Interne notater

STATISTISK SENTRALBYRÅ

IN 83/4 8. februar 1983

ESTIMERING AV VEKTENE TIL EN KOMBINERT ESTIMATOR FOR FYLKESTALL

Av

Erling Siring

INNHOLD

1. INNLEDNING

I Siring og Thomsen (1981) er det beskrevet forskjellige metoder for estimering av fylkestall når en har data fra en utvalgsundersøkelse. Tre av disse metodene er:

(i) Direkte estimering
(ii) Syntetisk estimering
(iii) Kombinasjon av (i) og (ii).

Metode (i) går ut på at en anvender observasjoner fra et fylke direkte. Metode (ii) går ut på at en bruker observasjoner fra hele landet, og justerer for at fordelingene med hensyn til visse demografiske variable er annerledes i et enkelt fylke enn i hele landet.

La D betegne den direkte estimator og S en syntetisk estimator for et fylkestall. Metode (iii) går ut på at en kombinerer (i) og (ii) på følgende måte:

$$K = cS + (1-c)\,D$$

K kaller vi her for en kombinert estimator. Det er vist at en kan oppnå en reduksjon av bruttovariansen med nærmere 50 prosent i visse tilfeller ved å bruke K i stedet for S eller D som estimator for fylkestall. Gevinsten ved å anvende K er bl.a. avhengig av valget av c-verdi. Den verdi av c som minimerer bruttovariansen til K, er en funksjon av bruttovariansene til S og D som er ukjente når en kun har data fra en utvalgsundersøkelse. I dette notatet skal vi behandle problemet med å estimere den optimale c-verdi, som heretter vil bli kalt c*.

I kap. 2 skal vi studere c* og egenskapene til den kombinerte estimatoren K. I kap. 3 skal vi se på vanskelighetene ved å estimere c*.

En framgangsmåte for å kunne estimere c* presenteres i kap. 4. Videre blir det i dette kapitlet presentert konkrete forslag til estimatorer for c* når en har en ett-trinns utvalgsplan. I avsnitt 5.3. presenteres estimatorer for c* som kan brukes ved Byråets utvalgsplan.

For å kunne vurdere kvaliteten til den kombinerte estimatoren i forhold til D og S, har vi foretatt noen simuleringsforsøk. Disse forsøkene er blitt gjort ved hjelp av data fra folketellingen i 1970. Resultatene presenteres i kap. 5.

Folketellingene gir bl.a. tall for sysselsetting i forskjellige næringer. Vi har her konsentrert oss om estimering av andelen av befolkningen som er sysselsatt innen enkelte næringer. Dette betyr at estimatorene for c* er utviklet for dikotome variable. Ved å modifisere dem noe kan de imidlertid også brukes når en har kontinuerlige variable. Dette blir behandlet i kap. 6.

2. DE OPTIMALE VEKTENE OG ROBUSTHETSEGENSKAPENE TIL DEN KOMBINERTE ESTIMATOREN

La p_i betegne den parameteren som skal estimeres for fylke nr. i, og la D_i, S_i og K_i betegne henholdsvis den direkte, den syntetiske og den kombinerte estimatoren for fylke nr. i.

$$K_i = c_i S_i + (1-c_i)\,D_i.$$

Vi forutsetter at c-verdien kan variere fra fylke til fylke. Vi lar c_i^* betegne den c_i-verdien som minimerer $E\,(K_i-p_i)^2$, dvs. bruttovariansen til K_i.

I Siring og Thomsen (1981) er det vist at:

$$(1) \qquad c_i^* = \frac{E\,(D_i-p_i)^2 - E\,(D_i-p_i)\,(S_i-p_i)}{E\,(D_i-p_i)^2 + E\,(S_i-p_i)^2 - 2E\,(D_i-p_i)\,(S_i-p_i)}$$

c_i^* kan også skrives på følgende form:

$$(2) \qquad c_i^* = \frac{\operatorname{var} D_i + (ED_i-p_i)^2 - \operatorname{cov}\,(D_i, S_i) - (ED_i-p_i)\,(ES_i-p_i)}{\operatorname{var} D_i + \operatorname{var} S_i + (ES_i-p_i)^2 + (ED_i-p_i)^2 - 2\operatorname{cov}\,(D_i,S_i) - 2(ES_i-p_i)\,(ED_i-p_i)}$$

Fra (1) ser vi at hvis bruttovariansen til D_i er mye større enn bruttovariansen til S_i, vil c_i^* ligge nær 1 og K_i vil bli tilnærmet lik S_i. Omvendt ser vi at hvis bruttovariansen til S_i er mye større enn bruttovariansen til D_i, vil c_i^* ligge nær 0, og K_i vil bli tilnærmet lik D_i. Fra (2) ser

vi at skjevhetene til D_i og S_i inngår i formlene for c_i^*.

Som nevnt i innledningen har vi her konsentrert oss om estimering av andelen av befolkningen som er sysselsatt innen forskjellige næringer. Den direkte estimatoren, D_i, blir da den observerte frekvensen for fylke nr. i i utvalget. Som representant for klassen av syntetiske estimatorer har vi her valgt å bruke den aller enkleste av de syntetiske estimatorene, nemlig frekvensen i hele det landsomfattende utvalget. Denne estimatoren kaller vi $\hat{p}$, og forventningen til $\hat{p}$ kaller vi p.

Den kombinerte estimatoren blir da:

$$K_i = c_i\hat{p} + (1-c_i)\, D_i.$$

Vi kan nevne tre årsaker til at vi har brukt $\hat{p}$ i stedet for en mer "avansert" syntetisk estimator eller en regresjonsestimator som i Fay og Herriot (1979):

(i) Det har medført mindre regnearbeid.

(ii) I Laake og Langva (1976) er det presentert resultater som indikerer at det har moderat effekt å utnytte data om kjønns- og aldersfordelingen i hvert fylke ved estimering av sysselsettingstall.

(iii) Vi har lagt mer vekt på å finne fram til en metode for å kunne estimere c_i^* enn å søke etter en "optimal" estimator for p_i.

Til tross for (ii) tror vi at en i mange tilfeller vil kunne finne en langt bedre syntetisk estimator enn $\hat{p}$. En vil da kunne konstruere en bedre kombinert estimator enn den som blir presentert her. Gevinsten ved å anvende en kombinert estimator i stedet for en direkte estimator vil derfor kunne bli større enn resultatene i dette notatet indikerer.

For å forenkle regnearbeidet og for i det hele tatt å muliggjøre simuleringsforsøkene som er beskrevet i kap. 5, skal vi nå forutsette at vi har en utvalgsplan der vi trekker enkle, tilfeldige utvalg fra hvert fylke. Videre skal vi forutsette at antall personer som trekkes fra hvert fylke, er proporsjonalt med folketallet i fylket. Forutsetningene som er beskrevet over, gir:

$$(3)\qquad \begin{aligned} &ED_i = p_i \\ &\operatorname{cov}(D_i, \hat{p}) = \frac{n_i}{n} \operatorname{var} D_i, \text{ der} \end{aligned}$$

n_i er antall observasjoner fra fylke nr. i og n er antall observasjoner fra hele landet.

Videre har vi at $\operatorname{var} D_i \approx p_i\frac{(1-p_i)}{n_i}$ og at $\operatorname{var} \hat{p} \approx \frac{p(1-p)}{n}$.

