttige kommentarer under bearbeidelsen av manuskriptet.
Kjetil Sørlie er ansvarlig for programmeringen av estimatorene, og jeg vil derfor rette en spesiell takk til han for en meget effektiv hjelp og et godt samarbeid.

Sammendrag

Ved Byråets arbeidskraftundersøkelser (AKU) skiftes en del av utvalget ut hvert kvartal i følge en spesiell rotasjonsplan. I dette notatet har en med utgangspunkt i denne rotasjonsplanen, utledet en estimator for variansen til estimert endring i tallene fra en undersøkelse til en annen.

Ved Intervjukontoret ble det våren 1974 laget et program for estimering av varianser til estimater for nivåtall. Dette programmet er utvidet med en subrutine som beregner tilsvarende variansestimater for endringstall basert på den estimatoren vi har utledet.

Det utvidede variansberegningsprogrammet er anvendt til beregning av estimater for nivåtall, endringstall samt deres varianser. Videre har en funnet estimater for "design-effekten" og for variansenkomponentene som skyldes variansen mellom utvalgsområdene og innen utvalgsområdene.

1. Innledning

Mens variansestimatorene ved utvalgsundersøkelser der utvalget er trukket rent tilfeldig er forholdsvis enkle og kan beregnes for hånd, er dette en temmelig arbeidskrevende regneoperasjon når utvalget er trukket etter Byråets utvalgsplan.

Variansestimering på grunnlag av slike kompliserte utvalgsplaner har ofte vært basert på grove tilnærmingsformler en har funnet blant annet ved empiriske undersøkelser i andre land.

Ved Intervjukontoret ble det våren 1974 laget et variansberegningsprogram på grunnlag av en forventningsrett variansestimator for to-trinns utvalg. (Laake et al. 1974). Dette programmet estimerer variansen til estimert gjennomsnitt og samlet antall av en variabel, ved en enkelt undersøkelse. Den formelle utledningen av denne variansestimatoren er gjengitt blant annet hos Des Raj 1968 og hos Hoem 1973. I AKU benyttes en roterende utvalgsplan som er slik at en fjerdedel av utvalget skiftes ut ved hvert kvartal og hvor halvparten av et ordinært kvartalvis utvalg er felles ved ett kvartal ett år og samme kvartal neste år. Intervjukontorets opprinnelige variansberegningsprogram kan ikke uten videre benyttes til å beregne variansestimater til estimatene for endringene i tallene fra en undersøkelse til en annen.

Vi har imidlertid utledet en variansestimator basert på en slik rotasjonsplan. Denne estimatoren har en slik form at de numeriske beregningene kan gjøres ved hjelp av en subrutine til det opprinnelige variansberegningsprogram.

I Byrået er det tidligere også utledet en slik variansestimator men denne er basert på grove tilnærmelser (Østerlund Petersen 1972).

I den første delen av notatet presenteres standardavvikestimater for nivåtallestimater og endringstallestimater. Videre gis estimater for "designeffekten" og varianskomponentene mellom og innen utvalgsområdene. I den andre delen presenteres den formelle utledningen av estimatoren for variansen til endringstallestimatoren, dvs. differensen mellom de respektive nivåtallsestimatorene.

2. Estimater for nivåtall og endringstall, samt deres standardavvik. Estimater for "design-effekten" og varianskomponentene mellom og innen utvalgsområdene

2.1. Estimatorene

De variansestimatorene som er programmert gjelder egentlig bare når nivåtallene estimeres ved vanlig oppblåsning av tallene i utvalget. I Byråets AKU er imidlertid estimeringsmetoden en annen idet en blåser opp tallene i utvalget for hvert kjønn og ettårige aldersklasser og deretter summerer disse. Nå gir variansberegningsprogrammet også utskrift av nivåtallene estimert ved vanlig oppblåsning av tallene fra utvalget slik at det er mulig å sammenlikne de to estimeringsmetodene (og eventuelt avgjøre om resultatene er signifikant forskjellige). Det følger forøvrig av Dagsvik 1974 (setning 3.7 (iv)) at variansen til Byråets AKU-estimater er større enn variansen til estimater beregnet ved vanlig oppblåsning.

Variansberegningsprogrammet gir endelig estimater for "design-effekten". Denne er definert som forholdet mellom variansen til den estimator som benyttes og variansen til den tilsvarende estimator dersom utvalget var trukket enkelt tilfeldig.

I Byråets AKU er annentrinns trekkeenheter de enkelte husholdninger. I den matematiske teorien inngår antall husholdninger i de utrukne utvalgsområdene i beregningen av den totale utvalgsbrøk som benyttes i variansestimatorene og ved oppblåsingen. Imidlertid inneholder ikke intervjukontorets utvalgsregister husholdning som enhet slik at vi må nøye oss med å estimere den totale utvalgsbrøk på grunnlag av kjennskapet til antall personer i bestanden. Vi innfører dermed en ny feilkilde i tillegg til de mer tradisjonelle feilkilder som frafall, utvalgsfeil og registerfeil. Vi har imidlertid antatt at denne feilkilden har liten betydning sammenliknet med de andre feilkildene.

