# MANAGEMENTUL DATELOR LIPSĂ

**Balázsi Robert**[1]
Universitatea Babeş-Bolyai, Cluj-Napoca

Indiferent de domeniul în care este utilizat, demersul statistic inferențial are ca scop oferirea unor informații asupra parametrilor populației (Radu şi colab., 1993). Atingerea acestui scop poate fi semnificativ afectată de prezența datelor lipsă. Prin definiție, datele lipsă (DL) nu se referă la rata non-răspunsurilor. Strategiile de cuantificare a ratei de răspuns, respectiv a impactului acesteia asupra parametrilor estimați, presupun utilizarea unor metode diferite decât cele descrise în acest articol (Domuța şi colab., 2003).

Datele lipsă se referă la caracterul incomplet al setului de date, la lipsa unui răspuns doar la anumite întrebări sau anumiți itemi al unui chestionar sau test standardizat. Ignorarea acestor date lipsă, şi efectuarea procesărilor statistice pe datele existente va afecta serios procesul inferențial, rezultând estimări distorsionate ale parametrilor populației (Cohen & Cohen, 1983).

Scopul acestui articol este de a oferi o imagine globală a procedurilor statistice utilizate în managementul DL, de a prezenta raționamentul ce stă la baza acestora fără însă a detalia aspectele lor computaționale. Pentru o aprofundare mai bună a cunoştințelor în acest sens, sugerăm parcurgerea studiilor lui Rubin (1976), sau Schafer (1997), acestea reprezentând lucrări de referință în domeniu.

Înainte de a trece la discuția propriu-zisă asupra managementului DL este foarte important să clarificăm caracteristicile acestora. În funcție de mecanismul care explică prezența lor, conform Little şi Rubin (1987), datele lipsă se împart în trei categorii:

a. *lipsă complet aleatoare* (missing completly at random): considerăm că lipsa datelor este complet aleatoare în cazul în care lipsa nu corelează cu alte valori existente ale variabilei respective, sau ale altor variabile. Dacă privim setul de date în ansamblul lui, putem afirma că datele lipsă sunt complet aleatoare în cazul în care oricare dintre date are aceeaşi probabilitate de a lipsi. În acest caz valorile prezente nu diferă în mod sistematic de valorile existente în baza de date, ca urmare singura problemă serioasă o reprezintă puterea redusă a testelor statistice (Wayman, 2003).

b. *lipsă aleatoare* (missing at random): considerăm că datele lipsă sunt aleatoare în cazul în care acestea depind de valori cunoscute ale altor variabile prezente în baza de date, şi se manifestă aleator doar în cadrul unui strat definit al eşantionului. Pentru a înțelege mai bine această problemă apelăm la un exemplu a lui Howel (1998). Persoanele depresive sunt mai puțin tentate să ofere informații legate de venitul lor. S-ar putea ca persoanele depresive să aibă în general un venit mai redus, față de populație în general. În acest caz, o rată crescută a lipsei informațiilor asupra venitului la depresivi va fi considerată aleatoare în contextul în care nu există o relație sistematică între raportarea/non-raportarea venitului în cadrul acestui grup. Lipsa datelor va fi asociată cu prezența depresiei, dar în cadrul acestei categorii se va manifesta aleator.

c. *lipsă non-aleatoare* (missing not at random): considerăm că lipsa datelor nu este aleatoare în cazul în care acestea sunt legate în mod sistematic de alte variabile şi evenimente care însă nu au fost incluse în studiu, ca urmare nici nu au fost măsurate.

Cele trei tipuri de mecanisme care generează date lipsă au fost împărțite de Graham & Donaldson (1993) în mecanisme „accesibile" (lipsa complet aleatoare şi lipsa aleatoare) şi mecanisme „non-accesibile" (lipsa non-aleatoare). Aceşti autori consideră că în practică, puține sunt acele cazuri în care datele sunt în întregime inaccesibile, de cele mai multe ori lipsa este rezultatul unui amestec a celor două tipuri de mecanisme.

Strategiile statistice de management al DL se împart în: strategii moderne (suplinire multiplă[2]) şi strategii tradiționale (înlocuirea valorilor lipsă cu media sau înlocuirea valorilor lipsă cu valoarea expectată a analizei de regresie). Utilitatea celor din urmă a fost frecvent chestionată de metodologi, dar în ciuda acestui fapt acestea sunt folosite în continuare în cercetare.

[1] Adresa de contact: robertbalazsi@psychology.ro

[2] Suplinire multiplă – termenul original din limba engleză este cel de *multiple imputation*.

Vom începe analiza cu o altă categorie, cea a strategiilor inacceptabile de management al DL (ştergerea totală a subiecților cu date lipsă sau ştergerea perechilor de date a subiectului cu date lipsă), subliniind că aceste metode ascund o serie de neajunsuri, ceea ce pune sub semnul întrebării utilizarea lor pe scară largă, în ciuda faptului că de multe ori acestea sunt singurele incluse în pachetele statistice uzuale.

a. *Ştergerea totală din baza de date a subiecților care au date lipsă*, incluzând în analiză doar subiecții cu date complete. Această strategie care reduce, mai mult sau mai puțin, numărul subiecților din baza de date, va reduce implicit şi puterea testelor statistice (Cohen & Cohen, 1983). Din această procedură rezultă parametri nedistorsionați doar în cazul în care datele lipsesc complet aleator, în alte cazuri inferențele noastre vor fi distorsionate.

b. *Ştergerea din calcul a subiectului care la una dintre variabilele incluse în corelație prezintă o valoare lipsă*. Această strategie caută să conserve numărul eşantionului şi implicit puterea testului. Singura problemă este că o matrice de corelație calculată pe diferite subgrupe ale eşantionului oferă informații eronate asupra eşantionului (şi a populației), în cazul în care datele nu lipsesc complet aleator. Dar chiar şi dacă lipsa este complet aleatoare, mai există o problemă, o astfel de matrice de corelație nu poate fi utilizată în prelucrări statistice avansate (analiză de cale, analiză factorială confirmatorie, etc) deoarece este dificilă (dacă nu imposibilă) stabilirea gradelor de libertate (Schafer, 1997).