Ved å sette dette inn i (2) og regne litt finner en så:

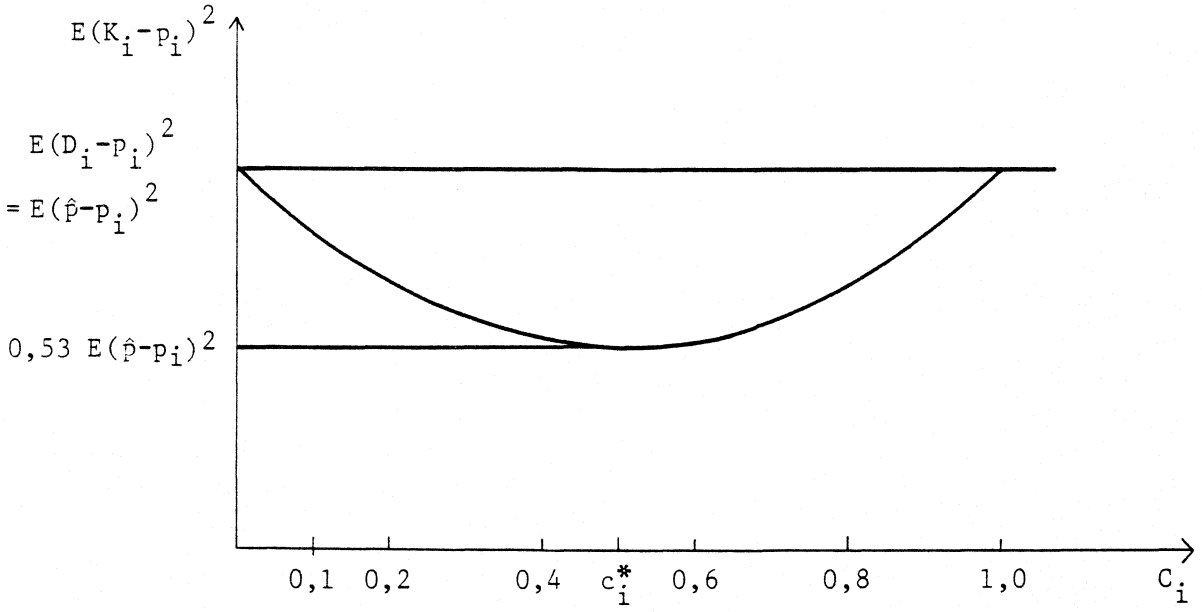
$$(4)\qquad C_i^* \approx \frac{(1-\frac{n_i}{n})\,\frac{1}{n_i}\,p_i(1-p_i)}{(1-2\,\frac{n_i}{n})\,\frac{1}{n_i}\,p_i(1-p_i) + \frac{1}{n}\,p(1-p)+(p-p_i)^2}$$

Når en skal estimere c_i^*, er det viktig å vite hvor robust kvaliteten til K_i er for forskjellige valg av c_i. Figur 1, 2 og 3 viser hvordan bruttovariansen til K_i varierer med c_i for forskjellige relasjoner mellom bruttovariansene til D_i og $\hat{p}$. De horisontale linjene i figurene markerer bruttovariansene til D_i og $\hat{p}$.

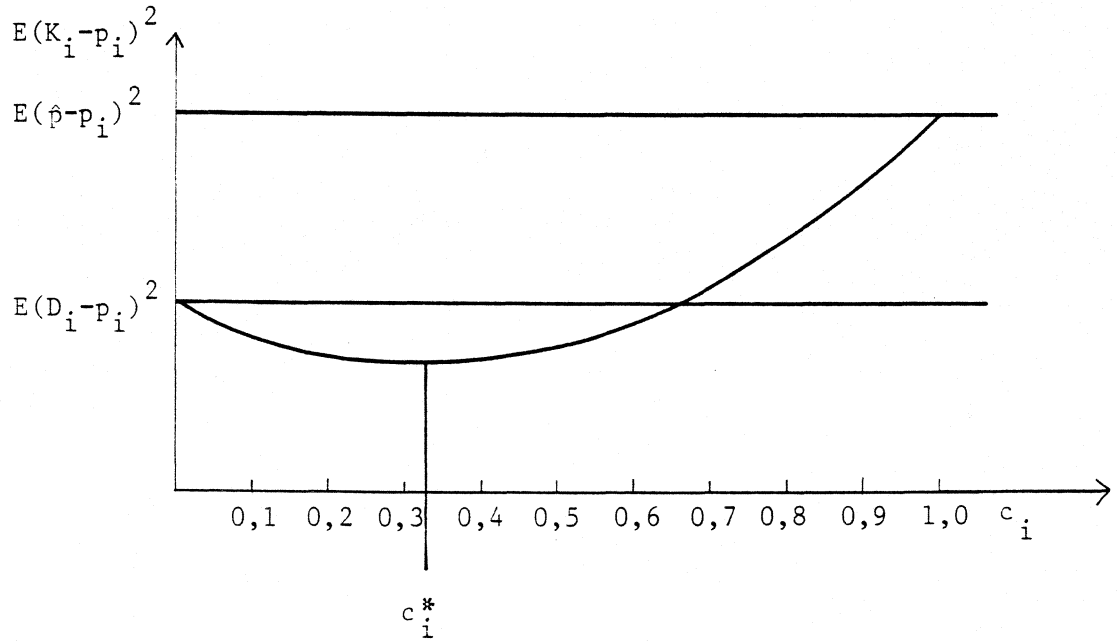
Bruttovariansen til K_i er bl.a. en funksjon av kovariansen til D_i og $\hat{p}$. I figurene har vi satt $\operatorname{cov}(D_i,\hat{p}) = \frac{1}{19}\operatorname{var} D_i$ siden gjennomsnittsverdien til $\frac{n_i}{n} = \frac{1}{19}$ (jfr. (3)). Hvis vi hadde forutsatt at $\operatorname{cov}(D_i,\hat{p})$ hadde vært større, ville kurvene ha vært noe flatere. Dette betyr at når $\operatorname{cov}(D_i,\hat{p})$ er større enn forutsatt her, er gevinsten ved å anvende en kombinert estimator mindre enn figurene antyder.

Dersom vi her hadde brukt en annen syntetisk estimator enn $\hat{p}$, ville kurvene ha vært de samme så sant kovariansene hadde vært like store som forutsatt. Dersom vi hadde forutsatt at Byråets utvalgsplan hadde vært brukt, ville også kurvene ha vært omtrent de samme.

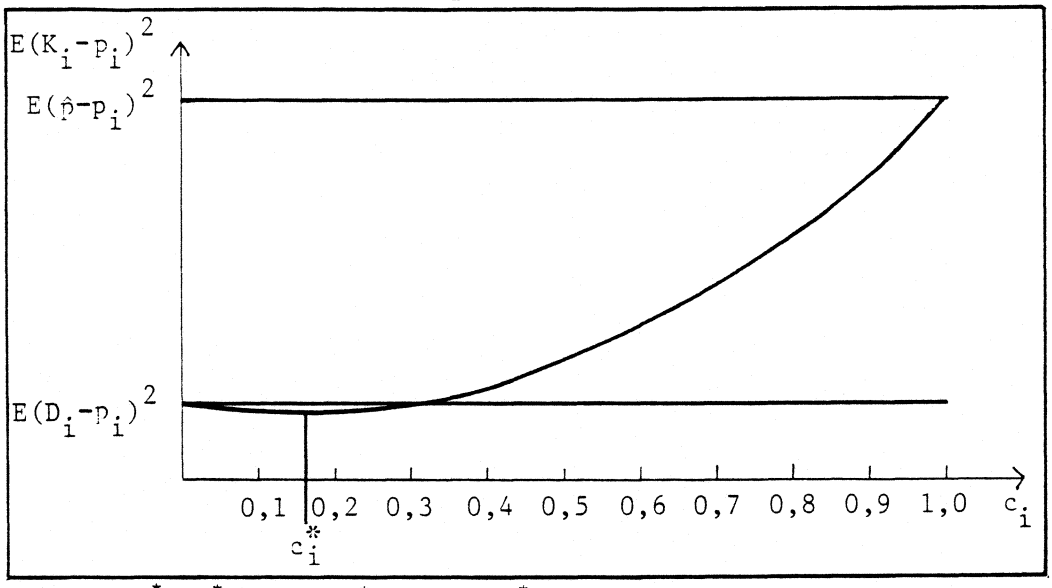
Figur 1. Bruttovariansen til K_i som funksjon av c_i når bruttovariansen til D_i er lik bruttovariansen til $\hat{p}$ ($c_i^* = 0,5$)



Figur 2. Bruttovariansen til den kombinerte estimatoren som funksjon av c_i når $E(\hat{p}-p_i)^2 = 2E(D_i-p_i)^2$



Figur 3. Bruttovariansen til den kombinerte estimatoren som funksjon av c_i når $E(\hat{p}-p_i)^2 = 5\ E(D_i-p_i)^2$



La $K_i^* = c_i^* \hat{p} + (1 - c_i^*) D_i$, dvs. K_i^* er den K_i (som funksjon av c_i) som har minst bruttovarians.

