AURELIAN MUNTEAN ANDREEA STANCEA

DATE CANTITATIVE ȘI CALITATIVE

INTRODUCERE ÎN METODE ȘI TEHNICI DE ANALIZĂ A DATELOR, CU APLICAȚII ÎN R, STATA, EXCEL ȘI NVIVO

PRESA UNIVERSITARĂ CLUJEANĂ

Aurelian Muntean **Andreea Stancea**

•

DATE CANTITATIVE ŞI CALITATIVE

Introducere în metode şi tehnici de analiză a datelor,
cu aplicaţii în R, STATA, EXCEL şi NVIVO

Aurelian Muntean este lector universitar la Facultatea de Științe Politice din cadrul SNSPA București, unde predă cursuri de sociologie și științe politice. Este directorul masteratului Labour Studies, singurul program cu această specializare din România și singurul din Estul Europei membru în consorțiul European Master in Labour Studies Network. A publicat în reviste precum Electoral Studies, European Journal of Industrial Relations, Europe-Asia Studies, Government and Opposition, American Journal of Economics and Sociology, Romanian Journal of Society and Politics, Social Change Review, Journal for the Study of Religion and Ideologies, dar și în diverse volume colective din țară și străinătate. A coordonat cercetări cantitative și calitative la nivel regional sau național în România, pe teme precum clientelism electoral, partide politice, participare civică, relații de muncă și dialog social, inegalități sociale, politici de sănătate, relația biserică-stat.

Andreea Stancea este doctorandă în sociologie la SNSPA București cu o teză care analizează impactul crizelor asupra procesului de luare a deciziilor. Interesele ei de cercetare sunt pe subiecte cum ar fi comportament electoral, comunicare, responsabilitate politică și personalizarea campaniei. A publicat în American Journal of Economics and Sociology.

Aurelian Muntean **Andreea Stancea**

DATE CANTITATIVE ŞI CALITATIVE

Introducere în metode şi tehnici de analiză a datelor,
cu aplicaţii în R, STATA, EXCEL şi NVIVO

PRESA UNIVERSITARĂ CLUJEANĂ
2022

Referenți științifici:

Cercetător științific 1 Bogdan Voicu, *Academia Română*

Prof. univ. dr. Adrian Dușa, *Universitatea din București*

Conf. univ. dr. Andrei Gheorghiță, *Universitatea „Lucian Blaga" din Sibiu*

Fotografia pentru copertă: **© Aurelian Muntean**

ISBN 978-606-37-1721-5

Universitatea Babeş-Bolyai
Presa Universitară Clujeană
Director: Codruța Săcelean
Str. Hasdeu nr. 51
400371 Cluj-Napoca, România
Tel./fax: (+40)-264-597.401
E-mail: editura@editura.ubbcluj.ro
http://www.editura.ubbcluj.ro/

CUPRINS

LISTĂ DE TABELE ȘI GRAFICE

Introducere

Cartea de față se adresează studenților la sociologie, științe politice, comunicare, relații internaționale, administrație publică, comunicare, jurnalism și management, precum și utilizatorilor sau consumatorilor de date care preferă o introducere prietenoasă în analiza datelor, cu un limbaj cât mai simplu. Cartea prezintă fundamente teoretice, centrându-se pe discuția despre utilizarea metodelor statistice. Oferim relativ puține detalii matematice, pentru a căror argumentare vom face trimiteri la alte manuale de statistică, preponderent în limba română, pentru a facilita lectura directă a acestora.

Date noi sunt generate în fiecare secundă. Fie că utilizăm biblioteci online pentru documentare, fie că realizăm tranzacții financiare online sau doar comandăm mâncare, internetul colectează toate aceste date pe care ulterior le utilizează în analize complexe de generare a strategiilor de marketing, a campaniilor electorale sau produselor bancare. În ultima decadă, utilizarea rețelelor sociale, a cumpărăturilor online și a serviciilor de video *streaming* a crescut și mai mult cantitatea de date generate. Toate datele colectate parcurg un complex ciclu de procesare care are rolul de a transforma datele în informații utilizabile.

