Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Dep, Oslo 1. Tlf.*(02) 41 38 20

IO 77/23 9. juni 1977

METODEHEFTE NR. 22

# Notat om varianser for endringstall som er estimert ved bruk av Byråets nye utvalgsplan

INNHOLD

# FORORD

## Metodehefter i serien Arbeidsnotater

I tilknytning til mange prosjekter i Statistisk Sentralbyrå utarbeides det mindre, upretensiøse notater for avklaring av spørsmål av metodisk interesse. Det kan dreie seg om utvalgsteknikk, alternative spørsmålsformuleringer, presentasjonsmetoder, begrepsavklaringer, diskusjon av "funn" i data, systemidéer eller andre temaer. Selv om mange slike notater bare har begrenset interesse i ettertid, vil det blant dem være noen som kunne fortjene å bli alminnelig tilgjengelig. Det kan også være nyttig å ha dem registrert sentralt slik at det blir lettere å få oversikt over det stoffet som foreligger, og lettere å referere tilbake til det. Byrået publiserer derfor leilighetsvis et passende antall notater av dette slaget samlet i metodehefter i serien Arbeidsnotater.

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

Forskningssjef Per Sevaldson er redaktør av metodeheftene. Førstekontorfullmektig Liv Hansen er redaksjonssekretær. Medarbeidere i Byrået som lager stoff som kan være aktuelt, bes sende dette til redaksjonen etter hvert som det blir ferdig. Retningslinjer for utformingen av inserater i metodeheftene finnes på side 46 til side 47 i Metodehefte nr. 9 (ANO IO 73/36).

# VARIANSER FOR ENDRINGSTALL SOM ER ESTIMERT VED BRUK AV BYRÅETS NYE UTVALGSPLAN

av

Hans Viggo Sæbø

INNHOLD

## 1. INNLEDNING

Byråets nye utvalgsplan ble tatt i bruk i 1975. Oppbyggingen av denne er beskrevet i Thomsen og Rideng (1974), mens Laake (1974) har gjort rede for estimering av nivåtall ved bruk av en slik plan. Her er det også funnet uttrykk for variansene til slike estimatorer, og det er foreslått en metode for å estimere disse variansene.

For å vurdere den nye utvalgsplanen har Sæbø (1976) beregnet eksakte varianser for estimerte sysselsettingstall. Grunnlaget for disse beregningene er data fra Folke- og Boligtellingen 1970 (FoB-70).

I dette notatet vil vi studere variansene til de endringstall som måles ved sammenlikning mellom forskjellige undersøkelser. Vi vil skille mellom undersøkelser med panelutvalg som arbeidskraftundersøkelsene (AKU), og undersøkelser med utvalg trukket uavhengig av hverandre i annet trinn.

Det trekkes ikke helt uavhengige utvalg etter Byråets utvalgsplan. Dette skyldes at trekkingen foregår i to trinn. I første trinn trekkes et område (som oftest kommune) i hvert av til sammen 102 strata. Denne trekkingen av utvalgsområder er foretatt en gang for alle, slik at bare annet trinns trekking av personer eller husholdninger fra disse områdene må gjøres for hver ny undersøkelse.

Dagsvik (1974) har funnet uttrykk for variansene til endringstall i AKU målt ved bruk av Byråets forrige utvalgsplan. Også i denne planen ble det først trukket utvalgsområder, men disse ble trukket med lik sannsynlighet innen hvert stratum. I den nye utvalgsplanen er det trukket ett område i hvert stratum med en sannsynlighet proporsjonal med folketallet.

I utledningene i dette notatet vil vi først anta at vi trekker et eller flere utvalgsområder fra hvert stratum med ulike sannsynligheter. En slik generalisering er også gjort av Laake (1974). Ved å sette inn for trekkesannsynlighetene og ta hensyn til at bare ett område blir trukket fra hvert stratum, vil vi forenkle de generelle uttrykkene og finne varianser for endringstall estimert ved bruk av Byråets nye utvalgsplan.

Et endringstall er differansen mellom to nivåtall. Dersom vi betrakter nivåtallene $a$ og $b$ med estimatorer $\hat{a}$ og $\hat{b}$, blir variansen til endringsestimatoren

$$\operatorname{var}(\hat{b}-\hat{a}) = \operatorname{var}\hat{a} + \operatorname{var}\hat{b} - 2\operatorname{cov}(\hat{a},\hat{b}). \qquad (1.1)$$