2.2. Estimater for nivåtall og deres standardavvik

Nivåtallene ved første og annet kvartal 1973 og 1974 for sysselsatte i hovednæringer estimert ved de to omtalte metodene samt standardavvikestimater for oppblåste tall er gjengitt i tabell 1. Denne viser at forskjellen mellom estimatene beregnet ved de to estimeringsmetodene er liten. Forskjellen er størst for jordbruk, men heller ikke her er forskjellen større enn ett standardavvik. Dersom standardavvik brukes som signifikansmål (Thomsen 1973) finner vi at resultatene fra de to

Tabell 1. Sysselsatte etter næring. 1 000

Næringer	1. kvartal 1973			2. kvartal 1973			1. kvartal 1974			2. kvartal 1974		
	AKU tall	Oppblåste tall	Standard avvik	AKU tall	Oppblåste tall	Standard avvik	AKU tall	Oppblåste tall	Standard avvik	AKU tall	Oppblåste tall	Standard avvik
Jordbruk	150	156	10,5	163	167	11,5	121	129	9,0	142	151	10,2
Skogbruk	19	19	3,0	14	15	2,7	14	14	2,6	13	14	2,4
Fiske og fangst	31	31	4,2	21	22	3,4	35	35	4,8	24	23	3,6
Bergverksdrift	10	10	3,7	14	14	4,7	12	11	3,0	8	8	2,9
Industri	378	378	14,5	390	391	15,5	378	378	14,4	399	400	14,6
Kraft og vannforsyning	13	13	2,0	16	16	2,3	17	18	2,3	16	17	2,4
Bygg og anlegg	135	135	7,3	142	143	7,4	140	140	6,7	149	150	7,3
Varehandel, hotell og restaurant	269	270	11,9	279	276	11,8	281	282	11,7	274	273	11,8
Transport, lagring, post og telekomm. ...	159	157	7,4	161	160	8,6	159	156	7,6	164	162	8,2
Bank og finansieringsvirksomhet, eiendomsdrift og forr. tjenesteyting	67	66	4,8	63	60	4,8	67	66	5,8	66	64	5,3
Offentlig, sosial og privat tjenesteyting.	411	411	15,1	393	392	14,4	420	417	15,2	391	388	14,7
Sysselsatte i alt, medregnet uoppgitt ..	1 645	1 649	24,1	1 657	1 663	24,3	1 645	1 652	24,5	1 649	1 643	25,0

estimeringsmetodene ikke er signifikant forskjellige. Studerer vi standardavvikestimatene ser vi at de varierer forholdsvis lite over tid for samme næring mens de varierer en del mellom enkelte næringer hvor antall sysselsatte er av samme størrelsesorden.

2.3. Varianskomponentene mellom og innen utvalgsområdene

Variansen kan oppdeles i komponenter som måler henholdsvis variasjonen innen og mellom utvalgsområdene. Varianskomponenten mellom utvalgsområdene er derfor et mål på hvor god stratifiseringen er. Dessverre gir ikke det opprinnelige variansberegningsprogrammet utskrift av disse varianskomponentene for nivåtallestimatene. Subrutinen som beregner variansestimatene for endringstallestimatene gir derimot utskrift av tilsvarende komponenter. Disse er gjengitt i tabell 2. Tallene tyder på at variansen mellom utvalgsområdene er av samme størrelsesorden som variansen innen utvalgsområdene for jordbruk, industri, varehandel m.m., og offentlig sosial og privat tjenesteyting, mens variansen mellom utvalgsområdene er vesentlig mindre enn variansen innen utvalgsområdene for kraft og vannforsyning, og bygg og anlegg. I Sverige er det gjort tilsvarende beregninger og disse resultatene viser gjennomgående at varianskomponenten innen utvalgsområdene er vesentlig større enn mellom utvalgsområdene for de fleste næringer. Dette tyder på at stratifiseringen ved den svenske utvalgsplanen er bedre enn ved den norske.

Tabell 2. Varianskomponentene innen- og mellom utvalgsområdene for 2. kvartal 1973 og 1974 i prosent av total varians

Næringer		Variansen innen utvalgs-områdene	Variansen mellom utvalgs-områdene
Jordbruk	1973	50	50
	1974	56	44
Industri	1973	51	49
	1974	53	47
Kraft og vann-forsyning	1973	64	36
	1974	74	26
Bygg og anlegg	1973	76	24
	1974	81	19
Varehandel, hotell og restaurant	1973	54	46
Off. sosial og privat tjenesteyting	1973	57	43
	1974	52	48

2.4. Estimater for endringstall og deres standardavvik

Tabell 3 viser estimerte endringer fra første og annet kvartal 1973 til henholdsvis første og annet kvartal 1974 for noen hovednæringer, samt estimert standardavvik for differensen mellom de respektive oppblåste tall. Vi legger merke til at bortsett fra jordbruk er endringstallestimatene i annet kvartal mindre enn standardavvikestimatene. Altså er endringstallestimatene i annet kvartal signifikante bare for jordbruket.