Ținând cont de complexitatea procesării datelor este important să înțelegem ce sunt datele. De la inventarea calculatoarelor, oamenii au folosit termenul de date pentru a se referi la informații computerizate transmise sau stocate. De asemenea, datele pot fi texte sau cifre scrise pe hârtie, octeți și biți din memoria dispozitivelor electronice sau pot fi fapte care sunt stocate în mintea unei persoane. Pentru ca datele să poată oferi informații utile este necesar ca acestea să fie colectate într-un format organizat, de regulă sub forma unei baze de date. Baza de date este o colecție organizată de informații structurate și stocate electronic. Programele statistice și de analiză statistică sunt instrumente folosite de analiști și cercetători pentru a colecta și analiza date cu scopul de a oferi informații bazate pe știință asupra tiparelor și tendințelor realității exterioare.

Acum douăzeci sau treizeci de ani oferta de programe de analiză statistică era dominată de programe oferite contra-cost, precum SAS, SPSS, Stata, Matlab sau Statistica. Dezvoltarea în prima jumătate a anilor 1990 a unor programe și limbaje de programare, *open source* și accesibile în mod gratuite, specializate în analiza datelor, cum ar fi R și Python, a condus la diversificarea opțiunii de învățare a utilizării și programării unui program specializat. Pe baza acestor din urmă limbaje de programare s-au dezvoltat în ultimii ani multe programe sau interfețe grafice de analiză, bazate pe limbajul R, precum Deducer, R Commander, BlueSky Statistics, RStudio, R-Instat, Rattle, JASP, jamovi; sau bazate pe limbajul Python, precum PyQt5, Tkinter, PyGUI, Kivy sau PySimple. Printre cele mai solicitate pe piața muncii (în domenii precum cercetare sau analiză de date), dar și cunoscute programe de statistică se numără: R, Python, Tableau, SAS, Stata, SPSS, Excel, Matlab.

Datele în forma lor brută pot conține informații utile oricărei organizații. Aceasta trebuie să aibă capacitatea de a le prelucra pentru extrage mai multe informații utile activității sale. Prelucrarea datelor este metoda de colectare a datelor brute și de traducere a acestora în informații utilizabile. De obicei, prelucrarea datelor se realizează de către o echipă de cercetători și ingineri de date. Datele brute sunt colectate, filtrate, sortate, procesate, analizate, stocate și apoi prezentate într-un format care poate fi înțeles de publicul țintă. Astfel, capacitatea de a procesa date este esențială pentru organizațiile care vor să elaboreze strategii mai bune, să își crească avantajul competitiv sau să ofere servicii de calitate.

Este important să înțelegem ce presupune procesarea datelor și care sunt pașii pe care trebuie să-i parcurgem. Procesarea datelor se bazează pe o serie de pași ciclici în care datele brute (în engleză *input*) sunt introduse într-un sistem pentru a produce informații și rezultate (în engleză *output*). Fiecare pas este realizat într-o anumită ordine, dar întregul proces se repetă într-o manieră ciclică. Ieșirea primului ciclu de procesare a datelor poate fi stocată și alimentată ca intrare pentru următorul ciclu. În general, există șase etape principale în ciclul de prelucrare a datelor:

Pasul 1: Colectarea datelor. Colectarea datelor brute este primul pas al ciclului de prelucrare a datelor. Tipul de date brute colectate are un impact uriaș asupra rezultatului final. Prin urmare, datele brute ar trebui adunate din surse definite și

precise, astfel încât constatările ulterioare să fie valide şi utilizabile. Datele brute pot include cifre monetare, cookie-uri pentru site-uri web, declaraţii de profit/pierdere ale unei companii, comportamentul utilizatorului sau rata de participare la vot în funcţie de judeţ.