Skal vi finne uttrykk for denne variansen, må vi finne uttrykk for

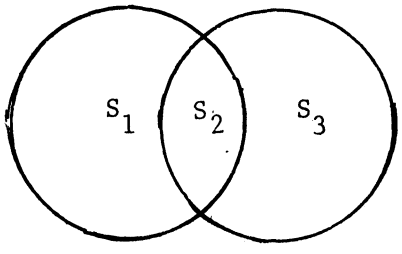
$\text{cov}(\hat{a},\hat{b})$, hvor $\hat{a}$ og $\hat{b}$ er målt i to forskjellige undersøkelser. Slike kovariansformler er utledet i avsnitt 3-5. Det er skilt mellom tilfellene med to utvalg som er trukket uavhengig i annet trinn, og to delvis overlappende utvalg som i AKU. Uttrykkene for variansen til endringsestimatorene blir satt opp i avsnitt 6. I sammendraget i avsnitt 7 er det satt opp forenklede uttrykk for variansen til endringstall.

Notasjonen i dette notatet følger stort sett notasjonen i Hoem (1973). Et notat som omhandler variansene til endringstall estimert i AKU, er tidligere skrevet av Østerlund Petersen (1973). Framstillingen der er forenklet, og en har ikke tatt hensyn til at utvalget trekkes i to trinn.

## 2. MODELL, NOTASJON OG DEFINISJONER

Vi vil i dette notatet betrakte to utvalgsundersøkelser. Utvalgene trekkes i to trinn etter samme utvalgsplan. I første trinn trekkes utvalgsområder fra strata som består av et eller flere slike områder. Utfallet av denne trekkingen er det samme i begge undersøkelser. Annet trinns trekking består i å trekke personer eller husholdninger fra de først uttrukne utvalgsområder.

Vi antar at utvalgene til de to undersøkelser trekkes med en på forhånd bestemt overlapping (panel). Utvalgene kan til sammen betraktes som 3 utvalg $S_1$, $S_2$ og $S_3$. $S_1$ er bare med i første undersøkelse, panelutvalget $S_2$ er med i begge undersøkelser, mens $S_3$ bare er med i siste undersøkelse. Utvalgene er vist skjematisk under.




Vi vil i det følgende anta at de 3 utvalgene er trukket uavhengig av hverandre i annet trinn, men fra de samme primære utvalgsområder.

Vi er interessert i å måle to nivåtall a og b, slik at disse estimeres i to forskjellige undersøkelser. Både a og b kan f.eks. være antall sysselsatte, men ved to forskjellige tidspunkter. Nivåtallet a måles i utvalget $(S_1 \cup S_2)$, mens b måles i $(S_2 \cup S_3)$.

Vi antar at j-te område i i-te stratum har $N_i(j)$ trekkenheter. Den k-te trekkenheten i område (i,j) har verdiene $a_i(j,k)$ og $b_i(j,k)$.

Vi lar

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

$$N = \sum_i N_i,$$

$$a_i(j) = \sum_k a_i(j,k),$$

$$a_i = \sum_j a_i(j),$$

og

$$a = \sum_i a_i.$$

Uttrykk for $b_i(j)$, $b_i$ og $b$ settes opp helt tilsvarende.

Vi betrakter første trinns trekking. La $\Pi_i(j)$ være sannsynligheten for at område j i stratum i blir trukket ut, og la $\Pi_i(j,k)$ være sannsynligheten for at både utvalgsområde j og k i stratum i skal bli trukket ut. I den nye utvalgsplanen er

$$\Pi_i(j,k) = 0 \quad \text{for } j \neq k,$$

og

$$\Pi_i(j,j) = \Pi_i(j) = N_i(j) / N_i. \tag{2.1}$$

Den siste likheten gjelder dersom trekkenhetene er personer.

La

$$I_{ij} = \begin{cases} 1 & \text{dersom område j i stratum i er i utvalget,} \\ 0 & \text{ellers.} \end{cases}$$

Vi har da

$$E\, I_{ij} = \pi_i(j),$$

$$\text{var } I_{ij} = \pi_i(j)(1 - \pi_i(j)),$$

og

$$\text{cov}(I_{ij}, I_{ik}) = \pi_i(j,k) - \pi_i(j)\,\pi_i(k) \quad \text{for } j \neq k \tag{2.2}$$

I Byråets utvalgsplan blir

$$\text{cov}(I_{ij}, I_{ik}) = -\pi_i(j)\,\pi_i(k) \quad \text{for } j \neq k \tag{2.3}$$

Dessuten har vi

$$\text{cov}(I_{ij}, I_{rk}) = 0 \quad \text{for} \quad i \neq r. \tag{2.4}$$

Dette følger av at områdene i to forskjellige strata trekkes uavhengig av hverandre. La $\underset{\sim}{J}$ betegne vektoren som består av numrene på alle i første trinn uttrukne områder.

