Arbeidsnotater

STATISTISK SENTRALBYRÅ

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

IO
) 76/1

13. januar 1976

VARIANSER OG DESIGNEFFEKTER FOR SYSSELSETTINGSTALL ESTIMERT VED BRUK AV BYRÅETS NYE UTVALGSPLAN

av

Hans Viggo Sæbø

INNHOLD

1. Innledning

Dette notatet er det første i en serie hvor en ønsker å studere forskjellige sider ved den nye utvalgsplanen. Her skal vi se på variansene til noen viktige sysselsettingstall, og sammenlikne dem med tilsvarende varianser i tidligere arbeidskraftundersøkelser, hvor den gamle utvalgsplan er benyttet.

I Byråets nye utvalgsplan er landet delt opp i 102 strata. Tettsteder med over 30 000 innbyggere utgjør egne strata. I alt er det 24 slike. De øvrige er satt sammen av kommuner, slik at de er mest mulig homogene med hensyn på geografisk beliggenhet, sentralitet og næringsstruktur. For en grundigere beskrivelse av utvalgsplanen henvises til Thomsen og Rideng (1974).

Trekking av utvalg foregår i to trinn. Først trekkes et område (primær utvalgsenhet) fra hvert stratum med sannsynlighet proporsjonal med folketallet. En primær utvalgsenhet består av en kommune eller i noen tilfeller av to eller flere mindre kommuner. Tettsteder med over 30 000 innbyggere trekkes altså med sannsynlighet lik 1. De uttrukne områder holdes fast ved utvalgsundersøkelser i Byrået. Disse områdene utgjør et basisutvalg. Annet trinns trekking består i å trekke familier eller personer direkte herfra. Antallet intervjuenheter som trekkes fra hvert område i basisutvalget, er proporsjonalt med folketallet i vedkommende stratum. Hver intervjuenhet får samme "oppblåsningsfaktor", og det endelige utvalg blir selvveiende.

Beregning av varianser er viktig når en skal vurdere en utvalgsplan. Ved en to-trinns trekkemetode blir ikke variansene ved undersøkelser lik variansene en hadde fått ved rent lotterisk trekking fra hele befolkningen. Trekkingen av kommuner i første trinn vil øke variansen i forhold til en slik metode, mens stratifiseringen vil redusere den igjen. (Se f.eks. Cochran (1963), kapittel 5 og 11). Det har vært særlig mye diskusjon om varianser og standardavvik i forbindelse med arbeidskraftundersøkelsene (AKU). De viktigste variablene her er sysselsettingstallene totalt og for de enkelte næringer. Folke- og Boligtellingen 1970 (FoB-70) gir sysselsettingstall for alle kommuner. Dataene herfra gir grunnlag for beregning av eksakte varianser og standardavvik for de aktuelle estimatorene. Det gjelder også beregning av den såkalte designeffekten, som for hver variabel gir et mål for utvalgsplanens "effekt". (Se avsnitt 3A).

Formålet med dette notatet er å presentere resultatene fra en slik beregning av varianser, standardavvik og designeffekter. En har

vurdert virkningen av trekking av primære utvalgsområder med hensyn på de forskjellige komponentene variansene består av.

Dagsvik (1974) har estimert variansene i AKU ved bruk av den forrige utvalgsplanen. Det er først og fremst naturlig å sammenlikne resultatene med disse estimatene. Beregninger av eksakte varianser og designeffekter er ellers gjort av Dahmstrøm og Tillgren (1972) for en tilsvarende utvalgsplan i Sverige. En sammenlikning med disse resultatene har derfor også interesse for oss.

2. Beregning av eksakte varianser og standardavvik

2.A. Utledning av variansformlene

Generell utledning av variansformler ved trekking av utvalg i to trinn er gjort av Laake (1974). Notasjonen i det følgende er lik notasjonen her.

Vi antar at j-te kommune i i-te stratum har $N_i(j)$ trekkenheter. Den k-te trekkenheten har verdien $a_i(j,k)$, og vi antar at denne bare kan anta verdien 0 eller 1.

