Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 74/7 4. februar 1974

METODEHEFTE NR. 11

Notater om metoder ved utvalgsundersøkelser

INNHOLD

FORORD

Metodehefter i serien Arbeidsnotater

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

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

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

ESTIMERING AV VARIANSEN TIL ESTIMATOREN FOR POPULASJONSVERDIEN a_o FOR OSLO I BYRÅETS INTERVJUUNDERSØKELSER

Av

PETTER LAAKE

I Byråets standard utvalgsplan er utvalgsstørrelsen n fastlagt på forhånd. For å oppnå at utvalget blir selvveiende, lar vi (Hoem, 1973, side 14)

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

og

$$b_i(\underset{\sim}{J}) = b(\underset{\sim}{J}) \frac{M_i}{m_i} .$$

Altså blir

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

Her er

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

Vi omtaler vanligvis $b(\underset{\sim}{J})$ som "total utvalgsbrøk". For særtilfellet Oslo (Hoem, 1973, side 16-17) lar vi antall trekkenheter i populasjonen være N_o. Av disse trekkes et utvalg på

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

rent lotterisk. En forventningsrett estimator for totalen a_o er da

$$\hat{a}_o = N_o \bar{X}_o.$$

Videre er

$$\text{var}\,(\bar{X}_o | \underset{\sim}{J} = \underset{\sim}{j}) = \frac{\sigma_o^2}{n_o(\underset{\sim}{j})} \left(1 - \frac{n_o(\underset{\sim}{j})}{N_o}\right),$$

og dermed

$$E\,\text{var}\,(\bar{X}_o | \underset{\sim}{J}) = \frac{\sigma_o^2}{N_o} \left(E \frac{1}{b(\underset{\sim}{J})} - 1\right) = \frac{\sigma_o^2}{N_o} \left(\frac{N}{n} - 1\right).$$

Vi finner da at

$$\text{var}\,\hat{a}_o = \frac{\sigma_o^2}{n} N_o\,(N-n).$$

En forventningsrett estimator for σ_o^2 er gitt ved

$$S_o^2 = \frac{1}{n_o(\underset{\sim}{J})-1} \sum_s (X_{os} - \bar{X}_o)^2.$$

Hoem (1973, side 17) foreslår som en forventningsrett estimator for variansen:

$$\text{lst}_1 \text{ var } \hat{a}_o = S_o^2 N_o (\frac{N}{n} - 1).$$

Dessverre er N sjelden eller aldri eksakt kjent fra våre registre. Vi skal derfor finne en annen estimator for var $\hat{a}_o$.

Vi lar

$$\hat{N} = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}),$$

og setter

$$\text{est}_2 \text{ var } \hat{a}_o = S_o^2 N_o (\frac{\hat{N}}{n} - 1).$$

Denne er da forventningsrett for var $\hat{a}_o$. Det vises slik:
Siden

$$\hat{N} = n/b(\underset{\sim}{J}),$$

blir

$$\text{est}_2 \text{ var } \hat{a}_o = S_o^2 N_o (\frac{1}{b(\underset{\sim}{J})} - 1).$$

Dermed er

$$\begin{aligned} E \text{ est}_2 \text{ var } \hat{a}_o &= E\{S_o^2 N_o(\frac{1}{b(\underset{\sim}{J})} - 1)\} \\ &= EE\{S_o^2 N_o(\frac{1}{b(\underset{\sim}{J})} - 1)|\underset{\sim}{J}\} = N_o\sigma_o^2 E\{\frac{1}{b(\underset{\sim}{J})} - 1\} \\ &= N_o\sigma_o^2 (\frac{N}{n} - 1) = \text{var } \hat{a}_o. \end{aligned}$$

Referanse:

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

ESTIMERING AV POPULASJONSGJENNOMSNITT OG ESTIMERING AV VARIANSEN TIL ESTIMATOREN I BYRÅETS INTERVJUUNDERSØKELSER

Av

PETTER LAAKE

INNHOLD

1. Innledning

Dette notatet har som formål å angi en estimator for et gjennomsnitt i en endelig populasjon og å angi variansen til denne estimatoren. For variansen til denne primære estimatoren har vi videre funnet en estimator som er tilnærmet forventningsrett under en gitt forutsetning. Vi har også angitt estimatorer for varianser til estimatorene for totalen i populasjonen og for totalt antall trekkeenheter i den. Disse **varians**-estimatorene er eksakt forventningsrette uansett om den nevnte forutsetningen er oppfylt eller ikke. I avsnittene 2 til 5 har vi tenkt oss at de enhetene som er trukket ut (trekkeenhetene), skal være grunnlag for analysen. I praksis vil ofte dette være en uheldig innskrenking. I avsnittene 6 til 10 har vi derfor tillatt andre enheter å være grunnlag for analysen. Dersom det er akkurat én analyseenhet i hver trekkeenhet, faller trekke- og analyseenhetene sammen. Det betyr at avsnittene 2 til 5 diskuterer et spesialtilfelle av teorien i avsnittene 6 til 10. Avsnittene 2 til 5 er likevel tatt med for å gi en enhetlig oversikt over problemområdet.

Framstillingen i notatet er i det vesentlige en anvendelse av teorien hos Hoem (1973), og vi skal bruke samme symboler som ham. Dette notatet atskiller seg likevel fra Hoem (1973) på to punkter. Hoem (1973) antar at totalt antall trekkeenheter i populasjonen er kjent. Dessverre er dette antallet sjelden kjent i Byråets undersøkelser. I dette notatet regner vi derfor antallet som ukjent, og vi estimerer det. Vi innfører videre en tilnærmelse. For de tilfellene der tilnærmelsen er god, har vi angitt en tilnærmet forventningsrett estimator for variansen til estimatoren.

2. Matematisk grunnlag når utvalgsstørrelsen er fastlagt

Vi antar at populasjonen er delt i strata, og at i-te stratum har M_i utvalgsområder. Utvalgsområde j har $N_i(j)$ trekkeenheter, og trekkeenhet k har verdien

$$a_i(j,k)$$

på den variable vi måler. Totalen i populasjonen er da

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

På første trinn trekkes m_i utvalgsområder rent lotterisk. I r-te uttrukne utvalgsområde i stratum i trekkes $n_{ir}(J)$ trekkeenheter rent lotterisk.

Vår estimator for totalen er da

$$\hat{a} = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir})\,\bar{X}_{ir}. \tag{2.1}$$

La

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

Da er

$$\hat{a} = \sum_i \frac{M_i}{m_i} \sum_r V_{ir}. \tag{2.2}$$

I Hoem (1973) er det vist at $\hat{a}$ er forventningsrett for a. Vi lar nå

$$b(\underset{\sim}{J}) = n/\sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}), \tag{2.3}$$

og

$$b_i(\underset{\sim}{J}) = b(\underset{\sim}{J})\,\frac{M_i}{m_i}. \tag{2.4}$$

Videre velger vi

$$n_{ir}(\underset{\sim}{J}) = b_i(\underset{\sim}{J})\,N_i(J_{ir}). \tag{2.5}$$

og oppnår da at utvalget blir selvveiende. Da er altså

$$\hat{a} = \frac{1}{b(\underset{\sim}{J})} \sum_i \sum_r \sum_s X_{irs}. \tag{2.6}$$

Siden vi regner totalt antall trekkeenheter N i populasjonen som ukjent, er vi også interessert i å estimere N. En forventningsrett estimator er

$$\hat{N} = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}) = \frac{n}{b(\underset{\sim}{J})}. \tag{2.7}$$

3. Estimering av et gjennomsnitt i populasjonen

I avsnitt 2 har vi angitt at $\hat{a}$ og $\hat{N}$ er forventningsrette estimatorer for henholdsvis a og N. En naturlig estimator for gjennomsnittet i populasjonen er da

$$\hat{\bar{a}} = \frac{\hat{a}}{\hat{N}}. \tag{3.1}$$

Vi skal nå finne tilnærmingsformler for forventning og varians til denne estimatoren. Ved å bruke metoden gitt i Sverdrup (1964), side 143-144, får vi at

$$E\hat{\bar{a}} \approx \frac{a}{N} = \bar{a}, \tag{3.2}$$

og

$$\text{var } \hat{\bar{a}} \approx \frac{1}{N^2} \{\text{var } \hat{a} + \bar{a}^2 \text{ var } \hat{N} - 2\bar{a} \text{ cov } (\hat{a}, \hat{N})\}. \qquad (3.3)$$

Ved å bruke et resonnement tilsvarende det en kan bruke til å vise at $\hat{\bar{a}}$ er tilnærmet forventningsrett, kan vi vise at en tilnærmet forventningsrett estimator for var $\hat{\bar{a}}$ er gitt ved

$$\text{est var } \hat{a} = \frac{1}{\hat{N}^2} \{\text{est var } \hat{a} + \hat{\bar{a}}^2 \text{ est var } \hat{N} - 2\,\hat{\bar{a}} \text{ est cov } (\hat{a},\hat{N})\}. \qquad (3.4)$$

En betingelse for at tilnærmelsene i (3.2) og (3.3) skal være gode, er at var $\hat{a}$ og var $\hat{N}$ er små.

