ARTIKLER 58

STATISTISK SENTRALBYRÅS UTVALGSUNDERSØKELSER: ELEMENTER AV DET MATEMATISKE GRUNNLAGET

Av Jan M. Hoem

THE SAMPLE SURVEYS OF THE CENTRAL BUREAU OF STATISTICS OF NORWAY: BASIC MATHEMATICAL ELEMENTS

OSLO 1973

STATISTISK SENTRALBYRÅ

ARTIKLER FRA STATISTISK SENTRALBYRÅ NR. 58

STATISTISK SENTRALBYRÅS UTVALGSUNDERSØKELSER: ELEMENTER AV DET MATEMATISKE GRUNNLAGET

Av Jan M. Hoem

THE SAMPLE SURVEYS OF THE CENTRAL BUREAU OF STATISTICS OF NORWAY: BASIC MATHEMATICAL ELEMENTS

OSLO 1973

ISBN 82-537-0259-0

FORORD

I de senere år har Statistisk Sentralbyrå gjennomført en rekke undersøkelser der data er hentet inn ved hjelp av den intervjuorganisasjon som ble opprettet i 1967. Innsamling og analyse av slike data må bygge på metoder fra matematisk statistikk. Denne artikkelen gir noe av det elementære matematiske grunnlaget for slike metoder.

Statistisk Sentralbyrå, Oslo, 17. april 1973

Petter Jakob Bjerve

PREFACE

Since the Bureau Interview Survey Division was established in 1967, the Central Bureau of Statistics of Norway has carried out a number of sample surveys. The collecting and analysis of this type of data must be based on methods of mathematical statistics. The present article provides some of the basic mathematics used in such cases.

Central Bureau of Statistics, Oslo 17 April 1973

Petter Jakob Bjerve

INNHOLD

CONTENTS

1. Innledning.

1.1. Statistisk Sentralbyrå gjennomfører regelmessig undersøkelser der dataene er samlet inn gjennom den intervjuorganisasjon som ble opprettet i 1967. Det er vel kjent at data samlet inn på denne måten, er beheftet med samplingfeil. Det er også vel kjent at det er nødvendig å bruke metoder fra matematisk statistikk både når man skal planlegge innsamlingen og når dataene skal analyseres.

Byrået har en standard utvalgsplan som brukes ved de fleste undersøkelser en gjennomfører. Ifølge denne har en plukket ut nærmere 300 primære utvalgsområder, som ligger spredt over hele landet. Hvem som skal intervjues i en konkret undersøkelse, bestemmes ved at en trekker et rent tilfeldig utvalg av husholdninger eller personer (alt avhengig av karakteren av undersøkelsen) blant dem som er bosatt i det enkelte utvalgsområde. (Vi skal gå litt inn på utvalgsplanen i kapittel 2 nedenfor.)

1.2. Denne artikkelen inneholder det mest elementære av den matematiske teorien en bygger på når en skal gjennomføre en enkelt utvalgsundersøkelse ved hjelp av Byråets standard utvalgsplan. Den antyder også forandringer som blir nødvendige i de tilfelle der det er hensiktsmessig å bruke som utvalgsområder de kommuner Byråets primære utvalgsområder ligger i, istedenfor å bruke de primære utvalgsområdene selv.

Artikkelen inneholder nok enkelte resonnementer og formler som ikke har vært skrevet ut akkurat på denne måten før. Det er allikevel ikke noe prinsipielt nytt her.

1.3. Artikkelen behandler utelukkende spørsmål vedrørende samplingfeil ved de vanligste estimatorene benyttet i standardsituasjonen, det vil si estimatorer for totaler og gjennomsnitt i den samlede bestand, og estimatorer for varianser til disse estimatorene. Vi kommer ikke inn på andre analysemetoder, slik som regresjonsanalyse på utvalgsmaterialet o.l. Vi har også sett bort fra andre typer av feil enn samplingfeil, så vi behandler ikke slikt som frafall, responsfeil, registerfeil, osv.

Vi har heller ikke tatt med teori for undersøkelser som gjentas periodisk, slik som for eksempel arbeidskraftundersøkelsene. Disse bygger på roterende utvalg, som ikke diskuteres i denne framstilling.

1.4. En utgir en artikkel med et så pass begrenset siktemål, fordi det er nyttig å få samlet på ett sted det viktigste av det formelverk som trenges på dette nivået, i en form som er skreddersydd for Byråets behov, Stoffet er omfangsrikt nok allikevel. En har dessuten ikke kunnet finne noen tidligere samlet framstilling som passer for formålet.

En bihensikt med artikkelen har vært å gjennomføre en symbolbruk som er bedre enn den vanlige i teorien for utvalgsundersøkelser. På dette punkt har Sverdrup (1964, kapittel XI) vært vår inspirasjonskilde.

1.5. Artikkelen er skrevet med utgangspunkt i en serie preliminære stensilnotater kalt "Variansberegninger ved intervjuundersøkelser" (Statistisk Sentralbyrå, 1972). Jeg vil takke Stein Østerlund Petersen og John Dagsvik for god hjelp med manuskriptkorrektur, både på tidligere notater og (i Petersens tilfelle) på denne artikkelen. Underkapittel 9.7 og kapittel 10 ble opprinnelig skrevet av Dagsvik (Statistisk Sentralbyrå, 1972, Notat IX).

Det er mange muligheter for å begå algebraiske småfeil i dette stoffet. Jeg håper at de viktigste av dem vi har begått, nå er luket vekk. I alle fall er jeg alene ansvarlig for små og store svakheter ved framstillingen her.

2. Kort om Byråets standard utvalgsplan.

Byråets standard utvalgsplan bygger på en to-trinns trekking av dem som skal intervjues. På første trinn har en trukket ut visse faste geografiske områder, som vi skal kalle primære utvalgsområder. De som skal intervjues i det enkelte tilfelle, trekkes så fra et register over dem som bor i disse områdene.

Følgende prinsipper ble valgt for utvelgelsen av primære utvalgsområder.

Hele landet ble delt opp i 1 502 områder av passende størrelse, stort sett slik at hvert område hadde et folketall i nærheten av 2 000 ved folketellingen i 1960. (Oslo utgjorde ett område for seg.) Grunnlaget for avgrensningen av disse områdene var tellingskretsene fra 1960. Tellingskretser ble slått sammen eller delt på passende måte. En søkte i størst mulig utstrekning å sørge for at tett- og spredtbygde strøk kom i hver sine områder. Disse områdene ble så inndelt i strata etter geografisk beliggenhet og næringsstruktur. En valgte å la Oslo utgjøre et stratum for seg. Bergen og Trondheim hadde tre strata hver, og hvert av dem hadde 15 - 18 områder. Utenom Oslo, Bergen og Trondheim var det 40 strata, hvert med 30 - 38 områder.

Fra hvert av strataene i Bergen og Trondheim har man trukket ut tre utvalgsområder rent tilfeldig. På samme måte har man trukket ut seks utvalgsområder fra hvert av strataene utenom de tre største byene. Sammen med Oslo utgjør de uttrukne områdene de primære utvalgsområdene.

Opprinnelig hadde man bare tredjeparten så mange områder i hvert stratum utenom Oslo. De øvrige utvalgsområder er tilføyet senere. I visse situasjoner velger man fortsatt å bruke bare én eller to tredjedeler av de primære utvalgsområdene (utenom Oslo) i en undersøkelse. Størrelsen av utvalget fra Oslo fastsettes da tilsvarende.

(Til å begynne med ble Oslo behandlet noe annerledes enn slik vi har beskrevet det her. Vi skal imidlertid se bort fra dette. Ytterligere detaljer om den opprinnelige utvalgsplan finnes hos Tamsfoss, 1964.)

3. Grunnelementene i den matematiske teorien.

3.1. Grunnleggende symboler og konvensjoner. I Byråets standard utvalgsplan utgjør Oslo et særtilfelle, idet dette primære utvalgsområdet utgjør et stratum for seg selv. Vi skal diskutere dette særskilt i kapittel 5 nedenfor. I resten av framstillingen skal vi se bort fra dette særtilfellet, og skal late som om hele bestanden er delt opp i strata der hvert enkelt stratum har flere utvalgsområder.

La da i-te stratum ha $M_i > 1$ utvalgsområder, og la det j-te av disse ha $N_i(j)$ trekkeenheter. (En trekkeenhet kan være en person eller en husholdning, alt etter undersøkelsens karakter.) Den k-te av disse trekkeenhetene har verdien $a_i(j,k)$ på en størrelse som måles gjennon untervjuet.

Vi lar

$$N_i = \sum_j N_i(j) \qquad \text{(= antall trekkeenheter i stratum } i\text{)},$$

$$N = \sum_i N_i \qquad \text{(= antall trekkeenheter i alt)},$$

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

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

$$a = \sum_i a_i = \sum_i \sum_j \sum_k a_i(j,k).$$

Vår estimand er totalverdien a.

Av de M_i utvalgsområdene i stratum i trekkes m_i stykker rent lotterisk. La numrene på de områdene som trekkes ut, være J_{i1}, J_{i2}, ..., J_{im_i}, og la

$$\underset{\sim}{J}_i = (J_{i1}, J_{i2}, \ldots, J_{im_i}), \quad \underset{\sim}{J} = (\underset{\sim}{J}_1, \underset{\sim}{J}_2, \ldots).$$

Fra hver av de uttrukne utvalgsområdene skal vi trekke et antall trekkeenheter rent lotterisk. Hvor mange trekkeenheter vi trekker fra utvalgsområdene i stratum i, skal få avhenge av hvilke områder vi har trukket ut i alle strataene, men ikke av hvilken rekkefølge de er trukket ut i. Vi betegner antall trekkeenheter som trekkes fra det r-te uttrukne området i stratum i med $n_{ir}(\underset{\sim}{J})$. Vi antar at alle $n_{ir}(J) > 1$.

Numrene på de intervjuenhetene som trekkes ut fra området J_{ir}, vil vi betegne K_{ir1}, K_{ir2}, ..., og vi vil la

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

$$\overline{X}_{ir} = \sum_s X_{irs}/n_{ir}(\underset{\sim}{J}).$$

Vår estimator for a skal da være

$$\hat{a} = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir})\overline{X}_{ir}. \tag{3.1.1}$$

<u>Sats 3.1.2</u>: $\hat{a}$ er forventningsrett.

<u>Bevis</u>: Vi innfører

$$V_{ir} = N_i(J_{ir})\overline{X}_{ir}$$

og

$$\overline{V}_i = \sum_r V_{ir}/m_i .$$

Da er

$$\hat{a} = \sum_i M_i\overline{V}_i. \tag{3.1.3}$$

Ved å anvende den elementære teorien for rent lotteriske utvalg (Sverdrup, 1964, kapittel XI) får vi direkte at

$$E(X_{irs} \mid \underset{\sim}{J}_i = \underset{\sim}{j}_i) = E(\overline{X}_{ir} \mid \underset{\sim}{J}_i = \underset{\sim}{j}_i) = \overline{a}_i(j_{ir}),$$

der j_{ir} er det r-te elementet i vektoren $\underset{\sim}{j}_i$. Videre blir

$$E(V_{ir} \mid \underset{\sim}{J}_i = \underset{\sim}{j}_i) = a_i(j_{ir}) \tag{3.1.4}$$

og følgelig

$$EV_{ir} = Ea_i(J_{ir}) = \overline{a}_i, \tag{3.1.5}$$

idet

$$P\{J_{ir} = j_{ir}\} = 1/M_i.$$

Satsen følger umiddelbart av dette. $\square$

<u>3.2. Varianser.</u> Vi lar

$$\sigma_i^2(j) = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k) - \overline{a}_i(j)\}^2,$$

$$\tau^2_{ir}(\underset{\sim}{j}) = \text{var}\ (\bar{X}_{ir} \mid \underset{\sim}{J} = \underset{\sim}{j}) = \frac{\sigma^2_i\ (j_{ir})}{n_{ir}(\underset{\sim}{j})}\ \frac{N_i(j_{ir}) - n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})},$$

$$\sigma^2_i = \frac{1}{M_i - 1} \sum_j \{a_i(j) - \bar{a}_i\}^2,$$

og

$$\gamma^2_i = \sum_{\underset{\sim}{j}} N^2_i(j_{ir})\ \tau^2_{ir}(\underset{\sim}{j}) \ / \ \pi[M\ (M-1)\ \ldots\ (M - m + 1)].$$

Da er

$$\text{var}\ (X_{irs} \mid \underset{\sim}{J} = \underset{\sim}{j}) = \sigma^2_i(j_{ir})\{N_i(j_{ir}) - 1\}/N_i(j_{ir}).$$

Hvis

$$S^2_{ir}(\underset{\sim}{j}) = \frac{1}{n_{ir}(\underset{\sim}{j}) - 1} \sum_s \{X_{irs} - \bar{X}_{ir}\}^2,$$

er

$$E\ \{S^2_{ir}\ (\underset{\sim}{J}) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \sigma^2_i\ (j_{ir}).$$

Vi skal nå vise at

(3.2.1) $$\text{cov}\ (V_{ir}, V_{is}) = -\ \sigma^2_i/M_i,$$

(3.2.2) $$\text{var}\ \bar{V}_i = \frac{1}{m_i}\ \{\gamma^2_i + (1 - \frac{m_i}{M_i})\ \sigma^2_i\},$$

og

(3.2.3) $$\text{var}\ \hat{a} = \sum_i M^2_i\ \text{var}\ \bar{V}_i.$$

Størrelsen γ^2_i er uavhengig av r på grunn av våre forutsetninger om $n_{ir}(\underset{\sim}{j})$. Formelen (3.2.2) representerer en oppspaltning av var $\bar{V}_i$ i ett ledd som gir variasjonen i a-verdiene <u>mellom trekkeenhetene</u> (i stratum i), og ett ledd som gir variasjonen i a-verdiene <u>mellom utvalgsområdene</u> (i stratumet).

B e v i s e r: Vi får at

$$E(V_{ir}V_{is}) = EE(V_{ir}V_{is} \mid \underset{\sim}{J}) = E[E(V_{ir} \mid \underset{\sim}{J})\ E(V_{is} \mid \underset{\sim}{J})]$$

$$= E[a_i(J_{ir})\ a_i(J_{is})] \text{ ifølge (3.1.4)}$$

$$= \sum_{j \neq k} a_i(j) a_i(k)\ P\ \{J_{ir} = j,\ J_{is} = k\}$$

$$= \frac{1}{M_i(M_i - 1)} \sum_j a_i(j)\ \{\sum_k a_i(k) - a_i(j)\}$$

$$= \frac{1}{M_i(M_i-1)} \{M_i^2 \bar{a}_i^2 - \sum_j a_i^2(j)\}.$$

Herav følger (3.2.1).

Vi bemerker så at

$$\text{var } V_{ir} = E \text{ var } (V_{ir}|\underset{\sim}{J}) + \text{var } E(V_{ir}|\underset{\sim}{J}).$$

Nå er

$$E \text{ var } (V_{ir}|\underset{\sim}{J}) = E[N_i(J_{ir})^2 \, \tau_{ir}^2(\underset{\sim}{J})] = \gamma_i^2,$$

idet

$$P\,\{\underset{\sim}{J}_i = \underset{\sim}{j}_i\} = 1/[M_i(M_i-1) \dots (M_i-m_i+1)]$$

og $\underset{\sim}{J}_1, \underset{\sim}{J}_2, \dots$ er stokastisk uavhengige. Videre er

$$\text{var } E(V_{ir}|\underset{\sim}{J}) = \text{var } a_i(J_{ir}) = E\, a_i^2(J_{ir}) - \bar{a}_i^2 = \sigma_i^2 \frac{M_i-1}{M_i}.$$

Altså blir

(3.2.4) $$\text{var } V_{ir} = \gamma_i^2 + \sigma_i^2(1-\frac{1}{M_i}).$$

Men

$$\text{var } \bar{V}_i = \frac{1}{m_i^2}\{\sum_r \text{var } V_{ir} + \sum_{r \neq s} \text{cov } (V_{ir}, V_{is})\}$$

$$= \frac{1}{m_i^2}\{m_i(\gamma_i^2 + \sigma_i^2 \frac{M_i-1}{M_i}) - m_i(m_i-1)\frac{\sigma_i^2}{M_i}\},$$

og herav følger (3.2.2). For å bevise (3.2.3) er det nok å vise at

(3.2.5) $$\text{cov } (\bar{V}_i, \bar{V}_k) = 0 \qquad \text{når } i \neq k.$$

Selv om $\bar{V}_i$ og $\bar{V}_k$ kan være stokastisk avhengige, er de altså i hvert fall ukorrelerte. Beviset for (3.2.5) går slik:

Når $\underset{\sim}{J}$ er gitt, er V_{ir} og V_{ks} stokastisk uavhengige (for $i \neq k$). Altså er

$$E\,(V_{ir} V_{ks}|\underset{\sim}{J} = \underset{\sim}{j}) = E\,(V_{ir}|\underset{\sim}{J} = \underset{\sim}{j})\; E\,(V_{ks}|\underset{\sim}{J} = \underset{\sim}{j})$$
$$= a_i(j_{ir})\; a_k(j_{ks}),$$

og følgelig blir

$$E\, V_{ir} V_{ks} = E\, a_i(J_{ir})\; a_k(J_{ks}) = E\, a_i(J_{ir}) \cdot E\, a_k(J_{ks})$$

$$= E\, V_{ir} \cdot E\, V_{ks}.$$

Siden $E\, \bar{V}_i \bar{V}_k = \sum_r \sum_s E\, V_{ir} V_{ks} = \sum_r \sum_s E\, V_{ir} \cdot E\, V_{ks} = E\, \bar{V}_i \cdot E\, \bar{V}_k$, gjelder (3.2.5). $\square$

3.3. Estimering av variansene. La

$$T_{ir}^2(j) = \frac{S_{ir}^2(j)}{n_{ir}(j)} \frac{N_i(j_{ir}) - n_{ir}(j)}{N_i(j_{ir})},$$

og la

$$G_i^2 = \frac{1}{m_i} \sum_r N_i^2(J_{ir})\, T_{ir}^2(j).$$

Da ser vi umiddelbart at G_i^2 er en forventningsrett estimator for γ_i^2.

