Arbeidsnotater

STATISTISK SENTRALBYRÅ

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

) 74/36 20. august 1974

Sammensatt estimering ved roterende utvalg.
Matematisk - statistiske problemer knyttet til arbeidskraftundersøkelsene.

av Steinar Bjerve*

INNHOLD

* Jeg vil takke Jan M. Hoem, som satte i gang denne undersøkelsen og ga verdifull bistand under arbeidet.

Innhold

Forord

Manuskriptet til dette notatet ble i alt vesentlig utarbeidet våren 1971. Det ble lagt til side da forfatteren dro til University of California for å studere. Det er nå tatt fram igjen og sluttført. Det publiseres i Statistisk Sentralbyrås serie av Arbeidsnotater fordi en mener det vil komme til nytte under det fortsatte arbeidet med matematisk - statistiske problemer knyttet til arbeidskraftundersøkelsene.

Jan M. Hoem

1. Innleiing

De norske arbeidskraftundersøkelser foretas på utvalgsbasis etter rotasjonsprinsippet. Det vil si at en med jamne mellomrom (kvartalsvis) trekker utvalg slik at én og samme utvalgsenhet er med i utvalget et visst antall ganger etter et gitt rotasjonsmønster. Den byttes så ut med en ny, som trekkes tilfeldig. Med en slik utvalgsplan viser det seg at en under visse forutsetninger får et bedre grunnlag for estimering av arbeidskraftens bevegelse og sammensetning enn ved uavhengige og tilfeldige utvalg.

Befolkningen tenkes delt opp i grupper etter kjennetegn som for praktiske formål kan betraktes som konstante over det aktuelle tidsrom (f.eks. kjønn, alder, utdanning). Hvert individ tilhører en arbeidskraftkategori, eksempelvis "arbeidssøkere uten arbeidsinntekt", "i arbeidsstyrken", "fullt sysselsatt". Vi vil tenke oss at en tar for seg ei gruppe og en kategori, la oss si "menn" og "arbeidssøkende uten arbeidsinntekt". Anta at tallet på menn som observeres er det samme til enhver tid. For hver observasjon registrerer vi om mannen er arbeidssøker uten arbeidsinntekt eller ikke. For mann nummer j i undersøkelsen og for hvert undersøkelsestidspunkt t definerer vi da en stokastisk variabel X_{jt} som tar verdiene 1 og 0 ettersom han på tidspunktet t er arbeidssøkende uten arbeidsinntekt eller ikke. Kall samlingen av undersøkelsestidspunkter for T. Observasjonene $\{X_{jt};\ t\varepsilon T\}$ er en realisasjon av en stokastisk prosess, $\{X_t;\ t\varepsilon T\}$.

La oss tenke oss at en skal estimere følgende parametre:

i) Nivåer,

(1.1) $$\alpha_t = EX_t = P(X_t = 1).$$

ii) Kvartalsvise endringer,

(1.2) $$\delta_t = \alpha_t - \alpha_{t-1}.$$

iii) Årsgjennomsnitt,

(1.3) $$\beta_t = \frac{1}{4}(\alpha_t + \ldots + \alpha_{t+3}).$$

iv) Overgangssannsynligheter,

$$(1.4) \qquad p_{ij} = P(X_{t+1} = i \mid X_t = j);\ i, j \in \{0, 1\}.$$

Korrelasjonen mellom observasjoner på ulike tidspunkter vil i det følgende dels betraktes som kjent, dels inngå som en forstyrrende parameter. Korrelasjonen betegnes

$$\rho_{th} = \operatorname{cov}(X_t, X_{t+h})/(\operatorname{var} X_t \cdot \operatorname{var} X_{t+h})^{\frac{1}{2}}.$$

I dette notatet vil vi bruke varians- og effisiensberegninger til å studere egenskaper ved estimerings- og rotasjonsmetoder. Vi vil særlig bygge på resultater som finnes i litteraturreferansene [1] - [5]. Spesielt vil vi se på disse problemene når vi antar at X_t er en homogen markovkjede, altså slik at p_{ij} i (1.4) ovenfor er uavhengig av t. Vi antar hele tida at populasjonen er uendelig.

Den mest utførlige litteraturlista finnes i referanse [2].

2. Estimering ved roterende utvalg

(a) Notasjoner

Vi kaller realisasjonene av prosessen $\{X_t;\ t \in T\}$ for sampelstier og lar $\{X_{it};\ i \in I_t\}$ være de sampelstier som observeres på tidspunkt t. Vårt observasjonsmateriale vil da være

$$\{X_{it};\ i \in I_t \quad \text{og} \quad t \in T\}.$$

Klassen $\{I_t: t \in T\}$ bestemmer rotasjonsmønsteret i undersøkelsen. Sett[1] $n = \#(I_t)$ og la (for i og j = o, 1)

$$_tN_1 = \sum_{k \in I_t} X_{kt},$$

$$_tN_o = n - {}_tN_1,$$

$$_tN_{ij} = \#\{k \in I_t \cap I_{t+1}: X_{kt} = i,\ X_{k,t+1} = j\},$$

---

1) $\#(A)$ er antall elementer i mengden A.

$$ {}_tN_{i\cdot} = \sum_j {}_tN_{ij}, \quad {}_tN_{\cdot j} = \sum_i {}_tN_{ij}, $$

$$ {}_tN'_i = {}_tN_i - {}_tN_{i\cdot}, \quad {}_tN''_i = {}_tN_i - {}_{t-1}N_{\cdot i}. $$

Her er ${}_tN'_i$ og ${}_tN''_i$ antall stier som er i tilstand i på tidspunkt t og som ikke er observert på etterfølgende, henholdsvis foregående observasjonstidspunkt. Sett videre

$$ n_{rs} = \#\#(I_r \cap I_s), $$

$$ T = \{1, 2, \ldots, N\}, $$

$$ \lambda_t = n_{t,\,t+1}/n \quad \text{og} \quad \mu_t = 1 - \lambda_t. $$

(b) Rotasjonsmønstre

Vi kan ha en situasjon der μ_t er positiv bare på tidspunkter som er multipla av et helt tall c. c kaller vi da rotasjonsavstanden. Dersom[1] $\mu_{c\cdot\lfloor t/c \rfloor} \equiv \mu$, vil vi kalle μ for rotasjonsraten. Vi vil anta at hver sti bare observeres i en sekvens av påfølgende tidspunkter. I litteraturen er det også studert rotasjonsmønstre der hver sti får "gjenbesøk", det vil si at de observeres i ekvidistante sekvenser. En kan innføre gjenbesøksavstanden, n. I de månedlige stikkprøver som tas av det amerikanske Bureau of the Census' Current Population Survey (C.P.S.), er eksempelvis m = 12, μ = 1/4 og c = 1.

(c) Optimale estimatorer

Når korrelasjonsfunksjonen ρ_{th} er kjent, er den beste lineære forventningsrette estimator (Best Linear Unbiased estimator eller BLU-estimator) for vektoren $\alpha = (\alpha_1, \ldots, \alpha_N)'$ gitt ved

$$ (2.1) \qquad \hat{\alpha}' = (\hat{\alpha}_1, \ldots, \hat{\alpha}_N) = (\underset{\sim}{Y}\, \underset{\sim}{C}^{-1}\, \underset{\sim}{Y}')^{-1}\, \underset{\sim}{Y}\, \underset{\sim}{C}^{-1} \underset{\sim}{X}. $$

Her er $\underset{\sim}{X}$ en søylevektor sammensatt av observasjonene. $\underset{\sim}{X}$ har K elementer, hvor

$$ K = \sum_{t \in T} \#(I_t). $$

1) $\lfloor x \rfloor$ er det største heltall som ikke overstiger x.

$\underset{\sim}{C}$ er kovariansmatrisen til $\underset{\sim}{X}$. $\underset{\sim}{C}$ er gitt ved korrelasjonsfunksjonen $\rho_{t,h}$. $\underset{\sim}{Y}$ er en N x K matrise med binære elementer gitt ved at

$$(2.2) \quad E\,\underset{\sim}{X}' = \alpha'\,\underset{\sim}{Y}.$$

Vi har videre

$$(2.3) \quad \text{var}\,\hat{\alpha} = (\underset{\sim}{Y}\,\underset{\sim}{C}^{-1}\,\underset{\sim}{Y}')^{-1}.$$

BLU-estimatoren til en vilkårlig lineær - kombinasjon av α-ene, a'α, er gitt ved a'$\hat{\alpha}$. Beregning av $\hat{\alpha}$ ved hjelp av (2.1) når $\underset{\sim}{C}$ er kjent, innebærer invertering av K x K matrisen $\underset{\sim}{C}$. Siden K vanligvis er svært stor, blir dette umulig i praksis. I [6] utleder Prabhu - Ajgaonkar et rekursivt beregnbart uttrykk for $\hat{\alpha}_N$ når $\rho_{t,h}$ oppfyller

$$(2.4) \quad \rho_{t,h} = \rho_{t,t+1}\,\rho_{t+1,t+2} \cdots \rho_{h-1,h}.$$

Under forutsetningen $\rho_{t,\,t+h} = \rho_h$ for alle t får (2.4) formen

$$(2.5) \quad \rho_{t,\,t+h} = \rho^h, \text{ der } \rho = \rho_1.$$

Med en rotasjonsrate på μ, en konstant varians, var $X_t = \sigma^2$ og en rotasjonsavstand på 1, får rekursjonsformelen for beregning av estimatoren $\hat{\alpha}_N$, utseendet[1] (Des Raj, [1], side 161, og Dahmstrøm, [2], side 36.)

$$\overset{\nu}{\alpha}_k = A_k\,\bar{N}_1'' + (1 - A_k)\left[{}_{k-1}\bar{N}_{\cdot 1} + \rho(\overset{\nu}{\alpha}_{k-1} - {}_{k-1}\bar{N}_1)\right]$$

for k = 2, ..., N,

med $\overset{\nu}{\alpha}_1 = {}_1\bar{N}_1$,

$A_1 = \mu$

og

$$A_k = 1 - \{1 + \mu(1-\rho^2)/\lambda + A_{k-1}\,\rho\}^{-1} \text{ for } k = 2, \ldots, N.$$

---

1) En strek over en observator indikerer at den er lik den tilsvarende observator uten strek, dividert med det antall sampelstier den er utledet av.