Det gjenstår nå å finne estimatorer for var $\hat{a}$, var $\hat{N}$ og cov $(\hat{a},\hat{N})$. Hoem (1973) har vist at

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

I Byråets utvalgsplan er ofte $m_i = 2$ og

$$\frac{m_i}{M_i} \approx 17.$$

Hvis vi benytter tilnærmelsen

$$1 - \frac{m_i}{M_i} \approx 1, \qquad (3.5)$$

øker vi i så fall faktoren med ca. 6%. Fra empiriske undersøkelser vet vi at variansen mellom utvalgsområdene utgjør høyst 20% av den totale variansen. Ved å sammenlikne feilen vi gjør ved å innføre tilnærmelsen (3.5), med formelen for var $\hat{a}$, finner vi at vi gjør en feil som er mindre enn 2%.

La oss derfor bruke (3.5). Hensikten med å innføre denne tilnærmelsen er at vi ønsker å forenkle problemene som blir behandlet i Hoem (1973), side 13.

Når vi anvender (3.5), blir

$$\text{var } \hat{a} \approx \sum_i \frac{M_i^2}{m_i} \{\gamma_i^2 + \sigma_i^2\}.$$

Hoem (1973) har vist at

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

har forventning

$$EU_i^2 = \gamma_i^2 + \sigma_i^2.$$

Da er altså en tilnærmet forventningsrett estimator for variansen gitt ved

$$\begin{aligned}\text{est var } \hat{a} &= \sum_i \frac{M_i^2}{m_i} U_i^2 \\ &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \sum_r (V_{ir} - \bar{V}_i)^2 \\ &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \frac{1}{b_i(\underset{\sim}{J})^2} \sum_r \left[\sum_s X_{irs} - \frac{1}{m_i} \sum_\nu \sum_\mu X_{i\nu\mu}\right]^2 \end{aligned} \qquad (3.6)$$

dersom $m_i > 1$ for alle i.

Vi studerer så estimatoren

$$\hat{N} = \sum_i \frac{M_i}{m_i} \sum_r N_i(J_{ir}).$$

Denne er av samme form som estimatoren for totalen a. Ved å erstatte X_{irs} med 1 over alt i formlene ovenfor finner vi at en tilnærmet forventningsrett estimator for var $\hat{N}$ er gitt ved

$$\begin{aligned}\text{est var } \hat{N} &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \sum_r \left[N_i(J_{ir}) - \frac{1}{m_i} \sum_s N_i(J_{is})\right]^2 \\ &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \frac{1}{b_i(\underset{\sim}{J})^2} \sum_r \left[n_{ir}(\underset{\sim}{J}) - \frac{1}{m_i} \sum_s n_{is}(\underset{\sim}{J})\right]^2 \end{aligned} \qquad (3.7)$$

dersom $m_i > 1$ for alle i.

Til slutt skal vi angi en estimator for cov $(\hat{a},\hat{N})$. Denne estimatoren bygger også på tilnærmelsen (3.5). Ved en metode analog til den i avsnitt 6.4 hos Hoem (1973) finner vi

$$\begin{aligned}\text{est cov } (\hat{a},\hat{N}) &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \sum_r \left[V_{ir} - \frac{1}{m_i} \sum_s V_{is}\right]\left[N_i(J_{ir}) - \frac{1}{m_i} \sum_s N_i(J_{is})\right] \qquad (3.8) \\ &= \sum_i \frac{M_i^2}{m_i} \frac{1}{m_i-1} \frac{1}{b_i(\underset{\sim}{J})^2} \sum_r \left[\sum_s X_{irs} - \frac{1}{m_i} \sum_\nu \sum_\mu X_{i\nu\mu}(\underset{\sim}{J}) - \frac{1}{m_i} \sum_\nu n_{i\nu}(J)\right]. \end{aligned}$$

dersom $m_i > 1$ for alle i.