Fra hvert uttrukne utvalgsområde trekkes i annet trinn et gitt antall enheter rent lotterisk. La utvalgene $(S_1 \cup S_2)$, $S_2$ og $(S_2 \cup S_3)$ bestå av henholdsvis $n_{ij}(\underset{\sim}{J})$, $n'_{ij}(\underset{\sim}{J})$ og $m_{ij}(\underset{\sim}{J})$ enheter fra område j i stratum i Vi definerer $n_{ij}(\underset{\sim}{J}) = n'_{ij}(\underset{\sim}{J}) = m_{ij}(\underset{\sim}{J}) = 0$ for $I_{ij} = 0$. Antall uttrukne enheter totalt betegnes tilsvarende med $n(\underset{\sim}{J})$, $n'(\underset{\sim}{J})$ og $m(\underset{\sim}{J})$.

Numrene på de enhetene som blir trukket ut fra område (i,j) i utvalget $(S_1 \cup S_2)$ betegner vi med $K_{ij1}$, $K_{ij2}$, ...., og vi lar

$$\begin{aligned} X_{ijs} &= a_i(j, K_{ijs}) \\ \text{og} \\ \bar{X}_{ij} &= \sum_s X_{ijs} / n_{ij}(\underset{\sim}{J}) \end{aligned} \tag{2.5}$$

Gjennomsnittsverdien $\bar{X}_{ij}$ er bare definert for $I_{ij} = 1$, og er en forventningsrett estimator for $a_i(j) / N_i(j)$. I panelet $S_2$ definerer vi $\bar{X}'_{ij}$ ved å summere over enhetene som er med her og dividere på $n'_{ij}(\underset{\sim}{J})$. Vi definerer tilsvarende estimatorene $\bar{Y}_{ij}$ i utvalget $(S_2 \cup S_3)$ og $\bar{Y}'_{ij}$ i $S_2$ Forventningsverdien for disse er $b_i(j) / N_i(j)$.

La

$$\begin{aligned} V_{ij} &= N_i(j)\, \bar{X}_{ij}\ , \\ \text{og} \\ W_{ij} &= N_i(j)\, \bar{Y}_{ij} \end{aligned} \tag{2.6}$$

Etter Laake (1974) vil forventningsrette estimatorer $\hat{a}$ og $\hat{b}$ for a og b være gitt ved

$$\begin{aligned} \hat{a} &= \sum_i \hat{a}_i = \sum_i \sum_j \{ I_{ij} V_{ij} / \pi_i(j) \}, \\ \text{og} \\ \hat{b} &= \sum_i \hat{b}_i = \sum_i \sum_j \{ I_{ij} W_{ij} / \pi_i(j) \}. \end{aligned} \tag{2.7}$$

Vi definerer videre

$$\lambda_i(j) = \frac{1}{N_i(j)-1} \sum_k (a_i(j,k) - \bar{a}_i(j))(b_i(j,k) - \bar{b}_i(j)), \qquad (2.8)$$

hvor $\bar{a}_i(j) = a_i(j) / N_i(j)$ og $\bar{b}_i(j) = b_i(j) / N_i(j)$.

La

$$\gamma_{ij}(\underset{\sim}{J}) = \frac{\lambda_i(j)}{n'_{ij}(\underset{\sim}{J})} \frac{N_i(j) - n'_{ij}(\underset{\sim}{J})}{N_i(j)} . \qquad (2.9)$$

For gjennomsnittsverdiene innen utvalgsområder og strata vil vi i dette notatet bruke betegnelsene

$$p_i(j) = \bar{a}_i(j),$$
$$q_i(j) = \bar{b}_i(j),$$

og

$$p_i = a_i/N_i,$$
$$q_i = b_i/N_i.$$

## 3. GENERELL UTLEDNING AV KOVARIANSFORMELEN

Vi vil finne $\mathrm{cov}(\hat{a},\hat{b})$, og tar utgangspunkt i (2.7). Etter (2.4) vil

$$\mathrm{cov}(\hat{a}_i, \hat{b}_r) = 0 \quad \text{for} \quad i \neq r. \qquad (3.1)$$

Vi finner

$$\mathrm{cov}(\hat{a}_i, \hat{b}_i) = \sum_j \sum_k \frac{1}{\pi_i(j)\,\pi_i(k)} \mathrm{cov}(I_{ij} V_{ij}, I_{ik} W_{ik}) .$$

Dette kan skrives

$$\mathrm{cov}(\hat{a}_i, \hat{b}_i) = \sum_j \frac{1}{(\pi_i(j))^2} \mathrm{cov}(I_{ij}V_{ij}, I_{ij} W_{ij}) +$$

$$+ \sum_{j \neq k} \sum \frac{1}{\pi_i(j)\,\pi_i(k)} \mathrm{cov}(I_{ij}V_{ij}, I_{ik} W_{ik}) \qquad (3.2)$$