Vi får da

(2.6) $\hat{\alpha}_N = \overset{\nu}{\alpha}_N$.

Følgende resultater utledes i [1] og [2]:

$$\text{var } \hat{\alpha}_N = (\sigma^2/n \cdot \mu) \cdot A_N,$$

$$A = \lim A_N = (2/\lambda \cdot \rho^2) \cdot \left[\{(1 - \rho^2)\left[1 - \rho^2(1 - 4\mu\lambda)\right]\}^{\frac{1}{2}} - (1 - \rho^2) \right].$$

(d) Optimal rotasjon

En ser lett at

$$\lim_N \text{var } \hat{\alpha}_N = \sigma^2 \cdot A/n \cdot \mu$$

minimeres med hensyn på μ for $\mu = \frac{1}{2}$ fordi denne verdien samtidig minimerer telleren og maksimerer nevneren. $\mu = \frac{1}{2}$ vil altså minimere var $\hat{\alpha}_N$ i det lange løp. Eckler [8] hevder i denne sammenheng at det betyr lite for var $\hat{\alpha}_N$ om A_k erstattes med A i rekursjonsformelen over. Optimal μ for estimering av β_t og δ_t er henholdsvis 0 og 1 når $\rho_{t,\,t+h} > 0$. Når en skal velge rotasjonsrate ved estimering av α, β og δ, må en derfor foreta en prioritering mellom parametrene.

(e) Andre korrelasjonsmønstre

Rao og Graham foreslår i [9] et annet korrelasjonsmønster som kan være mer realistisk enn (2.5) når det er sesongvariasjoner, slik vi har i arbeidskraftdata. De setter

(2.7) $\rho_{4 \cdot i + j} = \rho_1^{\,j} \cdot \rho_2^{\,i}$ for $i = 0, 1, \ldots$ og $j = 0, 1, 2, 3$.

Rao og Graham sier at variansberegninger for korrelasjonsmønsteret (2.7) er foretatt, men de er ennå ikke publisert. Ideen med (2.7) er at ρ_1 skal representere den årvise korrelasjon og ρ_2 den kvartalsvise.

(f) Estimatorer

Dersom ρ er kjent, kan

$$\hat{\alpha}_t,\ \hat{\delta}_t = \hat{\alpha}_t - \hat{\alpha}_{t+1} \text{ og } \hat{\beta}_t = \frac{1}{4}(\hat{\alpha}_t + \cdots + \hat{\alpha}_{t+3})$$

brukes som estimatorer for α_t, β_t og δ_t. Korrelasjonene (2.5) innsettes da i matrisen $\underset{\sim}{C}$. Disse estimatorene er imidlertid ikke optimale, men Dahmstrøm [2, s. 38] gjengir beregninger som Patterson [7] har gjort og som tyder på at det er lite å vinne i varians ved å finne BLU-estimatorene for disse parametrene når $\hat{\alpha}_t$ er kjent. Vi forutsetter her $\mu_t = \mu$ fast.

Vanligvis vil ρ være ukjent og $\hat{\alpha}_t$ vil da ikke være noen estimator. $\hat{\alpha}_t$ vil derfor først og fremst tjene som sammenlikningsgrunnlag for andre estimatorer. Vi vil da kunne si noe om hvor mye som kan vinnes ved å bruke $\hat{\alpha}_t$ når ρ er kjent. Det ville være interessant å studere egenskapene til den estimatoren en får dersom ρ estimeres og verdien innsettes i uttrykket for $\hat{\alpha}_t$.

Den estimatoren som vil bli vurdert her, er CPS - estimatoren[1] (se e.g. Dahmstrøm):

$$(2.8)\quad \tilde{\alpha}_k = C\,(\tilde{\alpha}_{k-1} + {}_{k-1}\bar{N}_{\cdot 1} - {}_{k-1}\bar{N}_{1\cdot}) + (1-C)\,{}_k\bar{N}_1\,,$$

hvor $\tilde{\alpha}_1 = {}_1\bar{N}_1$. C er en konstant som fastsettes etter nærmere studier av $\tilde{\alpha}_k$.

En estimator som er foreslått av Waksburg og Pearl i [4] er

$$(2.9)\quad {}^{s}\alpha_k = C({}^{s}\alpha_{k-1} + {}_{k-1}\bar{N}_{\cdot 1} - {}_{k-1}\bar{N}_{1\cdot}) + D\,(\hat{\alpha}_{k-4} + {}_{k-4}\bar{M}_{\cdot 1} - {}_{k-4}\bar{M}_{1\cdot}) + (1-C-D)\,{}_k\bar{N}_1,$$

hvor

$${}_t M_{ij} = \#\{I_t \cap I_{t+4} : X_t = i,\ X_{t+4} = j\},$$

$${}_t M_{i\cdot} = {}_t M_{i1} + {}_t M_{i0}, \quad = \quad {}_t M_{\cdot j} = {}_t M_{\cdot j} + {}_t M_{0j},$$

$${}_t\bar{M}_{i\cdot} = \frac{1}{n_{t,\,t+4}}\,{}_t M_{i\cdot} \text{ og } {}_t\bar{M}_{\cdot j} = \frac{1}{n_{t,\,t+4}}\,{}_t M_{\cdot j}.$$

---

1) CPS = Central Population Survey.

C og D er konstanter. Denne siste estimatoren er foreslått på intuitivt grunnlag for å dra nytte av eventuell høy år - til - år korrelasjon. Det vil si at ρ_4 er stor i forhold til ρ_1^4. Waksburg og Pearl [4] foreslår dessuten en modifisering av (2.8) for å få en estimator som svarer mer til (2.6). Denne estimatoren kalles A- K estimatoren. Den er gitt ved

$$(2.10)\quad \overset{\circ}{\alpha}_k = \tilde{\alpha}_k + A(\mu \cdot {}_k\overline{N}_1'' - \lambda \cdot {}_{k-1}\overline{N}_{\cdot 1}).$$

A er en konstant.

(g) Generelle betraktninger.

Estimatorer for β og δ kan også avledes av estimatorene for α_N gitt i (2.8) og (2.9). Disse avledete estimatorene vil også bli vurdert her. For å identifisere estimatorene, vil de bli utstyrt med de samme toppsymboler som de estimatorene de er avledet fra. Den enkle (usammensatte) estimator for α_t er

$$\overline{\alpha}_t = \frac{1}{n}\, {}_tN_1 = {}_t\overline{N}_1.$$

I Rao og Graham [9] er omfattende effisiensberegninger foretatt for $\tilde{\alpha}_t$ og $\tilde{\delta}_t$ under rotasjon med gjenbesøk. De optimerer effisiensen m.h.p. C og tabeller optimal effisiens relativt til $\overline{\alpha}_t$ (henholdsvis $\overline{\delta}_t$) for ulike verdier av μ, m og C. I Waksburg og Pearl [4] og Gurney og Daly [3] er noen enkle effisiensberegninger for henholdsvis "$\sim$"-estimatorene og "s"-estimatorene foretatt på grunnlag av (2.3). Det er dessverre ikke gjort forsøk på å undersøke robusthet m.h.p. avvik i korrelasjonsmønster for disse estimatorene.

Av de estimatorene som er nevnt, er bare (2.8) i bruk. I C.P.S. i U.S.A. brukes (2.8) med C = 0.5 for alle grupper og kategorier og alle parametre. I Sverige brukes (2.8) med

$$C = \begin{cases} 0.5 & \text{for kategorien "i arbeid"} \\ 0.0 & \text{" "arbeidsløs"} \\ 0.7 & \text{" "i Arbeidskraften"} \end{cases}$$

C varieres ikke fra parameter til parameter. (Se Dahmstrøm og Malmberg [10] og [12].)

Et argument som fører til den svake grad av fleksibilitet i valg av C i eksemplet over, er kravet om samsvar de publiserte tabeller imellom. Som eksempel kan nevnes en to-dimensjonal tabell med arbeidskraftkategori kolonnevis og aldersgrupper linjevis. Med svenskenes system vil kolonnesummene og estimatet for den tilhørende arbeidskraftkategori falle sammen, men linjesummene vil ikke nødvendigvis falle sammen med populasjonsstørrelsen i den tilhørende aldersgruppe. I de amerikanske tabeller vil både linje- og kolonnesummer "stemme".

