

Facultatea de Științe Politice, Administrative și ale Comunicării
Specializarea Științe Politice
Semestrul II, Anul II

Metode cantitative de analiza in stiintele politice

Suport Curs Învățământ la Distanță¹

Titlul disciplinei: Metode cantitative de cercetare in științele politice

Codul: ULR1415

Numărul de credite: 5

Locul de desfășurare: Facultatea de Științe Politice, Administrative și ale Comunicării,
str. General Traian Moșoiu, nr. 71

Nume, titlul științific: Dr. Daniela Angi

Birou: Facultatea de Științe Politice, Administrative și ale Comunicării, str. General
Traian Moșoiu, nr. 71

Informații de contact (adresă e-mail): angi@fspac.ro

Ore de audiență: Luni 18.00 – 20.00

Condiționări și cunoștințe prerechizite:

Fără condiționări

Descrierea cursului:

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 învăța (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.

Formatul și tipul activităților implicate de curs:

Pentru unele dintre aplicatii va fi folosit calculatorul.

Studentii vor avea acces la materiale scrise disponibile pe platforma e-learning FSPAC.

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

Materiale bibliografice obligatorii:

Rotariu T. (coord.). Metode statistice aplicate in stiintele sociale. Polirom, 2000

- Rotariu T., Petru Ilut, Ancheta sociologica, Polirom, 1997
- Babbie, Earl. Learning from the field: A guide from experience. London: Sage Publication. 1984.
- Babbie, Earl. Survey Research Methods, 2nd ed. Belmont, CA: Wadsworth Publishing Co. 1990
- Dalton, Rusell. Citizen Politics: Public Opinion and Political Parties in Advanced Western Democracies. Chatham House Publishers. 1988.
- Johnson J., Joslyn R., Political science research methods, 1991
- King, G., R. Keohane, S. Verba, Designing Social Inquiry, 1994
- Nachmias C., Nachmias D., Research methods in the social sciences, 1996
- White L., Political Analysis. Technique and Practice, 1994

Elemente de deontologie academica

Notiunea de *plagiat* se defineste în conformitate cu normele deontologice definite la link-ul de mai 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 științele sociale.

Ghid de studiu:

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

Unitatea 1

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

Noțiuni cheie: teorie, ipoteza, design de cercetare, date și tipuri de date.

Etapele unui proces de cercetare.

Organizarea cercetării

Crearea teoriei

Scopul principal al acestui capitol este de a aduce în discuție o serie de elemente necesare configurării celei mai importante părți a unui proiect de cercetare: teoria care stă la baza abordării, asumțiilor și presupuzițiilor făcute; în funcție de construcțiile teoretice de la care se pleacă sunt construite ipotezele, este aleasă abordarea și metodele de culegere și analiză a datelor și sunt prezentate rezultatele la care se ajunge. Elementele care alcătuiesc o teorie sunt: conceptele, categoriile și propozițiile (Corbin și Strauss 1990, p.7).

Conceptele sunt unitățile de bază ale analizei sau abordării; de la modul în care sunt conceptualizate datele, și nu de la datele în sine, este dezvoltată o teorie. “Teoriile nu pot fi construite pornind de la evenimente actuale sau de la activități observate sau relatate, adică din “date brute”. Circumstanțele, evenimentele, faptele sunt luate ca și, sau analizate ca și, indicatori potențiali ai fenomenelor, fenomene cărora le sunt atribuite astfel etichete” (Corbin și Strauss 1990, p.7). Spre exemplu, dacă un respondent afirmă că face parte din sindicatul instituției în care își desfășoară activitatea profesională, atunci acesta poate fi etichetat ca fiind “membru al sindicatelor” și în analiză noastră, atunci când ne vom referi la el, îl vom desemna folosind eticheta și nu prin descrierea activității 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 și desemnarea cu aceeași etichetă a faptelor asemănătoare.

Al doilea element important al unei construcții teoretice îl constituie categoriile. Categoriile au un nivel mai ridicat și 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 comparațiilor și evidențierea similarităților și diferențelor. 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 instituției în care își desfășoară activitatea profesională, și pe care i-am etichetat ca fiind “membri ai sindicatelor”, vom identifica alți indivizi participă la activități ale partidelor politice, ale asociațiilor non-guvernamentale, etc, iar aceștia vor fi etichetați “membri ai partidelor politice” respectiv “membri ai asociațiilor non-guvernamentale”. Deși conceptele amintite sunt diferite în ceea ce privește forma, ele reprezintă activități legate de același proces și pot fi grupate într-o categorie etichetată “cei care iau parte la activități participative”.

Al treilea element al teoriei sunt propozițiile care pun în evidență relații între categorii și concepte sau între categorii diferite. Propozițiile sunt adeseori desemnate cu eticheta de “ipoteze” (Glaser și Strauss 1967). Termenul de “ipoteză” este însă considerat mai puțin adecvat întrucât aceasta implică relații care pot fi “măsurate” între concepte și categorii, ceea ce nu se întâmplă întotdeauna – spre exemplu cazul unor abordări calitative (Whetten 1989, p. 492).

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

În procesul de creare a teoriei literatura de specialitate amintește patru etape analitice, etape care nu sunt strict secvențiale: design-ul de cercetare, culegerea datelor, analiza datelor și compararea rezultatelor obținute 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 conține în general referiri la: datele sau informațiile care urmează a fi colectate și la modul în care acestea urmează a fi analizate pentru a răspunde la întrebările sau cerințele de bază ale cercetării. Rezultă de aici că primul pas în construcția unui design de cercetare îl constituie definirea sau formularea întrebărilor la care urmează a se răspunde în cercetare. Acestea trebuie formulate suficient de restrâns ținând cont de faptul că design-ul cercetării de obicei este acea parte a unei cercetări în care sunt anunțate intențiile de a cerceta o anumită problemă și 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 științele sociale unde fenomenele studiate sunt în continuă evoluție. O sursă importantă de “întrebări” o constituie literatura de specialitate (spre exemplu: rapoarte ale unor studii, înscrisuri cu conținut specific diferitelor domenii studiate, etc).

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

În general, o cercetare în științele sociale încearcă să descrie și / sau să interpreteze un anumit fenomen uman, cel mai adesea pornind de la comportamente ale indivizilor sau de la relații ale acestora cu privire la comportamente adoptate în diferite situații. Date fiind varietatea interacțiunilor umane și dinamica acestora, în construcția design-ului de cercetare trebuie ținut cont de distorsiunile care pot apărea, de presuposițiile făcute și de interpretările care se dau diferitelor comportamente analizate astfel încât cititorii să poată înțelege și interpreta rezultatele la care ajunge cercetarea. Așa 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 pretenția de a fi completă sau exhaustivă – cerință oricum greu de îndeplinit dată fiind, așa cum am amintit și mai sus, varietatea subiectului analizat și a constrângerilor care trebuie avute în vedere în analiza acestuia - ci mai degrabă încearcă să fie un punct de plecare pentru cercetătorii care încearcă să se decida asupra unei modalități de organizare a datelor și de comunicare a ideilor. În funcție de subiectul abordat, de datele disponibile cu privire la acesta, de metoda de cercetare utilizată și de teoria de la care se pornește, cerințele enumerate mai jos sunt sau nu sunt prezente în structura unui design de cercetare particular.

Structura unei cercetări

1. Introducere

- Porniți la drum cu un citat sau cu o scurtă povestire care să capteze atenția cititorului. Incercați să găsiți 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.
- Formulați propriile dumneavoastră întrebări sau nelămuriri cu privire la problematica abordată, descrieți contextul în care aceste întrebări sau nelămuriri au apărut și cum au evaluat. Ce ați dori să știți sau să vă lămuriti? Cum ați ajuns să fiți interesat de problemă?
- Amintiți și alți cercetători care consideră că este necesară o abordare a tematicii avute în vedere, prezentați rezultatele la care au ajuns aceștia, sau, dacă este cazul, atrageți atenția asupra faptului că o astfel de tematică nu trebuie ignorată.
- Justificați alegerea făcută. De ce este importantă o abordare a fenomenului respectiv în momentul de față (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).
- Specificați ceea ce urmăriți în cercetarea dumneavoastră (ex: lărgirea bazei de cunoaștere, deschiderea unor noi perspective de abordare, confirmarea unor rezultate anterioare, verificarea unor asumptii, etc).
- Descrieți publicul căruia vă adresați.

2. Paradigma care stă la baza abordării

- Această secțiune este necesară mai ales atunci când tematica abordată nu este suficient de bine cunoscută de publicul căruia vă adresați sau atunci când, indiferent de public, fie tematica, fie abordarea, fie amândouă sunt noi.
- Prezentați propria paradigmă și încercați să o înscrieți într-o anumită tendință de abordare (ex: fenomenologică, hermeneutică, etc). Amintiți alți cercetători care au definit paradigme asemănătoare în alte domenii ale științelor sociale. (Guba, E. 1990).
- Prezentați și explicați asumptiile și presupuzițiile pe care le formulați în legătură cu subiectul abordat. Explicați modul în care acestea pot distorsiona rezultatele la care se va ajunge.
- Dacă este cazul, mai ales pentru abordărilor calitative, prezentați poziția pe care se plasează cercetătorul în raport cu subiectul cercetat, spre exemplu: cercetător ca și membru complet, cercetător ca și membru activ, cercetător ca și membru periferic (Adler și Adler, 1994).
- Specificați criteriile adecvate pentru evaluarea rezultatelor cercetării. (Atkinson, Heath, și Chenail, 1991).
- Discutați modul în care experiența dumneavoastră anterioară influențează modul în care concepeți abordarea subiectului studiat. Prezentați pe scurt experiență profesională care vă apropie de tematică.

3. Metoda de cercetare

- Identificați și descrieți metoda pe care urmează să o utilizați (ex: analiză de caz; metoda comparativă, metoda etnografică, observație, experiment, etc.). Descrieți modul în care alți autori au utilizat metoda avută în vedere de dumneavoastră (Glaser, B., și Strauss, A. 1967).
- Descrieți în detaliu ceea ce urmează să faceți. Prezentați modalitatea de selectare a subiecților de la care vor fi culese informațiile necesare cercetării.
- Descrieți datele pe care intenționați să le culegeți sau pe care intenționați să le utilizați și 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) prezentați întrebările folosite (fie în context, fie atașate într-un appendix).

- Descrieți procedurile de culegere și analiză a datelor în ordinea cronologică a desfășurării lor.
- Descrieți procedurile de analiză pe care intenționați să le utilizați (codarea datelor, sortarea datelor, procedurile statistice cu ajutorul cărora sunt puse în evidență relațiile existente între date, etc). Prezentați, dacă este cazul, programele statistice utilizate pentru modelarea datelor.
- Interpretați rezultatele obținute în funcție de teoria, asumțiile și presuposițiile formulate la începutul cercetării.

4. Concluzii

- Reluați pe scurt problematica de la care s-a pornit. Amintiți asumțiile și presuposițiile făcute, metoda de cercetare și rezultatele la care s-a ajuns.
- Stabiliți legăturile existente între rezultatele cercetării dumneavoastră și literatura de specialitate care prezintă rezultate similare sau asemănătoare.
- Imaginați modul în care design-ul cercetării poate evolua de la rezultatele pe care le-ați obținut și ținând cont de evoluțiile ulterioare ale domeniului studiat. Specificați “deschiderile” lăsate de proiectul dumneavoastră și posibilele modalități de valorificare a informației acumulate ulterior.
- Discutați “validitatea” și “fidelitatea” procedurilor utilizate în culegerea și analiza datelor.
- Discutați posibilele distorsiuni generate fie de asumțiile și de presuposițiile făcute, fie de metodele de culegere și de analiză a datelor.
- Prezentați modul în care literatura de specialitate v-a influențat în modalitățile de abordare a subiectului cercetat.
- Discutați limitele cercetării dumneavoastră și amintiți limitele cu care se confruntă orice subiect asemănător abordat și în alte studii.

Culegerea datelor

O dată formulate întrebările la care se intenționează a se răspunde prin cercetare, următorul pas este alegerea “cazurilor” sau a “indivizilor” care urmează a fi investigați pentru a obține datele necesare confirmării sau infirmării propozițiilor referitoare la problematica cercetată. În alegerea “cazurilor” noastre putem avea un plan prestabilit, așa cum este cazul în cercetările cantitative, sau putem să ne selectăm cazurile pe măsură ce procesul de cercetare evoluează, așa cum este cazul în cele mai multe din cercetările calitative (Strauss și Corbin, 1990, p. 192).

În faza inițială de culegere a datelor, atunci când sunt stabilite categoriile este necesară o investigație extinsă și în profunzime a “cazurilor” pentru a obține date cât mai adecvate cu putință. Atunci când “cazurile” sunt foarte multe și nu pot fi investigate toate se alege o procedură de eșantionare a populației noastre de “cazuri”. Pentru a stabili cât de multe “cazuri” vor intra în atenția analizei noastre, cu alte cuvinte pentru a stabili unde ne oprim cu eșantionarea din punct de vedere teoretic, este nevoie să recurgem la teorie și la logica cercetării noastre. Ne oprim cu eșantionarea acolo unde nu mai este identificată informație suplimentară cu ajutorul cărei cercetătorul să dezvolte proprietăți sau caracteristici ale conceptelor sau categoriilor cu care lucrează (Glaser și Strauss 1967, p. 65). În alegerea cazurilor noastre trebuie ținut seama de faptul că nu toate cazurile au aceeași relevanță raportat la problematica cercetată și la teoria pe care se bazează cercetarea; astfel, în unele situații, este suficient un “caz” pentru a pune în evidență o anumită problematică, în alte situații este nevoie de mai multe “cazuri” pentru a face același lucru. Ca și 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 și categorii, b) alegerea unui “caz” pentru a reproduce rezultatele obținute 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 și Strauss 1967, p. 65). Diferite surse de date oferă cercetătorului perspective diferite asupra fenomenului studiat. Inițial abordarea unui fenomen poate avea la bază o singură tehnică de culegere a datelor, ulterior însă este recomandată identificarea și altor surse de date și 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 realității, de descompunere a fenomenelor în elementele lor componente și reasamblarea lor în modalități noi (Strauss și 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 și cei susțin abordările calitative.

Compararea rezultatelor obținute cu rezultatele din literatura de specialitate.

O dată datele culese, analizate și interpretate problema care se pune este aceea de a compara rezultatele obținute literatura de specialitate existentă și examinarea a ceea ce este similar și 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 obținute 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 și a problematicei eşantionării

Ghid de studiu:

- ◆ Cercetări selective: de la populație la eşantion
- ◆ Reprezentativitatea eşantioanelor: a alege câțiva pentru a îi reprezenta pe toți.
- ◆ Proceduri de eşantionare. Tipuri de eşantioane
- ◆ Aspecte matematice ale eşantionării. Teste de semnificație

Unitatea 1

Obiectiv: Prezentarea noțiunii de eşantionare

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

Eșantionarea

Cercetări selective: de la populație la eșantion

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

Atât în aceste situații, cât și în multe altele de acest fel, problema care se pune este aceea de a culege informațiile necesare pentru a analiza temele avute în vedere doar de la o parte din indivizii care compun o populație și nu de la întreaga populație. Din punct de vedere tehnic, grupul sau mulțimea de indivizi care constituie obiectul de studiu sau de interes al cercetătorului la un moment dat este denumit *populație*, iar grupul mai mic de indivizi de la care sunt culese informațiile necesare cercetării este denumit *eșantion*. “Setul de operații cu ajutorul cărora, din ansamblul *populației* vizate de cercetare, se extrage o parte, numită *eșantion*, parte ce va fi supusă nemijlocit investigației”² este desemnat ca fiind operația de *eșantionare*.