Figur 1, 2 og 3 illustrerer enkelte av de generelle egenskapene til den kombinerte estimatoren. Når bruttovariansene til D_i og $\hat{p}$ er like store, er bruttovariansen til K_i^* bare om lag halvparten så stor som bruttovariansene til D_i og $\hat{p}$. Gevinsten ved å bruke K_i^* i stedet for den beste av estimatorene D_i og $\hat{p}$ avtar når forskjellen i bruttovarians mellom D_i og $\hat{p}$ øker. Reduksjonen i bruttovarians ved å bruke K_i^* i stedet for den beste av estimatorene $\hat{p}$ og D_i ligger alltid et sted mellom 0 og 50 prosent, avhengig av forholdet mellom bruttovariansene til $\hat{p}$ og D_i.

I praksis vil en ikke kunne bruke estimatoren K_i^* siden c_i^* er ukjent. Figurene illustrerer imidlertid at en ikke er avhengig av å bruke den eksakte c_i^* for å kunne tjene på å bruke en kombinert estimator. I figur 1 ser vi at den kombinerte estimatoren er bedre enn både D_i og $\hat{p}$ for alle verdier av c_i mellom 0 og 1 når D_i og $\hat{p}$ har like stor bruttovarians. Generelt har den kombinerte estimatoren mindre bruttovarians enn både D_i og $\hat{p}$

$$\text{hvis } c_i \in \begin{cases} (0,2c_i^*) \text{ for } c_i^* \in (0,\tfrac{1}{2}] \\ (2c_i^* - 1,1) \text{ for } c_i^* \in (\tfrac{1}{2},1) \end{cases}$$

Bredden på disse intervallene blir kortere og kortere ettersom forskjellen i bruttovarians mellom $\hat{p}$ og D_i øker. Dette er illustrert i figurene. I Schaible (1978) og Royall (1979) er det mer om dette. I de nevnte artikler er det også vist at bruttovarianskurven til K_i er svært flat i området omkring c_i^*.

Dette betyr altså at kvaliteten til K_i er svært robust overfor forskjellige valg av c_i i området omrking c_i^*. Selv om en bruker en "dårlig" estimator for c_i^*, kan en altså i de fleste tilfeller regne med å oppnå en gevinst ved å ta i bruk en kombinert estiamtor. I Schaible, Broch og Schnack (1977) og i Schaible (1979) er det presentert empiriske resultater som indikerer at dette er riktig.

I enkelte situasjoner har det imidlertid ingen hensikt å anvende en kombinert estimator. Hvis f.eks. den ene av de to estimatorene D_i eller $\hat{p}$ har mye mindre bruttovarians enn den andre, kan den kombinerte estimatoren ved et uheldig valg av c_i bli dårligere enn den beste komponentestimatoren. Hvis en vet at en er i en slik situasjon, bør en derfor bruke den beste av de to komponentestimatorene framfor å bruke en kombinert estimator.

Den absolutt maksimale gevinst ved å erstatte den beste komponentestimatoren med en kombinert estimator er å få redusert bruttovariansen med 50 prosent. I de fleste tilfeller vil gevinsten være langt mindre. Hvis begge de to komponentestimatorene er av uakseptabel dårlig kvalitet, er det derfor tvilsomt om en kombinert estimator vil være av tilfredstillende kvalitet.

3. PROBLEMET VED Å ESTIMERE DE OPTIMALE VEKTENE

I dette kapitlet skal vi se på vanskelighetene ved å estimere c_i^*. Videre skal vi presentere enkelte mislykkede forslag til estimatorer for c_i^*.

Fra (2) ser vi at c_i^* er en funksjon av variansene og skjevhetene til D_i og S_i, samt av kovariansen mellom de to estimatorene. I en praktisk situasjon vil det som regel være mulig å finne brukbare estimatorer for var D_i, var S_i og cov (D_i,S_i). Derimot vil det by på problemer å finne en brukbar estimator for skjevheten til S_i, dvs. ES_i-p_i.

Vi skal igjen begrense oss til den situasjonen som er beskrevet i kap. 2, nemlig at vi bruker $\hat{p}$ som representant for de syntetiske estimatorene og at vi har en ett-trinns utvalgsplan. En forventningsrett estimator for $p-p_i$ er $\hat{p} - D_i$. Denne estimatoren har imidlertid høy varians. I tillegg er det egentlig ikke $p-p_i$ vi er interessert i å estimere, men $(p-p_i)^2$. Gjennom de forslag til estimatorer for c_i^* som vi skal vurdere, vil vi se at denne størrelsen er problematisk å estimere.

Fra (4) har vi:

$$c_i^* \approx \frac{(1-\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i)}{(1-2\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i) + \frac{1}{n}p(1-p) + (p-p_i)^2}$$

Et naturlig forslag til estimator for c_i^* blir da:

$$C_{i1} = \frac{(1-\frac{n_i}{n})\frac{1}{n_i}D_i(1-D_i)}{(1-2\frac{n_i}{n})\frac{1}{n_i}D_i(1-D_i) + \frac{1}{n}\hat{p}(1-\hat{p}) + (\hat{p}-D_i)^2}$$

Hvis vi ser på de enkelte leddene i C_{i1}, har vi at

$$E(1-\frac{n_i}{n})\frac{1}{n_i}D_i(1-D_i) \approx (1-\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i),$$

$$E(1-2\frac{n_i}{n})\frac{1}{n_i}D_i(1-D_i) \approx (1-2\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i),$$

$$E\frac{1}{n}\hat{p}(1-\hat{p}) \approx \frac{1}{n}p(1-p) \text{ og } E(\hat{p}-D_i)^2 =$$

$$E(\hat{p}-p_i+p_i-D_i)^2 = E(D_i-p_i)^2 + E(\hat{p}-p_i)^2 - 2E(D_i-p_i)(\hat{p}-p_i) =$$

hele nevneren i c_i^* (jfr. (1)). Vi har altså at $E(\hat{p}-D_i)^2 > (p-p_i)^2$.

Ved bruk av estimatoren C_{i1} ville sannsynligvis c_i^* bli kraftig underestimert.

Siden $E(\hat{p}-D_i)^2$ = nevneren i c_i^*, er det nærliggende å vurdere følgende estimator for c_i^*:

$$C_{i2} = \frac{(1-\frac{n_i}{n})\frac{1}{n_i}D_i(1-D_i)}{(\hat{p}-D_i)^2}$$

Estimatoren C_{i2} er tilnærmet forventningsrett.

$$\text{La } C_{i3} = \begin{cases} 1 & \text{hvis } C_{i2} > 1 \\ C_{i2} & \text{hvis } C_{i2} \leq 1 \end{cases}$$

Vi har testet estimatoren C_{i3} sammen med de foreslåtte estimatorene i kap. 4. Sammenlignet med disse estimatorene gjorde C_{i3} det dårlig i simuleringsforsøkene. C_{i3} hadde rett og slett for høy varians.

En annen mulighet vi har vurdert, er å estimere p_i i c_i^* med K_i, dvs. vi har vurdert å løse likning (5) m.h.p. c_i, og bruke løsningen som estimator for c_i^*.

$$(5) \qquad c_i = \frac{(1-\frac{n_i}{n})\frac{1}{n_i} K_i(c_i)[1-K_i(c_i)]}{(1-2\frac{n_i}{n})\frac{1}{n_i} K_i(c_i)[1-K_i(c_i)] + \frac{1}{n}\hat{p}(1-\hat{p}) + [K_i(c_i)-\hat{p}]^2}$$

(5) kan løses ved iterative metoder.