Videre er standardavvikestimatene for endringstallestimatene stort sett mindre enn standardavvikestimatene for de respektive nivåtallestimater. I samsvar med resultatene i kapitel 2.2 viser også standardavvikestimatene for endringstallestimatene stor stabilitet over tid.

I Byrået er det også gjort sammenlikninger med en annen metode for estimering av endringstall, nemlig ved å estimere endringene på grunnlag av den delen av utvalget som er felles ved begge undersøkelser. Siden denne delen av utvalget bare utgjør halvparten av et ordinært AKU-utvalg har følgelig estimatene større varians. En har kalt denne metoden den direkte estimeringsmetoden mens den andre metoden vi har brukt er blitt kalt den indirekte metoden. For endringene fra 1. kvartal 1973 til 1. kvartal 1974 gir de to metodene en nedgang på henholdsvis 27 000 og 15 000 for jordbruket. Det første estimatet har et standardavvik på ca. 10 000 og det andre estimatet har et standardavvik større enn 10 000. Begge konfidensintervallene for de respektive estimatene inneholder altså intervallet (17 000, 25 000) slik at vi ikke kan si at de to metodene gir signifikant forskjellig resultat i dette tilfelle. Tilsvarende betraktninger gjelder også for andre næringer.

Den store forskjellen mellom estimatene ved de to estimeringsmetodene gir likevel grunn til å undersøke om en kombinasjon av indirekte og direkte estimatorer kan gi bedre resultater. Slike "sammensatte estimater" er i bruk i andre land og har vist seg å gi presisjonsgevinster for mange variable.

Tabell 3. Endringstall 1973-1974. 1 000

Næringer	Endringer 1. kvartal 1973 til 1. kvartal 1974			Endringer 2. kvartal 1973 til 2. kvartal 1974		
	AKU tall	Oppblåste tall	Standardavvik	AKU tall	Oppblåste tall	Standardavvik
Jordbruk	-29	-27	9,7	-21	-16	9,8
Industri	0	0	12,7	9	9	13,6
Kraft og vannforsyning ..	4	5	2,6	0	-1	3,0
Bygg og anlegg	5	5	7,0	7	7	8,8
Varehandel, hotell og restaurant	12	12	11,7	-5	-3	7,0
Off. sosial og privat tjenesteyting	9	6	13,6	-2	-4	12,2

2.5. Estimater for "design-effekten"

"Design-effekten" er et mål for "effekten" av den spesielle utvalgsplanen som ligger til grunn relativt til en utvalgsplan der utvalget er trukket rent tilfeldig. Dersom et kjennetegn er jevnt fordelt i bestanden er det av mindre betydning hvilken utvalgsplan som ligger til grunn. Derfor er "design-effekten" også et mål på hvor homogen bestanden er med hensyn på de respektive kjennetegn.

Ved å studere tabell 4 ser vi at tre næringsgrupper skiller seg klart ut med stor "design-effekt", nemlig gruppene jordbruk, fiske og fangst, og bergverksdrift. Spesielt bergverksdrift har en svært stor "design-effekt". Dette er i samsvar med det kjennskap en har fra andre kilder om næringens geografiske fordeling. (Rideng 1970). Bortsett fra bergverksdrift og fiske og fangst viser disse estimatene stabilitet over tid.

For de 5 siste næringsgruppene i tabellen ligger estimatene stort sett i underkant av 1,5.

I Byrået har en ofte brukt en variansestimator basert på antagelsen at design-effekten er lik 1.5. Våre resultater viser at denne estimatoren ville ha underestimert variansen for de 4 første næringsgruppene og overestimert variansen for de resterende gruppene.

Tabell 4. Designeffekt

Næringer	1. kvartal 1973	2. kvartal 1973	1. kvartal 1974	2. kvartal 1974
Jordbruk	2,6	2,6	2,4	2,6
Skogbruk	1,7	1,8	1,7	1,6
Fiske og fangst	2,0	2,0	2,5	2,1
Bergverksdrift	4,8	5,8	3,1	4,2
Industri	1,6	1,8	1,5	1,5
Kraft og vannforsyning	1,2	1,3	1,1	1,2
Bygg og anlegg	1,3	1,4	1,2	1,2
Varehandel, hotell og restaurant .	1,5	1,5	1,3	1,3
Transport, lagring, post og telekomm.	1,3	1,5	1,3	1,3
Bank og finansieringsvirksomhet, eiendomsdrift og forr. tjenesteyting	1,3	1,4	1,7	1,5
Off. sosial og privat tjenesteyting	1,5	1,4	1,2	1,3

