Statistisk sentralbyrås generelle utvalgsplan

Jan F. Bjørnstad

Statistisk sentralbyrå, Seksjon for statistiske metoder og standarder, Postboks 8131 Dep., 0033 Oslo
Telefon: 21 09 47 58 E-post: jab@ssb.no

SAMMENDRAG

SSBs generelle utvalgsplan gjelder besøksundersøkelser, både husholdningsundersøkelser og personundersøkelser. Den opprinnelige versjonen av utvalgsplanen har vært i bruk siden 1975, men har gjennomgått en del justeringer blant annet på grunn av kommuneendringer. I 1995 ble det foretatt en større revidering, blant annet for å kunne gi fylkesbasert statistikk. Utvalgsplanen er en to-trinns utvalgstrekking for å redusere reise- og arbeidstidskostnader med et utvalg av essensielt kommuner på trinn 1 og tilfeldig utvalg av personer innen de uttrukne kommunene på trinn 2. Denne generelle utvalgsplanen ble tidligere brukt også ved telefon- og postale undersøkelser. Nå trekkes de fleste utvalg til slike undersøkelser i ett trinn.

1. INNLEDNING

SSBs generelle utvalgsplan for besøksundersøkelser brukes både for husholdningsundersøkelser og personundersøkelser i Statistisk sentralbyrå, bl.a. de årlige helseundersøkelsene, levekårsundersøkelsene og forbruksundersøkelsene.

Den opprinnelige versjonen av utvalgsplanen har vært i bruk siden 1975, men har gjennomgått en del justeringer blant annet på grunn av kommuneendringer. I 1995 ble det foretatt en større revidering, blant annet for å kunne gi fylkesbasert statistikk. Utvalgsplanen er en totrinnstrekking med et stratifisert utvalg på trinn 1 og enkelt tilfeldig utvalg på trinn 2. Landet er først delt i primære utvalgsenheter som stort sett består av kommuner. Disse kalles *primære utvalgsområder*. De primære utvalgsområdene er deretter stratifisert, dvs. gruppert, etter kommunetype og størrelse. Innen hvert stratum trekkes ett primært utvalgsområde proporsjonalt etter størrelse (antall innbyggere). På 2. trinn trekkes personer tilfeldig innen hvert utvalgte primæromåde. Denne generelle utvalgsplanen ble tidligere brukt også ved telefon- og postale undersøkelser. Nå trekkes de fleste utvalg til slike undersøkelser i ett trinn.

De viktigste forutsetningene ved revideringen av utvalgsplanen i 1995 var:

(i) Utvalgsplanen skal gi grunnlag for undersøkelser av varierende art.
(ii) Antall intervjuere som brukes i forbindelse med en undersøkelse skal være ca. 135, beregnet ut fra årlig arbeidsmengde.
(iii) Det viktigste er å kunne publisere tall for hele landet, men for større undersøkelser bør utvalget være slik at man gi tall for mindre geografiske områder, spesielt fylkestall.
(iv) Utvalget skal være selvveiende hvis det er mulig, dvs. at alle personer i populasjonen av interesse i undersøkelsen skal ha samme sannsynlighet for å bli trukket ut.

2. VALG OG TREKKING AV PRIMÆRE UTVALGSOMRÅDER

2.1. Generelle betraktninger

Økonomiske og praktiske hensyn tilsier at utvalget trekkes i to trinn. Besøksundersøkelser er kostbare å gjennomføre. Det er derfor viktig med to-trinns utvalg fordi en da kan konsentrere intervjukostnadene til avgrensede geografiske områder. På denne måten sparer en både reiseutgifter og arbeidstidsutgifter. Ved konstruksjonen av de primære utvalgsområdene (PU) på trinn 1 var følgende momenter avgjørende:

- Størrelsen på PU samt fordelingen av intervjuobjektene (IO) på intervjuerne er de viktigste valg.
- Ut fra ønsket om minst mulig varians bør en velge mange små PU. Ut fra et økonomisk og praktisk synspunkt er det derimot en fordel å konsentrere intervjuerne så mye som mulig.
- PU må være identifiserbare i de registre som er tilgjengelige.