În a doua categorie a strategiilor de management a datelor lipsă intră, înlocuirea datelor lipsă cu anumite valori calculate. Valorile suplinite pot fi: media pe variabila care include subiecți cu date lipsă sau valoarea expectată, obținută printr-o analiză de regresie a variabilei cu date lipsă la alte variabile complete.

a. *Suplinirea prin medie*, reprezintă o strategie relativ simplă şi foarte frecvent utilizată. Problema majoră a acesteia este că reduce într-o măsură mai mică, sau mai mare varianța variabilei. Aceasta va afecta în mod direct valoarea corelațiilor dintre variabila completată şi alte variabile, deoarece coeficientul de corelație este o formă de exprimare standardizată a covarianței.

b. *Suplinirea prin calculul unei valori expectate,* calculată pe baza analizei de regresie. În acest caz procedura este de a alege ca şi predictori variabile unde nu avem subiecți cu date lipsă, iar criteriul este variabila cu date lipsă. Valorile expectate rezultate în urma analizei de regresie sunt introduse în baza de date. Această strategie, deşi ia în considerare informații existente în baza de date (relația variabilei care conține date lipsă cu alte variabile) şi permite o estimare mai bună decât suplinirea prin medie, nu reduce problema erorii standard.

Un neajuns comun al demersurilor discutate este că generează senzația că cineva inventează date. Procesul de management al DL nu presupune generarea de date, ci elaborarea unui model statistic care să permită estimarea datelor lipsă, luând în considerare datele existente. Fiind un proces de estimare statistică, asemenea tuturor celorlalte demersuri de estimare (de exemplu, estimarea mediei în populație) trebuie să includă şi o componentă de eroare, adică o variabilitate a valorilor estimate.

Toate aceste neajunsuri pot fi evitate prin utilizarea unor proceduri ce țin de ultima categorie discutată a strategiilor de management al DL precum suplinirea multiplă care, deşi este acceptată în general de metodologii statisticieni, este mai puțin frecvent utilizată în practica cercetărilor. Acest lucru se datorează faptului că procedurile statistice aparținând acestei categorii nu sunt incluse în pachetele statistice uzuale. Mai mult, instrumentarul matematic al acestora este unul mult mai complicat, față de procedurile deja discutate, ceea ce reduce probabilitatea ca cineva să recurgă la efectuarea unor calcule cu creionul.

Suplinirea multiplă reprezintă una dintre cele mai atractive metode moderne care reuşeşte să stabilească un echilibru între simplitatea computațională şi calitatea estimării. În esență, suplinirea multiplă presupune derularea mai multor analize de regresie, de obicei între 3 şi 10. Predictorii utilizați sunt variabilele cu date complete, criteriul variabila cu date lipsă. Fiecare linie de regresie obținută oferă un set de valori, ce reprezintă estimări ale valorilor lipsă. Diferența dintre liniile de regresie rezultă din componenta eroare, inclusă în fiecare ecuație. Astfel pornind de la aceleaşi variabile predictor se obțin diferite linii de regresie, fiecare definind un posibil set al valorilor lipsă.

Ulterior procesarea statistică prevăzută în studiu (ex. analiză de varianță, comparații, analiză de regresie, etc) se aplică seturilor de date complete. Dacă am elaborat patru seturi de date, atunci vom derula procedura statistică pe toate cele patru baze de date.

În ultima fază a procedurii, valorile estimate ale parametrilor rezultați sunt coroborate, astfel încât se obține o valoare unică pentru toate bazele rezultate în urma suplinirii multiple. Bine înțeles există formule matematice care permit estimarea eficienței suplinirii, dar discuția

acestora nu reprezintă scopul acestui articol (pentru o lectură mai aprofundată vezi, Little & Rubin, 1987).

În concluzie, dorim să subliniem că problema răspunsurilor lipsă nu este una ce poate fi ignorată sau tratată cu superficialitate. Ignorarea sistematică, respectiv excluderea din baza de date a subiecților cu date lipsă poate afecta grav calitatea inferențelor efectuate. Actualmente există metode statistice care oferă soluții acceptabile la această problemă, crescând validitatea inferențelor efectuate.

**Bibliografie**


Cohen, J. & Cohen, P. (1983) Applied multiple regression/correlation analysis for the behavioral sciences. Lavrence Erlbaum Associates , London.

Domuța, A., Balazsi, R., Comşa, M., & Rusu, C. (2003) Standardizarea pe populația României a testului Matrici Progresive Raven Standard Plus. Psihologie Resurselor Umane, vol. 2, nr. 1, 50 – 57.

Graham, J.W. & Donaldson, S.I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology, 78:119-128.

Howell, D.C. (1998). Treatment of missing data [Online]. http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

Little, R.J.A. & Rubin, D.B. (1987) Statistical analysis with missing data. New York, Wiley.

Radu, I., Miclean, M., Albu, M., Moldovan, O., Nemşe, S. & Szamosközy (1993) Metodologie psihologică şi analiza datelor, Ed. Sincron, Cluj.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581–592.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. New York: Chapman & Hall.

Wayman, J. C. (2003). Multiple imputation for missing data: What is it and how can I use it? Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.