Decizia de a culege datele necesare unei cercetări de la un eșantion sau de la o populație depinde de o serie de aspecte practice. Astfel, în unele situații, dacă timpul, resursele financiare și umane nu constituie o problemă sau dacă populația țintă nu este foarte numeroasă, atunci este mult mai avantajoasă culegerea datelor de la toți indivizii care compun o populație vizată; în felul acesta se obține o imagine exactă a problematicii investigate. În alte situații există o serie de constrângeri care îl împiedică pe cercetător să ajungă la toți indivizii care compun o populație, aceste constrângeri se referă în primul rând la timp, resursele financiare și umane aflate la dispoziție, dispersarea geografică a populației care urmează a fi cercetată, iar soluția cea mai la îndemână pentru a culege informațiile necesare constă în selectarea unui eșantion și investigarea indivizilor care îl compun. Din acest punct de vedere am putea spune că eșantionarea este un compromis datorat insuficienței resurselor. Nu întotdeauna este însă vorba numai de imposibilitatea fizică de a culege informații de la toți membrii unei populații – neajuns care în unele situații poate fi depășit – ci și de o lipsă de eficiență practică – spre exemplu, în cazul cercetătorului din științele naturale, care, dacă ar tăia toate exemplarele unei specii de plante pentru a le analiza în laborator ar determina dispariția speciei respective. Pe de altă parte, concentrând resursele existente doar pentru analiza unei părți dintr-un întreg se pot obține rezultate mai bune decât analizând întregul, mai ales atunci când acest întreg este format din mulți 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 populației care urmează a fi studiată. În acest context, prin “populație” sunt desemnate toate elementele care pot sau trebuie să fie studiate.

² Traian Rotariu, Petre Iluț, Anchetă sociologică și sondajul de opinie, Ed. Polirom, Iași, 1997, p.122.

Elementele pot fi indivizi umani, dar în același timp pot fi gospodării, școli, spitale, întreprinderi economice, orașe, organizații sociale sau profesionale, ziare, articole de presă, discursuri ale unor oameni politici, etc. Indiferent însă de cine sau ce constituie elementele populației vizate, aceasta trebuie să fie atent delimitată în funcție de obiectivele cercetării, întrucât rezultatele finale vor depinde de acest punct de referință stabilit inițial. Spre exemplu, să ne imaginăm că într-un oraș se pune problema adoptării unui nou sistem de transport în comun, iar ceea ce ne interesează este acordul sau dezacordul cetățenilor cu privire la modul practic de realizare a acestuia. În acest caz, populația vizată este compusă doar din cei care locuiesc în orașul respectiv? sau trebuie avută în vedere și populația care nu locuiește în oraș, dar care într-o măsură sau alta beneficiază de transportul în comun din acel oraș? care este vârsta minimă și maximă a celor care vor fi chestionați?, etc. Întrebări asemănătoare acestora se ivesc de fiecare dată când se pune problema definirii unei populații care urmează a fi cercetate, iar de răspunsurile date vor depinde în mod evident rezultatele finale. Un prim pas în definirea unei populații vizate este acela de a stabili o *populație ideală*, adică toți cei care ipotetic ar trebui să fie luați în considerare atunci când se cercetează o problematică anume. După care în funcție de constrângerile practice identificate – spre exemplu, în cazul în care elementele populației sunt indivizi umani, astfel de constrângeri ar putea fi date de imposibilitatea de a îi investiga pe cei aflați în închisori, unități militare, spitale, hoteluri, în străinătate, etc – populația ideală poate fi restrânsă la o populație vizată care poate fi abordată în cadrul cercetării. Avantajele luării în considerare în faza inițială a unei populații ideale este acela că excluderea unor segmente din aceasta este explicită, iar neajunsurile rezultate de aici pot fi luate în considerare.

O dată stabilită populația vizată, poate fi pusă și problema alegerii unui eșantion. Pentru aceasta, elementele populației vizate sunt traspuse într-o listă numită *cadru de eșantionare*, listă din care vor fi extrase ulterior potrivit unor proceduri clar definite acele elemente care vor compune eșantionul. Spre exemplu, dacă se realizează o anchetă telefonică al cărui scop este investigarea modului în care dotarea cu utilități publice a unei localități acoperă necesitățile existente, populația ideală este constituită din toate gospodăriile care au acces la utilități publice, iar cadrul de eșantionare este format din toate gospodăriile care au acces la utilități publice și au telefon. Constrângerea în acest caz este dată de existența unui post telefonic în gospodărie. Gospodăriile care au acces la utilități dar care nu au telefon neputând fi investigate, populația vizată este formată doar din acele gospodării care au acces la utilități publice și au telefon (Figura 1.). În exemplul de față, astfel de liste care să se constituie în cadrul de eșantionare pot fi evidențe ale companiilor furnizoare de utilități publice și liste ale abonaților la servicii telefonice din localitatea avută în vedere.

De la caz la caz, în funcție de problema investigată, pot constitui cadrul de eșantionare: lista celor care sunt înscrși 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 abonați la o firmă furnizoare de servicii de televiziune prin cablu, lista celor abonați la o anumită publicație, 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 populației 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 inițial 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;
- referințe la grupuri de elemente și 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 și nu familii;

- existența unor elemente străine, adică existența în listă a unor elemente care din diferite motive nu fac obiectul de studiu la un moment dat;
- existența unor duplicate: când unele elemente ale populației apar de mai multe ori pe o listă.

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

O modalitate de a depăși aceste neajunsuri constă în redefinirea problematicii cercetate în așa fel încât elementele populației care nu pot fi identificate să nu facă obiectul unei anumite cercetări. Acest lucru evident nu este posibil în toate situațiile și nu este posibil mai ales în acele situații în care elementele care nu pot fi identificate constituie majoritatea elementelor unei populații.

Atunci când nu există liste care să cuprindă elementele unei populații vizate prin cadru de eșantionare poate fi desemnată orice altă procedură care să permită identificarea elementelor unei populații. Spre exemplu, o arie geografică poate juca rolul de cadru de eșantionare, situație în care elementele populației vizate sunt asociate cu un anumit spațiu natural. Astfel, aria geografică ocupată de o populație vizată poate fi împărțită 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 și așa mai departe până la ultimul stadiu când din anumite zone astfel selectate sunt investigate toate elementele.

Unitatea 2

Obiectiv: Prezentarea problematicii reprezentativității eșantioanelor

Reprezentativitatea eșantioanelor: a alege câțiva pentru a fi reprezenta pe toți.

Un eșantion “bun” este într-o oarecare măsură o versiune în miniatură a unei populații, un model al unei populații. Caracteristica cea mai importantă a unui eșantion bun este dat de reprezentativitatea acestuia. Un eșantion este considerat reprezentativ pentru populația din care este extras dacă “caracteristici importante sunt distribuite similar în amândouă grupurile”³ sau cu alte cuvinte, ținând cont de ordinea temporală a constituirii celor două grupuri, un eșantion trebuie să reproducă caracteristici importante ale populației din care este extras. Aceste caracteristici importante pot fi spre exemplu, vârsta, nivelul de educație, mediul de reședință, sexul, venitul, etc. Spre exemplu, dacă populația vizată este fi constituită în proporție de 51% din femei, dintre care 27% au studii medii, atunci un eșantion reprezentativ va fi compus în proporție de 51% din femei dintre care aproximativ 27% vor avea studii medii.

Un eșantion nu va reproduce niciodată cu exactitate toate caracteristicile unei populații, ca urmare aproximarea unei caracteristici existente în populație recurgând la măsurători efectuate pe un eșantion va produce o anumită eroare (d), iar încadrarea rezultatului obținut într-o marjă de eroare rezonabilă se face cu un anumit grad sau nivel de probabilitate (P). Eroarea obținută este rezultatul diferenței reale existente între o caracteristică A dintr-o populație și caracteristica corespunzătoare A^* măsurată pe un eșantion extras din acea populație. Nivelul de probabilitate este măsura în care eroarea pe care o facem aproximând o valoare A din populație prin valoarea corespunzătoare A^* măsurată pe un eșantion este mai mică decât o eroare maximă admisă. Reprezentativitatea unui eșantion este exprimată cantitativ de cele două valori d și P , valori care sunt determinate una de cealaltă. Un eșantion 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 eșantion, acesta reproduce mai mult sau mai puțin caracteristici ale populației din care este extras, motiv pentru care nu există eșantioane nereprezentative, ci doar eșantioane mai mult sau mai puțin reprezentative pentru o populație în funcție de măsura în care caracteristici ale populației respective sunt regăsite și în aceste eșantioane. Astfel, un eșantion care reproduce mai bine caracteristicile unei populații decât un alt eșantion, vom spune care mai reprezentativ. Mai mult, unele caracteristici pot fi mai bine reproduse de un eșantion iar altele mai puțin bine, ceea ce înseamnă ca reprezentativitatea unui eșantion este diferită în funcție de caracteristica care este avută în vedere. Cu alte cuvinte, un eșantion nu este reprezentativ

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

în general, ci are o anumită reprezentativitate în raport cu o anumită caracteristică și o altă reprezentativitate în raport cu o altă caracteristică.

Gradul de reprezentativitate al unui eșantion depinde de trei factori importanți - caracteristicile populației din care este extras, de mărimea eșantionului și de procedura de eșantionare – factori care au fost sintetizați de Rotariu și Iluț în lucrarea “Ancheta sociologică și sondajul de opinie” și pe care îi voi reaminti în cele ce urmează.

Cum am spus deja reprezentativitatea unui eșantion este dată de capacitatea acestuia de a reproduce o serie de caracteristici existente în populație. Dacă o caracteristică este mai omogen distribuită într-o populație un același eșantion va fi mai reprezentativ pentru acea caracteristică decât pentru o altă caracteristică care este distribuită mai eterogen în aceeași populație. Sau altfel spus, pentru a obține o aceeași reprezentativitate, pentru o caracteristică în raport cu care populația este mai omogenă este nevoie de un eșantion de mărime mai mică decât pentru o caracteristică în raport cu care populația este mai eterogenă.

Mărimea eșantionului se referă la numărul de elemente care îl compun și care trebuie investigate pentru a obține rezultate cât mai precise. Intuitiv, un eșantion este cu atât mai reprezentativ cu cât cuprinde mai multe elemente din populația vizată, în felul acesta obținându-se o reproducere mai bună a acesteia. Dar creșterea nivelului de reprezentativitate nu este direct proporțională cu creșterea numărului de elemente din populația vizată care sunt incluse în eșantion, adică nu avem o relație lineară între cele două componente, dimpotrivă această relație poate fi reprezentată sub forma unei curbe asemănătoare celei din figura de mai jos (Figura 2.). Astfel, dacă modificăm mărimea eșantionului cu o cantitate K de elemente, iar eșantionul cuprinde inițial un număr mic de elemente, modificarea gradului de reprezentativitate este mai mare decât dacă modificăm mărimea eșantionului cu aceeași cantitate K de elemente dar eșantionul este compus inițial dintr-un număr mare de elemente.

Mărimea eșantionului este independentă de mărimea populației din care este extras. Un eșantion de o anumită mărime și constituit după aceleași proceduri are același grad de reprezentativitate și atunci când este extras din populația unei țări și atunci când este extras din populația unui oraș. Consecința faptului că reprezentativitatea unui eșantion nu depinde de mărimea populației din care este extras este aceea că acesta are un anumit grad de reprezentativitate pentru întreaga populație, dar subeșantioanele în care se împarte și care respectă proporția diferitelor segmente ale populației nu mai au același grad de reprezentativitate pentru aceste segmente ca și eșantionul inițial.

În ceea ce privește procedura de eșantionare, aceasta influențează atât gradul de reprezentativitate al unui eșantion cât și posibilitatea exprimării numerice a acesteia. Din punct de vedere tehnic – matematic, calcularea reprezentativității unui eșantion este posibilă numai în cazul eșantioanelor probabilistice sau aleatoare. Un eșantion probabilistic este acel eșantion pentru care fiecare element din populația vizată are o șansă calculabilă și nonnulă de a fi selectat în eșantion. Posibilitatea calculării șansei ca un element din populație să fie selectat în eșantion permite calcularea unei marje de eroare (d) și a unui nivel de probabilitate (P) prin care să fie exprimată cantitativ reprezentativitatea eșantionului. În cazul eșantioanelor neprobabilistice, cele pentru care șansa unui element al populației de a face parte din eșantion nu este cunoscută, nu poate fi calculat gradul de reprezentativitate și prin urmare nici nu se poate vorbi de reprezentativitatea lor.

Unitatea 3

Obiectiv: Proceduri de eșantionare. Tipuri de eșantioane

Cuvinte cheie: selecție la întâmplare, eșantioane probabilistice, eșantioane neprobabilistice

Proceduri de eşantionare. Tipuri de eşantioane

Distincția clasică în ceea ce privește tipurile de eşantioane este aceea între eşantioanele probabilistice sau aleatoare și cele neprobabilistice.

În primul caz în procesul de selectare a unui element din populație pentru a face parte din eşantion se presupune că se face “la întâmplare” fără să intervină în vreun fel subiectivismul celui care aplică procedura de eşantionare și nici vreun alt fenomen care să afecteze șansa unuia sau unor indivizi de a fi selectați. Dacă fiind această constrângere, următoarele situații: alegerea *la întâmplare* a unui număr de oameni care intră într-o instituție 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 localități rurale ale unui județ, selectând tot *întâmplător* localități rurale care se află pe șoseaua care leagă două orașe ale județului respectiv, etc, nu vor duce la constituirea unor eşantioane probabilistice. Motivul pentru care nu vom obține în aceste cazuri eşantioane aleatoare este acela că în alegerea pe care o facem “la întâmplare” excludem fie intenționat, fie neintenționat o parte din elementele populației vizate. Astfel în primul caz, dacă vrem să alegem un eşantion reprezentativ pentru cei care frecventează o anumită instituție și vom selecta “la întâmplare” doar pe cei care intră în acea instituție în intervalul orar amintit îi vom exclude pe toți cei care la momentul respectiv nu au șansa de a intra în acea instituție, similar vom exclude fără să vrem gospodăriile care nu au șansa de a se găsi pe străzile pe care ne plimbăm sau localitățile rurale care nu au șansa de a se afla pe șoseaua care leagă cele două orașe între care ne deplasăm. Mai mult, nu putem calcula șansa pe care fiecare element din cele trei cazuri prezentate mai sus – persoane, gospodării, localități rurale – le are de fi selectat în eşantion. În toate aceste cazuri este clar că *întâmplarea* favorizează anumite elemente și anulează șansele altor elemente care sunt excluse *a priori* întrucât nu au șansa de a se afla la locul sau pe traseul pe care se deplasează cel care face selecția.

Pentru a evita aceste situații ar trebui să avem o situație clară a tuturor elementelor care compun o populație și să le putem identifica fără echivoc.

Așa cum am arătat deja în paragraful anterior, în cazul procedurilor de eşantionare probabilistice fiecare element care compune o populație trebuie să aibă o șansă diferită de zero și 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 și cea mai des utilizată în domeniul cercetărilor practice și este considerată procedura de referință, “ideală”, atunci când se pune problema stabilirii unor tehnici de eşantionare. Asumsiile de bază ale acestei tehnici sunt acelea că: fiecare element al populației vizate are exact aceeași șansă ca și oricare alt element al aceleiași populații de a fi selectat în eşantion, iar selectarea unui element în eşantion nu a influențat în nici un fel șansele altui element de a fi selectat. Tehnica tipică sau modelul de realizare al acestui tip de eşantionare este reprezentat de *metoda urnei*, situație în care fiecărui element dintr-o populație vizată îi corespunde o bilă; toate bilele corespunzătoare unor membrii ai populației vizate

sunt introduse într-o urnă după care sunt amestecate și se extrage pe rând câte una până se ajunge la un număr de bile egal cu numărul de elemente care vor compune eșantionul. Simplu de pus în practică din punct de vedere teoretic, procedura astfel definită întâmpină o serie de dificultăți mai ales atunci când se lucrează cu populații mari, situație 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 populația unei țări spre exemplu.

O a doua metodă de realizare a unei eșantionări simple aleatoare sunt *tabelele cu numere aleatoare*. Procedura constă în generarea unor șiruri de numere aleatoare și introducerea lor într-un tabel similar celui prezentat mai jos (Tabelul 1.). Fiecărui element din populația vizată, care trebuie identificat univoc, i se atribuie un număr de la 1 la N. Cel care realizează selecția, alege la întâmplare un număr din șirul de numere aleatoare și caută apoi în populația vizată elementul cu numărul de ordine reprezentat de numărul aleator respectiv, element care va face parte din eșantion. După care din tabelul de numere aleatoare este ales numărul următor și se identifică din nou în populația vizată elementul cu numărul de ordine identic cu numărul aleator, element care este și el introdus în eșantion. Procedura continuă în acest fel până la completarea numărului de elemente necesare constituirii eșantionului. În cazul în care unui număr aleator nu îi corespunde un număr atribuit unui element din populația vizată, acest nu este utilizat și 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 eșantion. Situație 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 funcție de necesități, din mai mult de două cifre de către cel care face eșantionarea 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 adaugă prima cifră din coloana a doua se obțin numerele: 672, 533, 452, 769, 230, 091.