Ut fra en samlet vurdering ble det bestemt å bruke kommuner som primære utvalgsområder. Små kommuner slås sammen med andre kommuner, slik at det (vanligvis) er minst 3000 personer i hver PU. Dette for å forhindre at IO innen et utvalgt PU blir belastet med å være med i mange forskjellige undersøkelser. Dessuten ble alle PU delt i 3 områder. I mindre undersøkelser trekkes utvalget i tre trinn. Først utvalgsområde, deretter del av utvalgsområdet og til slutt enhetene til undersøkelsen. På trinn 1 trekkes et stratifisert utvalg av to grunner:

(a) redusering av usikkerheten på estimatene (estimeringsvariansen) fra en undersøkelse.
(b) produsering av regional statistikk.

2.2. Stratifiseringen

Punkt (a) betyr at strataene bør være mest mulig homogene. De mest brukte stratifiseringsvariablene er

innbyggerantall og geografisk beliggenhet. Stratifiseringen bestemmer de minste områder en kan gi tall for. En hovedgrunn for revideringen i 1995 var å kunne gi fylkestall. Hvert fylke er derfor stratifisert for seg. Ut fra ønsket om å redusere estimeringsvariansen bør en lage så mange strata som mulig, og trekke ett primært utvalgsområde fra hvert stratum. En ulempe ved denne formen for stratifisert trekking er at det blir vanskelig å estimere variansen til estimatorer for populasjonstotaler som er det vanligste estimeringsproblemet. På tross av det valgte en å lage flest mulig strata og trekke ett PU innen hvert stratum.

De primære utvalgsområdene ble inndelt i 109 strata. Det minste stratum omfatter 10 242 personer og det største 477 781, med gjennomsnittlig stratumstørrelse lik 39 677 (innbyggertallene er pr. 1.1.94). Byer med flere enn 30 000 innbyggere, samt noen få i tillegg, er stratifisert slik at de hver for seg utgjør det eneste PU i stratumet. Disse PU er da trukket ut med sikkerhet. Ellers, fra hvert stratum ble en PU trukket proporsjonalt med størrelse. Det uttrukne PU skal normalt dekkes av én intervjuer. Trekkingen av PU ble foretatt i desember 1994 for en tiårsperiode. Intervjuerne ble ansatt etter desember 1994 til å dekke de uttrukne PU. Siden de fleste intervjuerne har ansvar for kun en PU blir reisekostnadene kraftig redusert i forhold til direkte trekking fra hele populasjonen.

For en fullstendig beskrivelse av de 109 strata og en mer utførlig beskrivelse av utvalgsplanen henvises det til Stålnacke et al. (1999). Som et eksempel på stratifiseringen i et fylke, skal vi se på Østfold som er delt inn i 6 strata (kommuner med «/» mellom er slått sammen til ett PU). Tallene i parentes er innbyggertall pr. 1.1.94 og antall intervjuere til å betjene stratumet.

Stratum 1: Fredrikstad/Hvaler (68 207, 2)
Stratum 2: Sarpsborg (46 381, 2)
Stratum 3: Halden (25 908, 1)
Stratum 4: Moss (25 071, 1)
Stratum 5: Spydeberg, Askim, Råde, Rygge, Våler, Hobøl (43 619, 1)
Stratum 6: Trøgstad, Eidsberg, Skiptvedt, Rakkestad, Aremark/Marker/Rømskog (29 526, 1)

Strataene 1-4 består alle av ett PU, og dermed er disse trukket ut. Fra stratum 5 ble Rygge trukket ut med trekksannsynlighet 12189/43619 = 0,28, og fra stratum 6 ble Eidsberg trukket ut med trekksannsynlighet 9156/29526 = 0,31. Rygge er nest størst (Askim har 12 826 innbyggere) og Eidsberg er størst innenfor sine respektive strata.

3. UTVALGSPLANEN

La N betegne størrelsen på populasjonen av interesse for undersøkelsen, som antas å være kjent, og n størrelsen på det totale utvalget av personer.