Leddet $[K_i(c_i)-\hat{p}]^2 = (1-c_i)^2(D_i-\hat{p})^2$ er tenkt å skulle estimere $(p-p_i)^2$. Hvis vi ser på forventningen til dette leddet finner vi:

$$E[K_i(c_i)-\hat{p}]^2 = (1-c_i)^2 \text{ x nevneren i } c_i^* \neq (p_i-p)^2$$

Vi har derfor at c_i-en(e) som tilfredstiller (5) har en forventning som ikke er garantert å ligge i nærheten av c_i^*. La C_{i4} være løsningen av (5). I simuleringsforsøkene har vi testet C_{i4}. Det viste seg at denne estimatoren fungerte dårlig.

Hvis en forutsetter at alle p_i-ene er forskjellige, og at det ikke er noen funksjonell sammenheng mellom dem, tviler vi på at det er mulig å finne en brukbar estimator for c_i^*. Vi skal derfor gå over til å betrakte det hele fra en annen synsvinkel.

4. FORSLAG TIL ESTIMATORER VED EN ETT-TRINNS UTVALGSPLAN

I dette kapitlet skal vi komme fram til estimatorer for c_i^* ved å gjøre modellforutsetninger om p_i-ene. Estimatorene som blir presentert i dette kapitlet, er utviklet under forutsetning av at en bruker den utvalgsplanen som er beskrevet i kap. 2, og har derfor begrenset praktisk betydning. Modellene og framgangsmåten som er beskrevet i dette kapitlet, er mer interessante enn estimatorene i seg selv.

Estimatorene er utviklet for å anvendes i simuleringsforsøkene som er beskrevet i neste kapittel, og er spesialtilfeller av de mer generelle estimatorene som er beskrevet i kap. 6.

I avsnitt 5.3 er det presentert utgaver av estimatorene som er beregnet til bruk ved Byråets utvalgsplan.

Vi skal nå komme fram til estimatorer for c_i^* ved å ta utgangspunkt i følgende to modeller:

Modell 1: $E(\hat{p}-p_i)^2$ er konstant for alle fylker

Modell 2: $E(\hat{p}-p_i)^2 + E(D_i-p_i)^2 - 2E(D_i-p_i)(\hat{p}-p_i)$ er konstant for alle fylker.

Ut fra erfaring kan en si at bruttovariansen til $\hat{p}$, $E(\hat{p}-p_i)^2$, i de fleste tilfeller er dominert av kvadratet av skjevheten, dvs. $(p-p_i)^2$. Modell 1 baseres da på forutsetningen om at absoluttverdien av skjevheten til $\hat{p}$, $|p-p_i|$, varierer lite fra fylke til fylke. En begrunnelse for denne forutsetningen kan være: Hvis en ikke har noen forhåndskunnskaper om noen av fylkene, er det liten grunn til å påstå at $|p-p_i|$ er større eller mindre enn $|p-p_j|$ for fylkene i og j.

Modell 1 er testet i Schaible (1979) og i Schaible, Brock og Schnack (1977). I nevnte artikler har en egentlig tatt utgangspunkt i den mer generelle modellen som er beskrevet i kap. 6. Ut fra modell 1 har en laget en kombinert estimator, og laget estimater for forskjellige variable ved hjelp av data fra de amerikanske helseundersøkelsene (1969-71). Disse estimatene er så blitt sammenlignet med folketellingsdata fra 1970.

La $\hat{p}_i$ betegne en estimator for parameteren p_i i område nr. i. Som mål for kvaliteten til estimatoren $\hat{p}_i$ har en brukt $Q = \sum_i (\hat{p}_i - p_i)^2$. Med hensyn på målet Q var den kombinerte estimatoren overlegent bedre enn komponentestimatorene. I Schaible (1979) er det i tillegg presentert resultater som viser at den kombinerte estimatoren er robust overfor avvik i modellen.

Når det gjelder modell 2, virker denne urealistisk, og det er vanskelig å gi noen teoretisk begrunnelse for den. Modell 2 er egentlig kommet til ved at vi laget en estimator for c_i^* som var enkel å beregne, og som har gjort det bra under de empiriske forsøkene som er beskrevet i kapittel 5.

Vi skal nå ta utgangspunkt i modell 1 og utvikle en estimator for c_i^*. Fra (4) har vi:

$$c_i^* \approx \frac{(1-\frac{n_i}{n})\frac{1}{n_i} p_i (1-p_i)}{(1-2\frac{n_i}{n})\frac{1}{n_i} p_i (1-p_i) + E(\hat{p}-p_i)^2}.$$

Vi skal lage en estimator for c_i^* ved å lage estimatorer for de tre leddene i c_i^* hver for seg. Som estimator for $\frac{1}{n_i} p_i (1-p_i)$ bruker vi $\frac{1}{n_i}\hat{p}(1-\hat{p})$. Grunnen til at vi ikke bruker estimatoren $\frac{1}{n_i} D_i (1-D_i)$, er at $\frac{1}{n_i} p_i (1-p_i)$ er robust overfor variasjoner i p_i, og at det er enklere beregningsmessig å bruke estimatoren $\frac{1}{n_i}\hat{p}(1-\hat{p})$. I tillegg har $\hat{p}$ ikke vært særlig dårligere enn D_i som estimator for p_i i simuleringsforsøkene som er beskrevet i kap. 5. Dette gjelder spesielt for små utvalgsstørrelser.

Siden $E(\hat{p}-D_i)^2$ = nevneren i c_i^*, har vi at $(\hat{p}-D_i)^2 - (1-2\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})$ er en tilnærmet forventningsrett estimator for $E(\hat{p}-p_i)^2$. Modell 1 gir at også

$$\frac{1}{19}\sum_{j=1}^{19}\left\{(\hat{p}-D_j)^2 - (1-2\frac{n_j}{n})\frac{1}{n_j}\hat{p}(1-\hat{p})\right\} \quad \text{vil være}$$

en tilnærmet forventningsrett estimator for $E(\hat{p}-p_i)^2$. Sistnevnte estimator har langt mindre varians enn førstnevnte.

En tilnærmet forventningsrett estimator for nevneren i c_i^* blir da:

$$\frac{1}{19}\sum_{j=1}^{19}\left\{(\hat{p}-D_j)^2 - (1-2\frac{n_j}{n})\frac{1}{n_j}\hat{p}(1-\hat{p})\right\} + (1-\frac{2n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})$$

$$= \frac{1}{19}\sum_{j=1}^{19}(\hat{p}-D_j)^2 + \hat{p}(1-\hat{p})\left(\frac{1}{n_i} - \frac{1}{19}\sum_j \frac{1}{n_j}\right)$$

En estimator for c_i^* under modell 1 blir:

$$\hat{c}_i^* = \frac{(1-\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})}{\frac{1}{19}\sum_j (D_j-\hat{p})^2 + \hat{p}(1-\hat{p})(\frac{1}{n_i} - \frac{1}{19}\sum_j \frac{1}{n_j})}$$

Hvis modell 1 gjelder, vil $\hat{c}_i^*$ være en tilnærmet forventningsrett estimator for c_i^*.

Under modell 2 vil følgende estimator for c_i^* være tilnærmet forventningsrett:

$$\hat{\hat{c}}_i^* = \frac{(1-\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})}{\frac{1}{19}\sum_j (D_j-\hat{p})^2}$$

$\hat{\hat{c}}_i^*$ har en del likhetstrekk med konstanten "c" i den klassiske James-Stein estimatoren (James og Stein (1961)). Anta at n_i-ene er like. La n_i=k for alle i. $\hat{\hat{c}}_i^*$ blir da

$$\hat{\hat{c}}_i^* = \frac{18\frac{1}{k}\hat{p}(1-\hat{p})}{\sum_j (D_j-\hat{p})^2}$$

c i James-Stein estimatoren ville være:

$$c = \frac{16\frac{1}{k}\hat{p}(1-\hat{p})}{\sum_j (D_j-\hat{p})^2}$$

c og $\hat{\hat{c}}_i^*$ er utviklet under forskjellige forutsetninger og en kan ikke konkludere med at $\hat{\hat{c}}_i^*$ generelt overestimerer c_i^*.