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

Spre exemplificare, să presupunem că populația vizată este formată din 5000 de gospodării, și dorim să constituim un eșantion format din 250 de gospodării. Pasul de eșantionare 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 eșantionului nostru. Următoarele gospodării care vor face parte din eșantion sunt cele cărora le corespund numerele: 52, 77, 102, 127, 152, 177 și așa mai departe până la selectarea a 250 de gospodării.

2. Eșantionarea prin stratificare

Eșantionarea prin stratificare are la bază tot o procedură de alegere aleatoare. Această metodă este utilizată atunci când cel care face eșantionarea are motive să creadă că populația vizată este compusă din mai multe subpopulații sau subgrupuri distincte, denumite tehnic straturi. Realizarea din punct de vedere practic a unui eșantion prin stratificare presupune următorii pași: populația vizată este împărțită în subpopulații în funcție de un anumit criteriu care este deja cunoscut, după care este constituit un eșantion care la rândul lui va fi compus din atâtea subeșantioane câte subpopulații există în populația vizată. În interiorul fiecărei subpopulații elementele care vor fi introduse în eșantion sunt selectate aleator.

Spre exemplu, să presupunem că 30% din populația unei regiuni locuiește în localități rurale, 20% locuiește în orașe cu până la 50.000 de locuitori, 15% locuiește în orașe care au de la 50.001 la 100.000 de locuitori, iar restul de 35% locuiește în orașe de peste 100.000 de locuitori. Un eșantion stratificat format din 1000 de persoane va cuprinde 300 de persoane din mediul rural, 200 de persoane care locuiesc în orașe cu până la 50.000 de locuitori, 150 de persoane care locuiesc în orașe care au între 51.001 și 100.000 de locuitori și 350 de persoane care locuiesc în orașe de peste 100.000 de locuitori.

Principiul de bază al acestui tip de eșantionare este acela că, cu cât o populație este mai omogenă cu atât este mai ușor să se extragă din aceasta un eșantion reprezentativ. De asemenea, cu cât o populație este mai omogenă în raport cu o caracteristică, cu atât mărimea eșantionului necesar pentru a reproduce cu o anumită acuratețe acea caracteristică este mai mică în comparație cu mărimea unui eșantion extras dintr-o populație care este mai eterogenă în raport cu aceeași caracteristică.

Mărimea subeșantioanelor poate să pătreze proporția subpopulațiilor, situație în care vom vorbi de eșantionare prin stratificare proporțională. În felul acesta se asigură pentru toate elementele populației vizate o șansă egală de a fi selectate în eșantion.

Există însă și situații în care este recomandat ca subeșantioanele să nu pătreze proporțiile subpopulațiilor. Acest lucru se întâmplă mai ales atunci când unele subpopulații sunt reduse din punct de vedere numeric și în consecință, dacă ar fi păstrate proporțiile, și subeșantioanele ar fi formate dintr-un număr mic de elemente care nu ar avea un nivel de reprezentativitate rezonabil. În această situație se recurge la o stratificare diproporționată a eșantionului sau o stratificare ponderată, prin suprareprezentarea în eșantion a subpopulațiilor mai puțin numeroase, urmând ca la prelucrarea datelor aceste "abateri" să fie corectate prin metode statistice. În această situație șansele elementelor aparținând diferitelor subpopulații de a intra în eșantion sunt diferite: elementele care provin din subpopulațiile mai puțin numeroase având șanse mai mari de fi selectați în eșantion decât elementele care provin din subpopulațiile mai numeroase.

Indiferent de modalitatea în care sunt constituite subeșantioanele, păstrând sau nu proporțiile, eșantionarea prin stratificare presupune existența în momentul inițial al punerii în practică a procedurii de eșantionare a unei informații suplimentare despre populația vizată în comparație cu situația în care este utilizată eșantionarea simplă aleatoare. Această informație poate fi obținută cu ajutorul altor studii sau din alte surse de informare cu privire la populația vizată.

În ceea ce privește gradul de reprezentativitate al eșantionelor realizate prin stratificare în comparație cu gradul de reprezentativitate al eșantioanelor simple aleatoare, se admite în general că

este mai bun. Mai clar spus, dintre două eșantioane de aceeași mărime unul obținut prin eșantionare prin stratificare iar altul prin eșantionare simplă aleatoare, se consideră că primul are o reprezentativitate mai bună, în situația în care criteriile pe baza cărora se face eșantionarea au o legătură de tip statistic cu caracteristicile care fac obiectul cercetării.

3. Eșantionarea multistadială

Până acum am prezentat situații în care există un anumit cadru de eșantionare - liste care să cuprindă elementele unei populații - și situații în care pe lângă faptul că există un anumit cadru de eșantionare cercetătorul mai are la îndemână și o serie de criterii pe baza cărora o populație poate fi împărțită în subpopulații sau grupuri. În această din urmă situație din fiecare grup este extras un subeșantion care va face parte din eșantionul final.

Există însă și situații în care nu există un cadru de eșantionare și nici nu este necesară crearea unui întrucât nu toate elementele acelei populații vor fi incluse în eșantion. Dacă populația poate fi considerată ca fiind formată din grupuri, iar între aceste grupuri există o anumită asemănare, atunci are sens să nu fie selectați în eșantion indivizi din toate grupurile ci numai indivizi din anumite grupuri. Procedura de eșantionare care are la bază acest principiu este denumită: eșantionare multistadială. În această situație populația vizată este împărțită în grupuri în funcție de un anumit criteriu, aceste grupuri la rândul lor pot fi considerate ca fiind formate din alte grupuri și așa mai departe. Date fiind aceste condiții, selectarea elementelor care vor compune eșantionul poate începe prin selectarea grupurilor din care fac parte aceste elemente. Astfel, într-o primă fază sunt selectate aleator o parte din grupurile populației vizate, după care din fiecare grup selectat în prima fază vor fi selectate tot aleator alte grupuri mai mici și așa mai departe până când se ajunge la nivelul elementului de bază din care este compusă populația vizată. Spre exemplu, dacă dorim să alegem un eșantion din populația unui oraș, într-o primă fază putem selecta cartiere din acel oraș, apoi străzi, blocuri, apartamente și în cele din urmă persoanele care ne interesează.

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

În ceea ce privește reprezentativitatea unui astfel de eșantion, se consideră în general că, la volum egal, este mai puțin reprezentativ în comparație cu un eșantion obținut prin stratificare sau în comparație cu un eșantion obținut prin procedee simple aleatoare. Reprezentativitatea mai scăzută este rezultatul eliminării la diferite nivele a unor grupuri de elemente din populația vizată. Cu cât aceste grupuri care sunt eliminate sunt mai mari și cu cât sunt mai diferite în comparație cu grupurile care nu au fost eliminate cu atât este mai mare riscul de a greși.

Eșantioane neprobabilistice

Alături de aceste proceduri de eșantionare probabilistice în practica de cercetare sunt utilizate și o serie de tehnici mai puțin riguroase în ceea ce privește selectarea celor care vor compune un eșantion. Lipsa de rigurozitate se referă mai ales la neacordarea unei atenții speciale calculării sau egalizării șanselor fiecărui individ din populația vizată de a face parte din eșantion. Eșantioanele obținute în acest fel sunt denumite eșantioane neprobabilistice. Astfel de eșantioane se constituie în următoarele situații :

- persoane care se oferă voluntar pentru a fi investigate;
- persoane care își desfășoară activitatea într-o instituție 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 alți participanți la emisiunile respective;
- persoane intervievate pe stradă sau în anumite spații publice;

În cadrul acestor tehnici de eșantionare neprobabilistică cele mai des utilizate sunt eșantionarea “pe cote” și eșantionarea tip “bulgăre de zăpadă”.

1. Eșantionarea pe cote

Eșantionarea pe cote este probabil cea mai des utilizată procedură de eșantionare neprobabilistică utilizată atunci când se lucrează cu populații numeroase. Din punct de vedere al realizării practice această procedură este similară eșantionării prin stratificare prin aceea că populația vizată este stratificată după o serie de criterii însă în interiorul straturilor nu sunt selectați aleator, ci selecția acestora este lăsată la latitudinea operatorilor de anchetă. Acestora le sunt indicate numai anumite “cote” care indică frecvența cu care să fie selectați subiecții care au anumite caracteristici. Spre exemplu dacă în populația vizată avem 49% bărbați și 51% femei și 20% au studii superioare iar restul de 80% nu au astfel de studii, iar eșantionul este format 1000 de persoane, atunci în cadrul acestuia vor fi cuprinși 490 de bărbați și 510 femei, 200 de persoane cu studii superioare și 800 de persoane care nu au absolvit învățământul superior. În această situație dacă sunt utilizați 10 operatori de interviu fiecare i se cere să chestioneze 49 de bărbați și 51 de femei, 20 de absolvenți de învățământ superior și 80 de persoane care au absolvit o formă de învățământ alta decât facultatea. Pentru a se limita subiectivitatea operatorilor în selecționarea celor care vor fi incluși în eșantion se recomandă stabilirea a cât mai multor criterii de stratificare a populației vizate.

Avantajul unui astfel de procedeu de stratificare este acela că nu necesită existența unui cadru de eșantionare, lucru care în unele situații este greu de realizat, iar munca operatorilor este mult ușurată prin aceea ce nu trebuie să caute o persoană anume ci au libertatea de a alege pe cine vor cu condiția deținerii anumitor caracteristici vizate de cercetare.

2. Eșantionarea tip “bulgăre de zăpadă”

Este o procedură de eșantionare utilizată în situația în care nu există informații suficiente pentru a identifica toți indivizii care compun o anumită populație, ci este posibilă doar identificare doar a câtorva astfel de indivizii. Date fiind aceste circumstanțe, analiza unei populații vizate începe cu investigarea indivizilor cunoscuți după care acestora li se cere să precizeze dacă este posibil și alte persoane care se presupune ca întrunesc caracteristici vizate de cercetare. Procedeu se desfășoară în acest fel până când sunt identificați atâția indivizi câți sunt necesari constituirii unui eșantion. Se utilizează acest procedeu în cazul în care populația vizată este formată spre exemplu din oameni care au anumite hobby-uri sau pasiuni, preocupări și despre care de obicei nu se cunosc în faza inițială multe informații și nu se știe nici câte astfel de persoane compun populația vizată.

Concluzie

Eșantionarea este un procedeu des utilizat în practica de cercetare în diferite domenii ale activității umane. De la medicul care face analize de laborator prelevând o probă de sânge de la un pacient și până la cei care sunt interesați de aspecte ale opiniei publice în diferite domenii precum: preferințele

electorale, acordul sau dezacordul cu anumite politici publice sau decizii administrative, etc. În funcție de tematica avută în vedere și de informațiile disponibile cu privire la populația vizată procedurile de eșantionare respectă mai mult sau mai puțin anumite rigori în ceea ce privește selectarea elementelor din populație care vor constitui eșantionul.

În practică procedurile de eșantionare prezentate pe parcursul acestui capitol suferă o serie de abateri și de *adaptări* sau *ajustări*. De cele mai multe ori acestea constau în combinarea mai multor tehnici de eșantionare în felul acesta sperându-se obținerea unor informații cât mai corecte și mai precise despre populația avută în vedere.

Întrebări:

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

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 științele sociale. Ed. Polirom, Iași, 1999.
6. Rotariu, T., Iluț P., Ancheta sociologică și sondajul de opinie. Ed. Polirom, Iași, 1997.
7. Schimdt, M., Understanding and using statistics. Basic concepts, Second Edition, Lexington, Massachusetts, Toronto, 1979.

Unitatea 4

Obiectiv: Prezentarea aspectelor matematice ale eșantionării

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

Aspecte matematice ale eșantionării. Teste de semnificație

Valori măsurate pe populație și pe eșantion. Intervale de confidență

Extrăgând un eșantion dintr-o populație și măsurând pe acesta valoarea medie a unei caracteristici sau variabile putem spune într-o oarecare măsură că această valoare aproximează o valoare a aceleiași caracteristici din populație. Cu toate acestea întrebarea care se ridică este: cât de siguri putem fi de rezultatele obținute dat fiind că eșantionul extras la un moment dat este doar unul din multele eșantioane care pot fi extrase dintr-o populație? Spre exemplu, dorim să estimăm nivelul de inteligență al elevilor unei școli și pentru aceasta extragem aleator un eșantion format din 25 de elevi cărora le aplicăm un test de inteligență și obținem o valoare medie a indicelui de inteligență de 108 și o abatere standard de 12. Bazându-ne pe aceste rezultate, ce putem spune despre nivelul de inteligență al elevilor școlii respective? Eșantionul de 25 de elevi este evident doar unul din eșantioanele care ar fi putut fi extrase și prin urmare și media de 108 obținută de cei care au făcut parte din eșantion este doar una din posibilele medii. Mai clar spus, 108 este doar una dintre mediile din distribuția de medii care ar putea fi obținută extrăgând multe eșantioane formate din 25 de elevi ai școlii respective. Problema este: cât de aproape este această medie de valoarea reală a indicelui de inteligență a tuturor elevilor acelei școli? și care este valoarea medie a indicelui de inteligență pentru întreaga populație 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 eșantioane suficient de mari distribuția mediilor măsurate pe aceste eșantioane este întotdeauna normală, chiar dacă valorile caracteristicii inițiale sunt sau nu normal distribuite într-o populație vizată. Mediile unei caracteristici măsurate pe multe eșantioane pot fi privite ca formând o nouă variabilă pentru care vom putea calcula evident o medie și o abatere standard. Valoarea medie a noii variabile (media mediilor măsurate pe eșantioanele extrase din populația vizată) este egală cu media valorii din populație a caracteristicii vizate, iar abaterea standard a acestei variabile, în cazul în care eșantioanele sunt extrase printr-o simplă aleatoare cu reintroducerea elementului extras în populație (acordând deci o șansă egală fiecărui element de a fi extras), este egală cu abaterea standard a variabilei urmărite măsurată pe un eșantion oarecare împărțită la rădăcină pătrată din mărimea eșantionului. Abaterea standard a noii variabile este denumită *eroare standard* (e):

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

Revenind la întrebarea din exemplul de mai sus: care este valoarea medie a indicelui de inteligență pentru întreaga populație de elevi vizată? Un răspuns exact nu poate fi dat întrucât nu a fost investigată întreaga populație. Știind însă că distribuția valorilor medii măsurate pe multe eșantioane extrase din populația de elevi vizată este normală putem calcula un interval despre care să spunem că, cu probabilitate de 95% include media indicelui de inteligență din întreaga populație vizată. Acest interval este cuprins între plus două și minus două erori standard ($e = 12/5 = 2,4$) în jurul valorii medii obținute pe un eșantion oarecare extras din acea populație, adică între $108 - 4,8$ și $108 + 4,8$. Intervalul astfel construit poartă denumirea de *interval de confidență*.

Teste de semnificație. Inferența statistică

Adeseori observăm diferențe între rezultatele obținute atunci când se fac măsurători pe două eșantioane diferite extrase din aceeași populație. Spre exemplu, 17% dintre cei chestionați în cadrul unei anchete sociale sunt de acord cu o anumită decizie a administrației publice locale la un anumit moment dat de timp, dar numai 11% au aceeași opinie la un alt moment de timp. Problema care se pune în această situație este: cât de reală sau de semnificativă este diferența între cele două grupuri – cei chestionați la un moment de timp și cei chestionați la un moment de timp ulterior? Este această diferență autentică sau este rezultatul fluctuațiilor firești ale eșantionării?

Similar ne putem întreba: ce se poate spune despre valoarea unei caracteristici dintr-o populație pe baza rezultatelor obținute atunci când este investigat un eșantion? vor fi rezultatele obținute atunci când se fac măsurători pe un eșantion identice cu rezultatele obținute atunci când se fac măsurători pe întreaga populație? iar dacă nu, diferențele identificate sunt semnificative sau nu? mărimea eșantionului influențează modul în care rezultatele obținute reflectă caracteristici ale populației?