Pasul 2: Pregătirea datelor. Pregătirea sau curăţarea datelor este procesul de sortare şi filtrare a datelor brute pentru a elimina datele inutile şi inexacte. Datele brute sunt verificate pentru erori, calcule greşite sau date lipsă şi transformate într-o formă adecvată pentru analiza şi procesarea ulterioară. Scopul acestui pas este de a elimina datele redundante, astfel încât datele de înaltă calitate să poată oferi rezultate valide.

Pasul 3: Introducerea datelor. La acest pas, datele brute sunt convertite într-o formă care poate fi citită de programul statistic şi introduse în unitatea de procesare.

Pasul 4: Prelucrarea datelor. La acest pas, datele brute sunt supuse diferitelor metode de procesare folosind algoritmi de învăţare automată şi inteligenţă artificială pentru a genera rezultatul dorit. Acest pas poate varia uşor de la proces la proces, în funcţie de sursa datelor care sunt procesate (baze de date online, dispozitive conectate etc.) şi de scopul final al rezultatelor.

Pasul 5: Rezultate. Datele sunt în cele din urmă transmise şi afişate utilizatorului într-o formă lizibilă, cum ar fi grafice, tabele, fişiere vectoriale, audio, video, documente etc. Această ieşire poate fi stocată şi procesată în continuare în următorul ciclu de procesare a datelor.

Pasul 6: Depozitarea datelor. Ultimul pas al ciclului de prelucrare a datelor este stocarea, unde datele sunt stocate pentru utilizare ulterioară. Acest lucru permite accesul şi regăsirea rapidă a informaţiilor ori de câte ori este nevoie şi, de asemenea, permite ca acestea să fie utilizate direct ca intrare în următorul ciclu de procesare a datelor. În aparență complex și anevoios, procesul de deprindere a abilităților pentru analiza datelor este un proces cu dificultate graduală, similar jocurilor video. Mai mult, este un proces care poate fi învățat aplicând trei concepte de bază: încercare, eroare și repetare.

În consecință, deprinderea abilităților de a lucra cu seturi mari de date și a obține informații valide în urma analizei acestor date este necesară pentru a rămâne conectați la realitatea exterioară.

Partea practică a acestei cărți își propune să ofere informațiile de bază necesare pentru cei interesați să folosească instrumente statistice în activitatea lor; de exemplu, studenți care au de redactat o lucrare academică, cercetători implicați într-un proiect de cercetare, angajați într-o instituție publică sau privată. Pentru toate aceste categorii, utilizarea datelor într-un format organizat este extrem de importantă.

Cartea este structurată după cum urmează. Capitolul 1 prezintă o serie de probleme importante pentru a înțelege rolul cercetării științifice, standardele metodologice de care ținem cont atunci când realizăm un studiu științific, dar și reguli de etică în cercetare. Capitolul 2 prezintă principalele tipuri de design de cercetare, utile atât pentru studenți și cercetători, cât și pentru practicieni din instituții publice sau private. Capitolul 3 prezintă metode de analiză cantitativă. Discutăm despre surse de obținere a datelor, modalități de citire, transformare și manipulare a datelor pentru a fi ulterior utilizate în analize statistice complexe. Prezentăm o serie de modele statistice, exemplificând pe câteva modele de regresie utile pentru cele mai des întâlnite tipuri de date și probleme de analizat. Pentru fiecare metodă exemplificam și arătăm în detaliu, parcurgând etapele științifice ale cercetării, cum se face analiza în R (capitolul 3 și 6), Stata (capitolul 4) sau Excel (capitolul 5). Ultimul capitol este dedicat metodelor de analiză calitativă. În el, vom discuta despre instrumente specifice acestui tip de analiză, cum ar fi interviul aprofundat, focus grupul, analiza de conținut și analiza de discurs. Exemplele oferite în acest capitol de analiză a datelor calitative sunt ilustrate folosind programul NVivo.

Programele de analiză utilizate în această carte sunt cele prezentate mai jos.

