Sample rectifying by post-stratification and calibration

Lect. Ileana Gabriela NICULESCU-ARON
Catedra de Statistică şi Previziune Economică, ASE Bucureşti

Insuring the representativity of the sample is a main concern for the person who organizes a survey. The methods for rectifying the samples have constantly improved. Gradually, by using certain algorithms, on which efficient software programs are based, a transition from generalized post-stratification to generalized calibration was made.
Keywords: *sample, post-stratification, calibration.*

Stabilirea planului de sondaj şi a procedeelor de selecţie a unităţilor în eşantion sunt foarte importante şi se realizează cu multă rigurozitate.

Întocmirea chestionarului pe baza obiectivelor stabilite la începutul studiului presupune şi definirea clară a variabilelor. În studiile realizate pe baza sondajelor statistice distingem următoarele tipuri de variabile:

- **variabile de interes**. Sunt cele care trebuie estimate şi ţin de obiectivele sondajului. În cazul anchetelor asupra forţei de muncă variabile de interes pot fi de exemplu statutul după participarea la activitatea economică (ocupat, şomer, inactiv) sau durata efectivă a săptămânii de lucru.
- **variabile auxiliare sau de identificare.** Sunt cele pe care le cunoaştem din alte surse şi nu formează obiectul studiului prezent dar sunt utilizate în stabilirea planului de sondaj şi ulterior în ameliorarea extrapolării.

În ciuda eforturilor de a asigura reprezentativitatea eşantionului, în anchetele de mare amploare, realizate la nivel naţional, de cele mai multe ori estimatorii variabilelor auxiliare sunt deplasaţi faţă de parametrii populaţiei. Datorită existenţei corelaţiei dintre aceste variabile şi variabilele de interes putem presupune că şi estimatorii acestora din urmă vor avea aceeaşi problemă.

Plecând de la rezultatele brute furnizate de respondenţi se estimează cât mai bine posibil, prin extrapolare, situaţia populaţiei. Această metodă constă în atribuirea unei anumite ponderi sau coeficient de extrapolare fiecărui respondent după cum el reprezintă o fracţiune din populaţie. Pentru a extrapola trebuie să dăm fiecărui respondent ponderea sa iniţială reprezentată de inversul probabilităţii cu care a fost selecţionat ($d_k = 1/\pi_k$ k=1,…n_r, n_r fiind numărul de respondenţi). Acest lucru nu este posibil deoarece[97]:

- unitatea selectată este gospodăria, nimic nu ne garantează că un individ din 500 va face parte din eşantion;
- structura populaţiei după variabilele auxiliare (vârstă, sex medii de rezidenţă etc.) se va regăsi în eşantion cu o oarecare aproximaţie;
- mai mult ca sigur că, datorită nonrăspunsurilor totale, numărul respondenţilor va fi inferior efectivului eşantionului selectat iniţial.

Datorită acestor cauze este necesară modificarea coeficienţilor de extrapolare iniţiali având drept scop asigurarea preciziei şi coerenţei valorilor extrapolate.

Până nu demult redresarea eşantioanelor în vederea extrapolării se realiza prin poststratificarea realizată pe baza variabilelor auxiliare.

Variabilele clasice utilizate de obicei în vederea post-stratificării eşantioanelor din sondajele realizate la nivel naţional sunt:

- ***REG***: regiunea de dezvoltare în care locuieşte respondentul. România este împărţită în 8 regiuni: Nord Est, Sud Est, Sud, Sud Vest, Vest, Nord Vest, Centru, Bucureşti.
- ***VÂRSTA:*** grupa de vârstă căreia îi aparţine respondentul. În ancheta asupra forţei de muncă se utilizează următoarele 6 grupe: 15-24 ani, 25-34 ani, 35-44 ani, 45-54 ani, 55-64

[97] Luminet D., Vanderhoeft C., Une méthode de calibrage appliquée aux statistiques de l´emploi, Carrefour de l´Economie 2003/7-8A, Bruxelles, pg. 4

ani şi peste 65 ani.

o ***MREZ***: mediul de rezidenţă Urban sau Rural.

o ***SEX***: masculin sau feminin.