Utvalgets størrelse har, selvsagt, en avgjørende innflytelse på kostnadene til en undersøkelse. Bestemmelse av n er nær knyttet sammen med formålet for undersøkelsen. Noen utvalgsundersøkelser klarer seg med færre enn 1000 enheter i utvalget, mens en undersøkelse som Arbeidskraftundersøkelsen krever så mye som 24000 pr. kvartal. Det er hovedsakelig tre forhold å ta hensyn til:

(a) *Ønsket nøyaktighet på resultatene.*
(b) *Homogenitet i populasjonen*. Trenger mindre utvalg hvis det er liten variasjon i populasjonen.
(c) *Oppsplittinger av utvalget for estimering i delpopulasjoner.*

Det er ofte (c) som setter de største kravene. Som en illustrasjon, ser vi litt nærmere på (a). Disse betraktningene kan også brukes på (c) som anslag på størrelsen til de nødvendige delutvalg. Anta at vi skal estimere en populasjonsandel p med rent tilfeldig utvalg. Med liten utvalgsandel n/N så er 95% konfidensintervall for p gitt ved, med $\hat{p}$ lik observert andel i utvalget:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

La oss si vi har bestemt oss for et nøyaktighetskrav på ± 5%:

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}05$$
$$\Rightarrow n = 400 \cdot 4\hat{p}(1-\hat{p}) \leq 400.$$

Estimatet $\hat{p}$ er ukjent i planleggingsfasen. Vi bruker derfor enten det konservative anslaget 400, eller en planleggingsverdi p_0 som gir

$$n = 1600 p_0(1-p_0).$$

Med et generelt nøyaktighetskrav d (istedenfor 0,05):

$$n = \frac{4p_0(1-p_0)}{d^2}$$

Når n er bestemt foregår trekkingen på følgende måte:

Trinn 1: Innenfor hvert stratum trekkes ett primært utvalgsområde proporsjonalt med størrelse. Hvis N_h er det totale innbyggertallet i stratum h, og $N_{h,i}$ er innbyggertallet for PU i i stratum h så trekkes PU i med sannsynlighet $\pi_{1i} = N_{h,i}/N_h$. La s_1 være utvalget av PU. I en tiårsperiode holdes s_1 fast. Det nåværende utvalg av PU ble som nevnt trukket i desember 1994 hvor $N_{h,i}$ og N_h er folketallene pr. 1.1.94.

Trinn 2: For hver PU $i \in s_1$: Trekker et tilfeldig utvalg s_i på n_i personer. Utvalgsstørrelsen n_i bestemmes slik at det totale utvalg på n personer blir selvveiende.

Trekksannsynligheten på trinn 2 for en vilkårlig person k i PU i for stratum h er $\pi_{k|i} = n_i/N_{h,i}$. Dermed has at (den ubetingede) trekksannsynligheten for person k blir

$$\pi_k = \pi_{1i}\pi_{k|i} = n_i/N_h.$$

Selvveiende utvalg betyr at alle personer har samme trekksannsynlighet $\pi_k = n/N$. Det gir at

$$\frac{n_i}{N_h} = \frac{n}{N} \quad \text{og} \quad n_i = \frac{n}{N} N_h \qquad (1)$$

proporsjonal med stratumstørrelsen. Det betyr at enhetene i de mindre utvalgte PU får høyere trekksannsynligheter på 2. trinn.

Kommentar

Trinn 1 er foretatt i desember 1994. For en senere undersøkelse, f.eks. i 2003, så vil innbyggertallene ha endret seg noe. I tillegg er vi vanligvis interessert i den delen av befolkningen som er over 15 år. Dermed blir trekksannsynlighetene på 2. trinn noe annerledes, og bestemmelsen av n_i kan bli endret litt. Som regel vil imidlertid (1) være en god nok tilnærming til å kunne brukes i praksis. For å se det, la

$$N^{03}, N_h^{03}, \ N_{h,i}^{03}$$

være henholdsvis totalen, stratumstørrelsene og PU-størrelsene for befolkningen av interesse i 2003. Trekksannsynlighetene på 2. trinn blir da