3. Den matematiske teori for estimering av variansen til endringstallestimater når utvalget er roterende

3.1. Varianser

Utgangspunktet for utledningen av en estimator for variansen til endringstallestimatoren er at det samlede utvalg som var med i en av undersøkelsene på tidspunkt 1 og tidspunkt 2 kan betraktes som ett utvalg trukket uten tilbakelegging ved første undersøkelsestidspunkt. Dette utvalget består derfor av 3 delutvalg, S_1 som bare er med i første undersøkelse, S_2 som er med i begge undersøkelser, og S_3 som er med i siste undersøkelse. I samsvar med gjeldende utvalgsplan antar vi at delutvalget $S_1 \cup S_2$ er like stort som delutvalget $S_2 \cup S_3$, nemlig et ordinært AKU utvalg på n trekkenheter. Størrelsen på delutvalget S_2 betegner vi med n'. I det følgende er symboler og konvensjoner definert i samsvar med Hoem 1973.

Bestanden er inndelt i et visst antall strata hvor stratum nr. i inneholder M_i primære utvalgsområder. Det j-te primære utvalgsområde

i stratum nr. i har $N_i(j)$ trekkeenheter. Den k-te trekkeenhet i j-te utvalgsområde i i-te stratum har verdien $a_i(j, k)$ ved tidspunkt 1 og verdien $c_i(j, k)$ ved tidspunkt 2. Vi antar at det er like mange trekkeenheter i de primære utvalgsområdene ved begge de to undersøkelsestidspunktene. Vi innfører notasjonene

$$N_i = \sum_j N_i(j), \quad N = \sum_j N_i,$$

$$a_i(j) = \sum_k a_i(j, k), \quad \bar{a}_i(j) = a_i(j) / N_i(j),$$

$$a_i = \sum_j a_i(j), \quad \bar{a}_i = a_i / M_i ,$$

$$a = \sum_i a_i ,$$

og tilsvarende notasjoner for c-verdiene.

Blant de M_i utvalgsområdene i stratum i trekkes m_i områder med nummerene $J_{i1}, J_{i2}, .., J_{im}$, rent lotterisk. Sannsynligheten for at et vilkårlig område i i-te stratum skal komme med i utvalget er altså $m_i/M_i = \pi_i$. Vi lar $\underset{\sim}{J}$ være vektoren med komponenter lik numrene på alle utvalgsområdene som trekkes. Disse områdene er de samme ved begge undersøkelsestidspunktene. På tidspunkt 1 trekkes 2 $n_{ir}(\underset{\sim}{J})$ trekkeenheter rent lotterisk fra utvalgsområde J_{ir} hvor S_1 består av de $n_{ir}(\underset{\sim}{J})-n'_{ir}(\underset{\sim}{J})$ første, S_2 består av de $n'_{ir}(\underset{\sim}{J})$ neste og S_3 består av de $n_{ir}(\underset{\sim}{J}) - n'_{ir}(\underset{\sim}{J})$ resterende trekkeenhetene, r = 1, 2, .., m_i, i = 1, 2, ... Numrene på de enhetene som trekkes ut fra område J_{ir} betegnes $K_{ir1}, K_{ir2}, \ldots$, og vi definerer indeksvariablene

$$U_{irs} = \begin{cases} 1 \text{ dersom } & s \leq n_{ir}(\underset{\sim}{J}) \\ 0 \text{ eller} & \end{cases}$$

og

$$V_{irs} = \begin{cases} 1 \text{ dersom } & s > n_{ir}(\underset{\sim}{J}) - n'_{ir}(\underset{\sim}{J}) \\ 0 \text{ ellers} & \end{cases}$$

Videre innfører vi

$$X_{irs} = a_i(J_{ir}, K_{irs}),\ Y_{irs} = c_i(J_{ir}, K_{irs}),$$

$$X_{ir} = \sum_s U_{irs} X_{irs},\ Y_{ir} = \sum_s V_{irs} Y_{irs},$$

$$\bar{X}_{ir} = X_{ir}/n_{ir}(\underset{\sim}{J}),\ \bar{Y}_{ir} = Y_{ir}/n_{ir}(\underset{\sim}{J}),$$

$$X_i = \sum_r X_{ir},\ Y_i = \sum_r Y_{ir},$$

$$\bar{X}_i = X_i/m_i,\ \bar{Y}_i = Y_i/m_i,$$

$$X = \sum_i X_i,\ Y = \Sigma Y_i,$$

$$b_i(\underset{\sim}{J}) = \pi_i b_i(\underset{\sim}{J}).$$

$b_i(\underset{\sim}{J})$ er utvalgsbrøken for stratum i og er den samme for alle utvalgsområdene innen stratumet, og $b(\underset{\sim}{J})$ er den totale utvalgsbrøk og er uavhengig av i. De estimatene vi tar utgangspunkt i er

$$\hat{a}_i = b^{-1}(\underset{\sim}{J})\, X_i,\ \hat{c}_i = b^{-1}(\underset{\sim}{J})\, Y_i$$