La nå: $$\delta_{1i} = \begin{cases} 0 & \text{hvis } \hat{c}_i^* < 0 \\ \hat{c}_i^* & \text{hvis } \hat{c}_i^* \in [0,1] \\ 1 & \text{hvis } \hat{c}_i^* > 1 \end{cases}$$

og $$\delta_{2i} = \begin{cases} \hat{\hat{c}}_i^* & \text{hvis } \hat{\hat{c}}_i^* \in [0,1] \\ 1 & \text{hvis } \hat{\hat{c}}_i^* > 1 \end{cases}$$

Vårt forslag til en kombinert estimator blir da under modell 1 og 2 henholdsvis:

$$K_{1i} = \delta_{1i}\hat{p} + (1-\delta_{1i})\, D_i$$

$$K_{2i} = \delta_{2i}\hat{p} + (1-\delta_{2i})\, D_i$$

Vi skal i det følgende sammenligne 5 estimatorer for p_i

$$D_i,\ \hat{p},\ K_i^*,\ K_{1i} \text{ og } K_{2i}$$

K_i^* er den kombinerte estimatoren med den optimale vekten c_i^*.

$$K_i^* = c_i^*\hat{p} + (1-c_i^*)\, D_i.$$

c_i^* er beregnet ved hjelp av folketellingsdata fra 1970.

Vi har sammenlignet de 5 estimatorene ved hjelp av simuleringsforsøk basert på folketellingsdata fra 1970. Opplegget for forsøkene blir beskrevet i neste avsnitt.

5. UTPRØVING AV ESTIMATORENE VED HJELP AV SIMULERINGSFORSØK

5.1. Opplegget for simuleringsforsøkene

Som nevnt i kapittel 2 forutsetter vi at vi anvender en utvalgsplan som er slik at det trekkes enkle, tilfeldige utvalg fra hvert fylke. Videre forutsetter vi at antall personer som trekkes fra hvert fylke er proporsjonalt med folketallet i fylket.

I en slik utvalgsplan vil D_i være tilnærmet normalfordelt med forventning p_i og varians $\frac{p_i(1-p_i)}{n_i}$. Fra folketellingen i 1970 har vi funnet fram "p_i-er" for sysselsettingen i enkelte næringer. Hvis vi hadde hatt "D_i-er" fra en utvalgsundersøkelse, kunne vi brukt disse estimatene og målt tilpasningen til de "sanne p_i-ene" for forskjellige estimatorer.

Siden vi ikke hadde passende data fra en utvalgsundersøkelse, har vi laget kunstige "D_i-er". Dette er gjort ved å trekke tall fra en standardisert normalfordeling. Tallene er trukket fra boken "A Million Random Digits with 100 000 Normal Deviates".

Anta at vi har trukket tallet X. Vår D_i er da beregnet på følgende måte

$$D_i = X\sqrt{\frac{p_i(1-p_i)}{n_i}} + p_i$$

En har da at $D_i \approx N\left(p_i,\ \frac{p_i(1-p_i)}{n_i}\right)$

For hver variabel, hver utvalgsstørrelse og hvert fylke har vi trukket tre forskjellige X-er, og altså "laget" tre forskjellige D_i-er (simuleringsforsøk I, II og III).

5.2. Presentasjon av resultatene

La $\hat{p}_i$ være en estimator for p_i. Som mål for estimatorens tilpasning til p_i-ene har vi brukt

$$Q = \sum_{i=1}^{19} (\hat{p}_i - p_i)^2.$$

Denne størrelsen har vi beregnet for estimatorene D_i, $\hat{p}$, K_i^*, K_{i1} og K_{i2} for forskjellige variable og forskjellige utvalgsstørrelser. Tabellene under viser resultatene.

Tabell 1. p_i = andelen sysselsatt totalt av personer over 15 år i fylke nr. i. Beregnet verdi av Q for forskjellige estimatorer og forskjellige utvalgsstørrelser

Estimator	Simuleringsforsøk nr. I	II	III	Gjennomsn. av I, II og III $(\frac{1}{3}(Q_I+Q_{II}+Q_{III}))$
		Utvalgsstørrelse hele landet: 2 000		
D_i	$9{,}357\cdot10^{-2}$	$4{,}858\cdot10^{-2}$	$4{,}806\cdot10^{-2}$	$6{,}340\cdot10^{-2}$
$\hat{p}$	$0{,}996\cdot10^{-2}$	$0{,}995\cdot10^{-2}$	$1{,}444\cdot10^{-2}$	$1{,}145\cdot10^{-2}$
K_{1i}	$2{,}066\cdot10^{-2}$	$0{,}928\cdot10^{-2}$	$1{,}444\cdot10^{-2}$	$1{,}479\cdot10^{-2}$
K_{2i}	$1{,}900\cdot10^{-2}$	$0{,}699\cdot10^{-2}$	$1{,}175\cdot10^{-2}$	$1{,}258\cdot10^{-2}$
K_i^*	$0{,}517\cdot10^{-2}$	$0{,}593\cdot10^{-2}$	$1{,}039\cdot10^{-2}$	$0{,}716\cdot10^{-2}$
		Utvalgsstørrelse hele landet: 5 000		
D_i	$2{,}783\cdot10^{-2}$	$2{,}931\cdot10^{-2}$	$2{,}016\cdot10^{-2}$	$2{,}577\cdot10^{-2}$
$\hat{p}$	$1{,}034\cdot10^{-2}$	$1{,}027\cdot10^{-2}$	$1{,}018\cdot10^{-2}$	$1{,}026\cdot10^{-2}$
K_{1i}	$0{,}711\cdot10^{-2}$	$0{,}746\cdot10^{-2}$	$0{,}511\cdot10^{-2}$	$0{,}656\cdot10^{-2}$
K_{2i}	$0{,}728\cdot10^{-2}$	$0{,}636\cdot10^{-2}$	$0{,}719\cdot10^{-2}$	$0{,}694\cdot10^{-2}$
K_i^*	$0{,}560\cdot10^{-2}$	$0{,}668\cdot10^{-2}$	$0{,}184\cdot10^{-2}$	$0{,}471\cdot10^{-2}$
		Utvalgsstørrelse hele landet: 20 000		
D_i	$0{,}680\cdot10^{-2}$	$0{,}686\cdot10^{-2}$	$0{,}271\cdot10^{-2}$	$0{,}546\cdot10^{-2}$
$\hat{p}$	$1{,}060\cdot10^{-2}$	$1{,}170\cdot10^{-2}$	$1{,}180\cdot10^{-2}$	$1{,}137\cdot10^{-2}$
K_{1i}	$0{,}406\cdot10^{-2}$	$0{,}430\cdot10^{-2}$	$0{,}407\cdot10^{-2}$	$0{,}414\cdot10^{-2}$
K_{2i}	$0{,}434\cdot10^{-2}$	$0{,}450\cdot10^{-2}$	$0{,}502\cdot10^{-2}$	$0{,}462\cdot10^{-2}$
K_i^*	$0{,}144\cdot10^{-2}$	$0{,}250\cdot10^{-2}$	$0{,}254\cdot10^{-2}$	$0{,}216\cdot10^{-2}$

Tabell 2. p_i = andelen sysselsatt i industri m.v. av personer over 15 år i fylke nr. i. Beregnet verdi av Q for forskjellige estimatorer og forskjellige utvalgsstørrelser