R este un limbaj de programare și un program gratuit, folosit pentru analiza datelor. Este administrat de Fundația R (R Core Team 2022b). Se bazează pe programarea liniilor de cod, dar este folosit și pentru dezvoltarea unor programe de analiză statistică ce nu necesită cunoașterea limbajului de programare R. A fost lansat în 1993, iar în 2022 a ajuns la versiunea 4.2.2.

RStudio este o interfață pentru limbajul de programare R, dezvoltată de compania Posit Software, PBC, (fostă RStudio, PBC) (RStudio Team 2022). Programul este oferit atât în versiune desktop gratuită, cât și în versiune contra cost, în cloud, pentru companii.

Stata este un program de analiză a datelor deținut de compania StataCorp LLC (StataCorp 2019). Poate fi folosit prin limbaj de programare, dar oferă și posibilitatea analizelor statistice din meniu predefinit. A fost lansat în 1985, iar în anul 2022 a ajuns la versiunea 17.

Excel este un program de calcul tabelar, care poate fi folosit și pentru anumite analize statistice sau pentru managementul datelor. Este folosit pentru analize și input de date prin meniu predefinit, dar permite analize statistice și prin intermediul formulelor e programare. Este deținut de compania Microsoft, și este oferit în suita Office (Microsoft Corporation 2021). A fost lansat în 1982, iar pentru desktop a ajuns in anul 2022 la versiunea 16.

NVivo este un program de analiză și gestiune a datelor calitative. Permite analizarea acestora prin intermediul meniului predefinit. Este deținut de compania QSR International (2020). A fost lansat în anul 1999, iar în anul 2022 a ajuns la versiunea 12.

Toate aplicațiile practice reprezentate în această carte se bazează pe datele colectate în cadrul European Social Survey 2020 (ESS)[1] (Ancheta Socială Europeană), dar și pe unele baze de date oferite în mod gratuit de asociația care gestionează programul **R**. Datele ESS pot fi descărcate accesând linkul https://ess-search.nsd.no/.

[1] ESS este un sondaj trans-național administrat în toată Europa încă din anul 2001. Sondajele sunt aplicate la fiecare doi ani, față în față, cu eșantioane transversale nou selectate. Sondajul măsoară atitudinile, credințele și modelele de comportament ale diverselor populații din peste treizeci de state, majoritatea fiind membre ale Uniunii Europene.

1. Fundamentele cercetării științifice

În acest capitol vom discuta despre o serie de probleme fundamentale pentru înțelegerea și utilizarea metodelor de analiză a datelor cantitative şi calitative. Aceste competențe sunt importante pentru producerea cunoașterii ştiinţifice. Vom prezenta standardele cercetării ştiinţifice bazată pe date empirice, vom discuta despre modele de cercetare empirică, dar şi despre constructele teoretice fundamentale pentru înțelegerea şi utilizarea tehnicilor şi metodelor de cercetare empirică.

Rezultatele analizelor empirice obținute prin studii științifice, deşi par a produce dovezi irefutabile, rămân însă incerte, din motive care ţin de cantitatea şi calitatea datelor, a instrumentelor folosite, sau de designul cercetării. Incertitudinea poate conduce, astfel, la neîncrederea în studiile ştiinţifice (Rosenberg et al. 2022). Incertitudinea nu este întâlnită doar în cercetările științifice, ci este un fenomen întâlnit frecvent în viața de zi cu zi. Astfel, cu toate că dorim eliminarea incertitudinii, aceasta trebuie acceptată ca fiind un fenomen normal care trebuie înțeles. Dacă traversăm strada, există probabilitatea de a fi accidentaţi. Atunci când cumpărăm o shaorma de la magazinul din faţa locului nostru de muncă există probabilitatea de a face toxiinfecţie alimentară. Traversarea străzii poate duce la accidentare atunci când, deşi încercăm să controlăm ceea ce ţine de noi (traversăm prin loc permis, de exemplu doar pe culoarea verde şi pe trecerea de pietoni), există variabile care nu depind de noi, dar pe care încercăm, totuşi, să le observăm şi vom ţine cont de existenţa lor, de exemplu: lipsa monitorizării traficului de către echipajele de poliţie, nivelul scăzut de aplicare a legii, nivelul scăzut al amenzilor rutiere, nivelul scăzut de educaţie și pregătire teoretică şi practică a conducătorilor auto, lipsa aplicării legilor privind controlul şi monitorizarea calităţii serviciilor de ambalare şi comercializare a shaormei, lipsa controlului igienico-sanitar al personalului care manipulează produsele în magazin. Toţi aceşti factori, pe care îi numim variabile, ne sunt cunoscuţi din informările publice (de exemplu din mass-media) sau din observaţiile proprii. Acestea sunt exemple de puzzle-uri foarte complexe de factori empirici, caracteristici, fenomene, comportamente, pe care cercetătorii din științele sociale încearcă să le explice.