$$\hat{a} = \sum_i \hat{a}_i \text{ og } \hat{c} = \sum_i \hat{c}_i,$$

som er estimatorer for henholdsvis a_i, c_i, a og c. Estimatoren $\hat{a} - \hat{c}$ er derfor en estimator for endringen a - c, av samlet variabelverdi. Vi skal utlede variansen til $\hat{a} - \hat{c}$ og angi en estimator for denne variansen. Siden Var $\{\hat{a}-\hat{c}\}$ = Var $\hat{a}$ + Var $\hat{c}$ - 2 cov $\{\hat{a}, \hat{c}\}$ og de to første leddene er variansene til nivåtallestimatorene på de to tidspunktene (som er kjente), er det nok å finne cov $\{\hat{a}, \hat{c}\}$. Denne kovariansen kan splittes opp på samme måte som variansen til nivåtallestimatorene, nemlig i komponenter som er et uttrykk for kovariansen henholdsvis innen og mellom utvalgsområdene.

Setning 1: Kovariansen mellom utvalgsområdene er gitt ved

$$\operatorname{cov}\{E(\hat{a}|\underset{\sim}{J}),\ E(\hat{c}|\underset{\sim}{J})\} = \sum_i \pi_i^{-1}(M_i - m_i)\xi_i$$

hvor

$$\xi_i = \sum_j (a_i(j) - \bar{a}_i)(c_i(j) - \bar{c}_i)/(M_i - 1).$$

Bevis: Vi har at

$$E\{\hat{a}_i|\underset{\sim}{J} = \underset{\sim}{j}\} = \sum_r \pi_i^{-1} a_i(j_{ir})$$

og

$$E\{\hat{c}_i|\underset{\sim}{J} = \underset{\sim}{j}\} = \sum_r \pi_i^{-1} c_i(j_{ir}).$$

Ved å anvende Sverdrup 1963 (side 328) får vi umiddelbart at

$$\operatorname{cov}\{E(\hat{a}_i|\underset{\sim}{J}),\ E(\hat{c}_i|\underset{\sim}{J})\} = \pi_i^{-1}(M_i - m_i)\xi_i.$$

Samme resonnement som Hoem 1973 (side 12) gir videre

$$\operatorname{cov}\{E(\hat{a}_i|\underset{\sim}{J}),\ E(\hat{c}_j|\underset{\sim}{J})\} = 0 \text{ når } i \neq j.$$

Dermed er setningen bevist. □

Merknad: Det fremgår av setningen at kovariansen mellom utvalgsområdene er uavhengig av størrelsen på delutvalget S_2.

Lemma 2:

(i)
$$\sum_s E\{U_{irs}V_{irs}X_{irs}Y_{irs}|\underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= n'_{ir}(\underset{\sim}{j})(N_i(j_{ir})-1)\xi_i(j_{ir}) \,/\, N_i(j_{ir}) + n'_{ir}(\underset{\sim}{j})\bar{a}_i(j_{ir})\bar{c}_i(j_{ir})$$

(ii)
$$E\{X_{ir}Y_{ir}|\underset{\sim}{J} = \underset{\sim}{j}\} = n^2_{ir}(\underset{\sim}{j})\bar{a}_i(j_{ir})\bar{c}_i(j_{ir})$$

$$+ \xi_i(j_{ir})\{n'_{ir}(\underset{\sim}{j}) - n^2_{ir}(\underset{\sim}{j}) \,/\, N_i(j_{ir})\}$$

hvor

$$\xi_i(j) = \sum_k (a_i(j,k) - \bar{a}_i(j))(c_i(j,k) - \bar{c}_i(j))/(N_i(j) - 1).$$

Bevis: For $\underset{\sim}{J} = \underset{\sim}{j}$ har hver trekkeenhet innen utvalgsområde nr. j_{ir} sannsynligheten $1/N_i(j_{ir})$ for å bli trukket ut i en trekning. Herav følger at

$$E\{X_{irs}Y_{irs}|\underset{\sim}{J} = \underset{\sim}{j}\} = \sum_k a_i(j_{ir},k)c_i(j_{ir},k)\ Pr\{K_{irs} = k|\underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= \sum_k a_i(j_{ir},k)c_i(j_{ir},k) \ / \ N_i(j_{ir}).$$

Siden S_2 inneholder $n'_{ir}(\underset{\sim}{j})$ enheter fra utvalgsområde nr. j_{ir} er (i) dermed bevist. □

(ii): For $\underset{\sim}{J} = \underset{\sim}{j}$ har to ulike enheter innen utvalgsområde nr. j_{ir} sannsynligheten $1/N_i(j_{ir})\ (N_i(j_{ir}) - 1)$ for å bli trukket ut i en trekning. Siden $U_{irs}\ V_{irp}$ antar verdien 1 for $n^2_{ir}(\underset{\sim}{j}) - n'_{ir}(\underset{\sim}{j})$ verdier av (s,p) når $s \neq p$, blir

$$\sum_{s\neq p} E\{U_{irs}V_{irp}X_{irs}Y_{irp}|\underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= \{n^2_{ir}(\underset{\sim}{j}) - n'_{ir}(\underset{\sim}{j})\} \sum_{k\neq h} a_i(j_{ir},k)c_i(j_{ir},h)/N_i(j_{ir})(N_i(j_{ir}) - 1)$$

Sammen med (i) gir dette det ønskede resultat. □

Det neste resultatet følger nå umiddelbart fra lemma 2.