Toate aceste întrebări sunt justificate întrucât, așa cum am arătat în capitolul dedicat eșantionării, eșantioanele nu reproduc exact caracteristicile unei populații, ci există o anumită diferență între valoarea unei caracteristici măsurată pe un eșantion și valoarea aceleiași caracteristici măsurată pe populația din care este extras eșantionul. Cu toate acestea de multe ori suntem puși în situația de a trage concluzii cu privire la starea unei populații pornind de la măsurători efectuate la nivelul unui eșantion, cu alte cuvinte se pune problema de a face inferențe de la eșantion la populație. Bazate pe numere utilizate pentru a sumariza, evalua sau analiza un set de informații cu privire la un fenomen analizat, numere care în literatura de specialitate sunt denumite *statistici*, inferențele de acest fel sunt și ele denumite inferențe statistice. Inferențele statistice, ca urmare a faptului că eșantioanele pe baza cărora sunt realizate constituie doar aproximări ale unei populații, prezintă neajunsul de a putea produce concluzii eronate. Prin urmare, atunci când se compară două valori ale unor caracteristici dintre care cel puțin una a fost obținută prin măsurători efectuate pe un eșantion, se pune problema semnificației diferenței dintre ele.

Din punct de vedere cantitativ, *vom spune că diferența între două valori, fie că una este măsurată pe un eșantion și alta pe o populație, fie că amândouă valorile sunt măsurate pe eșantioane, 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 diferență care nu este semnificativă potrivit definiției 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 stabilește dacă diferențele între valori sunt sau nu semnificative statistic. Fiecare set de astfel de reguli poartă denumirea de *test de semnificație* și are scopul de a ajuta la stabilirea unei *concluzii statistice* cu privire la starea unor caracteristici ale populației investigate. Testele de semnificație nu sunt probe absolute ale existenței sau non-existenței unei diferențe semnificative între două valori, ele doar permit estimarea în raport cu o ipoteză prealabilă a probabilității prezenței unei diferențe reale între valori. Cel mai adesea astfel de ipoteze în care sunt enunțate predicții cu privire la valorile unor caracteristici avute în vedere în cercetare iau forma *ipotezei nule*, adică a afirmării inexistenței unei diferențe semnificative între două valori comparate. Mai clar spus, ipoteza nulă este ipoteza care afirmă că două mărimi A și B măsurate pe eșantioane diferite sau una măsurată pe un eșantion și una pe o populație, sunt egale. Ținând cont de toate acestea un *test de semnificație poate fi definit ca fiind măsura diferenței dintre două valori în raport cu ipoteza nulă*.

Ipoteza nulă este testată în felul următor: dacă cu un anumit nivel de probabilitate diferența dintre cele două valori comparate este mai mare decât o valoare maximă prestabilită atunci ipoteza

nulă este respinsă și vom spune ca acea diferență este semnificativă. În caz contrar – cu un anumit nivel de probabilitate diferența între valori este mai mică decât o valoare maximă prestabilită – ipoteza nulă este susținută și vom spune că diferența respectivă nu este semnificativă. O întrebare firească 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 diferență care cu o probabilitate de 95% nu depășește o valoare maximă prestabilită că este semnificativă statistic.

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

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

Testul Z

Este un test de semnificație utilizat în cazul în care se compară valorile unor caracteristici măsurate pe eșantioane mari (de ordinul a sute sau mii de indivizi). Cele două valori comparate pot fi măsurate fie una pe o populație și una pe un eșantion, fie amândouă valorile sunt măsurate pe eșantioane diferite.

În prima situație, fie **a** și **b** cele două valori ale aceleiași caracteristici, dintre care valoarea **a** este măsurată pe o populație iar valoarea **b** este măsurată pe un eșantion și fie **e** eroarea standard a caracteristicii luate în considerare. Testul Z este definit după formula:

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

și exprimă de fapt diferența dintre valorile **a** și **b** în erori standard. Dacă valoarea testului Z este mai mare de 1.96 atunci diferența 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% diferența 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) și 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 cetățenii unei localități sunt de acord cu introducerea unui nou sistem de impozite. Cu toate acestea într-un sondaj de opinie realizat anterior referendumului pe un eșantion de 900 de persoane indică că doar 37% dintre cetățeni vor fi de acord cu noua grilă de impozitare. Este diferența între cele două valori autentică sau nu? Sau altfel spus, este diferența dintre cele două valori semnificativă?

Pentru a pune în evidență acest lucru calculăm:

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

și

sau

$$e = \frac{0,48}{\sqrt{900}} = 0,016$$

$$e = 1,6\%$$

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

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

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

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

unde n_1 și n_2 sunt mărimile celor două eșantioane, iar σ_1 și σ_2 sunt abaterile standard ale valorilor caracteristicii pentru fiecare dintre cele două eșantioane.

Testul Student (t)

Atunci când se pune problema de a compara valori ale unor caracteristici dintre care cel puțin una este obținută prin măsurători efectuate pe eșantioane de mărimi mici (până la 30 de indivizi) corespondentul 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 față de testul Z constă în modul de calcul al erorii standard (**e**) care se face după formula:

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

dacă una dintre valori este măsurată pe o un eșantion de mărime n și una pe o populație, și după formula:

$$e = \sqrt{\frac{\sum(x_i - \bar{x}_1)^2 + \sum(x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

dacă cele două valori sunt măsurate pe eșantioane ale căror mărimi sunt n_1 respectiv n_2 .

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

respectiv

$$\nu = n - 1$$

$$\nu = n_1 + n_2 - 2$$

Testul χ^2 (hi pătrat)

Testele Z și Student sunt utilizate pentru a testa ipoteze care se referă la valori, sau parametrii (medii sau proporții), măsurate pe populații sau pe eșantioane, motiv pentru care sunt adeseori cunoscute și sub denumirea mai largă de *teste parametrice*. Există însă multe situații în care ipotezele nu pot fi testate utilizând doar medii sau proporții. Acest lucru se întâmplă spre exemplu atunci când datele cu care se lucrează nu sunt de tip cantitativ. Există apoi și alte condiții care trebuiesc îndeplinite în cazul testelor parametrice - utilizarea unor eșantioane mari sau a unor eșantioane extrase din populații normal distribuite astfel încât și forma distribuției de eșantionare să fie cunoscută - condiții care nu întotdeauna pot fi îndeplinite.

Pentru a depăși acest tip de neajunsuri au fost construite și o serie de teste a căror mod de operare nu presupune existența unor asumții cu privire la populația vizată sau cu privire la datele pe care le avem la dispoziție 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 și Student și anume încearcă să ofere un răspuns întrebării: dată fiind o mulțime de valori observate ale unei caracteristici, modul de distribuire a acestor valori poate fi atribuit în întregime fluctuațiilor firești ale eșantionării sau există o serie de alți factori care influențează această distribuire? Și în acest caz, pentru a răspunde acestei întrebări, punctul de plecare este o ipoteză nulă care afirmă că nu există alți factori care să influențeze distribuția valorilor observate ale unei variabile.

Pentru a ilustra modul de operare al testului χ^2 (hi pătrat) vom utiliza exemplul următor. Fie următoarea situație ipotetică: 100 de funcționari ai unei instituții publice sunt întrebați cu privire la ceea ce îi nemulțumește cel mai mult la locul de muncă, răspunsurile oferite având următoarele frecvențe:

	frecvența
modul de organizare al activităților	24
modul în care sunt tratați de șef	10
existența unui program fix de lucru	27
lipsa unui spirit de echipă	11
lipsa unor rezultate vizibile	28

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

Ipoteza nulă în această situație ar fi aceea că fiecare dintre motivele enumerate mai sus nemulțumește în egală măsură pe funcționarii acelei instituții, adică fiecare dintre cele cinci

răspunsuri având aceeași probabilitate de a fi indicat de către respondenți. Din punct de vedere statistic aceasta ar însemna că frecvențele observate ale răspunsurilor primite pot fi considerate egale cu frecvențele așteptate. Dacă ipoteza nulă este susținută 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:

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

unde O_i reprezintă frecvențele observate, iar A_i reprezintă frecvențele așteptate (adică distribuția aleatoare a răspunsurilor pe cele cinci categorii ale caracteristicii analizate – motiv de nemulțumire).

În cazul nostru cele două frecvențe sunt:

	O_i	A_i	$O_i - A_i$
modul de organizare al activităților	24	20	4
modul în care sunt tratați de șef	10	20	-10
existența 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) obținem:

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

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

$$\chi^2 = 0.80 + 5.00 + 2.45 + 4.05 + 3.20$$

$$\chi^2 = 15.50$$

Valoarea obținută pentru χ^2 (15.50) se compară cu valorile critice ale distribuției *hi pătrat* (Tabelul 3.) pentru diferite nivele de probabilitate, dintre care cel mai des utilizat este și de această dată pragul de 0,95 (95%). Ca și în cazul testului Student compararea valorilor critice ale unei distribuții observate cu distribuția *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 $k = 5 - 1 = 4$ și căutând în tabelul cu valori critice ale lui *hi pătrat* găsim că pentru nivelul de probabilitate de 95% și 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 funcționari nu se

distribuie aleator ci există un anumit motiv de nemulțumire care este mai acut decât celelalte, iar acest rezultat nu este generat de fluctuațiile de eșantionare.

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

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

În această situație se poate pune întrebarea: există sau nu o preferință a unei anumite categorii de vârstă pentru schimbarea modului de alocare a veniturilor? Cu alte cuvinte există o relație între vârstă și acordul cu această schimbare? Pentru a răspunde la această întrebare trebuie să vedem cum ar trebui să arate distribuția în situația în care nu există asociere. Astfel, dacă nu ar exista o relație între variabile, atunci preferințele 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 și o treime să răspundă că “nu știu”. Acest lucru raportat la frecvențele din tabelul de mai sus ar însemna: 60 de persoane cu vârsta sub 25 de ani să fie de acord cu schimbarea (adică o treime din cele 180 de persoane cu vârsta sub 25 de ani cuprinse în eșantionul nostru), 66,67 persoane cu vârsta cuprinsă între 26 și 45 de ani și așa mai departe:

Frecvențe așteptate				
	Da	Nu	Nu știu	Total
cei cu vârsta sub 25	60	66	54	180
cei cu vârstă între 26 și 45 de ani	66,67	73,33	60	200
cei cu vârsta peste 45 de ani	73,33	80,67	66	220
Total	200	220	180	600

Calculându-l pe *hi pătrat* obținem:

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

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

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

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

unde *j* reprezintă numărul de rânduri ale tabelului în care sunt dispuse frecvențele și *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% și 4 grade de libertate îi corespunde valoarea 9,488, valoare mai mică decât valoarea calculată a lui χ^2 . În această situație vom spune că ipoteza potrivit

căreia nu există asociere între vârstă și preferința pentru schimbarea modului de alocare a veniturilor se respinge.

Teste parametrice sau non-parametrice?

Când utilizăm teste parametrice și când utilizăm teste non-parametrice pentru a analiza un set de date? Răspunsul la această întrebare nu este întotdeauna foarte tranșant.

Astfel, nu vom putea utiliza teste parametrice dacă datele pe care le avem la dispoziție sunt de tip calitativ, motivul este acela 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 situație 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 și aceasta pentru că modul lor de operare ia în considerare mai multă informație despre caracteristica avută în vedere. Dar acest lucru se face cu anumite asumții, dintre care cea mai importantă este distribuția normală a valorilor caracteristicii analizate.

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

Practica a demonstrat că amândouă tipurile de teste pot fi utilizate cu același succes cu condiția luării în calcul a avantajelor și dezavantajelor fiecăruia.

Probleme:

1. Să presupunem că 35,4% dintre cetățenii unei localități au votat partidul X la alegerile locale. Un sondajele de opinie realizat în perioada pre-electorală pe un eșantion de 1000 de persoane acordă însă acestui partid 39% dintre intențiile de vot ale electoratului. Este diferența între cele două valori autentică sau nu?
2. Dintre 200 de elevii ai unei școli intervievați cu privire la dificultățile de învățare pe care le întâmpină : 38 au răspuns că acestea își au originea în programul încărcat de la școala, 62 au răspuns că lipsa unei dotări adecvate a școlii le crează astfel de dificultăți, 56 au răspuns că modul de structurare a materiilor învățate este cauza dificultăților de învățare, iar 46 au pus că dificultățile de învățare se datorează unor cauze externe școlii. Există un motiv care să determine într-o mai mare măsură dificultăți de învățare pentru elevii școlii avute în vedere?
3. Testele de semnificație nu sunt probe absolute ale existenței sau non-existenței unei diferențe semnificative între două valori. Comentați această afirmație.

Tabelul 1. Proporția din aria totală (10.000) ce corespunde distanței dintre medie și 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	0755

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	4993,12 9									
3.3	4995,16 6									
3.4	4996,63 1									

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 și 0.01, în funcție 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 și 0.01, în funcție de numărul gradelor de libertate (ν)

ν	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 Științifică, București, 1969.

Bibliografie:

- 1 Rotariu, T., (coord.). Metode statistice aplicate în științele sociale. Ed. Polirom, Iași, 1999.
- 2 Rotariu, T., Iluț P., Ancheta sociologică și sondajul de opinie. Ed. Polirom, Iași, 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 științele sociale și modelelor de analiză a acestora

Ghid de studiu:

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

Unitatea 1

Obiectiv: Introducerea noțiunii de variabilă și a tipurilor de variabile

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

Variabile. Tipuri de variabile.

Caracteristicile populației despre care facem inferențe pe baza eșantionului se numesc parametri. Caracteristicile eșantionului pe baza cărora inferăm se numesc pur și simplu statistici. În exemplul de mai sus, 55% reprezintă o statistică descriptivă, deoarece ea descrie sintetic o caracteristică a eșantionului. Cele mai multe studii sunt însă interesate în aflarea parametrilor, care în general sunt necunoscuți (exemple: Câți săraci există în România? Care este procentul din populație de susținători ai unui partid? etc.). Eșantioanele și statisticile descriptive sunt utile în măsura în care ele pot oferi informații despre parametri de interes. Statistica inferențială este aceea care permite obținerea unei măsuri a acurateții statisticilor folosite pentru **estimarea** valorii parametrilor. În consecință, atunci când întreaga populație este cuprinsă într-un studiu, statistica inferențială nu este necesară.

În final ne vom opri asupra unei ultime noțiuni deosebit de importante pentru studiul statisticii, și anume asupra **variabilelor**. Vom defini variabila ca fiind orice caracteristică a membrilor unei populații sau unui eșantion care variază (în respectiva populație/eșantion). Astfel, culoarea părului indivizilor dintr-o populație este o variabilă în măsura în care indivizii care compun respectiva populație au păr de culori diferite. Dacă toți indivizii ar fi blonzi, să zicem, atunci culoarea părului ar fi constantă în respectiva populație. Cu cât o caracteristică are o variație mai mare, cu atât respectiva populație este mai eterogenă și, invers, cu cât o caracteristică dată are o variație mai mică, cu atât respectiva populație va fi mai omogenă, din perspectiva respectivei caracteristici. În exemplul de mai sus, valorile posibile ale variabilei "culoarea părului" ar fi "brunet", "blond", "roșcat" etc.. Fiecare individ (statistic) poate lua o singură valoare pentru o variabilă..

Variabilele pot fi clasificate în funcție de multe criterii. Una din distincțiile importante este aceea dintre **variabile discrete** și **variabile continue**. Atât variabilele discrete cât și variabilele continue pot lua o infinitate de valori. Diferența 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 înălțimea clădirilor unui oraș măsurată în metri, iar un exemplu de variabilă discretă îl reprezintă veniturile indivizilor dintr-o populație, măsurate în lei. În cazul primei variabile, între doua valori succesive ale acesteia (de exemplu 5 și 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 și 6 lei nemaexistând subdiviziuni.