Hvis vi lar

$$U_i^2 = \sum_r (V_{ir} - \overline{V}_i)^2 / (m_i - 1),$$

blir

$$EU_i^2 = E\,\{\sum_r (V_{ir} - \overline{a}_i)^2 - m_i\,(\overline{V}_i - \overline{a}_i)^2\}/(m_i - 1)$$

$$= m_i\,\{\text{var } V_{ir} - \text{var } \overline{V}_i\}/(m_i - 1) = \gamma_i^2 + \sigma_i^2.$$

Hvis $m_i > 1$ blir derfor

$$S_i^2 = U_i^2 - G_i^2$$

en forventningsrett estimator for σ_i^2. (Hvis dette gir $S_i^2 < 0$, bør høyre side her erstattes med 0. Tilfellet $m_i = 1$ behandles i kapittel 11.)

Altså blir

$$W_i^2 = \frac{1}{m_i}\,\{G_i^2 + (1 - \frac{m_i}{M_i})\, S_i^2\}$$

en forventningsrett estimator for var $\overline{V}_i$, og

(3.3.1) $$\sum_i M_i^2\, W_i^2$$

blir en forventningsrett estimator for var $\hat{a}$. Dersom $S_i^2 \geqq 0$, blir forøvrig

$$W_i^2 = \frac{G_i^2}{M_i} + (\frac{1}{m_i} - \frac{1}{M_i})\, U_i^2 .$$

Følgelig, hvis alle $S_i^2 \geqq 0$, blir

(3.3.2) $$\sum_i M_i^2\, W_i^2 = \sum_i M_i\, G_i^2 + \sum_i (\frac{M_i}{m_i} - 1)\, M_i\, U_i^2 .$$

3.4. Spesialtilfellet med binære a-verdier. Hvis alle $a_i(j,k)$ antar verdiene 0 eller 1, vil et par av de foregående uttrykkene forenkle seg en del. Vi innfører nå et par ekstra betegnelser, nemlig

$$p_i(j) = \bar{a}_i(j) \quad \text{og} \quad p = a/N.$$

Vår estimand vil være p, men transformasjonen fra a er triviell. Vi får nå

$$\sigma_i^2(j) = p_i(j)\,[1 - p_i(j)]\,\frac{N_i(j)}{N_i(j)-1},$$

og

$$\sum_s \{X_{irs} - \bar{X}_{ir}\}^2 = n_{ir}(\underset{\sim}{j})\,\bar{X}_{ir}(1-\bar{X}_{ir}).$$

4. Byråets spesielle valg av $n_{ir}(\underset{\sim}{J})$.

4.1. Selvveiende utvalg. I det foregående har vi ikke sagt noe særlig om hvordan $n_{ir}(\underset{\sim}{J})$ velges. Vi skal nå beskrive den metode som brukes i Byråets standard utvalgsplan. Denne metoden medfører en god del forenklinger i formelverket.

Det samlede antall trekkeenheter i utvalget blir

(4.1.1) $$n = \sum_i \sum_r n_{ir}(\underset{\sim}{J}).$$

En vil fastsette n ut fra overlegninger om den økonomiske ramme for undersøkelsen. Vi lar derfor n være en fiksert, ikke-stokastisk størrelse.

Ifølge (3.1.1) framkommer $\hat{a}$ som en veid sum av observasjonene X_{irs}, med vekter på formen

$$\frac{M_i}{m_i} \cdot \frac{N_i(J_{ir})}{n_{ir}(\underset{\sim}{J})}.$$

Generelt vil disse vektene være forskjellige. Vi vil si at utvalget er selvveiende hvis $n_{ir}(\underset{\sim}{J})$ velges slik at disse vektene er like og samtidig (4.1.1) gjelder. Dét oppnår vi hvis vi innfører

(4.1.2) $$b(\underset{\sim}{J}) = n/\sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}),$$

og

(4.1.3) $$b_i(\underset{\sim}{J}) = b(\underset{\sim}{J})\, M_i/m_i\,,$$

og lar

(4.1.4) $$n_{ir}(\underset{\sim}{J}) = b_i(\underset{\sim}{J})\, N_i(J_{ir}).$$

Da blir

(4.1.5) $$\hat{a} = \sum_i \sum_r \sum_s X_{irs}/b(\underset{\sim}{J}).$$

Det er vanlig å omtale $b(\underset{\sim}{J})$ som <u>total utvalgsbrøk</u> og $b_i(\underset{\sim}{J})$ som <u>utvalgsbrøken i stratum nr. i</u>.

(Se kapittel 12 vedrørende en annen måte å bestemme n_{ir} på.)

<u>4.2. Tolkning av utvalgsbrøkene</u>. Det er interessant å se hvordan man kan tolke disse brøkene. La oss innføre

$$I_{ij}(\underset{\sim}{J}_i) = \begin{cases} 1 & \text{hvis utvalgsområde nr. } j \text{ i stratum nr. } i \text{ trekkes ut som et primært utvalgsområde} \\ 0 & \text{ellers,} \end{cases}$$

og la B_{ijk} være den begivenhet at trekkeenhet nr. k i utvalgsområde nr. j i stratum nr. i trekkes ut for intervjuing. Da er altså $I_{ij} = 1$ hvis og bare hvis ett av elementene i $\underset{\sim}{J}_i$ er lik j. Vi kan ikke få B_{ijk} uten at $I_{ij} = 1$.

Da blir

$$b_i(\underset{\sim}{j}) = P\,\{B_{ijk} | I_{ij} = 1,\ \underset{\sim}{J} = \underset{\sim}{j}\} = \frac{n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})}.$$

Siden $P\{B_{ijk} | I_{ij} = 0,\ \underset{\sim}{J} = \underset{\sim}{j}\} = 0$, kan vi skrive

$$P\,\{B_{ijk} | \underset{\sim}{J} = \underset{\sim}{j}\} = I_{ij}(\underset{\sim}{j}_i)\ b(\underset{\sim}{j})\ \frac{M_i}{m_i}.$$

Idet $P\{I_{ij} = 1\} = m_i/M_i$, blir derfor

$$P\,\{B_{ijk}\} = E\,\{I_{ij}(\underset{\sim}{J}_i)\ b(\underset{\sim}{J})\}\ \frac{M_i}{m_i}.$$

(4.2.1) $$= E\,\{b(\underset{\sim}{J}) | I_{ij} = 1\}.$$

Vi må regne med at denne vanligvis avhenger av i (men ikke av j), <u>slik at sannsynligheten for uttrekking av en trekkeenhet varierer fra stratum til stratum</u>. Spesielt vil <u>ikke</u> $P\{B_{ijk}\}$ generelt være lik for eksempel $E\ b(\underset{\sim}{J})$.

Forøvrig ser vi lett at

(4.2.2) $$E\ \frac{1}{b(\underset{\sim}{J})} = \frac{N}{n}.$$

<u>4.3. Andre forenklinger i formelverket</u>. En del andre formler vil også forenkles når en bruker (4.1.4). En får

$$\tau_{ir}^2(\underset{\sim}{j}) = \frac{1-b_i(\underset{\sim}{j})}{b_i(\underset{\sim}{j})}\ \frac{\sigma_i^2(j_{ir})}{N_i(j_{ir})}\ ;$$

$$V_{ir} = X_{ir}/b_i, \quad \text{der } X_{ir} = \sum_s X_{irs};$$

$$\overline{V}_i = X_i/(b_i m_i), \quad \text{der } X_i = \sum_r X_{ir} = \sum_r \sum_s X_{irs};$$

$$T_i^2(\underset{\sim}{j}) = \frac{1-b_i(\underset{\sim}{j})}{b_i(\underset{\sim}{j})} \, \frac{S_i^2(\underset{\sim}{j})}{N_i(\underset{\sim}{j})},$$

der vi har foretatt naturlig forenkling av symbolene $S_{ir}^2(\underset{\sim}{j})$ og $T_{ir}^2(\underset{\sim}{j})$;

$$G_i^2 = \frac{1}{m_i} \, \frac{1-b_i}{b_i} \sum_r N_i(J_{ir}) \, S_i^2(\underset{\sim}{J});$$

og

$$U_i^2 = Z_i^2/b_i^2, \quad \text{der } Z_i^2 = \frac{1}{m_i-1} \sum_r \{X_{ir} - X_i/m_i\}^2.$$

Estimatoren for var $\hat{a}$ får formen

(4.3.1) $$\frac{1}{b^2(\underset{\sim}{J})} \sum_i \left[1-b_i(\underset{\sim}{J})\right] b_i(\underset{\sim}{J}) \sum_r N_i(J_{ir}) \, S_{ir}^2(\underset{\sim}{J}) + \frac{1}{b^2(\underset{\sim}{J})} \sum_i \left(1-\frac{m_i}{M_i}\right) m_i \, Z_i^2 .$$

4.4. Tilfellet $m_i = 2$. Når det bare er to utvalgsområder i et stratum, får noen av formlene en særlig enkel form. Da blir

$$U_i^2 = \tfrac{1}{2}(V_{i1} - V_{i2})^2, \quad Z_i^2 = \tfrac{1}{2}(X_{i1} - X_{i2})^2;$$

og estimatoren for var $\hat{a}$ blir

$$\frac{1}{b^2} \sum_i (1-b_i) \, b_i \sum_{r=1}^{2} N_i(J_{ir}) \, S_i^2(J_{ir}) + \frac{1}{b^2} \sum_i \left(1-\frac{2}{M_i}\right) (X_{i1} - X_{i2})^2 .$$

5. Særtilfellet Oslo.

Den foregående teorien gjelder for strataene utenom Oslo. Størrelsen a kan tolkes som samlet a-verdi for landet utenom Oslo, osv. Vi skal nå behandle Oslo separat.

Oslo har N_0 trekkeenheter. Av disse trekkes et rent lotterisk utvalg på

$$n_0(\underset{\sim}{J}) = b(\underset{\sim}{J}) \, N_0$$

trekkeenheter. (Nevneren i (4.1.2) forutsettes utvidet med N_0.) k-te trekkeenhet i Oslo har a-verdien $a_0(k)$. Vi lar

$$a_0 = \sum_k a_0(k) \quad \text{og} \quad \overline{a}_0 = a_0/N_0.$$

La de n_0 trekkeenhetene i Oslo-utvalget være nr. $K_{01}, K_{02}, \ldots, K_{0,n_0}$, og la

$$X_{0s} = a_0(K_{0s}) \text{ og } \overline{X}_0 = \sum_s X_{0s}/n_0.$$

Vår estimator for a_0 er

$$\hat{a}_0 = N_0 \overline{X}_0.$$

Denne er forventningsrett. Vi finner variansen slik:

$$\text{var } (\overline{X}_0 | \underset{\sim}{J} = \underset{\sim}{j}) = \frac{\sigma_0^2}{n_0(\underset{\sim}{j})} (1 - \frac{n_0(\underset{\sim}{j})}{N_0}),$$

og

$$E \text{ var } (\overline{X}_0 | \underset{\sim}{J}) = \sigma_0^2 (E \frac{1}{n_0(\underset{\sim}{J})} - \frac{1}{N_0})$$

$$= \frac{\sigma_0^2}{N_0} (E \frac{1}{b(\underset{\sim}{J})} - 1) = \frac{\sigma_0^2}{N_0} (\frac{N}{n} - 1).$$

Siden var $E (\overline{X}_0 | \underset{\sim}{J}) = 0$, blir derfor

$$\text{var } \hat{a}_0 = \frac{\sigma_0^2}{n} N_0 (N-n).$$

En forventningsrett estimator for σ_0^2 er

$$S_0^2 = \frac{1}{n_0 - 1} \sum_s \{X_{0s} - \overline{X}_0\}^2.$$

Vår estimator for var $\hat{a}_0$ blir derfor $S_0^2 N_0 (\frac{N}{n} - 1)$.

Samlet a-verdi for landet blir $\tilde{a} = a + a_0$. Vår estimator for denne er

$$\hat{\tilde{a}} = (\sum_i \sum_r \sum_s X_{irs} + \sum_s X_{0s})/b(\underset{\sim}{J}),$$

og var $\hat{\tilde{a}}$ = var $\hat{a}$ + var $\hat{a}_0$, idet cov $(\hat{a}, \hat{a}_0) = 0$.

I det følgende skal vi se bort fra dette særtilfellet, bortsett fra i underkapittel 12.4, som inneholder noen ytterligere kommentarer.

6. Estimering av prosentfordelinger.

6.1. Innledning. I analyser av data fra en intervjuundersøkelse vil man ofte ønske å studere hvordan trekkeenheter som har et gitt kjennetegn, fordeler seg etter et annet kjennetegn. (For eksempel kan man være interessert i fordelingen av pendlere etter hustypen de bor i.) Den naturlige framgangsmåten er da å telle opp hvor mange trekkeenheter det er som har det første kjennetegnet, og så å regne ut hvor mange prosent det er av disse som har hver av verdiene av det andre kjennetegnet. Det er viktig å være oppmerksom på at både teller og nevner er stokastiske i den brøken som gir et slikt prosenttall. En brøk av denne typen kan betraktes som en estimator for en tilsvarende parameter. Vi skal nå se nærmere på slike estimatorer.

6.2. En særlig enkel situasjon. Vi skal først betrakte en særlig enkel situasjon, der det ikke er noen stratuminndeling og der trekkeenhetene trekkes rent lotterisk fra originalbestanden. (Det er altså ikke tale om å trekke fra utvalgsområder.) Bestanden består av N trekkeenheter, hvorav n trekkes ut og utgjør utvalget. La antall trekkeenheter med et gitt kjennetegn (f.eks. antall pendlere) være M i bestanden og X i utvalget. Av de M har M_i verdien i på et gitt, annet kjennetegn (f.eks. bor det M_i pendlere i hustype nr. i), og X_i av disse er med i utvalget. Her er naturligvis $M = \Sigma M_i$ og $X = \Sigma X_i$. Vi lar

$$p = M/N \quad \text{og} \quad q_i = M_i/M,$$

og vi ønsker å estimere $q_1, q_2, \ldots$. Vi vet at

$$P\{X = x\} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

og

$$P\{X_i = x_i \mid X = x\} = \frac{\binom{M_i}{x_i}\binom{M-M_i}{x-x_i}}{\binom{M}{x}},$$

og vi antar at vi i begge tilfelle kan tillate oss å erstatte det hypergeometriske uttrykket med det tilsvarende binomiske, slik at vi med god tilnærmelse kan regne som om

(6.2.1) $$P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0,1,\ldots,n,$$

(6.2.2) $$P\{X_i = x_i \mid X = x\} = \binom{x}{x_i} q_i^{x_i} (1-q_i)^{x-x_i} \quad \text{for } x_i = 0,1,\ldots,x.$$

Den naturlige estimator for q_i er

(6.2.3) $$q_i^* = X_i/X,$$

medmindre X er "for liten". For små X vil man antakelig avstå fra å estimere q_i. La oss anta at (6.2.3) bare brukes når X er så stor at den numeriske forskjellen mellom q_i^* og

(6.2.4) $$\hat{q}_i = \frac{X_i}{X+1}$$

er neglisjerbar. Vi vil her basere våre resonnementer på $\hat{q}_i$. Vi ser at

$$E(\hat{q}_i \mid X = x) = \frac{x}{x+1} q_i,$$

så når $X = x$ og $x/(x+1) \approx 1$, er $\hat{q}_i$ tilnærmet forventningsrett. Under (6.2.2) blir videre

$$\text{var } (\hat{q}_i \mid X = x) = \frac{x}{(x+1)^2} q_i(1-q_i),$$

som er tilnærmet lik $q_i(1-q_i)/(x+1)$ når $x/(x+1) \approx 1$. Hvis n og p er så pass store at

(6.2.5) $$P\{\frac{X}{X+1} \approx 1\} \approx 1,$$

blir

(6.2.6) $$E\,\hat{q}_i \approx q_i$$

og

$$\text{var } \hat{q}_i \approx q_i(1-q_i)\, E\, \frac{1}{X+1}.$$

Nå er

$$E\,\frac{1}{X+1} = \sum_{x=0}^{n} \frac{1}{x+1} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \frac{1}{n+1} \sum_{y=1}^{n+1} \binom{n+1}{y} p^{y-1} (1-p)^{n+1-y}$$

$$= \frac{1}{p(n+1)} \{\sum_{y=0}^{n+1} \binom{n+1}{y} p^y (1-p)^{n+1-y} - (1-p)^{n+1}\}$$

$$= \frac{1-(1-p)^{n+1}}{p(n+1)} \approx \frac{1}{p(n+1)} \approx \frac{1}{np}.$$

Under (6.2.1), (6.2.2) og (6.2.5) blir derfor

(6.2.7) $$\text{var } \hat{q}_i \approx \frac{q_i(1-q_i)}{np}.$$

En opplagt estimator for høyre side i (6.2.7) er

$$q_i^* (1-q_i)^*/X$$

når X er så stor at en vil estimere q_i. Dette er den samme estimatoren som den en ville bruke dersom X var ikke-stokastisk. For <u>estimatene</u> for

$\mathrm{var}(\hat{q}_i)$ spiller det altså ingen rolle om en feilaktig oppfatter X som "antall forsøk" i en hypergeometrisk (eller tilnærmet binomisk) situasjon.

6.3. Grunnstørrelser når en bruker Byråets utvalgsplan. Hensikten med framstillingen i avsnitt 6.2 ovenfor var å bringe klart fram i dagen ett hovedpoeng, nemlig dét at "interne" prosentfordelinger som regnes ut i materialet, framkommer som forholdstall der både teller og nevner er stokastisk. I den særlig enkle situasjonen vi der studerte, var det relativt lett å behandle de spørsmål vi tok opp. Vi vender oss nå til den situasjon en står overfor når dataene samles inn etter Byråets standard utvalgsplan. Det viser seg at vi må bruke en teknikk som er nokså forskjellig fra det ovenstående.