Vi lar

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

$$N = \sum_i N_i,$$

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

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

og

$$a = \sum_i a_i.$$

Gjennomsnittsverdien innen hvert område skrives

$$p_i(j) = a_i(j)/N_i(j).$$

Tilsvarende settes

$$p_i = a_i/N_i$$

og

$$p = a/N.$$

I hvert stratum trekker vi først et utvalgsområde. La $\Pi_i(j)$ være sannsynligheten for at område j i stratum i blir trukket ut, og la $\Pi_i(j,k)$ være sannsynligheten for at både utvalgsområde j og k i stratum i skal bli trukket ut. I den nye utvalgsplanen er

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

og

$$\Pi_i(j,j) = \Pi_i(j) = N_i(j)/N_i.$$

La

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

La J_1, J_2, ... være numrene på de områdene som trekkes ut la $\underset{\sim}{J} = (J_1, J_2, \ldots)$ være vektoren som består av numrene på alle uttrukne områder. La videre utvalget fra område (i,j) bestå av $n_{ij}(\underset{\sim}{J})$ personer. Vi definerer $n_{ij}(\underset{\sim}{J}) = 0$ for $I_{ij} = 0$. Dersom $\bar{X}_{ij}$ er gjennomsnittet for de $n_{ij}(\underset{\sim}{J})$ uttrukne enhetene i område (i,j) blir etter Laake (1974) estimatoren $\hat{a}$ for totalverdien a gitt ved

$$\hat{a} = \sum_{ij}\sum \{I_{ij} N_i(j) \bar{X}_{ij}/\Pi_i(j)\} = \sum_i N_i \sum_j I_{ij}\bar{X}_{ij}.$$

Vi skal i det følgende komme fram til et enkelt uttrykk for variansen til denne estimatoren.

Når $a_i(j,k)$ er en binær variabel, kan variansen innen et område (i,j) skrives som

$$\sigma_i^2(j) = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k)-a_i(j)\}^2 = \frac{N_i(j)}{N_i(j)-1} p_i(j)\,(1-p_i(j)).$$

For $I_{ij} = 1$ defineres

$$\tau^2_{ij}(\underset{\sim}{J}) = \frac{\sigma^2_i(j)}{n_{ij}(\underset{\sim}{J})} \frac{N_i(j)-n_{ij}(\underset{\sim}{J})}{N_i(j)},$$

og

$$\eta_i(j) = E\{N^2_i(j)\, \tau^2_{ij}(\underset{\sim}{J}) \mid I_{ij} = 1\}.$$

Laake (1974) har angitt variansen til $\hat{a}$ iet generelt tilfelle ved trekking i to trinn som

$$\text{var}\,\hat{a} = \sum_i \Big[\sum\sum_{jk} \frac{\Pi_i(j,k)-\Pi_i(j)\Pi_i(k)}{\Pi_i(j)\Pi_i(k)}\, a_i(j)\, a_i(k) + \sum_j \{\eta_i(j)/\Pi_i(j)\}\Big]. \qquad (2.A.1)$$

Vi oppnår at utvalget blir selvveiende ved å velge

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

Innsetting for $\Pi_i(j)$ og $n_{ij}(\underset{\sim}{J})$ i (2.A.1) gir

$$\text{var}\,\hat{a} = \frac{N}{n} \sum\sum_{ij} N_i(j)\, \sigma^2_i(j) \left(1- \frac{N_i \cdot n}{N \cdot N_i(j)}\right) + \sum_i N_i \sum_j N_i(j)\, (p_i(j) - p_i)^2.$$

I utvalgsplanen vil en typisk verdi for $N_i(j)$ være 5 000, mens N_i er ca. 25 000. For et utvalg på 12 000 personer vil $n/N \approx 0{,}004$. Insatt disse verdiene blir

$$\frac{N_i \cdot n}{N \cdot N_i(j)} \approx 0{,}02.$$

I gjennomsnitt vil denne størrelsen bli noe mindre p.g.a. tettstedene som utgjør egne strata. Vi setter

$$1 - \frac{N_i \cdot n}{N \cdot N_i(j)} = 1,$$

og

$$\sigma^2_i(j) = p_i(j)\,(1-p_i(j)).$$

Dette medfører at

$$\text{var}\,\hat{a} = \text{var}_1\hat{a} + \text{var}_2\hat{a}, \text{ hvor}$$

$$\text{var}_1\,\hat{a} = \frac{N}{n}\sum_{ij} N_i(j)\, p_i(j)\,(1-p_i(j)) \qquad (2.A.2)$$

og

$$\text{var}_2\,\hat{a} = \sum_i N_i \sum_j N_i(j)\,(p_i(j) - p_i)^2. \qquad (2.A.3)$$

Den totale varians kan altså deles opp i to komponenter, der $\text{var}_1\hat{a}$ defineres som variansen innen utvalgsområdene, mens $\text{var}_2\hat{a}$ er variansen mellom utvalgsområdene.