Estimator	Simuleringsforsøk nr. I	II	III	Gjennomsn. av I, II og III $(\frac{1}{3}(Q_I,Q_{II}+Q_{III}))$
		Utvalgsstørrelse hele landet: 5 000		
D_i	$1{,}222\cdot10^{-2}$	$0{,}964\cdot10^{-2}$	$1{,}024\cdot10^{-2}$	$1{,}070\cdot10^{-2}$
$\hat{p}$	$2{,}090\cdot10^{-2}$	$2{,}310\cdot10^{-2}$	$2{,}204\cdot10^{-2}$	$2{,}200\cdot10^{-2}$
K_{1i}	$0{,}794\cdot10^{-2}$	$0{,}740\cdot10^{-2}$	$0{,}572\cdot10^{-2}$	$0{,}702\cdot10^{-2}$
K_{2i}	$0{,}793\cdot10^{-2}$	$0{,}779\cdot10^{-2}$	$0{,}577\cdot10^{-2}$	$0{,}717\cdot10^{-2}$
K_i^*	$0{,}656\cdot10^{-2}$	$0{,}632\cdot10^{-2}$	$0{,}376\cdot10^{-2}$	$0{,}555\cdot10^{-2}$
		Utvalgsstørrelse hele landet: 20 000		
D_i	$0{,}202\cdot10^{-2}$	$0{,}206\cdot10^{-2}$	$0{,}446\cdot10^{-2}$	$0{,}285\cdot10^{-2}$
$\hat{p}$	$2{,}107\cdot10^{-2}$	$2{,}094\cdot10^{-2}$	$2{,}162\cdot10^{-2}$	$2{,}121\cdot10^{-2}$
K_{1i}	$0{,}199\cdot10^{-2}$	$0{,}224\cdot10^{-2}$	$0{,}350\cdot10^{-2}$	$0{,}258\cdot10^{-2}$
K_{2i}	$0{,}199\cdot10^{-2}$	$0{,}222\cdot10^{-2}$	$0{,}346\cdot10^{-2}$	$0{,}256\cdot10^{-2}$
K_i^*	$0{,}137\cdot10^{-2}$	$0{,}183\cdot10^{-2}$	$0{,}258\cdot10^{-2}$	$0{,}193\cdot10^{-2}$

Tabell 3. p_i = andelen sysselsatt i varehandel av personer over 15 år i fylke nr. i. Beregnet verdi av Q for forskjellige estimatorer. Utvalgsstørrelse: 12 000 fra hele landet

Estimator	Simuleringsforsøk nr. I	II	III	Gjennomsn. av I, II og III $(\frac{1}{3}(Q_I+Q_{II}+Q_{III}))$
D_i	$1{,}751\cdot10^{-3}$	$1{,}056\cdot10^{-3}$	$1{,}617\cdot10^{-3}$	$1{,}475\cdot10^{-3}$
$\hat{p}$	$5{,}638\cdot10^{-3}$	$5{,}615\cdot10^{-3}$	$5{,}592\cdot10^{-3}$	$5{,}615\cdot10^{-3}$
K_{1i}	$0{,}926\cdot10^{-3}$	$1{,}481\cdot10^{-3}$	$1{,}429\cdot10^{-3}$	$1{,}279\cdot10^{-3}$
K_{2i}	$0{,}986\cdot10^{-3}$	$1{,}854\cdot10^{-3}$	$1{,}667\cdot10^{-3}$	$1{,}502\cdot10^{-3}$
K_i^*	$0{,}620\cdot10^{-3}$	$0{,}774\cdot10^{-3}$	$1{,}029\cdot10^{-3}$	$0{,}808\cdot10^{-3}$

Med hensyn på målet Q er K_i^* den estimatoren som har vært den beste i forsøkene. K_{1i} og K_{2i} som er omtrent like gode, er dårligere enn K_i^*, men bedre enn D_i og $\hat{p}$. Når det gjelder variabelen "sysselsetting i alt", er $\hat{p}$ bedre enn D_i for små utvalgsstørrelser, mens D_i er bedre enn $\hat{p}$ for store utvalgsstørrelser. Q-verdien til $\hat{p}$ varierer lite med utvalgsstørrelsen. Dette har utvilsomt sammenheng med at kvadratet av skjevheten til $\hat{p}$ dominerer bruttovariansen.

I samsvar med de teoretiske resultatene i kap. 2 ser vi at det er mest å tjene på å bruke en kombinert estiamtor når komponentestimatorene er omtrent like gode. I tilfeller der den ene komponentestimatoren er overlegent bedre enn den andre, bør en bruke den beste komponentestimatoren framfor å bruke en av de kombinerte estimatorene K_{1i} og K_{2i}.

K_{1i} og K_{2i} er bedre enn $\hat{p}$ unntatt ved estimering av andelen sysselsatte i alt med utvalgsstørrelse 2 000. Dette indikerer at en for utvalgsstørrelser større enn 2 000 bør foretrekke en kombinert estimator framfor en syntetisk estimator. Når en kommer opp i så store utvalgsstørrelser som 20 000, ser det ut til at det ikke er noe å tjene på å bruke en av estimatorene K_{1i} eller K_{2i} i stedet for D_i.

I forsøkene er det lite som skiller K_{1i} og K_{2i}. Resultatene indikerer imidlertid at K_{1i} kanskje er noe bedre enn K_{2i}. For å undersøke dette nærmere har vi sammenlignet $\Sigma_i(\delta_{1i}-c_i^*)^2$ med $\Sigma_i(\delta_{2i}-c_i^*)^2$ (jfr. kap. 4) for forskjellige variable og forskjellige utvalgsstørrelser. Den første kvadratsummen viste seg i alle tilfellene å være noe mindre enn den siste. Dette skulle indikere at K_{1i} er noe bedre enn K_{2i}. Grunnen til at det har vært så liten forskjell mellom K_{1i} og K_{2i} i forsøkene, er utvilsomt at K_i er robust overfor variasjoner i c_i.

5.3. Forslag til estimatorer ved Byråets utvalgsplan

Byråets utvalgsplan er en totrinns utvalgsplan som ikke er konstruert for å gi fylkestall. Hvis "D_i-ene" hadde framkommet ved Byråets utvalgsplan, ville de hatt større varians enn i forsøkene. I tillegg ville de ha vært forventningsskjeve.

Dette medfører at den "optimale c_i" blir en annen ved Byråets utvalgsplan enn i forsøkene. Vi skal i dette avsnittet presentere estimatorer for den "optimale c_i" som kan brukes ved Byråets utvalgsplan.

La nå E_B, var_B, cov_B og C_i^B betegne henholdsvis forventning, varians, kovarians og den "optimale C_i" ved Byråets utvalgsplan. Fra (1) har vi:

$$(6) \qquad C_i^B = \frac{E_B(D_i-p_i)^2 - E_B(D_i-p_i)(\hat{p}-p_i)}{E_B(D_i-p_i)^2+E_B(\hat{p}-p_i)^2 - 2E_B(D_i-p_i)(\hat{p}-p_i)}$$

Bruttovariansen til $\hat{p}$ som estimator for p_i er i de fleste tilfeller dominert av kvadratet av skjevheten. Siden skjevheten til $\hat{p}$ er den samme ved Byråets utvalgsplan som ved den utvalgsplanen som er beskrevet i kap. 2, vil vi i de fleste tilfeller ha at $E_B(\hat{p}-p_i)^2 \approx E(\hat{p}-p_i)^2$. Med E menes forventningen ved utvalgsplanen som er beskrevet i kap. 2. Vi skal forutsette at tilnærmelsen gjelder, og at $\text{cov}_B(D_i,\hat{p}) \approx \frac{n_i}{n} \text{var}_B D_i$. La $e_i = \frac{E_B(D_i-p_i)^2}{E(D_i-p_i)^2}$. Vi kan kalle e_i for designeffekten til D_i. Fra (4) og (6) har vi da:

$$C_i^B \approx \frac{e_i(1-\frac{n_i}{n})\frac{1}{n_i} p_i(1-p_i)}{e_i(1-2\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i) + \frac{1}{n} p(1-p) + (p-p_i)^2}$$