Et liknende problem gjelder for valg av C ved estimering av f.eks. α og δ. De optimale verdier for C er ulike ved estimering av α og δ (størst for δ). Dersom C-ene velges ulike, vil estimatoren for endringer i f.eks. sysselsettingen fra forrige kvartal ikke stemme med differansen mellom de publiserte tall for nivåer. Det er imidlertid mulig ved forskjellige valg av C, å revidere forrige kvartals estimater ved hjelp av estimatet for δ.

I det foregående er det ikke tatt hensyn til at utvalget er et to-trinns utvalg med forhånds- og etterhåndsstratifisering.

Under vurderingen av resultatene foran må en huske på at variansen er satt sammen av variansen innen de primære utvalgsenheter (p.u.e.) og variansen mellom primærenhetene. Dersom utvalgsstørrelsen i hver p.u.e. skal bevares må rotasjonen foregå innen p.u.e.. Men da er det bare variansen innen p.u.e. som reduseres ved bruk av sammensatt estimering. Det er da viktig å vite noe om hva det er som bidrar mest til den totale varians, variasjonen innen p.u.e. eller variasjonen mellom dem. Dette er et emne som bør studeres nærmere.

Et annet viktig moment er at det er nær sammenheng mellom flytting og skifte av arbeidskraftkategori, f.eks. ved at den som mister arbeid flytter for å få nytt arbeid. Dersom adresse som vanlig skal være trekkeenhet, vil den nye som flytter inn, tilsvarende ofte ha arbeid. På denne måten vil vi risikere å underestimere arbeidsløsheten. Folk som flytter, bør følges opp. Dersom dette ikke er mulig, er det kanskje best å regne enheten som tapt fra undersøkelsen.

Til slutt vil jeg nevne problemet med at individene ikke oppfører seg etter forutsetningen om konstant klassetilhøreghet. Et individ kan for eksempel gå fra en aldersgruppe til en annen under undersøkelsen. Anta dette skjer mellom tidspunktene k og k - 1 og at vi skal estimere arbeidskraftnivået i den siste aldersgruppen. Individet vil da være med i ${}_k\bar{N}_{1\cdot}$ og i ${}_{k-1}\bar{N}_{\cdot 1}$ men ikke i de andre komponentene av (2.8). En måte å rette dette på ville være å beregne alle foregående observatorer på nytt,

under den klassetilhørighet som gjelder på tidspunkt k. Dermed ville korrelasjons-egenskapene bevares. Omkostningene forbyr sannsynligvis en slik framgangsmåte. En ville da under estimeringen måtte ha tilgjengelig rådata for alle undersøkelsene, mens en ellers kan nøye seg med å ha estimatorene for forrige tidspunkt samt rådata fra forrige og nåværende tidspunkt tilgjengelige under beregningene. Svenskene har omtalt dette i [11]. Det er altså svært viktig å gruppere slik at det blir få individer som går fra en gruppe til den annen under undersøkelsen. En kunne la individet under hele undersøkelsen tilhøre den aldersgruppen det tilhørte ved inntredelsen. Dette vil imidlertid skape skjevhet i estimatene for siste tidspunkt.

3. Varians- og effisiens-beregninger for estimatorer for nivåer og endringer

(a) Innledende merknader

Jeg vil i dette avsnittet søke å vise hva som kan vinnes ved bruk av roterende utvalg og sammensatt estimering. Sammensatte estimatorer i motsetning til enkle estimatorer, brukes som fellesbetegnelse på estimatorer som gjør bruk av observasjoner fra flere tidspunkter ved estimering av parameterverdiene for et gitt tidspunkt. Gevinsten ved sammensatt estimering og ved rotasjon vil jeg belyse først og fremst ved tabeller over effisienser. Tallene er for en stor del hentet fra tilsvarende tabeller i [2], [3], [4] og [5]. Der hvor matematikken ikke blir for komplisert, vil jeg utlede formler for effisienser. Dette vil gjelde estimatorer hvor observasjoner fra inntil to tidspunkter er med.

(b) Enkel estimering

La oss først se på gevinsten ved rotasjon uten bruk av sammensatt estimering. En kan her åpenbart bare vinne noe ved estimering av endringer, δ. Vi har

$$\overline{\delta}_t = {}_t\overline{N}_1 - {}_{t-1}\overline{N}_1$$

og

$$(3.1)\quad \operatorname{var} \bar{\delta}_t = \operatorname{var}\, {}_t\bar{N}_1 + \operatorname{var}\, {}_{t-1}\bar{N}_1 - 2 \operatorname{cov}({}_t\bar{N}_1,\ {}_{t-1}\bar{N}_t)$$

$$= \frac{2\sigma^2}{n} - \frac{2(1-\mu)\rho\sigma^2}{n}$$

$$= \frac{2\sigma^2}{n}(1 - (1-\mu)\rho).$$

hvor μ er rotasjonshastigheten, $\sigma^2 = \operatorname{var} X_t$ og $\rho = \sigma^{-2} \operatorname{cov}(X_t, X_{t-1})$. Vi vil anta at (2.5) er oppfylt. Sett

$$\bar{\bar{\alpha}}_t = \bar{\alpha}_t\big|_{\mu=1} \text{ og } \bar{\bar{\delta}}_t = \bar{\bar{\alpha}}_t = \bar{\bar{\alpha}}_t - \bar{\bar{\alpha}}_{t-1}.$$

$\bar{\bar{\delta}}_t$ blir estimatoren for δ_t ved uavhengige utvalg. Vi får

$$(3.2)\quad e(\bar{\delta}_t, \bar{\bar{\delta}}_t) = \frac{\operatorname{var} \bar{\delta}_t}{\operatorname{var} \bar{\bar{\delta}}_t} = 1 - (1-\mu)\rho.$$

Vi ser at ved små μ og store ρ er vinsten ved rotasjon betydelig. Vi kan samtidig se at vi får et tilsvarende varianstap ved estimering av $\beta_t = \frac{\alpha_t + \alpha_{t+1}}{2}$ under rotasjon:

$$\operatorname{var} \bar{\beta}_t = \operatorname{var} \tfrac{1}{2}({}_t\bar{N}_1 + {}_{t-1}\bar{N}_1) = \frac{\sigma^2}{2n}(1 + (1-\mu)\rho)$$

og

$$(3.3)\quad e(\bar{\beta}_t, \bar{\bar{\beta}}_t) = (1 + (1-\mu)\rho),$$

hvor

$$\bar{\bar{\beta}}_t = \tfrac{1}{2}(\bar{\bar{\alpha}}_t + \bar{\bar{\alpha}}_{t+1}).$$

(c) Sammensatt estimering

La oss se på det enkleste av de sammensatte estimatorer, nemlig C.P.S-estimatoren for to tidspunkter,

$$(3.4)\quad \tilde{\alpha}_t = C({}_{k-1}\bar{N}_1 + {}_{k-1}\bar{N}_{\cdot 1} - {}_{k-1}\bar{N}_1) + (1-C)\,{}_k\bar{N}_1 .$$

Vi får, med $\lambda = 1 - \mu$:

$$(3.5)\quad \operatorname{var} \tilde{\alpha}_t = \Big[C^2\Big(\frac{1}{n} + \frac{1}{n\lambda} + \frac{1}{n\lambda} + \frac{\rho}{n} - \frac{1}{n} - \frac{\rho}{n\lambda}\Big)$$

$$+ (1 - C)^2 \cdot \frac{1}{n} + C(1 - C)\Big(\frac{\lambda\rho}{n} + \frac{1}{n} - \frac{\rho}{n}\Big)\Big]\sigma^2$$

$$= \frac{\sigma^2}{n}\Big[C^2\lambda^{-1}(2 - \rho\mu) + (1 - C)(1 - C\rho\mu)\Big].$$

CPS-estimatoren for endringer blir:

$$(3.6)\quad \tilde{\delta}_t = \tilde{\alpha}_t - \tilde{\alpha}_{t-1} = C({}_{t-1}\bar{N}_{\cdot 1} - {}_{t-1}\bar{N}_{1\cdot}) + (1 - C)({}_t\bar{N}_1 - {}_{t-1}\bar{N}_1).$$

Nå merker vi oss at

$${}_t\bar{N}_1 = \lambda\ {}_{t-1}\bar{N}_{\cdot 1} + \mu\ {}_t\bar{N}'_1,$$

og at ${}_t\bar{N}'_1$ og ${}_t\bar{N}_{\cdot 1}$ er uavhengige. Tilsvarende spaltes ${}_{t-1}\bar{N}_1$ opp i uavhengige addender:

$${}_{t-1}\bar{N}_1 = \lambda\ {}_{t-1}\bar{N}_{1\cdot} + \mu\ {}_{t-1}\bar{N}''_1.$$

Vi får

$$(3.7)\quad \operatorname{var} \tilde{\delta}_t = \operatorname{var}\Big(\frac{1}{n}(C + \lambda(1 - C))\ ({}_{t-1}\bar{N}_{\cdot 1} - {}_{t-1}\bar{N}_{1\cdot})$$

$$+ \frac{1}{n}(1 - C)\ ({}_t\bar{N}'_1 - {}_{t-1}\bar{N}''_1)\Big)$$

$$= \frac{2\sigma^2}{\lambda \cdot n}\Big[(1 - (1 - C)\mu)^2(1 - \rho) + (1 - C)^2\mu \cdot \lambda\Big].$$