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

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

Tabelul I.1 - Niveluri de măsurare a variabilelor

	a) ordonare	b) unitate de măsură	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 și exhaustive. Acest tip de variabile (respectiv scalele folosite în măsurare) indică numai faptul că există o diferență calitativă între categoriile studiate, nu și magnitudinea acestei diferențe. La limită, putem privi aceste variabile ca pe niște tipologii. Câteva exemple de variabile măsurate la nivel nominal sunt: statutul ocupațional al indivizilor (agricultor, salariat, mic întreprinzător, șomer etc.), religia (ortodox, romano-catolic, greco-catolic etc.) apartenența etnică (român, maghiar, rrom etc.), mediul de rezidență (rural, urban) ș.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) și în consecință problema "distanței" sau a intervalelor dintre valori nici nu poate fi pusă. Cu atât mai puțin putem discuta despre existența unui "zero absolut" (exemplu: fiecare individ are un statut ocupațional sau aparține unei etnii, sau altfel spus absența caracteristicilor "statut ocupațional" sau "apartenență etnică" este imposibilă).
2. Nivelul de măsurare **ordinal** implică nu numai clasificarea elementelor în categorii ci și posibilitatea ordonării acestora de la minim la maxim (existența tranzitivității: dacă $a > b$ și $b > c$, atunci $a > c$). Totuși, la acest nivel de măsurare nu este oferită nici o informație cu privire la "distanța" dintre valorile scalei de măsură. Cu alte cuvinte, diferența dintre prima valoare și cea de-a doua poate fi diferită de diferența dintre a patra și a cincea. Exemple de variabile măsurate la nivel ordinal sunt calificativele școlare (cu valorile "insuficient", "suficient", "bine" și "foarte bine"), satisfacția față de anumite aspecte (cu valorile "foarte nesatisfăcut", "nesatisfăcut", "satisfăcut", "foarte satisfăcut") etc..
3. Măsurarea la nivel **de interval**, oferă în plus față de nivel anterior (cel ordinal) și informație referitoare la distanța dintre valorile scalei și este caracterizată de existența unor intervale egale. Totuși, la acest nivel de măsurare nu există un zero absolut, ci mai degrabă unul convențional. Exemple de astfel de scale de măsurare sunt temperatura măsurată în grade Celsius (intervalele dintre valori sunt egale, dar punctul 0 este convențional ales ca fiind temperatura la care apa

ingheață), coeficientul de inteligență - IQ - (daca două persoane au scoruri de 100 și respectiv 150, putem spune ca diferența 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 absența inteligenței).

4. Măsurarea la nivel **de raport** include toate caracteristicile nivelurilor anterioare (ordonare și intervale egale), plus existența unei "origini" sau zero absolut. Acest lucru permite formularea unor afirmații în termeni de proporții (raporturi) între valori. De exemplu, vitezele de raspuns a doi subiecți la un același stimul pot fi comparate în termeni de "timpul de răspuns a fost de două ori mai mare" etc.. Exemple de variabile măsurate la acest nivel sunt vârsta, greutatea, înălțimea, distanța, numărul de copii din gospodărie etc.

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

În funcție 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 și de raport. Reducând cele patru clase la două, putem vorbi de *variabile calitative* (nivelurile nominal și ordinal) și *variabile cantitative* (interval și raport). Datorita caracterului "ierarhic" și cumulativ al nivelurilor de măsurare (de la multe restricții către nici o restricție în ceea ce privește operațiile permise, sau de la "calitativ" la "cantitativ"), vom putea întotdeauna trata o variabilă aflată la un nivel "superior" de măsurare ca și cum ar fi fost măsurată la un nivel "inferior". De exemplu, vârsta măsurată în ani de viață 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 excepție de la această regulă, tratând variabilele ordinale ca și cum ar fi măsurate la nivel de interval. Totuși, 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 științele sociale, Polirom, 1999
-capitolele 1-8
- Sandu, Dumitru, Statistică în științele sociale , Universitatea București, 1992
- capitolele 1, 2, 3, 6, 7

Unitatea 2

Obiectiv: prezentarea analizei univariante a datelor

Cuvinte cheie: tendința centrală, indicatori ai tendinței centrale, indicatori de dispersie sau variație

Analiza univariată a datelor

2.2 Tendința centrală, variația și forma distribuției

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

- tendința centrală (sau centrul distribuției) - adică valoarea "tipică" a acelei variabile
- variația variabilei - ca indicator al gradului de "împrăștiere" a datelor
- forma distribuției

2.2.1 Indicatori (măsuri) ai tendinței centrale

Pentru a descrie centrul unei distribuții, sau tendința centrală a unei variabile, există mai multe măsuri. În aceasta secțiune vor fi discutate cele mai des utilizate: **modul**, **mediana** și **media**.

- Modul este definit ca fiind valoarea cu frecvența cea mai mare a unei distribuții. Altfel spus, modul este acea valoare a variabilei care apare cel mai des într-un eșantion sau într-o populație.*

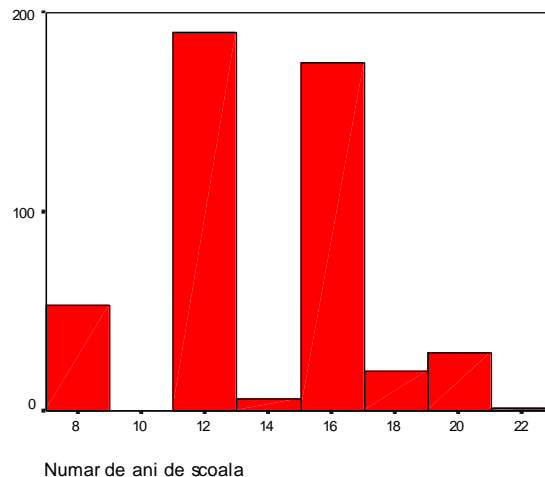
Termenul derivă din francezul "mode", adică modă. În cazul distribuției variabilei "starea civilă a capului gospodăriei" reprezentată în Graficul 1.2, modul este valoarea "căsătorit(ă)" (cu frecvența relativă 80%). De cele mai multe ori, pentru a simplifica lucrul cu datele, valorilor variabilelor nominale li se acordă convențional *coduri numerice*. De exemplu, pentru datele din Graficul 1.2, putem acorda codul 1 pentru valoarea "căsătorit(ă)", codul 2 pentru valoarea "uniune consensuală", codul 3 pentru valoarea "divorțat(ă)" etc.. Chiar dacă aceste coduri sunt numerice, ele trebuie privite ca niște simple simboluri convenționale. 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 - și anume categoria "2001-3000 locuitori", deoarece aceasta este "valoarea" (de fapt intervalul de valori) cu frecvența cea mai mare (651).

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

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

Graficul 1.3 Distribuție bimodală - histograma variabilei "nivel de educație", pentru angajații unei bănci



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

Să luăm de exemplu notele pe care 7 studenți 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 părți egale: 3 dintre studenți (observații) au note mai mici decât 8 și trei dintre ei au note mai mari. Nota 8 este exact la "mijlocul" seriei de date (după ordonare). Este important de reținut că ceea ce contează pentru stabilirea medianei este numărul de observații pe care se face analiza, și nu numărul de valori ale variabilei.

Calculul medianei este relativ simplu atunci când avem de-a face cu un număr mic și impar de observații. Lucrurile se complică puțin atunci când numărul de observații este par, sau dacă numărul de observații e foarte mare și e nevoie să apelăm la tabele de frecvențe. Lucrurile se complică și 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 observații nu va mai exista o singură valoare la mijlocul seriei de date, ci vom avea două valori. În această situație, mediana se află la mijlocul "distanței" dintre aceste valori, sau cu alte cuvinte, este media lor. Să presupunem că am dori să calculăm mediana pentru o serie de 8 studenți, deci un număr par de observații. După ordonare, datele arată astfel: 5, 5, 6, 7, 8, 9, 9, 10. La mijlocul seriei se află valorile 7 și 8. Mediana va fi deci 7,5.

Pentru situațiile în care suntem nevoiți să calculăm mediana pe baza datelor oferite de un tabel de frecvențe, vom utiliza frecvențele cumulate, și vom căuta acea valoare a variabilei sub care se află 50% din cazuri. Pentru datele din Tabelul 1.3, 28,75% din observații iau valoarea 6 sau o valoare mai mică, 46,25% iau valoarea 7 sau mai puțin, 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 studenți iau nota 7 sau mai puțin). Mediana va fi în consecință 8, deoarece, chiar dacă avem un număr par de observații, 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 frecvențe care conține date grupate în intervale de valori (așa 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 conține mediana
- *N* este numărul total de observații
- *nc* este frecvența absolută cumulată a tuturor categoriilor care preced intervalul care conține mediana (adică numărul de observații care iau valori mai mici decât *l*)
- *n* este frecvența intervalului care conține mediana
- *L* este lățimea sau mărimea intervalului care conține mediana

Tabelul 1.3 Distribuția notelor pentru 80 de studenți

Nota	Frecvențe absolute	Frecvențe relative (%)	Frecvențe 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	

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

Din tabel reiese ca mediana este conținută în intervalul 3001-4000 locuitori, deoarece frecvențele relative cumulate ale categoriilor precedente sunt mai mici de 50%, iar frecvența cumulată a intervalului 3001-4000 este aproximativ 63%. Limita inferioară a acestui interval este deci $l = 3001$. Observația careia îi corespunde mediana (numită și individ median) este observația care se află exact la mijlocul seriei ordonate de date, cu alte cuvinte este observația $N/2$, în cazul nostru observația cu numărul 1343. Dacă scădem din acest număr numărul total de observații care au valori mai mici decât 3001, obținem $1343 - 1084 = 259$, unde $1084 = 54 + 379 + 651$ este valoarea lui nc din formula mediane (obținut prin cumularea frecvențelor categoriilor precedente intervalului care conține mediana). Cu alte cuvinte, observația careia îi corespunde mediana este cea de-a 259-a observație din categoria "3001-4000 locuitori", categorie care apare cu frecvența $n = 602$. Am putea acum să ne întrebăm: dacă la 602 comune corespunde o creștere 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, conținută oarecum și în formula mediane: $\frac{259}{602} \times 1000 = 430,2$. 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 și deasupra căreia se află $(100-p)\%$ din cazuri. De exemplu, mediana este percentila 50. Cele mai cunoscute măsuri ale localizării sunt *quartilele*, *quintilele* și *decilele*. Quartilele sunt acele valori ale seriei de date care o împart în patru părți egale, quintilele sunt valorile care o împart în cinci părți 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 sfârșit, sub quartila 3 se află 75% din cazuri, iar deasupra ei se află 25% din cazuri (observații). Din această scurtă prezentare reiese că există numai 3 quartile (Q_1 , Q_2 și Q_3), deoarece pentru a împărți o serie de date în m părți egale sunt suficiente $m-1$ valori. În statistică quartilele, decilele etc. se referă la *valori ale variabilei*. Totuși, în științele sociale sunt folosite destul de des expresii cum ar fi "decila 10 de venituri", "cea mai săracă quintilă", "persoanele aparținând primei decile" etc. Aceste expresii se referă însă la observațiile care iau valori cuprinse între anumite percentile (quantile) și nu la valorile variabilei.

- *Media este probabil cea mai importantă și totodată cea mai populară măsură a tendinței centrale a unei distribuții. Ea se calculează ca sumă a tuturor valorilor observate ale seriei de date împărțită la numărul de observații:*

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

unde:

\bar{X} este media

x_i reprezintă valoarea variabilei pe care o ia observația i

N este numărul total de observații

Σ (sigma) este simbolul folosit pentru a indica o sumă

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

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

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

unde:

k este numărul de categorii (valori) ale variabilei

f_j reprezintă frecvența de apariție a categoriei j

x_j este valoarea categoriei j

N este numărul total de observații

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

$$\bar{X} = \frac{2 \times 3 + 4 \times 4 + 7 \times 5 + 10 \times 6 + 14 \times 7 + 23 \times 8 + 14 \times 9 + 6 \times 10}{80} = 7,31$$

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 $(1001 + 2000) / 2 = 1500,5$. Bineînțeles că, pentru un astfel de exemplu, la finalul calculului media se va rotunji, deoarece atunci când vorbim despre populația 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 și respectiv ultima categorie: în Tabelul 1.2, categoriile "1000 sau mai puțini locuitori", respectiv "peste 8000 de locuitori". Dacă se întâmplă ca valoarea minimă și 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 tendinței centrale?

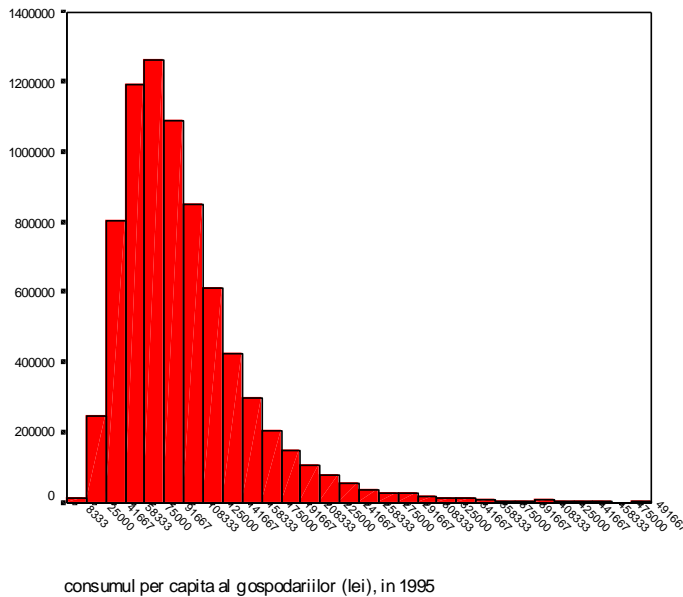
Decizia de a utiliza una sau alta dintre măsurile tendinței centrale este strâns legată în primul rând de nivelul de măsurare a variabilelor. Așa cum ne putem da seama, *modul poate fi utilizat pentru toate cele patru niveluri de măsurare. Mediana însă nu poate fi utilizată decât pentru nivelele care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

Un alt element important pentru a decide ce măsură a tendinței centrale merită folosită este existența observațiilor care au *valori extreme*. De fapt acest aspect este în strânsă legătură cu **forma distribuției**.

Să considerăm de exemplu distribuția consumului per capita al gospodăriilor, așa cum este ea reprezentată în Graficul 1.4. Media acestei distribuții este 103087 lei iar mediana este 87354 lei (valorile sunt exprimate în prețuri 1995). În ceea ce privește modul, valoarea exactă a acestuia nu are sens să fie calculată deoarece există relativ puține situații în care mai multe gospodării au

exact aceeași valoare a consumului per capita. Putem însă vorbi despre un interval modal, care se află undeva în jur de 72000 lei.

Graficul 1.4 Distribuția consumului per capita al gospodăriilor



Dacă dorim să aflăm valoarea "tipică" a consumului per capita într-o gospodărie pentru o distribuție 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 înțelegem noi în mod obișnuit prin "centrul distribuției": 50% dintre cazuri dedesubt și 50% deasupra. Mediana are avantajul de a nu fi influențată de valorile "extreme" ale seriei de date. Media seriei de date reprezentate în Graficul 1.4 este mai mare decât mediana tocmai datorită existenței 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 distribuții. El este util mai ales atunci când avem de-a face cu variabile măsurate la nivel nominal, dar și în cazurile în care distribuțiile studiate sunt bi- sau multi-modale. Mediana este indicată mai ales în cazurile în care dorim identificarea "valorilor tipice" ale unor distribuții *asimetrice* (vezi Graficul 1.5, b și 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 tendinței centrale. În plus ea mai are și alte proprietăți utile, care vor fi discutate în capitolele următoare.

2.2.2 Măsuri ale variației

Măsurile tendinței centrale sunt esențiale pentru descrierea unei caracteristici a unui eșantion sau a unei populații, însă ele nu sunt suficiente. Pentru descrierea completă a unei variabile este foarte important să știm deasemenea și cât de "împrăștiate" sunt valorile acesteia în jurul tendinței centrale sau, cu alte cuvinte, cât de omogenă respectiv eterogenă este populația (eșantionul) studiată în raport cu o anumită caracteristică. Să luăm ca exemplu performanța la o anumită materie a unei grupe de 80 studenți, măsurată cu note de la 1 la 10 (datele sunt prezentate în Tabelul 1.3). Nota medie a respectivei grupe este 7,31. Această informație însă pare a fi insuficientă pentru a ne putea pronunța asupra performanței respectivei grupe. Întrebarea pe care ne-o punem în mod natural este: cât de omogenă este respectiva grupă în ceea ce privește performanța școlară?

- Un prim răspuns la această întrebare îl putem da prin simpla examinare a intervalului în care sunt cuprinse notele respectivilor studenți, sau mai bine zis prin calcularea *amplitudinii* variabilei. *Amplitudinea unei variabile este diferența dintre valoarea maximă și valoarea minimă a acelei variabile.* Pentru exemplul nostru, amplitudinea este $10 - 3 = 7$ puncte. Deci, cei 80 de studenți sunt distribuiți de-a lungul unui interval de șapte puncte.

