Metode cantitative de analiza in stiintele politice

Suport curs Invatamant la Distanta[1]
2013-2014

Facultatea de Stiinte Politice, Administrative şi ale Comunicării
Specializarea Ştiinţe Politice
Anul universitar 2013 - 2014
Semestrul II

Informaţii generale

Date de identificare a cursului:
Titlul disciplinei: Metode cantitative de cercetare in ştiinţele politice
Codul: ULR1415
Numărul de credite: 5

Locul de desfăşurare: Facultatea de Ştiinţe Politice, Administrative şi ale Comunicării, str. General Traian Moşoiu, nr. 71

Nume, titlul ştiinţific: Dr. Daniela Angi
Birou: Facultatea de Ştiinţe Politice, Administrative şi ale Comunicării, str. General Traian Moşoiu, nr. 71

Informaţii de contact (adresă e-mail): angi@fspac.ro
Ore de audienţă: Luni 18.00 – 20.00

Condiţionări şi cunoştinţe prerechizite:
Fără condiţionări

Descrierea cursului:

[1] Suport de curs elaborate de catre Conf. Dr. Cosmin Marian – FSPAC, UBB Cluj si adaptat de dr. Daniela Angi, FSPAC, UBB Cluj.

Acest curs este o continuare a cursului "Metode de cercetare in stiintele sociale" din anul
Cursul pune accentul asupra invatarii unor metode de analiza cantitativa a datelor (metodele de culegere a datelor cantitative au fost predate in anul 1, si vor fi recapitulate sumar in acest semestru). Studentii vor învata (i) sa formuleze ipoteze de cercetare, (ii) sa operationalizeze concepte, (iii) sa realizeze o analiza descriptiva a datelor, si (iv) sa analizeze relatii de cauzalitate între fenomenele sociale. Structura cursului, precum si modalitatea de lucru in cadru acestuia, sunt alese astfel incit sa faciliteze interactiunea dintre profesor si studenti. Cursul va fi axat pe realizarea unor cercetari, avand deci un caracter aplicat pronuntat iar studentii vor fi incurajati sa lucreze independent, intr-un mod creativ.

Organizarea temelor în cadrul cursului:

1. Exemple de cercetări în ştiinţele politice. Etapele unui proces de cercetare.
2. Populaţie şi eşantion. Tipuri de eşantioane.
3. Aspecte matematice ale eşantionării. Teste de semnificaţie
4. Variabile. Tipuri de variabile.
5. Analiza univariată a datelor
6. Analiza univariată a datelor. Aplicaţii.
7. Analiza bivariată a datelor.
8. Analiza bivariată a datelor. Aplicaţii.
9. Regresia liniară.
10. Regresia liniară. Aplicaţii.
11. Regresia multiliniară
12. Regresia multiliniară. Aplicaţii.

Formatul şi tipul activităţilor implicate de curs:
Pentru unele dintre aplicatii va fi folosit calculatorul. Cursul va presupune comunicarea intre profesori si studenti cu ajutorul calculatorului: email, intra-net si internet. Studentii vor avea acces la materiale scrise care vor fi salvate pe server sau vor fi trimise fiecarui student prin e-mail. Deasemenea, lucrarile scrise de catre studenti vor fi predate profesorilor prin e-mail la urmatoarele adrese: ds_angi@yahoo.com

Materiale bibliografice obligatorii:
1. Babbie, Earl. Learning from the field: A guide from experience. London: Sage Publication. 1984.
2. Babbie, Earl. Survey Research Methods, 2nd ed. Belmont, CA: Wadsworth Publishing Co. 1990
3. Dalton, Rusell. Citizen Politics: Public Opinion and Political Parties in Advanced Western Democracies. Chatham House Publishers. 1988.
4. Johnson J., Joslyn R., Political science research methods, 1991
5. King, G., R. Keohane, S. Verba, Designing Social Inquiry, 1994
6. Nachmias C., Nachmias D., Research methods in the social sciences, 1996
7. Rotariu T., Metode si tehnici de cercetare in stiintele sociale
8. Rotariu T., Petru Ilut, Ancheta sociologica, Polirom, 1997
9. Rotariu T. (coord.). Metode statistice aplicate in stiintele sociale. Polirom, 2000
10.White L., Political Analysis. Technique and Practice, 1994

Materiale şi instrumente necesare pentru curs
Pentru unele dintre aplicatii va fi folosit calculatorul.

Calendar al cursului

Tema 1

Argumentare in stiintele sociale. Cercetare in stiintele politice vs. cercetare in alte stiinte sociale. Metode cantitative vs. metode calitative. Inferente descriptive si inferente cauzale.

Bibliografie:

King, R. Strategia cercetarii. Polirom. 2005. Cap. 1 si Cap. 2

sau

White, L. Political analysis. Technique and Practice. Cap. 2

Tema 2

Ipoteze de cercetare. Operationalizarea conceptelor. Variabile. Scale de masura. Indicatori multipli. Eroare de masurare.

Bibliografie:

King, R. Strategia cercetarii. p. 197-205

Rotariu & al. Metode statistice. Polirom. 1999. Cap. 2

Bibliografie optionala:

Culic, Irina. Metode avansate in cercetarea sociala. Polirom. 2005. p. 29-50

Tema 3

Statistica descriptiva. Indicatori ai tendintei centrale: media, mediana, modus. Indicatori de imprastiere: interval intercuartilic, abatere standar. Grafice: histograme, bar-charts, scatterplots.

Bibliografie:

Rotariu & al. Metode statistice. Cap. 16

Tema 4

Relatii intre variabile. Asociere/corelatie si cauzalitate. Relatii false (spurious relations). Modalitati de reprezentare grafica a asocierii.

Bibliografie:

King, R. Strategia cercetarii. Polirom. 2005. p. 71-84

Tema 5

Tabele de asociere cu doua dimensiuni. Indicatori de asociere pentru variabilele de tip nominal.

Rotariu & al. Metode statistice. p. 119-152

http://www.policy.hu/badescu/handbook.zip Cap. 6

Tema 6

Tabele de asociere cu doua dimensiuni. Indicatori de asociere pentru variabilele de ordinal si

de tip cantitativ.

Rotariu & al. Metode statistice. p. 119-152

http://www.policy.hu/badescu/handbook.zip Cap. 6

Tema 7

Analiza multivariata. Tabele de asociere cu mai mult de doua dimensiuni.

Bibliografie:

Rotariu & al. Metode statistice. p. 153-164

Tema 8
Studiul relatiilor între variabile cantitative. Corelatia. Regresia liniara.
Bibliografie:
http://www.policy.hu/badescu/handbook.zip Cap. 7
Rotariu & al. Metode statistice. Cap. 8

Tema 9
Regresia liniara (2).
Bibliografie:
http://www.policy.hu/badescu/handbook.zip Cap. 7
Rotariu & al. Metode statistice. Cap. 8

Tema 10
Teoria testarii. Elemente de baza ale testarii semnificatiei statistice.
Bibliografie:
King, R. Strategia cercetarii. p. 255-271
http://www.policy.hu/badescu/handbook.zip Cap. 5

Tema 11
Scrierea unui proiect de cercetare. Scrierea unui raport de cercetare.
Seminar:
Scrierea unui proiect de cercetare (1).
Bibliografie:
White, L. Political analysis. Technique and Practice. Cap. 14

Politica de evaluare şi notare:
Examen final: 100%.

Elemente de deontologie academica

Notiunea de *plagiat* se defineste în conformitate cu normele deontologice definite la link-ul demai jos http://fspac.ubbcluj.ro/resurse/formulare-regulamente/reguli-etice-si-deontologice/

Frauda la examenul final se pedepseste cu eliminarea de la examen.

Studenti cu dizabilitati

In cazul unor studenti cu dizabilitati motorii sau intelectuale pot fi contactat pe adresa de e-mail in vederea gasirii unei solutii in vederea oferirii de sanse egale acestora.

Modul 1

Obiectiv: Prezentarea etapelor unui proces de cercetare în ştiinţele sociale.

Ghid de studiu:

- Organizarea cercetării
- Exemple de cercetări în ştiinţele politice. Etapele unui proces de cercetare.
- Culegerea datelor
- Analiza şi interpretarea rezultatelor

Unitatea 1

Obiectiv: Detalierea obiectivelor propuse în acest modul. Prezentarea etapelor unui proces de cercetare în ştiinţele sociale.

Noţiuni cheie: teorie, ipoteza, design de cercetare, date si tipuri de date.

Etapele unui proces de cercetare.

Organizarea cercetării

Crearea teoriei

Scopul principal al acestui capitol este de a aduce în discuţie o serie de elemente necesare configurării celei mai importante părţi a unui proiect de cercetare: teoria care stă la baza abordării, asumpţiilor şi presupoziţiilor făcute; în funcţie de construcţiile teoretice de la care se pleacă sunt construite ipotezele, este aleasă abordarea şi metodele de culegere şi analiză a datelor şi sunt prezentate rezultatele la care se ajunge. Elementele care alcătuiesc o teorie sunt: conceptele, categoriile şi propoziţiile (Corbin şi Strauss 1990, p.7).

Conceptele sunt unităţile de bază ale analizei sau abordării; de la modul în care sunt conceptualizate datele, şi nu de la datele în sine, este dezvoltată o teorie. "Teoriile nu pot fi construite pornind de la evenimente actuale sau de la activităţi observate sau relatate, adică din "date brute". Circumstanţele, evenimentele, faptele sunt luate ca şi, sau analizate ca şi, indicatori potenţiali ai fenomenelor, fenomene cărora le sunt atribuite astfel etichete" (Corbin şi Strauss 1990, p.7). Spre exemplu, dacă un respondent afirmă că face parte din sindicatul instituţiei în care îşi desfăşoară activitatea profesională, atunci acesta poate fi etichetat ca fiind "membru al sindicatelor" şi în analiză noastră, atunci când ne vom referi la el, îl vom desemna folosind eticheta şi nu prin descrierea activităţii lui zilnice de a participa la activitatea unui sindicat, adică nu referindu-ne la evenimentele sau faptele observate. Conceptele sunt construite prin compararea faptelor brute şi desemnarea cu aceeaşi etichetă a faptelor asemănătoare.

Al doilea element important al unei construcţii teoretice îl constituie categoriile. Categoriile au un nivel mai ridicat şi sunt mai abstracte decât conceptele pe care le reprezintă (Corbin and Strauss 1990, p.7). Noile elemente teoretice sunt generate printr-un proces analitic similar celui prin care sunt generate conceptele: realizarea comparaţiilor şi evidenţierea similarităţilor şi diferenţelor. Pentru a ilustra modul în care conceptele sunt grupate pentru a forma categoriile vom continua exemplul de mai sus. Astfel, pe lângă cei care fac parte din sindicatul instituţiei în care îşi desfăşoară activitatea profesională, şi pe care i-am etichetat ca fiind "membri al sindicatelor", vom identifica alţi indivizi participă la activităţi ale partidelor politice, ale asociaţiilor non-guvernamentale, etc, iar aceştia vor fi etichetaţi "membrii ai partidelor politice" respectiv "membrii ai asociaţiilor non-guvernamentale". Deşi conceptele amintite sunt diferite în ceea ce priveşte forma, ele reprezintă activităţi legate de acelaşi proces şi pot fi grupate într-o categorie etichetată "cei care iau parte la activităţi participative".

Al treilea element al teoriei sunt propoziţiile care pun în evidenţă relaţii între categorii şi concepte sau între categorii diferite. Propoziţiile sunt adeseori desemnate cu eticheta de "ipoteze" (Glaser şi Strauss 1967). Termenul de "ipoteză" este însă considerat mai puţin adecvat întrucât aceasta implică relaţii care pot fi "măsurate" între concepte şi categorii, ceea ce nu se întâmplă întotdeauna – spre exemplu cazul unor abordări calitative (Whetten 1989, p. 492).

Formarea şi dezvoltarea conceptelor, categoriilor şi propoziţiilor este un proces continuu si mereu reluat / reînceput. Teoria nu este generată a priori şi ulterior testată, ci mai degrabă este "derivată inductiv din studierea fenomenelor pe care aceasta o reprezintă" (Strauss şi Corbin, 1990, p. 23). Teoria este descoperită, dezvoltată şi verificată prin colectare sistematică a datelor şi analiza acelor date care sunt legate de fenomenele studiate.

În procesul de creare a teoriei literatura de specialitate aminteşte patru etape analitice, etape care nu sunt strict secvenţiale: design-ul de cercetare, culegerea datelor, analiza datelor şi compararea rezultatelor obţinute cu rezultate similare din literatura de specialitate.

Design-ul cercetării

Design-ul cercetării, este definit ca fiind "configurarea generală a unei fragment de cercetare" (Easterby-Smith et al. 1990, p. 21) configurare care conţine în general referiri la: datele sau informaţiile care urmează a fi colectate şi la modul în care aceastea urmează a fi analizate pentru a răspunde la întrebările sau cerinţele de bază ale cercetării. Rezultă de aicia că primul pas în construcţia unui design de cercetare îl constituie definirea sau formularea întrebărilor la care urmează a se răspunde în cercetare. Acestea trebuie formulate suficient de restrâns ţinând cont de faptul că design-ul cercetării de obicei este acea parte a unei cercetări în care sunt anunţate intenţiile de a cerceta o anumită problemă şi nu e o cercetare dusă până la rezultatele finale, dar pe de altă parte acestea trebuie formulate suficient de larg pentru a permite o anumită flexibilitate necesară în cazul analizelor în ştiinţele sociale unde fenomenele studiate sunt în continuă evoluţie. O sursă importantă de "întrebări" o constituie literatura de specialitate (spre exemplu: rapoarte ale unor studii, înscrisuri cu conţinut specific diferitelor domenii studiate, etc).

Design-ul proiectelor de cercetare în ştiinţele sociale este destul de variat, depinzând de paradigma care stă la baza cercetării, de metodele utilizate pentru culegerea şi analiza datelor, şi de asumpţiile de la care porneşte cercetătorul în abordarea problematicii care urmează a fi cercetate.

În general, o cercetare în ştiinţele sociale încearcă să descrie şi / sau să interpreteze un anumit fenomen uman, cel mai adesea pornind de la comportamente ale indivizilor sau de la relatări ale acestora cu privire la comportamente adoptate în diferite situaţii. Date fiind varietatea interacţiunilor umane şi dinamica acestora, în construcţia design-ului de cercetare trebuie ţinut cont de distorsiunile care pot apărea, de presupoziţiile făcute şi de interpretările care se dau diferitelor comportamente analizate astfel încât cititorii să poată înţelege şi interpreta rezultatele la care ajunge

cercetarea. Aşa cum ne putem da seama din aceste problematici, nu există o configurare standard a proiectelor sau a rapoartelor de cercetare.

În cele ce urmează vom prezenta structura unui proiect de cercetare, care însă nu are pretenţia de a fi completă sau exhaustivă – cerinţă oricum greu de îndeplinit dată fiind, aşa cum am amintit şi mai sus, varietatea subiectului analizat şi a constrângerilor care trebuie avute în vedere în analiza acestuia - ci mai degrabă încearcă să fie un punct de plecare pentru pentru cercetătorii care încearcă să se decida asupra unei modalităţi de organizare a datelor şi de comunicare a ideilor. In funcţie de subiectul abordat, de datele disponibile cu privire la acesta, de metoda de cercetare utilizată şi de teoria de la care se porneşte, cerinţele enumerate mai jos sunt sau nu sunt prezente în structura unui design de cercetare particular.

Structura unei cercetări

1. Introducere
   - Porniţi la drum cu un citat sau cu o scurtă povestire care să capteze atenţia cititorului. Incercaţi să găsiţi un citat sau o povestire care să aiba legătură cu subiectul abordat, fie cu modalitatea de a pune problema, fie cu rezultatele la care se va ajunge.
   - Formulaţi propriile dumneavoastră întrebări sau nelămuriri cu privire la problematica abordată, descrieţi contextul în care aceste întrebări sau nelămuriri au apărut si cum au evaluat. Ce aţi dori să ştiţi sau sa vă lămuriţi? Cum aţi ajuns să fiţi interesat de problemă?
   - Amintiţi şi alti cercetători care consideră că este necesară o abordare a tematicii avute în vedere, prezentaţi rezultatele la care au ajuns aceştia, sau, dacă este cazul, atrageţi atenţia asupra faptului că o astfel de tematică nu trebuie ignorată.
   - Justificaţi alegerea făcută. De ce este importantă o abordare a fenomenului respectiv în momentul de faţă (ex: este un fenomen care se manifestă pentru prima dată într-o anumită societate sau într-un anumit context, fenomenul a dobândit o anumită amploare, etc).
   - Specificaţi ceea ce urmăriţi în cercetarea dumneavoastră (ex: lărgirea bazei de cunoaştere, deschiderea unor noi perspective de abordare, confirmarea unor rezultate anterioare, verificarea unor asumpţii, etc).
   - Descrieţi publicul căruia vă adresaţi.

2. Paradigma care stă la baza abordării
   - Această secţiune este necesară mai ales atunci când tematica abordată nu este suficient de bine cunoscută de publicul căruia vă adresaţi sau atunci când, indiferent de public, fie tematica, fie abordarea, fie amândouă sunt noi.
   - Prezentaţi propria paradigmă şi încercaţi să o înscrieţi într-o anumită tendinţă de abordare (ex: fenomenologică, hermeneutică, etc). Amintiţi alţi cercetători care au definit paradigme asemănătoare în alte domenii ale ştiinţelor sociale. (Guba, E. 1990).
   - Prezentaţi şi explicaţi asumpţiile şi presupoziţiile pe care le formulaţi în legătură cu subiectul abordat. Explicaţi modul în care acestea pot distorsiona rezultatele la care se va ajunge.

- Daca este cazul, mai ales pentru abordărilor calitative, prezentaţi poziţia pe care se plasează cercetătorul în raport cu subiectul cercetat, spre exemplu: cercetător ca şi membru complet, cercetător ca şi membru activ, cercetător ca şi membru periferic (Adler si Adler, 1994).
- Specificaţi criteriile adecvate pentru evaluarea rezultatelor cercetării. (Atkinson, Heath, şi Chenail, 1991).
- Discutaţi modul în care experienţa dumneavoastră anterioară influenţează modul în care concepeţi abordarea subiectului studiat. Prezentaţi pe scurt experienţă profesională care vă apropie de tematică.

3. Metoda de cercetare
   - Identificati şi descrieţi metoda pe care urmează să o utilizaţi (ex: analiză de caz; metoda comparativă, metoda etnografică, observaţie, experiment, etc.). Descrieţi modul în care alţi autori au utilizat metoda avută în vedere de dumneavoastră (Glaser, B., şi Strauss, A. 1967).
   - Descrieţi în detaliu ceea ce urmează să faceţi. Prezentaţi modalitatea de selectare a subiecţilor de la care vor fi culese informaţiile necesare cercetării.
   - Descrieţi datele pe care intentinaţi să le culegeţi sau pe care intenţionaţi să le utilizaţi si procedura de culegere a acestora (ex: baze de date statistice, note de teren, date provenite din examinarea unor documente, benzi audio sau video, etc). Dacă sunt utilizate interviuri (cum este cazul interviului individual, a celui de grup sau a anchetei, etc) prezentaţi întrebările folosite (fie în context, fie ataşate într-un appendix).
   - Descrieţi procedurile de culegere şi analiză a datelor în ordinea cronologică a desfăşurării lor.
   - Descrieţi procedurile de analiză pe care inteţionaţi să le utilizaţi (codarea datelor, sortarea datelor, procedurile statistice cu ajutorul cărora sunt puse in evidenţă relaţiile existe între date, etc). Prezentaţi, dacă este cazul, programele statistice utilizate pentru modelarea datelor.
   - Interpretaţi rezultatele obţinute în funcţie de teoria, asumpţiile şi presupoziţiile formulate la începutul cercetării.

4. Concluzii
   - Reluaţi pe scurt problematica de la care s-a pornit. Amintiţi asumpţiile şi presupoziţiile făcute, metoda de cercetare şi rezultatele la care s-a ajuns.
   - Stabiliţi legăturile existente între rezultatele cercetării dumneavoastră şi literatura de specialitate. care prezintă rezultate similare sau asemănătoare.
   - Imaginaţi modul în care design-ul cercetării poate evolua de la rezultatele pe care le-aţi obţinut şi ţinând cont de evoluţiile ulterioare ale domeniului studiat. Specificaţi "deschiderile" lăsate de proiectul dumneavoastră şi posibillee modalităţi de valorificare a informaţiei acumulate ulterior.
   - Discutaţi "validitatea" şi "fidelitatea" procedurilor utilizate în culegerea şi analiza datelor.
   - Discutaţi posibilelele distorsiuni generate fie de asumpţiile şi de presupoziţiile făcute, fie de metodele de culegere şi de analiză a datelor.
   - Prezentaţi modul în care literatura de specialitate v-a influeţat în modalităţile de abordare a subiectului cercetat.
   - Discutaţi limitele cercetării dumneavoastră şi amintiţi limitele cu care se confruntă orice subiect asemănător abordat şi în alte studii.

Culegerea datelor

O dată formulate întrebările la care se intenţionează a se răspunde prin cercetare, următorul pas este alegerea "cazurilor" sau a "indivizilor" care urmează a fi investigaţi pentru a obţine datele necesare confirmării sau infirmării propoziţiilor referitoare la problematica cercetată. In alegerea "cazurilor" noastre putem avea un plan prestabilit, aşa cum este cazul în cercetările cantitative, sau putem să ne selectăm cazurile pe măsură ce procesul de cercetare evoluează, aşa cum este cazul în cele mai multe din cercetările calitative (Strauss şi Corbin, 1990, p. 192).

In faza iniţială de culegere a datelor, atunci când sunt stabilite categoriile este necesară o investigare extinsă şi în profunzime a "cazurilor" pentru a obţine date cât mai adecvate cu putinţă. Atunci când "cazurile" sunt foarte multe şi nu pot fi investigate toate se alege o procedură de eşantionare a populaţiei noastre de "cazuri". Pentru a stabili cât de multe "cazuri" vor intra în atenţia analizei noastre, cu alte cuvinte pentru a stabili unde ne oprim cu eşantionarea din punct de vedere teoretic, este nevoie să recurgem la teorie şi la logica cercetării noastre. Ne oprim cu eşantionarea acolo unde nu mai este identificată informaţie suplimentară cu ajutorul cărei cercetătorul să dezvolte proprietăţi sau caracteristici ale conceptelor sau categoriilor cu care lucrează (Glaser şi Strauss 1967, p. 65). In alegerea cazurilor noastre trebuie ţinut seama de faptul că nu toate cazurile au aceeaşi relevanţă raportat la problematica cercetată şi la teoria pe care se bazează cercetarea; astfel, în unele situaţii, este suficient un "caz" pentru a pune în evidenţă o anumită problematică, în alte situaţii este nevoie de mai multe "cazuri" pentru a face acelaşi lucru. Ca şi regulă generală, alegem atâtea "cazuri " câte sunt necesare pentru a avea o imagine completă a problemei cercetate din perspectivele relevante pentru cercetarea noastră. Adaugarea unui nou "caz" trebuie să servească unor scopuri specifice ale cercetării (Yin 1989, p. 53-54), iar aceste scopuri specifice pot fi: a) identificarea unor concepte şi categorii, b) alegerea unui "caz" pentru a reproduce rezultatele obţinute în alt "caz", c) alegerea unui caz opus celui sau celor studiate până în acel moment.

Pentru o cât mai bună "acoperire" a unui fenomen sau fapt social este necesară utilizarea unor surse multiple de date. Nu există o singură categorie de date sau o singură tehnică de culegere a datelor care poate fi etichetată ca "adecvată" (Glaser şi Strauss 1967, p. 65). Diferite surse de date oferă cercetătorului perspective diferite asupra fenomenului studiat. Iniţial abordarea unui fenomen poate avea la bază o singură tehnică de culegere a datelor, ulterior însă este recomandată identificarea şi altor surse de date şi a altor tehnici de investigare a acestor date. Utilizarea unor surse multiple de date consolidează validitatea abordării.

Analiza datelor

Analiza datelor reprezintă etapa cea mai importantă în dezvoltarea, confirmarea, extinderea sau reproducerea unei teorii. Această analiză, pentru fiecare caz particular, implică în primul rând generarea unor concepte printr-un proces de observare a realităţii, de descompunere a fenomenelor în elementele lor componente şi reasamblarea lor în modalităţi noi (Strauss şi Corbin, 1990). Analiza datelor este în literatura de specialitate subiect al unei vii dispute între cei care consideră că aceasta trebuie făcută prin metode cantitative şi cei susţin abordările calitative.

Compararea rezultatelor obţinute cu rezultatele din literatura de specialitate.

O dată datele culese, analizate şi interpretate problema care se pune este aceea de a compara rezultatele obţinute literatura de specialitate existentă şi examinarea a ceea ce este similar şi a ceea ce este diferit. Compararea unei teorii noi cu ceea ce deja există va consolida validitatea internă, va

consolida de asemenea gradul de generalizare al rezultatelor obţinute pentru cazurile studiate (Eisenhardt, 1989, p. 545).

Bibliografie:

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. Qualitative Sociology, 13, 3-21.

Easterby-Smith, M., Thorpe, R., & Lowe, A. (1991). Management research: An introduction. London: Sage.