Vi trenger et matematisk verktøy, og starter med å innføre noen ytterligere symboler.

Siden vi er interessert i om den enkelte intervjuenhet har et visst kjennetegn eller ikke, skal vi la $a_i(j,k)$ være en binær variabel. Eksempelvis lar vi $a_i(j,k) = 1$ hvis intervjuenhet nr. (i,j,k) er pendler, mens $a_i(j,k) = 0$ ellers.

Vi skal jo studere hvordan intervjuenheter som har et gitt kjennetegn, fordeler seg etter verdiene av et gitt annet kjennetegn. La oss si at dette andre kjennetegnet har v verdier. Hvis $a_i(j,k) = 1$ og hvis intervjuenhet nr. (i,j,k) har verdien ℓ av dette andre kjennetegnet, lar vi $a_i(j,k,\ell) = 1$. Ellers lar vi $a_i(j,k,\ell) = 0$. Hvis $a_i(j,k) = 0$, er altså alle $a_i(j,k,\ell) = 0$, mens nøyaktig én av dem er lik 1 hvis $a_i(j,k) = 1$. Da blir

(6.3.1) $$a_i(j,k) = \sum_\ell a_i(j,k,\ell).$$

For hver ℓ har vi nå en teori analog med den i de foregående notatene. Vi innfører symboler parallelle med dem vi har fra før, og legger bare til argumentet ℓ der det trengs. Således lar vi

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

$$\bar{a}_i(j,\cdot,\ell) = a_i(j,\cdot,\ell)/\,N_i(j),$$

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

$$a(\ell) = \sum_i a_i(\cdot,\ell) = \sum_i \sum_j \sum_k a_i(j,k,\ell),$$

$$\sigma_i^2(j,\ell) = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k,\ell) - \bar{a}_i\,(j,\cdot,\ell)\}^2$$

$$\tau_{ir}^2(\underset{\sim}{j},\ell) = \frac{\sigma_i^2(j_{ir},\ell)}{n_{ir}(\underset{\sim}{j})} \; \frac{N_i(j_{ir}) - n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})}$$

$$\gamma_i^2(\ell) = \sum_{\underset{\sim}{j}} N_i^2(j_{ir}) \, \tau_{ir}^2(\underset{\sim}{j},\ell) \Big/ \prod_k \left[M_k(M_k-1) \ldots (M_k - m_k + 1)\right],$$

og

$$\sigma_i^2(\ell) = \frac{1}{M_i-1} \sum_j \{a_i(j,\cdot,\ell) - \bar{a}_i(\cdot,\ell)\}^2.$$

(Vi har i noen tilfelle innført $\cdot$-notasjonen for å forhindre forveksling med tidligere symboler.)

Våre primære stokastiske variable blir nå observasjonene

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

og observatorene

$$\bar{X}_{ir}(\ell) = \sum_s X_{irs}(\ell) / n_{is}(\underset{\sim}{J}),$$

$$V_{ir}(\ell) = N_i(J_{ir}) \, \bar{X}_{ir}(\ell),$$

$$\bar{V}_i(\ell) = \sum_r V_{ir}(\ell) / m_i,$$

$$S_{ir}^2(\underset{\sim}{J},\ell) = \frac{1}{n_{ir}(\underset{\sim}{J}) - 1} \sum_s \{X_{irs}(\ell) - \bar{X}_{ir}(\ell)\}^2,$$

$$T_{ir}^2(\underset{\sim}{J},\ell) = \frac{S_{ir}^2(\underset{\sim}{J},\ell)}{n_{ir}(\underset{\sim}{J})} \; \frac{N_i(J_{ir}) - n_{ir}(\underset{\sim}{J})}{N_i(J_{ir})},$$

$$G_i^2(\ell) = \frac{1}{m_i} \sum_r N_i^2(J_{ir}) \, T_{ir}^2(\underset{\sim}{J},\ell),$$

$$U_i^2(\ell) = \frac{1}{m_i-1} \sum_r \{V_{ir}(\ell) - \bar{V}_i(\ell)\}^2,$$

$$S_i^2(\ell) = \max\,\{0, U_i^2(\ell) - G_i^2(\ell)\},$$

og

$$W_i^2(\ell) = \frac{1}{m_i} \{G_i^2(\ell) + (1 - \frac{m_i}{M_i}) \, S_i^2(\ell)\}.$$

Vi får

$$E\,\{\bar{X}_{ir}(\ell) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \bar{a}_i(j_{ir},\cdot,\ell),$$

(6.3.2) $\quad \text{var}\,\{\bar{X}_{ir}(\ell) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \tau_{ir}^2(\underset{\sim}{j},\ell),$

(6.3.3) $$\hat{a}(\ell) = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir})\, \overline{X}_{ir}(\ell),$$

$$\text{est var } \hat{a}(\ell) = \sum_i M_i^2\, W_i^2(\ell),$$

osv. Nå vil ikke $\overline{X}_{ir}(\ell)$ og $\overline{X}_{ir}(\ell')$ være uavhengige når $\ell \neq \ell'$, så vi skal finne noen kovarianser. La oss da innføre

$$\sigma_i(j,\ell,\ell') = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k,\ell) - \overline{a}_i(j,\cdot,\ell)\}\{a_i(j,k,\ell') - \overline{a}_i(j,\cdot,\ell')\},$$

$$\tau_{ir}(\underset{\sim}{j},\ell,\ell') = \frac{\sigma_i(j_{ir},\ell,\ell')}{n_{ir}(\underset{\sim}{j})} \cdot \frac{N_i(j_{ir}) - n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})},$$

$$\gamma_i(\ell,\ell') = \sum_{\underset{\sim}{j}} N_i^2(j_{ir})\, \tau_{ir}(\underset{\sim}{j},\ell,\ell') / \prod_k \left[M_k(M_k-1)\ldots(M_k-m_k+1)\right],$$

og

$$\sigma_i(\ell,\ell') = \frac{1}{M_i-1} \sum_j \{a_i(j,\cdot,\ell) - \overline{a}_i(\cdot,\ell)\}\{a_i(j,\cdot,\ell') - \overline{a}_i(\cdot,\ell')\}.$$

Her blir $\sigma_i^2(j,\ell) = \sigma_i(j,\ell,\ell)$, $\sigma_i^2(\ell) = \sigma_i(\ell,\ell)$, osv.

Ifølge Sverdrup (1964, side 328, formel (35)) blir

(6.3.4) $$\text{cov }\{\overline{X}_{ir}(\ell), \overline{X}_{ir}(\ell') \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \tau_{ir}(\underset{\sim}{j},\ell,\ell'),$$

noe som generaliserer (6.3.2). Ved å bruke samme teknikk som Sverdrup, nemlig ved å se at vi har

(6.3.5) $$\text{cov }\{\hat{a}(\ell), \hat{a}(\ell')\} = \tfrac{1}{2}\{\text{var}\left[\hat{a}(\ell) + \hat{a}(\ell')\right] - \text{var } \hat{a}(\ell) - \text{var } \hat{a}(\ell')\},$$

får vi tilsvarende at

(6.3.6) $$\text{cov }\{\hat{a}(\ell), \hat{a}(\ell')\} = \sum_i \frac{M_i^2}{m_i}\{\gamma_i(\ell,\ell') + (1-\frac{m_i}{M_i})\,\sigma_i(\ell,\ell')\}.$$

Denne formelen gjelder både for $\ell = \ell'$ og $\ell \neq \ell'$.

Siden $\hat{a} = \sum_{\ell'} \hat{a}(\ell')$, blir

(6.3.7) $$\text{cov }\{\hat{a}(\ell), \hat{a}\} = \sum_{\ell'} \text{cov }\{\hat{a}(\ell), \hat{a}(\ell')\}.$$

Hvis vi innfører

$$\sigma_i(j,\ell,\cdot) = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k,\ell) - \bar{a}_i(j,\cdot,\ell)\} \{a_i(j,k) - \bar{a}_i(j)\},$$

$$\sigma_i(\ell,\cdot) \quad = \frac{1}{M_i-1} \sum_j \{a_i(j,\cdot,\ell) - \bar{a}_i(\cdot,\ell)\} \{a_i(j) - \bar{a}_i\},$$

og tilsvarende $\tau_{ir}(\underset{\sim}{j},\ell,\cdot)$ og $\gamma_i(\ell,\cdot)$, får vi derfor

$$\text{cov}\ \{\hat{a}(\ell),\ \hat{a}\} = \sum_i \frac{M_i^2}{m_i} \{\gamma_i(\ell,\cdot) + (1 - \frac{m_i}{M_i})\ \sigma_i(\ell,\cdot)\}.$$

6.4. Prosentfordelingen. De prosenter som regnes ut, er

$$\hat{q}(\ell) = \hat{a}(\ell)/\hat{a}$$

(multiplisert med 100), og de betraktes som estimatorer for

$$q(\ell) = a(\ell)/a.$$

En tilnærmelse til var $\hat{q}(\ell)$ finnes ved et resonnement analogt til det en bruker ved brøkestimering (Sverdrup, 1964, side 328-329):

$$\text{(6.4.1)} \qquad \text{var}\ \hat{q}(\ell) \approx q^2(\ell)\ \{\frac{\text{var}\ \hat{a}(\ell)}{a^2(\ell)} + \frac{\text{var}\ \hat{a}}{a^2} - 2\ \frac{\text{cov}\ [\hat{a}(\ell),\ \hat{a}]}{a(\ell)\ a}\}\ .$$

Høyre side her estimeres ved

$$\text{(6.4.2)} \qquad \text{est var}\ \hat{q}(\ell) = \hat{q}^2(\ell)\ \{\frac{\text{est var}\ \hat{a}(\ell)}{\hat{a}^2(\ell)} + \frac{\text{est var}\ \hat{a}}{\hat{a}^2} - 2\ \frac{\sum_{\ell'} \text{est cov}\ [\hat{a}(\ell),\ \hat{a}(\ell')]}{\hat{a}(\ell)\ \hat{a}}\}\ .$$

Det gjenstår derfor bare å finne en estimator for cov $\{\hat{a}(\ell),\ \hat{a}(\ell')\}$. Den greieste måten å gjøre dette på er antakelig å estimere var $\hat{a}(\ell)$, var $\hat{a}(\ell')$, og var $[\hat{a}(\ell) + \hat{a}(\ell')]$ hver for seg med de forventningsrette estimatorene vi har utledet tidligere, og så bruke (6.3.5).

Fra et beregningsteknisk synspunkt lønner det seg neppe å trekke sammen de tre separate variansestimatorene. Vi skal gjøre det allikevel, for det kan være greit å ha resultatet for senere referanse.

La da

$$S_{ir}(\underset{\sim}{J},\ell,\ell') = \frac{1}{n_{ir}(\underset{\sim}{J})-1} \sum_s \{X_{irs}(\ell) - \bar{X}_{ir}(\ell)\} \{X_{irs}(\ell') - \bar{X}_{ir}(\ell')\},$$

og la

$$S^2_{ir}(\underset{\sim}{J},\ell,\ell,+) = \frac{1}{n_{ir}(\underset{\sim}{J})-1} \Sigma \{X_{irs}(\ell) + X_{irs}(\ell') - \overline{X}_{ir}(\ell) - \overline{X}_{ir}(\ell')\}^2.$$

La $T_{ir}(\underset{\sim}{J},\ell,\ell')$ og $T^2_{ir}(\underset{\sim}{J},\ell,\ell',+)$ være definert tilsvarende. La

$$G_i(\ell,\ell') = \frac{1}{m_i} \sum_r N^2_{ir}(J_{ir})\ T_{ir}(\underset{\sim}{J},\ell,\ell'),$$

$$G^2_i(\ell,\ell',+) = \frac{1}{m_i} \sum_r N^2_{ir}(J_{ir})\ T^2_{ir}(\underset{\sim}{J},\ell,\ell',+),$$

$$U_i(\ell,\ell') = \frac{1}{m_i-1} \sum_r \{V_{ir}(\ell) - \overline{V}_i(\ell)\}\{V_{ir}(\ell') - \overline{V}_i(\ell')\},$$

$$U^2_i(\ell,\ell',+) = \frac{1}{m_i-1} \sum_r (V_{ir}(\ell) + V_{ir}(\ell') - \overline{V}_i(\ell) - \overline{V}_i(\ell')^2,$$

$$S_i(\ell,\ell') = [U_i(\ell,\ell') - G_i(\ell,\ell')]^+,$$

$$S^2_i(\ell,\ell',+) = [U^2_i(\ell,\ell',+) - G^2_i(\ell,\ell',+)]^+,$$

$$W_i(\ell,\ell') = \frac{1}{m_i}\{G_i(\ell,\ell') + (1 - \frac{m_i}{M_i})\ S_i(\ell,\ell')\},$$

og

$$W^2_i(\ell,\ell',+) = \frac{1}{m_i}\{G^2_i(\ell,\ell',+) + (1- \frac{m_i}{M_i})\ S^2_i(\ell,\ell',+)\}.$$

Her er $x^+ = \max(x,0)$. Da blir

$$S^2_{ir}(\underset{\sim}{J},\ell,\ell',+) = S^2_{ir}(\underset{\sim}{J},\ell) + S^2_{ir}(\underset{\sim}{J},\ell') + 2\ S_i(\underset{\sim}{J},\ell,\ell'),$$

$$W^2_i(\ell,\ell',+) = W^2_i(\ell) + W^2_i(\ell') + 2\ W_i(\ell,\ell'),$$

osv. Altså blir

(6.4.3) $$\text{est cov}\ \{\hat{a}(\ell),\ \hat{a}(\ell') = \sum_i M^2_i\ W_i(\ell,\ell').$$

<u>6.5. Noen forenklinger</u>. I de ovenstående formler har vi ikke egentlig brukt at $a_i(j,k,\ell)$-verdiene bare er 0 og 1. Vi får et par forenklinger når vi nå benytter oss av at vi bare har binære variable.

Hvis c_1, c_2, ..., c_n er n tall som alle har verdien 0 eller 1, og hvis $\overline{c} = \Sigma\ c_i/n$, er som kjent

$$\sum_i (c_i - \overline{c})^2 = n\ \overline{c}(1-\overline{c}).$$

Derfor blir

$$\sigma_i^2(j,\ell) = \frac{N_i(j)}{N_i(j)-1}\, \bar{a}_i(j,\cdot,\ell)\, \{1 - \bar{a}_i(j,\cdot,\ell)\},$$

og

$$S_{ir}^2(\underset{\sim}{J},\ell) = \frac{n_{ir}(\underset{\sim}{J})}{n_{ir}(\underset{\sim}{J})-1}\, \bar{X}_{ir}(\ell)\, \{1-\bar{X}_{ir}(\ell)\}.$$

7. Brøkestimering og etterhåndsstratifisering.

7.1. Innledning. I kapitlene 3 til 5 studerte vi bl.a. situasjoner der det til trekkeenheten (i,j,k) knyttes én skalar verdi $a_i(j,k)$. I disse kapitlene utnyttet vi ikke opplysninger som man eventuelt måtte ha om de observerte trekkeenhetene ut over kjennskapet til denne skalarverdien og utvalgsplanen. Nå skal vi se litt på enkelte teknikker som kan benyttes når en til en vilkårlig trekkeenhet (i,j,k) ikke bare knytter en skalar, men en hel vektor

$$\{a_i(j,k,1),\ a_i(j,k,2),\ \ldots\ \}.$$

Vi skal gå ut fra at en for et gitt formål er spesielt interessert i verdiene av $a_i(j,k,1)$-ene, og at $a_i(j,k,\ell)$-ene for $\ell \geq 2$ bare er hjelpevariable. Vår estimand er således

$$a(1) = \sum_i \sum_j \sum_k a_i(j,k,1).$$

I virkeligheten stod vi overfor en slik situasjon alt i kapittel 6. Riktignok var de mulige verdiene av $a_i(j,k,\ell)$ der bare 0 og 1, og dessuten kunne summen

$$a_i(j,k) = \sum_\ell a_i(j,k,\ell)$$

også bare ha verdiene 0 og 1, men vi benyttet ikke dette i avsnittene 6.3 og 6.4. Først i avsnitt 6.5 utnyttet vi den binære karakteren til $a_i(j,k,\ell)$-ene, og da bare til å forenkle et par variansformler.

Dette betyr naturligvis at formlene i avsnittene 6.3 og 6.4 fortsatt gjelder selv om $a_i(j,k,\ell)$-ene ikke trenger være binære. Vi skal gjøre bruk av dette nedenfor.

7.2. Brøkestimering. Vår estimand er altså $a(1)$. La oss først anta at $a(\ell)$ er kjent for en gitt $\ell > 1$. Vi kan da danne følgende brøkestimator for $a(1)$:

$$(7.2.1) \qquad \hat{a}^{(\ell)}(1) = \hat{a}(1)\,\frac{a(\ell)}{\hat{a}(\ell)},$$

såframt $\hat{a}(\ell) > 0$. En tilnærmelse av vanlig type til var $\hat{a}^{(\ell)}$ (1) vil være

$$(7.2.2) \qquad \text{var}\ \hat{a}^{(\ell)}(1) \approx a^2(1)\ \{\frac{\text{var}\ \hat{a}(1)}{a^2(1)} + \frac{\text{var}\ \hat{a}(\ell)}{a^2(\ell)} - 2\frac{\text{cov}\ \{\hat{a}(1),\ \hat{a}(\ell)\}}{a(1)\ a(\ell)}\}.$$

En estimator for høyre side her er

$$\text{est var}\ \hat{a}(1) + \frac{\hat{a}^2(1)}{a^2(\ell)}\ \text{est var}\ \hat{a}(\ell) - 2\ \frac{\hat{a}(1)}{a(\ell)}\ \text{est cov}\ \{\hat{a}(1),\ \hat{a}(\ell)\},$$

der varians- og kovariansestimatorene er dem vi utledet i kapittel 6.

Hvis $a(\ell)$ er kjent for flere enn én $\ell \neq 1$, kan man danne et tilsvarende antall brøkestimatorer. Man kan også forsøke å kombinere disse og lage en sammensatt estimator.