- O măsură a variației mai rafinată decât amplitudinea o reprezintă *abaterea interquartilă*, care se calculează ca diferență între quartila 3 și quartila 1. *Abaterea interquartilă măsoară împrăștierea celor 50% din observații aflate la mijlocul distribuției*. Ea are practic aceleași avantaje pe care le are și mediana ca măsură a tendinței centrale, și anume nu este influențată de existența cazurilor extreme.
- De cele mai multe ori suntem însă interesați să folosim o măsură a variației unei variabile care să includă toate observațiile, nu numai două dintre ele ca în cazul amplitudinii și abaterii interquartile. În plus, suntem interesați să examinăm variația în raport cu o măsură a tendinței centrale. De obicei, măsurile care satisfac aceste două cerințe sunt bazate pe *abaterile observațiilor de la medie*. *Abaterea de la medie a unei observații este diferența dintre valoarea pe care o ia respectiva observație și media variabilei* ($x_i - \bar{X}$). Una din proprietățile mediei este

însă aceea că suma tuturor abaterilor individuale de la medie este egală cu 0: $\sum_{i=1}^n (x_i - \bar{X}) = 0$

(sau cu alte cuvinte, abaterile pozitive se vor anula cu cele negative). În consecință, pentru a obține o măsură a variației la nivelul întregului eșantion sau a întregii populații 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:*

$$AMA = \frac{1}{N} \sum |x_i - \bar{X}|$$

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

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

Din motive teoretice care nu vor fi expuse în acest manual, pentru calcularea varianței la nivel de eșantion se folosește formula:

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

iar pentru date grupate în tabele de frecvențe (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 frecvența absolută de apariție a lui x_j

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

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

Din formula abaterii standard reiese clar că abaterea standard va fi cu atât mai mare cu cât valorile pe care le iau observațiile 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 și 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 performanței e mai mare.

În exemplul de mai sus am comparat două grupe de subiecți din punct de vedere al omogenității pentru o aceeași caracteristică. Însă atunci când trebuie analizăm omogenitatea unei singure populații sau a unui eșantion apar întrebări 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 populații pentru două variabile diferite?". Practic, răspunsul la prima întrebare depinde în mare măsură și de alte caracteristici ale distribuției. Pentru un anumit tip de distribuții interpretarea magnitudinii abaterii standard este mai ușoară, și 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 variație*, calculat ca raport între abaterea standard și media unei variabile:

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

Prin modul de calcul, coeficientul de variație are avantajul de a fi o măsură adimensională (fără unitate de măsură), deoarece unitatea de măsură a abaterii standard este aceeași cu cea a mediei. În consecință, el este foarte util în compararea variației a două variabile măsurate pe aceeași populație/eșantion. Putem astfel trage concluzii de tipul: "populația A este mai eterogenă în privința caracteristicii X decât în privința caracteristicii Y", concluzii imposibil de formulat numai cu ajutorul abaterii standard deoarece abaterea standard este o măsură dimensională și deci nu putem compara "mere cu pere" (de exemplu abaterea standard a performanței școlare cu abaterea standard a veniturilor familiei). Prin modul său de calcul coeficientul de variație indică practic cât la sută din medie corespunde unei abaterii standard, ceea ce face mai ușor de evaluat gradul de omogenitate a populației studiate. O populație cu o abatere standard egală sau mai mare decât media poate fi considerată în cele mai multe cazuri o populație eterogenă, în timp ce o populație a cărei abatere standard reprezintă 0,3 (30%) din medie poate fi considerată o populație relativ omogenă. Este important însă de reținut faptul că acest coeficient nu poate fi calculat decât în cazul variabilelor măsurate la nivel de raport, deoarece în cazul variabilelor nominale și ordinale abaterea standard nu poate fi calculată, iar în cazul variabilelor măsurate la nivel de interval media este una convențională, ceea ce face posibilă transformarea variabilei prin adunarea unei constante la valorile acesteia, fără ca semnificația 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 aceeași caracteristică am putea calcula coeficienți de variație diferiți ca valoare.

Exerciții și probleme

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

- Ce se poate spune despre simetria distribuției veniturilor?
- Care este venitul întregii țări (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 bivariante a datelor

Cuvinte cheie: intensitatea relațiilor dintre variabile, reducere proporțională a erorii, ranguri

Analiza bivariată a datelor.

Intensitatea relațiilor dintre variabilele calitative

În secțiunea anterioară am văzut cum putem testa ipoteza existenței unei relații (de asociere) între două variabile calitative. Testul χ^2 ne oferă însă informații numai despre existența/inexistența unei relații de asociere între două variabile, dar nu și despre intensitatea respectivei relații, atunci când ea există. Pentru a răspunde la întrebarea "Cât de puternică e relația de asociere dintre două variabile?" avem nevoie de măsuri specifice. Două dintre acestea vor fi prezentate în secțiunea care urmează.

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

Să ne întoarcem la datele din Tabelul 6.1 și să presupunem de această dată că nu cunoaștem decât distribuția marginală a atitudinii față de schimbarea modului de alocare a bugetului (cu alte cuvinte nu știm decât că 200 de indivizi sunt pentru, 220 sunt împotriva, iar 180 sunt nehotărâți). Dacă vom încerca să prezicem atitudinea unui individ oarecare, vom spune firește că respectivul individ va fi împotriva schimbării modului de alocare a bugetului, deoarece cu o astfel de predicție

avem cele mai reduse șanse de a greși. Cu alte cuvinte, ne-am bazat predicția pe frecvența modală (cea mai mare frecvență). În cazul în care am face o astfel de afirmație pentru fiecare din cei 600 de indivizi, predicția noastră ar fi corectă pentru 220 dintre ei (37%), și falsă pentru ceilalți 380. Să presupunem acum că la un moment dat primim o informație în plus, și anume distribuția atitudinilor față de schimbarea modului de alocare a bugetului în funcție de grupele de vârstă de care aparțin indivizii (adică exact informația prezentată în Tabelul 6.1). Să zicem că vom considera ca plauzibilă ipoteza în care atitudinile față 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 față de schimbarea modului de alocare a bugetului se va numi variabilă dependentă. Să zicem acum că vom repeta raționamentul de mai sus (predicția atitudinii unui individ pe baza frecvenței modale) pentru fiecare grupă de vârstă în parte. Vom avea deci, din nou, un număr de predicții corecte și un număr de predicții eronate. *Coeficientul λ reprezintă tocmai proporția cu care se reduce numărul de erori prin introducerea variabilei independente.* Să calculăm acum λ pentru datele din Tabelul 6.1:

Tabelul 6.1 Relația dintre două variabile categoriale

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

Așa cum am arătat, în absența variabilei independente, numărul de erori e_1 a fost 380. Să vedem acum câte erori am făcut prezicând variabila dependentă pe baza valorilor variabilei independente (pentru a ușura urmărirea 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 frecvenței modale în 110 cazuri, și vom face erori în 70 de cazuri.
- pentru grupa de vârstă 26 - 45 de ani, vom prezice corect pe baza frecvenței modale în 100 cazuri, și vom face erori în alte 100 de cazuri.
- pentru grupa de vârstă peste 45 de ani, vom prezice corect pe baza frecvenței modale în 90 cazuri, și vom face erori în 130 de cazuri.

Deci totalul erorilor făcute este $e_2 = 70 + 100 + 130 = 200$.

Să îl calculăm acum pe λ , după o formulă utilizată și pentru calculul altor măsuri ale asocierii și cunoscută sub numele de "*reducere proporțională a erorii*":

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

Coeficientul λ poate lua, prin modul de construcție numai valori între 0 și 1, 0 însemnând absența oricărei relații între variabile, adică independență, iar 1 însemnând intensitate maximă a asocierii (asociere puternică). El este o măsură asimetrică (avem o variabilă independentă pe baza căreia se fac predicții și o variabilă dependentă, ale cărei valori sunt prezise), însă există formule de calcul și pentru varianta simetrică a acestui coeficient. Avantajul lui constă în modul relativ ușor și intuitiv de calcul. Principalul dezavantaj al acestei măsuri este faptul că în condițiile în care o categorie a unei variabile conține 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, așa cum am văzut în introducerea acestui manual, există posibilitatea de ordonare a valorilor variabilelor, și în consecință există posibilitatea de a da *ranguri* indivizilor în funcție de valorile pe care aceștia 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 și 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 și pentru variabila Y, iar indivizii cu ranguri mici pe variabila X au de asemenea ranguri mici și pentru Y. asocierea negativă apare atunci când indivizii cu rang mare pentru variabila X tind să aibă ranguri mici pentru Y și 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 relație 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 noțiuni de bază care se referă la măsurile de asociere între variabile ordinale și la a arăta modul de calcul pentru o astfel de măsură.

O pereche de observații se numește *concordantă* dacă individul care are un rang mai înalt pe o variabilă are un rang mai înalt și pe a doua variabilă.

O pereche de observații se numește *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, ierarhizați 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 și perechile discordante, și pentru aceasta să începem cu toate perechile de observații 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) și perechea AD (concordantă). Să trecem acum la perechile lui B: Acestea sunt BC (concordantă) și 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 și 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 puțin mai complicat, pentru că deseori apar ceea ce se numesc ranguri "legate" sau egale. Acest lucru complică destul de mult calculele și formulele,

însă principiul rămâne același, al comparării numărului de perechi concordante cu numărul de perechi discordante.

Exerciții și probleme

1. Într-un studiu asupra modului în care ocupația se asociază cu educația, s-a realizat următorul eșantion aleator de 500 de bărbați anagajați.

Educația	Ocupația			
	<i>Funcționari</i>	<i>Muncitori în fabrică</i>	<i>Angajați în servicii</i>	<i>Agricultori</i>
4 sau mai mulți ani de liceu (incluzând și formarea vocațională)	194	146	27	10
Mai puțin de patru ani de liceu	18	79	18	8

- Explicitați în cuvinte ipoteza de nul H_0
- Calculați χ^2 și valoarea p pentru H_0

2. Se da tabelul:

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

Care din propozițiile următoare sunt adevărate?

- 30% dintre femeii sunt somere
 - 30% dintre someri sunt barbati
 - 70% din totalul populației se afla în somaj
 - probabilitatea ca o persoană din populație să fie în somaj este de 0.3
- nici una dintre acestea

3. Dacă variabilele nominale x și y nu sunt independente statistic atunci este de așteptat ca:

- Distribuțiile condiționate ale lui y funcție de x să fie diferite de distribuția marginală a lui y
- Distribuțiile condiționate ale lui y , funcție de x să fie egale între ele
- Corelația Bravais-Pearson dintre x și y să fie semnificativ diferită de 0
- Statistica test chi-patrat să difere semnificativ de 0
- Răspunsurile 1,2,3,4 să fie incorecte

Modulul 4

Obiectiv: prezentarea problematicii regresiei lineare în analiza datelor

Ghid de studiu:

- ◆ Regresia lineară simplă
- ◆ Construcția dreptei de regresie
- ◆ Regresia lineară multiplă
- ◆ Interpretarea coeficienților 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 determinație și coeficientul de corelație Pearson

Regresia lineară simplă

Fiind cunoscute valorile a două variabile cantitative pentru o mulțime de unități de analiză, este posibil să reprezentăm complet această informație 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 distanță de axa verticală proporțională cu valoarea variabilei independente luată de acea unitate, și la o distanță de axa orizontală proporțională 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, și în care șomajul a scăzut cu două procente, se află cu două unități deasupra axei orizontale și la cinci unități în dreapta axei verticale.

Foarte adesea, informația 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 informația reprezentată de distribuția sa la o singură valoare, cea a tendinței centrale, exprimată prin medie, mediană sau un alt indicator, și în cazul considerării simultane a două variabile ar fi de folos să putem descrie într-un mod cât mai succint relația dintre acestea.

O soluție 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 secțiunea care urmează, 7.1.1., vom arăta cum poate fi construită o astfel de dreaptă, numită *dreaptă de regresie*. Vom prezenta apoi interpretarea coeficienților prin care este descrisă dreapta de regresie. În secțiunea 7.1.2. vor fi definiți indicatori prin care poate fi apreciat gradul de acuratețe prin care o dreaptă de regresie descrie relația dintre două variabile. În ultima secțiune a acestei părți, 7.1.3., va fi discutate una dintre condițiile mai importante care trebuie îndeplinită pentru ca modelele de regresie să poată fi aplicate.

Construcția dreptei de regresie

Fie un grafic pe care sunt reprezentați mai mulți indivizi statistici, în funcție de valorile a două variabile cantitative, X și Y , și fie o dreaptă dusă la întâmplare pe acest grafic. Poziția fiecărui individ i este fixată de valorile pe care iau cele două variabile, notate cu x_i și y_i .

Poziția dreptei în raport cu cele două axe ale graficului este complet precizată de următoarea relație:

$$Y' = a + bX.$$

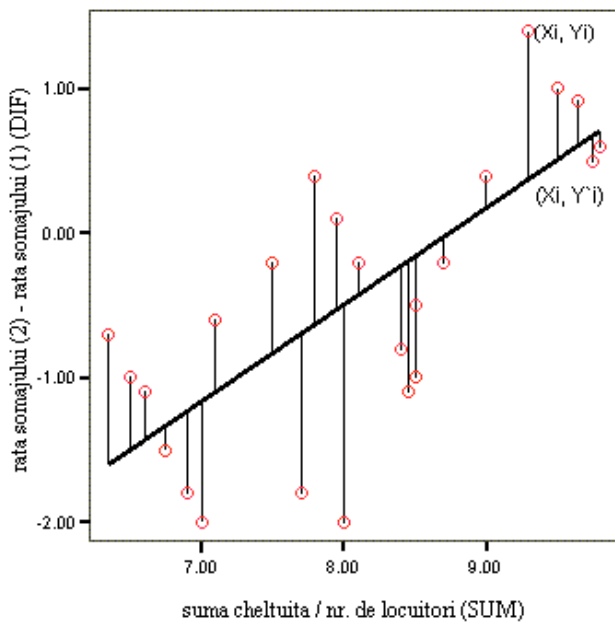
(1)

Relația exprimă faptul că orice punct k de pe dreaptă, are coordonatele x_k și y'_k astfel încât $y'_k = a + bx_k$. Mai mult, orice punct de pe grafic pentru care are loc relația 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 , și 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 mulțimii de puncte, atunci ar fi obținută o simplificare remarcabilă a modului în care este descrisă relația.

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

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



Dacă pentru două variabile cantitative am putea construi o dreaptă astfel încât toate punctele care corespund unităților 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 și dreaptă ar avea lungimea zero.

Este clar că în exemplul considerat aici nu există o astfel de dreaptă, care să descrie perfect relația dintre cele două variabile. Ar fi de dorit atunci, să fie determinată acea dreaptă pentru care lungimile segmentelor verticale dintre puncte și dreaptă să fie cât mai apropiate de zero.

Prin definiție, *dreapta cu proprietatea că pătratele lungimilor segmentelor dintre puncte și 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** și **Y**, care iau valori pentru **n** unități de analiză, pot fi determinate în mod unic constantele **a** și **b** astfel încât dreapta

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

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

Y' este o variabilă care se obține din intersecția segmentelor verticale care trec prin punctele (x_i, y_i) de pe grafic și dreapta de regresie, iar punctele de intersecție sunt de forma (x_i, 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, și 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:

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

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

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 și dreapta de regresie.

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

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

Interpretarea coeficienților dreptei de regresie

Coeficientul **b** este numit **panta asociată variabilei X** și, așa cum se poate vedea din expresia dreptei de regresie, *reprezintă numărul de unități cu care variază Y' atunci când X crește cu o unitate*:

dacă avem două puncte (x_1, y'_1) și (x_2, y'_2) , $x_2 = x_1 + 1$, și ambele puncte sunt pe dreapta

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

atunci, înlocuind în formula dreptei se obține

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

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

Semnul plus al lui b indică faptul că între X și Y are loc o relație 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 semnaleză prezența unei relații negative.

$b_1 = 0$ se obține atunci când forma norului de puncte nu poate fi aproximată printr-o dreaptă. O situație de acest gen apare atunci când cele două variabile estimează fenomene independente, fără legătură, dar și în cazul în care variabilele sunt într-o relație 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 relație.