Eisenhardt, K. M. (1989). Building theories from case study research. Academy of Management Review, 14, 532-550.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory. Chicago: Aldine.

Lee, R. M., & Fielding, N. G. (1991). Computing for qualitative research: Options, problems and potential. In N. G. Fielding & R. M. Lee. (Eds.), Using computers in qualitative research (pp. 1-13). London: Sage.

Martin, P. Y., & Turner, B. A. (1986). Grounded theory and organisational research. Journal of Applied Behavioural Science, 22, 141-157.

Muhr, T. (1993) ATLAS Release 1.1E Users Manual. Berlin: Technical University of Berlin.

Pandit, N. R. (1995). Towards a grounded theory of corporate turnaround: A case study approach. Unpublished doctoral thesis, University of Manchester, UK.

Pettigrew, A. M. (1987). Researching strategic change. In A. M. Pettigrew (Ed.), The management of strategic change (pp. 1-14). Oxford: Blackwell.

PROMT users manual. (1989). Cleveland, OH: Predicasts.

Strauss, A. & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. London: Sage.

Tesch, R. (1991). Software for qualitative researchers: Analysis needs and program capabilities. In N. G. Fielding & R. M. Lee (Eds.), Using computers in qualitative research (pp. 16-37). London: Sage.

Textline reference guide. (1993). London: Reuters.

Turner, B. A. (1983). The use of grounded theory for the qualitative analysis of organisational behaviour. Journal of Management Studies, 20, 333-348.

Whetten, D. A. (1989). What constitutes a theoretical contribution? Academy of Management Review, 14, 490-495.

Yin, R. K. (1989). Case study research: Design and methods. London: Sage.

Modulul 2

Obiective: Introducerea notiunii de eşantionare şi a problematicii eşantionării

Ghid de studiu:

- Cercetări selective: de la populaţie la eşantion
- Reprezentativitatea eşantioanelor: a alege câţiva pentru a îi reprezenta pe toţi.
- Proceduri de eşantionare. Tipuri de eşantioane
- Aspecte matematice ale eşantionării. Teste de semnificaţie

Unitatea 1

Obiectiv: Prezentarea noţiunii de eşantionare

Cuvinte cheie: populaţie, eşantion, cadru de eşantionare, populaţie ideală

Eşantionarea

Cercetări selective: de la populaţie la eşantion

Cine va câştiga alegerile prezidenţiale sau parlamentare? Sunt femeile o minoritate defavorizată în societăţile moderne? O politică publică sau o decizie administrativă produce modificări ale comportamentelor indivizilor vizaţi de acea politică publică sau de acea decizie? Cine este pentru şi cine este împotriva introducerii unor noi măsuri fiscale? Cât de populară este măsura luată de autorităţile dintr-o anumită unitate administrativă de a construi o nouă zonă industruială? Toate aceste întrebări au în comun o caracteristică importantă şi anume: se referă la populaţii atât de largi încât este practic imposibil de obţinut informaţii cu privire la toate elementele care le compun. Cu situaţii asemanătoare - imposibilitatea cuprinderii tuturor elementelor care compun un întreg - se confruntă şi medicul care face analize de sânge şi care nu poate extrage tot sângele aflat în organismul unui pacient pentru a îl supune unei investigaţii în laborator, cei care fac analize ale unor elemente din mediul natural pentru a stabili nivelul de poluare, sau cercetătorul din ştiintele naturale care taie un exemplar dintr-o specie de plante în scopul efectuării unor analize în laborator.

Atât în aceste situaţii, cât şi în multe altele de acest fel, problema care se pune este aceea de a culege informaţiile necesare pentru a analiza temele avute în vedere doar de la o parte din indivizii care compun o populaţie şi nu de la întreaga populaţie. Din punct de vedere tehnic, grupul sau mulţimea de indivizi care constituie obiectul de studiu sau de interes al cercetătorului la un moment dat este denumit *populaţie,* iar grupul mai mic de indivizi de la care sunt culese informaţiile necesare cercetării este denumit *eşantion.* "Setul de operaţii cu ajutorul cărora, din ansamblul *populaţiei* vizate de cercetare, se extrage o parte, numită *eşantion*, parte ce va fi supusă nemijlocit investigaţiei"[2] este desemnat ca fiind operaţia de *eşantionare*.

Decizia de a culege datele necesare unei cercetări de la un eşantion sau de la o populaţie depinde de o serie de aspecte practice. Astfel, în unele situaţii, dacă timpul, resursele financiare şi

[2] Traian Rotariu, Petre Iluţ, Ancheta sociologică şi sondajul de opinie, Ed. Polirom, Iaşi, 1997, p.122.

umane nu constituie o problemă sau dacă populaţia ţintă nu este foarte numeroasă, atunci este multe mai avantajoasă culegerea datelor de la toţi indivizii care compun o populaţie vizată; în felul acesta se obţine o imagine exactă a problematicii investigate. În alte situaţii există o serie de constrângeri care îl impiedică pe cercetător să ajungă la toţi indivizii care compun o populaţie, aceste constrângeri se referă în primul rând la timp, resursele financiare şi umane aflate la dispoziţie, dispersarea georgrafică a populaţiei care urmează a fi cercetată, iar soluţia cea mai la îndemână pentru a culege informaţiile necesare constă selectarea unui eşantion şi investigarea indivizilor care îl compun. Din acest punct de vedere am putea spune ca eşantionarea este un compromis datorat insuficienţei resurselor. Nu întotdeauna este însă vorba numai de imposibilitatea fizică de a culege informaţii de la toţi membrii unei populaţii – neajuns care în unele situaţii poate fi depăşit – ci şi de o lipsă de eficienţă practică – spre exemplu, în cazul cercetătorului din ştiinţele naturale, care, dacă ar tăia toate exemplarele unei specii de plante pentru a le analiza în laborator ar determina dispariţia speciei respective. Pe de altă parte, concentrând resursele existente doar pentru analiza unei părţi dintr-un întreg se pot obţine rezultate mai bune decât analizând întregul, mai ales atunci când acest întreg este format din mulţi indivizi a căror investigare implică utilizarea unui personal auxiliar numeros care datorită lipsei de specializare poate genera erori mai grave decât dacă ar fi analizată o parte din acel întreg utilizând un personal specializat.

Unul dintre primele aspecte care trebuie luate în considerare atunci când se pune problema realizării unor cercetări practice este aceea a delimitării populaţiei care urmază a fi studiată. În acest context, prin "populaţie" sunt desemnate toate elementele care pot sau trebuie să fie studiate. Elementele pot fi indivizi umani, dar în acelaşi timp pot fi gospodării, şcoli, spitale, intreprinderi economice, oraşe, organizaţii sociale sau profesionale, ziare, articole de presă, discursuri ale unor oameni politici, etc. Indiferent însă de cine sau ce constituie elementele populaţiei vizate, aceasta trebuie să fie atent delimitată în funcţie de obiectivele cercetării, întrucât rezultatele finale vor depinde de acest punct de referinţă stabilit iniţial. Spre exemplu, să ne imaginăm că într-un oraş se pune problema adoptării unui nou sistem de transport în comun, iar ceea ce ne interesează este acordul sau dezacordul cetăţenilor cu privire la modul practic de realizare a acestuia. În acest caz, populaţia vizată este compusă doar din cei care locuiesc în oraşul respectiv? sau trebuie avută în vedere şi populaţia care nu locuieşte în oraş, dar care într-o măsură sau alta beneficiază de transportul în comun din acel oraş? care este vârsta minimă şi maximă a celor care vor fi chestionaţi?, etc. Intrebări asemănătoare acestora se ivesc de fiecare dată când se pune problema definirii unei populaţii care urmează a fi cercetate, iar de răspunsurile date vor depinde în mod evident rezultatele finale. Un prim pas în definirea unei populaţii vizate este acela de a stabili o *populaţie ideală*, adică toţi cei care ipotetic ar trebui să fie luaţi în considerare atunci când se cercetează o problematică anume. După care în funcţie de constrângerile practice identificate – spre exemplu, în cazul în care elementele populaţiei sunt indivizi umani, astfel de constrângeri ar putea fi date de imposibilitatea de a îi investiga pe cei aflaţi în închisori, unităţi militare, spitale, hoteluri, în străinătate, etc – populaţia ideală poate fi restrânsă la o populaţie vizată care poate fi abordată în cadrul cercetării. Avantajele luării în considerare în faza iniţială a unei populaţii ideale este acela că excluderea unor segmente din aceasta este explicită, iar neajunsurile rezultate de aici pot fi luate în considerare.

O dată stabilită populaţia vizată, poate fi pusă şi problema alegerii unui eşantion. Pentru aceasta, elementele populaţiei vizate sunt traspuse într-o listă numită *cadru de eşantionare*, listă din care vor fi extrase ulterior potrivit unor proceduri clar definite acele elemente care vor compune eşantionul. Spre exemplu, dacă se realizează o anchetă telefonică al cărui scop este investigarea modului în care dotarea cu utilităţi publice a unei localităţi acoperă necesităţile existente, populaţia ideală este constituită din toate gospodăriile care au acces la utilităţi publice, iar cadrul de eşantionare este format din toate gospodăriile care au acces la utilităţi publice şi au telefon. Constrângerea în acest caz este dată de existenţa unui post telefonic în gospodărie. Gospodăriile

care au acces la utilităţi dar care nu au telefon neputând fi investigate, populaţia vizată este formată doar din acele gospodării care au acces la utilităţi publice şi au telefon (Figura 1.). În exemplul de faţă, astfel de liste care să se constituie în cadru de eşantionare pot fi evidenţe ale companiilor funizoare de utilităţi publice şi liste ale abonaţilor la servicii telefonice din localitatea avută în vedere.

De la caz la caz, în funcţie de problema investigată, pot constitui cadru de eşantionare: lista celor care sunt înscrişi la un medic de familie sau la medicii de familie care operează într-o anumită arie care urmează a fi acoperită de cercetare, lista celor abonaţi la o firmă furnizioare de servicii de televiziune prin cablu, lista celor abonaţi la o anumită publicaţie, lista celor care figurează în registrul auto, lista celor care figurează în registrul de carte funciară, etc. Ideal toate aceste liste ar trebui să includă fiecare element al populaţiei vizate doar o singură dată. În realitate însă există o serie de neajunsuri printre care cele mai importante sunt:

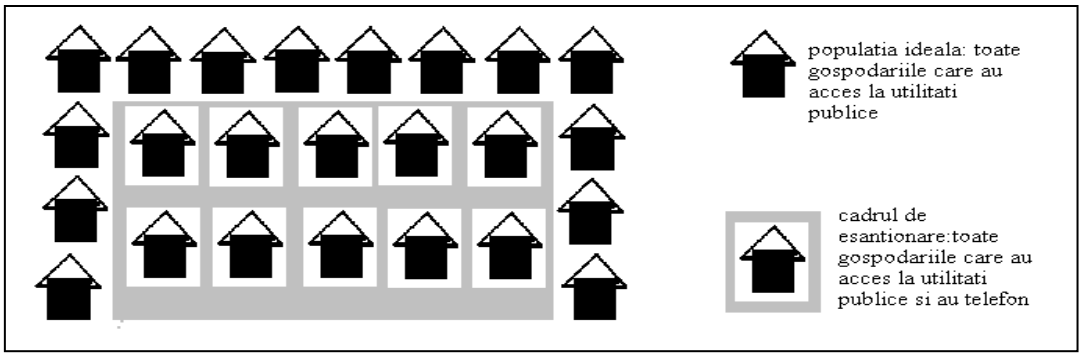
- lipsa unor elemente: fie lista este *inadecvată* în sensul în care iniţial nu a fost concepută pentru a include toate elementele care pot face la un moment dat obiectul de interes al cercetătorului, fie este *incompletă*, adică nu include din diferite motive toate elementele care se presupune care că ar trebui să le includă;
- referinţe la grupuri de elemente şi nu la elemente individuale; spre exemplu, liste care nu se referă la numărul de persoane ci la numărul de familii care locuiesc într-o gospodărie, dar cercetarea vizează persoane şi nu familii;
- existenţa unor elemente straine, adică existenţa în listă a unor elemente care din diferite motive nu fac obiectul de studiu la un moment dat;
- exitenţa unor duplicate: când unele elemente ale populaţiei apar de mai multe ori pe o listă.

Remedierea tuturor acestor neajunsuri va duce în mod evident la obţinerea unui eşantion mai bun. De la caz la caz prin remediere se înţelege: identificarea elementelor lipsă şi introducerea lor în lista care constituie cadrul de eşantionare, identificarea tuturor elementelor care fac parte dintr-un grup, eliminarea unor elemente străine care nu au legătură cu tematica cercetării, eliminarea duplicatelor şi păstrarea pe o listă a unui element doar o singura dată.

O modalitate de a depăşi aceste neajunsuri constă în redefinirea problematicii cercetate în aşa fel încât elementele populaţiei care nu pot fi identificate să nu facă obiectul unei anumite cercetări. Acest lucru evident nu este posibil în toate situaţiile şi nu este posibil mai ales în acele situaţii în care elementele care nu pot fi identificate constituie majoritatea elementelor unei populaţii.

Atunci când nu există liste care să cuprindă elementele unei populaţii vizate prin cadru de eşantionare poate fi desemnată orice altă procedură care să permită idetificarea elementelor unei populaţii. Spre exemplu, o arie geografică poate juca rolul de cadru de eşantionare, situaţie în care elementele populaţiei vizate sunt asociate cu un anumit spaţiu natural. Astfel, aria geografică ocupată de o populaţie vizată poate fi împărţită în zone mai mici din care sunt alese aleator câteva, care la rândul lor sunt divizate în arii mai mici dintre care vor fi selectate aleator câteva şi aşa mai departe până la utimul stadiu când din anumite zone astfel selectate sunt investigate toate elementele.

Figura 1. Cadrul de eşantionare pentru selectarea unui eşantion în vederea investigării printr-o anchetă telefonică a gradului de satisfacere de către utilităţile publice a nevoilor populaţiei unei localităţi (exemplu ipotetic).



Obiectiv: Prezentarea problematicii reprezentativităţii eşantioanelor

Reprezentativitatea eşantioanelor: a alege câţiva pentru a îi reprezenta pe toţi.

Un eşantion "bun" este într-o oarecare măsură o versiune în miniatură a unei populaţii, un model al unei populaţii. Caracteristica cea mai importantă a unui eşantion bun este dat de reprezentativitatea acestuia. Un eşantion este considerat reprezentetativ pentru populaţia din care este extras dacă "caracteristici importante sunt distribuite similar în amândouă grupurile"[3] sau cu alte cuvinte, ţinând cont de ordinea temporală a constituirii celor două grupuri, un eşantion trebuie să reproducă caracteristici importante ale populaţiei din care este extras. Aceste caracteristici importante pot fi spre exemplu, vârsta, nivelul de educaţie, mediul de reşedinţă, sexul, venitul, etc. Spre exemplu, dacă populaţia vizată este fi constituită în proporţie de 51% din femei, dintre care 27% au studii medii, atunci un eşantion reprezentativ va fi compus în proporţie de 51% din femei dintre care aproximativ 27% vor avea studii medii.

Un eşantion nu va reproduce niciodată cu exactitate toate caracteristicile unei populaţii, ca urmare aproximarea unei caracteristici existente în populaţie recurgând la măsurători efectuate pe un eşantion va produce o anumită *eroare (**d**)*, iar încadrarea rezultatului obţinut într-o marjă de eroare rezonabilă se face cu un anumit *grad sau nivel de probabilitate (**P**).* Eroarea obţinută este rezultatul diferenţei reale existente între o caracteristică **A** dintr-o populaţie şi caracteristica corespunzătoare **A*** măsurată pe un eşantion extras din acea populaţie. Nivelul de probabilitate este măsura în care eroarea pe care o facem aproximând o valoare **A** din populaţie prin valoarea corespunzătoare **A*** măsurată pe un eşantion este mai mică decât o eroare maximă admisă. Reprezentativitatea unui eşantion este exprimată cantitativ de cele două valori **d** şi **P**, valori care sunt determinate una de cealaltă. Un eşantion este cu atât mai reprezentativ cu cât eroarea pe care o facem este mai mică iar nivelul de probabilitate este mai mare.

Indiferent de modul în care selectat un eşantion, acesta reproduce mai mult sau mai puţin caracteristici ale populaţiei din care este extras, motiv pentru care nu există eşantioane nereprezentative, ci doar eşantioane mai mult sau mai puţin reprezentative pentru o populaţie în funcţie de măsura în care caracteristici ale populaţiei respective sunt regăsite şi în acete eşantioane. Astfel, un eşantion care reproduce mai bine caracteristicile unei populaţii decât un alt eşantion, vom spune care mai reprezentativ. Mai mult, unele caracteristici pot fi mai bine reproduse de un eşantion iar altele mai puţin bine, ceea ce înseamnă ca reprezentativitatea unui eşantion este diferită în

[3] Arlene Fink, How to Sample in Surveys, Sage Publications, Thousands Oaks, London, New York, 1995, p.1.

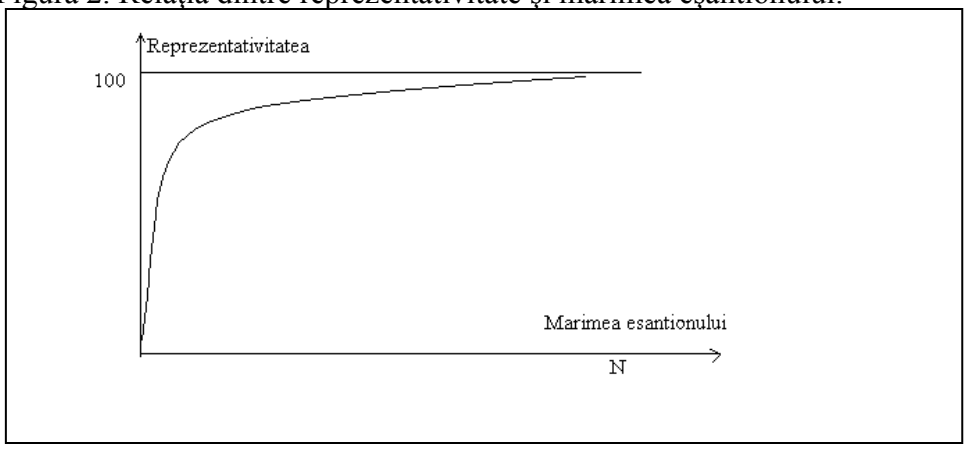
funcţie de caracteristica care este avută în vedere. Cu alte cuvinte, un eşantion nu este reprezentativ în general, ci are o anumită reprezentativitate în raport cu o anumită caracteristică şi o altă reprezentativitate în raport cu o altă caracteristică.

Gradul de reprezentativitate al unui eşantion depinde de trei factori importanti - caracteristicile populaţiei din care este extras, de mărimea eşantionului şi de procedura de eşantionare – factori care au fost sintetizaţi de Rotariu şi Iluţ în lucrarea "Ancheta sociologică şi sondajul de opinie" şi pe care îi voi reaminti în cele ce urmează.

Cum am spus deja reprezentativitatea unui eşantion este dată de capacitatea acestuia de a reproduce o serie de caracterisitici existente în populaţie. Dacă o caracteristică este mai omogen distribuită într-o populaţie un acelaşi eşantion va fi mai reprezentativ pentru acea caracteristică decât pentru o altă caracteristică care este distribuită mai eterogen în aceeaşi populaţie. Sau altfel spus, pentru a obţine o aceeaşi reprezentativitate, pentru o caracteristică în raport cu care populaţia este mai omegenă este nevoie de un eşantion de mărime mai mică decât pentru o caracteristică în raport cu care populaţia este mai eterogenă.

Mărimea eşantionului se referă la numărul de elemente care îl compun şi care trebuie investigate pentru a obţine rezultate cât mai precise. Intuitiv, un eşantion este cu atât mai reprezentativ cu cât cuprinde mai multe elemente din populaţia vizată, în felul acesta obţinându-se o reproducere mai bună a acesteia. Dar creşterea nivelul de reprezentativitate nu este direct proporţională cu creşterea numărului de elemente din populaţia vizată care sunt incluse în eşantion, adică nu avem o relaţie lineară între cele două componente, dimpotrivă această relaţie poate fi reprezentată sub forma unei curbe asemănătoare celei din figura de mai jos (Figura 2.). Astfel, dacă modificăm mărimea eşantionului cu o cantitate K de elemente, iar eşantionul cuprinde iniţial un număr mic de elemente, modificarea gradului de reprezentativitate este mai mare decât dacă modificăm mărimea eşantionului cu aceeaşi cantitate K de elemente dar eşantionul este compus iniţial dintr-un număr mare de elemente.

Figura 2. Relaţia dintre reprezentativitate şi mărimea eşantionului.



Mărimea eşantionului este indepedentă de mărimea populaţiei din care este extras. Un eşantion de o anumită mărime şi constituit după aceleaşi proceduri are aceelaşi grad de reprezentativitate şi atunci când este extras din populaţia unei ţări şi atunci când este extras din populaţia unui oraş. Consecinţa faptului că reprezentativitatea unui eşantion nu depinde de mărimea popolaţie din care este extra este aceea că acesta are un anumit grad de reprezentativitate pentru întreaga populaţie, dar subeşantioanele în care se împarte şi care respectă proporţia diferitelor

segmente ale populaţiei nu mai au acelaşi grad de reprezentativitate pentru aceste segmente ca şi eşantionul iniţial.

În ceea ce priveşte procedura de eşantionare, aceasta influenţează atât gradul de prezentativitate al unui eşantion cât şi posibilitatea exprimării numerice a acesteia. Din punct de vedere tehnic – matematic, calcularea reprezentativităţii unui eşantion este posibilă numai în cazul eşantioanelor probabilistice sau aleatoare. Un eşantion probabilistic este acel eşantion pentru care fiecare element din populaţia vizată are o şansă calculabilă şi nonnulă de a fi selectat în eşantion. Posibilitatea calcularii şansei ca un element din populaţie să fie selectat în eşantion permite calcularea unei marje de eroare (d) şi a unui nivel de probabilitate (P) prin care să fie exprimată cantitativ reprezentativitatea eşantionului. În cazul eşantioanelor neprobabilistice, cele pentru care şansa unui element al populaţiei de a face parte din eşantion nu este cunoscută, nu poate fi calculat gradul de reprezentativitate şi prin urmare nici nu se poate vorbi de reprezentativitatea lor.

Unitatea 3

Obiectiv: Proceduri de eşantionare. Tipuri de eşantioane

Cuvinte cheie: selecţie la întâmplare, eşantioane probabilistice, eşantioane neprobabilistice

Proceduri de eşantionare. Tipuri de eşantioane

Distincţia clasică în ceea ce priveşte tipurile de eşantioane este aceea între eşatioanele probabilistice sau aleatoare şi cele neprobabilistice.

În primul caz în procesul de selectare a unui element din populaţie pentru a face parte din eşantion se presupune că se face "la întâmplare" fără să intervină în vreun un fel subiectivismul celui care aplică procedura de eşantionare şi nici vreun alt fenomen care să afecteze şansa unuia sau unor indivizi de a fi selectaţi. Dată fiind această constrângere, următoarele situaţii: alegerea *la întâmplare* a unui număr de oameni care intră într-o instituţie de la orele 8.00 până la orele 10.00 ale unei zile, constituirea unui eşantion de gospodării alese *la întâmplare* atunci când ne plimbăm pe câteva străzi dintr-o localitate, sau constituirea unui eşantion format din localităţi rurale ale unui judeţ, selectând tot *întâmplător* localităţi rurale care se află pe şoseau care leagă două oraşe ale judeţului respectiv, etc, nu vor duce la constituirea unor eşantioane probabilistice. Motivul pentru care nu vom obţine în aceste cazuri eşantioane aleatoare este acela că în alegerea pe care o facem "la întâmplare" excludem fie intenţionat, fie neintenţionat o parte din elementele populaţiei vizate. Astfel în primul caz, dacă vrem să alegem un eşantion reprezentativ pentru cei care frecventează o anumită instituţie şi vom selecta "la întâmplare" doar pe cei care intră în acea instituţie în intervalul orar amintit îi vom exclude pe toţi cei care la momentul respectiv nu au şansa de a intra în acea instituţie, similar vom exlude fără să vrem gospodăriile care nu au şansa de a se găsi pe străzile pe care ne plimbăm sau localităţile rurale care nu au şansa de a se afla pe şoseaua care leagă cele două oraşe între care ne deplasăm. Mai mult, nu putem calcula şansa pe care fiecare element din cele trei cazuri prezentate mai sus – persoane, gospodării, localităţi rurale – le are de fi selectat în eşantion. În toate aceste cazuri este clar că *întâmplarea* favorizează anumite elemente şi anulează şansele altor elemente care sunt exluse *a priori* întrucât nu au şansa de a se afla la locul sau pe traseul pe care se deplasează cel care face selecţia.

Pentru a evita aceste situaţii ar trebui să avem o situaţie clară a tuturor elementelor care compun o populaţie şi să le putem identifica fără echivoc.