7.3. Etterhåndsstratifisering når en kjenner antall trekkeenheter i alle etterhåndsstrata fra hvert forhåndsstratum og utvalgsområde. I enkelte tilfelle vet man hvordan bestanden av trekkeenheter fordeler seg etter kjennetegn som ikke er brukt i Byråets standard utvalgsopplegg, og som man kunne ønske å bruke i stratifiseringsøyemed. Eksempelvis ville det ofte vært nyttig om man kunne stratifisere intervjuede personer etter kjønn og alder, men Byråets standard utvalgsplan gjør ikke dette. En teknikk som da kan brukes, er å observere verdiene av de ekstra kjennetegnene (f.eks. kjønn og alder) hos trekkeenhetene i utvalget, og så stratifisere utvalget når dataene foreligger. Dette kalles etterhåndsstratifisering. Teknikken brukes i tilknytning til arbeidskraftsundersøkelsene på utvalgsbasis.

Vi skal nå gi en formell beskrivelse av denne framgangsmåten, og begrenser oss da til ett ekstra kjennetegn. (Flere kjennetegn kan alltid kombineres og "omkodes" så man formelt sett får ett kjennetegn å arbeide med.

Ved siden av den verdien $a_i(j,k,1)$ som vi "egentlig" er interessert i hos trekkeenhet nr. (i,j,k) tenker vi oss at den tilordnes en verdi $b_i(j,k)$, som bare kan anta et endelig antall verdier, la oss si $\{1,2,\ \ldots,\ v\}$. Vi transformerer $b_i(j,k)$ til en vektor med binære variable,

$$(7.3.1) \qquad \{a_i(j,k,2),\ a_i(j,k,3),\ \ldots,\ a_i(j,k,v+1)\},$$

der

$$a_i(j,k,\ell+1) = \begin{cases} 1 & \text{hvis } b_i(j,k) = \ell, \\ 0 & \text{ellers,} \end{cases}$$

for $\ell = 1,2, \ldots, v$. Dette kan også skrives slik:

$$\begin{cases} a_i \{j,k,b_i(j,k) + 1\} = 1, \\ a_i(j,k,\ell+1) = 0 \text{ for } \ell \neq b_i(j,k). \end{cases}$$

Naturligvis er

$$(7.3.2) \qquad \sum_{\ell=2}^{v+1} a_i(j,k,\ell) = 1.$$

Trekkeenhetene i utvalget etterhåndsstratifiseres etter verdien av $b.(j,k)$. Dette betyr at de enhetene som har 1-tallet plassert i samme posisjon i vektoren (7.3.1), samles i samme etterhåndsstratum. Antall enheter fra forhåndsstratum i og utvalgsområde j som tilhører etterhåndsstratum nr. ℓ ($\ell = 1,2, \ldots, v$) blir $a_i(j,\cdot, \ell+1)$. Hvis denne er kjent for alle (i,J_{ir},ℓ), kan etterhåndsstratifiseringen foregå innen hvert forhåndsstratum og utvalgsområde.

La oss først studere en estimator for $a(1)$ når dette er tilfelle.

Når vi skal bruke etterhåndsstrataene til å konstruere en estimator analog til den som benyttes ved stratumlotteriske utvalg (Sverdrup, 1964, s. 338), merker vi oss først at

$$X_{ir}(\ell+1) = \sum_s X_{irs}(\ell+1)$$

er antall enheter observert i etterhåndsstratum ℓ i utvalgsområde J_{ir} i forhåndsstratum i i utvalget, og at deres gjennomsnittlige verdi av $a_i(j,k,1)$ er

$$(7.3.3) \qquad \tilde{X}_{ir}(1,\ell+1) = \sum_s X_{irs}(1)\, X_{irs}(\ell+1)/X_{ir}(\ell+1).$$

(Vi tolker 0/0 som 0.) Estimatoren for $a(1)$ blir derfor

$$(7.3.4) \qquad \tilde{a}(1) = \sum_{\ell=1}^{v} \sum_i \frac{M_i}{m_i} \sum_r a_i(J_{ir},\cdot,\ell+1)\, \tilde{X}_{ir}(1,\ell+1).$$

La nå $\underset{\sim}{X} = \{X_{ir}(\ell+1) : i = 1,2 \ldots; r = 1,2, \ldots; \ell = 1,2, \ldots, v\}$ være en opplisting av $X_{ir}(\ell+1)$-verdiene. Hvis det er gitt at utvalget er blitt slik at $\underset{\sim}{J} = \underset{\sim}{j}$ og

$$\underset{\sim}{X} = \underset{\sim}{x} = \{x_{ir}(\ell+1) : i = 1,2, \ldots; r = 1,2, \ldots; \ell = 1,2, \ldots, v\},$$

for spesifiserte $x_{ir}(\ell+1)$, kan vi betrakte trekkeenhetene i etterhåndsstratum ℓ i utvalgsområde j_{ir} i forhåndsstratum i i utvalget, som et rent lotterisk utvalg på $x_{ir}(\ell+1)$ trekkeenheter fra de $a_i(j,\cdot,\ell+1)$ enhetene som finnes der. Hvis vi lar

$$\bar{a}_{i}^{(\ell)}(j,1) = \frac{\sum_{k} a_{i}(j,k,1)\ a_{i}(j,k,\ell+1)}{a_{i}(j,\cdot,\ell+1)},$$

blir derfor

(7.3.5) $$E\{\sum_{s} X_{irs}(1)\ X_{irs}(\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j},\ \underset{\sim}{X} = \underset{\sim}{x}\} = x_{ir}(\ell+1)\ \bar{a}_{i}^{(\ell)}(j_{ir},1),$$

og følgelig blir

$$E\{\tilde{X}_{ir}(1,\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j},\ \underset{\sim}{X} = \underset{\sim}{x}\} = \varepsilon\{x_{ir}(\ell+1)\}\ \bar{a}_{i}^{(\ell)}(j_{ir},1),$$

med

$$\varepsilon(x) = \begin{cases} 0 & \text{for } x=0, \\ 1 & \text{for } x\neq 0. \end{cases}$$

Men da er

(7.3.6) $$E\{\tilde{X}_{ir}(1,\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \phi_{ir}(\underset{\sim}{j},\ell+1)\ \bar{a}_{i}^{(\ell)}(j_{ir},1),$$

med

$$\phi_{ir}(\underset{\sim}{j},\ell+1) = P\{X_{ir}(\ell+1) > 0 \mid \underset{\sim}{J} = \underset{\sim}{j}\}.$$

Altså blir

(7.3.7) $$E\ \tilde{a}(1) = \frac{\sum_{\ell=1}^{v} \sum_{i} \frac{M_{i}}{m_{i}} \sum_{r} \sum_{\underset{\sim}{j}} a_{i}(j_{ir},\cdot,\ell+1)\ \bar{a}_{i}^{(\ell)}(j_{ir},1)\ \phi_{ir}(j,\ell+1)}{\prod_{\nu} \left[M_{\nu}(M_{\nu}-1)\ \ldots\ (M_{\nu}-m_{\nu}+1)\right]}.$$

Hvis vi hadde hatt $\phi_{ir}(\underset{\sim}{j},\ell+1) \equiv 1$, ville høyre side her redusert seg til

$$\sum_{\ell} \sum_{i} \frac{M_{i}}{m_{i}} \sum_{j} \frac{m_{i}}{M_{i}}\ a_{i}(j,\cdot,\ell+1)\ \bar{a}_{i}^{(\ell)}(j,1)$$

$$= \sum_{i} \sum_{j} \sum_{k} a_{i}(j,k,1) \sum_{\ell=1}^{v} a_{i}(j,k,\ell+1) = a(1).$$

I så fall ville altså $\tilde{a}(1)$ blitt forventningsrett. Idet $\phi_{ir}(\underset{\sim}{j},\ell+1) < 1$, <u>vil $\tilde{a}(1)$ underestimere $a(1)$ noe</u>.

Vi ser forøvrig at

$$\phi_{ir}(\underset{\sim}{j},\ell+1) = 1-P\{X_{ir}(\ell+1) = 0 \mid \underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= 1 - \frac{\binom{N_{i}(j_{ir}) - a_{i}(j_{ir},\cdot,\ell+1)}{n_{ir}(\underset{\sim}{j})}}{\binom{N_{i}(j_{ir})}{n_{ir}(\underset{\sim}{j})}}$$

$$\underset{\sim}{\sim} 1 - \{1 - \frac{a_i(j_{ir},\cdot,\ell+1)}{N_i(j_{ir})}\}^{n_{ir}(\underset{\sim}{j})} .$$

Hvis $n_{ir}(\underset{\sim}{j})$ er relativt stor, trenger derfor ikke $\phi_{ir}(j,\ell+1)$ være så mye mindre enn 1.

Så skulle vi gjerne funnet en eksakt formel for var $\tilde{a}(1)$. Imidlertid viser det seg at dette leder til uttrykk som er alt for kompliserte til at vi antakelig kan ha noen glede av dem. For estimering av var $\tilde{a}(1)$ må vi derfor gå fram på en helt annen måte enn det som har vært nødvendig i tidligere notater. Det finnes en del "empiriske" metoder for variansestimering, og det er betydningsfullt å få anvendt en slik teknikk når vi etterhåndsstratifiserer. (Se f.eks. Lansing og Morgan, 1971, s. 87-90, og deres referanser. Se også Kish, 1957.) Alternativt kan vi forsøke å utlede tilnærmelsesformler analoge til den i avsnitt 7.2 ovenfor. (Sml. Dahmström 1968, side 19.)

Vi skal imidlertid ikke ta opp dette her. La oss bare avslutte dette underkapitlet med å gjengi noen av de variansformlene som vi kan få tak i, så langt at vi overbeviser oss om at formelen for var $\tilde{a}(1)$ må bli komplisert.

Hvis vi lar

$$[\sigma_i^{(\ell)}(j,1)]^2 = \frac{\sum_k \{a_i(j,k,1)\ a_i(j,k,\ell+1) - \bar{a}_i^{(\ell)}(j,1)\}^2}{a_i(j,\cdot,\ell+1) - 1},$$

blir

$$\text{var}\ \{\sum_s X_{irs}(1)\ X_{irs}(\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j},\ \underset{\sim}{X} = \underset{\sim}{x}\} =$$

$$x_{ir}(\ell+1)\ \frac{a_i(j_{ir},\cdot,\ell+1) - x_{ir}(\ell+1)}{a_i(j_{ir},\cdot,\ell+1)}\ [\sigma_i^{(\ell)}(j_{ir},1)]^2,$$

og følgelig

$$\text{var}\ \{\tilde{X}_{ir}(1,\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j},\ \underset{\sim}{X} = \underset{\sim}{x}\}$$

$$= \frac{\varepsilon\{x_{ir}(\ell+1)\}}{x_{ir}(\ell+1)} \cdot \frac{a_i(j_{ir},\cdot,\ell+1) - x_{ir}(\ell+1)}{a_i(j_{ir},\cdot,\ell+1)}\ [\sigma_i^{(\ell)}(j_{ir},1)]^2.$$

Altså blir

$$\text{var } \{\tilde{X}_{ir}(1,\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j}\} =$$

$$[\bar{a}_i^{(\ell)}(j_{ir},1)]^2 \; \pi_{ir}(\underset{\sim}{j},\ell+1) \; [1-\pi_{ir}(\underset{\sim}{j},\ell+1)]$$

$$+ \frac{[\sigma_i^{(\ell)}(j_{ir},1)]^2}{a_i(j_{ir},\cdot,\ell+1)} \; E \; \{\frac{\varepsilon[X_{ir}(\ell+1)]}{X_{ir}(\ell+1)} \; [a_i(j_{ir},\cdot,\ell+1)$$

$$- X_{ir}(\ell+1)] \mid \underset{\sim}{J} = \underset{\sim}{j}\}.$$

Sammen med (7.3.6) setter dette oss i stand til å finne

$$\text{var } \{a_i(J_{ir},\cdot,\ell+1) \; \tilde{X}_{ir}(1,\ell+1)\}$$

i prinsippet, men resultatet blir opplagt meget lite pent. I tillegg til dette trenger vi så

$$\text{cov } \{a_i(J_{ir},\cdot,\ell+1) \; \tilde{X}_{ir}(1,\ell+1), \; a_i(J_{ir},\cdot,\ell'+1) \; \tilde{X}_{ir}(1,\ell'+1)\}$$

for $\ell \neq \ell'$. Det er neppe bryet verd å forsøke å regne.

7.4. Etterhåndsstratifisering når en kjenner samlet antall trekkeenheter i hvert etterhåndsstratum, men ikke deres fordeling på forhåndsstrata og utvalgsområder. Det samlede antall intervjuenheter i etterhåndsstratum ℓ i totalbestanden blir $a(\ell+1)$. Hvis $a_i(j,\cdot,\ell+1)$-ene ikke er kjent, mens $a(2)$, ..., $a(v+1)$ er det, kan en inndele utvalget i v etterhåndsstrata på tvers av inndelingen i forhåndsstrata og utvalgsområder. Dette gjøres i arbeidskraftsundersøkelsene, siden en kjenner fordelingen av den samlede befolkning på kjønn og alder med tilstrekkelig nøyaktighet, men ikke deres geografiske fordeling på et tidspunkt som er aktuelt nok. Vi skal nå se hvordan slik etterhåndsstratifisering utnyttes i estimeringen av $a(1)$. Vi nøyer oss da med å etablere estimatoren og studere dens forventning. Spørsmålet om dens varians lar vi ligge.

(Det er beklagelig at vi her kommer så kort når det gjelder variansproblematikken. Ett viktig motiv for etterhåndsstratifisering er å oppnå variansreduksjon. Når vi nå ikke etablerer en metode til estimering av estimatorvariansen, er vi imidlertid foreløpig avskåret fra å vurdere effekten i denne retningen.)

Vi merker oss at

$$X(\ell+1) = \sum_i \sum_r \sum_s X_{irs}(\ell+1)$$

er antall enheter observert i etterhåndsstratum nr. ℓ i utvalget, og at deres gjennomsnittlige verdi av $a_i(j,k,1)$ er

$$\tilde{X}(1,\ell+1) = \sum_i \sum_r \sum_s X_{irs}(1)\, X_{irs}(\ell+1)/X(\ell+1). \tag{7.4.1}$$

Estimatoren blir da

$$a^{*}(1) = \sum_{\ell=1}^{v} a(\ell+1)\, \tilde{X}(1,\ell+1). \tag{7.4.2}$$

I forrige underkapittel klarte vi å finne et uttrykk for forventningsverdien til $\tilde{a}(1)$ fordi det var mulig å finne forventningen til $\hat{a}_{ir}(1,\ell+1)$ når $\underset{\sim}{J}$ og $X_{ir}(\ell+1)$ var gitt. I den nåværende situasjon er jeg ikke i stand til å gjøre noe tilsvarende. Spesielt ser det ikke ut til at vi kommer noen vei ved å betinge med hensyn på $X(\ell+1)$ (eller $X(2)$, ..., $X(v+1)$ og $\underset{\sim}{J}$).

La oss derfor nøye oss med å diskutere $\tilde{X}(1,\ell+1)$ som en estimator for

$$E\,\{\sum_i \sum_r \sum_s X_{irs}(1)\, X_{irs}(\ell+1)\} \;/\; E\, X(\ell+1). \tag{7.4.3}$$

Denne estimatoren vil da sannsynligvis være noe forventningsskjev, men la nå dét ligge.

Vi får av (7.3.5) at

$$E\,\{\sum_i \sum_r \sum_s X_{irs}(1)\, X_{irs}(\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j},\ \underset{\sim}{X} = \underset{\sim}{x}\}$$

$$= \sum_i \sum_r X_{ir}(\ell+1)\, \bar{a}_i^{(\ell)}(j_{ir}, 1).$$

Når $\underset{\sim}{J} = \underset{\sim}{j}$, er $X_{ir}(\ell+1)$ multinomisk fordelt, og

$$E\{X_{ir}(\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = n_{ir}(\underset{\sim}{j})\, \frac{a_i(j_{ir}, \cdot, \ell+1)}{N_i(j_{ir})}. \tag{7.4.4}$$

Altså blir

$$E\{\sum_s X_{irs}(1)\, X_{irs}(\ell+1) \mid \underset{\sim}{J} = \underset{\sim}{j}\}$$

$$= \frac{n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})} \sum_k a_i(j_{ir},k,1)\, a_i(j_{ir},k,\ell+1).$$

La oss nå som vanlig anvende at

$$\frac{n_{ir}(\underset{\sim}{j})}{N_i(j_{ir})} = b(j)\, \frac{M_i}{m_i}.$$

Da blir

$$E\{\sum_s X_{irs}(1)\ X_{irs}(\ell+1)\ |\ J_{ir} = j_{ir}\}$$

$$= \frac{M_i}{m_i} \sum_k a_i(j_{ir},k,1)\ a_i(j_{ir},k,\ell+1)\ \beta(j_{ir}),$$

der

(7.4.5) $$\beta_i(j_{ir}) = E\{b(\underset{\sim}{J})\ |\ J_{ir} = j_{ir}\}.$$

Hvis vi nå kaller teller og nevner i (7.4.3) for henholdsvis $A(\ell)$ og $B(\ell)$, blir derfor

$$A(\ell) = \sum_i \sum_j \beta_i(j) \sum_k a_i(j,k,1)\ a_i(j,k,\ell+1).$$

Vi får $B(\ell)$ ved å erstatte $X_{irs}(1)$ og $a_i(j,k,1)$ med 1 overalt i resonnementet. Dette gir

$$B(\ell) = \sum_i \sum_j \beta_i(j)\ a_i(j,\cdot,\ell+1).$$

Vi kan derfor oppfatte $a^{*}(1)$ som en estimator for

$$\alpha = \sum_{\ell=1}^{v} a(\ell+1)\ A(\ell)\ /\ B(\ell).$$

Hvis alle $\beta_i(j)$ hadde vært like, la oss si lik β, ville $A(\ell)$ blitt lik

$$\beta \sum_i \sum_j \sum_k a_i(j,k,1)\ a_i(j,k,\ell+1),$$

og $B(\ell)$ ville blitt lik $\beta\ a(\ell+1)$. Altså ville α blitt lik $a(1)$, slik vi gjerne skulle se var tilfelle. Ved Byråets standard utvalgsplan kommer imidlertid variasjonene i $\beta_i(j)$ inn som et forstyrrende element. Hvis disse variasjonene ikke er store, slik at $\beta_i(j)$-ene stort sett ligger på samme nivå, vil naturligvis α tilsvarende være tilnærmet lik $a(1)$.