Nå kan $V_{ij}$ etter (2.5) og (2.6) skrives

$$V_{ij} = N_i(j)\,\bar{X}_{ij} = N_i(j)\,\frac{1}{n_{ij}(\underset{\sim}{J})} \sum_{s\in(S_1 \cup S_2)} X_{ijs} =$$

$$= N_i(j)\,\frac{1}{n_{ij}(\underset{\sim}{J})} \left( \sum_{s\in S_1} X_{ijs} + \sum_{s\in S_2} X_{ijs} \right). \tag{3.3}$$

$W_{ij}$ kan tilsvarende spaltes opp i en del hvor vi summerer over $S_2$ og en del hvor vi summerer over $S_3$ Da $S_1$, $S_2$ og $S_3$ antas trukket uavhengig av hverandre, vil $V_{ij}$ og $W_{ij}$ bare avhenge av hverandre gjennom enhetene i det felles utvalget $S_2$ Nå er

$$\sum_{s\in S_2} X_{ijs} = n'_{ij}(\underset{\sim}{J})\,\bar{X}'_{ij},$$

og

$$\sum_{s\in S_2} Y_{ijs} = n'_{ij}(\underset{\sim}{J})\,\bar{Y}'_{ij}$$

Sammen med (3.3) gir dette

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij} \mid I_{ij} = 1, \underset{\sim}{J} = \underset{\sim}{j}) =$$

$$N_i^2(j)\,\frac{(n'_{ij}(\underset{\sim}{J}))^2}{n_{ij}(\underset{\sim}{J})\, m_{ij}(\underset{\sim}{J})} \operatorname{cov}(\bar{X}'_{ij}, \bar{Y}'_{ij} \mid I_{ij} = 1, \underset{\sim}{J} = \underset{\sim}{j}). \tag{3.4}$$

Vi definerer

$$d_{ij}(\underset{\sim}{J}) = \frac{(n'_{ij}(\underset{\sim}{J}))^2}{n_{ij}(\underset{\sim}{J})\, m_{ij}(\underset{\sim}{J})} \tag{3.5}$$

Etter Sverdrup (1973, side 366) vil

$$\operatorname{cov}(\bar{X}'_{ij}, \bar{Y}'_{ij} \mid I_{ij} = 1, \underset{\sim}{J} = \underset{\sim}{j}) = \gamma_{ij}(\underset{\sim}{J}) \tag{3.6}$$

Dette gir

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij} \mid I_{ij} = 1) = E\{N_i^2(j)\, d_{ij}(\underset{\sim}{J})\, \gamma_{ij}(\underset{\sim}{J}) \mid I_{ij} = 1\},$$

og

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij} \mid I_{ij}) = I_{ij}\, \xi_i(j), \tag{3.7}$$

hvor

$$\xi_i(j) = E\{N_i^2(j)\, d_{ij}(\underset{\sim}{J})\, \gamma_{ij}(\underset{\sim}{J}) \mid I_{ij} = 1\}. \tag{3.8}$$

Nå kan vi skrive

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij}) = E \operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij} \mid I_{ij}) +$$

$$+ \operatorname{cov}\{E(I_{ij} V_{ij} \mid I_{ij}), E(I_{ij} W_{ij} \mid I_{ij})\}. \tag{3.9}$$

Etter Laake (1974, side 3) vil

$$E(I_{ij} V_{ij} \mid I_{ij}) = I_{ij}\, a_i(j),$$

og tilsvarende får vi

$$E(I_{ij} W_{ij} \mid I_{ij}) = I_{ij}\, b_i(j) \tag{3.10}$$

Innsetting av (3.7) og (3.10) i (3.9) og bruk av (2.2) gir da

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ij} W_{ij}) = \pi_i(j)\, \xi_i(j) + \pi_i(j)\,(1-\pi_i(j))\, a_i(j)\, b_i(j). \tag{3.11}$$

Da $V_{ij}$ og $W_{ik}$ er uavhengige for $j \neq k$, vil i dette tilfellet

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ik} W_{ik} \mid I_{ij}, I_{ik}) = 0 \tag{3.12}$$

For $j \neq k$ kan vi sette opp en formel som svarer til (3.9). Innsetting i denne fra (3.10) og (3.12) og bruk av (2.2) gir nå

$$\operatorname{cov}(I_{ij} V_{ij}, I_{ik} W_{ik}) = (\pi_i(j,k) - \pi_i(j)\, \pi_i(k))\, a_i(j)\, b_i(k). \tag{3.13}$$