Når en skal vurdere en estimator, er det viktig både å se på hvor mye en vinner og hvor mye en kan oppnå ved å forbedre den. Etter [1], s. 157 og 158, har vi

$$(3.8)\quad \operatorname{var} \hat{\alpha}_t = \frac{1 - \rho^2\mu}{1 - \rho^2\mu^2} \cdot \frac{\sigma^2}{n},$$

$$(3.9)\quad \operatorname{var} \hat{\delta}_t = \frac{1 - \rho}{1 - \mu\rho} \cdot \frac{2\sigma^2}{n}.$$

Effisiensene av de optimale relativt til CPS-estimatene blir:

$$(3.10)\quad e(\hat{\alpha}, \tilde{\alpha}) = \frac{(1 - \rho^2\mu)\,\lambda}{(1 - \rho^2\mu^2)\left[C^2(2 - \rho\mu) + (1 - C)\,\lambda\,(1 - (\rho\mu)\right]},$$

$$(3.11)\quad e(\hat{\delta}, \tilde{\delta}) = \frac{\lambda(1 - \rho)}{(1 - \mu\rho)\left[(\lambda + \mu C)^2\,(1 - \rho) + (1 - C)^2\,\mu\lambda\right]}.$$

For å se hvor mye som kan oppnås i forhold til de enkle estimatorer, ser vi på effisiensen til de optimale relativt til de enkle. Av (3.8) og (3.9) får vi (var $\bar{\alpha} = \sigma^2/n$):

$$(3.12)\quad e(\hat{\alpha}, \bar{\alpha}) = \frac{1 - \rho^2\mu}{1 - \rho^2\mu^2},$$

$$(3.13)\quad e(\hat{\delta}, \bar{\delta}) = \frac{1 - \rho}{(1 - \mu\rho)\,(1 - \lambda\rho)} = \left(1 + \frac{\lambda\mu\rho}{1 - \rho}\right)^{-1}.$$

Av (3.12) kan vi finne den optimale rotasjonsrate ved estimering av α (se [1] s. 157):

$$(3.14)\quad \mu_{opt} = \left(1 + (1 - \rho^2)^{\frac{1}{2}}\right)^{-1} \geq 1/2.$$

Det som vinnes ved å bruke CPS-estimatorer istedenfor enkle, uttrykkes ved

$$(3.14a)\quad e(\tilde{\alpha}, \alpha) = \operatorname{var} \tilde{\alpha} \cdot \frac{n}{\sigma^2},$$

$$(3.14b)\quad e(\tilde{\delta}, \delta) = \operatorname{var} \tilde{\delta} \cdot \frac{n}{2\,\sigma^2(1 - \lambda\rho)}.$$

Det viser seg at ved optimalt valg av C i C.P.S.-estimatoren, $\tilde{\delta}$, så blir $\tilde{\delta} = \hat{\delta}$. Er vi altså heldig med valg av C, vil estimatoren $\tilde{\delta}$ ikke kunne forbedres når vi begrenser oss til bruk av data fra to observasjonstidspunkter. Sett $C' = 1 - C$:

$$\frac{d}{dc'}\left[\operatorname{var} \tilde{\delta}\right] = \left[-2\mu(1 - C'\mu)\,(1 - \rho) + 2\,C'\mu\lambda\right]\frac{2\sigma^2}{n}.$$

Setter vi dette uttrykket lik o og løser m h p C får vi:

$$C'_{opt} = \frac{1 - \rho}{\mu(1 - \rho) + \lambda},$$

eller

$$(3.15)\quad C_{opt} = 1 - C'_{opt} = \frac{\lambda\rho}{\mu(1-\rho)+\lambda}\,.$$

Setter vi inn i uttrykket for var $\tilde{\delta}_t$, så finner vi at

$$(3.16)\quad \inf_c (\text{var}\ \tilde{\delta}) = \text{var}\ \hat{\delta},$$

noe som bekrefter påstanden over. Tabell 1 og 3 gir effisienser for CPS-estimatorene, $\tilde{\alpha}$ og $\tilde{\delta}$, relativt til $\hat{\alpha}$, $\hat{\delta}$, $\bar{\alpha}$ og $\bar{\delta}$ for ulike verdier av C, μ og ρ. Ved å lese loddrett, vil robusthet overfor variasjon i ρ for ulike C og μ kunne avleses. I Rao og Graham [5], s. 498 og s. 503, står formler for var $\tilde{\alpha}$ og var $\tilde{\delta}$ når alle observasjoner fra en løpende undersøkelse brukes. $\tilde{\alpha}$ og $\tilde{\delta}$ beregnes da etter formel (2.8). I tabell 1 og 3 er også $e(\tilde{\alpha}, \bar{\alpha})$ og $e(\tilde{\delta}, \bar{\delta})$ beregnet på grunnlag av Rao og Grahams formler. Disse tallene kan også brukes til å se hvor mye som vinnes ved å ta med observasjoner fra fler enn to tidspunkter. Anta at C velges separat for estimering av α og δ. Som nevnt i avsnitt 2 (g), vil en da kunne revidere foregående estimat for α ved

$$(3.17)\quad \tilde{\alpha}_{t-1} = \tilde{\alpha}_t - \tilde{\delta}_t.$$

Det foreligger lite om kvaliteten av estimatorene med toppsymbol o og s. (Se (2.9) og (2.10).) Tabell 5 i [3] gir effisienser for $\overset{o}{\alpha}$ og $\overset{o}{\delta}$ relativt til $\bar{\alpha}$ og $\bar{\delta}$. Tabell 2 i [4] viser at det er bare for estimering av årlige endringer en vinner noe særlig ved bruk av estimatoren gitt i (2.6). I disse tabellene er følgende korrelajsonsmønster antatt å gjelder

	Korrelasjon mellom påfølgende observasjoner	Årlig korrelasjon
"Civilian labour force"	0.8	0.7
"Agricultural employment"	0.9	0.7
"Unemployed"	0.5	0.3

I [12], tabell 1, side 9, går det fram at den største gevinsten de regner med å oppnå ved bruk av sammensatt estimering, er ved estimering av kvartalsvise endringer i arbeidskraften. Der er gevinsten 11 prosent i forhold til $\overline{\delta}$. For de to første norske undersøkelsene (i oktober og desember 1971) ville det vært lite tjenlig å bruke sammensatt estimering av parametrene for desember. En vet nemlig lite eller ingenting om korrelasjonsmønstret for observasjonene. Det er imidlertid viktig at det blir lagt vekt på estimeringen av korrelasjonene slik at disse estimatene kan danne grunnlag for å avgjøre eventuelt hvordan de sammensatte estimatorer bør se ut.

Tabell 1. Tre effisienser for estimatorer for nivåer tabellert etter ulike verdier av rotasjonsraten, μ, konstanten i CPS-estimatoren, C, og korrelasjonen, ρ. Effisiensene er regnet ovenfra: CPS-estimatoren for to tidspunkter relativt til henholdsvis den optimale og den simple estimator, og CPS-estimatoren for ∞ mange tidspunkter relativt til den simple estimator.

μ =			1/2					1/4					1/6					1/8		
C =	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8
	.96	1.18	1.41	1.71	2.57	1.06	1.64	2.15	2.80	4.55	1.16	2.11	2.89	3.88	6.49	1.27	2.58	3.64	4.96	8.41
ρ = 0.5	.90	1.10	1.31	1.60	2.40	1.00	1.55	2.03	2.65	4.30	1.11	2.02	2.77	3.72	6.21	1.23	2.50	3.52	4.80	8.13
	.95	1.07	1.25	1.60	3.80	.97	.99	1.04	1.15	1.93	.98	.99	1.02	1.08	1.53	.98	.99	1.01	1.05	1.36
	.99	1.19	1.41	1.72	2.58	1.07	1.62	2.12	2.76	4.47	1.16	2.06	2.81	3.77	6.30	1.26	2.50	3.51	4.78	8.11
0.6	.89	1.07	1.27	1.55	2.33	.98	1.48	1.94	2.52	4.10	1.08	1.92	2.62	3.52	5.88	1.19	2.36	3.32	4.52	7.67
	.92	.97	1.10	1.36	3.08	.95	.95	.98	1.06	1.72	.97	.96	.98	1.03	1.41	.98	.97	.98	1.01	1.28
	1.02	1.21	1.44	1.75	2.62	1.09	1.62	2.11	2.75	4.46	1.17	2.03	2.76	3.70	6.18	1.25	2.44	3.41	4.65	7.88
0.7	.88	1.04	1.24	1.50	2.26	.95	1.42	1.84	2.40	3.89	1.05	1.82	2.48	3.32	5.55	1.15	2.23	3.12	4.25	7.21
	.89	.88	.95	1.12	2.36	.94	.91	.92	.97	1.46	.96	.94	.94	.96	1.26	.97	.95	.95	.97	1.18
	1.07	1.26	1.48	1.80	2.70	1.14	1.66	2.15	2.80	4.54	1.21	2.05	2.78	3.71	6.21	1.28	2.43	3.39	4.61	7.82
0.8	.86	1.02	1.20	1.46	2.18	.93	1.35	1.75	2.27	3.69	1.01	1.72	2.33	3.12	5.21	1.10	2.10	2.92	3.98	6.74
	.86	.79	.80	.88	1.64	.93	.87	.85	.86	1.15	.95	.91	.89	.89	1.07	.96	.93	.92	.92	1.04
	1.14	1.32	1.56	1.89	2.83	1.25	1.78	2.30	2.98	4.83	1.32	2.18	2.94	3.93	6.57	1.39	2.56	3.56	4.83	8.19
0.9	.85	.99	1.16	1.41	2.11	.90	1.29	1.66	2.15	3.48	.98	1.62	2.19	2.92	4.88	1.06	1.96	2.73	3.70	6.28
	.83	.69	.65	.64	.92	.91	.82	.77	.73	.78	.94	.88	.84	.80	.80	.96	.91	.88	.85	.83