(Se ytterligere kommentarer i underkapittel 12.5.)

<u>7.5. Problemet med tomme etterhåndsstrata</u>. I de ovenstående to avsnittene (7.3 og 7.4) har vi tolket 0/0 som null. (Se f.eks. rett etter (7.3.3).) I avsnitt 7.3 kan en slik tolkning bli nødvendig fordi vi kan ha $X_{ir}(\ell+1) = 0$, dvs. det kan være at utvalget ikke kommer til å inneholde intervjuenheter i ℓ-te etterhåndsstratum i utvalgsområde j_{ir} i forhåndsstratum nr. i. Konsekvensen av denne tolkningen er at et slikt etterhåndsstratum ikke bidrar med noe i beregningen av $\tilde{a}(1)$, og vi kompenserer ikke dette på noen måte, f.eks. ved å gi observasjoner fra øvrige etterhåndsstrata øket vekt. (Dette gav også opphav til forventningsskjevheten til

$\tilde{a}(1)$ og medfører underestimering av $a(1)$.)

En annen måte å gå fram på, er å slå det tomme etterhåndsstratumet sammen med et annet etterhåndsstratum der en har observasjoner. (Sml. kapittel 11.) En velger da gjerne å slå sammen strata som ligger nær hverandre i en eller annen forstand, enten geografisk, næringsøkonomisk, eller på annen måte. Dette betyr i realiteten at en i en viss utstrekning lar dataene bestemme hvilken etterhåndsstratifisering en skal bruke, mens en jo egentlig tok sikte på å fastlegge etterhåndsstratifiseringen <u>før</u> en fikk se observasjonene.

La oss se på hva det vil bety for $\tilde{a}(1)$ om en slo sammen etterhåndsstrataene ℓ og ℓ' i utvalgsområde j_{ir} i forhåndsstratum i, når etterhåndsstratum ℓ' er tomt mens stratum ℓ ikke er det. Uten sammenslåing vil de to strataene til sammen bidra

$$a_i(J_{ir},\cdot,\ell+1)\ \tilde{X}_{ir}(1,\ell+1)$$

til den innerste summen i (7.3.4). Med sammenslåing blir bidraget i stedet

$$\{a_i(J_{ir},\cdot,\ell+1) + a_i(J_{ir},\cdot,\ell'+1)\}\ \tilde{X}_{ir}(1,\ell+1). \tag{7.5.1}$$

Det kompenseres altså for det tomme etterhåndsstratumet. Vi ser at sammenslåingen i realiteten innebærer at vi bruker $\tilde{X}_{ir}(1,\ell+1)$ som anslag for gjennomsnittlig $a_i(j,k,1)$-verdi også i det tomme etterhåndsstratumet.

En annen mulighet, som bygger på en idé av Fuller (1966), er å erstatte krøllparantesen i (7.5.1) med en annen vekt, slik at det totale bidraget fra de to etterhåndsstrataene blir en forventningsrett estimator for det tilsvarende bidraget til $a(1)$ i grunnbestanden når $\underset{\sim}{J}$ og $\underset{\sim}{X}$ er gitt. (Størrelsen i (7.5.1) har ikke denne egenskapen.)

Uansett hva man gjør, opererer man jo med en statistisk estimeringsprosedyre, og det er ønskelig å få studert dens statistiske egenskaper. Det viser seg imidlertid at det er svært vanskelig å gjennomføre dette med vanlige, analytiske hjelpemidler, særlig når utvalgsplanen er så komplisert som den Byrået bruker. Fuller (1966) har studert tilfellet med rent lotterisk trekking av utvalget, slik at det altså ikke er tale om forhåndsstratifisering og totrinnstrekning. Han ser i detalj på hva som skjer når en har bare to etterhåndsstrata, og han får raskt ganske stygg analyse. Han antyder også hvordan dette kan generaliseres til flere etterhåndsstrata, men synes vesentlig å oppnå å demonstrere hvor komplisert det hele da blir.

Tilsvarende betraktninger gjelder for avsnitt 7.4.

7.6. Avslutningsmerknader.

A. De fleste lærebøker i statistisk teori for utvalgsundersøkelser (f.eks. Hansen, Hurwitz og Madow; Sukhatme; Kish; Raj) har noe stoff om teorien for etterhåndsstratifisering, og det finnes noen tidsskriftsartikler. Den samlede stoffmengde jeg har klart å finne, er imidlertid ikke stor. Behandlingen teorien gir, er svært knapp og i alt vesentlig begrenset til enkle utvalgsplaner, oftest til rent lotteriske utvalg. Tomme etterhåndsstrata behandles gjerne svært lettvint.

Dahmström (1968) har gitt en oversikt over det meste av det stoffet som var kommet fram til 1967. (Han synes å ha oversett Fuller, 1966.) Jeg vet ikke om at det er kommet betydningsfulle bidrag til teorien siden den tid. Derimot er det kommet en del interessante beskrivelser av anvendelser, f.eks. Dahmström og Wahlström (1969) og Sverige (1970/1971/1972).

B. Klassikerne Hansen, Hurwitz og Madow (1953) har en bemerkning som gjør det viktig å studere effekten av etterhåndsstratifiseringen ganske nøye. De framhever nemlig (bind 1, side 232 og bind 2, side 139) at etterhåndsstratifisering (av et rent lotterisk utvalg) kan resultere i øking av samplingvariansen hvis en ikke har stort nok forhold mellom utvalgsstørrelse og antall etterhåndsstrata. Et stort antall etterhåndsstrata er altså ikke i og for seg noe gode hvis ikke utvalget er stort.

Vi skal sitere hva de tre forfatterne sier. La da n være utvalgsstørrelsen, L antall etterhåndsstrata, og $\bar{n} = n/L$. Sitat:

"... if $\bar{n}$ is large, say 25-50 or more, one might achieve nearly all the gains of initial stratification, but if the strata are made too small so that $\bar{n}$ is small, one might lose considerably by the introduction of stratification after sampling."

C. Det er antakelig en vanlig oppfatning blant dem som har hørt om metoden at den virker til å eliminere estimeringsskjevhet. Våre studier ovenfor viser at dette ikke er riktig. Tvert imot kan etterhåndsstratifiseringen bidra til å innføre forventningsskjevhet. Hensikten med å bruke metoden er vel hovedsakelig ønsket om variansreduksjon, og så er håpet det at forventningsskjevheten er ubetydelig i forhold.

Dahmström (1969, s. 25) nevner et annet moment av samme type: Hvis frafallet er systematisk større i noen etterhåndsstrata enn i andre, vil naturligvis dét i seg selv bidra til forventningsskjevhet. Det kan ofte være grunn til å regne med selektivt frafall på denne måten.

Vi har ikke behandlet frafall som problem i denne artikkelen, og skal fortsatt la temaet ligge.

8. Trekkeenhet og analyseenhet.

8.1. Terminologien. I de foregående kapitlene har vi operert med trekkeenheter, og vi har antatt at et utvalg av slike enheter har dannet grunnlaget for analysene. Disse trekkeenhetene har da kunnet være individer, husholdninger, bedrifter, eller noe annet.

Det er imidlertid mulig at andre, mindre enheter legges til grunn for analysen. I så fall vil vi kalle disse for analyseenheter. Selv om en f.eks. har husholdninger som trekkeenheter, hender det ofte at en vil bruke de enkelte husholdningsmedlemmer som analyseenheter. Dette er ikke uproblematisk fra et statistisk synspunkt. Selv om trekkeenhetene framkommer ved et stratumlotterisk to-trinnsutvalg, slik som i Byråets standard utvalgsplan, vil ikke analyseenhetene gjøre det. Situasjonen der hver trekkeenhet kan omfatte flere enn én eneste analyseenhet, krever derfor særskilt behandling.

Vi skal nå antyde hvordan denne situasjonen kan behandles. Det viser seg at det først og fremst blir et spørsmål om å gi en passende tolkning av den teorien vi tidligere har beskrevet elementer av.

8.2. Endringer i formelverket. I teorien foran er trekkeenhet nr. (i,j,k) (dvs. trekkeenhet nr. k i utvalgsområde nr. j i stratum nr. i) tilordnet verdien $a_i(j,k)$. En ønsker å estimere "samlet a-verdi"

$$a = \sum_i \sum_j \sum_k a_i(j,k).$$

La nå antall analyseenheter i trekkeenhet (i,j,k) være $v_i(j,k)$. Vi skal se bort fra det trivielle tilfelle da alle $v_i(j,k)$ er lik 1, for da faller jo trekke- og analyseenhet sammen. Vi antar derfor at ikke alle $v_i(j,k)$ er like. La analyseenhet nr. (i,j,k,ℓ) (dvs. analyseenhet nr. ℓ i trekkeenhet nr. (i,j,k)) ha a-verdien $a_i(j,k,\ell)$, og la

$$a_i(j,k) = \sum_{\ell=1}^{v_i(j,k)} a_i(j,k,\ell).$$

Når vi skal estimere a, trenger vi da ikke ta hensyn til at trekkeenhetene er oppdelt i analyseenheter. Vi kan som før la

$$\hat{a} = \sum_i \frac{M_i}{m_i} \sum_r \frac{N_i(J_{ir})}{n_{ir}(J)} \sum_s X_{irs},$$

der X_{irs}-ene fortsatt er de observerte a-verdiene for trekkeenhetene. Således er

$$X_{irs} = \sum_{\ell=1}^{v_i(J_{ir},K_{irs})} X_{irs\ell},$$

med

$$X_{irs\ell} = a_i(J_{ir},K_{irs},\ell).$$

Her er $N_i(J_{ir})$ og $n_{ir}(\underset{\sim}{J})$ antall av <u>trekkeenheter</u>, <u>ikke</u> analyseenheter. Om vi lar

$$N'_i(j) = \sum_{k=1}^{N_i(j)} v_i(j,k)$$

og

$$n'_{ir}(\underset{\sim}{J}) = \sum_{s=1}^{n_{ir}(\underset{\sim}{J})} v_i(J_{ir},K_{irs}),$$

blir $N'_i(j)$ og $n'_{ir}(\underset{\sim}{J})$ det antall <u>analyseenheter</u> som de $N_i(j)$ og $n_{ir}(\underset{\sim}{J})$ trekkeenhetene inneholder til sammen. Vi ser at $N'_i(j)$ ikke blir stokastisk, mens $n'_{ir}(\underset{\sim}{j})$ blir en stokastisk variabel selv når det er gitt at $\underset{\sim}{J}=\underset{\sim}{j}$. Dette er intuitivt opplagt. Selv om utvalgsområdene er gitt, vil trekkeenhetene velges ut rent tilfeldig, så en kan ikke på forhånd være sikker på hvor mange analyseenheter en vil få.

Med utgangspunkt i analyseenhetene kunne en tenkes å bruke estimatoren

$$\hat{\hat{a}} = \sum_i \frac{M_i}{m_i} \sum_r \frac{N'_i(J_{ir})}{n'_{ir}(\underset{\sim}{J})} \sum_s \sum_\ell X_{irs\ell}.$$

Denne vil naturligvis ikke falle sammen med $\hat{a}$, og den vil ha andre statistiske egenskaper. Vi skal se litt på disse nedenfor.

<u>8.3. Bruk av antall analyseenheter</u>. La oss først merke oss at $v_i(j,k)$ er et kjennetegn tilknyttet trekkeenhet nr. (i,j,k) på samme måte som $a_i(j,k)$ er det. Den generelle teorien i de tidligere kapitlene gjelder altså like mye for v-verdiene som for a-verdiene.

Vi får derfor f.eks. at

$$E\{v_i(J_{ir},K_{irs}) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = \bar{v}_i(j_{ir}) = \frac{1}{N_i(j_{ir})} \sum_{k=1}^{N_i(j_{ir})} v_i(j,k)$$

$$= \frac{N'_i(j_{ir})}{N_i(j_{ir})}.$$

Altså blir

$$(8.3.1) \qquad E\,\{n'_{ir}(\underset{\sim}{J}) \mid \underset{\sim}{J} = \underset{\sim}{j}\} = n_{ir}(\underset{\sim}{j})\,\frac{N'_i(j_{ir})}{N_i(j_{ir})}\,.$$

Vi kan derfor betrakte brøken $N'_i(J_{ir})\ /\ n'_{ir}(\underset{\sim}{J})$ i formelen for $\hat{\hat{a}}$ som en estimator for brøken $N_i(J_{ir})\ /\ n_{ir}(\underset{\sim}{J})$ i formelen for $\hat{a}$. Som estimator vil den første brøken systematisk overestimere den siste. Vi skal nemlig vise at

$$(8.3.2) \qquad E\,\{\frac{N'_i(J_{ir})}{n'_{ir}(\underset{\sim}{J})} \mid \underset{\sim}{J} = \underset{\sim}{j}\} > \frac{N_i(j_{ir})}{n_{ir}(\underset{\sim}{j})}\,.$$

<u>Bevis</u>: Dette følger direkte av (8.3.1) og av Jensens ulikhet, som, anvendt på den aktuelle situasjon, sier at

$$E\,\{\frac{1}{n'_{ir}(\underset{\sim}{J})} \mid \underset{\sim}{J} = \underset{\sim}{j}\} > \frac{1}{E\,\{n'_{ir}(\underset{\sim}{J}) \mid \underset{\sim}{J} = \underset{\sim}{j}\}}\,. \quad \square$$

Nå må en ikke av (8.3.2) feilaktig slutte direkte at $\hat{\hat{a}}$ systematisk overestimerer a. Dét ville $\hat{\hat{a}}$ gjøre hvis

$$(8.3.3) \qquad E\,\{\frac{N'_i(J_{ir})}{n'_{ir}(\underset{\sim}{J})}\sum_s X_{irs} \mid \underset{\sim}{J} = \underset{\sim}{j}\}$$

var lik

$$(8.3.4) \qquad E\,\{\frac{N'_i(J_{ir})}{n'_{ir}(\underset{\sim}{J})} \mid \underset{\sim}{J} = \underset{\sim}{j}\} \cdot E\{\sum_s X_{irs} \mid \underset{\sim}{J} = \underset{\sim}{j}\}.$$

Imidlertid vet vi ikke om uttrykket i (8.3.3) faktisk er lik det i (8.3.4), for $n'_{ir}(\underset{\sim}{J})$ og $\sum_s X_{irs}$ er stokastisk avhengige, også når $\underset{\sim}{J} = \underset{\sim}{j}$ er gitt. Vi kan derfor ikke umiddelbart slutte noe om estimeringsskjevheten til $\hat{\hat{a}}$ av (8.3.2).

<u>8.4. a-verdi pr. analyseenhet.</u> Hvis

$$v = \sum_i \sum_j \sum_k v_i(j,k)$$

er kjent, vil a-verdien pr. analyseenhet, altså a/v, enkelt kunne estimeres ved $\hat{a}/v$.

Hvis derimot v <u>ikke</u> er kjent, vil det falle naturlig å estimere a/v med $\hat{a}/\hat{v}$, der

$$\hat{v} = \sum_i \frac{M_i}{m_i} \sum_r \frac{N_i(J_{ir})}{n_{ir}(\underset{\sim}{J})} \sum_s v_i(J_{ir}, K_{irs}).$$

Variansen til $\hat{a}/\hat{v}$ estimeres slik vi pleier å estimere variansen til en brøk.

8.5. Et praktisk problem. Intervjukontorets utvalgsregister inneholder ikke egentlig husholdning som enhet. Derimot har man en enhet vi kan kalle "adresse". De klynger som trekkes som trekkeenheter ved en husholdningsundersøkelse, avviker derfor en del fra de vanlige husholdningsbegrepene.

En vil vanligvis regne med at det er minst én analyseenhet i hver trekkeenhet, slik at alle $v_i(j,k) \geq 1$. Når trekkeenheten er en adresse, kan det imidlertid hende at en får $v_i(j,k) = 0$, fordi enkelte boliger står tomme, eller fordi boligen er revet.

Dette har imidlertid ingen betydning for den statistiske teorien her.

9. Kommunene som utvalgsområder.

9.1. Innledning. I noen undersøkelser gir man slipp på de straturninndelte utvalgsområdene i intervjukontorets utvalgsplan og bruker de kommunene som har utvalgsområder, til primærområder i undersøkelsene. Dette har man gjort i forbindelse med en boligfinansieringsundersøkelse (Tamsfoss, 1969), og ved prøveundersøkelsen om flyttemotiver våren 1972 (Dagsvik, 1972). Formelverket blir da noe annerledes enn i de foregående kapitlene. Vi skal her vise hvordan formlene blir i den nye situasjonen.

9.2. Symboler og konvensjoner. Vi tenker oss kommunene i landet nummerert, og lar $N(j)$ være antall intervjuenheter i kommune nr. j. Den k-te av disse har en a-verdi på $a(j, k)$. Vi lar

$$\bar{a}(j) = \sum_k a(j,k)/N(j),$$

$$a = \sum_j \sum_k a(j,k) = \sum_j N(j)\, \bar{a}(j),$$

og

$$\bar{a} = a/N,$$

der $N = \sum_j N(j)$. Vi betrakter a som vår estimand.

La kommunene nr. L_1, L_2, ... ha utvalgsområder. Mens nummereringen av de uttrukne utvalgsområdene i et stratum har en egen mening i de tidligere kapitlene, har ikke nummereringen av kommunene det. Hvis $\underset{\sim}{L} = (L_1, L_2, \ldots)$, er altså $\underset{\sim}{L}$ en ren opplisting av numrene på de uttrukne kommunene uten at nummerrekkefølgen betyr noe.