Vi har nå funnet tilnærmet forventningsrette estimatorer for var $\hat{a}$, var $\hat{N}$ og cov $(\hat{a},\hat{N})$. Vi setter disse inn i (3.4) og får at

$$\begin{aligned}\text{est var } \hat{\bar{a}} = \frac{1}{\{\sum_j \sum_r n_{jr}(\underset{\sim}{J})\}^2} \sum_i \frac{m_i}{m_i-1} \sum_r \{&\left[\sum_s X_{irs} - \frac{1}{m_i} \sum_\nu \sum_\mu X_{i\nu\mu}\right] \\ &- \hat{\bar{a}} \left[n_{ir}(\underset{\sim}{J}) - \frac{1}{m_i} \sum_s n_{is}(\underset{\sim}{J})\right]\}^2. \end{aligned} \qquad (3.9)$$

Dette er altså en tilnærmet forventningsrett estimator for var $\hat{\bar{a}}$.

4. Særtilfellet Oslo

Oslo har N_o trekkeenheter. (N_o antas å være kjent.) Av disse trekkes et utvalg på

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

Vi lar $a_o(k)$ være verdien på den variable vi måler for trekkeenhet k. En forventningsrett estimator for totalen a_o er da

$$\hat{a}_o = N_o\bar{X}_o.$$

Laake (1974) har vist at en forventningsrett estimator for var $\hat{a}_o$ er gitt ved

$$\text{est var } \hat{a}_o = N_o\left(\frac{\hat{N}}{n} - 1\right)\frac{1}{n_o - 1}\sum_s (X_{os} - \bar{X}_o)^2$$

$$= \frac{n_o(\underset{\sim}{J})}{n_o(\underset{\sim}{J})-1}\,\frac{1 - b(\underset{\sim}{J})}{b(\underset{\sim}{J})^2}\sum_s (X_{os} - \bar{X}_o)^2. \qquad (4.1)$$

5. Estimering av gjennomsnitt for hele landet

Ved å bruke estimatorene $\hat{a}$ og $\hat{a}_o$ får vi at estimert totalverdi for hele landet er

$$\hat{\underset{\sim}{a}} = \hat{a} + \hat{a}_o,$$

mens estimert antall trekkeenheter er

$$\hat{\underset{\sim}{N}} = \hat{N} + N_o.$$

Da er

$$\hat{\underset{\sim}{\bar{a}}} = \frac{\hat{a} + \hat{a}_o}{\hat{N} + N_o}$$

en estimator for gjennomsnittet i hele populasjonen. Ved å bruke samme resonnement som i avsnitt 2, bare ved å erstatte $\hat{a}$ med $\hat{a} + \hat{a}_o$ og $\hat{N}$ med $\hat{N} + N_o$, får vi at

$$E\hat{\underset{\sim}{a}} \approx \frac{a + a_o}{N + N_o} = \hat{\underset{\sim}{\bar{a}}}.$$

En tilnærmet forventningsrett estimator for var $\hat{\underset{\sim}{\bar{a}}}$ er gitt ved

$$\text{est var } \hat{\bar{a}} = \frac{1}{\{\hat{N} + N_o\}^2} \{\text{est var } (\hat{a} + \hat{a}_o) + \hat{\bar{a}}^2 \text{ est var } (\hat{N} + N_o) - 2\hat{\bar{a}} \text{ est cov } (\hat{N} + N_o, \hat{a} + \hat{a}_o)\}.$$

Siden N_o er ikke-stokastisk, finner vi at

$$\text{est var } \hat{\bar{a}} = \frac{1}{\{\hat{N}+N_o\}^2} \{\text{est var } \hat{a} + \text{est var } \hat{a}_o + \hat{\bar{a}}^2 \text{ est var } \hat{N} - 2\hat{\bar{a}} \text{ est cov } (\hat{a}, \hat{N})\}. \qquad (5.2)$$

6. Trekkeenheter og analyseenheter

I de foregående avsnittene har vi latt trekkeenhetene være grunnlaget for analysen. Som nevnt innledningsvis vil dette ofte være en uheldig innskrenking. I enkelte tilfeller vil man være interessert i mindre analyseenheter. Vi tenker oss da at hver trekkeenhet består av én eller flere analyseenheter. I det tilfellet at hver trekkeenhet har akkurat én analyseenhet, gjelder teorien som er utviklet i avsnittene 2 til 5.