Tabell 2. Optimale verdier for konstanten C, i CPS-estimatoren for nivåer, basert på observasjoner på to tidspunkter

T =	1/2	1/3	1/4	1/5	1/6	1/7	1/8
ρ =							
.0	.13	.08	.06	.05	.04	.04	.03
.1	.14	.09	.07	.06	.05	.04	.04
.2	.15	.10	.08	.06	.05	.05	.04
.3	.16	.12	.09	.07	.06	.05	.05
.4	.18	.13	.10	.08	.07	.06	.05
.5	.19	.14	.11	.09	.08	.07	.06
.6	.21	.16	.13	.10	.09	.08	.07
.7	.23	.18	.14	.12	.10	.09	.08
.8	.25	.20	.16	.13	.11	.10	.09
.9	.27	.22	.18	.15	.13	.11	.10

Tabell 3. Tre effisienser for estimatorer for endringer, tabellert etter ulike verdier av rotasjonsraten, μ, konstanten i CPS-estimatoren C og korrelasjonen, ρ. Effisiensene er regnet ovenfra: CPS-estimatoren for to tidspunkter relativt til henholdsvis den optimale og den simple estimator og CPS-estimatoren for ∞ mange tidspunkter relativt til den simple estimator.

μ = :			1/2					1/4					1/6					1/8		
C = :	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8	0.2	0.4	0.5	0.6	0.8
ρ = 0.5	1.02	1.01	1.03	1.08	1.25	1.00	1.09	1.21	1.38	1.84	1.01	1.22	1.43	1.71	2.47	1.03	1.37	1.67	2.06	3.10
	.82	.80	.82	.86	1.00	.84	.92	1.02	1.16	1.55	.89	1.08	1.26	1.50	2.17	.93	1.23	1.51	1.86	2.80
	0.90*)	-	-	-	-	-	0.65*)	-	-	-	-	0.60*)	-	-	-	-	0.58*)	-	-	-
0.6	1.06	1.00	1.01	1.04	1.17	1.01	1.04	1.12	1.24	1.63	1.00	1.12	1.28	1.50	2.12	1.01	1.23	1.46	1.77	2.63
	.77	.73	.73	.75	.85	.79	.81	.87	.97	1.27	.83	.93	1.06	1.24	1.76	.87	1.06	1.26	1.52	2.26
	-	0.81*)	-	-	-	-	-	0.58*)	-	-	-	-	-	0.50*)	-	-	-	-	0.47*)	-
0.7	1.16	1.03	1.00	1.01	1.10	1.06	1.00	1.04	1.12	1.42	1.02	1.04	1.14	1.30	1.78	1.00	1.11	1.26	1.49	2.16
	.73	.65	.63	.63	.69	.74	.70	.72	.78	.99	.77	.79	.86	.98	1.35	.80	.88	1.01	1.19	1.72
	-	-	-	0.70*)	-	-	-	-	-	0.44*)	-	-	-	-	0.40*)	-	-	-	-	0.35*)
0.8	1.39	1.13	1.05	1.01	1.03	1.22	1.02	1.00	1.02	1.22	1.11	1.00	1.03	1.11	1.44	1.06	1.01	1.09	1.22	1.69
	.70	.56	.52	.50	.52	.69	.59	.57	.59	.69	.71	.64	.66	.71	.93	.73	.71	.76	.85	1.17
	-	-	-	-	0.55*)	-	-	-	-	0.44*)	-	-	-	-	0.28*)	-	-	-	-	0.24*)
0.9	2.16	1.53	1.31	1.14	1.00	1.77	1.27	1.12	1.03	1.04	1.50	1.12	1.03	1.00	1.13	1.34	1.05	1.00	1.02	1.23
	.66	.47	.40	.35	.31	.66	.47	.42	.38	.39	.67	.50	.46	.44	.50	.68	.53	.51	.51	.62
	-	-	-	-	0.33*)	-	-	-	-	0.17*)	-	-	-	-	0.14*)	-	-	-	-	0.12*)

*) Disse tallene er beregnet etter tabellene til Rao & Graham [5].

Tabell 4. Optimale verdier for konstanten C i CPS-estimatoren for endringer basert på observasjoner på to tidspunkter

μ =							
ρ = .0	0	0	0	0	0	0	0
.1	.05	.04	.03	.02	.02	.02	.01
.2	.11	.08	.06	.05	.04	.03	.03
.3	.18	.13	.10	.08	.07	.06	.05
.4	.25	.18	.14	.12	.10	.09	.08
.5	.33	.25	.20	.17	.14	.13	.11
.6	.43	.33	.27	.23	.20	.18	.16
.7	.54	.44	.37	.32	.28	.25	.23
.8	.67	.57	.50	.44	.40	.36	.33
.9	.82	.75	.69	.64	.60	.56	.53

4. Markovkjedemodellen. Variansberegninger for estimatoren $p_{ij} = N_{ij}/N_{i\cdot}$

(a) Modellen

Vi skal nå se på følgende modell: X_t er en totilstands homogen markovprosess som observeres etter en rotasjonsplan på tidspunktene $T = \{0, 1, \ldots, N\}$. Som før, sett

X_{it} = observasjonen av i-te sampelsti på tidspunk t; $i \in I_t$, $t \in T$.

Sampelstiene er nummerert slik at[1)]

$$I_t = \{J_t + 1, \ldots, J_t + n\} \ , J_t = n \cdot \mu \cdot \left[\frac{t}{c}\right].$$

c er rotasjonsavstanden og μ er rotasjonsraten.

1) $[x]$ = det største hele tall $\leq x$.

En har videre

$$(4.1) \quad \alpha_{t+1} = \alpha_t(p_{11} - p_{01}) + p_{01} \; ; \; t \in T;$$

hvor $\alpha_t = EX_t = P(X_t = 1)$ og X_t har overgangsmatrise,

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} .$$

Etter appendiks 1 har den lineære differenslikning (4.1), generell løsning

$$(4.2) \quad \alpha_t = \alpha_0 \rho^t + p_{01} \frac{1 - \rho^t}{1 - \rho}, \text{ hvor } \rho = p_{11} - p_{01}$$

Med notasjonen fra avsnitt 2 og under forutsetning av at sampelstiene observeres uavhengig av hverandre, kan likelihooden for observasjonene nå skrives:

$$(4.3) \quad \Lambda = \alpha_0^{{}_0N_1} (1 - \alpha_0)^{{}_0N_0} \prod_{t=1}^{N} \alpha_t^{{}_tN''_1} (1 - \alpha_t)^{{}_tN''_0} \cdot \prod_{i,j \in E} p_{ij}^{N_{ij}}$$

hvor $N_{ij} = \sum_{t \in T'} {}_tN_{ij}$.

Vi har

$${}_0N_0 = n - {}_0N_1$$

$${}_tN''_0 = n \cdot \mu - {}_tN''_1 \qquad t - 1 \in T' \; ; \; T' = \{0, 1, \ldots, N - 1\},$$

og $\qquad (E = \{0, 1\})$

$$(4.4) \quad \sum_{i,j \in E} N_{ij} = \sum_{t \in T'} \# \{I_t \cap I_{t+1}\} = (N - \tfrac{N}{c}) \cdot n + (\tfrac{N}{c} - 1) \cdot n \cdot \lambda$$

hvor $\lambda = 1 - \mu$. Vi antar at $\frac{N}{c}$ er et helt tall. Vi har

$$(4.5) \quad n_t = \# \{I_t \cap I_{t+1}\} = \begin{cases} n \text{ når } \left[\frac{t+1}{c}\right] \neq \frac{t+1}{c}, & t \in T'. \\ n \cdot \lambda \text{ ellers.} \end{cases}$$

Et suffisient observatorsett er

$$(4.6) \quad ({}_{o}N_{1}, {}_{t+i}N''_{1};\ t \in T',\ N_{oo},\ N_{o1},\ N_{11}).$$

Vi setter

$$N_{i\cdot} = \sum_{T'} {}_{t}N_{i} = N_{il} + N_{io}.$$

(b) Taylor-utvikling av X/Y.

La oss utvikle funksjonen $f(X, y) = x/y$ i en Taylorrekke om (EX, EY). X og Y er stokastiske variable:

$$\frac{X}{Y} = \frac{EX}{EY} + \frac{1}{EY}(X - EX) - \frac{EX}{(EY)^2}(Y - EY)$$

$$\frac{1}{\{EY + \theta_1 (Y - EY)\}^2}(Y - EY)(X - EX) + \frac{2\{EX + \theta_2 (X - EX)\}}{\{EY + \theta_1 (Y - EY)\}^3}$$

$$(Y - EY),^2$$

der θ_1 , $\theta_2 \in [o, 1]$. Hvis vi nå ser bort[1)] fra de to siste leddene, får vi tilnærmet:

$$E\frac{X}{Y} = \frac{EX}{EY},$$

$$(4.7) \quad \operatorname{var}\frac{X}{Y} = E\left(\frac{X}{Y} - \frac{EX}{EY}\right)^2$$

$$= \frac{1}{EY^2}\operatorname{var} X + \frac{EX^2}{EY^4}\operatorname{var} Y - 2\frac{EX}{EY^3}\operatorname{cov}(X, Y).$$

Dersom vi skal bruke dette til å finne et tilnærmet uttrykk for $\operatorname{var}(N_{11}/N_{1\cdot})$, så får vi altså å beregne $\operatorname{var} N_{11}$, $\operatorname{var} N_{1\cdot}$, og $\operatorname{cov}(N_{11}, N_{1\cdot})$.