Innsetting i (3.2) og bruk av (3.1) gir

$$\operatorname{cov}(\hat{a}, \hat{b}) = \sum_i \sum_j \frac{1}{\pi_i(j)} [\xi_i(j) + (1 - \pi_i(j))\, a_i(j)\, b_i(j)] +$$

$$+ \sum_i \sum_{j \neq k} \sum \frac{\pi_i(j,k) - \pi_i(j)\, \pi_i(k)}{\pi_i(j)\, \pi_i(k)}\, a_i(j)\, b_i(k) =$$

$$= \sum_i \sum_j \sum_k \frac{\pi_i(j,k) - \pi_i(j)\,\pi_i(k)}{\pi_i(j)\,\pi_i(k)}\, a_i(j)\, b_i(k)$$

$$+ \sum_i \sum_j \xi_i(j) \,/\, \pi_i(j)\,, \qquad (3.14)$$

hvor $\xi_i(j)$ er gitt i (3.8). Dette resultatet gjelder generelt for en to-trinns utvalgsplan hvor det i første trinn trekkes et eller flere områder innen hvert stratum med ulike sannsynligheter. Dersom vi setter $b_i(j) = a_i(j)$ og $n_{ij}(J) = m_{ij}(J) = n'_{ij}(J)$ for alle $(i, j)$, får vi $\text{var}\,\hat{a}$ som vist av Laake (1974). Det første leddet kaller vi kovariansen <u>mellom</u> utvalgsområdene og det andre kovariansen <u>innen</u> utvalgsområdene. Denne terminologien er brukt av Dagsvik (1974), og betegnelsene er helt analoge til de tilsvarende betegnelser for variansen brukt av Sæbø (1976). I Byråets utvalgsplan gjelder (2.1). Innsetting herfra gir for kovariansen mellom utvalgsområdene

$$\text{cov}_2(\hat{a},\hat{b}) = \sum_i N_i \sum_j N_i(j)\,(p_i(j) - p_i)\,(q_i(j) - q_i) \qquad (3.15)$$

Formelen gjelder dersom $N_i(j)$ betegner antall <u>personer</u> i område $(i,j)$.

## 4. KOVARIANSER I TO UTVALG TRUKKET UAVHENGIG AV HVERANDRE I ANNET TRINN

Vi tenker oss at nivåtallene a og b estimeres i to undersøkelser med utvalg trukket uavhengig av hverandre i annet trinn. $N_i(j)$ regnes å være fast for alle $(i,j)$.

Formelen (3.14) gjelder generelt for to undersøkelser med et panel-utvalg $S_2$. I dette tilfellet har vi ikke noe panel, og vi kan sette $n'_{ij}(J) = 0$ for alle $(i,j)$. Dette fører til at $d_{ij}(J) = 0$ i (3.5), og vi får dermed også $\xi_i(j) = 0$ for alle $(i,j)$. Kovariansen innen utvalgsområdene faller bort, og vi har

$$\text{cov}\,(\hat{a},\hat{b}) = \sum_i \sum_j \sum_k \frac{\pi_i(j,k) - \pi_i(j)\,\pi_i(k)}{\pi_i(j)\,\pi_i(k)}\, a_i(j)\, b_i(k)\,, \qquad (4.1)$$

eller kovariansen mellom utvalgsområdene. Som vi har sett, kan denne skrives som i (3.15) ved bruk av Byråets utvalgsplan.

Forutsetningen for dette resultatet er at utvalgene $S_1$ og $S_3$ i de to undersøkelsene er trukket uavhengig av hverandre i annet trinn. I praksis innretter vi oss slik at personer som er trukket ut i en undersøkelse, ikke blir trukket ut i en annen. Dette fører til at kovariansen innen utvalgsområdene får et negativt bidrag. Når utvalgsstørrelsen avtar kan utvalgene betraktes som uavhengige (trukket med tilbakelegging), og dette bidraget går mot 0 Vi regner ikke med at denne kovariansen har noen betydning i våre undersøkelser.

## 5. KOVARIANSER I TO AKU-UTVALG

AKU-utvalgene trekkes etter en rotasjonsplan, slik at ca. halvparten av utvalget i et kvartal også er med neste kvartal. Vi vil finne et forenklet uttrykk for kovariansen mellom to estimatorer fra to på hverandre følgende undersøkelser. I Sæbø (1976) er det beregnet eksakte varianser til noen sysselsettingstall målt i AKU, og det er derfor naturlig å finne kovariansen mellom slike tall. Vi antar derfor at $a_i(j,k)$ og $b_i(j,k)$ for alle $(i,j,k)$ bare kan anta verdiene 0 og 1. En person får f.eks. verdien 1 dersom han registreres som sysselsatt og 0 ellers. Vi regner videre at utvalget er trukket som personutvalg. Vi kan nå skrive

$$\lambda_i(j) = \frac{1}{N_i(j)-1}\left\{\sum_k a_i(j,k)\, b_i(j,k) - \frac{a_i(j)\, b_i(j)}{N_i(j)}\right\} =$$

$$= \frac{N_i(j)}{N_i(j)-1}\left[u_i(j) - p_i(j)\, q_i(j)\right], \qquad (5.1)$$

hvor $u_i(j)$ er andelen i område $(i,j)$ av personer som har verdien 1 i begge undersøkelser.

