Opvejningsproblemet i stikprøve-undersøgelser

Af Carsten Stig Poulsen*)

Resumé

Ved udtagelse af stikprøver ved simpel tilfældig udvælgelse vil der ofte forekomme skævheder i stikprøvens fordeling på en række kriterier i forhold til populationens sammensætning. I det omfang man kender denne sammensætning vil det være nærliggende at søge at korrigere disse skævheder for derved at opnå en bedre »repræsentativitet« af stikprøven. Denne fremgangsmåde er da også ganske udbredt blandt analyseinstitutter og andre praktikere og kendes under betegnelsen opvejning.

Flemming Hansen har i en artikel i dette tidsskrift[1] gennemgået den konkrete metode ved et eksempel i forbindelse med bortfald. Jeg har selv tidligere hævdet det synspunkt[2], at spørgsmålet om opvejning burde ses som et selvstændigt problem, som ikke nødvendigvis var forbundet med bortfald. Jeg skal derfor i denne artikel belyse fremgangsmåden fra en mere teoretisk statistisk synsvinkel.

I næste afsnit vil nogle resultater fra stikproveteorien blive gennemgået for at etablere den nødvendige baggrund for det efterfølgende mere centrale afsnit, hvor opvejning karakteriseres som ex post stratifikation med tilfældig allokering af stikprøven. Med dette som udgangspunkt demonstreres det, at opvejningen undertiden kan føre til ringere, dvs. mere usikre estimater, og betingelserne herfor påpeges. Den væsentligste konklusion er følgelig, at opvejning som procedure ikke bør anvendes automatisk, men at man i hvert tilfælde må sikre sig, at de nødvendige forudsætninger herfor er til stede.

*) Seniorstipendiat, Handelshøjskolen i Århus, p.t. Wharton School, University of Pennsylvania, USA. Artiklen modtaget juni 1981.

Opvejningsproblemet i stikprøve-undersøgelser

Af Carsten Stig Poulsen*)

Resumé

Ved udtagelse af stikprøver ved simpel tilfældig udvælgelse vil der ofte forekomme skævheder i stikprøvens fordeling på en række kriterier i forhold til populationens sammensætning. I det omfang man kender denne sammensætning vil det være nærliggende at søge at korrigere disse skævheder for derved at opnå en bedre »repræsentativitet« af stikprøven. Denne fremgangsmåde er da også ganske udbredt blandt analyseinstitutter og andre praktikere og kendes under betegnelsen opvejning.

Flemming Hansen har i en artikel i dette tidsskrift[1] gennemgået den konkrete metode ved et eksempel i forbindelse med bortfald. Jeg har selv tidligere hævdet det synspunkt[2], at spørgsmålet om opvejning burde ses som et selvstændigt problem, som ikke nødvendigvis var forbundet med bortfald. Jeg skal derfor i denne artikel belyse fremgangsmåden fra en mere teoretisk statistisk synsvinkel.

I næste afsnit vil nogle resultater fra stikproveteorien blive gennemgået for at etablere den nødvendige baggrund for det efterfølgende mere centrale afsnit, hvor opvejning karakteriseres som ex post stratifikation med tilfældig allokering af stikprøven. Med dette som udgangspunkt demonstreres det, at opvejningen undertiden kan føre til ringere, dvs. mere usikre estimater, og betingelserne herfor påpeges. Den væsentligste konklusion er følgelig, at opvejning som procedure ikke bør anvendes automatisk, men at man i hvert tilfælde må sikre sig, at de nødvendige forudsætninger herfor er til stede.

*) Seniorstipendiat, Handelshøjskolen i Århus, p.t. Wharton School, University of Pennsylvania, USA. Artiklen modtaget juni 1981.