Aşa cum am arătat deja în paragraful anterior, în cazul procedurilor de eşantionare probabilistice fiecare element care compune o populaţie trebuie să aibă o şansă diferită de zero şi

calculabilă de a face parte din eşantion. Acesta este criteriul de bază în stabilirea dihotomiei: eşantioane probabilistice - eşantioane neprobabilistice

De-a lungul timpului au fost dezvoltate o larg varietate de tehnici de eşantionare, dintre acestea cele mai frecvent întâlnite sunt: eşantionarea simplă aleatoare, eşantionarea aleatoare prin stratificare, eşantionarea aleatoare multistadială sau cluster, eşantionarea pe cote, eşantionarea tip "bulgăre de zăpadă". În cele ce urmează vom prezenta o serie de aspecte legate de modul de aplicare al fiecăreia dintre aceste proceduri de eşantionare.

Eşantioane probabilistice

1. Eşantionarea simplă aleatoare

Eşantionarea simplă aleatoare este probabil procedura cea mai importantă şi cea mai des utilizată în domeniul cercetărilor practice şi este considerată procedura de referinţă, "ideală", atunci când se pune problema stabilirii unor tehnici de eşantionare. Asumpţiile de bază ale acestei tehnici sunt acelea că: fiecare element al populaţie vizate are exact aceeaşi şansă ca şi oricare alt element al aceleiaşi populaţii de a fi selectat în eşantion, iar selectarea unui element în eşantion nu a influenţat în nici un fel şansele altui element de a fi selectat. Tehnica tipică sau modelul de realizare al acestui tip de eşantionare este reprezentat de *metoda urnei*, situaţie în care fiecarui element dintr-o populaţie vizată îi corespunde o bilă; toate bilele corespunzătoare unor membrii ai populaţiei vizate sunt introduse într-o urnă după care sunt amestecate şi se extrage pe rând câte una până se ajunge la un număr de bile egal cu numărul de elemente care vor compune eşantionul. Simplu de pus în practică din punct de vedere teoretic, procedura astfel definită întâmpină o serie de dificultăţi mai ales atunci când se lucrează cu populaţii mari, situaţie în care este practic imposibil de conceput o urnă în care să poate fi introdus un număr de bile egal cu numărul de indivizi care compun populaţia unei ţări spre exemplu.

O a doua metodă de realizare a unei eşantionări simple aleatoare sunt *tabelele cu numere aleatoare*. Procedura constă în generarea unor şiruri de numere aleatoare şi introducerea lor într-un tabel similar celui prezentat mai jos (Tabelul 1.). Fiecărui element din populaţia vizată, care trebuie identificat univoc, i se atribuie un număr de la 1 la N. Cel care realizează selecţia, alege la întâmplare un număr din şirul de numere aleatoare şi caută apoi în populaţia vizată elemetul cu numărul de ordine reprezentat de numărul aleator respectiv, element care va face parte din eşantion. După care din tabelul de numere aleatoare este ales numărul următor şi se identifică din nou în populaţia vizată elementul cu numărul de ordine identic cu numărul aleator, element care este şi el introdus în eşantion. Procedura continuă în acest fel până la completarea numărului de elemente necesare constituirii eşantionului. În cazul în care unui număr aleator nu îi corespunde un număr atribuit unui element din populaţia vizată, acest nu este utilizat şi se trece la următorul.

Tabelul 1. Tabel de numere aleatoare (exemplu ipotetic).

67	21	03	17	03	89	73	81
53	32	75	72	77	33	10	01
45	27	41	98	86	05	40	50
76	90	83	78	26	92	77	13
23	07	47	63	19	94	11	43
09	11	23	49	15	28	48	85

Neajunsul acestei metodei tabelelor aleatoare constă în posibilitatea ca un element să fie selectat de mai multe ori în eşantion. Situaţie care este evitată în cazul utilizării metodei urnei, dacă o bilă o dată extrasă nu mai este introdusă înapoi în urnă.

Numerele aleatoare pot fi compuse din start din mai mult de două cifre, sau pot fi compuse, în funcţie de necesităţi, din mai mult de două cifre de către cel care face eşantionarea prin adăugarea la o coloană a câte cifre este nevoie din coloana sau coloanele alăturate. Astfel, în exemplul de mai sus dacă la prima coloana se adaugată prima cifră din coloana a doua se obţin numerele: 672, 533. 452, 769, 230, 091.

O altă procedură de punere în practică a unei eşantionări simple aletoare este cunoscută sub denumirea de *metoda pasului.* În această situaţie este necesară o listă care să cuprindă toate elementele populaţiei vizate, fiecărui element fiindu-i atribuit un număr de la 1 la N. După care se stabileşte un pas de eşantionare, de obicei egal cu raportul dintre mărimea populaţiei (N) şi mărime eşantionului (n): N/n. Se alege la întâmplare un număr din lista care cuprinde toate elementele populaţiei vizate, elementul corespunzător acelui număr fiind primul element al eşantionului, după care începând de la acel element tot al N/n –lea element din populaţie este selectat în eşantion. Pasul de N/n se aplică de câte ori este nevoie pentru a selecta numărul de elemente care trebuie să fac parte din eşantion. Procedura pasului mai este cunoscută şi sub denumirea de *eşantionare simplă sistematică*.

Spre exemplificare, să presupunem că populaţia vizată este formată din 5000 de gospodării, şi dorim să constituim un eşantion format din 250 de gospodării. Pasul de eşantionare în acest caz va fi 5000/250 = 20. Gospodăriile sunt ordonate pe o listă, fiecăreia atribuindu-i-se un număr de la 1 la 5000. Se alege la întâmplare o un număr de pe listă, să spunem că acest număr este 27, iar gospodăria căreia i-a fost atribuit acest număr este primul element al eşantionului nostru. Următoarele gospodării care vor face parte din eşantion sunt cele cărora le corespund numerele: 52, 77, 102, 127, 152, 177 şi aşa mai departe până la selectarea a 250 de gospodării.

2. Eşantionarea prin stratificare

Eşantionarea prin stratificare are la bază tot o procedură de alegere aleatoare. Această metodă este utilizată atunci când cel care face eşantionarea are motive să creadă că populaţia vizată este compusă din mai multe subpopulaţii sau subgrupuri distincte, denumite tehnic straturi. Realizarea din punct de vedere practic a unui eşantion prin stratificare presupune următorii paşi: populaţia vizată este împarţită în subpopulaţii în funcţie de un anumit criteriu care este deja cunoscut, după care este constituit un eşantion care la rândul lui va fi compus din atâtea subeşantioane câte subpopulaţii există în populaţia vizată. În interiorul fiecărei subpopulaţii elementele care vor fi introduse în eşantion sunt selectate aleator.

Spre exemplu, să presupunem că 30% din populaţia unei regiuni locuieşte în localităţi rurale, 20% locuieşte în oraşe cu până la 50.000 de locuitori, 15% locuieşte în oraşe care au de la 50.001 la 100.000 de locuitori, iar restul de 35% locuieşte în oreşe de peste 100.000 de locuitori. Un eşantion stratificat format din 1000 de persoane va cuprinde 300 de persoane din mediul rural, 200 de persoane care locuiesc în oraşe cu pânâ la 50.000 de locuitori, 150 de persoane care locuiesc în oraşe care au între 51.001 şi 100.000 de locuitori şi 350 de persoane care locuiesc în oraşe de peste 100.000 de locuitori.

Principiul de bază al acestui tip de eşantionare este acela că, cu cât o populaţie este mai omogenă cu atât este mai uşor să se extragă din aceasta un eşantion reprezentativ. De asemenea, cu cât o populaţie este mai omogenă în raport cu o caracteristică, cu atât mărimea eşantionului necesar pentru a reproduce cu o anumită acurateţe acea caracteristică este mai mică în comparaţie cu mărimea unui eşantion extras dintr-o populaţie care este mai eterogenă în raport cu aceeaşi caracteristică.

Mărimea subeşantioanelor poate să pătreze proporţia subpopulaţiilor, situaţie în care vom vorbi de eşantionare prin stratificare proporţională. În felul acesta se asigură pentru toate elementele populaţiei vizate o şansă egală de a fi selectate în eşantion.

Există însă şi situaţii în care este recomandat ca subeşantioanele să nu pătreze proporţiile subpopulaţiilor. Acest lucru se întâmplă mai ales atunci când unele subpopulaţii sunt reduse din punct de vedere numeric şi în consecinţă, dacă ar fi păstrate proporţiile, şi subeşantioanele ar fi formate dintr-un număr mic de elemente care nu ar avea un nivel de reprezentativitate rezonabil. În această situaţie se recurge la o stratificare diproporţionată a eşantionului sau o stratificare ponderată, prin suprareprezentarea în eşantion a subpopulaţiilor mai puţin numeroase, urmând ca la prelucrarea datelor aceste "abateri" să fie corectate prin metode statistice. În această situaţie şansele elementelor aparţinând diferitelor subpopulaţii de a intra în eşantion sunt diferite: elementele care provin din subpopulaţiile mai puţin numeroase având şanse mai mari de fi selectaţi în eşantion decât elementele care provin din subpopulaţiile mai numeroase.

Indiferent de modalitatea în care sunt constituite subeşantioanele, păstrând sau nu proporţiile, eşantionarea prin stratificare presupune existenţa în momentul iniţial al punerii în practică a procedurii de eşantionare a unei informaţii suplimentare despre populaţia vizată în comparaţie cu situaţia în care este utilizată eşantionarea simplă aleatoare. Această informaţie poate fi obţinută cu ajutorul altor studii sau din alte surse de informare cu privire la populaţia vizată.

În ceea ce priveşte gradul de reprezentativitate al eşantionelor realizate prin stratificare în comparaţie cu gradul de reprezentativitte al eşantioanelor simple aleatoare, se admite în general că este mai bun. Mai clar spus, dintre două eşantioane de aceeaşi mărime unul obţinut prin eşantionare prin stratificare iar altul prin eşantionare simplă aleatoare, se consideră că primul are o reprezentativitate mai bună, în situaţia în care criteriile pe baza cărora se face eşantionarea au o legătură de tip statistic cu caracteristicile care fac obiectul cercetării.

3. Eşantionarea multistadială

Până acum am prezentat situaţii în care există un anumit cadru de eşantionare - liste care să cuprindă elementele unei populaţii - şi situaţii în care pe lângă faptul că există un anumit cadru de eşantionare cercetătorul mai are la îndemână şi o serie de criterii pe baza cărora o populaţie poate fi împărţită în subpopulaţii sau grupuri. În această din urmă situaţie din fiecare grup este extras un subeşantion care va face parte din eşantionul final.

Există însă şi situaţii în care nu există un cadru de eşantionare şi nici nu este necesară creerea unuia întrucât nu toate elementele acelei populaţii vor fi incluse în eşantion. Dacă populaţia poate fi considerată ca fiind formată din grupuri, iar între aceste grupuri există o anumită asemănare, atunci are sens să nu fie selectaţi în eşantion indivizi din toate grupurile ci numai indivizi din anumite grupuri. Procedura de eşantionare care are la bază acest princpiu este denumită: eşantionare multistadială. În această situaţie populaţia vizată este împarţită în grupuri în funcţie de un anumit criteriu, aceste grupuri la rândul lor pot fi considerate ca fiind formate din alte grupruri şi aşa mai departe. Date fiind aceste condiţii, selectarea elementelor care vor compune eşantionul poate începe prin selectarea grupurilor din care fac parte aceste elemente. Astfel, într-o primă fază sunt selectate aleator o parte din grupurile populaţiei vizate, după care din fiecare grup selectat în prima fază vor fi selectate tot aleator alte grupruri mai mici şi aşa mai departe până când se ajunge la nivelul elementului de bază din care este compusă populaţia vizată. Spre exemplu, dacă dorim să alegem un eşantion din populaţia unui oraş, într-o primă fază putem selecta cartiere din acel oraş, apoi străzi, blocuri, apartamente şi în cele din urmă persoanele care ne interesează.

Avantajul unei astfel de proceduri de eşantionare îl constituie costurile reduse în raport cu celelalte proceduri prezentate până acum, în sensul în care efortul şi timpul necesar identificării unui element care va fi inclus în eşantion este mult mai redus.

În ceea ce priveşte reprezentativitatea unui astfel de eşantion, se consideră în general că, la volum egal, este mai puţin repezentativ în comparaţie cu un eşantion obţinut prin stratificare sau în comparaţie cu un eşantion obţinut prin procedee simple aleatoare. Reprezentativitatea mai scăzută este rezultatul eliminării la diferite nivele a unor grupuri de elemente din populaţia vizată. Cu cât aceste grupuri care sunt eliminate sunt mai mari şi cu cât sunt mai diferite în comparaţie cu grupurile care nu au fost eliminate cu atât este mai mare riscul de a greşi.

Eşantioane neprobabilistice

Alături de aceste proceduri de eşantionare probabilistice în practica de cercetare sunt utilizate şi o serie de tehnici mai puţin riguroase în ceea ce priveşte selectarea celor care vor compune un eşantion. Lipsa de rigurozitate se referă mai ales la neacordarea unei atenţii speciale calculării sau egalizării şanselor fiecărui individ din populaţia vizată de a face parte din eşantion. Eşantioanele obţinute în acest fel sunt denumite eşantioane neprobabilistice. Astfel de eşantioane se constituie în următoarele situaţii :

- persoane care se oferă voluntar pentru a fi investigaţe;
- persoane care îşi desfăşoara activitatea într-o instituţie anume care prezintă interes pentru cel care efectuează cercetarea;
- persoane care răspund la chestionare publicate în ziare;
- persoane care apelează telefonic un post de radio sau de televiziune pentru a răspunde la întrebările care sunt formulate de moderatorii unor emisiuni sau de alţi participanţi la emisiunile respective;
- persoane intervievate pe stradă sau în anumite spaţii publice;

În cadrul acestor tehnici de eşantionare neprobabilistică cele mai des utilizate sunt eşantionarea "pe cote" şi eşantionarea tip "bulgăre de zăpadă".

1. Eşantionarea pe cote

Eşantionarea pe cote este probabil cea mai des utilizată procedură de eşantionare neprobabilistică utilizată atunci când se lucrează cu populaţii numeroase. Din punct de vedere al realizării practice această procedură este similară eşantionării prin stratificare prin aceea că populaţia vizată este stratificată după o serie de criterii însă în interiorul straturilor nu sunt selectaţi aleator, ci selecţia acestora este lăsată la latitudinea operatorilor de anchetă. Acestora le sunt indicate numai anumite "cote" care indică frecvenţa cu care să fie selectaţi subiecţii care au anumite caracteristici. Spre exemplu dacă în populaţia vizată avem 49% bărbaţi şi 51% femei şi 20% au studii superioare iar restul de 80% nu au astfel de studii, iar eşantionul este format 1000 de persoane, atunci în cadrul acetuia vor fi cuprinşi 490 de barbaţi şi 510 femei, 200 de persoane cu studii superioare şi 800 de persoane care nu au absolvit învăţământul superior. În această situaţie dacă sunt utilizaţi 10 operatori de interviu fiecăria i se cere să chestioneze 49 de bărbaţi şi 51 de femei, 20 de absolvenţi de învăţământ superior şi 80 de persoane care au absolvit o formă de învăţământ alta decât facultatea. Pentru a se limita subiectivitatea operatorilor în selecatarea celor care vor fi incluşi în eşantion se recomandă stabilirea a cât mai multor criterii de stratificare a populaţiei vizate.

Avantajul unui astfel de procedeu de stratificare este acela că nu necesită existenţa unui cadru de eşantionare, lucru care în unele situaţii este greu de realizat, iar munca operatorilor este mult uşurată prin aceea ce nu trebuie să caute o persoană anume ci au libertatea de a alege pe cine vor cu condiţia deţinerii anumitor caracteristici vizate de cercetare.

2. Eşantionarea tip "bulgăre de zăpadă"

Este o procedură de eşantionare utilizată în situaţia în care nu există informaţii suficiente pentru a identifica toţi indivizii care compun o anumită populaţie, ci este posibilă doar identificare doar a câtorva astfel de indivizie. Date fiind aceste circumstanţe, analiza unui populaţii vizate începe cu investigarea indivizilor cunoscuţi după care acestora li se cere să precizeze dacă este posibil şi alte persoane care se presupune ca întrunesc caracteristici vizate de cercetare. Procedeul se desfăşoară în acest fel până când sunt identificaţi atâţia indivizi câţi sunt necesari constituirii unui eşantion. Se utilizează acest procedeu în cazul în care populaţia vizată este formată spre exemplu din oameni care au aumite hobby-uri sau pasiuni, preocupări şi despre care de obicei nu se cunosc în faza iniţială multe informaţii şi nu se ştie nici câte astfel de persoane compun populaţia vizată.

Concluzie

Eşantionarea este un procedeu des utilizat în practica de cercetare în diferite domenii ale activităţii umane. De la medicul care face analize de laborator prelevând o probă de sânge de la un pacient şi până la cei care sunt interesaţi de aspecte ale opinei publice în diferite domenii precum: preferinţele electorale, acordul sau dezacordul cu anumite politici publice sau decizii administrative, etc. În funcţie de tematica avută în vedere şi de informaţiile disponibile cu privire la populaţia vizată procedurile de eşantionare respectă mai mult sau mai puţin anumite rigori în ceea ce priveşte selectarea elementelor din populaţie care vor constitui eşantionul.

În practică procedurile de eşantionare prezentate pe parcursul acestui capitol suferăo serie de abateri şi de *adaptări* sau *ajustări* . De cele mai multe ori acestea constau în combinarea mai multor tehnici de eşantionare în felul acesta sperându-se obţinerea unor informaţii cât mai corecte şi mai precise despre populaţia avută în vedere.

Întrebări:

1. Cât de multe elemente trebuie să cuprindă un eşantion extras dintr-o populaţie perfect omogenă?
2. Între un eşantion simplu aleator şi unul prin stratificare este mai reprezentativ: a) cel simplu aleator, b) cel prin stratificare c) amândouă eşantioanele au acelaşi nivel de reprezentativitate.
3. Să presupunem că se realizează un eşantion utilizând “metoda pasului”. Mărimea eşantionului este de 200 de elemente iar cea a populaţiei vizate este de 2800 de elemente. Care este mărimea pasului utilizat? De la al câtelea element al populaţiei poate începe punerea în practică a pasului de eşantionare?

Bibliografie:

1. Babbie, E. Survey Research Methods, Belmont, Calif. Wadsworth, 1973
2. Fink, A., How to sample in surveys, Sage Publications, Thousands Oaks, London, New York, 1995.

3. Johnson, J., Joslyn, R., Political science research methods, CQ Press, Washington, 1995.
4. Kalton, G., Introduction to survey sampling, Sage University Press, 1983.
5. Rotariu, T., (coord.). Metode statistice aplicate în ştiinţele sociale. Ed. Polirom, Iaşi, 1999.
6. Rotariu, T., Iluţ P., Ancheta sociologică şi sondajul de opinie. Ed. Polirom, Iaşi, 1997.
7. Schimdt, M., Understanding and using statistics. Basic concepts, Second Edition, Lexington, Massachusetts, Toronto, 1979.

Unitatea 4

Obiectiv: Prezentarea aspectelor matematice ale eşantionării

Cuvinte cheie: intervale de conficenţă, teste de seminificaţie, testul t, testul Z, testul χ^2 (hi pătrat)

Aspecte matematice ale eşantionării. Teste de semnificaţie

Valori măsurate pe populaţie şi pe eşantion. Intervale de confidenţă

Extrăgând un eşantion dintr-o populaţie şi măsurând pe acesta valoarea medie a unei caracteristici sau variabile putem spune într-o oarecare măsură că această valoare aproximează o valoare a aceleiaşi caracteristici din populaţie. Cu toate acestea întreabarea care se ridică este: cât de siguri putem fi de rezultatele obţinute dat fiind că eşantionul extras la un moment dat este doar unul din multele eşantioane care pot fi extrase dintr-o populaţie? Spre exemplu, dorim să estimăm nivelul de inteligenţă al elevilor unei şcoli şi pentru aceasta extragem aleator un eşantion format din 25 de elevi cărora le aplicăm un test de inteligenţă şi obţinem o valoare medie a indicelui de inteligenţă de 108 şi o abatere standard de 12. Bazându-ne pe aceste rezultate, ce putem spune despre nivelul de inteligenţă al elevilor şcolii respective? Eşantionul de 25 de elevi este evident doar unul din eşantioanele care ar fi putut fi extrase şi prin urmare şi media de 108 obţinută de cei care au făcut parte din eşantion este doar una din posibilele medii. Mai clar spus, 108 este doar una dintre mediile din distribuţia de medii care ar putea fi obţinută extrăgând multe eşantioane formate din 25 de elevi ai şcolii respective. Problema este: cât de aproape este această medie de valoarea reală a indicelui de inteligenţă a tuturor elevilor acelei şcoli? şi care este valoarea medie a indicelui de inteligenţă pentru întreaga populaţie de elevi vizată? - valoare evident necunoscută, altfel ce rost ar mai avea să facem cercetarea!

Pentru a răspunde la această întrebare trebuie să facem apel la o teoremă statistică, denumită *teorema limitei centrale*, care afirmă că pentru eşantioane suficient de mari distribuţia mediilor măsurate pe aceste eşantioane este întotdeuna normală, chiar dacă valorile caracteristicii iniţiale sunt sau nu normal distribuite într-o populatie vizată. Mediile unei caracterisitici măsurate pe multe eşantioane pot fi privite ca formând o nouă variabilă pentru care vom putea calcula evident o medie şi o abatere standard. Valoarea medie a noii variabile (media mediilor măsurate pe eşantioanele extrase din populaţia vizată) este egală cu media valorii din populaţie a caracteristicii vizate, iar abaterea standard a acestei variabile, în cazul în care eşantioanele sunt extrase printr-o simplă aleatoare cu reintroducerea elementului extras în populaţie (acordând deci o şansă egală fiecărui element de a fi extras), este egală cu abaterea standard a variabilei urmărite măsurată pe un eşantion

$$e = \frac{\sigma}{\sqrt{n}}$$

oarecare împărţită la rădăcină pătrată din mărimea eşantionului. Abaterea standard a noii variabile este denumită *eroare standard* (**e**):

Revenind la întrebarea din exemplul de mai sus: care este valoarea medie a indicelui de inteligenţă pentru întreaga populaţie de elevi vizată? Un răspuns exact nu poate fi dat întrucât nu a fost investigată întreaga populaţie. Ştiind însă că distribuţia valorilor medii măsurate pe multe eşantioane extrase din populaţia de elevi vizată este normală putem calcula un interval despre care să spunem că, cu probabilitate de 95% include media indicelui de inteligenţă din întrega populaţie vizată. Acest inteval este cuprins între plus două şi minus două erori standard (e = 12/5 = 2,4) în jurul valorii medii obţinute pe un eşantion oarecare extras din acea populaţie, adică între 108 – 4,8 şi 108 + 4,8. Intervalul astfel construit poartă denumirea de *interval de confidenţă.*

Teste de semnificaţie. Inferenţa statistică

Adeseori observăm diferenţe între rezultatele obţinute atunci când se fac măsurători pe două eşantioane diferite extrase din aceeaşi populaţie. Spre exemplu, 17% dintre cei chestionaţi în cadrul unei anchete sociale sunt de acord cu o anumită decizie a administraţie publice locale la un anumit moment dat de timp, dar numai 11% au aceeaşi opinie la un alt moment de timp. Problema care se pune în această situaţie este: cât de reală sau de semnificativă este diferenţa între cele două grupuri – cei chestionaţi la un moment de timp şi cei chestionaţi la un moment de timp ulterior? Este această diferenţă autentică sau sau este rezultatul fluctuaţiilor fireşti ale eşantionării?

Similar ne putem întreba: ce se poate spune despre valoarea unei caracteristici dintr-o populaţie pe baza rezultatelor obţinute atunci când este investigat un eşantion? vor fi rezultatele obţinute atunci când se fac măsurători pe un eşantion identice cu rezultatele obţinute atunci când se fac măsurători pe întreaga populaţie? iar dacă nu, diferenţele identificate sunt semnificative sau nu? mărimea eşantionului influenţează modul în care rezultatele obţinute reflectă caracteristici ale populaţiei?

Toate aceste întrebări sunt justificate întrucât, aşa cum am arătat în capitolul dedicat eşantionării, eşantioanele nu reproduc exact caracteristicile unei populaţii, ci există o anumită diferenţă între valoarea unei caracteristici măsurată pe un eşantion şi valoarea aceleiaşi caracteristici măsurată pe populaţia din care este extras eşantionul. Cu toate acestea de multe ori suntem puşi în situaţia de a trage concluzii cu privire la starea unei populaţii pornind de la măsurători efectuate la nivelul unui eşantion, cu alte cuvinte se pune problema de a face inferenţe de la eşantion la populaţie. Bazate pe numere utilizate pentru a sumariza, evalua sau analiza un set de informaţii cu privire la un fenomen analizat, numere care în literatura de specialitate sunt denumite *statistici*, inferenţele de acest fel sunt şi ele denumite inferenţe statistice. Inferenţele statistice, ca urmare a faptului că eşantioanele pe baza cărora sunt realizate constituie doar aproximări ale unei populaţii, prezintă neajunsul de a putea produce concluzii eronate. Prin urmare, atunci când se compară două valori ale unor caracteritici dintre care cel puţin una a fost obţinută prin măsurători efectuate pe un eşantion, se pune problema semnificaţiei diferenţei dintre ele.