Dersom vi f.eks. måler antall sysselsatte i to forskjellige kvartaler, er $u_i(j)$ andelen som var sysselsatt i begge kvartaler, mens $p_i(j)$ og $q_i(j)$ betegner andelene som var sysselsatt i første henholdsvis andre kvartal.

Vi regner at AKU-utvalgene er like store med gitt utvalgsstørrelse $n$, og at halvparten er med i begge kvartaler, slik at $n' = n/2$. Vi antar

videre at utvalgene er så små at vi kan sette

$$1 - \frac{n_{ij}(J)}{N_i(j)} = 1 \qquad (5.2)$$

i alle områder. Dette er samme tilnærmelse som er gjort av Sæbø (1976, side 4). Som her trekker vi selvveiende utvalg. I Byråets utvalgsplan oppnås dette ved å velge

$$n_{ij}(J) = \frac{N_i}{N} n \quad \text{for} \quad I_{ij} = 1. \qquad (5.3)$$

Vi skriver kovariansen som summen av kovarianser innen og mellom utvalgsområdene:

$$\text{cov}(\hat{a},\hat{b}) = \text{cov}_1(\hat{a},\hat{b}) + \text{cov}_2(\hat{a},\hat{b}),$$

hvor vi nå får

$$\text{cov}_1(\hat{a},\hat{b}) = \frac{N}{2n} \sum_i \sum_j N_i(j) \, (u_i(j) - p_i(j) \, q_i(j))$$

og

$$\text{cov}_2(\hat{a},\hat{b}) = \sum_i N_i \sum_j N_i(j) \, (p_i(j) - p_i) \, (q_i(j) - q_i) \qquad (5.4)$$

## 6. VARIANSEN TIL ESTIMERTE ENDRINGSTALL

Et nivåtall estimeres med $\hat{a}$ ved et tidspunkt og $\hat{b}$ ved et senere tidspunkt. Variansen til endringsestimatoren $(\hat{b} - \hat{a})$ blir da

$$\text{var}(\hat{b} - \hat{a}) = \text{var}\,\hat{a} + \text{var}\,\hat{b} - 2\,\text{cov}(\hat{a},\hat{b}). \qquad (6.1)$$

Vi kan skrive $\text{var}\,\hat{a}$ og $\text{var}\,\hat{b}$ som en sum av variansen innen og mellom utvalgsområdene. Etter Sæbø (1976) har vi for variansen mellom utvalgsområdene

$$\text{var}_2\,\hat{a} = \sum_i N_i \sum_j N_i(j) \, (p_i(j) - p_i)^2,$$

og

$$\text{var}_2\,\hat{b} = \sum_i N_i \sum_j N_i(j) \, (q_i(j) - q_i)^2.$$

Vi setter inn for $\text{cov}_2 \hat{a}$ fra (3.15) og får

$$\text{var}(\hat{b}-\hat{a}) = \text{var}_1 \hat{a} + \text{var}_1 \hat{b} - 2\text{cov}_1(\hat{a},\hat{b}) + \\ + \sum_i N_i \sum_j N_i(j) \left[(q_i(j) - p_i(j)) - (q_i - p_i)\right]^2 \qquad (6.2)$$

Variansen til endringsestimatoren kan altså også skrives som en sum av variansen innen og variansen mellom utvalgsområdene. Dersom de to undersøkelsene er trukket uavhengig av hverandre i annet trinn, er som vist i avsnitt 4 $\text{cov}_1(\hat{a},\hat{b}) = 0$ Vi får i dette tilfellet

$$\text{var}(\hat{b}-\hat{a}) = \text{var}_1 \hat{a} + \text{var}_1 \hat{b} + \sum_i N_i \sum_j N_i(j) \left(\Delta_i(j) - \Delta_i\right)^2, \qquad (6.3)$$

hvor

$$\Delta_i(j) = q_i(j) - p_i(j) \quad \text{og} \quad \Delta_i = q_i - p_i$$

Dersom endringen er lik i alle utvalgsområder i hvert stratum, altså $\Delta_i(j) = \Delta_i$ for alle (i,j), får vi

$$\text{var}(\hat{b}-\hat{a}) = \text{var}_1 \hat{a} + \text{var}_1 \hat{b} \qquad (6.4)$$