Setning 3: Kovariansen innen utvalgsområdene er gitt ved

$$E\ cov\{\hat{a}, \hat{c}|\underset{\sim}{J}\} =$$

$$E\{b^{-2}(\underset{\sim}{J}) \sum_{i,r} [n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J}) \ / \ N_i(J_{ir})]\ \xi_i(J_{ir})\}$$

Av setning 3 følger umiddelbart:

Korollar 4: Dersom $n'_{ir}(\underset{\sim}{J}) = n_{ir}(\underset{\sim}{J})/q$ hvor q er uavhengig av i og r for alle i og r er

$$E\ cov\{\hat{a}, \hat{c}|\underset{\sim}{J}\} =$$

$$E\{b^{-1}(\underset{\sim}{J}) \sum_{i,r} \pi_i^{-1}N_i(J_{ir})\ (q^{-1} - b_i(\underset{\sim}{J}))\ \xi_i(J_{ir})\}.$$

I Byråets rotasjonsplan er halvparten av et ordinært AKU utvalg felles ved en undersøkelse et kvartal og undersøkelsen samme kvartal året etter.

Ved å sette q = 2 får vi derfor kovariansen innen utvalgsområdene i dette tilfelle.

Merknad: Vi ser av setning 3 at dersom delutvalget S_2 blir lite nok skifter kovariansen innen utvalgsområdene fortegn.

Korollar 5: Når S_2 er tom er

$$\text{Cov}\{\hat{a}, \hat{c}\} = \sum_i \left[\pi_i^{-1}(M_i - m_i)\,\xi_i - \pi_i^{-1} \sum_j N_i(j)\,\xi_i(j)\right].$$

Forholdet mellom det negative leddet i korollar 5 og kovariansen innen utvalgsområdene er av størrelsesorden $2\pi_i^{-1}b(\underset{\sim}{J}) \simeq 12b(\underset{\sim}{J})$ når q = 2. I mange tilfeller kan vi derfor regne kovariansen tilnærmet lik kovarianskomponenten mellom utvalgsområdene når S_2 er tom.

3.2. Estimatorer

Vi skal nå behandle estimeringen av kovariansen mellom $\hat{a}$ og $\hat{c}$ i det tilfelle at delutvalget S_2 er ikke-tomt. De estimatorene vi kommer fram til kan imidlertid ikke brukes når S_2 er tom.

For å vise at estimatorene er forventningsrette trenger vi følgende hjelperesultater:

Lemma 6: Estimatoren

$$W_i = b^{-2}(\underset{\sim}{J}) \sum_r \frac{\{n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\}}{n'_{ir}(\underset{\sim}{J})\{1-n'_{ir}(\underset{\sim}{J})/n^2_{ir}(\underset{\sim}{J})\}} \sum_s U_{irs}V_{irs}(X_{irs}-\bar{X}_{ir})(Y_{irs}-\bar{Y}_{ir})$$

er forventningrett for

$$E\{b^{-2}(\underset{\sim}{J}) \sum_r \left[n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J}) / N_i(J_{ir})\right] \xi_i(J_{ir})\}.$$

Bevis: Ved å bruke lemma 2 får vi

$$E\{\sum_s U_{irs}V_{irs}(X_{irs}-\bar{X}_{ir})(Y_{irs}-\bar{Y}_{ir}) \mid \underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= E\{\sum_s U_{irs}V_{irs}X_{irs}Y_{irs} - n'_{ir}(\underset{\sim}{J})\,\bar{X}_{ir}\bar{Y}_{ir} \mid \underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= \xi_i(j_{ir})\, n'_{ir}(\underset{\sim}{j})\,\{1-n'_{ir}(\underset{\sim}{j})/n^2_{ir}(\underset{\sim}{j})\} \qquad \square\ .$$