Statistica socială este o ştiinţă aplicată în domeniul cunoaşterii societăţii şi a oamenilor, cuprinzând metodele prin care clasificăm şi interpretăm date cantitative în consens cu modul în care am formulat şi testat ipotezele. Despre testarea ipotezelor folosind instrumente specifice de analiză, vom discuta în detaliu în capitolele 4 şi 5. În general, se consideră că statistica socială poate să se rezume la simpla descriere şi evaluare a (proprietăţilor) datelor şi informaţiilor empirice. Aceasta mai este cunoscută şi sub numele de statistică descriptivă şi ne este utilă de exemplu atunci când indicăm numărul de copii înmatriculaţi într-o clasă, numărul şomerilor din România în anul 2022, sau procentul de femei care locuiesc în zona rurală şi folosesc internetul pentru a se informa. Aceasta utilizare a statisticii pentru a descrie lucrurile este și cea mai des şi mai larg utilizată formă de calcule statistice. Cu toţii ne amintim de modul în care ne calculam media anuală din, spre exemplu clasa a VII-a, sau calcularea şi compararea mediei de la o anumită materie cu media de la o altă materie. La orele de educaţie fizică profesorul foloseşte acest tip de măsurători statistice pentru a ordona copiii dintr-o clasă în funcţie de înălţime.

Uneori ne dorim să ştim mai mult decât un rezultat al unei simple numărări a unor copii, înălţimi, femei din zona rurală sau şomeri. Ne-am dori să ştim de ce unii copii sunt mai înalţi decât alţii (poate pentru că părinţii lor sunt mai înalţi, sau din cauza alimentaţiei, sau din cauza zonei în care locuiesc, sau, de ce nu, din cauza stării generale de sănătate mai bună); de ce diferă proporţia femeilor din zona rurală care folosesc internetul de cele din zona urbană (poate pentru că în zona urbană conexiunea la internet este mai bună, sau pentru că femeile din zona urbană au un venit mai ridicat decât cele din zona rurală), de ce şomajul în România este la nivelul observat în anul 2022, sau, cât de înalt voi fi atunci când voi fi adult, dat fiind că acum sunt printre cei mai scunzi din clasa mea. Răspunsurile la aceste întrebări nu le mai putem obţine folosind statistica descriptivă, ci doar făcând apel la statistica inferenţială. Prin urmare, cu ajutorul raţionamentului inferenţial putem testa relaţiile dintre caracteristicile observaţiilor pe care le facem în lumea empirică, deducând relaţiile dintre proprietăţile a două sau mai multe distribuţii de date.