I praksis har vi nesten alltid at $e_i \geq 1$, som igjen medfører at vi nesten alltid har at $C_i^B \geq c_i^*$, der c_i^* er som i (4). Dette kan tyde på at de foreslåtte estimatorene $\hat{c}_i^*$ og $\hat{c}_i^*$ underestimerer c_i^B ved Byråets utvalgsplan. Vi skal se nærmer på dette:

$$E_B\hat{c}_i^* = E_B\left\{\frac{(1-\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})}{\frac{1}{19}\Sigma(D_j-\hat{p})^2}\right.$$

$$\approx \frac{E_B(1-\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})}{E_B\frac{1}{19}\Sigma_j(D_j-\hat{p})^2}$$

$$\approx \frac{(1-\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i)}{\frac{1}{19}\sum_j\{E_B(D_j-p_j)^2 + E_B(\hat{p}-p_j)^2 - 2E_B(D_j-p_j)(\hat{p}-p_j)}$$

Hvis modell 2 gjelder, har vi at

$$E_B\hat{C}_i^* \approx \frac{(1-\frac{n_i}{n})\frac{1}{n_i}p_i(1-p_i)}{E_B(D_i-p_i)^2 + E_B(\hat{p}-p_i)^2 - 2E_B(D_j-p_j)(\hat{p}-p_j)}$$

$$\approx \frac{C_i^B}{e_i} \leq C_i^B$$

Når det gjelder forventningen til $\hat{C}_i^*$ ved Byråets utvalgsplan, er sammenhengen mellom denne og C_i^B litt mer komplisert. Det kan imdilertid vises at vi i de fleste tilfeller vil ha:

$$\frac{1}{e_i} C_i^B \leq E_B\hat{C}_i^* \leq C_i^B$$

e_i er en størrelse som varierer med fylke, variabel og utvalgsstørrelse. Vi har beregnet e_i for

Troms og Finnmark for forskjellige variable og utvalgsstørrelser. Resultatene er gjengitt i tabell 4 og 5.

Tabell 4. Designeffekten til D_i for Finnmark for forskjellige variable og utvalgsstørrelser

Næring	Utvalgsstørrelse hele landet		
	2 000	5 000	12 000
Sysselsetting i alt	1.19	1.21	1.36
Industri m.v.	1.40	1.73	2.59
Varehandel	1.28	1.33	1.56
Jord og skog	1.63	2.68	5.21
Fiske og hvalfangst	1.45	2.27	4.23
Bygg og anlegg	1.25	1.35	1.67
Samferdsel	1.23	1.26	1.41
Tjenesteytende	1.35	1.53	2.05

Tabell 5. Designeffekten til D_i for Troms for forskjellige variable og utvalgsstørrelser

Næring	Utvalgsstørrelse hele landet		
	2 000	5 000	12 000
Sysselsetting i alt	1.08	1.16	1.36
Industri m.v.	1.06	1.07	1.13
Varehandel	1.12	1.21	1.44
Jord og skog	1.14	1.39	1.99
Fiske og hvalfangst	1.22	1.63	2.60
Bygg og anlegg	1.09	1.20	1.47
Samferdsel	1.05	1.05	1.06
Tjenesteytende	1.36	1.88	3.08

I tabellene ser vi at e_i varierer ganske mye med både variabel og utvalgsstørrelse. Vi kan derfor ikke anbefale noe bestemt tall for e_i som kan brukes ved estimering av c_i^B.

Ved Byråets utvalgsplan foreslår vi følgende estimatorer for c_i^B under henholdsvis modell 1 og 2 (jfr. $\hat{c}_i^*$ og $\hat{\hat{c}}_i^*$ i kap. 4):

a) $$\hat{C}_i^B = \frac{e_i(1-\frac{n_i}{n})\frac{1}{n_i}\hat{p}(1-\hat{p})}{\frac{1}{19}\sum_j(D_j-\hat{p})^2 + e_i\hat{p}(1-\hat{p})(\frac{1}{n_i} - \frac{1}{19}\sum_j\frac{1}{n_j})}$$

under modell 1

b) $\hat{\hat{c}}_i^B = e_i\hat{\hat{c}}_i^*$ under modell 2.

e_i er et tall som en bør finne et anslag for i hvert enkelt tilfelle.

La nå

$$\delta_i' = \begin{cases} 0 & \text{hvis } \hat{c}_i^B < 0 \\ \hat{c}_i^B & \text{hvis } \hat{c}_i^B \in [0,1] \\ 1 & \text{hvis } \hat{c}_i^B > 1 \end{cases}$$

og

$$\delta_i'' = \begin{cases} \hat{c}_i^B & \text{hvis } \hat{c}_i^B \in [0,1] \\ 1 & \text{hvis } \hat{c}_i^B > 1 \end{cases}$$

Vi foreslår følgende kombinerte estimatorer under henholdsvis modell 1 og 2:

$$K_i' = \delta_i'\hat{p} + (1-\delta_i')\hat{p}_i$$

$$K_i'' = \delta_i''\hat{p} + (1-\delta_i'')\hat{p}_i$$

Et naturlig spørsmål å stille nå er hvordan simuleringsforsøkene i kap. 5 ville ha forløpt hvis vi hadde anvendt Byråets utvalgsplan og brukt estimatorene K_i' og K_i''. Som nevnt tidligere mener vi at $\hat{p}$ som estimator for p_i ville ha vært av omtrent samme kvalitet som i kap. 5, mens D_i ville ha vært en del dårligere. K_i' og K_i'' mener vi ville ha vært av noe dårligere kvalitet enn K_{1i} og K_{2i} var i forsøkene. Vi regner med at forringelsen i kvalitet ville ha vært mindre for de kombinerte estimatorene enn for den direkte estimatoren.

Vi kan oppsummere dette med at hvis Byråets utvalgsplan hadde dannet grunnlag for simuleringsforsøkene, tror vi at D_i ville ha kommet dårligere ut i forhold til $\hat{p}$ og de kombinerte estimatorene enn det som var tilfellet. Når det gjelder forholdet mellom $\hat{p}$ og de kombinerte estimatorene, tror vi at de kombinerte estimatorene ville ha gjort det litt dårligere enn i forsøkene.

6. EN GENERALISERING AV ESTIMATORENE

Anta at vi er interessert i å estimere gjennomsnittet y_i i populasjonen i fylke nr. i. La $\bar{X}_i = \frac{1}{n_{ij}} \Sigma X_{ij}$ være gjennomsnittet av observasjonene i en utvalgsundersøkelse fra fylke nr. i. X_{ij} er det som er observert for person nr. j i fylke nr. i. Vi tenker oss at X_{ij} kan være en dikotom variabel, en kategorisk variabel eller en kontinuerlig variabel. La S_i være en syntetisk estimator for fylke nr. i. En kombinert estimator er da:

$$K_i = c_i S_i + (1-c_i)\bar{X}i$$

Den optimale verdi av c_i, kaller vi som før c_i^*. Vi har:

$$c_i^* = \frac{E(\bar{X}_i-y_i)^2 - E(\bar{X}_i-y_i)(S_i-y_i)}{E(\bar{X}_i-y_i)^2 + E(S_i-y_i)^2 - 2E(\bar{X}_i-y_i)(S_i-y_i)}$$

La $\sigma^2(\bar{X}_i)$ = var $\bar{X}_i$ og $\sigma^2(S_i)$ = var S_i

Vi antar at $E\bar{X}_i \approx y_i$.

Vi har da:

$$c_i^* \approx \frac{\sigma^2(\bar{X}_i) - \text{cov}\,(\bar{X}_i, S_i)}{\sigma^2(\bar{X}_i) + \sigma^2(S_i) + (ES_i - y_i)^2 - 2\,\text{cov}\,(\bar{X}_i, S_i)}$$

La $s^2(\bar{X})$, $s^2(S_i)$ og $C(\bar{X}_i, S_i)$ være estimatorer for henholdsvis $\sigma^2(\bar{X}_i)$, $\sigma^2(S_i)$ og cov $(\bar{X}_i, S_i)$.

Vi skal anvende generaliserte utgaver av modellene 1 og 2 som er definert i kap. 4. La nå:

Modell 1: $E(S_i - y_i)^2$ er konstant for alle fylker

Modell 2: $E(S_i - y_i)^2 + E(\bar{X}_i - y_i)^2 - 2E(\bar{X}_i - y_i)(S_i - y_i)$ er konstant for alle fylker.