Pentru a realiza o post-stratificare după cele patru variabile auxiliare fiecare respondent va fi clasat într-una din cele 8x6x2x2=192 celule numite post-straturi. În fiecare celulă h (h=1,…192) vom avea un număr n_h de respondenţi iar $\sum_{h=1}^{192} n_h = n$ unde n este numărul total de respondenţi.

Pe baza informaţiilor din alte surse putem determina distribuţia pentru populaţia României cu vârsta 15 ani şi peste (N) după cele patru variabile: REG, VÂRSTA, MREZ, SEX reprezentată de frecvenţele N_h unde $\sum_{h=1}^{192} N_h = N$.

Frecvenţele relative ale eşantionului sunt date de raportul n_h / n iar ale populaţiei de $\frac{N_h}{N}$.

Din diferite motive vor exista celule pentru care $\frac{n_h}{n} \neq \frac{N_h}{N}$ eşantionul respondenţilor nereprezentând fidel populaţia României cu vârsta de 15 ani şi peste, anumite celule (post-straturi) fiind sub sau supraestimate. „ În sens strict matematic am putea spune că eşantionul nu este reprezentativ pentru populaţia din care a fost extras. Totuşi noţiunea (ne)reprezentativ este în general utilizată într-un sens mai puţin strict. În acest caz se poate pretinde că eşantionul este reprezentativ pentru populaţie dacă toate frecvenţele n_h sunt nenule şi suficient de mari”[98]. În felul acesta fiecare combinaţie dintre regiune, vârstă, mediu de rezidenţă şi sex este suficient reprezentată.

Existenţa unor diferenţe semnificative statistic între frecvenţele relative ale eşantionului de respondenţi şi populaţie poate avea consecinţe grave asupra calităţii estimatorilor.

Am considerat **y** o variabilă de interes (de exemplu durata săptămânii de lucru. Pe baza anchetei asupra forţei de muncă urmează să se estimeze durata medie efectivă a săptămânii de lucru. Această variabilă nu este auxiliară şi nu dispunem de informaţii în ceea ce o priveşte din alte surse. Calculând durata medie efective a săptămânii de lucru ($\bar{y}$) ca o medie neponderată pe baza datelor din eşantion vom obţine un estimator deplasat pentru media populaţiei din motivele de sub sau supra reprezentare prezentate anterior.

Estimatorul timpului total de muncă ($N\bar{y}$) care va prezenta aceleaşi neajunsuri se poate scrie sub forma: $N\bar{y} = N\frac{y_T}{n}$ unde y_T reprezintă timpul total de muncă din eşantion.

Din această relaţie se deduce că pentru a trece de la totalul eşantionului la totalul populaţiei trebuie să atribuim fiecărui respondent acelaşi coeficient de ponderare N/n. Se poate spune că, în medie, fiecare respondent al anchetei reprezintă N/n din populaţia României cu vârste de 15 ani şi peste.

Utilizând post-stratificarea această ponderare nu se va efectua uniform. Fiecare respondent din post-stratul h va primi un coeficient de ponderare de forma N_h / n_h cu h=1,…..192. În aceste condiţii timpul total de muncă din populaţie Y_T va fi estimat pe baza relaţiei: $\hat{Y}_T = \sum_{h=1}^{192} N_h \bar{y}_h$, iar durata medie efectivă a săptămânii de lucru se va estima pe baza relaţiei:

$\hat{\bar{Y}} = \sum_{h=1}^{192} N_h \bar{y}_h \Big/ N$. Practic, fiecare respondent din stratul h reprezintă N_h / n_h indivizi ce îndeplinesc aceleaşi caracteristici.