Vi innfører nå

$$I_j = \begin{cases} 1 \text{ hvis kommune nr. } j \text{ har minst ett utvalgsområde,} \\ 0 \text{ ellers.} \end{cases}$$

Da blir altså $I_j = 1$ hvis og bare hvis én av L_r-ene er lik j. La $n_j(\underset{\sim}{L})$ være det antall trekkeenheter som trekkes fra kommune nr. j. Hvis $I_j = 0$, blir også $n_j(\underset{\sim}{L}) = 0$, naturligvis. Hvis $L_r = j$, lar vi

$$X_{js} = a(j,K_{js}),$$

$$\bar{X}_j = \sum_s X_{js}/n_j(\underset{\sim}{L}),$$

der $K_{j1}, K_{j2}, \ldots$ er numrene på de trekkeenhetene som trekkes fra kommunen. $\bar{X}_j$ er nå definert bare for de j der $I_j = 1$. La Z_j være en vilkårlig stokastisk (eller ikke-stokastisk) variabel som er definert for slike j. Da skal vi la

$$I_j Z_j = \begin{cases} Z_j \text{ hvis } I_j = 1, \\ 0 \text{ hvis } I_j = 0, \end{cases}$$

slik at $I_j Z_j$ er definert også for $I_j = 0$ selv om Z_j ikke nødvendigvis er det.

La

$$\sigma^2(j) = \frac{1}{N(j) - 1} \sum_k \{a(j,k) - \bar{a}(j)\}^2,$$

slik at $\sigma^2(j)$ representerer variasjonen i a-verdier mellom trekkeenhetene i kommune nr. j. La videre

$$\tau^2_j(\underset{\sim}{L}) = \frac{\sigma^2(j)}{n_j(\underset{\sim}{L})} \cdot \frac{N(j) - n_j(\underset{\sim}{L})}{N(j)},$$

$$S^2_j(\underset{\sim}{L}) = \frac{I_j}{n_j(\underset{\sim}{L}) - 1} \sum_s \{X_{js} - \bar{X}_j\}^2,$$

og

$$T^2_j(\underset{\sim}{L}) = \frac{S^2_j(\underset{\sim}{L})}{n_j(\underset{\sim}{L})} \frac{N(j) - n_j(\underset{\sim}{L})}{N(j)}.$$

Endelig lar vi

$$f(j) = P\{I_j = 1\}, \ f(i,j) = P\{I_i = I_j = 1\},$$

$$g(i,j) = f(i,j) - f(i)\,f(j),$$

$$c(j) = N(j)/f(j),$$

$$\eta(j) = E\{\tau^2_j(\underset{\sim}{L}) \,/\, I_j = 1\},$$

$$\gamma^2 = \sum_j c^2(j)\,f(j)\,\eta(j),$$

og

$$\omega^2 = \sum_i \sum_j c(i)\,c(j)\,\bar{a}(i)\,\bar{a}(j)\,g(i,j).$$

Da blir altså $f(j)$ sannsynligheten for at kommune nr. j inneholder ett eller flere av intervjukontorets utvalgsområder, og $f(i,j)$ blir sannsynligheten for at både kommune nr. i og j gjør det. Vi skal vise hvordan de kan beregnes i underkapittel 9.7 nedenfor.

Vi merker oss at $f(j,j) = f(j)$, og at vi følgelig har

$$g(j,j) = f(j)\left[1 - f(j)\right].$$

9.3. Noen mellomresultater. Vi får

(9.3.1) $E\{\bar{X}_j \mid I_j = 1\} = \bar{a}(j),$

(9.3.2) $\text{var}\,\{\bar{X}_j \mid I_j = 1, \underset{\sim}{L}\} = \tau^2_j(\underset{\sim}{L}),$

(9.3.3) $E\{S^2_j(\underset{\sim}{L}) \mid I_j = 1, \underset{\sim}{L}\} = \sigma^2(j),$

(9.3.4) $E\{T^2_j(\underset{\sim}{L}) \mid I_j = 1, \underset{\sim}{L}\} = \tau^2_j(\underset{\sim}{L}),$

(9.3.5) $E\{\bar{X}_i\,\bar{X}_j \mid I_i = I_j = 1\} = \bar{a}(i)\,\bar{a}(j)$ for $i \neq j,$

og

(9.3.6) $\text{cov}\,\{\bar{X}_i, \bar{X}_j \mid I_i = I_j = 1\} = 0$ for $i \neq j.$

Altså blir

(9.3.7) $E\{I_j\,\bar{X}_j \mid I_j\} = I_j\,\bar{a}(j),$

(9.3.8) $\text{var}\,\{\bar{X}_j \mid I_j = 1\} = E\{\tau^2_j(\underset{\sim}{L}) \mid I_j = 1\} = \eta(j),$

(9.3.9) $\text{var}\,\{I_j\,\bar{X}_j \mid I_j\} = I_j\,\eta(j),$

og

(9.3.10) $E\{I_i\,\bar{X}_i\,I_j\,\bar{X}_j \mid I_i, I_j\} = I_i\,I_j\,\bar{a}(i)\,\bar{a}(j).$

Følgelig blir

(9.3.11) $E\{I_j \bar{X}_j\} = f(j)\ \bar{a}(j),$

$\text{var}\ \{I_j \bar{X}_j\} = f(j)\ \eta(j) + \bar{a}^2(j)\ f(j)\ \{1 - f(j)\}$

(9.3.12) $= f(j)\ \eta(j) + \bar{a}^2(j)\ g(j,j),$

$\text{cov}\ \{I_i \bar{X}_i,\ I_j \bar{X}_j\} = E\ \text{cov}\ \{I_i \bar{X}_i,\ I_j \bar{X}_j \mid I_i,\ I_j\}$

$+ \text{cov}\ \{E(I_i \bar{X}_i \mid I_i,\ I_j),\ E(I_j \bar{X}_j \mid I_i,\ I_j)\}$

$= E\{0\} + \text{cov}\ \{I_i\ \bar{a}(i),\ I_j\ \bar{a}(j)\}$

$= \bar{a}(i)\ \bar{a}(j)\ \{f(i,j) - f(i)\ f(j)\}$

(9.3.13) $= \bar{a}(i)\ \bar{a}(j)\ g(i,j)$ for $i \neq j,$

og

(9.3.14) $E\ \{I_i \bar{X}_i\ I_j \bar{X}_j\} = \bar{a}(i)\ \bar{a}(j)\ f(i,j)$ for $i \neq j.$

9.4. Estimatoren for a. Med bakgrunn i det ovenstående lar vi

(9.4.1) $\hat{a} = \sum_j c(j)\ I_j \bar{X}_j = \sum_{\{j:I_j=1\}} N(j)\ \bar{X}_j/f(j).$

Da er $\hat{a}$ forventningsrett, og

$$\sigma^2 = \text{var}\ \hat{a} = \gamma^2 + \omega^2.$$

Forventningsrettheten følger direkte av (9.3.11). Variansformelen følger slik av (9.3.12) og (9.3.13):

$$\text{var}\ \hat{a} = \sum_j c^2(j)\ \text{var}\ \{I_j \bar{X}_j\} + 2 \sum_{i<j} c(i)\ c(j)\ \text{cov}\ \{I_i \bar{X}_i,\ I_j \bar{X}_j\}$$

$$= \sum_j c^2(j)\ f(j)\ \eta(j) + \sum_j c^2(j)\ \bar{a}^2(j)\ g(j,j)$$

$$+ 2 \sum_{i<j} c(i)\ c(j)\ \bar{a}(i)\ \bar{a}(j)\ g(i,j) = \gamma^2 + \omega^2.$$

9.5. Estimering av variansen. Følgende estimatorer er forventningsrette:

(9.5.1) $\hat{\sigma}^2 = \sum_j c^2(j)\ f(j)\ T^2_j\ (\underset{\sim}{L})$

$+ \sum_i \sum_j c(i)\ c(j)\ g(i,j)\ I_i \bar{X}_i\ I_j \bar{X}_j\ /\ f(i,j)$

og

(9.5.2) $\hat{\gamma}^2 = \sum_j c^2(j)\, T^2_j\, (\underset{\sim}{L}).$

Bevis: Siden $E\, T^2_j\, (\underset{\sim}{L}) = E\, E\, \{T^2_j\, (\underset{\sim}{L}) \mid I_j, \underset{\sim}{L}\} = E\, \{I_j\, \tau^2_j\, (\underset{\sim}{L})\}$, blir

(9.5.3) $E\, T^2_j\, (\underset{\sim}{L}) = f(j)\, \eta(j).$

Det følger da umiddelbart at $E\hat{\gamma}^2 = \gamma^2$. Videre er

$E\{I_j \bar{X}_j^2\} = f(j)\, \eta(j) + \bar{a}^2(j)\, \{g(j,j) + f^2(j)\}$, eller

(9.5.4) $E\, \{I_j \bar{X}^2_j\} = f(j)\, \{\eta(j) + \bar{a}^2(j)\}.$

Altså blir

$$E\hat{\sigma}^2 = \sum_j c^2(j)\, \eta(j)\, f^2(j) + \sum_j c^2(j)\, [1 - f(j)]\, f(j)\, [\eta(j) + \bar{a}^2(j)]$$

$$+\, 2 \sum_{i<j} c(i)\, c(j)\, g(i,j)\, \bar{a}(i)\, \bar{a}(j) = \gamma^2 + \omega^2.$$

9.6. Selvveiende utvalg. Vi vil nå kalle et utvalg for selvveiende når estimatoren for a kan skrives på formen

$$\hat{a} = \sum_{\{j:I_j=1\}} \sum_s X_{js} \,/\, b(\underset{\sim}{L})$$

for en passende størrelse $b(\underset{\sim}{L})$. For å få dét til, vil vi kreve at

(9.6.1) $\dfrac{N(j)}{f(j)\, n_j\, (\underset{\sim}{L})} = \dfrac{1}{b(\underset{\sim}{L})}$ når $I_j = 1,$

samtidig som

(9.6.2) $\sum_j n_j\, (\underset{\sim}{L}) = n,$

der n er gitt. Kravet (9.6.1) gir

(9.6.3) $n_j\, (\underset{\sim}{L}) = b(\underset{\sim}{L})\, I_j\, N(j) \,/\, f(j),$

og kravet (9.6.2) gir

(9.6.4) $b(\underset{\sim}{L}) = n \,/\, \sum_j \{I_j\, N(j) \,/\, f(j)\}.$

Relasjonene (9.6.3) og (9.6.4) vil til sammen bestemme $n_j\, (\underset{\sim}{L})$-ene. (En annen måte å bestemme disse størrelsene på diskuteres i underkapittel 12.6.)

Siden vi nå får

$$I_j \cdot \frac{N(j) - n_j(\underset{\sim}{L})}{n_j(\underset{\sim}{L})\, N(j)} = f(j)\, I_j \{1 - b\, /\, f(j)\}\, /\, \{b\, N(j)\},$$

blir estimatoren for var $\hat{a}$ lik

$$\hat{\sigma}^2 = \frac{1}{b} \sum_{\{j:I_j=1\}} N(j)\, S^2_j(\underset{\sim}{L})\, \{1 - b\, /\, f(j)\}$$

$$+ \sum_i \sum_j c(i)\, c(j)\, I_i\, \bar{X}_i\, I_j\, \bar{X}_j\, g(i,j)\, /\, f(i,j).$$

Det er interessant å se hvordan man skal tolke b ($\underset{\sim}{L}$). La B_j være den begivenhet at en gitt intervjuenhet fra kommune nr. j kommer med i utvalget. Da er

$$P\{B_j \mid I_j = 1, \underset{\sim}{L}\} = n_j(\underset{\sim}{L})\, /\, N(j) = I_j\, b\, (\underset{\sim}{L}) / f(j)$$

ifølge (9.6.3). Altså blir

(9.6.6) $$P\{B_j\} = E\{I_j\, b\, (\underset{\sim}{L})\, /\, f(j)\} = E\{b\, (\underset{\sim}{L}) \mid I_j = 1\}.$$

Vi må regne med at denne avhenger av j.

Vi får endelig at

(9.6.7) $$E\, (\frac{1}{b(\underset{\sim}{L})}) = \frac{N}{n}.$$

Disse resultatene er analoge med dem i underkapittel 4.2.

9.7. Beregning av $f(j)$ og $f(i,j)$.[x)] La oss nå innføre betegnelsen $M_s(i)$ for antall primærområder fra stratum i som ligger i kommune j, og la

$$I_j(s) = \begin{cases} 1 \text{ hvis kommune nr. } j \text{ får minst ett utvalgsområde} \\ \quad \text{fra stratum } s, \\ 0 \text{ ellers.} \end{cases}$$

Det er hensiktsmessig å benytte notasjonene

$$f_s(j) = Pr\, \{I_j(s) = 0\},$$

$$f_s(i,j) = Pr\, \{I_j(s) = I_i(s) = 0\}.$$

Nå er

(9.7.1) $$f_s(j) = \binom{M_s - M_s(j)}{m_s} \Big/ \binom{M_s}{m_s}$$

og

x) Dette underkapitlet er skrevet av John Dagsvik.

$$(9.7.2) \qquad f_s(i,j) = \binom{M_s - M_s(i) - M_s(j)}{m_s} \Big/ \binom{M_s}{m_s}.$$

Siden

$$\{I_j = 0\} = \bigcap_s \{I_j(s) = 0\},$$

blir

$$f(j) = 1 - P\,\{I_j = 0\} = 1 - P\,\{\bigcap_s (I_j(s) = 0)\}.$$

Idet trekkingene i de ulike strata er uavhengige, blir derfor

$$(9.7.3) \qquad f(j) = 1 - \prod_s f_s(j).$$

Tilsvarende er

$$\{I_i = I_j = 0\} = \bigcap_s \{I_i(s) = I_j(s) = 0\},$$

noe som medfører at

$$P\,\{I_i = I_j = 0\} = \prod_s f_s\,(i,j).$$

Siden

$$f(i,j) = 1 - P\,(I_i = 0) - P\,(I_j = 0) + P\,(I_i = I_j = 0),$$

får vi den enkle formelen

$$(9.7.4) \qquad f(i,j) = f(i) + f(j) - 1 + \prod_s f_s(i,j).$$

Uttrykkene (9.7.1) og (9.7.2) kan finnes direkte ved hjelp av tabeller over den hypergeometriske fordeling (Lieberman og Owen, 1961).

10. Estimering av prosentfordelinger når kommunene er utvalgsområder.x)

10.1. Innledning. Som beskrevet i kapittel 6, er man ofte interessert i å finne hvordan en gruppe som har et gitt kjennetegn A er fordelt etter et annet kjennetegn B. Kapittel 6 behandler situasjonen når Byråets standard utvalgsplan brukes. Vi skal her se på hvilke forandringer som må til når utvalgsplanen er som i kapittel 9, dvs. når kommunene er primære utvalgsområder.

x) Dette kapitlet er i alt vesentlig skrevet av John Dagsvik.

10.2. Symboler og konvensjoner. La $a(j,k)$ være a-verdien for intervjuenhet nr. k i kommune nr. j. Vi skal spesielt la $a(j,k)=1$, dersom intervjuenheten har kjennetegnet A, og 0 ellers.

Når $a(j,k) = 1$, lar vi $a(j,k,\ell)$ være henholdsvis 1 eller 0 etter som verdien ℓ av kjennetegnet B opptrer eller ikke. Når $a(j,k) = 0$ lar vi $a(j,k,\ell) = 0$ for alle ℓ. La

$$a(j,\cdot,\ell) = \sum_k a(j,k,\ell),$$

$$\bar{a}(j,\cdot,\ell) = a(j,\cdot,\ell)/N(j),$$

$$a(\ell) = \sum_j a(j,\cdot,\ell),$$

$$\sigma(j,\ell,\ell') = \frac{1}{N(j)-1} \sum_k \{a(j,k,\ell) - \bar{a}(j,\cdot,\ell)\} \{a(j,k,\ell') - \bar{a}(j,\cdot,\ell')\},$$

$$\tau(\underset{\sim}{L},\ell,\ell') = \frac{\sigma(j,\ell,\ell')}{n_j(\underset{\sim}{L})} \frac{N(j) - n_j(\underset{\sim}{L})}{N(j)},$$

$$\eta(j,\ell,\ell') = E\,\{\tau_j(\underset{\sim}{L},\ell,\ell')/I_j = 1\},$$

$$\gamma(\ell,\ell') = \sum_j c^2(j)\, f(j)\, \eta(j,\ell,\ell'),$$

$$\omega(\ell,\ell') = \sum_{i,j}\sum c(i)\, c(j)\, \bar{a}(i,\cdot,\ell')\, \bar{a}(j,\cdot,\ell)\, g(i,j),$$

$$\sigma(\ell,\ell') = \gamma(\ell,\ell') + \omega(\ell,\ell'),$$

$$\sigma^2(j,\ell) = \sigma(j,\ell,\ell),$$

$$\tau_j^2(\underset{\sim}{L},\ell) = \tau_j(\underset{\sim}{L},\ell,\ell),$$

$$\gamma^2(\ell) = \gamma(\ell,\ell),$$

$$\omega^2(\ell) = \omega(\ell,\ell),$$

$$X_{js}(\ell) = a(j,K_{js},\ell) \qquad \text{når } L_r = j,$$

$$\bar{X}_j(\ell) = \sum_s X_{js}(\ell)/n_j(\underset{\sim}{L}),$$

$$S_j(\underset{\sim}{L},\ell,\ell') = \frac{I_j}{n_j(\underset{\sim}{L}) - 1} \sum_s \{X_{js}(\ell) - \bar{X}_j(\ell)\} \{X_{js}(\ell') - \bar{X}_j(\ell')\},$$

$$T_j(\underset{\sim}{L},\ell,\ell') = S_j(\underset{\sim}{L},\ell,\ell') \frac{N(j) - n_j(\underset{\sim}{L})}{N(j)\, n_j(\underset{\sim}{L})},$$

$$S_j^2(\underset{\sim}{L},\ell) = S_j(\underset{\sim}{L},\ell,\ell), \qquad T_j^2(\underset{\sim}{L},\ell) = T_j(\underset{\sim}{L},\ell,\ell).$$