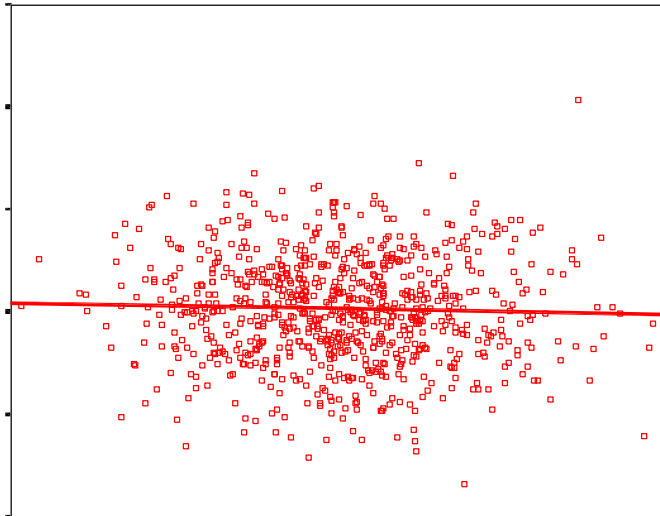
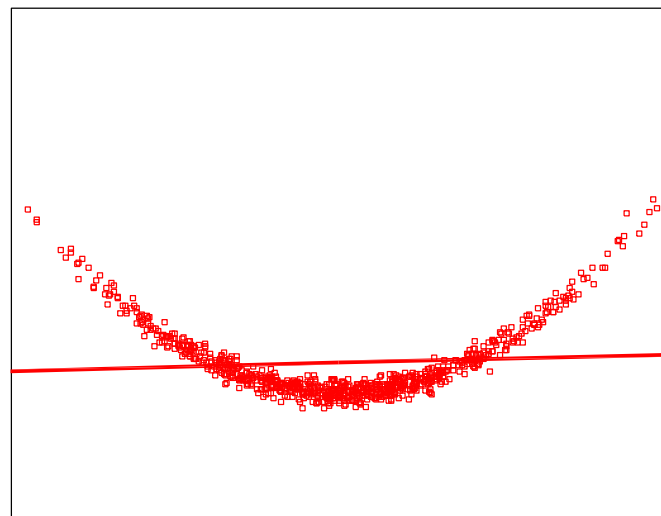


Figura 7.3. Exemplul a două variabile între care există o relație (de forma $Y = X^2$) care nu poate fi



aproximată printr-o dreaptă de regresie.

Coefficientul b are următoarea proprietate importantă: valoarea sa depinde de unitățile de măsură ale celor două variabile.

Astfel, dacă **SUM** din exemplul anterior ar fi exprimat în unități monetare / numărul de locuitori, adică într-o unitate de măsură de o mie de ori mai mică decât cea din exemplu, **b₁** 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 intensității relației dintre variabila dependentă și variabila independentă.

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

Indicatori ai intensității relației dintre două variabile cantitative: coeficientul de determinație și coeficientul de corelație Pearson

Dreapta de regresie asociată relației dintre două variabile cantitative oferă o imagine sintetică despre forma acestei relații, însă nu oferă informații despre cât de asemănătoare este această imaginea simplificată cu cea reală. Am întâlnit o situație similară în cazul mediei: acest indicator descrie succint tendința centrală a distribuției unei variabile cantitative, însă nu cuprinde informații 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 distribuția variabilei.

În Figura 7.4. și în Figura 7.5. sunt reprezentate relațiile dintre câte două perechi de variabile cantitative. În ambele cazuri ecuația dreptei de regresie este aceeași:

$$Y = 2 - 2,5 X.$$

Se observă însă că unitățile 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 relației dintre perechea de variabile cărora le corespunde.

Figura 7.4. Distribuția a două variabile cantitative și dreapta lor de regresie (A).

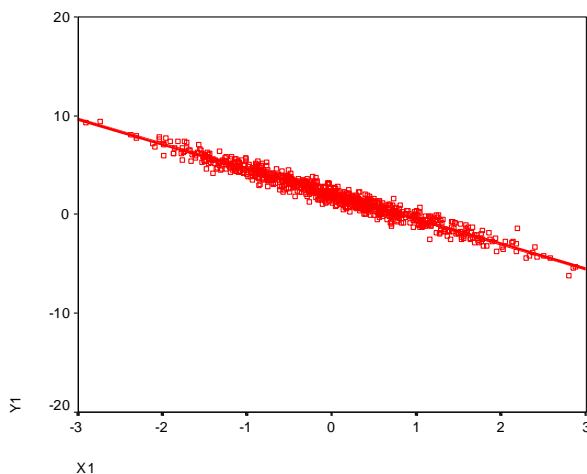
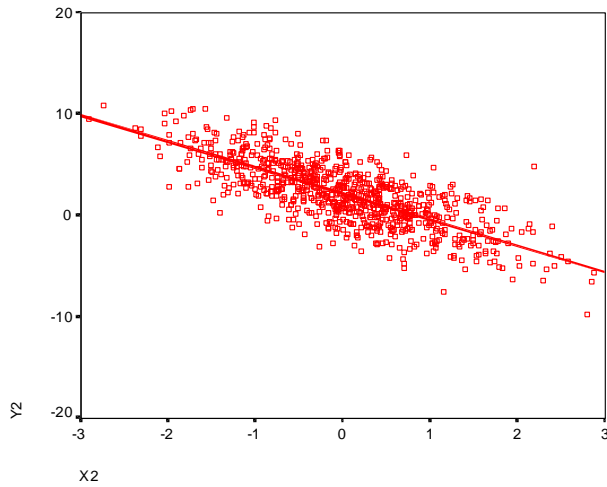


Figura 7.5. Distribuția a două variabile cantitative și dreapta lor de regresie (B).



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

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

Numărătorul expresiei reprezintă variația lui Y care este "explicată" de ecuația de regresie, în timp ce valoarea de la numitor este egală cu variația totală a lui Y . Deci, R^2 indică proporția din variația lui Y care este "explicată" de variabila independentă.

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

În exemplul din secțiunea anterioară $R^2 = 0,53$.

R^2 care corespunde relației reprezentate în Figura 7.4. are valoarea 0,95, în timp ce R^2 din Figura 7.5. are valoarea 0,58.

Un alt indicator al intensității relației dintre două variabile cantitative este **coeficientul de corelație Pearson**, notat cu r și definit prin următoarea formulă:

$$r = b \sigma_X / \sigma_Y.$$

σ_X și σ_Y reprezintă abaterea standard a variabilei X , respectiv abaterea standard a variabilei Y .

Coeficientul de corelație are două proprietăți din care poate fi dedus și modul său de interpretare:

1. $r^2 = R^2$ --coeficientul de corelație Pearson ridicat la pătrat este egal cu coeficientul de determinație.
2. r are același semn cu b , deoarece cele două abateri standard din definiția sa au întotdeauna semn pozitiv.

Astfel, din proprietatea (1) rezultă că r ia valori în intervalul $[-1, 1]$, iar valorile extreme sunt luate în același situații în care R^2 ia valoarea 1: atunci când relația dintre cele două variabile cantitative este de intensitate maximă și punctele care reprezintă grafic unitățile de analiză sunt distribuite pe o dreaptă. În mod similar, r ia valoarea 0 atunci când R^2 este nul, adică în situațiile în care distribuția unităților de analiză nu poate fi aproximată printr-o dreaptă (Figurile 2. și 3. ilustrează situații în care r este 0).

Din proprietatea (2) rezultă că r ia valori pozitive atunci când dreapta de regresie are o înclinație ascendentă de la stânga spre dreapta, și valori negative atunci când înclinația este descendentă.

Unitatea 2

Obiectiv: prezentarea problematicii regresiei lineare multiple

Cuvinte cheie: coeficienții de regresie standardizați, coeficient de determinație multiplă, multicoliniaritate, variabile "dummy".

Regresia lineară multiplă

Modelul de regresie simplă este folosit pentru a descrie relația dintre două variabile cantitative. În cazul în care sunt disponibile date despre mai mulți factori cu potențial explicativ, iar aceștia sunt estimați prin variabile cantitative, este de dorit ca analiza să cuprindă simultan toate variabilele și 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 situație de acest tip.

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

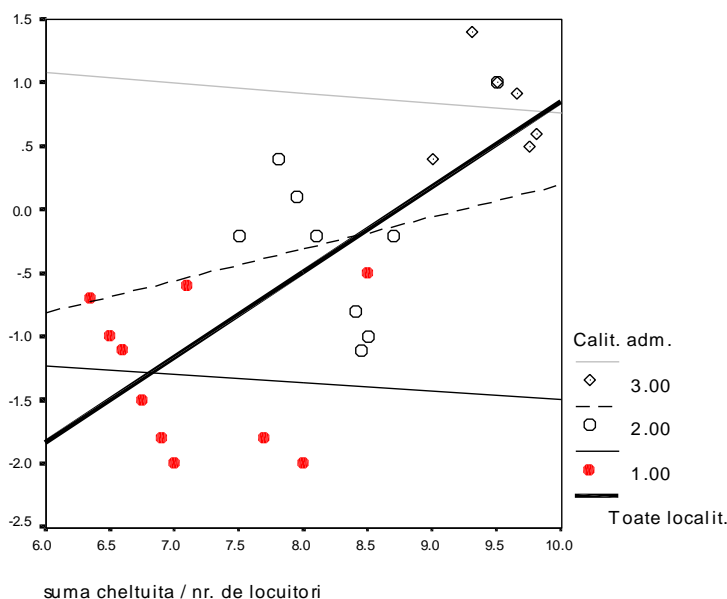
În Figura 7.7 sunt reprezentate localitățile cuprinse în studiu, în funcție de cele trei variabile. La fel ca și în Figura 7.8 valorile lui DIF sunt pe axa verticală, iar cele ale lui SUM pe axa orizontală. Marcarea localităților pe grafic se face prin simboluri grafice diferite în funcție de valorile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

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

Dacă, însă, relația dintre cele două variabile este studiată separat pe grupele de localități 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 variabilei CALIT. Valorile lui R^2 care se obțin sunt 0,006 pentru CALIT = 1, 0,005 pentru CALIT = 2, și 0,004 pentru CALIT = 3. Cele trei valori indică faptul că intensitatea relațiilor dintre DIF și SUM pentru fiecare dintre cele trei categorii de localități este foarte aproape de zero. Altfel spus, când sunt comparate localități care sunt asemănătoare din punctul de vedere al performanței administrării de programe, cheltuirea unei sume mari pe cap de locuitor nu este asociată, în medie, unei scăderi mai accentuate a ratei șomajului decât în localitățile în care suma a fost mai mică. Acest rezultat indică, contrar celui obținut din analiza doar a primelor două variabile, că programul de reducere a șomajului nu a fost eficient.

Figura 7.7. Relația dintre variabilele DIF, SUM, și CALIT pentru 25 de localități.



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, \dots sunt factori explicativi potențiali; Care este efectul independent al fiecărui A_i asupra lui **B**? Care este ierarhia importanței factorilor A_1, A_2, \dots 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 și de aceea cel mai ușor de interpretat sunt cele de regresie multiplă. Vom arăta modul în care acestea sunt definite (7.2.1), felul în care pot fi interpretate relațiile dintre variabilele cuprinse în model (7.2.2.) și cum poate fi evaluată eficiența de ansamblu a modelelor (7.2.3). În secțiunea (7.2.4.) vor fi discutate modalitățile de generalizare a rezultatelor obținute pe

un eșantion iar în secțiunea (7.2.5) va fi descrisă problema multicolarității. În secțiunea (7.2.6) va fi prezentată o extindere a modelelor de regresie pentru variabile nominale și ordinale.

Definirea modelelor de regresie multiplă

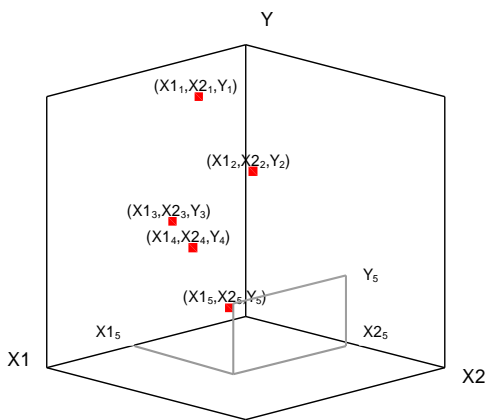
Fie Y , X_1 , X_2 , ..., X_m , variabile cantitative. Y este variabila a cărei variație încercăm să o explicăm iar X_1 , X_2 , ..., X_m , sunt variabilele independente. Putem scrie următoarea relație între variabile:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_mX_m + U \quad (3)$$

unde a , b_1 , b_2 , ..., b_m sunt numere iar U este o variabilă.

Se observă că pentru orice combinație de numere a , b_1 , b_2 , ..., b_m , relația (3) este adevărată, pentru că acestea împreună cu valorile lui Y , X_1 , X_2 , ..., X_m determină U .

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



Dacă $m=2$ relația (3) poate fi descrisă grafic printr-un desen tridimensional. Fiecărui individ statistic îi corespunde un punct de coordonate (X_1, X_2, Y) (Figura 7.8.), iar a , b_1 și b_2 definesc un plan descris de ecuația

$$Y' = a + b_1X_1 + b_2X_2.$$

Variabila U este determinată de acest plan și de punctele de forma (X_1, X_2, Y) într-un mod analog cazului cu două dimensiuni:

valoarea U_i care îi corespunde unui individ statistic care a luat valorile X_{1i} , X_{2i} , Y_i , este egală cu lungimea segmentului paralel cu axa OY care are la extremități punctul care îi corespunde în spațiu (X_{1i}, X_{2i}, Y_i) , respectiv punctul de intersecție cu planul (și care are coordonatele (X_{1i}, X_{2i}, Y'_i)).

Expresia (3) indică faptul că Y poate fi exprimată ca o combinație liniară de X_1 , X_2 , ..., X_m , și o variabilă U numită *variabilă reziduală*. Dacă fixăm a , b_1 , b_2 , ..., b_m atunci U poate fi exprimat în funcție de aceste numere și Y , X_1 , X_2 , ..., X_m :

$$U = Y - (a + b_1X_1 + b_2X_2 + \dots + b_mX_m) \quad (4)$$

Dacă notăm expresia din paranteza cu Y' atunci

$$U = Y - Y'.$$

Vom alege din mulțimea (infinită) a expresiilor de forma (3) acea combinație 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 șir 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 și în cazul bivariat, cel *al celor mai mici pătrate*:

Unei variabile U îi corespunde un număr u obținut din aplicarea formulei $u = u_1^2 + u_2^2 + \dots + u_n^2$, unde 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, b_1, b_2, \dots, b_m astfel încât u să fie minim. Este important de reținut că pentru orice număr de variabile independente $m, a, b_1, b_2, \dots, b_m$ sunt determinați în mod unic de condiția de a avea u minim.

Ecuția

$$Y' = a + b_1X_1 + b_2X_2 + \dots + b_mX_m, \quad (5)$$

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

Dacă avem o singură variabilă independentă ($m=1$) ecuația (5) descrie un model de regresie simplă.

În exemplul de mai sus, planul care aproximează cel mai bine - după criteriul celor mai mici pătrate -, distribuția punctelor din spațiu asociate celor 25 de unități de analiză (localități cuprinse în studiu) are ecuația

$$DIF = -2,61 + 0,05 \text{ SUM} + 1,00 \text{ CALIT.}$$

Interpretarea modelelor de regresie multiplă

Coeficientul b_i , numit panta asociată variabilei X_i , reprezintă numărul de unități cu care variază Y' atunci când X_i crește cu o unitate iar celelalte variabile independente sunt menținute constante. Altfel spus, b_i arată cum se modifică valoarea așteptată a variabilei dependente atunci când X_i variază iar X_j sunt constante, $j \neq i$.

Deasemenea, în măsura în care datele satisfac anumite proprietăți (dintre care un principiu important este cel al distribuțiilor apropiate de cele normale, iar un alt principiu, al multicolinearității, va fi discutat în secțiunea 7.2.5.) este corect să afirmăm că b_i indică variația în mediile valorilor lui Y care corespund punctelor de forma $(X_1, \dots, X_i, \dots, X_m)$ respectiv $(X_1, \dots, X_i + 1, \dots, X_m)$. La fel, a arată care este media lui Y atunci când $X_1 = X_2 = \dots = X_m = 0$.

Semnul plus al lui b_i indică faptul că între X_i și Y are loc o relație pozitivă în condiții de control al efectului celorlalte variabile, în timp ce semnul minus indică prezența unei relații negative.