Vi lar $v_i(j,k)$ være antall analyseenheter i trekkeenhet (i,j,k). Vi lar videre $a_i(j,k,l)$ være verdien på det vi målet for analyseenhet l i trekkeenhet (i,j,k). Totalen a estimeres ved

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

$$= \frac{1}{b(\underset{\sim}{J})} \sum_i \sum_r \sum_s X_{irs},$$

der X_{irs}-ene er de totale observerte a-verdier for trekkeenhetene. Vi har da ikke tatt hensyn til at trekkeenhet (i,j,k) består av flere analyseenheter. En estimator for totalt antall analyseenheter er

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

$$= \frac{1}{b(\underset{\sim}{J})} \sum_i \sum_r \sum_s v_i(J_{ir}, K_{irs})$$

Som tidligere er $N_i(J_{ir})$ og $n_{ir}(\underset{\sim}{J})$ antall trekkeenheter henholdsvis i populasjonen og i utvalget.

7. Estimering av gjennomsnitt pr. analyseenhet i populasjonen

Vi skal nå angi en estimator for gjennomsnittet pr. analyseenhet. i populasjonen. Vi lar $\bar{\bar{a}} = a/v$ og

$$\hat{\bar{\bar{a}}} = \frac{\hat{a}}{\hat{v}}.$$

10. Forandring i teorien når antagelsen om at $1 - \frac{m_i}{M_i} \approx 1$ ikke er oppfylt

Dersom tilnærmelsen $1 - \frac{m_i}{M_i} \approx 1$ ikke er god nok, kan vi ikke bruke variansuttrykkene som er angitt i avsnittene foran. Vi bruker da teorien for variansestimatorene fra Hoem (1973). La

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

og

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

Sett så

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

La U_i^2 være definert som i avsnitt 3. Da er en eksakt forventningsrett estimator for var $\hat{a}$ gitt ved

$$\text{est var } \hat{a} = \sum_i \frac{M_i}{m_i} G_i^2 + \sum_i M_i \left(\frac{M_i}{m_i} - 1\right) S_i^2.$$

Tilsvarende finner vi en eksakt forventningsrett estimator for var $\hat{v}$. Ved å bruke en teori som den i avsnitt 6.4 i Hoem (1973) finner vi også en eksakt forventningsrett estimator for cov $(\hat{a}, \hat{v})$. Vi har i avsnittene 4 og 8 funnet estimatorer for var $\hat{a}_o$ og var $\hat{v}_o$. Disse uttrykkene setter vi inn i (9.1) og får dermed et generelt uttrykk for estimatoren til variansen til estimatoren for et gjennomsnitt.

De to variansuttrykkene vi har presentert, bør sammenliknes numerisk. Byrået har for tida et variansberegningsprogram under utarbeiding. Det beregner begge disse estimatorene. Ved å bruke dette programmet kan man få vurdert forskjellene mellom de to estimatorene for noen variable i Byråets utvalgsundersøkelser. Vi planlegger å gjennomføre slike beregninger og å utgi resultatene i et senere notat.

11. Forandring i teorien når "total utvalgsbrøk" er fastlagt

Avsnittene 2 til 10 er bygd på antagelsen om at utvalgsstørrelsen er fast. En del annen teori bygger på at man fikserer "total utvalgsbrøk" b og lar

$$b_i = b \frac{M_i}{m_i}$$

Da får vi ved et resonnement tilsvarende det i avsnitt 3 at $\hat{\bar{\bar{a}}}$ er tilnærmet forventningsrett for $\bar{\bar{a}}$, og at en tilnærmet forventningsrett estimator for var $\hat{\bar{\bar{a}}}$ er gitt ved

$$\text{est var } \hat{\bar{\bar{a}}} = \frac{1}{\hat{b}^2}\{\text{est var } \hat{a} + \hat{\bar{\bar{a}}}^2 \text{ est var } \hat{v} - 2\hat{\bar{\bar{a}}} \text{ est cov } (\hat{a}, \hat{v})\}, \quad (7.1)$$