10.3. Prosentfordelingen. En er interessert i et anslag for

(10.3.1) $$q(\ell) = \frac{a(\ell)}{a} .$$

$a(\ell)$ estimeres ved

(10.3.2) $$\hat{a}(\ell) = \sum_j c(j)\, I_j \bar{X}_j(\ell) = \sum_j N(j)\, I_j \bar{X}_j(\ell)/f(j),$$

og en lar

(10.3.3) $$\hat{a} = \sum_\ell \hat{a}(\ell).$$

Den naturlige estimator for q er da

(10.3.4) $$\hat{q}(\ell) = \frac{\hat{a}(\ell)}{\hat{a}} .$$

Et tilnærmet uttrykk for variansen til $\hat{q}(\ell)$ gitt ved (6.4.1). Vi får

(10.3.5) $$\text{var}\ \hat{a}(\ell) = \gamma^2(\ell) + \omega^2(\ell);$$

(10.3.6) $$\text{cov}\ \{\bar{X}_j(\ell),\ \bar{X}_j(\ell')\,|\,\underset{\sim}{L},\ I_j = 1\} = \tau_j(\underset{\sim}{L},\ell,\ell');$$

(10.3.7) $$\text{cov}\ \{\hat{a}(\ell),\ \hat{a}(\ell')\} = \tfrac{1}{2}\{\text{var}\ [\hat{a}(\ell) + \hat{a}(\ell')] - \text{var}\ \hat{a}(\ell) - \text{var}\ \hat{a}(\ell')\}$$
$$= \gamma(\ell,\ell') + \omega(\ell,\ell');$$
$$\text{cov}\ (\hat{a}(\ell),\hat{a}) = \sum_{\ell'} \text{cov}\ (\hat{a}(\ell),\ \hat{a}(\ell')).$$

Ved å bruke betegnelsene

$$\gamma(\ell,\cdot) = \sum_{\ell'} \gamma(\ell,\ell'),$$

$$\omega(\ell,\cdot) = \sum_{\ell'} \omega(\ell,\ell'),$$

får vi at

(10.3.8) $$\text{cov}\ (\hat{a}(\ell),\hat{a}) = \gamma(\ell,\cdot) + \omega(\ell,\cdot).$$

10.4. Estimering av varianser. Estimatorene

(10.4.1) $$\hat{\gamma}(\ell,\ell') = \sum_j c^2(j)\ T_j(\underset{\sim}{L},\ell,\ell')$$

og

(10.4.2) $\hat{\sigma}(\ell,\ell') = \sum_j c^2(j)\ f(j)\ T_j(\underset{\sim}{L},\ell,\ell')$

$$+ \sum_i \sum_j c(i)\ c(j)\ g(i,j)\ I_i \bar{X}_i(\ell)\ I_j\ X_j(\ell')/f(i,j)$$

er forventningsrette for henholdsvis $\gamma(\ell,\ell')$ og $\sigma(\ell,\ell')$.

Bevis: Ved å bruke (10.3.7) får vi

$$E\{T_j(\underset{\sim}{L},\ell,\ell')|\underset{\sim}{L},I_j\} = I_j\ \tau_j(\underset{\sim}{L},\ell,\ell'),$$

slik at

$$E\ T_j(\underset{\sim}{L},\ell,\ell') = f(j)\ \eta(j,\ell,\ell'),$$

og altså

$$E\ \hat{\gamma}(\ell,\ell') = \gamma(\ell,\ell').$$

Videre er

$$E\ \{\bar{X}_j(\ell)\ \bar{X}_j(\ell')|\underset{\sim}{L},\ I_j = 1\} = \text{cov}\ \{\bar{X}_j(\ell),\ \bar{X}_j(\ell')|\underset{\sim}{L},\ I_j = 1\}$$

$$+ E\ \{\bar{X}_j(\ell)|\underset{\sim}{L},\ I_j = 1\}\ E\ \{\bar{X}_j(\ell')|\underset{\sim}{L},\ I_j = 1\} = \tau_j(\underset{\sim}{L},\ell,\ell')$$

$$+ \bar{a}(j,\cdot,\ell)\ \bar{a}(j,\cdot,\ell').$$

Dette medfører at

$$E\ \{I_j\ \bar{X}_j(\ell)\ \bar{X}_j(\ell')\} = f(j)\ \eta(j,\ell,\ell') + f(j)\ \bar{a}(j,\cdot,\ell)\ \bar{a}(j,\cdot,\ell)$$

og

$$E\ \hat{\sigma}(\ell,\ell') = \sum c^2(j)\ f(j)^2\ \eta(j,\ell,\ell')$$

$$+ \sum_j c^2(j)\ g(j,j)\ \eta(j,\ell,\ell') + \sum_j c^2(j)\ g(j,j)\ \bar{a}(j,\cdot,\ell)\ \bar{a}(j,\cdot,\ell')$$

$$+ \sum_{i \neq j} c(i)\ c(j)\ g(i,j)\ \bar{a}(i,\cdot,\ell)\ \bar{a}(j,\cdot,\ell') = \gamma(\ell,\ell') + \omega(\ell,\ell'). \quad \square$$

10.5. Forenklinger når utvalget er selvveiende. Når utvalget er slik at

(10.5.1) $n_j(\underset{\sim}{L}) = \dfrac{N(j)}{f(j)}\ b(\underset{\sim}{L})\ I_j$

(sml. (9.6.3)), blir

(10.5.2) $\tau_j(\underset{\sim}{L}\ \ell,\ell') = \sigma(j,\ell,\ell')\ \dfrac{f(j) - b(\underset{\sim}{L})}{N(j)b(\underset{\sim}{L})}$.

Videre blir

$$(10.5.3) \qquad E\{\tfrac{1}{b} \mid I_j = 1\} = \frac{1}{n}\sum_i \frac{N(i)}{f(i)} P\{I_i = 1 \mid I_j = 1\}$$

$$= \frac{1}{n}\sum_i N(i) \frac{f(i,j)}{f(i)\,f(j)} = \frac{1}{n}\sum_i c(i)\, f(i,j) \,/\, f(j)$$

slik at

$$(10.5.4) \qquad \eta_j(\ell,\ell') = \frac{\sigma(j,\ell,\ell')}{n\,N(j)} \left(\frac{1}{n}\sum_i c(i)\, f(i,j) - 1\right).$$

Forventningsrette estimatorer for cov $\{\hat{a}(\ell), \hat{a}(\ell')\}$ og $\gamma(\ell,\ell')$ blir nå

$$(10.5.5) \qquad \hat{\sigma}(\ell,\ell') = \frac{1}{b}\sum_j N(j)\, S_j(\underset{\sim}{L},\ell,\ell')\,\{1 - b/f(j)\}$$

$$+ \sum_{i,j} c(i)\, c(j)\, I_i\, \bar{X}_i(\ell)\, I_j\, \bar{X}_j(\ell')\, g(i,j) \,/\, f(i,j),$$

$$(10.5.6) \qquad \hat{\gamma}(\ell,\ell') = \frac{1}{b}\sum_j N(j)\, S_j(\underset{\sim}{L},\ell,\ell')\,\{1 - b/f(j)\} \,/\, f(j).$$

11. Tilfellet $m_i = 1$.

11.1. Innledning. Som vi nevnte på slutten av kapittel 2, hender det at man bare bruker en tredjedel av de primære utvalgsområdene i hvert stratum (utenom Oslo) i en undersøkelse. Hvert stratum i Bergen og Trondheim representeres da med ett av sine utvalgsområder. Strataene utenom de tre største byene skal i prinsippet representeres med to utvalgsområder hver. Noen ganger vil imidlertid ett av områdene falle bort, f.eks. ved sykdom i intervjustaben, og da får en igjen tilfellet $m_i = 1$. I den foregående teorien har vi bygd på en forutsetning om at $m_i > 1$ ved variansestimeringen. (Se f.eks. underkapittel 3.3.) Vi skal nå se på et par forslag til hvordan variansen kan estimeres i et stratum der $m_i = 1$. Forslagene går ut på

(i) å slå sammen to strata med ett utvalgsområde hver, og

(ii) å splitte opp dataene fra et utvalgsområde.

Disse idéene kan føres videre på mange måter, men vi skal ikke gjøre det her. Vår hensikt er først og fremst å illustrere problemstillingen.

11.2. Sammenslåing av to strata med ett utvalgsområde hver. En måte å behandle tilfellet $m_i = 1$ på, er å slå stratum i sammen med et annet stratum. Man må da sondre mellom det tilfellet at det andre stratumet har to eller flere utvalgsområder, og det tilfellet at også det andre stratumet bare har ett utvalgsområde. Vi skal her se på det siste tilfellet.

Anta at stratum nr. 1 og 2 har ett utvalgsområde hver, slik at $m_1 = m_2 = 1$. G_1^2 og G_2^2 er da fortsatt forventningsrette estimatorer for γ_1^2 og γ_2^2, men det er ikke umiddelbart gitt hvordan vi skal estimere σ_1^2 og σ_2^2.

La nå $U_{1,2}^2 = (\bar{V}_1 - \bar{V}_2)^2$. (Når $m_i = 1$, blir naturligvis $\bar{V}_i = V_{i1}$.) Siden $\bar{V}_1$ og $\bar{V}_2$ jo er stokastisk uavhengige, blir

$$EU_{1,2}^2 = E\{(\bar{V}_1 - E\bar{V}_1) + (E\bar{V}_1 - E\bar{V}_2) - (\bar{V}_2 - E\bar{V}_2)\}^2$$

$$= \text{var } \bar{V}_1 + \text{var } \bar{V}_2 + (E\bar{V}_1 - E\bar{V}_2)^2$$

$$(11.2.1) \qquad = \gamma_1^2 + (1 - \frac{1}{M_1})\sigma_1^2 + \gamma_2^2 + (1 - \frac{1}{M_2})\sigma_2^2 + (\bar{a}_1 - \bar{a}_2)^2.$$

Hvis

$$(11.2.2) \qquad \bar{a}_1 = \bar{a}_2 \text{ og } \sigma_1^2 = \sigma_2^2,$$

blir derfor

$$S_{1,2}^2 = \frac{U_{1,2}^2 - G_1^2 - G_2^2}{2 - \frac{1}{M_1} - \frac{1}{M_2}}$$

en forventningsrett estimator for σ_1^2 $(= \sigma_2^2)$. Hvis (11.2.2) ikke nødvendigvis gjelder, blir

$$(11.2.3) \qquad ES_{1,2}^2 = \sigma_1^2 + \frac{(1 - \frac{1}{M_2})(\sigma_2^2 - \sigma_1^2) + (\bar{a}_1 - \bar{a}_2)^2}{2 - \frac{1}{M_1} - \frac{1}{M_2}}.$$

Hensikten med å ha to strata opprinnelig, istedenfor å slå dem sammen til ett stratum for trekking av utvalgsområdene, må vel i første rekke være at $\bar{a}_1 \neq \bar{a}_2$. Dette betyr at $S_{1,2}^2$ vil få en tendens til å overestimere σ_1^2 (hvis $\sigma_1^2 \leqq \sigma_2^2$).

La oss anta at vi i alle tilfelle bruker $S_{1,2}^2$ som estimator for σ_1^2 og σ_2^2. Bidraget til estimatoren for var $\hat{a}$ fra disse to strataene blir da i alt

$$M_1^2[G_1^2 + (1 - \frac{1}{M_1})S_{1,2}^2] + M_2^2[G_2^2 + (1 - \frac{1}{M_2})S_{1,2}^2]$$

$$(11.2.4) \quad = \frac{M_1M_2[M_1(M_1-1) + M_2(M_2-1)]U_{1,2}^2 + (M_1^2-M_2^2)\ [M_1(M_2-1)G_1^2-M_2(M_1-1)G_2^2]}{M_1(M_2-1) + M_2(M_1-1)}.$$

Hvis $M_1 = M_2$, blir dette lik

$$(11.2.5) \quad M_1^2\ U_{1,2}^2.$$

Dersom de to strataene hadde vært slått sammen til ett før trekkingen av utvalgsområdene, og ren lotterisk trekking av to områder fra det sammenslåtte stratumet gav de to samme utvalgsområdene som dem vi hadde ovenfor, ville bidraget til estimatoren for var $\hat{a}$ blitt

$$\{1 - \frac{1}{2}\ b(M_1+M_2)\}\ (\frac{M_1}{1-bM_1}\ G_1^2 + \frac{M_2}{1-bM_2}\ G_2^2) + \frac{1}{2}\ (M_1+M_2)(M_1+M_2-2)\ U_{1,2}^2.$$

Med $M_1 = M_2$ reduserer dette seg til

$$M_1(G_1^2 + G_2^2) + 2M_1(M_1-1)\ U_{1,2}^2.$$

Vi ser at disse uttrykkene er vesentlig annerledes enn dem i (11.2.4) og (11.2.5). <u>Sammenslåing av to strata etter at utvalgsområdene er trukket, gir altså helt andre resultater enn sammenslåing før trekking.</u>

<u>11.3. Oppsplitting av dataene fra et utvalgsområde.</u> En alternativ framgangsmåte når $m_i = 1$, er å dele materialet fra det ene utvalgsområde man har, i to deler og utnytte dette i variansestimeringen. En slik oppdeling av materialet faller jo naturlig hvis man har hatt to intervjuere i aktivitet i området, men oppdelingen kan naturligvis gjennomføres uavhengig av dette. Vi skal her bare anta at det foreligger en eller annen regel som, når $J_{i1} = j_{i1}$, deler opp de $n_{i1}(j_{i1})$ uttrukne intervjuenhetene i én gruppe på $n_i'(j_{i1}) \geqq 2$ stykker og en annen gruppe på $n_i''(j_{i1}) = n_{i1}(j_{i1}) - n_i'(j_{i1}) \geqq 2$ stykker, f.eks. henholdsvis de $n_i'(j_{i1})$ først uttrukne intervjuenhetene og de $n_i''(j_{i1})$ siste.

Det er fortsatt ingen problemer med estimeringen av γ_i^2. Vi bruker

$$G_i^2 = N_i^2(J_{i1})\ T_{i1}^2(J_{i1})$$

som estimator.

For å estimere σ_i^2, danner vi

$$\bar{X}'_i = \sum_{s=1}^{n'_i} X_{i1s}/n'_i, \quad \bar{X}''_i = \sum_{s=1}^{n''_i} X_{i,1,n'_i+s}/n''_i,$$

og

$$U_i^2 = \frac{1}{2}(\bar{X}'_i - \bar{X}''_i)^2.$$

Da blir $E\bar{X}'_i = E\bar{X}''_i = \bar{a}_i$ og

$$\text{var } \bar{X}'_i = (\gamma'_i)^2 + (1 - \frac{1}{M_i})\sigma_i^2,$$

der

$$(\gamma'_i)^2 = \sum_j N_i^2(j) \; [\tau'_i(j)]^2 \; / \; M_i,$$

med

$$[\tau'_i(j)]^2 = \frac{\sigma_i^2(j)}{n'_i(j)} \frac{N_i(j) - n'_i(j)}{N_i(j)} .$$

Formelen for var $\bar{X}''_i$ blir analog, og

$$(11.3.1) \qquad c = \text{cov } (\bar{X}'_i, \bar{X}''_i) = -\frac{1}{M_i} \sum_j \frac{\sigma_i^2(j)}{N_i(j)},$$

som vi skal vise nedenfor. Videre blir

$$EU_i^2 = \tfrac{1}{2} E \{(\bar{X}'_i - E\bar{X}'_i) - (\bar{X}''_i - E\bar{X}''_i)\}^2$$

$$= \tfrac{1}{2} \text{ var } \bar{X}'_i + \tfrac{1}{2} \text{ var } \bar{X}''_i - \text{cov } (\bar{X}'_i, \bar{X}''_i)$$

$$= \tfrac{1}{2}\{(\gamma'_i)^2 + (\gamma''_i)^2\} + (1 - \frac{1}{M_i})\sigma_i^2 - c$$

$$(11.3.2) \qquad = \frac{1}{M_i} \sum_j \frac{\sigma_i^2(j)}{N_i(j)} \{N_i^2(j) \left[\frac{1}{2} N_i(j) (\frac{1}{n'_i(j)} + \frac{1}{n''_i(j)}) - 1\right] + 1\}$$

$$+ (1 - \frac{1}{M_i})\sigma_i^2.$$

La

$$Y_i^2 = \frac{S_{i1}^2(J_{i1})}{N_i(J_{i1})} \{N_i^2(J_{i1}) \left[\frac{1}{2} N_i(J_{i1}) \left(\frac{1}{n'_i(J_{i1})} + \frac{1}{n''_i(J_{i1})}\right) - 1\right] + 1\}.$$

Da er Y_i^2 en forventningsrett estimator for den første summen i (11.3.2), og

$$\frac{M_i}{M_i - 1}(U_i^2 - Y_i^2)$$

er en forventningsrett estimator for σ_i^2. Det bidrag som stratum i gir til estimatoren for var $\hat{a}$, blir derfor

$$M_i^2\ (G_i^2 + U_i^2 - Y_i^2).$$

Siden

$$G_i^2 = N_i(J_{i1})\ S_{i1}^2(J_{i1})\ \{N_i(J_{i1}) - n_i(J_{i1})\}\ /\ n_{i1}(J_{i1}),$$

kan bidraget til estimatoren også skrives som

$$M_i^2 S_i^2(J_{i1})\Big[\frac{1}{2}\,N_i^2(J_{i1})\ \Big(\frac{2}{n_i(J_{i1})} - \frac{1}{n_i'(J_{i1})} - \frac{1}{n_i''(J_{i1})}\Big) - \frac{1}{N_i(J_{i1})}\Big]$$

$$+\ M_i^2 U_i^2.$$

11.4. Bevis for formel (11.3.1). Vi får at

$$E(\bar{X}_i'\ \bar{X}_i''\ |J_{i1} = j) = \frac{1}{n_i'(j)\ n_i''(j)} \sum_{r=1}^{n_i'(j)} \sum_{s=n_i'(j)+1}^{n_{i1}(j)} E(X_{i1r}X_{i1s}\ |\ J_{i1} = j)$$

$$= \sum_{k \neq \ell} a_i(j,k)\ a_i(j,\ell)\ /\ [N_i(j)\ (N_i(j)-1)]$$

$$= \sum_k a_i(j,k)\Big[\sum_\ell a_i(j,\ell) - a_i(j,k)\Big]/[N_i(j)\ (N_i(j) - 1)]$$

$$=\{N_i(j)\ \bar{a}_i^2(j) - \frac{1}{N_i(j)} \sum_k a_i^2(j,k)\}/[N_i(j) - 1].$$

Idet $E(\bar{X}_i'\ |\ J_{i1} = j)\ =\ E(\bar{X}_i''\ |J_{i1} = j)\ =\ \bar{a}_i(j)$, blir derfor

$$\text{cov}\ \{\bar{X}_i',\bar{X}_i''\ |J_{i1} = j\} = -\ \sigma_i^2(j)/N_i(j).$$

Formel (11.3.1) følger da umiddelbart.