Din punct de vedere cantitativ, v*om spune că diferenţa între două valori, fie că una este măsurată pe un eşantion şi alta pe o populaţie, fie că amândouă valorile sunt măsurate pe eşantioane, este semnificativă atunci când nu poate fi încadrată cu un anumit nivel de probabilitate acceptabil într-o limită maximă prestabilită*. Pe de altă parte, o diferenţă care nu este semnificativă potrivit definiţiei de mai sus nu înseamnă în mod automat că nu poate fi reală, ci doar că nu se poate spune cu un nivel de probabilitate acceptabil că este reală.

Pentru a facilita munca în domeniul practic au fost elaborate seturi de reguli pe baza cărora se stabileşte dacă diferenţele între valori sunt sau nu semnificative statistic. Fiecare set de astfel de reguli poartă denumirea de *test de semnificaţie* şi are scopul de a ajuta la stabilirea unei *concluzii statistice* cu privire la starea unor caracteristici ale populaţiei investigate. Testele de semnificaţie nu sunt probe absolute ale existenţei sau non-existenţei unei diferenţe semnificative între două valori, ele doar permit estimarea în raport cu o ipoteză prealabilă a probabilităţii prezenţei unei diferenţe reale între valori. Cel mai adesea astfel de ipoteze în care sunt enunţate predicţii cu privire la valorile unor caracteristici avute în vedere în cercetare iau forma *ipotezei nule,* adică a afirmării inexistenţei unei diferenţe semnificative între două valori comparate. Mai clar spus, ipoteza nulă este ipoteza care afirmă că două mărimi A şi B măsurate pe eşantioane diferite sau una măsurată pe un eşantion şi una pe o populaţie, sunt egale. Ţinând cont de toate aceste un *test de semnicaţie poate fi definit ca fiind măsura diferenţei dintre două valori în raport cu ipoteza nulă.*

Ipoteza nulă este testată în felul următor: dacă cu un anumit nivel de probabilitate diferenţa dintre cele două valori comparate este mai mare decât o valoare maximă prestabilită atunci ipoteza nulă este respinsă şi vom spune ca acea diferenţă este semnificativă. În caz contrar – cu un anumit nivel de probabilitate diferenţa între valori este mai mică decât o valoare maximă prestabilită – ipoteza nulă este susţinută şi vom spune că diferenţa respectivă nu este semnificativă. O întrebare fierească este: cât de mare trebuie să fie nivelul de probabilitate pentru a accepta sau respinge ipoteza nulă? Alegerea depinde în general de ipoteza care urmează a fi testată. Practica a consacrat însă ca nivel de probabilitate cel mai des utilizat pragul de 0.95 (95%) spunându-se despre o diferenţă care cu o probabilitate de 95% nu depăşeşte o valoare maximă prestabilită că este semnificativă statistic.

Valorile comparate pot fi după caz: medii, proporţii, sau orice alte măsuri. Una din valorile avute în vedere în cazul în care sunt efectuate teste de semnificaţie poate fi zero, ceea ce înseamnă că practic testăm semnificaţia unei singure mărimi în comparaţie cu valoarea zero.

În funcţie de mărimea grupurile pe care sunt măsurate valorile caracteristicilor urmărite şi de modul de măsurare a acestora avem mai multe teste de semnificaţie. În cele ce urmează vom prezenta testul Z, testul Student (t), şi testul χ^2 (hi pătrat).

Testul Z

Este un test de semnificaţie utilizat în cazul în care se compară valorile unor caracteristici măsurate pe eşantioane mari (de ordinul a sute sau mii de indivizi). Cele două valori comparate pot fi măsurate fie una pe o populaţie şi una pe un eşantion, fie amândouă valorile sunt măsurate pe eşantioane diferite.

În prima situaţie, fie **a** şi **b** cele două valori ale aceleiaşi caracteristici, dintre care valoarea **a** este măsurată pe o populaţie iar valoarea **b** este măsurată pe un eşantion şi fie **e** eroarea standard a caracteristicii luate în considerare. Testul Z este definit după formula:

$$Z = \frac{|a-b|}{e}$$

şi exprimă de fapt diferenţa dintre valorile **a** şi **b** în erori standard. Dacă valoarea testului Z este mai mare de 1.96 atunci diferenţa dintre cele două valori este semnificativă din punct de vedere statistic la un nivel de probabilitate de 0,95 (95%). Sau altfel spus, cu o probabilitate de 95%

diferenţa între cele două valori este semnificativă din punct de vedere statistic. Alături de nivelul de probabilitate de 0.95 mai sunt utilizate nivelele de probabilitate de 0.99 (Z=2,6) şi 0,999 (Z=3,3). Valorile pragurilor de probabilitate pentru testul Z sunt prezentate în Tabelul 1.

Pentru a ilustra modul de aplicare a testului Z vom utiliza un exemplu. Să presupunem că în cadrul unui referendum 42% dintre cetăţenii unei localităţi sunt de acord cu introducerea unui nou sistem de impozite. Cu toate acestea într-un sondaj de opinie realizat anterior referendumului pe un eşantion de 900 de persoane indică că doar 37% dintre cetăţeni vor fi de acord cu noua grilă de impozitare. Este diferenţa între cele două valori autentică sau nu? Sau altfel spus, este diferenţa dintre cele două valori semnificativă?

Pentru a pune în evidenţă acest lucru calculăm:

$\sigma^2 = 0{,}37(1 - 0{,}37) = 0{,}2331$ şi $\sigma = 0{,}48$

şi sau

$$e = \frac{0{,}48}{\sqrt{900}} = 0{,}016 \qquad e = 1{,}6\%$$

înlocuind în formula lui Z obţinem: (42 - 37)/1,6 = 3,12

Cautând în tabel pragurile de probabilitate ale lui Z (Tabelul 1.) în dreptul lui 3,1 şi pe coloana 0,02 (cea care indică sutimile numărului 3,12) găsim numărul 4991 care redus la unitate devine 0,4991 şi reprezintă jumătate din nivelul de probabilitate cautat (este de fapt jumătate din aria determinată de curba normală). Înmulţind cu 2 obţinem numărul 0,9982 (P = 0,9982) care ne spune că sunt aproximativ 99,8% şanse ca diferenţă dintre cele două valori să fie reală.

În cazul în care cele două valori ale unei caracteristici sunt măsurate pe două eşantioane distincte formula testului Z este aceeaşi cu precizarea că eroarea standard se calculează după formula

$$e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

unde $\boldsymbol{n_1}$ şi $\boldsymbol{n_2}$ sunt mărimile celor două eşantioane, iar $\boldsymbol{\sigma_1}$ şi $\boldsymbol{\sigma_2}$ sunt abaterile standard ale valorilor caracteristicii pentru fiecare dintre cele două eşantioane.

Testul Student (t)

Atunci când se pune problema de a compara valori ale unor caracterisitici dintre care cel puţin una este obţinută prin măsurători efectuate pe eşantioane de mărimi mici (pâna la 30 de indivizi) corespondetul testului Z este testul **Student (t).** Formula de calcul a testului Student este identică cu aceea a testului Z:

$$t = \frac{|a - b|}{e}$$

Deosebirea faţă de testul Z constă în modul de calcul al erorii standard (**e)** care se face după formula:

$$e = \frac{\sqrt{\frac{\sum\left(x-\overline{x}\right)^2}{n-1}}}{\sqrt{n}}$$

dacă una dintre valori este măsurată pe o un eşantion de mărime ***n*** şi una pe o populaţie, şi după formula:

$$e = \sqrt{\frac{\sum\left(x_i-\overline{x_1}\right)^2+\sum\left(x_i-\overline{x_2}\right)^2}{n_1+n_2-2}}$$

dacă cele două valori sunt măsurate pe eşantioane ale căror mărimi sunt $\boldsymbol{n_1}$ respectiv $\boldsymbol{n_2}$.

La fel ca şi în cazul testului Z şi pentru testul Student sunt utilizate diferite praguri de probabilitate (Tabelul 2.) care reclamă şi specificarea numărului de grade de libertate, care se calculează după formulele:

respectiv

$$\upsilon = n-1 \qquad \upsilon = n_1+n_2-2$$

Testul χ^2 (hi pătrat)

Testele Z si Student sunt utilizate pentru a testa ipoteze care se referă la valori, sau parametrii (medii sau proporţii), măsurate pe populaţii sau pe eşantioane, motiv pentru care sunt adeseori cunoscute şi sub denumirea mai largă de *teste parametrice.* Există însă multe situaţii în care ipotezele nu pot fi testate utilizând doar medii sau proporţi. Acest lucru se întâmpla spre exemplu atunci când datele cu care se lucrează nu sunt de tip cantitativ. Există apoi şi alte condiţii care trebuiesc îndeplinite în cazul testelor parametrice - utilizarea unor eşantioane mari sau a unor eşantioane extrase din populaţii normal distribuite astfel încât şi forma ditribuţiei de eşantionare să fie cunoscută - condiţii care nu întotdeauna pot fi îndeplinite.

Pentru a depăşi acest tip de neajunsuri au fost construite şi o serie de teste a căror mod de operare nu presupune existenţă unor asumpţii cu privire la populaţia vizată sau cu privire la datele pe care le avem la dispoziţie cu privire la aceasta. Acest tip de teste sunt denumite *teste non-parametrice.* Unul dintre cele mai des utilizate teste de acest fel este testul **χ^2 (hi pătrat).**

Scopul principal al acestui test este similar testelor Z şi Student şi anume încearcă să ofere un răspuns întrebării: dată fiind o mulţime de valori observate ale unei caracteristici, modul de distribuire a acestor valori poate fi atribuit în întregime fluctuaţiilor fireşti ale eşantionării sau există o serie de alţi factori care influenţează această distribuire? Şi în acest caz, pentru a răspunde acestei întrebări, punctul de plecare este o ipoteză nulă care afirmă că nu există alţi factori care să influenţeze distribuţia valorilor observate ale unei variabile.

Pentru a ilustra modul de operare al testului χ^2 (hi pătrat) vom utiliza exemplu următor. Fie următoarea situaţie ipotetică: 100 de funcţionari ai unei instituţii publice sunt întrebaţi cu privire la

ceea ce îi nemulţumeşte cel mai mult la locul de muncă, răspunsurile oferite având următoarele frecvenţe:

	frecvenţa
modul de organizare al activităţilor	24
modul în care sunt trataţi de şef	10
existenţa unui program fix de lucru	27
lipsa unui spirit de echipă	11
lipsa unor rezultate vizibile	28

Întrebarea care se pune în această situaţie este: există un motiv de nemulţumire care este mai acut decât altele?

Ipoteza nulă în această situaţie ar fi aceea că fiecare dintre motivele enumerate mai sus nemultumeşte în egală măsură pe funcţionarii acelei instituţii, adică fiecare dintre cele cinci răspunsuri având aceeaşi probabilitate de a fi indicat de către respondeţi. Din punct de vedere statistic aceasta ar însemna că frecvenţele observate ale răspunsurilor primite pot fi considerate egale cu frecvenţele aşteptate. Dacă ipoteza nulă este susţinută atunci răspunsurile ar trebui să fie distribuite aleator pe cele cinci categorii de răspunsuri luate în considerare.

Pentru a testa această ipoteză să utilizăm testul χ^2 (hi pătrat) care este definit după formula:

$$\chi^2 = \sum_i^n \frac{(O_i - A_i)}{A_i}$$

unde O_i reprezintă frecvenţele observate, iar A_i reprezintă frecvenţele aşteptate (adică distribuţia aleatoare a răspunsurilor pe cele cinci categorii ale caracteristicii analizate – motiv de nemulţumire).

În cazul nostru cele două frecvenţe sunt:

	O_i	A_i	$O_i - A_i$
modul de organizare al activităţilor	24	20	4
modul în care sunt trataţi de şef	10	20	-10
existenţa unui program fix de lucru	27	20	7
lipsa unui spirit de echipă	11	20	-9
lipsa unor rezultate vizibile	28	20	8

înlocuind în formula lui χ^2 (hi pătrat) obţinem:

$$\chi^2 = \frac{(24-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(11-20)^2}{20} + \frac{(28-20)^2}{20}$$

$$\chi^2 = \frac{4^2}{20} + \frac{10^2}{20} + \frac{7^2}{20} + \frac{9^2}{20} + \frac{8^2}{20}$$

$$\chi^2 = 10.50 + 5.00 + 2.45 + 4.05 + 3.20$$

Valoarea obţinută pentru χ^2 (15.50) se compară cu valorile critice ale distibuţiei *hi pătrat* (Tabelul 3.) pentru diferite nivele de probabilitate, dintre care cel mai des utilizat este şi de această dată pragul de 0,95 (95%) .Ca şi în cazul testului Student compararea valorilor critice ale unei distribuţii observate cu ditribuţia *hi pătrat* reclamă specificarea numărului de grade de libertate, număr care se calculează după formula *df* = k –1, unde k reprezintă numărul de categorii ale caracteristicii analizate. În exemplul de mai sus k = 5-1 = 4 şi căutând în tabelul cu valori critice ale lui *hi pătrat* găsim că pentru nivelul de probabilitate de 95% şi 4 grade de libertate valoarea critică este 9,488. Cum 15.50 este mai mare decât această valoare critică, vom spune că ipoteza nulă se respinge cu o probabilitate de 95% sau, cu alte cuvinte, răspunsurile date de funcţionari nu se distribuie aleator ci există un anumit motiv de nemulţumire care este mai acut decât celelalte, iar acest rezultat nu este generat de fluctuaţiile de eşantionare.

Hi pătrat poate fi utilizat şi pentru a testa dacă două variabile sunt sau nu asociate. Fie, spre exemplu, următoarea situaţie ipotetică: 600 de locuitori ai unei localităţi sunt întrebaţi dacă vor sprijini sau nu o schimbare a modului de alocare a veniturilor bugetare ale localităţii lor au răspuns după cum urmează:

Frecvenţe observate				
	Da	Nu	Nu ştiu	Total
cei cu vârsta sub 25	110	40	30	180
cei cu vârstă între 26 şi 45 de ani	40	100	60	200
cei cu varsta peste 45 de ani	50	80	90	220
Total	200	220	180	600

În această situaţie se poate pune întrebarea: există sau nu o preferinţă a unei anumite categorii de vârstă pentru schimbarea modului de alocare a veniturilor? Cu alte cuvinte există o relaţie între vârstă şi acordul cu această schimbare? Pentru a răspunde la această întrebare trebuie să vedem cum ar trebui să arate distribuţia în situaţia în care nu există asociere. Astfel, dacă nu ar exista o relaţie între variabile, atunci preferinţele ar trebui să se distribuie uniform pentru fiecare categorie de vârstă în parte; cu alte cuvinte, o treime dintre indivizii din fiecare categorie de vârstă să fie de acord cu schimbarea, o treime să nu fie de acord şi o treime să răspundă că "nu ştiu". Acest lucru raportat la frecvenţele din tabelul de mai sus ar însemna: 60 de persoane cu vârsta sub 25 de ani să fie de acord cu schimbare (adică o treime din cele 180 de persoane cu vărsta sub 25 de ani cuprinse în eşantionul nostru), 66,67 persoane cu vărsta cuprinsă între 26 şi 45 de ani şi aşa mai departe:

Frecvenţe aşteptate				
	Da	Nu	Nu ştiu	Total
cei cu vârsta sub 25	60	66	54	180
cei cu vârstă între 26 şi 45 de ani	66,67	73,33	60	200
cei cu varsta peste 45 de ani	73,33	80,67	66	220
Total	200	220	180	600

Calculându-l pe *hi pătrat* obţinem:

$$\chi^2 = \frac{(110-60)^2}{60} + \frac{(40-66)^2}{66} + \ldots + \frac{(40-66{,}7)^2}{66{,}7_{28}} + \frac{(100-73{,}33^2)}{73{,}33} + \ldots + \frac{(90-66)^2}{66}$$

Numărul gradelor de libertate în acest caz se calculează după formula:

$$df = (j-1)(k-1)$$

$\chi^2 = 99{,}11$

unde j reprezintă numărul de rânduri ale tabelului în care sunt dispuse frecvenţele şi k reprezintă numărul de coloane. În acest caz df = 4. Cautând în tabelul cu valori critice pentru χ^2 observăm că unui nivel de probabilitate de 95% şi 4 grade de libertate îi corespunde valoarea 9,488, valoare mai mică decât valoarea calculată a lui χ^2. În această situaţie vom spune că ipoteza potrvit căreia nu există asociere între vârstă şi preferinţa pentru schimbarea modului de alocare a veniturilor se respinge.

Teste parametrice sau non-parametrice?

Când utilizăm teste parametrice şi când utilizăm teste non-parametrice pentru a analiza un set de date? Răspunsul la această întrebare nu este întotdeună foarte tranşant.

Astfel, nu vom putea utiliza teste parametrice dacă datele pe care le avem la dispoziţie sunt de tip calitativ, motivul este acelă că testele parametrice operează de cele mai multe ori cu valori medii, valori care evident nu pot fi calculate pentru date de tip calitativ. În această situaţie un test non-parametric este singura alternativă posibilă. Pe de altă parte testele parametrice sunt considerate a avea o putere statistică mai mare decât testele non-parametrice şi aceasta pentru că modul lor de operare ia în considerare mai multă informaţie despre caracteristica avută în vedere. Dar acest lucru se face cu anumite asumpţii, dintre care cea mai importantă este distribuţia normală a valorilor caracteristicii analizate.

Cât de puternice sunt testele paramentrice în raport cu cele non-parametrice? Raspunsul trebuie şi de această dată nuanţat. Puterea statistică a unui test este de fapt probabiliatea de a respinge ipoteza nulă atunci când aceasta nu este adevărată. Dar şi în acest caz situaţiile depind de modul de formulare a ipotezei nule şi de mărimea eşntionului extras. Dacă una dintre aceste două variabile suferă modificări şi puterea statistică a unui test este afectată.

Practica a demonstrat că amândouă tipurile de teste pot fi utilizate cu acelaşi succes cu condiţia luării în calcul a avantajelor şi dezavantajelor fiecăruia.

Probleme:

1. Să presupunem că 35,4% dintre cetăţenii unei localităţi au votat partidul X la alegerile locale. Un sondajele de opine realizat în perioada pre-electorală pe un eşantion de 1000 de persoane acorda însă acestui partid 39% dintre intenţiile de vot ale electoratului. Este diferenţa între cele două valori autentică sau nu?

2. Dintre 200 de elevii ai unei şcoli intevievaţi cu privire la dificultăţile de învăţăre pe care le întâmpină : 38 au răspuns că acestea îşi au originea în programul încărcat de la şcoala, 62 au răspuns că lipsa unei dotări adcvate a şcolii le crează astfel de dificultăţi, 56 au răspuns că modul de structurare a materiilor învăţăte este cauza dificultăţilor de învăţăre, iar 46 au pus că dificultăţile de învăţăre se datorează unor cauze externe şcolii. Există un motiv care să determine într-o mai mare măsură dificulăţi de învăţăre pentru elevii şcolii avute în vedere?

3. Testele de semnificaţie nu sunt probe absolute ale existenţei sau non-existenţei unei diferenţe semnificative între două valori. Comentaţi această afirmaţie.

Tabelul 1. Proporţia din aria totală (10.000) ce corespunde distanţei dintre medie şi Z abateri standard de la medie (Valorile pragurilor de probabilitate pentru testul Z).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0735
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0.7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1.6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2.0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2.9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3.0	4986,5	4986	4987	4988	4988	4988	4989	4989	4989	4990
3.1	4990,0	4991	4991	4991	4992	4992	4992	4992	4993	4993
3.2	4993,129									
3.3	4995,166									
3.4	4996,631									

Sursa: Mohr, L.B., Understanding Significance Testing. Sage Publications, Newbury Park/London, New Delhi, Sage Publications, 1990.

Tabelul 2. . Valorile critice pentru testul **Student (t),** pentru nivelurile de probabilitate de 0.05, 0.02 şi 0.01, în funcţie de numărul gradelor de libertate (**v**)

v	**p=0.05**	**p=0.02**	**p=0.01**
1	12.71	31.82	63.66
2	4.30	6.97	9.93
3	3.18	4.54	5.84
4	2.78	3.75	4.60
5	2.57	3.37	4.03
6	2.45	3.14	3.71
7	2.73	3.00	3.50
8	2.31	2.90	3.36
9	2.26	2.82	3.25
10	2.23	2.76	3.17
11	2.20	2.72	3.11
12	2.18	2.68	3.06
13	2.16	2.65	3.01
14	2.15	2.62	2.98
15	2.13	2.60	2.95
16	2.12	2.58	2.98
17	2.11	2.57	2.90
18	2.10	2.55	2.88
19	2.09	2.54	2.86
20	2.09	2.53	2.85
21	2.08	2.52	2.83
22	2.07	2.51	2.82
23	2.07	2.50	2.81
24	2.06	2.49	2.80
25	2.06	2.49	2.79
26	2.06	2.48	2.78
27	2.05	2.47	2.77
28	2.05	2.47	2.76
29	2.05	2.46	2.75
30	2.04	2.46	2.75
∞	1.96	2.33	2.58

Sursa: Pinty, J.J., Gaultier Claude, Dictionnaire pratique de mathématiques et statistiques en sciences humaines, Édition Universitaire, Paris, 1971.

Tabelul 3. Valorile critice pentru testul χ^2, pentru nivelurile de probabilitate de 0.05, 0.02 şi 0.01, în funcţie de numărul gradelor de libertate (**v**)

v	**p=0.05**	**p=0.02**	**p=0.01**
1	3.84	5.41	6.64
2	5.99	7.82	9.21
3	7.82	9.84	11.35
4	9.49	11.67	13.28
5	11.07	13.39	15.09
6	12.59	15.03	16.81
7	14.07	16.62	18.48
8	15.51	18.17	20.09
9	16.92	19.68	21.67
10	18.31	21.16	23.21
11	19.68	22.62	24.72
12	21.03	24.05	26.22
13	22.36	25.47	27.69
14	23.69	26.87	29.14
15	25.00	28.26	30.58
16	26.30	29.63	32.00
17	27.59	31.00	33.41
18	28.87	32.35	34.81
19	30.14	33.69	36.19
20	31.41	35.02	37.57
21	32.67	36.34	38.93
22	33.92	37.66	40.29
23	35.17	38.97	41.64
24	36.42	40.27	42.98
25	37.65	41.57	44.31
26	38.89	42.86	45.64
27	40.11	44.14	46.96
28	41.34	45.42	48.28
29	42.56	46.69	49.59
30	43.77	47.96	50.89

Sursa: Yule, G.U., Kendall, M.G. Introducere în teoria statisticii, Editura Ştiinţifică, Bucureşti, 1969.

Bibliografie:

1 Rotariu, T., (coord.). Metode statistice aplicate în ştiinţele sociale. Ed. Polirom, Iaşi, 1999.

2 Rotariu, T., Iluţ P., Ancheta sociologică şi sondajul de opinie. Ed. Polirom, Iaşi, 1997.

3 Schimdt, M., Understanding and using statistics. Basic concepts, Second Edition, Lexington, Massachusetts, Toronto, 1979.

4 Freedman D., Pisani R., Purves R., Adhikari A., Statistics, Second Edition, New York, London , 1991;

Modulul 3

Obiectiv: prezentarea tipurilor de variabile utilizate în ştiinţele sociale şi modelelor de analiză a acestora

Ghid de studiu:

- Variabile. Tipuri de variabile.
- Analiza univariată a datelor
- Analiza bivariată a datelor

Unitatea 1

Obiectiv: Introducerea noţiunii de variabilă şi a tipurilor de variabile

Cuvinte cheie: parametrii, variabile, estimare, variabile continue, variabile discrete

Variabile. Tipuri de variabile.

Caracteristicile populaţiei despre care facem ineferenţe pe baza eşantionului se numesc parametrii. Caracteristicile eşantionului pe baza cărora inferăm se numesc pur şi simplu statistici. În exemplul de mai sus, 55% reprezintă o statistică descriptivă, deoarece ea descrie sintetic o caracteristică a eşantionului. Cele mai multe studii sunt însă interesate în aflarea parametrilor, care în general sunt necunoscuţi (exemple: Câţi săraci exista în România? Care este procentul din populaţie de susţinători ai unui partid? etc.). Eşantioanele şi statisticile descriptive sunt utile în măsura în care ele pot oferi informaţii despre parametrii de interes. Statistica inferenţială este aceea care permite obţinerea unei masuri a acurateţei statisticilor folosite pentru **estimarea** valorii parametrilor. În consecinţă, atunci când întreaga populaţie este cuprinsă într-un studiu, statistica inferenţială nu este necesară.

În final ne vom opri asupra unei ultime noţiuni deosebit de importante pentru studiul statisticii, şi anume asupra **variabilelor**. *Vom defini variabila ca fiind orice caracteristică a membrilor unei populaţii sau unui eşantion care variază (în respectiva populaţie/eşantion).* Astfel, culoarea părului indivizilor dintr-o populaţie este o variabilă în măsura în care indivizii care compun respectiva populaţie au păr de culori diferite. Dacă toţi indivizii ar fi blonzi, să zicem,

atunci culoarea părului ar fi constantă în respectiva populaţie. Cu cât o caracteristică are o variaţie mai mare, cu atît respectiva populaţie este mai *eterogenă* şi, invers, cu cât o caracteristică dată are o variaţie mai mică, cu atât respectiva populaţie va fi mai *omogenă*, din perspectiva respectivei caracteristici. În exemplul de mai sus, valorile posibile ale variabilei "culoarea părului" ar fi "brunet", "blond", "roşcat" etc.. Fiecare individ (statistic) poate lua o singură valoare pentru o variabilă..