Vårt forslag til estimator for c_i^* under modell 1 blir da:

$$\hat{c}_i^* = \frac{s^2(\bar{X}_i) - C(\bar{X}_i, S_i)}{s^2(\bar{X}_i) - 2C(\bar{X}_i, S_i) + \frac{1}{19}\sum_{j=1}^{19}\{(S_j - \bar{X}_j)^2 - s^2(\bar{X}_j) - 2C(\bar{X}_j, S_j)\}}$$

Under modell 2 har vi følgende forslag til estimator for c_i^*:

$$\hat{c}_i^* = \frac{s^2(\bar{X}_i) - C(\bar{X}_i, S_i)}{\sum_{j=1}^{19}(S_j - \bar{X}_j)^2}$$

La nå $\delta_i' = \begin{cases} 0 & \text{hvis } \hat{c}_i^* < 0 \\ c_i^* & \text{hvis } 0 \le \hat{c}_i^* \le 1 \\ 1 & \text{hvis } \hat{c}_i^* > 1 \end{cases}$

og la $\delta_i'' = \begin{cases} \hat{c}_i^* & \text{hvis } 0 \le \hat{c}_i^* \le 1 \\ 1 & \text{hvis } \hat{c}_i^* > 1 \end{cases}$

Vårt forslag til en kombinert estimator under modell 1 og 2 blir da henholdsvis

$$K_i' = \delta_i' S_i + (1 - \delta_i')\bar{X}_i$$

$$K_i'' = \delta_i'' S_i + (1 - \delta_i'')\bar{X}_i$$

Vi tror at estimatoren K_i' er noe bedre enn estimatoren K_i''. K_i'' har imidlertid den fordel at den er lettere å beregne. P.g.a. robusthetsegenskapene til den kombinerte estimatoren, mener vi at det vil være liten forskjell mellom K_i' og K_i'' i praksis.

7. DIFFERANSEN MELLOM FYLKESTALL

Anta at vi er interessert i å sammenlikne to fylker, og at vi er interessert i å estimere differansen p_i-p_j. Siden den kombinerte estimatoren delvis er basert på observasjoner fra hele landet, er det naturlig å spørre seg om de kombinerte estimatorene er egnet ved sammenlikning av fylkestall. Vi ønsker å sammenlikne estimatoren K_i-K_j med D_i-D_j som estimator for p_i-p_j.

Vi ser på følgende differans:

$$\Delta = E[D_i-D_j-(p_i-p_j)]^2 - E[K_i-K_j-(p_i-p_j)]^2 = E(D_i-p_i)^2 + E(D_j-P_j)^2 - 2E(D_i-p_i)\,(D_j-p_j) -$$

$$E(K_i-p_i)^2 - E(K_j-p_j)^2 + 2E(K_i-p_i)(K_j-p_j)$$

Leddet $E(D_i-p_i)(D_j-p)$ vil være lite i forhold til de andre leddene. I de fleste tilfellene vil det være 0, så vi setter det lik 0.

For leddet $E(K_i-p_i)(K_j-p_j)$ har vi:

$$E(K_i-p_i)(K_j-p_j) = E(K_i-EK_i)(K_j-EK_j) + (EK_i-p_i)(EK_j-p_j) \approx \text{cov}\,(K_i,K_j) + c_ic_j(ES_i-p_i)(ES_j-p_j)$$

Vi kan skrive Δ på følgende form:

$$\Delta = [E(D_i-p_i)^2 - E(K_i-p_i)^2] + [E(D_j-p_j)^2 - E(K_j-p_j)^2] + \text{cov}\,(K_i,K_j) + c_ic_j(ES_i-p_i)(ES_j-p_j)$$

Dersom c_i ligger nær c_i^* og c_j ligger nær c_j^*, har vi:

$$\Delta \geq \text{cov}\,(K_i,K_j) + c_ic_j(ES_i-p_i)(ES_j-p_j)$$

Konklusjonen vi kan trekke av dette er at dersom de syntetiske estimatorene S_i og S_j begge overestimerer eller begge underestimerer sine respektive estimander (p_i og p_j), vil K_i-K_j være en bedre estimator for p_i-p_j enn D_i-D_j. Dersom den ene av de syntetiske estimatorene underestimerer fylkesandelen og den andre overestimerer fylkesandelen, er det usikkert hvilken av estimatorene K_i-K_j eller D_i-D_j som er best.

8. KONKLUSJONER OG OPPSUMMERING

I kap. 2 studerte vi egenskapene til den kombinerte estimatoren som viste seg å være robust overfor valg av vekter. I kap. 3 ble konklusjonen at det likevel er problematisk å estimere den optimale vekten, c_i^*, direkte.

I dette notatet har vi funnet fram til en framgangsmåte for å estimere c_i^* ved å ta utgangspunkt i to alternative modeller. Framgangsmåten i sin mest generelle form er beskrevet i kap. 6. Den modellen, som vi har kalt modell 1, har vi egentlig hentet fra Schaible, Brock og Schnack (1977). I nevnte artikkel og i Schaible (1979) er det presentert resultater som indikerer at det fungerer bra i praksis å anvende denne modellen.

Ved å ta utgangspunkt i modellene har vi kommet fram til forslag til estimatorer for c_i^* både ved en ett-trinns utvalgsplan, ved Byråets utvalgsplan og ved mer generelle situasjoner.

For hver estimator for c_i^* har vi laget en kombinert estimator. Ved å gjennomføre noen simuleringsforsøk har vi sammenlignet de kombinerte estimatorene med den direkte og den syntetiske estimatoren, $\hat{p}$. Som mål for estimatorenes kvalitet har vi brukt målet Q som er definert i kap. 5. Med hensyn på dette målet var våre foreslåtte kombinerte estimatorer bedre i forsøkene enn den direkte og

den syntetiske estimatoren.

I forsøkene våre måtte vi av praktiske grunner forutsette at vi hadde en ett-trinns utvalgsplan, mens Byråets utvalgsplan er en totrinns utvalgsplan. Vi mener likevel at forsøkene indikerer at en ved "vanlige" utvalgsstørrelser vil oppnå en gevinst ved å bruke en av våre foreslåtte kombinerte estimatorer framfor å bruke den direkte- eller en syntetisk estimator ved estimering av fylkestall. De generelle egenskapene til de kombinerte estimatorene indikerer det samme. Hvis en primært er interessert i å estimere forskjeller mellom fylker, viser resultatene i kap. 7 at det i mange tilfeller er usikkert om en vil oppnå en gevinst ved å bruke en kombinert estimator framfor den direkte estimatoren.

9. LITTERATUR

Carter, G.M. and Rolph, J.E. (1974): "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities". Journal of the American Statistical Association, 69, 880-885.

Fay, R.E. and Herriot, R.A. (1979): "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data". Journal of the American Statistical Ass., 74, p. 269-277.

Laake, P. and Longva, H. K. (1976): "Estimering av sysselsetting i geografiske regioner, om estimatorenes skjevhet, varians og bruttovarians". Statistisk Sentralbyrå. Artikler 88.

Royall, R. M. (1979): "Prediction Models in Small Area Estimation". NIDA Resarh Monograph 24. Papers presented at a workshop conducted by Response Analysis, Princeton, New Jersey, under NIDA Contract No. 271-77-3425.

Schaible, W. L., Brock, D. B. and Schnack, G. A.: "An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics". Proceedings of the American Statistical Association, Social Statistics Section: 1017-1021, 1977.

Schaible, W. L. (1978): "Choosing Weights for Composite Estimators for Small Area Statistics". Reprint from the 1978 Proceedings of the Section on Survey Research Methods.

Schaible, W. L. (1979): "A Composite Estimator for Small Area Statistics". NIDA Research Monograph 24. Papers presented at a workshop conducted by Response Analysis, Princeton, New Jersey, under NIDA Contract No. 271-77-3425.

Siring, E. og Thomsen, I. (1981): "Metoder for estimering av tall for fylker vha. utvalgsundersøkelser". Rapporter 81/6. Statistisk Sentralbyrå.