Calitatea estimatorilor obţinuţi prin post-stratificare depinde de măsura în care variabilele auxiliare regiune, vârstă, mediu de rezidenţă şi sex explică variaţia variabilelor de interes. Este clar că cele patru variabile sunt explicative pentru un număr foarte mare de variabile de interes totuşi, în cazul în care se doreşte o analiză detaliată sau estimarea unor

[98] Luminet D., Vanderhoeft C., Une méthode de calibrage appliquée aux statistiques de l´emploi, Carrefour de l´Economie 2003/7-8A, Bruxelles, pg. 13.

variabile de interes specifice, ele nu mai sunt suficiente. Acest fapt a determinat abandonarea tehnicii clasice de post-stratificare. Modelul de post-stratificare prezentat anterior poate fi prezentat succint sub forma: REG x VÂRSTĂ x MREZ x SEX.

Pentru aceasta, pornind de la modelul inițial REG x VÂRSTĂ x MREZ x SEX putem adăuga alte variabile ajungând la un model detaliat de forma: REG x VÂRSTĂ x MREZ x SEX x X_1 x X_2…… Noile variabile introduse vor permite ameliorarea estimațiilor. În acest model variabilele utilizate nu se vor mai numi variabile de post-stratificare ci **variabile de calibrare**.

În momentul în care numărul variabilelor de calibrare este foarte mare nu se vor putea calcula izolat ponderile pentru fiecare celulă rezultată din încrucișarea variabilelor. Literatura de specialitate propune diferiți algoritmi pentru determinarea unei soluții.

Într-un model de post-stratificare este posibil, ca pentru fiecare post-strat h să se definească o ecuație exprimată ca sumă a greutăților (ce trebuie calculate) acordate respondenților din post-stratul h şi a căror sumă trebuie să corespundă cu efectivul populației N_h: astfel: $\sum_{k\in h} w_k = N_h$ unde w_h este greutatea acordată respondentului h.

Pentru fiecare post-strat stabilim o astfel de ecuație în final obținând un sistem de h ecuații liniare ce ar trebui rezolvat. Rezolvarea lui duce la mai multe soluții deoarece, pe de o parte acest sistem cuprinde mai multe necunoscute decât ecuații iar pe de altă parte fiecare individ k nu este cuprins într-o singură ecuație. O soluție particulară este obținută impunând ca toți indivizii din același post-strat h să aibă aceeași greutate w_h astfel încât fiecare va fi tratat în aceeași manieră. Ecuația post-stratului h poate fi formulată astfel:

$n_h w_k = N_h$ deci $w_k = N_h / n_h$

Deoarece sistemul de ecuații de calibrare are mai multe soluții putem alegea acea soluție care modifică cel mai puțin **coeficienții de extrapolare inițiali.** Practic vom căuta acei coeficienți de redresare w_k care să verifice sistemul de ecuații de calibrare şi în același timp să fie cât mai apropiați posibil de coeficienții de extrapolare ce rezultă direct din planul de eşantionare ($d_k = 1/\pi_k$ k=1,…n_r).

Problema generală a calibrării constă în ajustarea coeficienților de extrapolare inițiali $d_k = 1/\pi_k$ şi obținerea unor coeficienți de redresare calibrați de forma : $w_k = g_k d_k$ unde g_k reprezintă factorul de ajustare.

Este vorba de următoarea problemă de optimizare:

$$\begin{cases} \sum_{k=1}^{n_r} d_k G\left(\frac{w_k}{d_k}\right) \text{ minim} \\ \sum_{k=1}^{n_r} w_k x_{kj} = T_j (j = 1,...m) \text{ restrictiile de calibrare} \end{cases}$$

unde:

- m reprezintă numărul de variabile de calibrare iar x_{kj} este valoarea variabilei de calibrare j pentru respondentul k
- T_j reprezintă totalul populației pentru variabila de calibrare j.
- *G este funcția distanțelor* definită pe vecinătatea convexă a lui 1 şi care verifică următoarele condiții:

  G ≥0;
  G este strict convexă;
  G este de două ori continuu derivabilă;
  G(1)= 0
  G′(1)= 0;
  G′′(1)= 1.

Teorema funcțiilor implicite afirmă că funcția reciprocă F a lui G′, definită şi continuu derivabilă în vecinătatea lui 0 satisface condițiile F(0)=1 şi F′(0)=1. F se numeşte **funcție de calibrare**.''[99]

Pentru simplificarea explicației presupun că problema de optimizare are doar o soluție şi o singură restricție şi anume:

$\sum_{k=1}^{n_r} x_k w_k = T$ unde x_k este variabila de calibrare iar T reprezintă totalul populației pentru variabila de calibrare.