Variabilele pot fi clasificate în funcţie de multe criterii. Una din distincţiile importante este aceea dintre **variabile discrete** şi **variabile continue**. Atât variabilele discrete cât şi variabilele continue pot lua o infinitate de valori. Diferenţa dintre ele constă în faptul că în timp ce în cazul variabilelor continue între două valori succesive ale variabilei pot exista o infinitate de valori, în cazul variabilelor discrete acest lucru nu se întâmplă. Un exemplu de variabilă continuă este înălţimea clădirilor unui oraş masurată în metri, iar un exemplu de variabilă discretă îl reprezintă veniturile indivizilor dintr-o populaţie, măsurate în lei. În cazul primei variabile, între doua valori succesive ale acesteia (de exemplu 5 şi 6 m) există o infinitate de alte valori deoarece metrii se subdivid în centimetri, apoi în milimetri etc., în cazul veniturilor acest lucru nu mai este posibil, între 5 lei şi 6 lei nemaiexistând subdiviziuni.

Nivelul de măsurare al variabilelor este un alt criteriu de clasificare a acestora, de o mare importanţă pentru studiul statisticii. Putem distinge între patru niveluri de măsurare (*nominal, ordinal, de interval* şi *de raport*), în funcţie de trei criterii:

a) posibilitatea de a ordona valorile variabilei,
b) egalitatea intervalelor dintre valorile variabilei (sau altfel spus existenţa unei unităţi de măsură),
c) existenţa unei "origini" a variabilei sau, cu alte cuvinte, a unui "zero absolut".

Tabelul I.1 - Niveluri de masurare a variabilelor

	a) ordonare	b) unitate de masură	c) zero absolut
Nominal	nu	nu	nu
Ordinal	da	nu	nu
De interval	da	da	nu
De raport	da	da	da

1. Nivelul de măsurare **nominal** presupune clasificarea unor atribute, caracteristici, fenomene etc. în categorii care trebuie să fie distincte, mutual exclusive şi exhaustive. Acest tip de variabile (respectiv scalele folosite în măsurare) indică numai faptul că exista o diferenţă calitativă între categoriile studiate, nu şi magnitudinea acestei diferenţe. La limită, putem privi aceste variabile ca pe nişte tipologii. Câteva exemple de variabile măsurate la nivel nominal sunt: statutul ocupaţional al indivizilor (agricultor, salariat, mic întreprinzator, şomer etc.), religia (ortodox, romano-catolic, greco-catolic etc.) apartenenţa etnică (român, maghiar, rrom etc.), mediul de rezidenţă (rural, urban) ş.a.m.d.. Valorile acestui tip de variabile nu pot fi ordonate, sau cu alte cuvinte nu există o ierarhie (decât eventual conform unor criterii extrinseci) şi în consecinţă problema "distanţei" sau a intervalelor dintre valori nici nu poate fi pusă. Cu atât mai puţin putem discuta despre existenţa unui "zero absolut" (exemplu: fiecare individ are un statut ocupaţional sau aparţine unei etnii, sau altfel spus absenţa caracteristicilor "statut ocupaţional" sau "apartenenţă etnică" este imposibilă).
2. Nivelul de masurare **ordinal** implică nu numai clasificarea elementelor în categorii ci şi posibilitatea ordonării acestora de la minim la maxim (existenţa tranzitivităţii: dacă a>b şi b>c, atunci a>c). Totuşi, la acest nivel de măsurare nu este oferită nici o informaţie cu privire la

"distanţa" dintre valorile scalei de măsură. Cu alte cuvinte, diferenţa dintre prima valoare şi cea de-a doua poate fi diferită de diferenţa dintre a patra şi a cincea. Exemple de variabile măsurate la nivel ordinal sunt calificativele şcolare (cu valorile "insuficient", "suficient", "bine" şi "foarte bine"), satisfacţia faţă de anumite aspecte (cu valorile "foarte nesatisfăcut", "nesatisfăcut", "satisfăcut", "foarte satisfăcut") etc..

3. Măsurarea la nivel **de interval**, oferă în plus faţa de nivel anterior (cel ordinal) şi informaţie referitoare la distanţa dintre valorile scalei şi este caracterizată de existenţa unor intervale egale. Totuşi, la acest nivel de măsurare nu există un zero absolut, ci mai degrabă unul convenţional. Exemple de astfel de scale de măsurare sunt temperatura masurată în grade Celsius (intervalele dintre valori sunt egale, dar punctul 0 este convenţional ales ca fiind temperatura la care apa ingheaţă), coeficientul de inteligenţă - IQ - (daca două persoane au scoruri de 100 şi respectiv 150, putem spune ca diferenţa dintre cei doi este de 50 de puncte, dar nu putem spune că cel de-al doilea este cu 1/2 mai inteligent decât primul sau că scorul 0 semnifică absenţa inteligenţei).
4. Măsurarea la nivel **de raport** include toate caracteristicile nivelurilor anterioare (ordonare şi intervale egale), plus existenţa unei "origini" sau zero absolut. Acest lucru permite formularea unor afirmaţii în termeni de proporţii (raporturi) între valori. De exemplu, vitezele de raspuns a doi subiecţi la un acelaşi stimul pot fi comparate în termeni de "timpul de răspuns a fost de două ori mai mare" etc.. Exemple de variabile masurate la acest nivel sunt vârsta, greutatea, înălţimea, distanţa, numărul de copii din gospodărie etc.

Corecta identificare a nivelului de măsurare utilizat este foarte importantă în alegerea procedurilor satistice de analiză. După cum se poate observa din descrierea de mai sus, pentru fiecare nivel exista operaţii matematice permise şi operaţii interzise. Astfel, la primul nivel, cel nominal nu sunt permise nici ordonarea, nici adunarea/scăderea şi nici înmulţirea/împărţirea. La nivelul ordinal este permisă numai ordonarea, la cel de interval sunt permise în plus şi operaţiile de adunare/scădere, iar la ultimul nivel, cel de raport sunt permise toate operaţiile.

În funcţie de nivelul de măsurare, vom vorbi despre variabile măsurate la nivel nominal, variabile măsurate la nivel ordinal etc., sau, mai pe scurt, variabile nominale, ordinale, de interval şi de raport. Reducând cele patru clase la două, putem vorbi de *variabile calitative* (nivelurile nominal şi ordinal) şi *variabile cantitative* (interval şi raport). Datorita caracterului "ierarhic" şi cumulativ al nivelurilor de măsurare (de la multe restricţii către nici o restricţie în ceea ce priveşte operaţiile permise, sau de la "calitativ" la "cantitativ"), vom putea întotdeauna trata o variabilă aflată la un nivel "superior" de măsurare ca şi cum ar fi fost măsurată la un nivel "inferior". De exemplu, vârsta masurata în ani de viaţă va putea oricând fi tratată ca o variabilă ordinală, dacă îi grupăm valorile (sub 20, 21-30, 31-50, peste 50). Niciodată însă nu vom putea trata o variabilă aflată la un nivel "inferior" ca pe una aflată "mai sus" în ierarhie. (Câteodată, cercetătorii fac excepţie de la această regulă, tratând variabilele ordinale ca şi cum ar fi măsurate la nivel de interval. Totuşi, o dată cu dezvoltarea unor noi tehnici de analiză, dedicate special nivelelor de măsurare "calitativă", aceste practici devin din ce în ce mai rare.)

Bibliografie:

- Clocotici V., Stan, A., Statistică aplicată în psihologie, Polirom, 2000
  - capitolele 1-8
- Rotariu Traian (coordonator), Metode statistice aplicate în ştiinţele sociale, Polirom, 1999
  -capitolele 1-8
- Sandu, Dumitru, Statistică în ştiinţele sociale , Universitatea Bucureşti, 1992
  - capitolele 1, 2, 3, 6, 7

Unitatea 2

Obiectiv: prezentarea analizei univariate a datelor

Cuvinte cheie: tendinţa centrală, indicatori ai tendinţei centrale, indicatori de dispersie sau variaţie

Analiza univariată a datelor

2.2 Tendinţa centrală, variaţia şi forma distribuţiei

În general, o descriere completă a unei variabile se face urmărind trei caracteristici ale acesteia:

a) tendinţa centrală (sau centrul distribuţei) - adică valoarea "tipică" a acelei variabile
b) variaţia variabilei - ca indicator al gradului de "împraştiere" a datelor
c) forma distribuţiei

2.2.1 Indicatori (măsuri) ai tendinţei centrale

Pentru a descrie centrul unei distribuţii, sau tendinţa centrala a unei variabile, există mai multe măsuri. În aceasta secţiune vor fi discutate cele mai des utilizate: **modul**, **mediana** şi **media**.

- *Modul este definit ca fiind valoarea cu frecvenţa cea mai mare a unei distribuţii. Altfel spus, modul este acea valoare a variabilei care apare cel mai des într-un eşantion sau într-o populaţie.*

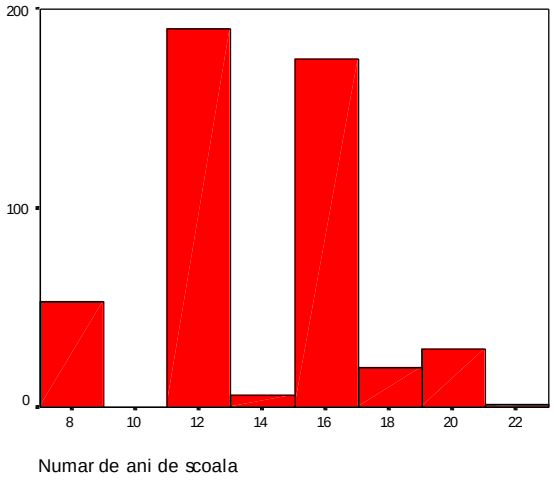
Termenul derivă din francezul "mode", adică modă. În cazul distribuţiei variabilei "starea civilă a capului gospodăriei" reprezentată în Graficul 1.2, modul este valoarea "căsatorit(ă)" (cu frecvenţa relativă 80%). De cele mai multe ori, pentru a simplifica lucrul cu datele, valorilor variabilelor nominale li se acordă convenţional *coduri numerice*. De exemplu, pentru datele din Graficul 1.2, putem acorda codul 1 pentru valoarea "căsatorit(ă)", codul 2 pentru valoarea "uniune consensuală", codul 3 pentru valoarea "divorţat(ă)" etc.. Chiar dacă aceste coduri sunt numerice, ele trebuie privite ca nişte simple simboluri convenţionale. Utilizarea lor nu înseamnă că valorile pot fi ordonate sau că intervalele dintre valori sunt egale. În cazul în care valorile variabilei "stare civilă" ar fi fost codificate ca mai sus, modul ar fi fost valoarea (codul) 1.

Pentru datele din Tabelul 1.2, care prezintă date grupate în intervale, vom vorbi despre un interval modal - şi anume categoria "2001-3000 locuitori", deoarece aceasta este "valoarea" (de fapt intervalul de valori) cu frecvenţa cea mai mare (651).

Grafic, modul este valoarea variabilei căreia îi corespunde "vârful" distribuţiei.

Deşi simplu de obţinut, modul nu este întotdeauna cea mai bună măsură a tendinţei centrale, deoarece de multe ori depinde de gruparea arbitrară a datelor (de exemplu, pentru datele din Tabelul 1.2 am fi obţinut un alt mod dacă datele ar fi fost altfel grupate). De asemenea, nu rareori se întâlnesc *distribuţii bimodale,* în care există două valori diferite ale variabilei care apar cu o aceeaşi "cea mai mare" frecvenţă. Grafic, o distribuţie bimodală este o distribuţie cu două "vârfuri" (Graficul 1.3).

Graficul 1.3 Distribuţie bimodală - histograma variabilei "nivel de educaţie", pentru angajaţii unei bănci



- *Mediana este acea valoare a unei variabile care împarte seria ordonată de date în două părţi egale, astfel încât 50% din observaţii se vor situa deasupra valorii mediane iar 50% dedesubtul ei.*

Să luăm de exemplu notele pe care 7 studenţi le primesc la examenul de statistică (după ce le-am ordonat în prealabil de la minim la maxim): 5, 5, 6, 8, 9, 9, 10. Mediana acestei serii de date este 8, deoarece ea divide seria de date în două parţi egale: 3 dintre studenţi (observaţii) au note mai mici decât 8 şi trei dintre ei au note mai mari. Nota 8 este exact la "mijlocul" seriei de date (după ordonare). Este important de reţinut că ceea ce conteaza pentru stabilirea medianei este numărul de observaţii pe care se face analiza, şi nu numărul de valori ale variabilei.

Calculul medianei este relativ simplu atunci când avem de-a face cu un număr mic şi impar de observaţii. Lucrurile se complică puţin atunci când numărul de observaţii este par, sau dacă numărul de observaţii e foarte mare şi e nevoie să apelăm la tabele de frecvenţe. Lucrurile se complică şi mai mult dacă datele de care dispunem sunt date grupate în intervale, ca în Tabelul 1.2.

În cazul în care avem de-a face cu un număr par de observaţii nu va mai exista o singură valoare la mijlocul seriei de date, ci vom avea două valori. În această situaţie, mediana se află la mijlocul "distanţei" dintre aceste valori, sau cu alte cuvinte, este media lor. Să presupunem că am dori sa calculăm mediana pentru o serie de 8 studenţi, deci un număr par de observaţii. După ordonare, datele arată astfel: 5, 5, 6, 7, 8, 9, 9, 10. La mijlocul seriei se află valorile 7 şi 8. Mediana va fi deci 7,5.

Pentru situaţiile în care suntem nevoiţi să calculăm mediana pe baza datelor oferite de un tabel de frecvenţe, vom utiliza frecvenţele cumulate, şi vom căuta acea valoare a variabilei sub care se află 50% din cazuri. Pentru datele din Tabelul 1.3, 28,75% din observaţii iau valoarea 6 sau o valoare mai mică, 46,25% iau valoarea 7 sau mai puţin, iar 75% iau valoarea 8 sau o valoare mai mică. Rezultă de aici că nota mediană nu poate fi 7 sau altă notă mai mică (deoarece numai 46,25% dintre studenţi iau nota 7 sau mai puţin). Mediana va fi în consecinţă 8, deoarece, chiar dacă avem un număr par de

Tabelul 1.3 Distribuţia notelor pentru 80 de studenţi

Nota	Frecvenţe absolute	Frecvenţe relative (%)	Frecvenţe relative cumulate (%)
3	2	2,5	2,5
4	4	5	7,5
5	7	8,75	16,25
6	10	12,5	28,75
7	14	17,5	46,25
8	23	28,75	75
9	14	17,5	92,5
10	6	7,5	100
Total	80	100	

observaţii, ambele valori care se găsesc la mijlocul seriei de date sunt egale cu 8.

În cazul în care avem de-a face cu un tabel de frecvenţe care conţine date grupate în intervale de valori (aşa cum este Tabelul 1.2), valoarea medianei poate fi calculată cu ajutorul formulei:

$$Me = l + \frac{\frac{N}{2} - nc}{n} \times L$$

unde:

- *Me* este mediana,
- *l* este limita inferioară a intervalului care conţine mediana
- *N* este numărul total de observaţii
- nc este frecvenţa absolută cumulată a tuturor categoriilor care preced intervalul care conţine mediana (adică numărul de observaţii care iau valori mai mici decât *l*)
- *n* este frecvenţa intervalului care conţine mediana
- *L* este lărgimea sau mărimea intervalului care conţine mediana

Exemplu de calcul al medianei pe baza datelor din Tabelul 1.2:

Din tabel reiese ca mediana este conţinută în intervalul 3001-4000 locuitori, deoarece frecvenţele relative cumulate ale categoriilor precedente sunt mai mici de 50%, iar frecvenţa cumulată a intervalului 3001-4000 este aproximativ 63%. Limita inferioară a acestui interval este deci *l* = 3001. Observaţia căreia îi corespunde mediana (numită şi individ median) este observaţia care se află exact la mijlocul seriei ordonate de date, cu alte cuvinte este observaţia *N*/2, în cazul nostru observaţia cu numărul 1343. Dacă scădem din acest număr numărul total de observaţii care au valori mai mici decât 3001, obţinem 1343 - 1084 = 259, unde 1084 = 54 + 379 + 651 este valoarea lui *nc* din formula medianei (obţinut prin cumularea frecvenţelor categoriilor precedente intervalului care conţine mediana). Cu alte cuvinte, observaţia căreia îi corespunde mediana este cea de-a 259-a observaţie din categoria "3001-4000 locuitori", categorie care apare cu frecvenţa *n* = 602. Am putea acum să ne întrebăm: dacă la 602 comune corespunde o creştere a numărului de locuitori cu *L*=1000 (de la 3001 la 4000), atunci la 259 de comune cât va corespunde? Răspunsul e dat de regula de trei simplă, conţinută oarecum şi în formula medianei: $\frac{259}{602} \times 1000 = 430{,}2$. Cu alte cuvinte, mediana este egala cu 3001 + 430 = 3431 locuitori.

Mediana este un caz special de *măsură a localizării.* Măsurile localizării sunt de obicei cunoscute sub numele de *percentile* sau *quantile.* Pentru cazul general, numim *percentila p* acea valoare sub care se află *p*% din cazuri şi deasupra căreia se află *(100-p)*% din cazuri. De exemplu, mediana este percentila 50. Cele mai cunoscute măsuri ale localizării sunt *quartilele, quintilele şi decilele.* Quartilele sunt acele valori ale seriei de date care o împart în patru părţi egale, quintilele sunt valorile care o împart în cinci părţi egale, iar decilele în 10. Sub quartila 1 se află 25% din cazuri, iar deasupra ei 75%. Sub quartila 2 se afla 50% din cazuri, de unde reiese ca această quartilă este chiar mediana. În sfârşit, sub quartila 3 se află 75% din cazuri, iar deasupra ei se află 25% din cazuri (observaţii). Din această scurtă prezentare reiese că există numai 3 quartile (Q1, Q2 şi Q3), deoarece pentru a împărţi o serie de date în *m* părţi egale sunt suficiente *m-1* valori. În statistică quartilele, decilele etc. se referă la *valori ale variabilei*. Totuşi, în ştiinţele sociale sunt folosite destul de des expresii cum ar fi "decila 10 de venituri", "cea mai săracă quintilă", "persoanele aparţinând primei decile" etc. Aceste expresii se referă însă la observaţiile care iau valori cuprinse între anumite percentile (quantile) şi nu la valorile variabilei.

- *Media este probabil cea mai importantă şi totodată cea mai populară măsură a tendinţei centrale a unei distribuţii. Ea se calculează ca sumă a tuturor valorilor observate ale seriei de date împărţită la numărul de observaţii*:

$$\overline{X} = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

unde:
$\overline{X}$ este media
x_i reprezintă valoarea variabilei pe care o ia observaţia *i*
N este numărul total de observaţii
Σ (sigma) este simbolul folosit pentru a indica o sumă

De exemplu, pentru cei 7 studenţi de mai sus, cu notele 5, 5, 6, 8, 9, 9, 10, suma notelor este 52, numărul total de observaţii este 7, iar media va fi 52 împărţit la 7, adică 7,43.

În cazul în care media trebuie calculată pe baza unui tabel de frecvenţe, formula devine:

$$\overline{X} = \frac{\sum_{j=1}^{k} f_j x_j}{N}$$

unde:
k este numărul de categorii (valori) ale variabilei
f_j reprezintă frecvenţa de apariţie a categoriei j
x_j este valoarea categoriei j
N este numărul total de observaţii

De exemplu, pentru datele din Tabelul 1.3, media este:

$$\overline{X} = \frac{2\times3+4\times4+7\times5+10\times6+14\times7+23\times8+14\times9+6\times10}{80} = 7{,}31$$

Pentru cazurile în care media trebuie calculată pentru date grupate în intervale, ca în Tabelul 1.2, se aplică formula de mai sus, considerându-se ca "valori ale variabilei" centrele de interval. Exemplu: pentru categoria "1001-2000 locuitori", centrul de interval este (1001 + 2000) / 2 = 1500,5. Bineînţeles că, pentru un astfel de exemplu, la finalul calculelor media se va rotunji, deoarece atunci când vorbim despre populaţia unei comune nu o putem exprima decât în numere întregi. Atunci când avem de-a face cu date grupate în intervale, probleme pot apărea la calculul centrului de interval pentru prima şi respectiv ultima categorie: în Tabelul 1.2, categoriile "1000 sau mai puţini locuitori", respectiv "peste 8000 de locuitori". Dacă se întîmplă ca valoarea minimă şi respectiv cea maximă a seriei de date să fie cunoscute, atunci nu există practic nici o problemă. Dacă aceste valori nu sunt cunoscute, rămâne la latitudinea cercetătorului să decidă ce valori urmează să atribuie respectivelor centre de interval.

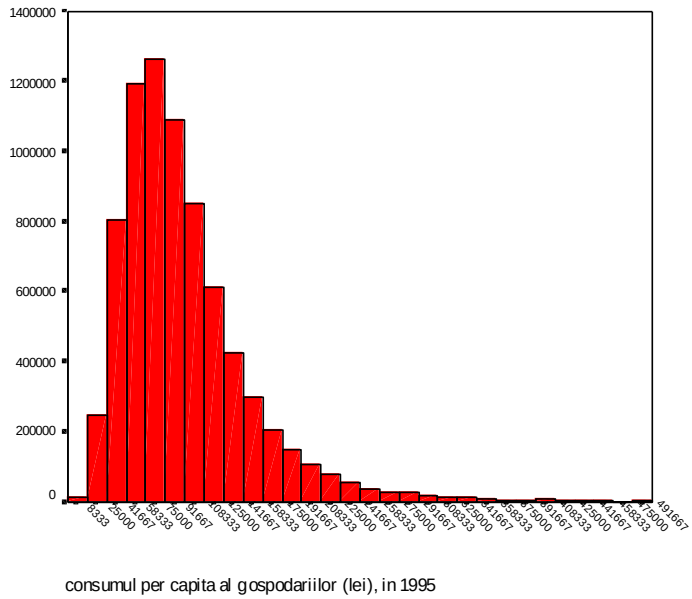
Când folosim una sau alta dintre măsurile tendinţei centrale?

Decizia de a utiliza una sau alta dintre măsurile tendinţei centrale este strâns legată în primul rând de nivelul de măsurare a variabilelor. Aşa cum ne putem da seama, *modul poate fi utilizat pentru toate cele patru niveluri de măsurare*. *Mediana însă nu poate fi utilizată decât pentru nivelele care permit o ordonare prealabilă a datelor*, *adică numai pentru variabilele ordinale, de interval şi de raport*. *În ceea ce priveşte media, aceasta poate fi calculată numai pentru variabilele masurate la ultimele două nivele, adică cel de interval şi respectiv cel de raport, deoarece în cazul celorlalte nivele operaţiile de adunare/scădere a valorilor variabilelor nu sunt permise.*

Un alt element important pentru a decide ce măsură a tendinţei centrale merită folosită este existenţa observaţiilor care au *valori extreme*. De fapt acest aspect este în strânsă legatură cu ***forma distribuţiei***.

Să considerăm de exemplu distribuţia consumului per capita al gospodăriilor, aşa cum este ea reprezentată în Graficul 1.4. Media acestei distribuţii este 103087 lei iar mediana este 87354 lei lei (valorile sunt exprimate în preţuri 1995). În ceea ce priveşte modul, valoarea exactă a acestuia nu are sens să fie calculată deoarece există relativ puţine situaţii în care mai multe gospodării au exact aceeaşi valoare a consumului per capita. Putem însă vorbi despre un interval modal, care se află undeva în jur de 72000 lei.

Graficul 1.4 Distributia consumului per capita al gospodăriilor



Dacă dorim să aflăm valoarea "tipică" a consumului per capita într-o gospodărie pentru o distribuţie ca cea din Graficul 1.4, este mai indicat să utilizăm mediana, deoarece modul de calcul al acesteia este mai apropiat în acest caz de ceea ce înţelegem noi în mod obişnuit prin "centrul distribuţiei": 50% dintre cazuri dedesubt şi 50% deasupra. Mediana are avantajul de a nu fi influenţată de valorile "extreme" ale seriei de date. Media seriei de date reprezentate în Graficul 1.4 este mai mare decât mediana tocmai datorită existenţei unui număr relativ mic de gospodării cu valori foarte mari ale consumului per capita, valori care "trag" media spre dreapta (sau cu alte cuvinte conduc către o valoare mai ridicată a acesteia în raport cu mediana).

În concluzie, putem afirma că modul nu e o măsură foarte adecvată a centrului unei distribuţii. El este util mai ales atunci când avem de-a face cu variabile măsurate la nivel nominal, dar şi în cazurile în care distribuţiile studiate sunt bi- sau multi-modale. Mediana este indicată mai ales în cazurile în care dorim identificarea "valorilor tipice" ale unor distribuţii *asimetrice* (vezi Graficul 1.5, b şi c), care au valori extreme. Media, pe de altă parte, prezintă marele avantaj de a lua în calcul toate valorile unei serii de date. Aceasta este unul din motivele pentru care ea continuă să fie cea mai utilizată măsură a tendinţei centrale. În plus ea mai are şi alte proprietăţi utile, care vor fi discutate în capitolele următoare.