der est var $\hat{a}$ er gitt ved (3.6), mens

$$\text{est var } \hat{v} = \frac{1}{b(\underset{\sim}{J})^2} \sum_i \frac{m_i}{m_i - 1} \sum_r \{\sum_s v_i(J_{ir}, K_{irs}) - \frac{1}{m_i} \sum_\nu \sum_\mu v_i(J_{i\nu}, K_{i\nu\mu})\}^2, \quad (7.2)$$

$$\text{est cov } (\hat{a},\hat{v}) = \frac{1}{b(\underset{\sim}{J})^2} \sum_i \frac{m_i}{m_i - 1} \sum_r \{\sum_s X_{irs} - \frac{1}{m_i} \sum_\nu \sum_\mu X_{i\nu\mu}\}^2 X$$

$$\{\sum_s v_i(J_{ir}, K_{irs}) - \frac{1}{m_i} \sum_\nu \sum_\mu v_i(J_{i\nu}, K_{i\nu\mu})\} \quad (7.3)$$

dersom $m_i > 1$ for alle i.

8. Særtilfellet Oslo

Vi lar v_{os} være antall analyseenheter i uttrukken trekkeenhet s i Oslo og lar X_{osr} være observert verdi av en variabel på analyseenhet r i denne trekkeenheten. Vi trekker

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

trekkeenheter rent lotterisk. En estimator for totalen er fortsatt

$$\hat{a}_o = N_o\bar{X}_o$$

$$= \frac{1}{b(\underset{\sim}{J})} \sum_s X_{os},$$

der

$$X_{os} = \sum_r X_{Osr}.$$

En estimator for totalt antall analyseenheter i Oslo er tilsvarende

$$\hat{v}_o = \frac{1}{b(\underset{\sim}{J})} \sum_s v_{os}.$$

Da er forventningsrette estimatorer for var $\hat{a}_o$ og var $\hat{v}_o$ gitt ved formel (4.1) og

$$\text{est var } \hat{v}_o = \frac{1-b(\underset{\sim}{J})}{b(\underset{\sim}{J})^2} n_o(\underset{\sim}{J}) \frac{1}{n_o(\underset{\sim}{J})-1} \sum_s (v_{os} - \hat{\bar{v}}_o)^2, \quad (8.1)$$

der $\hat{\bar{v}}_o = \hat{v}_o/N_o$. Siden

$$\text{var}\,(\hat{a}_o + \hat{v}_o) = \text{var}\,\hat{a}_o + \text{var}\,\hat{v}_o + 2\,\text{cov}\,(\hat{a}_o, \hat{v}_o),$$

finner vi at en forventningsrett estimator for kovariansen er

$$\text{est cov}\,(\hat{a}_o, \hat{v}_o) = \frac{1-b(\underset{\sim}{J})}{b(\underset{\sim}{J})^2}\, n_o(\underset{\sim}{J})\, \frac{1}{n_o(\underset{\sim}{J})-1} \sum_s (v_{os} - \hat{\bar{v}}_o)(X_{os} - \bar{X}_o). \qquad (8.2)$$

9. Estimering av et gjennomsnitt for hele landet

Vi skal til slutt finne en estimator for gjennomsnitt pr. analyseenhet for hele landet. Vi betrakter da estimatoren

$$\hat{\underset{\sim}{\bar{a}}} = \frac{\hat{a} + \hat{a}_o}{\hat{v} + \hat{v}_o}.$$

Igjen finner vi at $\hat{\underset{\sim}{\bar{a}}}$ er tilnærmet forventningsrett for gjennomsnittet, og et tilnærmet uttrykk for variansen er gitt ved

$$\text{var}\,\hat{\underset{\sim}{\bar{a}}} \approx \frac{1}{(v+v_o)^2}\,\{\text{var}\,(\hat{a} + \hat{a}_o) + \hat{\underset{\sim}{\bar{a}}}^2\,\text{var}\,(\hat{v} + \hat{v}_o) - 2\hat{\underset{\sim}{\bar{a}}}\,\text{cov}\,(\hat{a} + \hat{a}_o, \hat{v} + \hat{v}_o)\}.$$