For å finne variansen til endringsestimatoren er det i dette tilfellet nok å estimere variansene til de respektive nivåtall innen utvalgsområdene. Det er vanskelig å si noe generelt om størrelsen på det siste leddet i (6.3), men vi kan anta at $|\Delta_i(j) - \Delta_i|$ gjennomgående er mindre enn $|p_i(j) - p_i|$ Dette gjelder i hvert fall i de strata hvor $\text{var}_2 \hat{a}_i$ bidrar vesentlig til $\text{var}\,\hat{a}_i$. I Sæbø (1976) er det beregnet eksakte varianser til estimerte sysselsettingstall på grunnlag av data fra FoB-70. I strata med relativt høy verdi for variansen mellom utvalgsområdene avviker $p_i(j)$ fra $p_i$ med inntil 5 - 10 %. For totalt antall sysselsatte er f.eks. $p_i \approx 0.50$, mens $p_i(j)$ varierer mellom 0.45 og 0.55. Ved moderate endringer, f.eks. $\Delta_i = 0.02$, vil $\Delta_i(j) \ll 0.10$. Vi regner derfor at

$$\sum_i N_i \sum_j N_i(j) \left(\Delta_i(j) - \Delta_i\right)^2 \ll \sum_i N_i \sum_j N_i(j) \left(p_i(j) - p_i\right)^2,$$

eller

$$\text{var}_2(b-\hat{a}) \ll \text{var}_2 \hat{a}. \qquad (6.5)$$

Etter Sæbø (1976) representerte $\text{var}_1 \hat{a}$ det største bidraget til $\text{var}\, \hat{a}$ for sysselsettingstall. Vi får derfor $\text{var}_2 (\hat{b} - \hat{a}) \ll \text{var}_1 \hat{a}$, og formelen (6.4) kan brukes for slike tall.

La oss til slutt sette opp et tilnærmet uttrykk for variansen til et endringstall målt i AKU mellom to kvartaler. Uttrykket gjelder under de samme forutsetninger som i avsnitt 5. Vi antar at vi kan se bort fra variansen til endringsestimatoren mellom utvalgsområdene. Etter det som er vist i avsnitt 5 og av Sæbø (1976), har vi da

$$\text{var}\,(\hat{b} - \hat{a}) = \text{var}_1 \hat{a} + \text{var}_1 \hat{b} - 2\,\text{cov}_1 (\hat{a}, \hat{b}) =$$

$$= \frac{N}{n} \sum_i \sum_j N_i(j)\, p_i(j)\,(1 - p_i(j)) + \frac{N}{n} \sum_i \sum_j N_i(j)\, q_i(j)\,(1 - q_i(j))$$

$$- \frac{N}{n} \sum_i \sum_j N_i(j)\ (u_i(j) - p_i(j)\, q_i(j) \qquad (6.6)$$

La oss anta at endringen og bruttostrømmene er små i forhold til de tilsvarende nivåtall. I AKU gjelder dette som regel alltid for nettoendringen, og Vannebo (1975) har vist at det samme kan antas for bruttostrømmene mellom to kvartaler for de aller fleste sysselsettingstall. Gruppen "arbeidssøkere" representerer et unntak, idet bare ca. 10 % av de som er med her i et kvartal er med i det neste.

La $\varepsilon_i(j)$ være maksimum av $(p_i(j) - u_i(j),\ q_i(j) - u_i(j))$.
Vi får da

$$\text{var}_1 \hat{b} - 2\,\text{cov}_1 (\hat{a}, \hat{b}) =$$

$$\frac{N}{n} \sum_i \sum_j N_i(j)\, [q_i(j)\,(1 - q_i(j)) - u_i(j) + p_i(j)\, q_i(j)] =$$

$$\frac{N}{n} \sum_i \sum_j N_i(j)\, [(q_i(j) - u_i(j)) + q_i(j)\,(p_i(j) - q_i(j))] \leq$$

$$\frac{N}{n} \sum_i \sum_j N_i(j)\, [\varepsilon_i(j) + q_i(j)\, \varepsilon_i(j)] \leq \frac{2N}{n} \sum_i \sum_j N_i(j)\ \varepsilon_i(j) \qquad (6.7)$$

Dette uttrykket går mot 0 når $\varepsilon_i(j) \to 0$

For

$$\sum_i \sum_j N_i(j)\ \varepsilon_i(j) \ll \frac{1}{2} \sum_i \sum_j N_i(j)\ p_i(j)\ (1 - p_i(j))$$

kan vi etter (6.7) sette

$$\mathrm{var}(\hat{b} - \hat{a}) \approx \mathrm{var}_1\, \hat{a} \tag{6.8}$$

Forutsatt små bruttostrømmer kan vi altså regne at variansen til et endringstall i AKU er lik variansen innen utvalgsområdene til et av nivåtallene.