2.2.2 Măsuri ale variaţiei

Măsurile tendinţei centrale sunt esenţiale pentru descrierea unei caracteristici a unui eşantion sau a unei populaţii, însă ele nu sunt suficiente. Pentru descrierea completă a unei variabile este foarte important să ştim deasemenea şi cât de "împrăştiate" sunt valorile acesteia în jurul tendinţei centrale sau, cu alte cuvinte, cât de omogenă respectiv eterogenă este populaţia (eşantionul) studiată în raport cu o anumită caracteristică. Să luăm ca exemplu performanţa la o anumită materie a unei grupe de 80 studenţi, măsurată cu note de la 1 la 10 (datele sunt prezentate în Tabelul 1.3). Nota medie a respectivei grupe este 7,31. Această informaţie însă pare a fi insuficientă pentru a ne putea pronunţa asupra performanţei respectivei grupe. Întrebarea pe care ne-o punem în mod natural este: cât de omogenă este respectiva grupă în ceea ce priveşte performanţa şcolară?

- Un prim răspuns la această întrebare îl putem da prin simpla examinare a intervalului în care sunt cuprinse notele respectivilor studenţi, sau mai bine zis prin calcularea *amplitudinii* variabilei. *Amplitudinea unei variabile este diferenţa dintre valoarea maximă şi valoarea minimă a acelei variabile.* Pentru exemplul nostru, amplitudinea este 10 - 3 = 7 puncte. Deci, cei 80 de studenţi sunt distribuiţi de-a lungul unui interval de şapte puncte.
- O măsură a variaţiei mai rafinată decât amplitudinea o reprezintă *abaterea interquartilă*, care se calculează ca diferenţă între quartila 3 şi quartila 1. *Abaterea interquartilă măsoară împrăştierea celor 50% din observaţii aflate la mijlocul distribuţiei.* Ea are practic aceleaşi avantaje pe care le are şi mediana ca măsură a tendinţei centrale, şi anume nu este influenţată de existenţa cazurilor extreme.
- De cele mai multe ori suntem însă interesaţi să folosim o măsură a variaţiei unei variabile care să includă toate observaţiile, nu numai două dintre ele ca în cazul amplitudinii şi abaterii interquartile. În plus, suntem interesaţi să examinăm variaţia în raport cu o măsură a tendinţei centrale. De obicei, măsurile care satisfac aceste două cerinţe sunt bazate pe *abaterile observaţiilor de la medie. Abaterea de la medie a unei observaţii este diferenţa dintre valoarea pe care o ia respectiva observaţie şi media variabilei ($x_i - \overline{X}$).*Una din proprietăţile mediei este însă aceea că suma tuturor abaterilor individuale de la medie este egală cu 0: $\sum_{i=1}^{n}(x_i - \overline{X}) = 0$ (sau cu alte cuvinte, abaterile pozitive se vor anula cu cele negative). În consecinţă, pentru a obţine o măsură a variaţiei la nivelul întregului eşantion sau a întregii populaţii trebuie utilizată fie suma valorilor absolute ale abaterilor individuale de la medie, fie suma pătratelor acestor abateri.
- *Abaterea medie absolută este definită ca medie aritmetică a abaterilor individuale absolute (ignorând semnul acestora) de la media variabilei:*

$$AMA = \frac{1}{N}\sum\left|x_i - \overline{X}\right|$$

- O altă măsură, mult mai răspândită, este *varianţa* variabilei. *Varianţa (sau dispersia) se defineşte ca fiind media aritmetică a pătratelor abaterilor individuale de la medie:*

$$Varianta = \frac{1}{N}\sum(x_i - \overline{X})^2$$

Din motive teoretice care nu vor fi expuse în acest manual, pentru calcularea varianţei la nivel de eşantion se foloseşte formula:

$$s^2 = \frac{1}{N-1}\sum(x_i - \overline{X})^2,$$

iar pentru date grupate în tabele de frecvenţe (ca în Tabelul 1.3):

$$s^2 = \frac{1}{N-1}\sum (x_j - \bar{X})^2 f_j$$

unde:
x_j este valoarea variabilei pe care o ia grupa j
f_j este frecvenţa absolută de apariţie a lui x_j

- Deoarece varianţa, datorită ridicării la pătrat, este destul de dificil de interpretat, cea mai utilizată măsură a variaţiei unei variabile, pentru scopuri descriptive, este *abaterea standard, definită ca radical de ordinul doi (rădăcină pătrată) din varianţă:*

$$s = \sqrt{s^2}$$

Din formula abaterii standard reiese clar că abaterea standard va fi cu atât mai mare cu cât valorile pe care le iau observaţiile se abat mai mult de la medie. Să considerăm de exemplu notele la o materie a două grupe mici de elevi, ambele serii de date având media 6 şi amplitudinea 8:

Grupa 1: 2, 4, 6, 6, 8, 10
Grupa 2: 2, 2, 5, 7, 10, 10

Întrebarea pe care ne-o putem pune este: cât de omogene sunt cele două grupe? Calculul abaterilor standard arată că în prima grupă $s_1 = 2{,}8$, iar în a doua $s_2 = 3{,}6$. Este clar deci că prima grupă e mai omogenă decât a doua, în care variabilitatea performanţei e mai mare.

În exemplul de mai sus am comparat două grupe de subiecţi din punct de vedere al omogenităţii pentru o aceeaşi caracteristică. Însă atunci când trebuie analizăm omogenitatea unei singure populaţii sau a unui eşantion apar întrebari al căror răspuns e mai dificil de dat: "cum interpretăm magnitudinea abaterii standard?", "când putem spune că avem o abatere standard mică sau una mare?", "cum putem compara omogenitatea unei populaţii pentru două variabile diferite?". Practic, răspunsul la prima întrebare depinde în mare măsură şi de alte caracteristici ale distribuţiei. Pentru un anumit tip de distribuţii interpretarea magnitudinii abaterii standard este mai uşoară, şi acest lucru va fi tratat în Capitolul 2 al acestui manual. În cazul celorlalte două întrebări un răspuns satisfăcător poate fi dat cu ajutorul unei alte măsuri, numite *coeficient de variaţie*, calculat ca raport între abaterea standard şi media unei varibile:

$$CV = \frac{s}{\bar{X}}$$

Prin modul de calcul, coeficientul de variaţie are avantajul de a fi o măsură adimensională (fără unitate de măsură), deoarece unitatea de măsură a abaterii standard este aceeaşi cu cea a mediei. În consecinţă, el este foarte util în compararea variaţiei a două variabile măsurate pe aceaşi populaţie/eşantion. Putem astfel trage concluzii de tipul: "populaţia A este mai eterogenă în privinţa caracteristicii X decât în privinţa caracteristicii Y", concluzii imposibil de formulat numai cu ajutorul abaterii standard deoarece abaterea standard este o măsură dimensională şi deci nu putem compara "mere cu pere" (de exemplu abaterea standard a performanţei şcolare cu abaterea standard a veniturilor familiei). Prin modul său de calcul coeficientul de variaţie indică practic cât la sută din medie corespunde unei abateri standard, ceea ce face mai uşor de evaluat gradul de omogenitate a populaţiei studiate. O populaţie cu o abatere standard egală sau mai mare decât media poate fi considerată în cele mai multe cazuri o populaţie eterogenă, în timp ce o populaţie a cărei abatere standard reprezintă 0,3 (30%) din medie poate fi considerată o populaţie relativ omogenă. Este important însă de reţinut faptul că acest coeficient nu poate fi calculat decât în cazul variabilelor

măsurate la nivel de raport, deoarece în cazul variabilelor nominale şi ordinale abaterea standard nu poate fi calculată, iar în cazul variabilelor măsurate la nivel de interval media este una convenţională, ceea ce face posibilă transformarea variabilei prin adunarea unei constante la valorile acesteia, fără ca semnificaţia valorilor variabilei să se modifice O astfel de transformare ar lăsa nemodificată abaterea standard (lucru care poate fi demonstrat matematic) însă ar modifica media variabilei. Ori aceasta înseamnă că pentru aceeaşi caracteristică am putea calcula coeficienţi de variaţie diferiţi ca valoare.

Exerciţii şi probleme

1. Veniturile gospodăriilor locuitorilor ţării Alfa, care cuprinde 87 de milioane de gospodării, sunt distribuite în jurul unei valori medii de 27000 Alfa-lei şi o mediană de 22000 Alfa-lei.

a. Ce se poate spune despre simetria distribuţiei veniturilor?
b. Care este venitul întregii ţări (toate cele 87 de milioane de gospodării)?

Pentru următoarele întrebări, să se încercuiască varianta corectă /variantele corecte:

2. Decila 5 este o masura a:
   1. tendintei centrale
   2. variatiei
   3. formei distributiei
   4. nici una dintre acestea

3. Valoarea sub care se afla 50% dintre cazurile seriei de date ordonate de la minim la maxim este:
   1. media
   2. quartila 2
   3. modul
   4. abaterea standard
   5. nici una dintre acestea

4. Valorile variabilei ocupatie, intr-un grup de 5 persoane, sunt: 1, 3, 3, 4, 5. Tendinta centrala in acest grup, pentru variabila ocupatie, poate fi descrisa prin:
   1. media egala cu 3,2
   2. mod egal cu 3
   3. mod egal cu 2
   4. mediana egala cu 3
   5. nici una dintre acestea

Unitatea 3

Obiectiv: prezentarea analizei bivariate a datelor

Cuvinte cheie: intensitatea relaţiilor dintre variabile, reducere proporţională a erorii, ranguri

Analiza bivariată a datelor.

Intensitatea relaţiilor dintre variabilele calitative

În secţiunea anterioară am văzut cum putem testa ipoteza existenţei unei relaţii (de asociere) între două variabile calitative. Testul χ^2 ne oferă însă informaţii numai despre existenţa/inexistenţa unei relaţii de asociere între două variabile, dar nu şi despre intensitatea respectivei relaţii, atunci când ea există. Pentru a răspunde la întrebarea "Cât de puternică e relaţia de asociere dintre două variabile?" avem nevoie de măsuri specifice. Două dintre acestea vor fi prezentate în secţiunea care urmează.

- *Cazul variabilelor nominale - coeficientul λ (lambda)*

Să ne întoarcem la datele din Tabelul 6.1 şi să presupunem de această dată că nu cunoaştem decât distribuţia marginală a atitudinii faţă de schimbarea modului de alocare a bugetului (cu alte cuvinte nu ştim decât că 200 de indivizi sunt pentru, 220 sunt împotrivă, iar 180 sunt nehotărâţi). Dacă vom încerca să prezicem atitudinea unui individ oarecare, vom spune firește că repsectivul individ va fi împotriva schimbării modului de alocare a bugetului, deoarece cu o astfel de predicţie avem cele mai reduse şanse de a greşi. Cu alte cuvinte, ne-am bazat predicţia pe frecvenţa modală (cea mai mare frecvenţă). În cazul în care am face o astfel de afirmaţie pentru fiecare din cei 600 de indivizi, predicţia noastră ar fi corectă pentru 220 dintre ei (37%), şi falsă pentru ceilalţi 380. Să presupunem acum că la un moment dat primim o informaţie în plus, şi anume distribuţia atitudinilor faţă de schimbarea modului de alocare a bugetului în funcţie de grupele de vârstă de care aparţin indivizii (adică exact informaţia prezentată în Tabelul 6.1). Să zicem că vom considera ca plauzibilă ipoteza în care atitudinile faţă de modificarea modului de alocare a bugetului sunt dependente de grupa de vârstă a individului. În acest caz, variabila vârstă se va numi variabilă independentă, iar atitudinea faţă de schimbarea modului de alocare a bugetului se va numi variabilă dependentă. Să zicem acum că vom repeta raţionamentul de mai sus (predicţia atitudinii unui individ pe baza frecvenţei modale) pentru fiecare grupă de vârstă în parte. Vom avea deci, din nou, un număr de predicţii corecte şi un număr de predicţii eronate. *Coeficientul λ reprezintă tocmai proporţia cu care se reduce numărul de erori prin introducerea variabilei independente*. Să calculăm acum λ pentru datele din Tabelul 6.1:

Tabelul 6.1 Relaţia dintre două variabile categoriale

Frecvenţe observate				
	Da	Nu	Nu ştiu	Total
cei cu vârsta sub 25	110	40	30	180
cei cu vârstă între 26 şi 45 de ani	40	100	60	200
cei cu varsta peste 45 de ani	50	80	90	220
Total	200	220	180	600

Aşa cum am arătat, în absenţa variabilei independente, numarul de erori e_1 a fost 380. Să vedem acum câte erori am facut prezicând variabila dependentă pe baza valorilor variabilei independente (pentru a uşura urmarirea calculelor, am copiat încă o dată mai jos datele Tabelului 6.1):

- pentru grupa de vârstă sub 25 de ani, vom prezice corect pe baza frecvenţei modale în 110 cazuri, şi vom face erori în 70 de cazuri.
- pentru grupa de vârstă 26 - 45 de ani, vom prezice corect pe baza frecvenţei modale în 100 cazuri, şi vom face erori în alte 100 de cazuri.
- pentru grupa de vârstă peste 45 de ani, vom prezice corect pe baza frecvenţei modale în 90 cazuri, şi vom face erori în 130 de cazuri.

Deci totalul erorilor făcute este $e_2 = 70 + 100 + 130 = 200$.
Să îl calculăm acum pe lambda, după o formulă utilizată şi pentru calculul altor măsuri ale asocierii şi cunoscută sub numele de "*reducere proporţională a erorii*":

$$\lambda = \frac{e_1 - e_2}{e1} = \frac{380 - 200}{380} = 0{,}47$$

Coeficientul λ poate lua, prin modul de construcţie numai valori între 0 şi 1, 0 însemnând absenţa oricărei relaţii între variabile, adică independenţă, iar 1 însemnând intensitate maximă a asocierii (asociere puternică). El este o măsură asimetrică (avem o variabilă independentă pe baza căreia se fac predicţii şi o variabilă dependentă, ale cărei valori sunt prezise), însă există formule de calcul şi pentru varianta simetrică a acestui coeficient. Avantajul lui constă în modul relativ uşor şi intuitiv de calcul. Principalul dezavantaj al acestei măsuri este faptul că în condiţiile în care o categorie a unei variabile conţine un număr foarte mare de indivizi, λ poate fi egal cu 0 chiar dacă cele două variabile nu sunt independente.

- *Cazul variabilelor ordinale*

În cazul variabilelor ordinale, aşa cum am văzut în introducerea acestui manual, există posibilitatea de ordonare a valorilor variabilelor, şi în consecinţă există posibilitatea de a da *ranguri* indivizilor în funcţie de valorile pe care aceştia le au pentru o variabilă. Măsurile Ca urmare, în analiza acestui tip de variabile vom putea vorbi de un semn al asocierii (sau sensul asocierii). Măsurile de asociere a variabilelor ordinale pot lua valori cuprinse între -1 şi 1. La modul general vorbind, o măsură a asocierii dintre două variabile ordinale va fi pozitivă dacă un individ cu un rang mare pentru variabila X tinde să aibă un rang mare şi pentru variabila Y, iar indivizii cu ranguri mici pe variabila X au de asemenea ranguri mici şi pentru Y. asocierea negativă apare atunci când indivizii cu rang mare pentru variabila X tind să aibă ranguri mici pentru Y şi invers. Dacă o măsură a asocierii dintre două variabile ordinale ia valoarea 0, atunci vom spune că cele două variabile sunt independente. Cu cât o relaţie de asociere între două variabile ordinale va fi mai puternică, cu atât măsura asocierii va fi mai mare în valoare absolută (mai aproape de 1). În cele ce urmează ne vom rezuma la a prezenta câteva noţiuni de bază care se referă la măsurile de asociere între variabile ordinale şi la a arăta modul de calcul pentru o astfel de măsură.

O pereche de observaţii se numeşte *concordantă* dacă individul care are un rang mai înalt pe o variabilă are un rang mai înalt şi pe a doua variabilă.

O pereche de observaţii se numeşte *discordantă* dacă individul care are un rang mai înalt pe o variabilă are un rang mai coborât pe cealaltă variabilă.

Să presupunem că avem 4 elevi, ierarhizaţi după calificativele la două materii:

Elevii	Materia X	Materia Y
A	Foarte bine	Bine
B	Bine	Foarte bine
C	Satisfăcător	Satisfăcător
D	Suficient	Suficient

Să încercăm acum să numărăm perechile concordante şi perechile discordante, şi pentru aceasta să începem cu toate perechile de observaţii pe care le putem forma cu elevul A: Acestea sunt: perechea AB (discordantă, deoarece A are un rang mai înalt decât B pe variabila X, dar un rang mai coborât decât B pe variabila Y), perechea AC (concordantă) şi perechea AD (concordantă).

Să trecem acum la perechile lui B: Acestea sunt BC (concordantă) şi BD (concordantă). În fine, trecem acum la perechile lui C, adică la CD (concordantă). În total am avut 6 perechi, din care una discordantă iar 5 concordante. Să calculăm acum o măsură simplă de asociere între cele două variabile (calificativele la materiile X şi Y), numită coeficientul τ_a al lui Kendall:

$$\tau_a = \frac{nc - nd}{nt}$$

unde
nt este numărul total de perechi
nc este numărul de perechi concordante
nd este numărul de perechi discordante

În concluzie, pentru exemplul nostru (care este unul pur didactic), $\tau_a = 4/6 = 0{,}66$.

Aceasta a fost practic cea mai simplă ilustrare de măsură de asociere a două variabile ordinale. În practică însă, lucrurile stau puţin mai complicat, pentru că deseori apar ceea ce se numesc ranguri "legate" sau egale. Acest lucru complică destul de mult calculele şi formulele, însă principiul rămâne acelaşi, al comparării numărului de perechi concordante cu numărul de perechi discordante.

Exerciţii şi probleme

1. Într-un studiu asupra modului în care ocupaţia se asociază cu educaţia, s-a realizat următorul eşantion aleator de 500 de bărbaţi anagajaţi.

	Ocupaţia			
Educaţia	*Funcţionari*	*Muncitori în fabrică*	*Angajaţi în servicii*	*Agricultori*
4 sau mai mulţi ani de liceu (incluzând şi formarea vocaţională)	194	146	27	10
Mai puţin de patru ani de liceu	18	79	18	8

a. Explicitaţi în cuvinte ipoteza de nul H_0
b. Calculaţi χ^2 şi valoarea p pentru H_0

2. Se da tabelul:

		somaj		
		da	nu	total
sex	femei	30%	70%	100%
	barbati	30%	70%	100%
	total	30%	70%	100%

Care din propozitiile urmatoare sunt adevarate?
1. 30% dintre femei sunt somere
2. 30% dintre someri sunt barbati
3. 70% din totalul populatiei se afla in somaj

4. probabilitatea ca o persoana din populatie sa fie in somaj este de 0.3
nici una dintre acestea

3. Dacă variabilele nominale x şi y nu sunt independente statistic atunci este de aşteptat ca:
1. Distribuţiile condiţionate ale lui y funcţie de x să fie diferite de distribuţia marginală a lui y
2. Distribuţiile condiţionate ale lui y , funcţie de x să fie egale între ele
3. Corelaţia Bravais-Pearson dintre x şi y să fie semnificativ diferită de 0
4. Statistica test chi-patrat să difere semnficativ de 0
5. Răspunsurile 1,2,3,4 să fie incorecte

Modulul 4

Obiectiv: prezentarea problematicii regresiei lineare în analiza datelor

Ghid de studiu:

- Regresia lineară simplă
- Construcţia dreptei de regresie
- Regresia lineară multiplă
- Interpretarea coeficienţilor dreptei de regresie

Unitatea 1

Obiectiv: prezentarea problematicii regresiei lineare simple

Cuvinte cheie: dreaptă de regresie, criteriul celor mai mici pătrate, panta asociată variabilei independente, coeficientul de determinaţie şi coeficientul de corelaţie Pearson

Regresia lineară simplă

Fiind cunoscute valorile a două variabile cantitative pentru o mulţime de unităţi de analiză, este posibil să reprezentăm complet această informaţie printr-un grafic. Variabilei dependente îi corespunde axa verticală, iar celei independente îi corespunde axa orizontală. Fiecare unitate de analiză este reprezentată printr-un punct care se află la o distanţă de axa verticală proporţională cu valoarea variabilei independente luată de acea unitate, şi la o distanţă de axa orizontală proporţională cu valoarea variabilei dependente. Astfel, în exemplul precizat anterior, dacă variabila DIF are valorile exprimate în valori procentuale, iar variabila SUM este exprimată în mii de lei, o localitate în care s-au cheltuit 5000 de lei pe cap de locuitor, şi în care şomajul a scăzut cu două procente, se află cu două unităţi deasupra axei orizontale şi la cinci unităţi în dreapta axei verticale.

Foarte adesea, informaţia cuprinsă într-un grafic de acest tip este prea bogată pentru a putea fi analizată direct. La fel cum în cazul unei singure variabile este util să reducem informaţia

reprezentată de distribuţia sa la o singură valoare, cea a tendinţei centrale, exprimată prin medie, mediană sau un alt indicator, şi în cazul considerării simultane a două variabile ar fi de folos să putem descrie într-un mod cât mai succint relaţia dintre acestea.

O soluţie simplă este aceea de a înlocui norul de puncte de pe grafic printr-o singură dreaptă care să îi aproximeze forma cât mai bine. În secţiunea care urmează, 7.1.1., vom arăta cum poate fi construită o astfel de dreaptă, numită *dreaptă de regresie*. Vom prezenta apoi interpretarea coeficienţilor prin care este descrisă dreapta de regresie. În secţiunea 7.1.2. vor fi definiţi indicatori prin care poate fi apreciat gradul de acurateţe prin care o dreaptă de regresie descrie relaţia dintre două variabile. În ultima secţiune a acestei părţi, 7.1.3., va fi discutate una dintre condiţiile mai importante care trebuie îndeplinită pentru ca modelele de regresie să poată fi aplicate.

Construcţia dreptei de regresie

Fie un grafic pe care sunt reprezentaţi mai mulţi indivizi statistici, în funcţie de valorile a două variabile cantitative, **X** şi **Y**, şi fie o dreaptă dusă la întâmplare pe acest grafic. Poziţia fiecărui individ **i** este fixată de valorile pe care iau cele două variabile, notate cu $\mathbf{x_i}$ şi $\mathbf{y_i}$.
Poziţia dreptei în raport cu cele două axe ale graficului este complet precizată de următoarea relaţie:

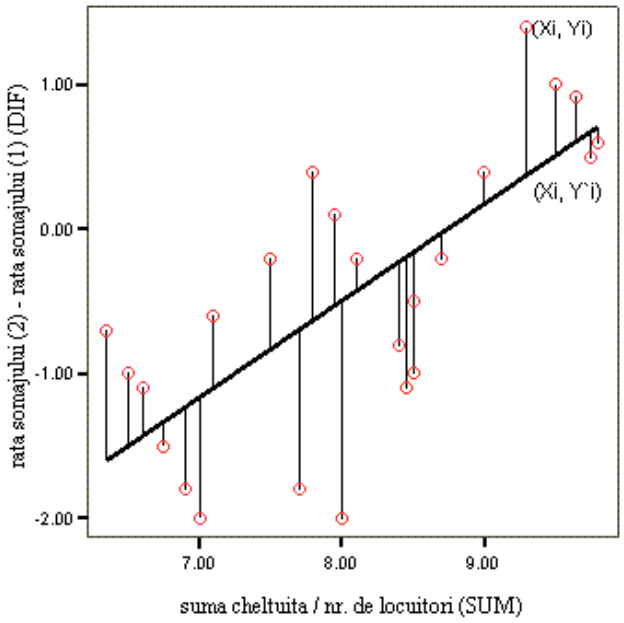
$$\mathbf{Y' = a + bX.} \qquad \mathbf{(1)}$$

Relaţia exprimă faptul că orice punct **k** de pe dreaptă, are coordonatele $\mathbf{x_k}$ şi $\mathbf{y'_k}$ astfel încât $\mathbf{y'_k = a + bx_k}$. Mai mult, orice punct de pe grafic pentru care are loc relaţia anterioară între coordonatele sale, se află pe dreaptă.

De aici rezultă faptul că orice dreaptă este identificată complet prin doar două valori, cea a constantei **a**, şi cea a constantei **b**. Dacă ar fi posibilă înlocuirea unui nor de **n** puncte, care oferă o reprezentare precisă a **n** perechi de valori, printr-o dreaptă care să indice forma de ansamblu a mulţimii de puncte, atunci ar fi obţinută o simplificare remarcabilă a modului în care este descrisă relaţia.

În Figura 7.1 sunt reprezentate localităţile din exemplul discutat anterior, caracterizate de valorile variabilei dependente DIF, respectiv a variabilei independente SUM. Pe grafic este trasată şi o dreaptă (d) precum şi o mulţime de segmente verticale, fiecare fiind construit astfel încât să unească punctul care corespunde unei localităţi cu dreapta (d).

Figura 7.1. Reprezentarea grafică a variabilelor DIF şi SUM, care iau valori pentru 25 de localităţi.