1. Nogle resultater fra samplingteorien

Vi antager, at middelværdien $\bar{X}$ af en kvantitativ størrelse i en endelig population med N elementer ønskes estimeret ved udtagelse af en stikprøve på n<N enheder. To velkendte statistiske udvælgelsesmetoder er:

(a) simpel tilfældig udvælgelse
(b) stratificeret tilfældig udvælgelse

med estimatorerne[3]

$$(a):\ \bar{x} = \frac{1}{N} \sum_i x_i \tag{1}$$

$$(b):\ \bar{x}_{str} = \sum_j \frac{N_j}{N} \bar{x}_j,\ \bar{x}_j \triangleq \frac{1}{n_j} \sum_i x_{ij} \tag{2}$$

hvor j=1,...,L betegner stratum. Det kan vises[4], at

$$E(\bar{x}) = \bar{X} \quad \text{og} \quad V(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \approx \frac{\sigma^2}{n} \left(1-\frac{n}{N}\right) \tag{3}$$

$$E(\bar{x}_{str}) = \bar{X} \text{ og } V(\bar{x}_{str}) = \sum_j \left(\frac{N_j}{N}\right)^2 \cdot V(\bar{x}_j);$$

$$V(\bar{x}_j) \approx \frac{\sigma_j^2}{n_j} \left(1 - \frac{n_j}{N_j}\right) \tag{4}$$

Her betegner σ^2 (σ_j^2) variationen i kendetegnet for populationen (j'te stratum), målt ved variansen på X. Begge estimatorer er middelrette, og en sammenligning af de to procedurer kan ske ved at betragte differencen mellem deres varianser:

$$V(\bar{x}) - V(\bar{x}_{str}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_j \frac{N_j}{N} (\bar{X}_j - \bar{X})^2$$

$$+ \frac{1}{n} \sum_j \frac{N_j}{N} (1-W_j) \sigma_j^2 \ ;\ W_j \triangleq \frac{N_j/N}{n_j/n} \tag{5}$$

Af (5) ses, at stratifikationsgevinsten kan være positiv, nul eller negativ. Første led på højresiden af (5) måler variationen i stratummiddelværdier og er altid ikke-negativ. Fortegnet for andet led er derimod ubestemt, idet det afhænger af *fordelingen* af stikprøven på strata. Antag eksempelvis, at strata'erne *er* fuldstændig ens: $\bar{X}_j$ og σ_j^2 er konstante over j. Da reduceres (5) til:

$$V(\bar{x})-V(\bar{x}_{str})=\frac{1}{n}\sigma^2\sum_j \frac{N_j}{N}\cdot(1-W_j) \qquad (6)$$

og gevinsten ved stratifikation (som altså her er overflødig på forhånd) er *negativ,* såfremt

$$\sum_j \frac{N_j}{N}\cdot W_j>1$$

Ved stratificeret tilfældig udvælgelse er stikprøvens fordeling på strata derfor afgørende. En ofte benyttet metode er *proportional allokering,* hvor fordelingen af stikprøven sker efter relativ stratumstørrelse N_j/N, dvs. $W_j=1$ for alle j. Det ses, at stratifikationsgevinsten i dette tilfælde alene afhænger af variationen i stratummiddeltallene $\bar{X}_j$, idet andet led på højresiden i (5) reduceres til nul. Med proportional allokering kan gevinsten aldrig blive negativ.

2. Stratifikation efter sampling: Opvejning

Det kan ofte være umuligt eller yderst vanskeligt at stratificere populationen, før stikprøven er udtaget. Fordelingen af populationen på generelle kriterier som køn, alder, civilstand osv. kan være tilgængelig fra eksterne kilder (officiel statistik), og det er da nærliggende at udnytte denne information til *efterfølgende* at stratificere den udtagne stikprøve[6]. Det er denne fremgangsmåde, som går under betegnelsen *opvejning* og som med et finere udtryk kan kaldes *ex post stratifikation* til forskel fra den sædvanlige stratifikation beskrevet ovenfor, som foregår *ex ante,* dvs. før stikprøven udtages.

Rent intuitivt er det rimeligt, at stratifikation ex post er mindre effektiv (mindre variansreduktion) end ex ante stratifikation. Fælles for de to samplingmetoder er udnyttelsen af ekstern information om den betragtede population, men ved stratifikation ex ante inddrages denne information allerede ved *planlægningen af stikprøven,* mens den ved ex post stratifikation kun kan udnyttes ved den *realiserede* stikprøve.