Lemma 7:

$$E\{b^{-2}(\underset{\sim}{J})\sum_r (X_{ir}-\bar{X}_i)(Y_{ir}-\bar{Y}_i)\}\, m_i/(m_i-1)$$

$$= \sum_r E\{b^{-2}(J)\left[n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\right]\xi_i(J_{ir})\}$$

$$+ M_i\pi_i^{-1}\,\xi_i.$$

Bevis: Lemma 2 gir oss

$$\sum_r \operatorname{cov}\{b^{-1}(\underset{\sim}{J})X_{ir},\ b^{-1}(\underset{\sim}{J})Y_{ir}\} =$$

$$\sum_r E\{X_{ir}b^{-1}(\underset{\sim}{J}),\ Y_{ir}b^{-1}(\underset{\sim}{J})\} - a_ic_i/m_i$$

$$= \sum_r E\{b^{-2}(\underset{\sim}{J})\left[n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\right]\xi_i(J_{ir})\}$$

$$+ \pi_i^{-1}\,(M_i-1)\,\xi_i.$$

Sammen med setning 1 og setning 3 får vi derfor

$$\sum_{r,i} E\{(X_{ir} - \bar{X}_i)(Y_{ir} - \bar{Y}_i)\, b^{-2}(\underset{\sim}{J})\}m_i$$

$$= \sum_{i,r} \operatorname{cov}\{X_{ir}b^{-1}(\underset{\sim}{J}),\ Y_{ir}b^{-1}(\underset{\sim}{J})\}m_i$$

$$- \sum_i \operatorname{cov}\{X_i b^{-1}(\underset{\sim}{J}),\ Y_i b^{-1}(\underset{\sim}{J})\}$$

$$= \sum_i \Big[(m_i-1) \sum_r E\{b^{-2}(\underset{\sim}{J}) \big[n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\big] \xi_i(J_{ir})\}$$

$$+ M_i(M_i-1)\,\xi_i - \pi_i^{-1}(M_i-m_i)\,\xi_i\Big] .$$

Da de to siste leddene reduserer seg til $M_i\pi_i^{-1}(m_i-1)\,\xi_i$, er lemmaet bevist. □.

Setning 8: Estimatoren

$$H = b^{-2}(\underset{\sim}{J}) \sum_i (1-\pi_i)m_i \sum_r (X_{ir}-\bar{X}_i)(Y_{ir}-\bar{Y}_i)/(m_i-1)$$

$$+ \sum_i \pi_i W_i$$

er forventningsrett for $\operatorname{cov}\{\hat{a}, \hat{c}\}$.

Bevis: Av lemma 6 og lemma 7 følger det at

$$E\{H \mid \underset{\sim}{J} = \underset{\sim}{j}\} = E \sum_i \Big[\pi_i b^{-2}(\underset{\sim}{J}) \sum_r \big[n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\big] \xi_i(J_{ir})$$

$$+ (1-\pi_i)b^{-2}(\underset{\sim}{J}) \sum \{n'_{ir}(\underset{\sim}{J}) - n^2_{ir}(\underset{\sim}{J})/N_i(J_{ir})\} \xi_i(J_{ir})$$

$$+ (1-\pi_i)\,M_i\pi_i^{-1}\xi_i\Big] = E \operatorname{cov}\{\hat{a}, \hat{c} \mid \underset{\sim}{J}\}$$

$$+ \operatorname{cov}\{E(\hat{a} \mid \underset{\sim}{J}), E(\hat{c} \mid \underset{\sim}{J})\} = \operatorname{cov}\{\hat{a}, \hat{c}\} \quad \square.$$

Korollar 9: Dersom $n'_{ir}(\underset{\sim}{J}) = n_{ir}(\underset{\sim}{J})/q$ for alle i og r reduseres estimatoren H til

$$H = b^{-2}(\underset{\sim}{J}) \sum_i (1-m_i/M_i)m_i \sum_r (X_{ir}-\bar{X}_i)(Y_{ir}-\bar{Y}_i)/(m_i-1)$$

$$+ qb^{-1}(\underset{\sim}{J}) \sum_i \{1-qb_i(\underset{\sim}{J})\} \sum_{r,s} N_i(J_{ir})U_{irs}V_{irs}(X_{irs}-\bar{X}_{ir})(Y_{irs}-\bar{Y}_{ir})/(qn_{ir}(\underset{\sim}{J})-1).$$

3.3. Spesialtilfellet Oslo

I Oslo trekkes et utvalg på $2n_o(\underset{\sim}{J}) - n_o'(\underset{\sim}{J})$ trekkeenheter rent lotterisk fra en bestand på N_o trekkeenheter. Vi lar

$$a_o = \sum_k a_o(k) \text{ og } c_o = \sum c_o(k)$$

være samlet variabelverdi ved henholdsvis første og andre undersøkelsestidspunkt. Analogt til notasjonene i avsnittet foran lar vi K_{o1}, K_{o2}, .., være numrene på de enhetene som trekkes ut og vi innfører videre notasjonene

$$U_{os} = \begin{cases} 1 & \text{når } s \le n_o(\underset{\sim}{J}) \\ 0 & \text{ellers} \end{cases}$$

$$V_{os} = \begin{cases} 1 & \text{når } s > n_o(\underset{\sim}{J}) - n_o'(\underset{\sim}{J}) \\ 0 & \text{ellers} \end{cases},$$

$$X_{os} = a_o(K_{os}),\ Y_{os} = c_o(K_{os}),$$

$$X_o = \sum_s U_{os}X_{os},\ Y_o = \sum_s V_{os}Y_{os},$$

$$\bar{X}_o = X_o/n_o(\underset{\sim}{J}) \text{ og } \bar{Y}_o = Y_o/n_o(\underset{\sim}{J}).$$