Pentru a afla răspunsuri la aceste tipuri de întrebări folosim statistica. Dezvoltată de matematicieni, statistica este frecvent folosită în ştiinţele sociale pentru a analiza realitatea empirică. Aşadar, ea reprezintă o sumă de metode prin care

clasificăm şi interpretăm datele cantitative în concordanţă cu formularea şi testarea ipotezelor, folosind calcule matematice. O putem folosi pentru a descrie, dar şi pentru a explica realitatea. Putem diferenţia între statistica descriptivă şi cea explicativă. Statistica descriptivă este folosită pentru descrierea şi evaluarea caracteristicilor sau proprietăţilor datelor (de exemplu, pentru a identifica şi raporta procentul de femei din zona rurală care folosesc internetul pentru a se informa, sau procentul de alegători care afirmă că au încredere în partidul A). Statistica explicativă este numită şi statistică inferenţială şi este folosită pentru testarea relaţiilor de cauzalitate dintre caracteristicile observaţiilor şi deducerea relaţiilor dintre proprietăţile a două sau mai multe distribuţii de date (de exemplu, pentru stabilirea legăturilor cauzale între genul indivizilor, zona lor de rezidenţă, gradul de informare, educaţie, respectiv preferinţa electorală pentru un anumit candidat).

Cercetarea științifică are la bază canoane științifice. În științele sociale aceste standarde folosite se referă la: formularea întrebărilor de cercetare, a ipotezelor, operaționalizarea conceptelor și măsurarea variabilelor, selecția cazurilor studiate, selecția unităților de analiză, culegerea datelor și informațiilor empirice, analiza acestora. Științele sociale studiază relațiile umane (instituțiile, indivizii și regulile care guvernează societățile și comunitățile) și se bazează pe studiul variabilelor identificând tipare de cauză și efect în realitatea empirică. Cercetarea empirică se bazează pe testarea ipotezelor pe baza datelor culese, folosind raționamentul inferențial cauză → efect (de exemplu: Nori → Precipitații; Zona de rezidență → Gradul de informare politică). Astfel, explicăm un fenomen prin intermediul unui alt fenomen pe care îl cunoaștem și pe care putem să-l explicăm. Cauzalitatea nu este un concept statistic ci mai degrabă unul teoretic. O relație este cauzală nu pentru că datele statistice ne arată acest lucru, ci pentru că persoana care realizează analiza se bazează pe interpretarea datelor empirice pentru a fundamenta argumentul privind cauzalitatea. Lanțul inferențial cauză → efect este prezentat în mod plauzibil în acest fel folosindu-ne de argumente logice, contextualizări, fundamente empirice.

O primă regulă din acest canon ştiinţific pe care îl folosim în ştiinţele sociale poartă numele de **falsificabilitate**. A fost teoretizată de Karl Popper (1981). Falsificabilitatea se referă la capacitatea unui enunţ de a nu fi ştiinţific dacă nu poate fi demonstrat sub nici un chip a fi fals. Pentru simplitatea explicaţiei utilităţii acestui

principiu, enunțul este o teorie care se referă la o realitate pe care o putem observa și măsura empiric. Astfel, Popper (1981, 83) argumenta că „[putem] considera ca empirice sau ştiinţifice numai acele sisteme care pot fi testate (controlate) prin experienţă. […] Un sistem al ştiinţelor empirice trebuie să poată eşua în confruntarea cu experienţa". Desigur, acest lucru nu înseamnă că o teorie este falsă sau că urmărim să dovedim falsitatea acesteia, ci că o teorie care are pretenţia a fi dovedită empiric adevărată şi, prin urmare, este acceptată şi validată empiric de comunitatea ştiinţifică, poate fi respinsă oricând ca fiind falsă, pe baza unor studii reproductibile, care folosesc alte date, la un alt moment de timp. O teorie care afirmă că există o legătură între satisfacţia de la locul de muncă şi nivelul salarizării este falsificabilă. Ea poate fi testată empiric. Chiar dacă pe un anumit set de date (observaţii), un cercetător poate identifica o legătură între aceste două caracteristici (satisfacţie şi salariu), oricând pot apărea noi observaţii, culese de alţi cercetători, care să arate că nu există nici o legătură între aceste caracteristici, prin urmare teoria poate fi dovedită a fi falsă. În schimb, un enunţ conform căruia creşterea cheltuielilor publice pentru cercetări poate conduce la identificarea de schelete de câini galbeni pe planeta Pluto, nu este un enunţ falsificabil.