Det er måske mindre indlysende, at ex post stratifikation eller opvejning som metode kan være direkte *skadelig* for præcisionen af estimatorerne. Dette forhold vil blive formelt belyst i det følgende.

Fra et teoretisk statistisk synspunkt er den væsentligste forskel mellem stratifikation ex ante og ex post, at allokeringen af stikprøven kun i første tilfælde er under planlæggerens kontrol. Ved ex post stratifikation er fordelingen af stikprøven på strata derimod tilfældig og må betragtes som resultatet af en stokastisk proces som primært styres af de relative stratumstørrelser N_j/N. På denne baggrund er det ikke overraskende, at den fulde (ex ante) stratifikationsgevinst ikke kan opnås ex post. Med en *given* allokering n_j kan formlerne i (4) stadig anvendes, men n_j og dermed også W_j er nu stokastiske variabler, og vi skal derfor vurdere [7]

$$E(\bar{x}_{str*}) = E(E(\bar{x}_{str}|n_j)) \tag{7}$$

$$V(\bar{x}_{str*}) = E(V(\bar{x}_{str}|n_j)) + V(E(\bar{x}_{str}|n_j)) \tag{8}$$

Da $\bar{x}_{str}$ er en middelret estimator for $\bar{X}$ for alle n_j, vil det samme gælde $\bar{x}_{str*}$. Det medfører videre, at andet led i (8) falder bort. Tilbage at betragte bliver:

$$E(V(\bar{x}_{str}|n_j)) = \left(\frac{N_j}{N}\right)^2 \cdot \sigma_j^2 \left(E\left(\frac{1}{n_j}\right) - \frac{1}{N}\right) \tag{9}$$

og den *forventede* gevinst ved ex post stratifikation sammenholdt med simpel tilfældig udvælgelse er:

$$V(\bar{x}) - V(\bar{x}_{str*}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_j \frac{N_j}{N} \cdot (\bar{X}_j - \bar{X})^2 + \frac{1}{n} \sum_j \frac{N_j}{N} (1 - E(W_j)) \sigma_j^2 \tag{10}$$

hvor

$$E(W_j) \triangleq E\left(\frac{N_j/N}{n_j/n}\right) = \frac{N_j}{N} \cdot n \cdot E\left(\frac{1}{n_j}\right) \tag{11}$$

Af Jensens ulighed [8] fås nu:

$$E\left(\frac{1}{n_j}\right) > \frac{1}{E(n_j)}$$

n_j kan tilnærmelsesvis betragtes som multinomisk fordeling med antalsparameter n og sandsynlighedsparametre N_j/N. Derfor fås:

$$E(n_j) = n \cdot \frac{N_j}{N}$$

og dermed

$$E(W_j) > \frac{N_j}{N} \cdot n \cdot \frac{N/N_j}{n} = 1$$

Andet led på højresiden af (10) er altså negativt. Antager vi derfor igen, at $\overline{X}_j$ ikke varierer over strata, er første led nul og ex post stratifikationsgevinsten er negativ. Det samme resultat vil gælde selv ved mindre forskelle mellem stratummiddelværdierne.

3. Afsluttende bemærkninger

Den generelle konklusion af den gennemførte analyse er, at inddragelse af ekstern information om populationen gennem opvejning ikke er omkostningsfri. Den nødvendiggør estimering af flere parametre, her de betingede middelværdier $\overline{X}_j$ inden for hvert stratum[9]. Opvejning som metode forudsætter klare forskelle mellem strata m.h.t. det undersøgte kendetegn, hvilket bør testes inden metoden anvendes, om muligt på et selvstændigt datasæt. Dertil kommer, at stikprøvens størrelse sætter en grænse for hvor finmasket en opvejning der kan gennemføres. Ved små datasæt vil estimationen af stratummiddelværdierne være meget ustabil, og tomme celler vil kunne forekomme, hvorved en vurdering af usikkerheden på estimatorerne umuliggøres. En analytisk belysning af sammenhængen mellem disse centrale faktorer udover den allerede gennemførte forekommer vanskelig. Det vil dog være en enkel sag ved hjælp af numerisk simulation at klarlægge de nævnte forholds kvantitative og dermed praktiske betydning.