$$\pi_{k|i} = n_i / N_{h,i}^{03}.$$

Et selvveiende utvalg betyr dermed at:

$$\frac{N_{h,i}}{N_h} \cdot \frac{n_i}{N_{h,i}^{03}} = \frac{n}{N^{03}}, \quad \text{dvs.}$$

$$n_i = \frac{n}{N^{03}} \cdot N_h \frac{N_{h,i}^{03}}{N_{h,i}} = n \frac{N_h}{N} \cdot \left[\frac{N_{h,i}^{03} / N_{h,i}}{N^{03} / N} \right]. \qquad (2)$$

Vanligvis så has at

$$\frac{N_{h,i}^{03} / N_{h,i}}{N^{03} / N} \approx 1,$$

slik at (1) kan benyttes. F.eks., hvis

$$N_{h,i}^{03} / N_{h,i} = c$$

for alle *i,h* så er for alle *h,*

$$N_h^{03} = \sum_i N_{h,i}^{03} = c \sum_i N_{h,i} = cN_h,$$

og dermed $N^{03} = cN$, dvs.
$N^{03} / N = c$ og $n_i = (n/N)N_h$.

Man kan også bruke

$$n_i = n \frac{N_h^{03}}{N^{03}} \qquad (3)$$

som en tilnærming til (2), siden fra (2)

$$n_i = n \frac{N_h^{03}}{N^{03}} \cdot \left[\frac{N_{h,i}^{03} / N_{h,i}}{N_h^{03} / N_h} \right] \quad \text{og} \quad \frac{N_{h,i}^{03} / N_{h,i}}{N_h^{03} / N_h} \approx 1.$$

Eksempel. La oss si at vi skal foreta en undersøkelse i 2003 hvor vi skal trekke et landsutvalg på 5000 personer over 15 år (16 år og eldre). Pr. 1.1.03 besto denne populasjonen av 3 585 558 personer. Betrakt stratum 6 i Østfold ovenfor. Pr. 1.1.03 hadde stratum 6 i alt 31 027 innbyggere. Av disse var 24 679 over 15 år. Det uttrukne PU, Eidsberg hadde 7917 innbyggere over 15 år pr. 1.1.03. Det totale innbyggertallet i Norge 1.1.94 var 4 324 815. Vi har dermed følgende tall:

2003-undersøkelsen	$N^{03} = 3\,585\,558$	$N_h^{03} = 24\,679$	$N_{h,i}^{03} = 7\,917$
1994-tall	$N = 4\,324\,815$	$N_h = 29\,526$	$N_{h,i} = 9\,156$

Tilnærmingen (1) gir $n_i = 5000 \cdot (29526/4324815) = 34{,}136 = 34$. Forholdet

$$\frac{N_{h,i}^{03} / N_{h,i}}{N^{03} / N} = \frac{0{,}8647}{0{,}8291} = 1{,}043 \cdot$$

Dermed blir eksakt bestemmelse lik:

$$n_i = 34{,}136 \cdot 1{,}043 = 35{,}60 = 36,$$

mens tilnærmingen (3) gir $n_i = 5000 \cdot (24679/3585558) = 34{,}41 = 34$. Begge valg av n_i gir tilnærmet et selvveiet utvalg. Selv om 36 gir en bedre tilnærming er forskjellen så liten at vi like godt kan bruke (1) eller (3) til å bestemme utvalgsstørrelsene innen hvert valgte PU.

I noen tilfeller vil, imidlertid, tilnærmingen (1) bli for grov. Anta at vi en tid etter at de primære utvalgsområdene ble trukket, får en betydelig flytting innen de enkelte strataene fra landlige primære utvalgsområder til de mer tettbygde. Da vil $N_{h,i}$ øke for tettbygde PU. Samtidig er π_{Ii} fast, og hvis da (1) benyttes til å bestemme utvalgsstørrelsene i de valgte PU vil det bety at

$$n_i = \frac{n}{N} N_h$$

er fast for hele stratumet. Dermed vil

$$\pi_k = \pi_{Ii} \cdot \frac{n_i}{N_{h,i}}$$

avta i tettbygde PU og vi har ikke lenger et selvveiende utvalg, $\pi_k < n/N$ for tettbygde PU og $\pi_k > n/N$ for de landlige PU. Estimatoren blir dermed skjev, og den eksakte bestemmelsen av n_i gitt ved (2) bør brukes.