---

1) Vi vil måtte vende tilbake til disse leddene for å finne graden av nøyaktighet i tilnærmingene.

(c) Momentberegninger

Vi har,

$$(4.8)\quad EN_{ij} = \sum_{T'} E_t N_{ij} = \sum_{T'} P_{ij}\, n_t\, \alpha_t = np_{ij}\, \alpha,$$

$$(4.9)\quad EN_{l\cdot} = \sum_{T'} E_t N_{l\cdot} = \sum_{T'} n_t \alpha_t = n\alpha,$$

hvor

$$\alpha = \sum_{T'} \frac{n_t}{n} \alpha_t.$$

Følgende resultater[1] utledes lett av resultatene i appendiks 1:

$$\operatorname{var}_t N_{l\cdot} = n_t\, \alpha_t(1 - \alpha_t),$$

$$(4.10)\quad \operatorname{cov}(X_r, X_s) = \alpha_k(1 - \alpha_k)\rho^{|lr - sl|},\ k = r \vee s.$$

Når vi skal beregne var N_{ll}, var $N_{l\cdot}$ og cov $(N_{ll}, N_{l\cdot})$, er følgende uttrykk mer bekvemme:

$$N_{ll} = \sum_{t \in T'} \sum_{K \in I_t \cap I_{t+l}} X_{kt} \cdot X_{k\,t+l}$$

og

$$N_{l\cdot} = \sum_{t \in T'} \sum_{k \in I_t \cap I_{t+l}} X_{kt}.$$

Vi får

$$(4.11)\quad \operatorname{var} N_{i\cdot} = \sum_{r,\, s \in T'} \sum_{k \in I_r \cap I_{r+l}} \sum_{l \in I_s \cap I_{s+l}} \operatorname{cov}(X_{kt}, X_{ls}).$$

---

1) $a \vee b = \sup(a, b)$, $a \wedge b = \inf(a, b)$

Setter vi K for den indeksmengden som hører til summen over, så kan vi videre skrive:

$$(4.12)\quad \operatorname{var} N_{11} = \sum_{K} \operatorname{cov}(X_{kr} \cdot X_{kr+1}, X_{ls} \cdot X_{ls+1}),$$

$$(4.13)\quad \operatorname{cov}(N_{11}, N_{1\cdot}) = \sum_{K} \operatorname{cov}(X_{kr} \cdot X_{kr+1}, X_{ls}).$$

Sett så

$$m_{rs} = \#\{I_r \cap I_{r+1} \cap I_s \cap I_{s+1}\}.$$

Siden stiene er uavhengige og identisk fordelte får vi nå:

$$(4.14)\quad \operatorname{var} N_{1\cdot} = \sum_{r,\, s \in T'} m_{rs} \operatorname{cov}(X_t, X_s),$$

$$(4.15)\quad \operatorname{var} N_{11} = \sum_{r,\, s \in T'} m_{rs} \operatorname{cov}(X_r \cdot X_{r+1}, X_s \cdot X_{s+1}),$$

$$(4.16)\quad \operatorname{cov}(N_{1\cdot}, N_{11}) = \sum_{r,\, s \in T'} m_{rs} \operatorname{cov}(X_r \cdot X_{r+1}, X_s).$$

La oss beregne de to siste kovariansene over. Vi har

$$E(X_t \cdot X_{t+1}) = P(X_t = X_{t+1} = 1) = P(X_{t+1} = 1 \mid X_t = 1)\, P(X_t = 1)$$

$$= \alpha_t \cdot P_{11}.$$

Anta at $r \leq s$. Da blir

$$E(X_r \cdot X_{r+1} \cdot X_s \cdot X_{s+1}) = P(X_r = X_{r+1} = X_s = X_{s+1} = 1)$$

$$= \begin{cases} P(X_{m+1} = 1 \mid X_m = 1) \cdot P(X_m = 1 \mid X_{k+1} = 1) \cdot \\ \cdot P(X_{k+1} = 1 \mid X_k = 1) \cdot P(X_k = 1) \\ = \alpha_k \cdot p_{11}^{\,2} \cdot p_{11}^{(m-k-1)}, & r \neq s, \\ = \alpha_r \cdot p_{11}, & r = s, \end{cases}$$

hvor $P^k = \{p_{ij}^{(k)}\}$, $k = r \wedge s$ og $m = r \vee s$.

Nå er

$$(4.17)\quad p_{11}^{(k)} - p_{01}^{(k)} = \rho^k.$$

Dette kan vises slik ved induksjon: (4.17) er riktig for $k = 1$. Anta at den er riktig for k. Da er

$$(4.18)\quad p_{11}^{(k+1)} - p_{01}^{(k+1)} = p_{10}^{(k)} \cdot p_{01} + p_{11}^{(k)} \cdot p_{11} - p_{00}^{(k)} \cdot p_{01} - p_{01}^{(k)} \cdot p_{11}$$

$$= p_{01} - p_{11}^{(k)} p_{01} + p_{11}^{(k)} p_{11} - p_{01} + p_{01}^{(k)} p_{01} - p_{01}^{(k)} p_{11}$$

$$= (p_{11} - p_{01})(p_{11}^{(k)} - p_{01}^{(k)}) = \rho \cdot \rho^k = \rho^{k+1}.$$

Analogt med (4.1), har vi når $s > r$:

$$(4.19)\quad \alpha_s = \alpha_r(p_{11}^{(s-r)} - p_{01}^{(s-r)}) + p_{01}^{(s-r)} = \alpha_r \rho^{s-r} + p_{01}^{(s-r)}.$$

Vi får nå ved å bruke (4.18) og (4.19),

$$(4.20)\quad \operatorname{cov}(X_{r+1} \cdot X_r, X_{s+1} \cdot X_s) = \alpha_r p_{11}^{\,2} p_{11}^{(s-r-1)} - \alpha_r \cdot \alpha_s \cdot p_{11}^{\,2}$$

$$= \alpha_k p_{11}^{\,2}(p_{11}^{(m-k-1)} - \alpha_{k+1}\rho^{m-k-1} - p_{01}^{(m-k-1)})$$

$$= \alpha_k(1 - \alpha_{k+1})\, p_{11}^{\,2}\, \rho^{m-k-1}, \quad k = r \wedge s < r \vee s = m$$

og

$$(4.21)\quad \operatorname{var}(X_t \cdot X_{t+1}) = \alpha_t \cdot p_{11}(1 - \alpha_t \cdot p_{11}).$$

Videre har vi

$$E\,X_r \cdot X_{r+1} \cdot X_s = \begin{cases} \alpha_r \cdot p_{11} \cdot p_{11}^{(s-r-1)}, & s \geq r+1, \\ \alpha_s \cdot p_{11} p_{11}^{(r-s)}, & s \leq r, \end{cases}$$

$$E\,X_r \cdot X_{r+1} \cdot E\,X_s = \alpha_r \cdot \alpha_s \cdot p_{11} = \begin{cases} \alpha_r p_{11}\left[\alpha_{r+1}\rho^{s-r-1} + p_{01}^{(s-r-1)}\right], & s \geq r+1 \\ \alpha_s p_{11}\left[\alpha_s \rho^{s-r} + p_{01}^{(s-r)}\right], & s \leq r. \end{cases}$$

Dette gir

$$(4.22)\quad \operatorname{cov}(X_r \cdot X_{r+1}, X_s) = \begin{cases} \alpha_r(1-\alpha_{r+1})\,p_{11}\,\rho^{s-r-1}, & s \geq r+1, \\ \alpha_s(1-\alpha_s)\,p_{11}\,\rho^{r-s}, & s \leq r. \end{cases}$$

(4.7) - (4.9), (4.14) - (4.16), (4.10) og (4.20) - (4.22) gir oss nå:

$$\operatorname{var}\frac{N_{11}}{N_{1\cdot}} \approx \sum_{r,\,s\,\in\,T'} m_{rs}\left[\frac{1}{n^2\alpha^2}\operatorname{cov}(X_{r+1}X_r, X_{s+1}X_s) + \frac{n^2 p_{11}^2 \alpha^2}{n^4\alpha^4}\operatorname{cov}(X_s, X_r) - \frac{np_{11}\alpha}{n^3\alpha^3}\operatorname{cov}(X_r X_{r+1}, X_s)\right]$$

$$= \sum_{s\,\in\,T'} \frac{m_{rs}}{n^2\alpha^2}\left[\alpha_s p_{11}(1-\alpha_s p_{11}) + p_{11}^2\,\alpha_s(1-\alpha_s) - p_{11}^2\alpha_s(1-\alpha_s)\right]$$

$$+ \sum_{s\,\neq\,r} \frac{m_{rs}}{n^2\alpha^2}\left[p_{11}^2\alpha_k(1-\alpha_{k+1})\,\rho^{|r-s|-1} + p_{11}^2\alpha_k(1-\alpha_k)\cdot\rho^{|r-s|}\right.$$

$$\left. - 2\,p_{11}^2\,\alpha_k(1-\alpha_{k+\delta(r,s)})\rho^{|r-s|-\delta(r,s)}\right],\quad k = r \wedge s,$$

$$= \sum_{s \in T'} \frac{m_{rs}}{n^2\alpha^2} \alpha_s p_{11}(1 - \alpha_s p_{11})$$

$$+ \sum_{s \neq r} \frac{m_{rs}}{n^2\alpha^2} \rho^{|r-s|} \left[\rho^{-1}(1 - \alpha_{r\wedge s} + 1) + (1 - \alpha_{r\wedge s})\right.$$

$$\left. - 2(1 - \alpha_{r\wedge s + \delta(r,s)}) \cdot \rho^{-\delta(r,s)}\right] \cdot p_{11}^2 \alpha_{r\wedge s}.$$