Pentru soluționarea problemei de minimizare ținând cont de restricție (extreme cu legături)

[99] Luminet D., L'enquête sur les Forces de travail: calibrage et autres développements, Statistics Belgium, Working Paper nr. 8, pg. 32

se recurge la funcţia Lagrange definită astfel:

$$L(w_{k,\lambda}) = \sum_{k=1}^{n_r} d_k G\left(\frac{w_k}{d_k}\right) - \lambda\left(\sum_{k=1}^{n_r} x_k w_k - T\right)$$

unde λ este valoarea multiplicatorului Lagrange.

Pentru a afla valorile minime anulăm derivatele parţiale şi obţinem următorul sistem:

$$\begin{cases} \sum_{k=1}^{n_r} x_k w_k = T \\ G'\left(\frac{w_k}{d_k}\right) = \lambda x_k \;\Rightarrow\; w_k = d_k F(\lambda(_k) \end{cases}$$

Introducând k_w în prima relaţie obţinem :

$$\sum_{k=1}^{n_r} x_k d_k F(\lambda(_k) = T$$

Pe baza acestei relaţii putem afla valoarea multiplicatorului Lagrange şi ulterior valorile w_k.

Avem posibilitatea de a alege un criteriu de apropiere între coeficienţii de extrapolare iniţiali şi coeficienţii de redresare calibraţi, deci de a alege o formă a funcţiei distanţelor G şi implicit a funcţiei de calibrare F.

Calibrarea generalizată a devenit în momentul de faţă o metodă cunoscută şi puternică în mediile de specialitate deoarece, utilizând informaţii auxiliare din diferite surse, reuşeşte să îmbunătăţească estimaţiile sondajelor prin creşterea preciziei estimatorilor obţinuţi. Mai mult decât atât, calibrarea este utilizată pentru corectarea deplasării produse de nonrăspunsurile totale.

Rezolvarea matematică a problemei de minimizare cu restricţii prezentată necesită un volum mare de calcule cu un nivel ridicat de dificultate. Literatura de specialitate propune diferiţi algoritmi pentru determinarea soluţiei ce stau la baza programelor software pe baza cărora se realizează calibrarea eşantioanelor. În con tinuare voi prezenta succint aceste programe:

Generalized Estimation System (GES). Este utilizat de Statstics Canada şi este realizat sub programul SAS. Are la bază estimarea pe baza regresiei generalizate (GREG) descrisă de Stärndal, Swensson şi Wretman în „Model Assisted Survey Sampling", 1992. Această metodă acoperă o clasă de estimatori calibraţi care cuprinde cei mai utilizaţi estimatori. Cu toate acestea metoda de calibrare generalizată introdusă de Deville şi Stärndal în 1992 este mai cuprinzătoare.

GES este utilizat împreună cu un alt soft realizat sub SAS, GSAM (Generalised Sampling System). Ambele programe acoperă diverse tipuri de sondaj simple sau complexe.

Avantajul GES este deci integrarea calibrării, a estimării pentru totaluri, medii, proporţii, rapoarte şi variaţia estimatorilor chiar dacă numai pe baza metodei GREG.

BASCULA a fost realizată în Delphi pentru Windows 95 de către Nieuwenbroek în 1997 şi este utilizat în Olanda. Ca şi GES se bazează pe metoda regresiei generalizate (GREG). Variaţia estimatorilor se bazează pe tehnica reeşantionării (balanced repeated sampling – BRR)

O particularitate a programului BASCULA este modul în care factorul de ajustare g_k este limitat în metoda liniară. Contrar procedurii utilizate de CALMAR şi g-CALIB factorii de ajustare nu sunt trunchiaţi ci mai degrabă redimensionaţi printr-o procedură iterativă. Aceasta este considerată o tehnică de netezire limitată faţă de tehnica trunchierii iterative.