Dacă pentru două variabile cantitative am putea construi o dreaptă astfel încât toate punctele care corespund unităţilor de analiză să se afle pe dreaptă, atunci dreapta ar oferi o descriere completă a formei norului de puncte. Într-un astfel de caz, fiecare dintre segmentele verticale dintre puncte şi dreaptă ar avea lungimea zero.
Este clar că în exemplul considerat aici nu există o astfel de dreaptă, care să descrie perfect relaţia dintre cele două variabile. Ar fi de dorit atunci, să fie determinată acea dreaptă pentru care lungimile segmentelor verticale dintre puncte şi dreaptă să fie cât mai apropiate de zero.

Prin definiţie, *dreapta cu proprietatea că pătratele lungimilor segmentelor dintre puncte şi dreaptă au suma minimă este numită* ***dreaptă de regresie.***

Datorită modului în care este definită, se spune despre dreapta de regresie că satisface *criteriul celor mai mici pătrate*.
Se poate demonstra matematic faptul că pentru două variabile date există o dreaptă unică de regresie, iar aceasta poate fi determinată. Cu alte cuvinte, oricare ar fi două variabile **X** şi **Y**, care iau valori pentru **n** unităţi de analiză, pot fi deteminate în mod unic constantele **a** şi **b** astfel încât dreapta

$$\mathbf{Y' = a + bX,} \qquad \mathbf{(2)}$$

să ofere o cea mai bună aproximare a relaţiei dintre **X** şi **Y**--din perspectiva criteriului *celor mai mici pătrate*--, dintre toate dreptele posibile.

Y' este o variabilă care se obţine din intersecţia segmentelor verticale care trec prin punctele $(\mathbf{x_i}, \mathbf{y_i})$ de pe grafic şi dreapta de regresie, iar punctele de intersecţie sunt de forma $(\mathbf{x_i}, \mathbf{y'_i})$. Datorită modului în care este construită variabila **Y'**, valorile sale sunt identice cu ale lui **Y** atunci când punctele sunt pe o dreaptă, şi sunt cu atât mai diferite de cele ale lui **Y** cu cât punctele sunt mai dispersate în jurul dreptei de regresie.

Un alt mod de a scrie expresia (2) este următorul:

$$\mathbf{Y = a + bX + U,}$$

unde **U = Y - Y'**.
U este o variabilă care pentru fiecare unitate de analiză ia o valoare egală cu lungimea segmentului vertical dintre punctul care îi corespunde pe grafic şi dreapta de regresie.

În exemplul anterior, **a = -5,86**, **b = 0,67**. Ecuaţia dreptei de regresie este

$$\mathbf{DIF = - 5,86 + 0,67SUM.}$$

Interpretarea coeficienţilor dreptei de regresie

Coeficientul **b** este numit **panta asociată variabilei X** şi, aşa cum se poate vedea din expresia dreptei de regresie, *reprezintă numărul de unităţi cu care variază **Y'** atunci când **X** creşte cu o unitate*:

dacă avem două puncte $\mathbf{(x_1, y'_1)}$ şi $\mathbf{(x_2, y'_2)}$, $\mathbf{x_2 = x_1 + 1}$, şi ambele puncte sunt pe dreapta

$$\mathbf{Y' = a + bX,}$$

atunci, înlocuind în formula dreptei se obţine

$$\mathbf{y'_2 = a + bx_2 = a + b(x_1 + 1) = a + bx_1 + b = y'_1 + b.}$$

În exemplul discutat anterior, valoarea lui **b** indică faptul că o creştere a sumei cheltuite pe cap de locuitor cu o mie de lei conduce în medie la o creştere a diferenţei cu 0,67, adică la o scădere a ratei şomajului cu 0,67 de puncte procentuale.

Semnul plus al lui **b** indică faptul că între **X** şi **Y** are loc o relaţie pozitivă--adică valorilor mici ale lui **X** tind să le corespundă valori mici ale lui **Y**, iar valorilor mari ale lui **X** tind să le corespundă valori mari ale lui **Y**--, în timp semnul minus semnalează prezenţa unei relaţii negative.

$\mathbf{b_i}$ = 0 se obţine atunci când forma norului de puncte nu poate fi aproximată printr-o dreaptă. O situaţie de acest gen apare atunci când cele două variabile estimează fenomene independente, fără legătură, dar şi în cazul în care variabilele sunt într-o relaţie a cărei formă nu este liniară (de exemplu, atunci când punctele sunt pe o curbă în formă de parabolă). Cele două cazuri sunt ilustrate în Figura 7.2., respectiv în Figura 7.3.

Figura 7.2. Exemplul a două variabile cantitative între care nu are loc o relaţie.

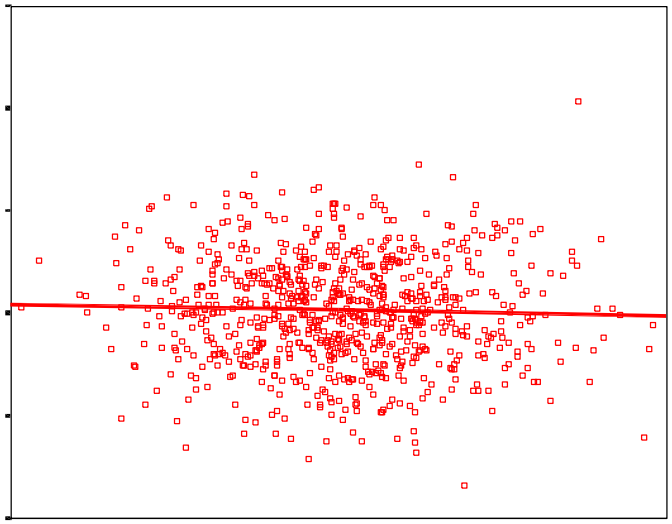
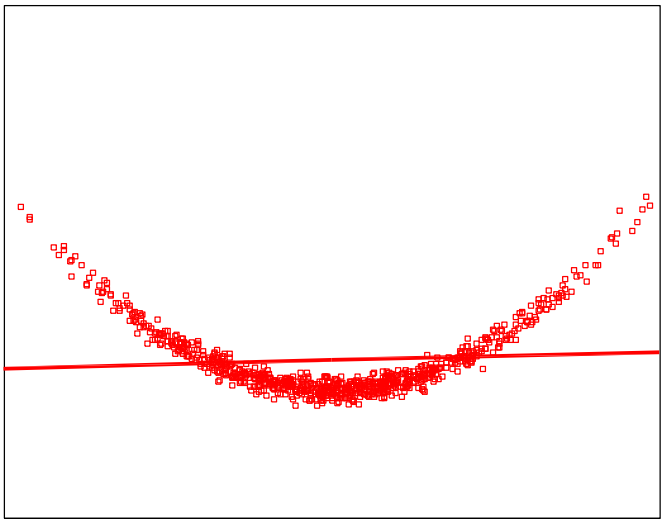


Figura 7.3. Exemplul a două variabile între care există o relaţie (de forma $Y' = X^2$) care nu poate fi



aproximată printr-o dreaptă de regresie.

*Coeficientul **b** are următoarea proprietate importantă: valoarea sa depinde de unităţile de măsură ale celor două variabile.*
Astfel, dacă **SUM** din exemplul anterior ar fi exprimat *în unităţi monetare / numărul de locuitori*, adică într-o unitate de măsură de o mie de ori mai mică decât cea din exemplu, $\mathbf{b_1}$ ar fi de 1000 de ori mai mic. În general, se poate arăta că,

dacă în loc de **X** avem **cX + d**, atunci în loc de **b** avem **b / c**.

Din această proprietate rezultă faptul că panta de regresie nu poate fi folosită drept un indicator al intensităţii relaţiei dintre variabila dependentă şi variabila independentă.

Constanta **a** din ecuaţia dreptei de regresie indică valoarea **y'** pe care o ia un punct pentru care **x = 0** şi care este aflat pe dreaptă.

Indicatori ai intensităţii relaţiei dintre două variabile cantitative: *coeficientul de determinaţie* şi *coeficientul de corelaţie Pearson*

Dreapta de regresie asociată relaţiei dintre două variabile cantitative oferă o imagine sintetică despre forma acestei relaţii, însă nu oferă informaţii despre cât de asemănătoare este această imaginea simplificată cu cea reală. Am întâlnit o situaţie similară în cazul mediei: acest indicator descrie succint tendinţa centrală a distribuţiei unei variabile cantitative, însă nu cuprinde informaţii despre cât de completă este această reprezentare. În acest caz, există un indicator care arată cât de dispersate sunt valorile luate de variabilă în jurul mediei: *abaterea standard*. Cu cât valorile sale sunt mai mici cu atât media descrie mai precis distribuţia variabilei.

În Figura 7.4. şi în Figura 7.5. sunt reprezentate relaţiile dintre câte două perechi de variabile cantitative. În ambele cazuri ecuaţia dreptei de regresie este aceeaşi:

Y = 2 - 2,5 X.

Se observă însă că unităţile de analiză din Figura 7.5. sunt mai dispersate în raport cu dreapta de regresie decât cele din Figura 7.4. Acest fapt arată că dintre cele două drepte, cea din Figura 7.4. oferă reprezentarea cea mai precisă a relaţiei dintre perechea de variabile cărora le corespunde.

Figura 7.4. Distribuţia a două variabile cantitative şi dreapta lor de regresie (A).

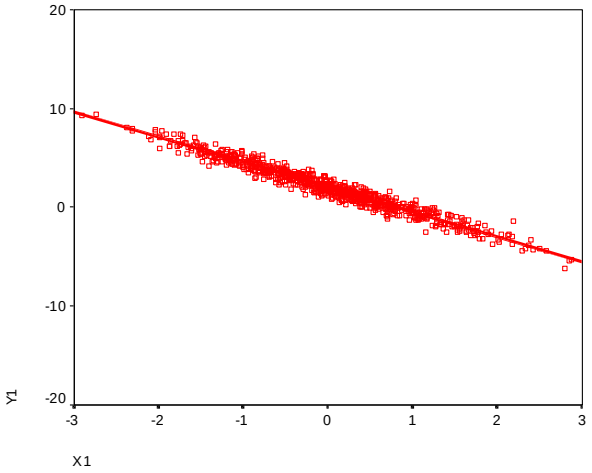
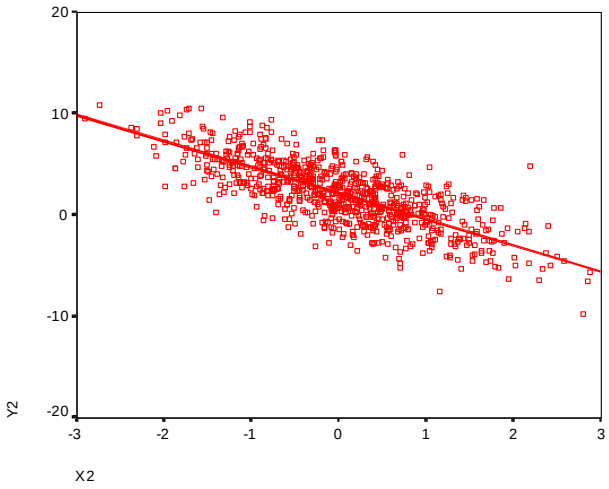


Figura 7.5. Distribuţia a două variabile cantitative şi dreapta lor de regresie (B).



Puterea explicativă a unui model de regresie simplă poate fi evaluată cu ajutorul mai multor indicatori. Coeficientul $\mathbf{R^2}$, numit **coeficient de determinaţie**, este definit de următoarea formulă:

$$R^2 = \frac{\sum (Y' - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

Numărătorul expresiei reprezintă variaţia lui **Y** care este "explicată" de ecuaţia de regresie, în timp ce valoarea de la numitor este egală cu variaţia totală a lui **Y**. Deci, $\mathbf{R^2}$ indică proporţia din variaţia lui **Y** care este "explicată" de variabila independentă.

Din modul în care este definit rezultă că $\mathbf{R^2}$ poate să ia valori între 0 şi 1. $\mathbf{R^2}$ este egal cu 1 atunci când distribuţia punctelor se face după o dreaptă. Valoarea sa este zero în situaţii cum sunt cele ilustrate în Figura 7.2. şi în Figura 7.3., adică atunci când distribuţia punctelor nu poate fi aproximată printr-o dreaptă. În general, cu cât valorile lui $\mathbf{R^2}$ sunt mai apropiate de 1, cu atât relaţia dintre cele două variabile este mai intensă iar reprezentarea sa grafică este mai apropiată de o dreaptă.

În exemplul din secţiunea anterioară $\mathbf{R^2 = 0,53}$.
$\mathbf{R^2}$ care corespunde relaţiei reprezentate în Figura 7.4. are valoarea 0,95, în timp ce $\mathbf{R^2}$ din Figura 7.5. are valoarea 0,58.

Un alt indicator al intensităţii relaţiei dintre două variabile cantitative este **coeficientul de corelaţie Pearson**, notat cu **r** şi definit prin următoarea formulă:

$$\mathbf{r = b\ \sigma_X / \sigma_Y.}$$

$\boldsymbol{\sigma_X}$ şi $\boldsymbol{\sigma_Y}$ reprezintă abaterea standard a variabilei **X**, respectiv abaterea standard a variabilei **Y**.

Coeficientul de corelaţie are două proprietăţi din care poate fi dedus şi modul său de interpretare:

1. $\mathbf{r^2 = R^2}$ --coeficientul de corelaţie Pearson ridicat la pătrat este egal cu coeficientul de determinaţie.
2. **r** are acelaşi semn cu **b**, deoarece cele două abateri standard din definiţia sa au întotdeauna semn pozitiv.

Astfel, din proprietatea (1) rezultă că **r** ia valori în intervalul [-1, 1], iar valorile extreme sunt luate în acelaşi situaţii în care $\mathbf{R^2}$ ia valoarea 1: atunci când relaţia dintre cele două variabile cantitative este de intensitate maximă şi punctele care reprezintă grafic unităţile de analiză sunt distribuite pe o dreaptă. În mod similar, **r** ia valoarea 0 atunci când $\mathbf{R^2}$ este nul, adică în situaţiile în care distribuţia unităţilor de analiză nu poate fi aproximată printr-o dreaptă (Figurile 2. şi 3. ilustrează situaţii în care **r** este 0).

Din proprietatea (2) rezultă că **r** ia valori pozitive atunci când dreapta de regresie are o înclinaţie ascendentă de la stânga spre dreapta, şi valori negative atunci când inclinaţia este descendentă.

Unitatea 2

Obiectiv: prezentarea problematicii regresiei lineare multiple

Cuvinte cheie: coeficienţii de regresie standardizaţi, coeficient de determinaţie multiplă, multicoliniaritate, variabile "dummy".

Regresia lineară multiplă

Modelul de regresie simplă este folosit pentru a descrie relaţia dintre două variabile cantitative. În cazul în care sunt disponibile date despre mai mulţi factori cu potenţial explicativ, iar aceştia sunt estimaţi prin variabile cantitative, este de dorit ca analiza să cuprindă simultan toate variabilele şi nu doar două dintre acestea. Utilizarea regresiei simple într-un astfel de caz, prin ignorarea unora dintre variabilele independente, ori prin aplicarea succesivă pentru fiecare dintre variabilele independente, poate să conducă la rezultate eronate.

Exemplul următor ilustrează o situaţie de acest tip.

Să presupunem că în evaluarea unui program prin care s-a urmărit reducerea şomajului se cunoaşte variaţia ratei şomajului (DIF), suma cheltuită raportată la numărul de locuitori (SUM), şi, în plus, faţă de exemplu similar descris în secţiunea precedentă, fiecare localitate este descrisă de un indicator global al calităţii administrării programelor locale, altele decât cel evaluat aici. Acest din urmă indicator, notat CALIT, este de tip cantitativ, si are trei valori: 1 desemnează un nivel scăzut, 2 un nivel mediu, iar 3 un nivel ridicat al calităţii administrării programelor.

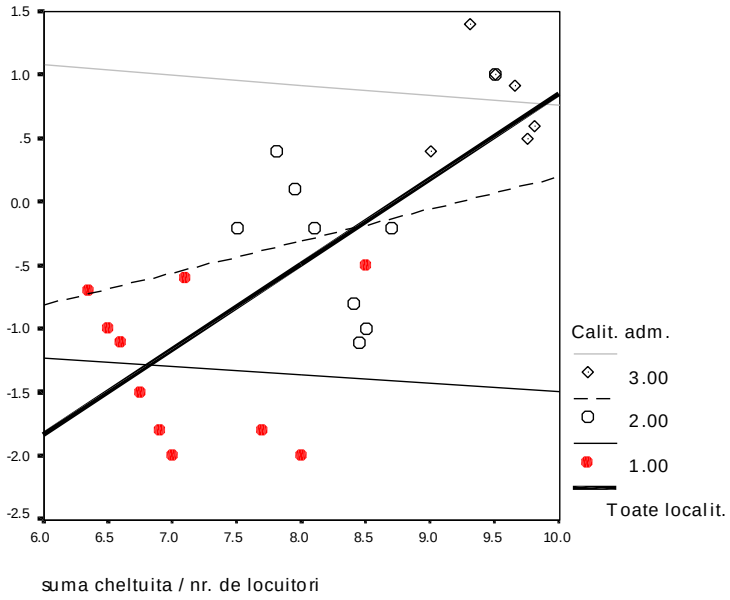
În Figura 7.7 sunt reprezentate localităţile cuprinse în studiu, în funcţie de cele trei variabile. La fel ca şi în Figura 7.8 valorile lui DIF sunt pe axa verticală, iar cele ale lui SUM pe axa orizontală. Marcarea localităţilor pe grafic se face prin simboluri grafice diferite în funcţie de valorile celei de a treia variabile.

Analiza legăturii dintre DIF şi SUM printr-o regresie simplă conduce la concluzia că relaţia dintre cele două variabile este directă, şi destul de intensă ($R^2 = 0,53$).

Dacă, însă, relaţia dintre cele două variabile este studiată separat pe grupele de localităţi desemnate prin valorile celei de-a treia variabile, concluzia este diferită. În locul unui singur model, vom urmări parametrii a trei modele de regresie simplă, câte unul pentru fiecare dintre valorile variablei CALIT. Valorile lui R^2 care se obţin sunt 0,006 pentru CALIT = 1, 0,005 pentru CALIT = 2, şi 0,004 pentru CALIT = 3. Cele trei valori indică faptul că intensitatea relaţiilor dintre DIF şi SUM pentru fiecare dintre cele trei categorii de localităţi este foarte aproape de zero. Altfel spus,

când sunt comparate localităţi care sunt asemănătoare din punctul de vedere al performanţei administrării de programe, cheltuirea unei sume mari pe cap de locuitor nu este asociată, în medie, unei scăderi mai accentuate a ratei şomajului decât în localităţile în care suma a fost mai mică. Acest rezultat indică, contrar celui obţinut din analiza doar a primelor două variabile, că programul de reducere a şomajului nu a fost eficient.

Figura 7.7. Relaţia dintre variabilele DIF, SUM, şi CALIT pentru 25 de localităţi.



Problema generală pe care încercăm să o rezolvăm prin modelare statistică poate fi redusă adesea la următoarea exprimare:

***B** este un fenomen care trebuie explicat iar **A_1, A_2**, ... sunt factori explicativi potenţiali; Care este efectul independent al fiecărui **A_i** asupra lui B? Care este ierarhia importanţei factorilor **A_1, A2**, ... în explicarea lui **B**?*

Exemplul de mai sus arată faptul că numai prin modele care cuprind simultan toate variabilele relevante pentru fenomenul studiat poate fi evaluat efectul independent al fiecăreia. Modelele multivariate cele mai simple şi de aceea cel mai usor de interpretat sunt cele de regresie multiplă. Vom arăta modul în care acestea sunt definite (7.2.1), felul în care pot fi interpretate relaţiile dintre variabilele cuprinse în model (7.2.2.) şi cum poate fi evaluată eficienţa de ansamblu a modelelor (7.2.3). În secţiunea (7.2.4.) vor fi discutate modalitaţile de generalizare a rezultatelor obţinute pe un eşantion iar în sectiunea (7.2.5) va fi descrisă problema multicoliniarităţii. În secţiunea (7.2.6) va fi prezentată o extindere a modelelor de regresie pentru variabile nominale şi ordinale.

Definirea modelelor de regresie multiplă

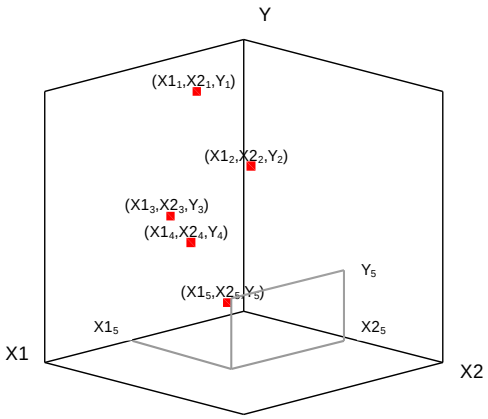
Fie **Y**, $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_m}$, variabile cantitative. **Y** este variabila a cărei variaţie incercăm să o explicăm iar $\mathbf{X_1}$, $\mathbf{X_2}$,..., $\mathbf{X_m}$, sunt variabilele independente. Putem scrie următoarea relaţie între variabile:

$$\mathbf{Y = a + b_1X_1 + b_2X_2 + ... + b_mX_m + U} \qquad \mathbf{(3)}$$

unde **a**, $\mathbf{b_1}$, $\mathbf{b_2}$, ..., $\mathbf{b_m}$ sunt numere iar **U** este o variabilă.

Se observă că pentru orice combinaţie de numere **a**, $\mathbf{b_1}$, $\mathbf{b_2}$, ..., $\mathbf{b_m}$, relaţia (3) este asevărată, pentru că acestea împreună cu valorile lui **Y**, $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_m}$ determină **U**.

Figura 7.8. Reprezentarea grafică în trei dimensiuni a unor cazuri caracterizate de trei variabile.



Dacă **m=2** relaţia (3) poate fi descrisă grafic printr-un desen tridimensional. Fiecărui individ statistic îi corespunde un punct de coordonate $\mathbf{(X_1,X_2,Y)}$ (Figura 7.8.), iar **a**, $\mathbf{b_1}$ şi $\mathbf{b_2}$ definesc un plan descris de ecuaţia

$$\mathbf{Y' = a + b_1X_1 + b_2X_2.}$$

Variabila **U** este determinată de acest plan şi de punctele de forma $\mathbf{(X_1,X_2,Y)}$ într-un mod analog cazului cu două dimensiuni:
valoarea $\mathbf{U_i}$ care îi corespunde unui individ statistic care a luat valorile $\mathbf{X_{1i}}$, $\mathbf{X_{2i}}$,$\mathbf{Y_i}$, este egală cu lungimea segmentului paralel cu axa **OY** care are la extremităţi punctul care îi corespunde în spaţiu $\mathbf{(X_{1i},X_{2i},Y_i)}$, respectiv punctul de intersecţie cu planul (si care are coordonatele $\mathbf{(X_{1i},X_{2i},Y'_i)}$).

Expresia (3) indică faptul că **Y** poate fi exprimată ca o combinaţie liniară de $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_m}$, şi o variabilă **U** numita *variabilă reziduală*. Dacă fixăm **a**, $\mathbf{b_1}$, $\mathbf{b_2}$,..., $\mathbf{b_m}$ atunci **U** poate fi exprimat în funcţie de aceste numere şi **Y,** $\mathbf{X_1}$, $\mathbf{X_2}$,...$\mathbf{X_m}$:

$$\mathbf{U = Y - (a + b_1X_1 + b_2X_2 + ... + b_mX_m)} \qquad \mathbf{(4)}$$

Dacă notăm expresia din paranteza cu **Y'** atunci

$$\mathbf{U = Y - Y'.}$$

Vom alege din mulţimea (infinită) a expresiilor de forma (3) acea combinaţie liniară pentru care **U** (determinat din (4)) are valori minime. Pentru că **U** este o variabilă, atunci când expresia (3) se aplică unui număr de **n** indivizi statistici, **U** este un şir de **n** numere. Avem nevoie să definim un criteriu după care variabilele **U** să poată fi comparate astfel încât să putem alege un **U** având valorile cele mai mici. Principiul folosit pentru modelele de regresie liniară multiplă este, la fel ca şi în cazul bivariat, cel *al celor mai mici pătrate*:

Unei variabile **U** îi corespunde un număr **u** obţinut din aplicarea formulei
$\mathbf{u = u_1^2 + u_2^2 + ... + u_n^2}$, unde $\mathbf{u_i}$ este valoarea luată de **U** pentru cazul statistic **i**;
este ales **U** pentru care **u** este cel mai mic. Din (3) rezultă că problema este echivalentă cu determinarea valorilor **a**, $\mathbf{b_1}$, $\mathbf{b_2}$, ..., $\mathbf{b_m}$ astfel încât **u** să fie minim. Este important de reţinut că pentru orice număr de variabile independente **m**, **a**, $\mathbf{b_1}$, $\mathbf{b_2}$, ..., $\mathbf{b_m}$ sunt determinaţi în mod unic de condiţia de a avea **u** minim.

Ecuaţia

$$\mathbf{Y' = a + b_1X_1 + b_2X_2 + ... + b_mX_m,} \qquad \mathbf{(5)}$$

în care a, b_1, b_2, ..., b_m sunt determinate în urma aplicării principiului celor mai mici pătrate este un model de regresie multiplă.

Dacă avem o singura variabilă independentă (**m=1**) ecuaţia (5) descrie un model de regresie simplă.