Vi har for $r < s$:

$$(4.23)\quad \rho^{-1}(1 - \alpha_{r\wedge s + 1}) + (1 - \alpha_{r\wedge s}) - 2\rho^{-\delta(r,s)}(1 - \alpha_{r\wedge s + \delta(r,s)})$$

$$= \rho^{-1}(1 - \alpha_{r+1}) + 1 - \alpha_r - 2\rho^{-1}(1 - \alpha_{r+1})$$

$$= -\rho^{-1}\left[(1 - \alpha_{r+1}) - \rho(1 - \alpha_r)\right].$$

Ved å anvende resultater fra appendiks 1 på den kjeden vi får når tilstandene 0 og 1 byttes om, ser vi at vi får følgende uttrykk for (4.23) når $r < s$:

$$-\rho^{-1}(1 - p_{11})$$

For $r > s$ kan (4.23) skrives slik:

$$\rho^{-1}(1 - \alpha_{s+1}) + (1 - \alpha_s) - 2(1 - \alpha_s)$$

$$= \rho^{-1}\left[(1 - \alpha_{s+1}) - \rho(1 - \alpha_s)\right]$$

$$= \rho^{-1}(1 - p_{11}).$$

Vi får nå:

$$(4.24)\quad \operatorname{var}\frac{Nu}{N_{1.}} \approx \sum_{r=1}^{N-1} \frac{np_{11}(1 - \alpha_r p_{11})}{\alpha^2} \alpha_r + \sum_{r<s}^{N} m_{rs} \frac{p_{11}^2(1 - p_{11})}{\alpha^2\rho} \rho^{|r-s|}(\alpha_s - \alpha_r).$$

Her er jo $m_{rs} = 0$ når $|r - s| > \frac{1}{\mu} \cdot c$, hvor c er rotasjonsavstanden.
Med $c = 1$ og rotasjonsrate μ blir

$$m_{rs} = \begin{cases} (1 - |r - s| \cdot \mu)n & \text{for } |r - s| \leq \frac{1}{\mu}, \\ 0 & \text{ellers} \end{cases}$$

Bruken av denne markovmodellen krever at (4.2) er en god tilpassing til virkeligheten for det tidsrom som vi bruker under estimeringen. De estimatorene vi vil komme fram til, vil kunne kalles glidende estimatorer i det de bruker observasjonene fra tidspunktene t - s + 1, ..., t - 1, t. s kan kalles estimatorens lengde.

5. Konklusjoner

(a) Hvor mye kan vinnes?

Ved å ta i bruk den sammensatte estimator $\tilde{\alpha}_t$, slik den er gjengitt i (2.8), kan større presisjon oppnås ved estimering av kvartalsvise endringer og nivåer. Gevinsten kan bli stor ved estimering av endringer, mens det er mindre å vinne ved estimering av nivåer. Tabell 1 og 3 gir oss de gevinstene som kan oppnås. Avvik fra forutsetningene for beregningen av effisiensene svekker imidlertid fordelene ved sammensatt estimering noe. Slike avvik vil bli behandlet nedenfor. Det er ikke dokumentert tilstrekkelige fordeler ved de andre estimatorer som ble foreslått i avsnitt 2 (se (2.9) og (2.10)) til at de vil bli foreslått brukt. Nærmere studier av disse estimatorer anbefales.

Ved fastsetting av rotasjonsraten er det tidligere nevnt at det trengs en prioritering av de interessante parametrene.* Dersom en vil ha observasjoner av en og samme sampelsti med ett års mellomrom, må imidlertid rotasjonsraten være mindre enn 1/4. Dette er nødvendig hvis en vil beregne den årvise korrelasjon for å kunne ta hensyn til sesongvariasjonene under estimeringen. Estimatoren (2.9) vil ikke kunne brukes uten at observasjonen av samme sti foretas med ett års mellomrom. Årvise endringer vil kunne estimeres sikrere med en rotasjonsrate mindre enn 1/4. Fastsetting av rotasjonsraten påvirkes også av momenter som ligger utenfor rammen av dette notatet. Selv uten sammensatt estimering, er det en betydelig gevinst ved selve roteringen når endringer skal estimeres. Effisiensen til den enkle estimator for endringer ved rotering relativt til samme estimator uten rotering er gitt ved

(5.1) $\quad 1 - (1 - \mu)\,\rho$

Av tabell 1 ser vi at gevinsten ved sammensatt estimering av nivåer aldri er betydelig unntagen ved en så høy korrelasjon som 0.9 og rotasjon 1/2. Dette vil neppe opptre i praksis. Derimot kan gevinster

*) Som nevnt er optimal rotasjonsrate 1/2 for sammensatt estimering av nivåer, 0 for endringer og 1 for årsgjennomsnitt.

på opp til 40 prosent oppnås allerede ved korrelasjon 0.5 under estimering av endringer. Siden det bare er ved estimering av endringer større gevinster oppnås, er det rimelig at rotasjonen tilpasses estimering av endringer. Anta at en vil estimere endringer i arbeidsstokken. La oss si at korrelasjonen er 0.7 og at en har bestemt seg for en rotasjonsrate på 1/6. Gevinsten som da kan oppnås, er 60 prosent dersom en velger konstanten C lik 0.7. Dersom en bruker sammensatt estimering av nivået, med denne konstanten, ser en at en risikerer varianstap. (C = 0.6 gir en effisiens på 0.96 og C = 0.8 gir 1.26).

(b) Årsaker til reduksjon i variansgevinsten

Fra avsnitt 2 husker vi at følgende momenter vil gjøre variansgevinsten mindre enn det som ble nevnt i avsnitt 5(a).

1) Roteringen skjer innafor de p.u.e. (primære utvalgsenheter). Dersom variansen mellom p.u.e. dominerer, vil det bli liten variansgevinst.

2) Frafall. (Individer som flytter, må følges opp.)

3) Endring av gruppetilhørighet under undersøkelsen. (Individer kan grupperes etter kjennetegn ved inntredelsen dersom dette ikke skaper for stor forventningsskjevhet).

4) Optimalt valg av konstanten C begrenses av kravet om konsistens i de publiserte tabeller.

Virkningene av punktene 1 til 3 ovenfra vil svare til at korrelasjonen, ρ, vil reduseres. En vil derfor av tabellene kunne se hvordan effisiensene påvirkes. Variansgevinsten beholdes for endringer mens det fort oppstår varianstap for nivåer.

(c) Valg av konstanten C

Endringer estimeres ved $\tilde{\delta}_k = \tilde{\alpha}_k - \tilde{\alpha}_{k-1}$, hvor $\tilde{\alpha}_k$ er gitt ved (2.8). C i (2.8) bør kunne velges fleksibelt etter beregning av korrelasjoner for de ulike grupper og kjennetegn. Svenskene har funnet at korrelasjonen først og fremst varierer etter kjennetegn. Det er nesten ingen korrelasjon for kjennetegnet "arbeidsløs", men høy korrelasjon for kjennetegnet "i arbeid". (Se [11].) De lar derfor C variere med disse kjennetegnene, mens konstanten ikke varierer mellom de ulike alders- og utdanningsgrupper. På denne måten vil kolonnesummene stemme i de publiserte tabellene, mens linjesummene ikke vil stemme med totalen. (Jeg

tenker meg her at vi har "arbeidsløs" osv. kolonnevis og "15-20 år" osv. linjevis.)

(d) Reviderte tall for nivåer

En kan gi avkall på sammensatt estimering av nivåer og publisere reviderte tall for forrige to kvartal ved formelen

$$\tilde{\alpha}_{k-1,\,\text{revidert}} = \tilde{\alpha}_k - \tilde{\delta}_k.$$

På denne måten unngår en at de publiserte endringer ikke stemmer med differansen mellom de publiserte nivåer. Dessuten vil $\tilde{\alpha}_{k-1,\,\text{revidert}}$ sannsynligvis være et sikrere estimat enn $\tilde{\alpha}_{k-1}$.

I [12] er det beregnet at merkostnadene ved innføring av estimatoren gitt ved (2.8) er S.kr. 30 000 pr. år, mens en reduksjon av variansen på 2 - 3 prosent ved å øke utvalgsstørrelsen med 800-1 000 individer, vil koste S.kr. 100 000 pr. år. Det må nevnes at svenskene bruker månedsvise undersøkelser.

Overgangssannsynligheter vil måtte estimeres med estimatoren $N_{ij}/N_{i.}$. Det er vanskelig, av (4.24) å si hvordan variansen påvirkes av rotasjonsplanen.

Referanser

[1] Des Raj (1968): "Sampling Theory." Mc. Graw-Hill, New York.

[2] Dahmstrøm, P. (1971): "Sampling and estimation for Surveys on several occations - a review." Statistisk tidskrift, 9 (1), 25-47.