CALMAR (Calage sur Marges). Acest program a fost propus de membrii INSEE (Institut National de la Statistique des Etudes Economiques – Franţa). CALMAR este realizat sub SAS şi are la bază metoda de calibrare generalizată introdusă de Deville şi Stärndal în 1992 concentrată pe calcularea coeficienţilor de calibrare w_h şi a factorilor de ajustare g_h. Un instrument central al metodei îl constituie funcţia distanţelor G. Din punct de vedere practic, CALMAR este considerat superior sistemelor prezentate anterior deoarece permite utilizatorilor să limiteze în mai multe moduri flexibilitatea coeficienţilor de redresare w_k. Din punct de vedere teoretic, deoarece se bazează pe metoda calibrării generalizate, sfera estimatorilor calibraţi este mult mai cuprinzătoare.

Şi în România Institutul Naţional de Statistică utilizează acest pachet program pentru calcularea coeficienţilor de ponderare în vederea creşterii gradului de precizie al estima-

ţiilor şi pentru tratarea non-răspunsurilor totale atât în Ancheta asupra forţei de muncă AMIGO cât şi în celelalte anchete în care unitatea de selecţie este gospodăria sau întreprinderea.

g-CALIB. A fost introdus de către INS (Institut National de Statistique) – Belgia, având la bază pachetul statistic SPSS. Prima versiune a acestui program a fost realizată de Vanderhoeft şi a devenit un instrument performant, aplicabil în situaţii diverse şi capabil să rezolve probleme complicate de redresare a eşantioanelor.Ca şi CALMAR, are la bază metoda de calibrare generalizată introdusă de Deville şi Stärndal în 1992 concentrată pe calcularea coeficienţilor de calibrare w_h şi a factorilor de ajustare g_h.

Pachetele program **g-CALIB şi CALMAR** sunt comparabile din punct de vedere al fundamentării teoretice şi al performanţelor. Cu toate acestea, din anumite puncte de vedere CALMAR este privit ca fiind în prezent cel mai bun soft în acest domeniu din următoarele motive:

- Interfaţa este foarte prietenoasă nefiind necesar ca utilizatorul să fie un expert în metoda calibrării generalizate;
- CALMAR este astfel conceput ca variabilele cantitative şi calitative de calibrare să fie transformate automat într-o matrice proiectată de program. Acest lucru reduce foarte mult munca de pregătire a fişierelor input de către utilizator. Totuşi, pentru calibrarea variabilelor cantitative utilizatorul CALMAR trebuie să realizeze o transformare a acestora astfel încât să obţină un format standard pentru fişierul de intrare.
- Detectarea şi raportarea erorilor se realizează într-un mod mai precis şi mai eficient decât celelalte pachete program similare.

Concluzii

Pentru ca estimatorii variabilelor de interes obţinuţi în urma prelucrării datelor din eşantion să nu fie deplasaţi faţă de parametrii populaţiei este recomandat să fie aplicată una din metodele de ameliorare a eşantioanelor prezentate în această lucrare..

Deşi calibrarea generalizată este în momentul de faţă o metodă puternică pentru că reuşeşte să îmbunătăţească estimaţiile sondajelor prin creşterea preciziei estimatorilor obţinuţi corectând într-o oarecare măsură şi deplasările produse de non-răspunsurile totale, ea nu este suficient cunoscută în România. Unul dintre motive este şi costul foarte ridicat al programelor software pe baza cărora se poate implementa această metodă. Consider că orice institut ce are pretenţia realizării unui eşantion reprezentativ la nivel naţional ar trebui să poată realiza calibrarea acestuia.

Bibliografie

Luminet D., Vanderhoeft C., Une méthode de calibrage appliquée aux statistiques de l´emploi, Carrefour de l´Economie 2003/7-8A, Bruxelles

Luminet D., Vanderhoeft C., Une méthode de calibrage appliquée aux statistiques de l´emploi, Carrefour de l´Economie 2003/7-8A, Bruxelles

Droesbeke J-J.,Théorie et Practiques de l'échantillonnage, Formation des statisticiens européens, Eurostat, Support de cours