4. Noter

1) Se Hansen (1977).

2) Poulsen (1980).

3) Det er almindeligt i stikprøveteori (mere præcist: i teorien for udtagelse af stikprøver fra endelige populationer) at lade store (latinske) bogstaver angive populationsværdier, mens de tilsvarende små bogstaver betegner stikprøveværdier. Vi følger denne konvention her.

4) Se f.eks. Jensen (1974).

5) Ved såkaldt *optimal allokering* inddrages også forskelle i samplingomkostninger og stratumvarians σ_j^2. Princippet kan yderligere generaliseres til at omfatte forskelle i bortfaldssandsynlighed, se Poulsen (1980).

6) Der gøres opmærksom på, at i det omfang de anvendte kriterier er statistisk afhængige må opvejningen ske på grundlag af den simultane fordeling (den fuldstændige krydstabel). Ofte foreligger kun oplysninger om marginale fordelinger (frekvenstabeller) af kriterierne umiddelbart tilgængelige.

7) I udtrykket $E(E(\bar{x}_{str}|n_j))$ går den inderste middelværdidannelse på $\bar{x}_{str}$, givet n_j, mens den yderste henføres til variationen i n_j. Formlen for variansen kan findes i f.eks. Jensen (1974) p. 59.

8) Denne ulighed udtrykker, at den forventede værdi af en konveks (»opad hul«) funktion af en stokastisk variabel, er større end middelværdien indsat i funktionen. Formelt:

$$E(f(x)) > f(E(x)) \text{ for } f \text{ konveks.}$$

9) Ved vurderingen af usikkerheden kræves tillige skøn over stratumspredningerne σ_j.

Litteratur:

Flemming Hansen, Sampling teori og anvendt statistik, s. 215-227, Erhvervsøkonomisk Tidsskrift 4, 1977.

N. E. Jensen (red.), Stikprøveteori, København, 1974.

Carsten Stig Poulsen, Bortfaldsproblemet i stikprøveundersøgelser, s. 227-240, Erhvervsøkonomisk Tidsskrift 4, 1980.

4. Noter

1) Se Hansen (1977).

2) Poulsen (1980).

3) Det er almindeligt i stikprøveteori (mere præcist: i teorien for udtagelse af stikprøver fra endelige populationer) at lade store (latinske) bogstaver angive populationsværdier, mens de tilsvarende små bogstaver betegner stikprøveværdier. Vi følger denne konvention her.

4) Se f.eks. Jensen (1974).

5) Ved såkaldt *optimal allokering* inddrages også forskelle i samplingomkostninger og stratumvarians σ_j^2. Princippet kan yderligere generaliseres til at omfatte forskelle i bortfaldssandsynlighed, se Poulsen (1980).

6) Der gøres opmærksom på, at i det omfang de anvendte kriterier er statistisk afhængige må opvejningen ske på grundlag af den simultane fordeling (den fuldstændige krydstabel). Ofte foreligger kun oplysninger om marginale fordelinger (frekvenstabeller) af kriterierne umiddelbart tilgængelige.

7) I udtrykket $E(E(\bar{x}_{str}|n_j))$ går den inderste middelværdidannelse på $\bar{x}_{str}$, givet n_j, mens den yderste henføres til variationen i n_j. Formlen for variansen kan findes i f.eks. Jensen (1974) p. 59.

8) Denne ulighed udtrykker, at den forventede værdi af en konveks (»opad hul«) funktion af en stokastisk variabel, er større end middelværdien indsat i funktionen. Formelt:

$$E(f(x)) > f(E(x)) \text{ for } f \text{ konveks.}$$

9) Ved vurderingen af usikkerheden kræves tillige skøn over stratumspredningerne σ_j.

Litteratur:

Flemming Hansen, Sampling teori og anvendt statistik, s. 215-227, Erhvervsøkonomisk Tidsskrift 4, 1977.

N. E. Jensen (red.), Stikprøveteori, København, 1974.

Carsten Stig Poulsen, Bortfaldsproblemet i stikprøveundersøgelser, s. 227-240, Erhvervsøkonomisk Tidsskrift 4, 1980.