Dagsvik (1975) har anslått autokorrelasjonen innen utvalgsområdene mellom to kvartaler for ulike variable. Denne korrelasjonen kan med vår notasjon skrives

$$\rho = \frac{\sum_i \sum_j N_i(j)\ (u_i(j) - p_i(j)\, q_i(j))}{\sqrt{\left[\sum_i \sum_j N_i(j)\ p_i(j)\,(1 - p_i(j))\right]\left[\sum_i \sum_j N_i(j)\ q_i(j)\ (1 - q_i(j))\right]}},$$

eller ved innsetting fra (5.4) som

$$\rho = \frac{2\ \mathrm{cov}_1\,(\hat{a},\hat{b})}{\sqrt{\mathrm{var}_1\,\hat{a}\ \ \mathrm{var}_1\,\hat{b}}} \tag{6.9}$$

For sysselsettingstall (unntatt "arbeidssøkere") har denne en verdi på 0.8 - 0.9, og dette viser at (6.8) underestimerer variansen til endringstallet med anslagsvis 10 - 20 %

## 7. SAMMENDRAG

Rammen for dette notatet er en to-trinns utvalgsplan hvor det i første trinn trekkes områder med ulike sannsynligheter. For en slik utvalgsplan er det utledet formler for kovarianser mellom nivåtall estimert i to undersøkelser. Utvalgene til begge undersøkelser regnes trukket fra de samme områder.

Formlene er forenklet slik at de gjelder for Byråets nye utvalgsplan, og det er funnet uttrykk for variansene til endringstall estimert ved bruk av denne planen.

Dersom endringstallet er målt i to utvalg trukket uavhengig av hver-

andre i annet trinn, kan vi som en "tommelfingerregel" bruke

$$\text{var}\,(\hat{b} - \hat{a}) \;=\; \text{var}_1\,\hat{a} + \text{var}_1\,\hat{b}\;, \tag{7.1}$$

hvor $\text{var}_1\,\hat{a}$ og $\text{var}_1\,\hat{b}$ er variansene til $\hat{a}$ og $\hat{b}$ innen utvalgsområdene. Disse variansene kan lett estimeres. Formelen gjelder dersom endringene er små i forhold til de respektive nivåtall. For endringer i sysselsettingstall målt i AKU kan vi bruke

$$\text{var}\,(\hat{b} - \hat{a}) \;\approx\; \text{var}_1\,\hat{a} \tag{7.2}$$

Dette anslaget er inntil 20 % for lavt, og det gjelder bare ved sammenlikning mellom to etterfølgende kvartaler, slik at halvparten av utvalget er felles i de to undersøkelsene.

REFERANSER


Dagsvik, J. (1974): "Variansestimering for nivåtallsestimater og endringstallsestimater ved Byråets Arbeidskraftundersøkelser." Statistisk Sentralbyrå, Arbeidsnotat (IO 74/50).

Dagsvik, J. (1975): "Presisjonsgevinst ved bruk av sammensatt estimering i Byråets arbeidskraftundersøkelser." Statistisk Sentralbyrå, Arbeidsnotat (IO 75/20).

Hoem, Jan M. (1973): "Statistisk Sentralbyrås utvalgsundersøkelser. Elementer av det matematiske grunnlaget." Statistisk Sentralbyrå, Artikkel nr. 58.

Laake, P. (1974): "Estimering av totaler med en to-trinns utvalgsplan der de primære utvalgsområder trekkes med ulik sannsynlighet i første trinn." Statistisk Sentralbyrå, Arbeidsnotat (IO 74/49).

Sverdrup, E. (1973): "Lov og tilfeldighet." Bind I. Universitetsforlaget, Oslo.

Sæbø, H.V. (1976): "Varianser og designeffekter for sysselsettingstall estimert ved bruk av Byråets nye utvalgsplan." Statistisk Sentralbyrå, Arbeidsnotat (IO 76/1).

Thomsen, Ib og Rideng, A. (1974): "Oversikt over arbeidet med ny utvalgsplan." Statistisk Sentralbyrå, Arbeidsnotat (IO 74/25).

Vannebo, O. (1975): "Bruttostrømmer i arbeidsmarkedet 1.kvartal 1972 - 1.kvartal 1974." Statistisk Sentralbyrå, Arbeidsnotat (IO 75/24).

Østerlund Petersen, S. (1973): "Arbeidskraftundersøkelsene. Om endringer i tallene fra en undersøkelse til en annen." (SØP/IH, 23/11-72.) Statistisk Sentralbyrå, Metodehefte 2 (IO 73/6).