12. En annen måte å bestemme n_{ir} på: Fiksér b og la n være stokastisk.

12.1. Innledning. Vi diskuterte Byråets spesielle valg av $n_{ir}(\underset{\sim}{J})$ i kapittel 4. Et viktig aspekt ved denne måten å bestemme utvalgsstørrelsene i strataene på, var at en tok n som fiksert og bestemte total utvalgsbrøk b av (4.1.2), dvs.

$$(12.1.1) \qquad b = n/\sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}).$$

Størrelsen b blir da en stokastisk variabel. Enkelte trekk ved teorien i de senere kapitlene er sterkt preget av dette.

En kunne imidlertid snu bak fram på den prosedyren som gav (12.1.1), nemlig ved å fiksere verdien av b, og bruke (12.1.1) til å bestemme n, som da blir en stokastisk variabel. I analogi med underkapittel 4.1 lar vi da

(12.1.2) $\quad b_i = b\, M_i/m_i,$

(12.1.3) $\quad n_{ir}(J_{ir}) = b_i\, N_i(J_{ir})$

og

(12.1.4) $\quad n(\underset{\sim}{J}) = \sum_i \sum_r n_{ir}(J_{ir}).$

(Funksjonsargumentene angir her hva de enkelte variable i realiteten avhenger av.)

En stor del av den teorien man finner i tidligere litteratur, bygger på et opplegg av denne typen. Dette gjelder f.eks. for Byråets artikkel nr. 37 (Tamsfoss, 1970).

12.2. Tolkning av utvalgsbrøkene. Siden $EN_i(J_{ir}) = N_i/M_i$, blir $En(\underset{\sim}{J}) = bN$, der N fortsatt er det samlede antall trekkeenheter i landet. Utvalgsbrøken b har altså tolkningen

(12.2.1) $\quad b = E\,\dfrac{n(\underset{\sim}{J})}{N},$

dvs. b er den forventede utvalgsandelen i undersøkelsen.

Hvis B_{ijk} fortsatt er den begivenhet at trekkeenhet nr. k i utvalgsområde nr. j i stratum nr. i trekkes ut for intervjuing, slik som i underkapittel 4.2, får vi nå at

$$b_i = P\,\{B_{ijk} \mid I_{ij} = 1,\ \underset{\sim}{J} = j\},$$

og følgelig at

$$P\,\{B_{ijk} \mid I_{ij}\} = I_{ij}\, b_i.$$

Siden $E\, I_{ij} = m_i/M_i$, blir derfor

$$P\,\{B_{ijk}\} = b.$$

I den nåværende situasjon er altså b lik sannsynligheten for at en vilkårlig trekkeenhet skal bli trukket ut for intervjuing. Denne sannsynligheten er derfor den samme i alle strata.

Selv om svært mange av resultatene i kapitlene foran blir de samme når b fikseres som når n fikseres, er det likevel enkelte slående eksempler på at resultatene forenkles når b fikseres istedenfor n, i tillegg til dem vi så ovenfor i dette underkapitlet. Vi skal nevne noen slike eksempler i underkapitlene nedenfor.

12.3. Endringer i underkapittel 4.3 når b fikseres. En får nå følgende formel for variansen til $\hat{a}$:

$$\text{var } \hat{a} = \frac{1}{b} \sum_i (1-b_i) \sum_j N_i(j)\, \sigma_i^2(j) + \frac{1}{b} \sum_i (M_i - m_i)\, \sigma_i^2\, b_i \,; \tag{12.3.1}$$

og estimatoren for var $\hat{a}$ blir

$$\frac{1}{b^2} \sum_i (1-b_i)\, b_i \sum_r N_i(J_{ir})\, S_i^2(J_{ir}) + \frac{1}{b^2} \sum_i \left(1 - \frac{m_i}{M_i}\right) m_i Z_i^2. \tag{12.3.2}$$

I estimeringsformelen får en altså bare den trivielle forandringen at $b(\underset{\sim}{J})$ og $b_i(\underset{\sim}{J})$ erstattes med henholdsvis b og b_i.

12.4. Endringer i kapittel 5. Når b er fiksert, blir $n_0 = bN_0$ ikke- stokastisk, og

$$\text{var } \hat{a}_0 = \frac{\sigma_0^2}{n_0} N_0 (N_0 - n_0) = \frac{1-b}{b} N_0 \sigma_0^2.$$

12.5. Endringer i underkapittel 7.4. Når b fikseres, blir naturligvis alle $\beta_i(j_{ir})$ lik b, og $\alpha = a(1)$. $a^*(1)$ blir derfor en (trolig noe forventningsskjev) estimator for den estimand en der er interessert i.

12.6. Endringer i underkapittel 9.6. Når kommunene brukes som utvalgsområder, består den framgangsmåten som er analog til den ovenstående, i å fiksere b, bestemme $n_j(\underset{\sim}{L})$ av

$$n_j(\underset{\sim}{L}) = I_j\, N(j)\, b/f(j), \tag{12.6.1}$$

og la den totale utvalgsstørrelsen være

$$n(\underset{\sim}{L}) = \sum_j n_j(\underset{\sim}{L}).$$

Da blir sannsynligheten for at en vilkårlig telleenhet fra kommune nr. j skal trekkes ut for intervjuing, lik b for alle j. Analogien med det ovenstående er slående.

Referanser.

[1] Dagsvik, John (1972): "Utvalgsplan for flyttemotivundersøkelsen 1972; prøveundersøkelse." Maskinskrevet notat JD/WA, 4/4-72.

[2] Dahmström, Per (1968): "Användning av supplementär information i skattningsprocessen vid stickprovsundersökingar av ändlige populationer." Stensiltrykk. Statistiska institutionen, Lunds universitet.

[3] Dahmström, Per og Staffan Wahlström (1969): "Några urvals- och estimationsmetoder tillämpade på politiska opinionsundersøkningar." Statistisk tidskrift III 7 (4) : 308-315.

[4] Fuller, Wayne A. (1966): "Estimation employing post strata." J. Am. Statist. Ass. 61 : 1172-1183.

[5] Hansen, Morris H.; William N. Hurwitz; og William G. Madow (1953): "Sample Survey Methods and Theory." 2 bind. Wiley.

[6] Kish, Leslie (1957): "Confidence intervals for clustered samples." Am. Sociol. Rev. 22 (2) : 154-165.

[7] Lansing, John B. og James N. Morgan (1971): "Economic Survey Methods." Survey Research Center, The University of Michigan, Ann Arbor.

[8] Lieberman, G.J. og D.B. Owen (1961): "Tables of the Hypergeometric Probability Distribution". Stanford University Press.

[9] Statistisk Sentralbyrå (1972): "Variansberegninger ved intervjuundersøkelser." Ti stensilnotater, hvorav åtte er skrevet av Jan M. Hoem og to av John Dagsvik.

[10] Sverdrup, Erling (1964): "Lov og tilfeldighet." Bind I. Universitetsforlaget, Oslo etc.

[11] Sverige. Statistiska Centralbyrån (1970/1971/1972): "Teori och metodik vid partisympatiundersökningar." Forskningsrapporter. Tre utkommet hittil, en fjerde er under utarbeidelse.

[12] Tamsfoss, Steinar (1970): "Om bruk av stikkprøver ved Kontoret for intervjuundersøkelser, Statistisk Sentralbyrå." SSB-artikkel nr. 37.

[13] Tamsfoss, Steinar (1969): "Utvalgsplan for undersøkelse om boligfinansiering." Maskinskrevet notat ST/KJ, 2/7-69.

Utkommet i serien ART

Issued in the series Artikler fra Statistisk Sentralbyrå (ART)

Nr. 1 Odd Aukrust: Investeringenes effekt på nasjonalproduktet *The Effects of Capital Formation on the National Product* 1957 28 s. Utsolgt

" 2 Arne Amundsen: Vekst og sammenhenger i den norske økonomi 1920 - 1955 *Growth and Interdependence in Norwegian Economy* 1957 40 s. Utsolgt

" 3 Statistisk Sentralbyrås forskningsavdeling: Skattlegging av personlige skattytere i årene 1947 - 1956 *Taxation of Personal Tax Payers* 1957 8 s. Utsolgt

" 4 Odd Aukrust og Juul Bjerke: Realkapital og økonomisk vekst 1900 - 1956 *Real Capital and Economic Growth* 1958 32 s. kr. 3,50

" 5 Paul Barca: Utviklingen av den norske jordbruksstatistikk *Development of the Norwegian Agricultural Statistics* 1958 23 s. kr. 2,00

" 6 Arne Amundsen: Metoder i analysen av forbruksdata *Methods in Family Budget Analyses* 1960 24 s. kr. 5,00

" 7 Arne Amundsen: Konsumelastisiteter og konsumprognoser bygd på nasjonalregnskapet *Consumer Demand Elasticities and Consumer Expenditure Projections Based on National Accounts Data* 1963 44 s. kr. 5,00 Utsolgt

" 8 Arne Øien og Hallvard Borgenvik: Utviklingen i personlige inntektsskatter 1952 - 1964 *The Development of Personal Income Taxes* 1964 30 s. kr. 5,00

" 9 Hallvard Borgenvik: Personlige inntektsskatter i sju vesteuropeiske land *Personal Income Taxes in Seven Countries in Western Europe* 1964 16 s. kr. 5,00

" 10 Gerd Skoe Lettenstrøm og Gisle Skancke: De yrkesaktive i Norge 1875 - 1960 og prognoser for utviklingen fram til 1970 *The Economically Active Population in Norway and Forecasts up to 1970* 1964 56 s. kr. 6,00 Utsolgt

" 11 Hallvard Borgenvik: Aktuelle skattetall 1965 *Current Tax Data* 1965 38 s. kr. 6,00 Utsolgt

" 12 Idar Møglestue: Kriminalitet, årskull og økonomisk vekst *Crimes, Generations and Economic Growth* 1956 63 s. kr. 7,00

" 13 Svein Nordbotten: Desisjonstabeller og generering av maskinprogrammer for granskning av statistisk primærmateriale *Decision Tables and Generation of Computer Programs for Editing of Statistical Data* 1965 11 s. kr. 4,00

" 14 Gerd Skoe Lettenstrøm: Ekteskap og barnetall - En analyse av fruktbarhetsutviklingen i Norge *Marriages and Number of Children - An Analysis of Fertility Trend in Norway* 1965 29 s. kr. 6,00

" 15 Odd Aukrust: Tjue års økonomisk politikk i Norge: Suksesser og mistak *Twenty Years of Norwegian Economic Policy: An Appraisal* 1965 38 s. kr. 6,00 Utsolgt

Nr. 16 Svein Nordbotten: Long-Range Planning, Progress- and Cost-Reporting in the Central Bureau of Statistics of Norway *Langtidsprogrammering, framdrifts- og kostnadsrapportering i Statistisk Sentralbyrå* 1966 17 s. kr. 4,00

" 17 Olav Bjerkholt: Økonomiske konsekvenser av nedrustning i Norge *Economic Consequences of Disarmament in Norway* 1966 25 s. kr. 4,00 Utsolgt

" 18 Petter Jakob Bjerve: Teknisk revolusjon i økonomisk analyse og politikk? *Technical Revolution in Economic Analysis and Policy?* 1966 23 s. kr. 4,00

" 19 Harold W. Watts: An Analysis of the Effects og Transitory Income on Expenditure of Norwegian Households 1968 28 s. kr. 5,00

" 20 Thomas Schiøtz: The Use of Computers in the National Accounts of Norway *Bruk av elektronregnemaskiner i nasjonalregnskapsarbeidet i Norge* 1968 28 s. kr. 5,00

" 21 Petter Jakob Bjerve: Trends in Quantitative Economic Planning in Norway *Utviklingstendensar i den kvantitative økonomiske planlegginga i Norge* 1968 29 s. kr. 5,00

" 22 Kari Karlsen og Helge Skaug: Statistisk Sentralbyrås sentrale registre *Registers in the Central Bureau of Statistics* 1968 24 s. kr. 3,50

" 23 Per Sevaldson: MODIS II A Macro-Economic Model for Short-Term Analysis and planning *MODIS II En makroøkonomisk modell for korttidsanalyse og planlegging* 1968 40 s. kr. 4,50

" 24 Olav Bjerkholt: A Precise Description of the System of Equations of the Economic Model MODIS III *Likningssystemet i den økonomiske modell MODIS III* 1968 30 s. kr. 4,50 Utsolgt

" 25 Eivind Hoffmann: Prinsipielt om måling av samfunnets utdanningskapital og et forsøk på å måle utdanningskapitalen i Norge i 1960 *On the Measurement of the Stock of Educational Capital and an Attempt to Measure Norway's Stock of Educational Capital in 1960* 1968 60 s. kr. 5,00

" 26 Hallvard Borgenvik: Aktuelle skattetall 1968 *Current Tax Data* 1969 40 s. kr. 7,00

" 27 Hallvard Borgenvik: Inntekts- og formuesskattlegging av norske kapitalplasseringer i utlandet *Income and Net Wealth Taxes of Norwegian Investment in Foreign Countries* 1969 40 s. kr. 7,00

" 28 Petter Jakob Bjerve og Svein Nordbotten: Automasjon i Statistikkproduksjonen *Automation of the Production of Statistics* 1969 30 s. kr. 7,00

" 29 Tormod Andreassen: En analyse av industriens investeringsplaner *An Analysis of the Industries Investment Plans* 1969 26 s. kr. 5,00

" 30 Bela Balassa og Odd Aukrust: To artikler om norsk industri *Two Articles on Norwegian Manufacturing Industries* 1969 40 s. kr. 5,00

" 31 Hallvard Borgenvik og Hallvard Flø: Virkninger av skattereformen av 1969 *Effects of the Taxation Reform of 1969* 1969 35 s. kr. 7,00 Utsolgt

Nr. 32 Per Sevaldson: The Stability of Input-Output Coefficients *Stabilitet i kryssløpskoeffisienter* 1969 40 s. kr. 7,00

" 33 Odd Aukrust og Hallvard Borgenvik: Inntektsfordelingsvirkninger av skattereformen av 1969 *Income Distribution Effects of the Taxation Reform of 1969* 1969 29 s. kr. 7,00

" 34 Odd Aukrust og Svein Nordbotten: Dataregistrering, dataarkiver og samfunnsforskning *Data Registration, Data Banks and Social Research* 1970 43 s. kr. 7,00

" 35 Odd Aukrust: PRIM I A Model of the Price and Income Distribution Mechanism of an Open Economy *PRIM I En modell av pris- og inntektsfordelingsmekanismen i en åpen økonomi* 1970 61 s. kr. 7,00

" 36 Arne Amundsen: Konsumets og sparingens langsiktige utvikling *Consumption and Saving in the Process of Long-Term Growth* 1970 18 s. kr. 5,00

" 37 Steinar Tamsfoss: Om bruk av stikkprøver ved kontoret for intervjuundersøkelser, Statistisk Sentralbyrå *On the use of Sampling Surveys by the Central Bureau of Statistics, Norway* 1970 46 s. kr. 7,00

" 38 Svein Nordbotten: Personmodeller, personregnskapssystemer og persondataarkiver *Population Models, Population Accounting Systems and Individual Data Banks* 1970 28 s. kr. 7,00

" 39 Julie E. Backer: Variasjoner i utviklingen hos nyfødte barn *Variations in the Maturity Level of New Born Infants* 1970 36 s. kr. 7,00

" 40 Svein Nordbotten: Two Articles on Statistical Data Files and Their Utilization in Socio-Demographic Model Building *To artikler om statistiske dataarkiver og deres bruk i sosio-demografisk modellbygging* 1971 30 s. kr. 7,00

" 41 Per Sevaldson: Data Sources and User Operations of MODIS, a Macro-Economic Model for Short Term Planning *Datagrunnlag og brukermedvirkning ved MODIS, en makroøkonomisk modell for planlegging på kort sikt* 1971 31 s. kr. 7,00

" 42 Erik Biørn: Fordelingsvirkninger av indirekte skatter og subsidier *Distributive Effects of Indirect Taxes and Subsidies* 1971 42 s. kr. 5,00

" 43 Hallvard Borgenvik og Inger Gabrielsen: Aktuelle skattetall 1970 *Current Tax Data* 1971 53 s. kr. 7,00

" 44 Vidar Ringstad: PRIM II En revidert versjon av pris- og inntektsmodellen *PRIM II A Revised Version of the Price and Income Model* 1972 43 s. kr. 7,00

" 45 Jan M. Hoem: Purged and Partial Markov Chains *Lutrede og partielle Markovkjeder* 1972 16 s. kr. 5,00

" 46 Jan M. Hoem: Two Articles on the Interpretation of Vital Rates *To artikler om tolking av befolkningsrater* 1972 33 s. kr. 7,00

" 47 Inger Gabrielsen: Aktuelle skattetall 1972 *Current Tax Data* 1972 58 s. kr. 8,00

Publikasjonen utgis i kommisjon hos
H. Aschehoug & Co., Oslo, og er til salgs hos alle bokhandlere
Pris kr. 8,00