În exemplul de mai sus, planul care aproximeaza cel mai bine - după criteriul celor mai mici pătrate -, distribuţia punctelor din spaţiu asociate celor 25 de unităţi de analiză (localităţi cuprinse în studiu) are ecuaţia

DIF = -2,61 + 0,05 SUM + 1,00 CALIT.

Interpretarea modelelor de regresie multiplă

Coeficientul $\mathbf{b_i}$, numit panta asociată variabilei $\mathbf{X_i}$, reprezintă numărul de unităţi cu care variază **Y'** atunci când $\mathbf{X_i}$ creşte cu o unitate iar celelalte variabile independente sunt menţinute constante. Altfel spus, $\mathbf{b_i}$ arată cum se modifică valoarea aşteptată a variabilei dependente atunci când $\mathbf{X_i}$ variază iar $\mathbf{X_j}$ sunt constante, $j \neq i$.
Deasemenea, în măsura în care datele satisfac anumite proprietăţi (dintre care un principiu important este cel al distribuţiilor apropiate de cele normale, iar un alt principiu, al multicoliniarităţii, va fi discutat în secţiunea 7.2.5.) este corect să afirmăm că $\mathbf{b_i}$ indică variaţia în mediile valorilor lui **Y** care corespund punctelor de forma $\mathbf{(X_1,...,X_i,...,X_m)}$ respectiv $\mathbf{(X_1,...,X_i + 1,...,X_m)}$. La fel, **a** arată care este media lui **Y** atunci când $\mathbf{X_1 = X_2 = ... = X_m = 0}$.
Semnul plus al lui $\mathbf{b_i}$ indică faptul că între $\mathbf{X_i}$ şi **Y** are loc o relaţie pozitivă în condiţii de control al efectului celorlalte variabile, în timp ce semnul minus indică prezenţa unei relaţii negative.

In exemplul anterior, $\mathbf{b_1 = 0,05}$ arată că dacă vom compara două localităţi unde diferenţa dintre sumele cheltuite în program / numărul de locuitori este 1000 lei, şi care sunt identice din perspectiva variabilei CALIT, ne aşteptăm ca, în medie, rata şomajului să fi scăzut cu 0,05 puncte procentuale în localitatea în care s-a suma / locuitor a fost mai mare. $\mathbf{b_2 = 1,00}$ arată că pentru aceeaşi valoare a lui SUM, localităţile cu o evaluare a calităţii administrării programelor mai bună

cu o unitate au, în medie, o scădere mai mare cu o unitate procentuală. Această interpretare este consistentă cu rezultatul obţinut după aplicarea de regresii simple pentru fiecare din subeşantioanele definite de **CALIT**, dar aduce un plus de precizie în exprimarea relaţiilor dintre variabila dependentă şi cele două variabile independente.

Din interpretarea coeficienţilor $\mathbf{b_i}$ se vede cum regresia multiplă permite compararea de perechi de grupe de indivizi statistici care sunt identici din perspectiva tuturor variabilelor independente cu excepţia unei singure variabile. Diferenţa observată în valorile variabilei dependente este atribuită variaţiei în variabila independentă care ia valori diferite pentru grupe diferite.

Asemeni pantei pentru cazul bivariat, coeficienţii modelului de regresie multiplă depind de unităţile de măsură ale variabilelor şi este adevărată proprietatea

dacă în loc de $\mathbf{X_i}$ avem $\mathbf{cX_i + d}$, atunci în loc de $\mathbf{b_i}$ avem $\mathbf{b_i / c}$. **(6)**

Această proprietate arată faptul că panta de regresie nu poate fi folosită drept un indicator al intensităţii relaţiei dintre variabila dependentă şi variabila independentă corespunzătoare, şi nici nu permite ierarhizarea variabilelor independente în funcţie de contribuţia fiecăreia la explicaţia variaţiei variabilei dependente.

Pentru a descrie nu doar forma ci şi intensitatea relaţiilor liniare între variabilele independente şi variabila dependentă sunt folosiţi **coeficienţii de regresie standardizaţi**. Modul în care sunt definiţi este intuitiv: variabilele $\mathbf{Y}$, $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_m}$ sunt standardizate folosind formula scorului z prezentată în Capitolul 1.

Noile variabile, obţinute în urma aplicării formulei, au media egală cu zero iar abaterea standard egală cu unitatea. Coeficienţii de regresie din modelul determinat de variabilele standardizate, se notează în mod obisnuit cu $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, ..., $\boldsymbol{\beta_m}$. Aplicând proprietatea (6) avem următoarea formulă pentru coeficienţii standardizaţi:

$$\beta_i = b_i \frac{\sigma_{X_i}}{\sigma_Y}$$

În cazul bivariat (m = 1), $\boldsymbol{\beta_1}$ coincide cu coeficientul de corelaţie Pearson. În cazul general, $\boldsymbol{\beta_i}$ arată cu câte abateri standard variază $\mathbf{Y}$ atunci când $\mathbf{X_i}$ creşte cu o abatere standard. Din formulă reiese şi faptul că $\boldsymbol{\beta_i}$ are acelaşi semn cu $\mathbf{b_i}$, iar $\boldsymbol{\beta_i = 0}$ este echivalent cu $\mathbf{b_i} = 0$.

Atunci când m > 1 coeficienţii de regresie standardizaţi au câteva proprietăţi diferite faţă de cazul bivariat:

1. $\boldsymbol{\beta_i}$ poate să ia valori şi în afara intervalului [-1, 1]. $\boldsymbol{\beta_i}$ în valoare absolută este supraunitar atunci când relaţia dintre $\mathbf{X_i}$ şi $\mathbf{Y}$ este foarte intensă şi în plus, există o relaţie liniară strânsă între $\mathbf{X_i}$ şi cel puţin una dintre celelalte variabile independente.
2. în timp ce în cazul bivariat U = 0 implică faptul că $\beta_1 = \pm 1$, atunci când m > 1, condiţia U = 0 nu restrânge valorile posibile pentru $\boldsymbol{\beta_i}$.

Coeficienţii de regresie standardizaţi permit ierarhizarea variabilelor independente în funcţie de importanţa pe care o are fiecare în explicarea variaţiei variabilei dependente printr-o relaţie directă. Sunt necesare două precizări privind limitele în utilizarea acestor coeficienţi:

1. Dacă într-un model teoretic în care $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_m}$ sunt variabile independente pentru **Y**, $\mathbf{X_i}$ este o variabilă explicativă şi pentru unul sau mai multi $\mathbf{X_j}$, j≠i, atunci modelul de regresie în care **Y** este variabila dependentă ne permite numai estimarea efectului direct pe care îl are $\mathbf{X_i}$ asupra lui **Y**, nu şi a celui mediat de alte variabile din model. De exemplu, variabila care exprimă proporţia celor care au absolvit liceul din populaţia unei localităţi poate avea un efect direct nul asupra numărului de infracţiuni pe cap de locuitor ($\mathbf{b_{liceu} = 0}$), însă un efect indirect substanţial, mediat de o altă variabilă independentă cuprinsă în modelul de regresie multiplă (de exemplu, venitul pe cap de locuitor).
2. Atunci când avem două modele cu aceleaşi variabile, dar care descriu date diferite, coeficienţii standardizaţi nu sunt comparabili între modele decât dacă variabilele care le corespund au dispersii asemănătoare. De aceea, este de preferat ca în comparaţiile dintre populaţii diferite să fie folosiţi coeficienţii nestandardizaţi, după ce în prealabil datele au fost transformate astfel încât variabilele $\mathbf{X_i}$ să aibă aceeaşi unitate de măsură în ambele modele.

Eficienţa unui model de regresie multiplă

La fel ca şi în cazul bivariat, puterea explicativă a unui model multivariat poate fi evaluată cu ajutorul unor indicatori. Coeficientul $\mathbf{R^2}$, numit **coeficient de determinaţie multiplă**, este definit la fel ca şi atunci când avem o singură variabilă independentă şi are o interpretare similară:

$$R^2 = \frac{\sum (Y^* - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

Numărătorul expresiei reprezintă variaţia lui **Y** care este "explicată" de ecuaţia de regresie, în timp ce valoarea de la numitor este egală cu variaţia totală a lui **Y**. Deci, $\mathbf{R^2}$ indică proporţia din variaţia lui **Y** care este "explicată" de toate variabilele independente din model. Din definiţie rezultă că $\mathbf{R^2}$ poate să ia valori între 0 şi 1.
În exemplul din această secţiune avem $\mathbf{R^2 = 0{,}69}$.

Desigur, valorile ridicate ale lui $\mathbf{R^2}$ sunt de dorit în locul celor scăzute pentru că implică faptul că explicaţia este în mai mare măsură completă. Totuşi, această afirmaţie necesită câteva precizări:

$\mathbf{R^2}$ are proprietatea de a creşte cu fiecare variabilă care este introdusă în model şi de aceea valorile lui $\mathbf{R^2}$ trebuie judecate şi în raport cu numărul de variabile independente. La limită, este posibil să avem $\mathbf{R^2 = 1}$ dacă avem un număr de variabile independente suficient de mare, chiar dacă acestea sunt generate aleator.

Concluzia care se desprinde este că alegerea variabilelor care urmează să fie incluse în model nu poate fi decisă folosind exclusiv informaţia de natură statistică (chiar dacă există procedee complexe prin care putem îmbogăţi această informaţie). Numai prin luarea în considerare şi a unor aspecte de natură teoretică poate fi decisă includerea sau eliminarea unor variabile în analiză.

Să vedem ce semnificaţie au valorile extreme pe care le poate lua $\mathbf{R^2}$, 0 şi 1, pentru că interpretarea valorilor intermediare este posibilă prin raportarea la situaţiile maximale.

$\mathbf{R^2}$ este egal cu 1 atunci când valorile lui **Y** sunt complet determinate de combinaţiile liniare ale valorilor variabilelor independente. În cazul bivariat, distribuţia punctelor se face după o dreaptă, iar atunci când sunt două variabile independente, după un plan.

La fel ca şi în cazul bivariat, $\mathbf{R^2 = 0}$ nu indică în mod necesar absenţa unor relaţii între variabilele independente şi variabila dependentă:
1. **Y** poate să fie determinată complet de variabilele independente prin relaţii neliniare iar $\mathbf{R^2}$ să fie egal cu zero. De exemplu, dacă

$$Y = \sqrt{10 - X^2_1 - X^2_2}),$$

se obţine distribuţia din Figura 7.9. în care toate punctele sunt pe o suprafata curbă (cele mai multe sunt pe o semisferă), iar $\mathbf{R^2 = 0}$.

2. Mai mult, este posibil să avem $\mathbf{R^2 = 0}$ chiar şi atunci când între una dintre variabilele independente şi **Y** exista o relaţie liniară, în condiţii de control, însă forma (panta) acestei relaţii nu este constantă pe categoriile celorlalte variabile.

Figura 7.9. Exemplul unei distribuţii în spaţiu pentru care variabila dependentă este reprezentată pe axa verticală şi $R^2 = 0$.

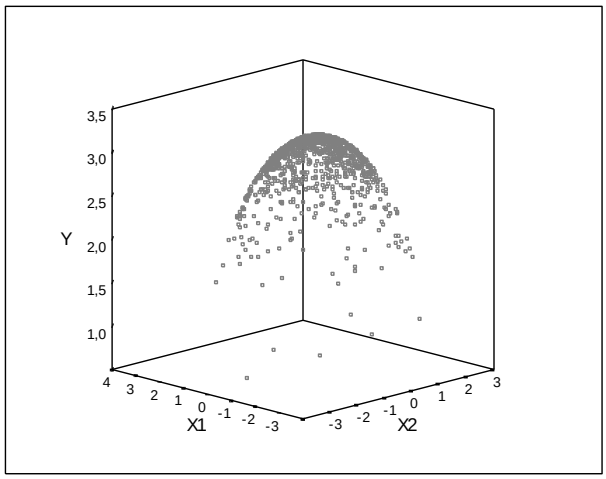
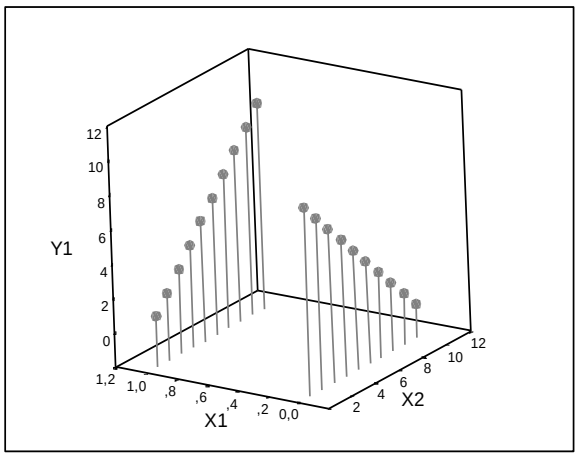


Figura 7.10. Exemplul unei distribuţii în spaţiu pentru care variabila dependentă este reprezentată pe axa verticală, au loc relaţii bivariate cu variabila dependentă de intensitate maximă şi $R^2 = 0$ pentru modelul multivariat.



Generalizarea rezultatelor obţinute pe eşantion (inferenţa)

Atunci când avem la dispozitie date dintr-un eşantion probabilistic şi construim pe acestea un model de regresie multiplă ne punem problema de a generaliza rezultatele obţinute pentru populaţia din care provine eşantionul. Să presupunem că am obţinut $\mathbf{b_i > 0}$ şi dorim să ştim dacă panta corespunzătoare variabilei $\mathbf{X_i}$ este pozitivă şi la nivelul întregii populaţii. Pentru a afla acest lucru vom construi un interval de încredere în jurul valorii lui $\mathbf{b_i}$. Determinarea intervalului de încredere presupune îndeplinirea de către datele analizate a unor proprietăţi, aceleaşi cu condiţiile specifice cazului bivariat, la care se adaugă condiţia de *absenţă a multicoliniarităţii* (pe care o vom defini şi analiza în sectiunea 6). În continuare, presupunem îndeplinite toate aceste proprietăţi. Pentru un nivel de incredere de 0,95 avem intervalul

$$\mathbf{(b_i - t_{n-m-1,\,0,975}\sigma_{bi},\ b_i + t_{n-m-1,\,0,975}\sigma b_i),}$$

unde **n** este numărul de cazuri în eşantion, **m** este numărul de variabile independente, numărul $\mathbf{t_{n-m,\,0,975}}$ poate fi găsit în tabelele pentru distributia **t** (pentru n - m - 1 = 60 ia valoarea 2,0, iar pentru un număr care tinde la infinit ia valoarea 1,96) iar $\boldsymbol{\sigma_{bi}}$ este eroarea standard a lui $\mathbf{b_i}$.

Dacă intervalul de încredere nu îl conţine pe zero atunci *ipoteza de nul*, adică afirmaţia conform căreia între $\mathbf{X_i}$ şi **Y** nu avem o relaţie liniară directă, poate fi respinsă (un mod mai riguros de a exprima ipoteza de nul în cazul regresiei este acela de a spune că parametrul - coeficientul de regresie din populaţie - este egal cu 0).

O altă modalitate prin care poate fi realizată generalizarea valorilor $\mathbf{b_i}$ constă în determinarea valorii maxime a nivelului de semnificatie statistică pentru care 0 aparţine intervalului de încredere. Dacă nivelul de semnificaţie observat este mai mic decât 0,05 atunci vom respinge ipoteza de nul, conform regulilor de testare a ipotezelor statistice.

Deasemenea, putem calcula valoarea **t** asociată lui $\mathbf{b_i}$

$$\mathbf{t = b_i / \sigma_{bi.}}$$

Putem găsi în tabelele statistice care este nivelul de încredere ce corespunde valorii determinate în acest fel.

Pentru exemplul din această secţiune, tabelul următor conţine coeficienţii de regresie, coeficienţii standardizaţi, erorile standard, valorile **t** şi nivelele de semnificatie statistică.

	Coeficienti ne-standardizati		Coeficienti standardizati		Nivel de semnif. stat.
	B	Eroare standard	Beta	t	
SUM	0,048	0,212	0,052	0,229	0,820
CALIT	0,994	0,291	0,786	3,412	0,002

Pentru a doua variabilă independentă din exemplu, CALIT, putem respinge ipoteza de nul ($p = 0,002<0,05$).

Un model de regresie multiplă poate fi folosit atit în explicaţie cit şi în predicţie. Astfel, din modelul anterior rezultă că, în medie, localităţile în care ar fi aplicat un program similar cu cel studiat iar suma cheltuită ar fi de 8000 de lei / locuitor, şi care ar fi evaluate prin scorul 3 pentru calitatea administrării programelor, vor avea *în medie* o scădere a ratei şomajului cu o valoare dată de următoarea formulă

$$\mathbf{DIF_0 = -2,61 + 0,05 * 8 + 1,00 * 3 = 0.79.}$$

Mai mult, putem determina cu o probabilitate **p** intervalul căruia îi apartine valoarea variabilei DIF pentru care cunoaştem valorile variabilelor **SUM** şi **CALIT**. Acesta este

$$\mathbf{(DIF_0 - t_{n-m-1,0,975}\sigma_{DIF}, DIF_0 + t_{n-m-1,0,975}\sigma_{DIF}),}$$

unde $\mathbf{DIF_0}$ este valoarea medie "prezisă" de model, σ_{DIF} este **eroarea standard** a valorii estimate $\mathbf{DIF_0}$.

În exemplul anterior, **DIF** este cu o probabilitate de 0,95 în intervalul de incredere
(0,79 - 2*0,59, 0,79 + 2*0,59) = (-1,57, 3,15).

Problema multicoliniarităţii

Situaţia în care o variabilă independentă poate fi exprimată ca o combinaţie liniară perfectă a celorlalte variabile independente, este numită **multicoliniaritate perfectă**:
De exemplu, dacă variabilele independente sunt $\mathbf{X_1}$, $\mathbf{X_2}$, şi $\mathbf{X_3}$, iar
X2 = 3X1 + 2X3, se spune că $\mathbf{X_2}$ este exprimat printr-o combinaţie liniară a variabilelor $\mathbf{X_1}$ şi $\mathbf{X_3}$, iar variabilele $\mathbf{X_1}$, $\mathbf{X_2}$ şi $\mathbf{X_3}$ sunt într-o relaţie de multicoliniaritate perfectă.

Atunci când variabilele independente sunt într-o situaţie de multicoliniaritate perfectă coeficienţii de regresie nu pot fi determinaţi, şi analiza de regresie nu poate fi aplicată.

În practică, o situaţie de acest tip este rar întâlnită şi este uşor de detectat. În schimb, sunt mai frecvente cazurile de **multicoliniaritate ridicată**, în care o variabila independentă poate fi exprimată aproape perfect printr-o combinatie liniară a celorlalte variabile independente. Când se intimplă acest lucru, coeficienţii pot fi determinaţi în mod unic însă sunt instabili: valoarea pantei unui anumit coeficient diferă foarte mult de la un eşantion la altul pentru o anumită populaţie. Din acest motiv, atât comparaţiile între valorile coeficienţilor dintr-un model cât şi comparaţiile pentru aceeaşi coeficienţi ai unor modele pe eşantioane diferite sunt nesigure.
Care este pragul peste care multicoliniaritatea este considerată a fi ridicată şi poate să ridice probleme în interpretarea modelului? O metodă frecvent folosită constă în realizarea de regresii în care, pe rând, fiecare dintre $\mathbf{X_i}$ este variabilă dependentă iar ceilalţi $\mathbf{X_j}$ sunt variabile independente. Valoarea cea mai ridicată pentru un $\mathbf{R^2}$ obţinut în acest fel este o măsură a nivelui de multicoliniaritate din model, iar limita convenţională sub care se consideră că multicoliniaritatea nu afectează interpretabilitatea modelului este **0,8**.

Atunci când este întâlnită o situaţie de multicoliniaritate ridicată sunt mai multe moduri prin care pot fi atenuate efectele ei:
1. Este mărit volumul eşantionului astfel încât ipoteza de nul să poată fi respinsă pentru o parte dintre coeficienţi.
2. Variabilele care sunt puternic corelate sunt combinate în indicatori unici. De exemplu, într-o analiză în care secţiile de vot sunt unităţi statistice, rata de participare în primul tur de scrutin al alegerilor din 1996 este o variabilă independentă şi rata de participare în al doilea tur de scrutin al alegerilor din 1996 este o altă variabilă independentă, coeficientul de corelaţie între cele două variabile este **r = 0,91**. Un model realizat pe un eşantion ales dintre secţiile de vot şi în care cele

două variabile sunt independente va fi afectat de o problemă de multicoliniaritate ridicată. O soluţie ar fi includerea în analiză a mediei în locul celor două variabile.
3. Sunt realizate mai multe modele fiecare având doar o parte dintre variabilele care produc multicoliniaritate. Pentru exemplul anterior, ar însemna considerarea a două modele, unul cu rata de participare pentru primul tur, al doilea cu rata de participare pentru al doilea tur.

Variabile "dummy"

Regula generală conform căreia analiza de regresie poate fi aplicată numai variabilelor de interval sau de rapoarte are o excepţie importantă: toate proprietăţile pe care le au valorile estimate ale unui model de regresie se păstrează şi în cazul în care una sau mai multe dintre variabilele independente sunt **dihotomice** (adică variabile care iau două valori).
Consecinţele acestei proprietăţi sunt importante deoarece permit nu doar estimarea efectelor unor variabilele care în mod obisnuit sunt dihotomice (exemplu *sexul*, *mediul de rezidentă* -- urban / rural, etc.) asupra variabilei dependente ci şi includerea într-o analiză de regresie a unor variabile nominale sau ordinale cu mai mult de două categorii. Acest lucru este posibil în urma transformării unei variabile cu **n** categorii în **n - 1** variabile dihotomice.
Ca o ilustrare, să presupunem că datele despre programul de reducere a şomajului din exemplul discutat în această secţiune cuprind informaţii despre încă o variabilă independentă: judeţul în care se află localitatea (JUDET). Mai presupunem că localităţile din studiu provin din trei judeţe, notate cu A, B, C. Variabila JUDET este transformată în două variabile dihotomice: JUDET1 şi JUDET2.
JUDET1 este definită astfel: localităţile care sunt în judeţul A au valoarea 1, iar toate celelalte au valoarea 0.
JUDET2 este definită asemănător: localităţile care sunt în judeţul B au valoarea 1, iar toate celelalte au valoarea 0.
Cunoscând valorile celor două variabile pentru o localitate, ştim sigur în ce judeţ se află aceasta, deci informaţia oferită de JUDET1 şi JUDET2 este egală cu cea oferită de variabila iniţială, JUDET.
Prin introducerea celor două variabile dihotomice în analiză putem verifica dacă scăderea şomajului a fost influenţată şi de factori care s-au manifestat la nivel de judeţ, independenţi de condiţiile de nivel local. Dacă coeficientul **b**, respectiv **beta**, care corespunde uneia dintre aceste variabile este diferit de 0, atunci rezultă că variaţia variabilei DIF poate fi explicată mai bine în urma includerii ei.

Exerciţii şi probleme

1. Presupunem cunoscute pentru mai multe localităţi următoarele două variabile: diferenţa între venitul pe cap de locuitor în ultimul an şi cel din anul anterior (VENIT)--măsurat în mii lei--, şi proporţia celor din localitate care au absolvit cel mult 10 clase (SCOALA)--estimată în procente.
O analiză de regresie prin care se încearcă explicarea efectului variabilei SCOALA asupra variabilei VENIT conduce la următoarele rezultate:

r = -0,55, **R^2 = 0,30**, iar ecuaţia de regresie este
VENIT = -7,5 - 12 SCOALA.

Care dintre următoarele afirmaţii este adevărată?
a. Localităţile în care SCOALA are valori mici au, în medie, valori mai mici ale variabilei VENIT.

b. Localităţile în care sunt 20% locuitori care nu au absolvit 10 clase au, în medie, diferenţa între veniturile anuale pe cap de locuitor (VENIT) cu 12 mii de lei mai mare decât localităţile în care sunt 30% locuitori care nu au absolvit 10 clase.

2. Presupunem că, în plus, avem şi date despre proporţia locuitorilor care au mai puţin de 18 ani (MINORI)-- estimată în procente. Ecuaţia de regresie multiplă care include variabilele SCOALA şi MINORI ca variabile independente este următoarea:

VENIT = 12 - 4 SCOALA - 0,2 MINORI.
β_{SCOALA} = -0,08, iar **β_{MINORI} = -0,25**.

Care dintre următoarele afirmaţii este adevărată?
a. Variabila MINORI explică mai puţin din variaţia variabilei dependente decât variabila SCOALA.
b. Conform modelului de regresie, localităţile în care SCOALA = 5, iar MINORI = 20, au avut în medie o scădere a venitului pe cap de locuitor cu 12 mii de lei.

Pentru următoarea întrebare, să se încercuiască varianta corectă /variantele corecte:

3. Se da ecuatia de regresie cu coeficienti nestandardizati (în paranteză sunt prezentate erorile standard ale coeficienţilor de regresie:
$Y' = -40 + 0.72x_1 + 1.29x_2 – 0.15x_3$
ES (0.13) (0.37) (0.16)
R=0.9

Care din coeficientii ecuatiei de regresie difera semnificativ de zero pentru un nivel de semnificatie p=0.05?
1. cel al lui x_1
2. cel al lui x_2
3. cel al lui x_3
4. niciunul