Det er uttrykkene (2.A.2) og (2.A.3) som ligger til grunn for beregningene av varianser. For kjent utvalgsstørrelse n trenger vi i alle kommuner folketallet $N_i(j)$ og sysselsettingstallet $a_i(j) = N_i(j)p_i(j)$ for en næringsgren for å kunne beregne variansen for dette sysselsettingstallet eksakt. Hefte II fra Folke- og Boligtellingen 1970 gir tall på kommunenivå for total sysselsetting og sysselsettingen i de enkelte næringer samt folketall. Disse tall ligger til grunn for beregningene i dette notatet.

I tabell 1 har en valgt å publisere standardavviket $s(\hat{a})$ istedenfor den totale variansen. Dette gis ved

$$s(\hat{a}) = \sqrt{\text{var}\,\hat{a}}.$$

2.B. Standardavvik i et AKU-utvalg

Definisjonene på næringsgruppene i FoB-70 er ikke de samme som de som brukes i AKU. Nivåtallene i tabellene kan derfor ikke sammenliknes direkte med AKU-tall, men en kan anta at effekten av utvalgsplanen på samme type variable i FoB-70 og AKU er den samme. Varianser, standardavvik og designeffekter er derfor sammenliknbare forutsatt at de respektive nivåtall er av noenlunde samme størrelse.

Som sysselsatte har en valgt å regne de som i FoB-70 hadde oppgitt arbeid som viktigste kilde til livsopphold. Disse er så blitt fordelt på de enkelte næringsgrupper. Dette fører til at nivåtallet for de fleste næringer vil være lavere enn i AKU, hvor en har et mer omfattende sysselsettingsbegrep.

I AKU trekkes husholdninger, mens intervjuenheten er person. Variansene er beregnet som om person skulle ha vært trekkenhet. Disse antas å avvike lite fra variansene ved trekking av familier. En har ved

beregningene heller ikke tatt hensyn til etterstratifiseringen som i AKU foregår etter alder og kjønn.

Tabell 1. Nivåtall med standardavvik for antall sysselsatte etter næring beregnet for et utvalg fra FoB-70 med størrelse 12 000 personer 16 år og over

Næring	Antall i 1 000	Standardavvik i 1 000
Jordbruk, skogbruk	143	7,3
Fiske, fangst	27	3,8
Industri m.v.[1]	416	10,7
Bygg, anlegg	129	5,8
Varehandel	199	6,8
Samferdsel	157	6,2
Tjenester m.v.[2]	389	9,7
Uoppgitt	2	-
Total sysselsetting[3]	1 462	13,7

1) Industri, bergverk, kraft- og vannforsyning. 2) Off. og privat tjenesteyting, administrasjon, forsvar, eiendomsdrift og finansvirksomhet. 3) Personer med inntekt av eget arbeid som viktigste kilde til livsopphold.

Dagsvik (1974) har presentert tabeller over estimerte standardavvik for AKU 1973 og 74. I den utvalgsplanen som er brukt i arbeidet til Dagsvik, trakk en seks utvalgsområder med lik sannsynlighet i hvert stratum i første trinn. I den nye utvalgsplanen har vi flere strata, og vi trekker ett utvalgsområde fra hvert stratum. Da vi også trekker kommunene med sannsynlighet proporsjonal med folketallet, kunne en totalt vente en nedgang i standardavvik (Cochran (1965), kapittel 11.2). Dette bekreftes av tallene i tabellen. Selv om en tar hensyn til de lavere nivåtallene her enn i AKU, ligger standardavvikene ca. 30 prosent lavere for de næringsgruppene som kan sammenliknes. Det eneste unntaket er fiske og fangst hvor standardavviket er det samme.

2.C. Varianser innen og mellom utvalgsområdene

Tabell 2 viser hvordan variansen fordeler seg innen og mellom kommunene (se formel 2.A.2 og 2.A.3).