4. ESTIMERING AV POPULASJONSTOTALER

Det vanligste problemet i utvalgsundersøkelser er estimering av populasjonstotalen *t* eller gjennomsnittet *t*/*N* for en variabel *y*,

$$t = \sum_{i=1}^{N} y_i.$$

Her er y_i verdien av y for *i*'te person i populasjonen. Siden vi har et selvveiende utvalg blir totrinnsestimatoren lik den såkalte ekspansjonsestimatoren,

$$\hat{t}_{2T} = N\bar{y}_s$$

hvor $\bar{y}_s$ er gjennomsnittet i det endelige utvalget s av personer. Vi skal sammenligne SSBs utvalgsplan med den tilsvarende selvveiende ett-trinns stratifiserte utvalgsplanen, dvs. hvor det tas et rent tilfeldig utvalg fra hvert av de 109 strata. Med basis i (1), så blir utvalgsstørrelsene i strataene igjen lik

$$n_h = \frac{n}{N} N_h .$$

Den stratifiserte estimatoren $\hat{t}_{st}$ for ett-trinnsplanen blir da også lik ekspansjonsestimatoren, $\hat{t}_{st} = N\bar{y}_s$.

Anta vi ønsker å estimere andelen p i befolkningen med et gitt kjennetegn A, for eksempel andelen med en bestemt sykdom. Andelene er da estimert ved $\hat{t}_{2T} / N$ og $\hat{t}_{st} / N$, dvs. $\bar{y}_s = \hat{p} =$ andelen med kjennetegn A i utvalget. La:

m_h = antall PU i stratum h,
p_h = andel med kjennetegn A i stratum h,
$p_{h,i}$ = andel med kjennetegn A i PU i i stratum h.

Anta $n_i / N_{h,i}$ er små. Da er variansene for de to utvalgsplanene tilnærmet gitt ved:

$$Var_{SSB}(\hat{p}) = V_1 + V_2$$

hvor

$$V_1 = \frac{1}{N^2} \sum_{h=1}^{109} N_h v_h , \quad \text{hvor } v_h = \sum_{i=1}^{m_h} N_{h,i} (p_{h,i} - p_h)^2$$

$$V_2 = \frac{1}{n} \cdot \frac{1}{N} \sum_{h=1}^{109} w_h , \quad \text{hvor } w_h = \sum_{i=1}^{m_h} N_{h,i} p_{h,i} (1 - p_{h,i}).$$

$$Var_{STRAT}(\hat{p}) = \frac{1}{n} \cdot \frac{1}{N} \sum_{h=1}^{109} N_h p_h (1 - p_h) .$$

Det kan vises at $V_2 \approx Var_{STRAT}(\hat{p})$. Vanligvis vil imidlertid $Var_{SSB}(\hat{p}) > Var_{STRAT}(\hat{p})$. V_1 er bidraget fra 1.trinntrekkingen. Den er mindre jo mindre variasjonen i andelene $p_{h,i}$ er innen strataene, dvs. når strata er homogene med hensyn på andel med kjennetegn A. Vi kan si at de to hovedegenskapene ved to-trinnstrekking er følgende:

Fordel: redusering av reiseutgifter og arbeidstidsutgifter
Ulempe: økt estimeringsvarians.

Det er viktig å danne homogene strata. Da vil den økte estimeringsvariansen, i forhold til et-trinns stratifisert utvalg, være liten og av mindre betydning enn kostnadsreduksjonen man oppnår med to-trinnstrekking.

REFERANSE

Stålnacke, M., J-A. Sigstad Lie og L.Solheim (1999). En analyse av SSBs generelle utvalgsplan fra 1995 basert på næringsvise sysselsettingstall. Notater 99/36. Statistisk sentralbyrå, 83s.