Størrelsen på utvalget er valgt slik at

$$n_o(\underset{\sim}{J}) = b(\underset{\sim}{J})N_o.$$

Vår estimator for endringstallene innen Oslo er $\hat{a}_o - \hat{c}_o$ hvor

$$\hat{a}_o = b^{-1}(\underset{\sim}{J})\ X_o \text{ og } \hat{c}_o = b^{-1}(\underset{\sim}{J})\ Y_o.$$

Tilsvarende beviset for setning 3 og lemma 6 finner vi

Setning 10:

$$\text{cov}\{\hat{a}_o, \hat{c}_o\} = E\{b^{-2}(\underset{\sim}{J})(n_o'(\underset{\sim}{J}) - n_o^2(\underset{\sim}{J})/N_o)\}\xi_o$$

hvor

$$\xi_o = \sum_k (a_o(k) - \bar{a}_o)(c_o(k) - \bar{c}_o)/(N_o-1).$$

Setning 11: Kovariansen mellom $\hat{a}_o$ og $\hat{c}_o$ estimeres ved

$$W_o = b^{-2}(J)\,\frac{(n_o'(J) - n^2(J)/N_o)}{n_o'(J)\{1 - n_o'(J)/n_o^2(J)\}}\,\Sigma_s U_{os} V_{os}\,(X_{os}-\bar{X}_o)(Y_{os}-\bar{Y}_o).$$

Korollar 12: Dersom $n_o'(J) = n_o(J)/q$ reduseres W_o til

$$W_o = qb^{-1}(J)\{1-qb(J)\}\,N_o\,\Sigma_s U_{os} V_{os}(X_{os} - \bar{X}_o)(Y_{os} - \bar{Y}_o)/(qn_o(J) - 1).$$

3.4. Omforminger og tilnærmelser i beregningsprogrammet

Under programmeringen har en brukt tilnærmelsen $2n_{ir}(J)/(2n_{ir}(J) - 1) \simeq 1$ for alle i og r, og $2n_o/(2n_o - 1) \simeq 1$.

Dermed forenkles estimatoren H til

$$H = b^{-2}(J)\,\Sigma_i\,(1- \pi_i)\,m_i\,\Sigma_r\,(X_{ir} - \bar{X}_i)(Y_{ir} - \bar{Y}_i)/(m_i - 1)$$

$$+ \Sigma_i\,\pi_i^{-1}\,\{1 - 2b_i(J)\}\,\Sigma_{r,s}\,U_{irs} V_{irs}(X_{irs} - \bar{X}_{ir})(Y_{irs} - \bar{Y}_{ir}).$$

Videre har en benyttet at

$$\Sigma_j\,2(X_j - \bar{X})(Y_j - \bar{Y})$$

$$= \Sigma_j\,\{(X_j + Y_j - \bar{X} - \bar{Y})^2 - (X_j - \bar{X})^2 - (Y_j - \bar{Y})^2\}.$$

Referanser

Dagsvik, J. 1974: Etterhåndsstratifisering og estimering innen delbestander. Statistisk Sentralbyrå, Artikkel 66.

Des Raj 1968: Sampling theory. McCraw-Hill, London.

Hoem, J. M. 1973: Statistisk Sentralbyrås utvalgsundersøkelser: Elementer av det matematiske grunnlaget. Statistisk Sentralbyrå, Artikkel 58.

Laake, P. Forsén, L. og Sørlie, K. 1974: Orientering om bruk av Byråets program for estimering av gjennomsnitt og estimering av variansen til estimatoren i Byråets intervjuundersøkelser. Stensil. (PL/TF, 24/4-1974).

Rideng, A. 1970: Typifisering av kommuner i Norge, på grunnlag av næringsstrukturen 1968. Meddelelser fra Geografisk Institutt, Universitetet i Oslo. Kulturgeografisk serie, nr. 2.

Sverdrup, E. 1964: Lov og tilfeldighet. Bind I. Universitetsforlaget, Oslo.

Thomsen, I. 1973: Hvor oppdelt kan en offentliggjøre resultatene fra en intervjuundersøkelse? Statistisk Sentralbyrå, Metodehefte nr. 2 (IO 73/6).

Østerlund Petersen, S. 1972: Arbeidskraftundersøkelsene. Om endringer i tallene fra en undersøkelse til en annen. Statistisk Sentralbyrå, Metodehefte nr. 2. (IO 73/6).

Østerlund Petersen, S. 1974: Arbeidskraftundersøkelsene. Estimering av endringstall. Statistisk Sentralbyrå, Metodehefte nr. 14. (IO 74/34).