Tabell 2. Relativ andel av varianskomponentene innen og mellom utvalgsområdene beregnet for et utvalg fra FoB-70 med størrelse 12 000 personer 16 år og over

Næring	Varianskomponentene i %	
	$\widehat{var}_1 a$ (innen)	$\widehat{var}_2 a$ (mellom)
Jordbruk, skogbruk	57	43
Fiske, fangst	43	57
Industri m.v.	72	28
Bygg, anlegg	87	13
Varehandel	94	6
Samferdsel	91	9
Tjenester m.v.	83	17
Total sysselsetting	91	9

En ser av (2.A.3) at variansen mellom kommunene er uavhengig av utvalgsstørrelsen n. Denne komponenten gir uttrykk for hvor forskjellige kommunene i hvert stratum er m.h.p. næringsfordelingen. I et homogent stratum vil den relative andelen sysselsatt i en næring være nesten den samme i alle kommuner innen vedkommende stratum, det vil si $p_i(j) \approx p_i$ for alle j. Det gir $\widehat{var}_2 a \approx 0$ for denne næringsgren, og trekkingen av en kommune fra dette stratum i første trinn gir ingen ekstra varians. De kommuner som utgjør egne strata bidrar ikke til $\widehat{var}_2 a$. Variansen mellom kommunene blir derfor liten for næringer som er sterkest utbredt i byene samt for homogent fordelte næringer. I tabell 2 sees dette ved at varehandel, samferdsel og total sysselsetting har lave relative verdier for denne varianskomponenten. At industri som er sterkt utbredt i byene har en noe større relativ verdi for $\widehat{var}_2 a$ skyldes at denne næringsgren har mer inhomogen fordeling mellom kommunene.

3. Beregning av designeffekter

3.A. Definisjon av designeffekt

Den enkleste (men i praksis dyreste) utvalgsplan består i å trekke hele utvalget som skal være med i en undersøkelse rent lotterisk fra hele befolkningen. Estimatoren $\hat{a}_o$ for en totalverdi vil da være $\hat{a}_o = N\bar{X}$ hvor $\bar{X}$ er middelverdien i utvalget. Ved å definere $p = a_o/N = a/N$ finner vi at variansen til $\hat{a}_o$ blir

$$\text{var } \hat{a}_o = \frac{N^2}{n} p(1-p). \qquad (3.A.1)$$

Designeffekten d defineres som forholdet mellom den variansen vi har ved bruk av utvalgsplanen vår og variansen vi ville ha hatt dersom vi hadde trukket utvalg på denne måten (Kish (1965), s. 258):

$$d = \frac{\text{var } \hat{a}}{\text{var } \hat{a}_o} = \frac{\text{var}_1 \hat{a}}{\text{var } \hat{a}_o} + \frac{\text{var}_2 \hat{a}}{\text{var } \hat{a}_o} . \qquad (3.A.2)$$

Designeffekten gir altså et mål for øking i varians ved en to-trinns trekkemetode i forhold til enkel tilfeldig trekking.

Innsetting for var $\hat{a}$ og var $\hat{a}_o$ gir

$$d = \frac{\sum_{ij} N_i(j)\, p_i(j)\, (1-p_i(j))}{Np(1-p)} + n \frac{\sum_i N_i \sum_j N_i(j)\, (p_i(j)-p_i)^2}{N^2 p(1-p)} . \qquad (3.A.3)$$

Uttrykket avhenger lineært av utvalgsstørrelsen n, og gir grunnlag for beregning av d for forskjellige verdier på n i tabell 3. Nå kan vi skrive

$$Np(1-p) = \sum_{ij} N_i(j)\, p_i(j)\, (1-p_i(j)) + \sum_{ij} N_i(j)\, (p_i(j) - p)^2 .$$

Av dette sees at første ledd i (3.A.3) alltid vil være mindre enn 1, og for små verdier av n vil vi kunne få $d < 1$. Dette kan observeres for små utvalgsstørrelser i tabell 3.

Vi så i avsnitt 2.C at det relative bidraget til variansen fra variansen mellom utvalgsområdene, $\text{var}_2 \hat{a}$, kunne betraktes som et mål for hvor homogene strataene våre er m.h.p. et sysselsettingstall. Siste ledd i (3.A.3) består av forholdet mellom $\text{var}_2 \hat{a}$ og var $\hat{a}_o$. Dette leddet kan for fast n også betraktes som et slikt mål. Dersom $\text{var}_1 \hat{a} \approx \text{var } \hat{a}_o$ gjelder dette hele designeffekten idet første ledd da er tilnærmet lik 1. Beregningene viser at $\text{var}_1 \hat{a}/\text{var } \hat{a}_o \approx 1$ for alle sysselsettingstallene.

Typiske verdier er 0,97 - 0,99, mens jordbruk gir lavest verdi på 0,94. Dette betyr at designeffekten for sysselsettingstall ved bruk av utvalgsplanen vår for fast utvalgsstørrelse først og fremst avhenger av hvor like kommunene innen hvert stratum er med hensyn på næringsstruktur.