[3] Gurney, M. and Daly, J. F. (1965): "A multivariate approach to estimation in periodic sample surveys." Proceedings of the Social Statistics Section, Amer. Statist. Assoc. 1965, 242-257.

[4] Waksberg, J. and Pearl, R. B. (1964): "The Current Population Survey: A case history in panel operations." Proceedings of the Social Statistics Section, Amer. Statist. Assoc. 1964, 217-231.

[5] Rao, J. N. K. and Graham, J. E. (1964): "Rotation designs for sampling on repeated occations." J. Amer. Statist. Assoc., 59, 492-509.

[6] Prahbu-Ajgaonkar, S. G. (1968): "The theory of univariate sampling on successive occations under the general correlation pattern." Austral. J. Statist., 10, 56-63.

[7] Patterson, H. D. (1950): "Sampling on successive occations with partial replacement of units." J. Roy. Statist. Soc. Ser. B, 12, 241,255.

[8] Eckler, A. R. (1955): "Rotation sampling." Ann. Math. Statist., 26, 664-685.

[9] Dahmstrøm, P. og Malmberg, S. (1970): "Forsøk med sammensatta skattningar i A.K.U." P. M. från Statistiska Centralbyrån, Utredningsinstitutet, 28. 7. 1970.

[10] Malmberg, S. (1970): "Korrelasjonsberäkningar i arbetskraftundersökningarna (A.K.U.) 1970." Resultatredovisning från Statistiska Centralbyrån, 6. 11. 1970.

[11] Dahmström, P. og Malmberg, S. (1971): "Utredningsarbetet rörande sammensatta skattningar i A.K.U. "P. M. från Statistiska Centralbyrån, Utredningsinstitutet, 25. 5. 1971.

[12] Kulldorf, G. (1963): "Some problems of optimum allocation for sampling on two occations." Rev. of the Internat. Statist. Inst., 31, 24-56.

Appendiks 1.

Noen utregninger for markovkjedemodellen.

X_t er en to - tilstands markovkjede med overgangsmatrise P og tilstandsrom, E = {0, 1}. Vi har

$$EX_t = \alpha_t = P(X_t = 1),$$

$$\text{var } X_t = \alpha_t(1 - \alpha_t),$$

(A.1,1)
$$\alpha_{t+1} = EX_{t+1} = EE(X_{t+1} \mid X_t)$$

$$= \alpha_t E(X_{t+1} \mid X_t = 1) + (1 - \alpha_t) E(X_{t+1} \mid X_t = 0)$$

$$= \alpha_t p_{11} + (1 - \alpha_t) p_{01}$$

$$= \alpha_t (p_{11} - p_{01}) + p_{01},$$

$$EX_t \cdot X_{t+1} = P(X_t = X_{t+1} = 1)$$

$$= \alpha_t p_{11},$$

$$\text{cov } (X_t X_{t+1}) = \alpha_t p_{11} - \alpha_t \cdot \alpha_{t+1}$$

$$= \alpha_t(1 - \alpha_t) (p_{11} - p_{01}),$$

$$\rho(X_t X_{t+1}) = (p_{11} - p_{01}) \sqrt{\frac{\alpha_t (1 - \alpha_t)}{\alpha_{t+1}(1-\alpha_{t+1})}}.$$

Differenslikningen (A.1,1) kan løses. Sett $p_{11} - p_{01} = \rho$. Differenslikningen blir da

(A.1,2)
$$\alpha_{t+1} = \alpha_t \rho + p_{01}.$$

Den homogene likning $\alpha_{t+1} = \alpha_t \rho$ har løsning $\alpha_t = \alpha_o \rho^t$. Dersom f(t) er en partikulær løsning av (A.1,2), vet vi nå at den generelle løsning av (A.1,2) er

$$\alpha_t = \alpha_o \rho^t + f(t).$$

Nå er

$$f(t) = p_{01} \frac{1 - \rho^t}{1 - \rho}$$

en partikulær løsning. Den generelle løsning blir derfor

$$\alpha_t = \alpha_t' \rho^t + p_{01} \frac{1 - \rho^t}{1 - \rho} .$$

Når $\rho < 1$, blir

$$\lim_t \alpha_t = \frac{p_{01}}{1 - \rho} .$$

Appendiks 2.

Forsøk på å finne BLU-estimatorer i markovprosessmodellen når prosessen observeres på to tidspunkter

Vi skal se på modellen i avsnitt 4 med T = {1, 2}. Av (4.3) ser vi at likelihooden for observasjonene kan skrives

$$\text{(A.2,1)}\quad \Lambda = \alpha_1^{{}_1N_1}(1-\alpha_1)^{{}_1N_0}\, p_{11}^{N_{11}}\, p_{10}^{N_{10}}\, p_{01}^{N_{01}}\, p_{00}^{N_{00}}\, \alpha_2^{{}_2N''_1}(1-\alpha_2)^{{}_2N'_0}$$

Etter appendiks 1 vet vi at $\alpha_2 = \alpha_1(p_{11} - p_{01}) + p_{01}$. Vi kan sette opp følgende sammenhenger mellom de observatorene som går inn i uttrykket for Λ:

$$\begin{aligned}
{}_1N_1 &= N_{1\cdot} + {}_1N'_1,\\
{}_1N_0 &= n - N_{1\cdot} - {}_1N'_1,\\
N_{10} &= N_{1\cdot} - N_{11},\\
N_{01} &= N_{\cdot 1} - N_{11},\\
N_{00} &= \lambda n - N_{11} - N_{01} - N_{10},\\
{}_2N''_0 &= \mu n - {}_2N''_1.
\end{aligned}\qquad \text{(A.2,2)}$$

Som før er μ rotasjonsraten, og $\lambda = 1 - \mu$. Vi har ved (A.2,2) uttrykt alle observatorer i likelihooden (A.2,1) ved observatorsettet

$$\text{(A.2,3)}\quad ({}_1N'_1,\ N_{1\cdot},\ N_{11},\ N_{\cdot 1},\ {}_2N''_1).$$

(A.2,3) er derfor et suffisient observatorsett for parametrene i modellen. På grunn av de tre lineært uavhengige releasjonene:

$$p_{i0} + p_{i1} = 1,\ i = 0, 1,$$

$$\text{(A.2,4)}\quad \alpha_2 = \alpha_1(p_{11} - p_{01}) + p_{01},$$

så kan våre seks parametre utledes av tre vilkårlige av dem. Vi skal se på BLU-estimering av α_1, p_{11} og α_2. Det må nevnes at estimater for p_{01} og p_{00} utledet av BLU-estimatorer for α_1, p_{11} og α_2, ikke nødvendigvis selv er BLU-estimater fordi sammenhengen (A.2,4) er ikke- lineær.

En BLU-estimator for en parameter eller en lineær kombinasjon av parametre i modellen må være en lineær kombinasjon av observatorene i det suffisiente settet. La oss se på BLU-estimatet $\hat{\alpha}_2$ av α_2. $\hat{\alpha}_2$ må ha formen:

$$\hat{\alpha}_2 = a_{1\,1}N_1' + a_2 N_{1\cdot} + b\, N_{11} + C_{1\,2}N_1'' + C_2 N_{\cdot 1}$$

At $\hat{\alpha}_2$ skal være forventningsrett, betyr at

$$(A.2,5)\quad E\hat{\alpha}_2 = \mu n a_1 \alpha_1 + \lambda n a_2 \alpha_1 + \lambda n b \alpha_1 p_{11} + \lambda n c_1 \alpha_2 + \mu n c_2 d_2$$

$$= n(\mu\alpha_1 \alpha_1 + \lambda a_2 \alpha_1 + \lambda b \alpha_1 p_{11} + \lambda C_1 \alpha_2 + \lambda C_2 \alpha_2)$$

$$= \alpha_2.$$

Her har jeg gjort bruk av:

$$EN_{11} = EE(N_{11} \mid N_{1\cdot}) = E(N_{1\cdot}\, p_{11}) = \lambda\alpha_1 p_{11}.$$

Dersom (A.2,5) skal gjelde for alle α_1, α_2 og p_{11}, ser vi at vi må ha

$$\mu a_1 = -\lambda a_2,\ \lambda C_1 = 1 - \mu C_2 \text{ og } b = 0.$$

Koeffisientene a_1 og C_1 kan lett bestemmes ved hjelp av en kjent setning som sier at en BLU-estimator er null-korrelert med enhver ikketriviell forventningsrett estimator av 0. ${}_1N_1' - N_{1\cdot}$ og ${}_2N_1'' - N_{\cdot 1}$ er slike ikke-trivielle estimatorer. Regningen for bestemmelse av a_1 og C_1 er gjort i f.eks. [1], s. 156 ff. Av (A.2,5) ser vi også at det ikke finnes noen BLU-estimator for p_{11}. Bytt nemlig ut $E\hat{\alpha}_2$ med $E\,\hat{p}_{11}$ og α_2 på høyre side med p_{11}, og bemerk at vi da må ha $\lambda b \alpha_1 = 1$ for alle α_1, noe som er umulig for fast p.

Ved lineær estimering kan vi altså ikke nyttiggjøre oss den informasjonen son ligger i N_{11} til å forbedre estimatorene. Dette skulle tyde på en vesentlig utilstrekkelighet ved lineære estimatorer. Resultatene over kan tyde på at dette gjelder når det er en ikke-lineær sammenheng mellom parametrene.