En tilnærmet forventningsrett estimator for var $\hat{\underset{\sim}{\bar{a}}}$ er gitt ved

$$\text{est var}\,\hat{\underset{\sim}{\bar{a}}} = \frac{1}{(\hat{v}+\hat{v}_o)^2}\,\{\text{est var}\,\hat{a} + \text{est var}\,\hat{a}_o + \hat{\underset{\sim}{\bar{a}}}^2\,\text{est var}\,\hat{v} + \hat{\underset{\sim}{\bar{a}}}\,\text{est var}\,\hat{v}_o - 2\hat{\underset{\sim}{\bar{a}}}\,\text{est cov}\,(\hat{a}, \hat{v}) - 2\hat{\underset{\sim}{\bar{a}}}\,\text{est cov}\,(\hat{a}_o, \hat{v}_o)\,, \qquad (9.1)$$

der est var $\hat{a}$, est var $\hat{a}_o$, est var $\hat{v}$, est var $\hat{v}_o$, est cov $(\hat{a}, \hat{v})$ og est cov $(\hat{a}_o, \hat{v}_o)$ er gitt ved (3.6), (4.1), (7.2), (8.1), (7.3) og (8.2).

Dersom vi antar at hver trekkeenhet har akkurat én analyseenhet, får vi at

$$\hat{v}_o = N_o$$

og

$$\hat{v} = \hat{N},$$

der N_o er ikke-stokastisk. Da vil (9.1) bli forenklet til (5.2).

Estimatorene for gjennomsnitt og variansformlene i avsnittene 2 til 5 er altså et spesialtilfelle av uttrykkene i avsnittene 6 til 9.

og

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

(Se f.eks. Tamsfoss, 1970, og Hoem, 1973, kapittel 12.) Da blir

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

en stokastisk variabel. Dette fører til visse forandringer i teorien i avsnittene foran. Variansformlene blir imidlertid som før, bortsett fra at vi erstatter $b(\underset{\sim}{J})$ med b, $n_o(\underset{\sim}{J})$ med n_o og $n_{ir}(\underset{\sim}{J})$ med $n_{ir}(J_{ir})$.

Referanser:

[1] Hoem, Jan M. (1973): "Statistisk Sentralbyrås utvalgsundersøkelser. Elementer av det matematiske grunnlaget". SSB-artikkel nr. 58.

[2] Laake, Petter (1974): "Estimering av variansen til estimatoren for populasjonsverdien a_o for Oslo i Byråets intervjuundersøkelser". Side 3 - 4 i dette Metodehefte.

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

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

EN TOMMELFINGER-REGEL FOR HVOR FINT EN KAN GRUPPERE I UTVALGSUNDERSØKELSER

av

Ronny Klæboe

Motivering

I publikasjonene Kommunevalget 1971, hefte II og Folkeavstemningen om EF, hefte II har en valgt å benytte parenteser rundt enkelte prosenttall. Parentesene varsler prosenttall som det knytter seg store relative standardavvik til. F.eks. er et standardavvik på 2 prosent lite i forhold til et prosenttall på 90; men stort i forhold til et prosenttall på 3.

Forekommer flere parenteser i tilknytning til samme prosentfordeling, er en i en situasjon der de observerte forskjeller og tendenser i prosenttallene ofte kan skyldes tilfeldighetene. Dette er en uønsket situasjon, og en vil søke å gruppere grovere.

Regelen

Nedenstående tabell er produsert ved hjelp av en beregning som viser den tilnærmet øvre grensen for antall grupper en bør ha ved forskjellige linjesummer. Kriteriet som er benyttet, er at en ikke vil ha parenteser rundt alle prosenttallene når like mange personer faller i hver av gruppene.

Tabell 1

Linjesummen	=	36	72	108	144	180	216	252	288	324	360	396	432	468	504
Antall grupper	≤	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Regelen kan nå formuleres:

Anslå hvor stor linjesummen vil bli og velg antall grupper mindre enn eller lik tallet som står oppført i tabellen under denne linjesummen.

Eksempelvis bør en benytte 3 aldersklasser eller færre når en forventer en linjesum på ca. 90 personer.

Ovenstående tabell kan selvsagt brukes uavhengig om en nytter parenteser eller ikke, men er beregnet på grunnlag av en slik praksis. Praktisk erfaring med regelen viser at den gir gode resultater i planleggingsfasen.

Tabellen og regelen er basert på et notat av Stein Østerlund Petersen (1972).

Referanse

Petersen, Stein Østerlund (1972): "Utvalgsundersøkelser. Tallet på sparte, estimerte prosenttall og nøyaktighetsgrad." Side 83-85 i Statistisk Sentralbyrå (1972): "Kommunevalget 1971. Hefte II." NOS A 503. Vedlegg 2.