În exemplul anterior, $b_1 = 0,05$ arată că dacă vom compara două localități unde diferența dintre sumele cheltuite în program / numărul de locuitori este 1000 lei, și care sunt identice din

perspectiva variabilei CALIT, ne așteptăm ca, în medie, rata șomajului să fi scăzut cu 0,05 puncte procentuale în localitatea în care s-a suma / locuitor a fost mai mare. $b_2 = 1,00$ arată că pentru aceeași valoare a lui SUM, localitățile cu o evaluare a calității administrării programelor mai bună cu o unitate au, în medie, o scădere mai mare cu o unitate procentuală. Această interpretare este consistentă cu rezultatul obținut după aplicarea de regresii simple pentru fiecare din subeșantioanele definite de CALIT, dar aduce un plus de precizie în exprimarea relațiilor dintre variabila dependentă și cele două variabile independente.

Din interpretarea coeficienților 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 excepția unei singure variabile. Diferența observată în valorile variabilei dependente este atribuită variației în variabila independentă care ia valori diferite pentru grupe diferite.

Asemeni pantei pentru cazul bivariat, coeficienții modelului de regresie multiplă depind de unitățile de măsură ale variabilelor și este adevărată proprietatea

$$\text{dacă în loc de } \mathbf{X}_i \text{ avem } c\mathbf{X}_i + d, \text{ atunci în loc de } b_i \text{ avem } b_i / c. \quad (6)$$

Această proprietate arată faptul că panta de regresie nu poate fi folosită drept un indicator al intensității relației dintre variabila dependentă și variabila independentă corespunzătoare, și nici nu permite ierarhizarea variabilelor independente în funcție de contribuția fiecăreia la explicația variației variabilei dependente.

Pentru a descrie nu doar forma ci și intensitatea relațiilor liniare între variabilele independente și variabila dependentă sunt folosiți **coeficienții de regresie standardizați**. Modul în care sunt definiți 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, obținute în urma aplicării formulei, au media egală cu zero iar abaterea standard egală cu unitatea. Coeficienții de regresie din modelul determinat de variabilele standardizate, se notează în mod obișnuit cu β_1 , β_2 , ..., β_m . Aplicând proprietatea (6) avem următoarea formulă pentru coeficienții standardizați:

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

În cazul bivariat ($m = 1$), β_1 coincide cu coeficientul de corelație Pearson. În cazul general, β_i arată cu câte abateri standard variază \mathbf{Y} atunci când \mathbf{X}_i crește cu o abatere standard. Din formulă reiese și faptul că β_i are același semn cu b_i , iar $\beta_i = 0$ este echivalent cu $b_i = 0$.

Atunci când $m > 1$ coeficienții de regresie standardizați au câteva proprietăți diferite față de cazul bivariat:

1. β_i poate să ia valori și în afara intervalului $[-1, 1]$. β_i în valoare absolută este supraunitar atunci când relația dintre \mathbf{X}_i și \mathbf{Y} este foarte intensă și în plus, există o relație liniară strânsă între \mathbf{X}_i și cel puțin una dintre celelalte variabile independente.
2. în timp ce în cazul bivariat $U = 0$ implică faptul că $\beta_1 = \pm 1$, atunci când $m > 1$, condiția $U = 0$ nu restrânge valorile posibile pentru β_i .

Coeficienții de regresie standardizați permit ierarhizarea variabilelor independente în funcție de importanța pe care o are fiecare în explicarea variației variabilei dependente printr-o relație directă. Sunt necesare două precizări privind limitele în utilizarea acestor coeficienți:

1. Dacă într-un model teoretic în care $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$ sunt variabile independente pentru \mathbf{Y} , \mathbf{X}_i este o variabilă explicativă și pentru unul sau mai mulți $\mathbf{X}_j, j \neq i$, atunci modelul de regresie în care \mathbf{Y} este variabila dependentă ne permite numai estimarea efectului direct pe care îl are \mathbf{X}_i asupra lui \mathbf{Y} , nu și a celui mediat de alte variabile din model. De exemplu, variabila care exprimă proporția celor care au absolvit liceul din populația unei localități poate avea un efect direct nul asupra numărului de infracțiuni pe cap de locuitor ($\mathbf{b}_{\text{liceu}} = \mathbf{0}$), însă un efect indirect substanțial, 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 aceleași variabile, dar care descriu date diferite, coeficienții standardizați nu sunt comparabili între modele decât dacă variabilele care le corespund au dispersii asemănătoare. De aceea, este de preferat ca în comparațiile dintre populații diferite să fie folosiți coeficienții nestandardizați, după ce în prealabil datele au fost transformate astfel încât variabilele \mathbf{X}_i să aibă aceeași unitate de măsură în ambele modele.

Eficiența unui model de regresie multiplă

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

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

Numărătorul expresiei reprezintă variația lui \mathbf{Y} care este "explicată" de ecuația de regresie, în timp ce valoarea de la numitor este egală cu variația totală a lui \mathbf{Y} . Deci, \mathbf{R}^2 indică proporția din variația lui \mathbf{Y} care este "explicată" de toate variabilele independente din model. Din definiție rezultă că \mathbf{R}^2 poate să ia valori între 0 și 1.

În exemplul din această secțiune 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 explicația este în mai mare măsură completă. Totuși, această afirmație necesită câteva precizări:

\mathbf{R}^2 are proprietatea de a crește cu fiecare variabilă care este introdusă în model și de aceea valorile lui \mathbf{R}^2 trebuie judecate și 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 informația de natură statistică (chiar dacă există procedee complexe prin care putem îmbogăți această informație). Numai prin luarea în considerare și a unor aspecte de natură teoretică poate fi decisă includerea sau eliminarea unor variabile în analiză.

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

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

La fel ca și în cazul bivariat, $R^2 = 0$ nu indică în mod necesar absența unor relații între variabilele independente și variabila dependentă:

1. Y poate să fie determinată complet de variabilele independente prin relații neliniare iar R^2 să fie egal cu zero. De exemplu, dacă

$$Y = \sqrt{10 - X_1^2 - X_2^2},$$

se obține distribuția din Figura 7.9. în care toate punctele sunt pe o suprafață curbă (cele mai multe sunt pe o emisferă), iar $R^2 = 0$.

2. Mai mult, este posibil să avem $R^2 = 0$ chiar și atunci când între una dintre variabilele independente și Y există o relație liniară, în condiții de control, însă forma (panta) acestei relații nu este constantă pe categoriile celorlalte variabile.

Figura 7.9. Exemplul unei distribuții în spațiu pentru care variabila dependentă este reprezentată pe axa verticală și $R^2 = 0$.

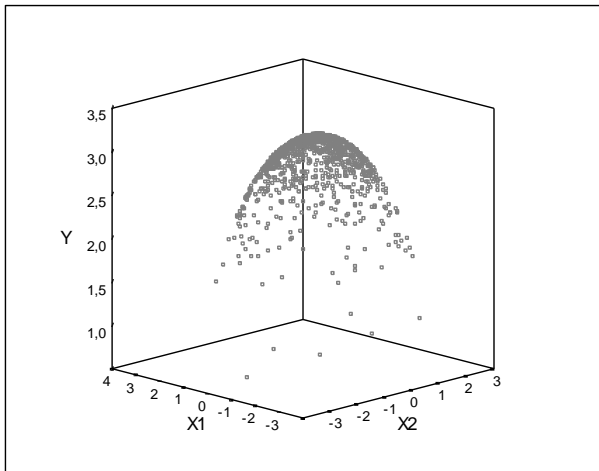
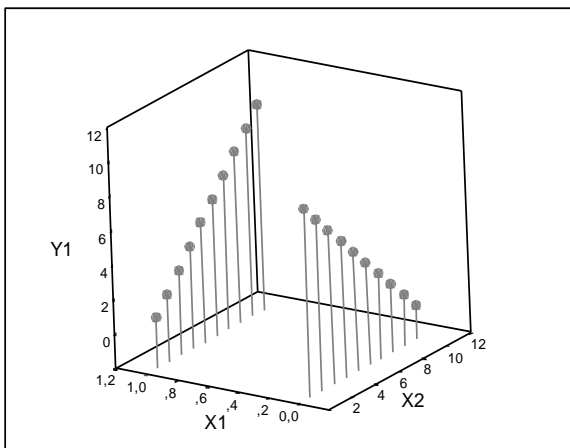


Figura 7.10. Exemplul unei distribuții în spațiu pentru care variabila dependentă este reprezentată pe axa verticală, au loc relații bivariate cu variabila dependentă de intensitate maximă și $R^2 = 0$ pentru modelul multivariat.



Generalizarea rezultatelor obținute pe eșantion (inferența)

Atunci când avem la dispoziție date dintr-un eșantion probabilistic și construim pe acestea un model de regresie multiplă ne punem problema de a generaliza rezultatele obținute pentru populația din care provine eșantionul. Să presupunem că am obținut $b_i > 0$ și dorim să știm dacă panta corespunzătoare variabilei X_i este pozitivă și la nivelul întregii populații. Pentru a afla acest lucru vom construi un interval de încredere în jurul valorii lui b_i . Determinarea intervalului de încredere presupune îndeplinirea de către datele analizate a unor proprietăți, aceleași cu condițiile specifice cazului bivariat, la care se adaugă condiția de *absență a multicolarității* (pe care o vom defini și analiza în secțiunea 6). În continuare, presupunem îndeplinite toate aceste proprietăți. Pentru un nivel de încredere de 0,95 avem intervalul

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

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

Dacă intervalul de încredere nu îl conține pe zero atunci *ipoteza de nul*, adică afirmația conform căreia între X_i și Y nu avem o relație 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 populație - este egal cu 0).

O altă modalitate prin care poate fi realizată generalizarea valorilor b_i constă în determinarea valorii maxime a nivelului de semnificație statistică pentru care 0 aparține intervalului de încredere. Dacă nivelul de semnificație 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 b_i

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

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

Pentru exemplul din această secțiune, tabelul următor conține coeficienții de regresie, coeficienții standardizați, erorile standard, valorile t și nivelele de semnificație statistică.

	Coeficienti ne-standardizati		Coeficienti standardizati	t	Nivel de semnif. stat.
	B	Eroare standard	Beta		
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 atât în explicație cât și în predicție. Astfel, din modelul anterior rezultă că, în medie, localitățile în care ar fi aplicat un program similar cu cel

studiat iar suma cheltuită ar fi de 8000 de lei / locuitor, și care ar fi evaluate prin scorul 3 pentru calitatea administrării programelor, vor avea *în medie* o scădere a ratei șomajului cu o valoare dată de următoarea formulă

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

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

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

unde **DIF₀** este valoarea medie "prezisă" de model, **σ_{DIF}** este **eroarea standard** a valorii estimate **DIF₀**.

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

Problema multicolarității

Situația în care o variabilă independentă poate fi exprimată ca o combinație liniară perfectă a celorlalte variabile independente, este numită **multicolaritate perfectă**:

De exemplu, dacă variabilele independente sunt **X₁**, **X₂**, și **X₃**, iar

X₂ = 3X₁ + 2X₃, se spune că **X₂** este exprimat printr-o combinație liniară a variabilelor **X₁** și **X₃**, iar variabilele **X₁**, **X₂** și **X₃** sunt într-o relație de multicolaritate perfectă.

Atunci când variabilele independente sunt într-o situație de multicolaritate perfectă coeficienții de regresie nu pot fi determinați, și analiza de regresie nu poate fi aplicată.

În practică, o situație de acest tip este rar întâlnită și este ușor de detectat. În schimb, sunt mai frecvente cazurile de **multicolaritate ridicată**, în care o variabilă independentă poate fi exprimată aproape perfect printr-o combinație liniară a celorlalte variabile independente. Când se întâmplă acest lucru, coeficienții pot fi determinați în mod unic însă sunt instabili: valoarea pantei unui anumit coeficient diferă foarte mult de la un eșantion la altul pentru o anumită populație. Din acest motiv, atât comparațiile între valorile coeficienților dintr-un model cât și comparațiile pentru aceeași coeficienți ai unor modele pe eșantioane diferite sunt nesigure.

Care este pragul peste care multicolaritatea este considerată a fi ridicată și poate să ridice probleme în interpretarea modelului? O metodă frecvent folosită constă în realizarea de regresii în care, pe rând, fiecare dintre **X_i** este variabilă dependentă iar ceilalți **X_j** sunt variabile independente. Valoarea cea mai ridicată pentru un **R²** obținut în acest fel este o măsură a nivelului de multicolaritate din model, iar limita convențională sub care se consideră că multicolaritatea nu afectează interpretabilitatea modelului este **0,8**.

Atunci când este întâlnită o situație de multicolaritate ridicată sunt mai multe moduri prin care pot fi atenuate efectele ei:

1. Este mărit volumul eșantionului astfel încât ipoteza de nul să poată fi respinsă pentru o parte dintre coeficienți.
2. Variabilele care sunt puternic corelate sunt combinate în indicatori unici. De exemplu, într-o analiză în care secțiile de vot sunt unități statistice, rata de participare în primul tur de scrutin al alegerilor din 1996 este o variabilă independentă și rata de participare în al doilea tur de scrutin al

alegerilor din 1996 este o altă variabilă independentă, coeficientul de corelație între cele două variabile este $r = 0,91$. Un model realizat pe un eșantion ales dintre secțiunile de vot și în care cele două variabile sunt independente va fi afectat de o problemă de multicolaritate ridicată. O soluție 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 multicolaritate. 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 excepție importantă: toate proprietățile pe care le au valorile estimate ale unui model de regresie se păstrează și în cazul în care una sau mai multe dintre variabilele independente sunt **dihotomice** (adică variabile care iau două valori).

Consecințele acestei proprietăți sunt importante deoarece permit nu doar estimarea efectelor unor variabile care în mod obișnuit sunt dihotomice (exemplu *sexul, mediul de rezidență* -- urban / rural, etc.) asupra variabilei dependente ci și 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 șomajului din exemplul discutat în această secțiune cuprind informații despre încă o variabilă independentă: județul în care se află localitatea (JUDET). Mai presupunem că localitățile din studiu provin din trei județe, notate cu A, B, C. Variabila JUDET este transformată în două variabile dihotomice: JUDET1 și JUDET2. JUDET1 este definită astfel: localitățile care sunt în județul A au valoarea 1, iar toate celelalte au valoarea 0.

JUDET2 este definită asemănător: localitățile care sunt în județul B au valoarea 1, iar toate celelalte au valoarea 0.

Cunoscând valorile celor două variabile pentru o localitate, știm sigur în ce județ se află aceasta, deci informația oferită de JUDET1 și JUDET2 este egală cu cea oferită de variabila inițială, JUDET.

Prin introducerea celor două variabile dihotomice în analiză putem verifica dacă scăderea șomajului a fost influențată și de factori care s-au manifestat la nivel de județ, independenți de condițiile de nivel local. Dacă coeficientul b , respectiv **beta**, care corespunde uneia dintre aceste variabile este diferit de 0, atunci rezultă că variația variabilei DIF poate fi explicată mai bine în urma includerii ei.

Exerciții și probleme

1. Presupunem cunoscute pentru mai multe localități următoarele două variabile: diferența între venitul pe cap de locuitor în ultimul an și cel din anul anterior (VENIT)--măsurat în mii lei--, și proporția 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 ecuația de regresie este
VENIT = -7,5 - 12 SCOALA.

Care dintre următoarele afirmații este adevărată?

- a. Localitățile în care SCOALA are valori mici au, în medie, valori mai mici ale variabilei VENIT.
- b. Localitățile în care sunt 20% locuitori care nu au absolvit 10 clase au, în medie, diferența între veniturile anuale pe cap de locuitor (VENIT) cu 12 mii de lei mai mare decât localitățile în care sunt 30% locuitori care nu au absolvit 10 clase.

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

$$\text{VENIT} = 12 - 4 \text{ SCOALA} - 0,2 \text{ MINORI.}$$

$$\beta_{\text{SCOALA}} = -0,08, \text{ iar } \beta_{\text{MINORI}} = -0,25.$$

Care dintre următoarele afirmații este adevărată?

- a. Variabila MINORI explică mai puțin din variația variabilei dependente decât variabila SCOALA.
- b. Conform modelului de regresie, localitățile 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 ecuația de regresie cu coeficienți nestandardizati (în paranteză sunt prezentate erorile standard ale coeficienților de regresie:

$$Y' = -40 + 0.72x_1 + 1.29x_2 - 0.15x_3$$

$$\text{ES} \quad (0.13) \quad (0.37) \quad (0.16)$$

$$R=0.9$$

Care din coeficienții ecuației de regresie difera semnificativ de zero pentru un nivel de semnificație $p=0.05$?

1. cel al lui x_1
2. cel al lui x_2
3. cel al lui x_3
4. niciunul