3.B. Designeffekt for forskjellige utvalgsstørrelser

I tabell 3 har en beregnet designeffekten for sysselsettingstall ved utvalgsstørrelser fra 1 000 til 12 000. Designeffektene for et utvalg på 12 000 kan sammenliknes med designeffekten i AKU. Dagsvik (1974) har estimert designeffektene i AKU 1973 og 74, hvor utvalgene ble trukket etter den gamle utvalgsplanen. Designeffektene i tabell 3 ligger under disse estimatene for alle næringer unntatt fiske, hvor verdien er den samme.

I 3.A vises hvordan designeffekten for en fast utvalgsstørrelse kan betraktes som et mål på hvor homogene strataene våre er med hensyn på de enkelte sysselsettingstall. Samme resonnement som i 2.B gir da næringer med homogen fordeling og/eller stor uttredelse i tettbygde strøk lav designeffekt. Dette gjelder som vi kan se av tabellen særlig varehandel og samferdsel, mens jordbruk og fiske har høy designeffekt. Særlig har jordbruket lav utbredelse i byene. Nedgangen fra designeffekt på 2,5 i den gamle utvalgsplanen er likevel klar.

Det har vært vanlig å regne med en designeffekt på 1,5 i Byråets utvalgsundersøkelser. Med denne antagelsen ville vi i dette eksperimentet overestimert variansen i alle hovednæringer unntatt jordbruk, skogbruk, fiske og fangst.

Hvis vi betrakter designeffekten for sysselsettingstallene for mindre utvalg enn det som brukes i AKU, viser det seg at d er ubetydelig forskjellig fra 1,0 for alle næringer unntatt fiske for $n \leq 2\,000$. I Sverige har en også en utvalgsplan som er bygd opp på samme måte som vår. Denne beskrives bl.a. av Dahmstrøm (1972). Designeffekter fra tilsvarende beregninger her av Dahmstrøm og Tillgren (1972) er for sammenliknbare næringer angitt i parenteser i tabell 3. Disse ligger i de fleste tilfelle nær våre resultater. Som i vår utvalgsplan har jordbruk størst designeffekt og varehandelen minst (ubetydelig fiske i Sverige).

Som antydet i 3.A kan designeffekten bli mindre enn 1 for små utvalg. For $n = 1\,000$ kan dette i vår utvalgsplan bare registreres for varehandel og tjenester.

Tabell 3. Designeffekt for forskjellige utvalgsstørrelser. Tall for svensk utvalgsplan i parentes

Næring	Utvalgsstørrelse									
	1 000		2 000		5 000		10 000		12 000
Jordbruk, skog-bruk	1,01	(0,97)	1,07	(1,04)	1,25	(1,24)	1,54	(1,58)	1,64
Fiske, fangst.	1,08	-	1,19	-	1,50	-	2,03	-	2,24
Industri m.v.	1,01	(1,00)	1,04	(1,04)	1,13	(1,14)	1,29	(1,32)	1,35
Bygg, anlegg.	1,00	(1,02)	1,01	(1,05)	1,05	(1,12)	1,11	(1,25)	1,13
Varehandel ..	0,99	(0,99)	0,99	(1,00)	1,01	(1,03)	1,04	(1,07)	1,05
Samferdsel ..	1,00	(1,01)	1,00	(1,02)	1,03	(1,06)	1,06	(1,13)	1,08
Tjenester m.v.	0,99	(1,00)	1,01	(1,01)	1,06	(1,07)	1,14	(1,17)	1,18
Sysselsetting	1,00	(1,01)	1,01	(1,02)	1,04	(1,07)	1,08	(1,14)	1,10

Referanser

Cochran, W. G. (1963): "Sampling Techniques". J. Wiley & Sons, New York.

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

Dahmstrøm, P. (1972): "Basurval". Statistiska Centralbyrån, SCB metodinformation (Nr. 72-2).

Dahmstrøm, P. og Tillgren, U. (1972): "Variansberäkningar vid SCB's basurval". Statistiska Centralbyrån, Metodmeddelande (Nr. IV:3).

Kish, L. (1965): "Survey Sampling". J. Wiley & Sons, New York.

Laake, P. (1964): "Estimering av totaler med en to-trinns utvalgsplan der de primære utvalgområder trekkes med ulik sannsynlighet i første trinn". Statistisk Sentralbyrå, Arbeidsnotat (IO 74/49).

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