Prin urmare, o teorie trebuie să permită posibilitatea logică şi empirică de a fi dovedită falsă, ulterior enunţării sale, chiar şi ulterior testării sale empirice. Adevărul absolut nu poate fi asumat de nici o persoană, cu atât mai puţin de cercetători în demersurile lor de explicare a realităţii. Această caracteristică este importantă nu doar din punct de vedere al delimitării enunţurilor ştiinţifice de cele neştiinţifice, ci şi pentru asigurarea reproductibilităţii ştiinţifice.

Această **regulă**, **a reproductibilităţii**, trebuie înţeleasă ca fiind fundamentală pentru utilitatea publică a demersurilor ştiinţifice. Activitatea ştiinţifică este publică, iar concluziile sale sunt incerte (pot fi oricând contrazise de alte noi cercetări). Rezultatele unei cercetări ştiinţifice trebuie să fie reproductibile deoarece în lipsa acestei caracteristici am fi lipsiţi de verificarea validităţii acelor concluzii. Nu în ultimul rând, în lipsa publicării acestor concluzii, atât reproductibilitatea, cât şi utilitatea publică a cercetării ştiinţifice ar fi imposibile. Să presupunem că cercetătorii care au descoperit vaccinul anti-coronavirus ar fi ţinut secrete aceste rezultate. Pe de o parte, am fi putut concluziona că pentru publicul larg, utilitatea cercetării lor ştiinţifice ar fi fost zero, neputând beneficia de efectele acestuia (accesul la un vaccin

care le poate salva viața). Pe de altă parte, în lipsa publicării rezultatelor studiului, am fi putut afirma că, din punct de vedere ştiinţific, nu putem verifica dacă într-adevăr vaccinul pe care ei l-au descoperit are efectul dorit împotriva coronavirusului şi, de asemenea, dacă există efecte adverse ale acelui vaccin. Odată publicate rezultatele cercetării, acestea pot fi supuse verificărilor de către alţi cercetători, folosind aceleaşi date sau date diferite, dar, de asemenea, pot conduce la avansul cunoaşterii ştiinţifice, aducând noi evidenţe empirice, contrazicând altele, sau generând noi teorii. În lipsa publicării rezultatelor şi în lipsa caracterului reproductibil al analizelor şi concluziilor, un studiu (de exemplu din ştiinţele sociale) nu poate fi etichetat ca fiind ştiinţific. Metodele şi tehnicile pe care le vom detalia în capitolele următoare ne vor ajuta să înţelegem cum putem să asigurăm reproductibilitatea analizelor noastre.

În plus, regula de **publicare a rezultatelor** cercetării ajută la avansarea cunoaşterii ştiinţifice despre acel subiect. E posibil ca acea problemă studiată să primească clarificări sau prin publicarea rezultatelor unei noi cercetări traiectorii de studiu empiric, linii de argumentare să fie închise sau deschise ca urmare a invalidării unor studii anterioare. Nu în ultimul rând, reproductibilitatea studiilor permite identificarea unor noi informaţii, a unor noi cazuri relevante, şi poate să contrazică concluzii anterioare ale unor studii cu mare impact. De exemplu, un studiu celebru al psihologei Amy Cuddy şi a colaboratorilor săi, cu privire la modul în care comportamentul indivizilor şi succesul lor este influenţat de starea mentală şi fiziologică a acestora, identifica, folosind un design experimental care a inclus 42 de subiecţi umani, legături pozitive între nivelul de testosteron crescut, și nivelul de cortizol scăzut, ca urmare a unor comportamente ce mimau poziţiile de putere, la fel crescând şi auto-percepţia gradului de siguranţă şi asumarea unor riscuri (Carney, Cuddy, și Yap 2010). Studiul, promovat prin capacitatea autoarei de comunicare facilă a rezultatelor cercetării ştiinţifice, a generat valuri de apreciere şi entuziasm. O înregistrare a prezentării publice a concluziilor studiului este disponibilă aici:

https://www.ted.com/talks/amy_cuddy_your_body_language_may_shape_who_you_are

Publicarea rezultatelor studiului a deschis o dezbatere foarte aprinsă cu privire la concluziile sale, a designului şi metodelor folosite, dar şi a rezultatelor analizelor

statistice. Publicarea acestor date şi informaţii metodologice a permis multor cercetători să încerce să replice aceste analize şi rezultate. Aceştia au replicat aceeaşi analiză pentru a obţine fundamentarea suplimentară a concluziilor sau respingerea acestora. Un asemenea studiu (Ranehill et al. 2015) care a inclus 200 de subiecţi umani nu a reuşit să ajungă la aceleaşi concluzii ca studiul lui Amy Cuddy, concluziile sale fiind că nu există legătură între poziţia de putere, comportamentul de risc şi starea mentală sau fiziologică a subiecţilor.

O altă regulă importantă este aceea potrivit căreia un studiu științific urmărește **culegerea de date sistematice**, nu întâmplătoare. Dacă vrem să explicăm de ce plouă, nu este suficient să ne uităm pe fereastră și să constatăm că plouă, afară fiind soare, o temperatură pozitivă, fără vânt. Asemenea observații sunt întâmplătoare. Pentru a putea explica de ce plouă, ar trebui să culegem informații (observații) sistematice. Prin urmare, ar trebui să facem observații în zile diferite, în anotimpuri diferite, în regiuni diferite, poate și la ore diferite, atunci când vântul are viteze diferite, atunci când umiditatea aerului este diferită, dar și atunci când sunt nori pe cer sau când e soare, pentru a putea identifica corect posibilele cauze ale apariției ploii. Aceste observații sistematice ne pot ajuta să diferențiem între cauze aparente și cauze reale ale fenomenului pe care dorim să îl explicăm.

1.1. Etapele cercetării

De obicei, studiile ştiinţifice pornesc de la o caracteristică specifică cercetătorilor: întrebarea determinată de dorinţa de a afla răspuns la o problemă din realitatea care ne înconjoară, (altfel spus, întrebarea de cercetare). Curiozitatea este o trăsătură pe care o putem observa încă de la cele mai fragede vârste. Cei mai mulţi copii pun o mulţime de întrebări părinţilor, fraţilor mai mari, apoi educatorilor şi învăţătorilor. Acest tip de curiozitate, deseori simplă (sau pentru probleme simple), este cea care îi va determina pe unii dintre aceşti copii să devină cercetători, intelectuali, oameni de ştiinţă, inventatori, antreprenori. Din păcate, de prea multe ori această curiozitate manifestată de la vârste fragede de cei mai mulţi copii, nu este

întreţinută nici de familie, nici în şcoală. De prea multe ori observăm că în scoală copiilor nu le mai este cultivată această curiozitate pe care să o manifeste prin cât mai multe întrebări, cât mai multe nedumeriri şi să fie îndrumaţi spre a le transforma în capacitatea de a formula întrebări, dorinţa de cunoaştere şi de a studia cât mai în amănunt problemele formulate în aceste întrebări.

Astfel, prima etapă a oricărei cercetări este cea în care ne punem întrebări și încercăm să găsim acea afirmație care poate primi un răspuns. De ce plouă? De ce unii oameni au o bunăstare mai ridicată decât alții? De ce unii studenți obțin note mai bune decât alții? De ce unele firme sunt mai competitive decât altele? De ce există diferențe de dezvoltare economică între județele din Moldova și cele din Transilvania? De ce unele state sunt mai stabile din punct de vedere politic, în vreme ce guvernarea din alte state se prăbușește? Asemenea întrebări și, desigur, multe altele, sunt exemple de probleme abordate în științele sociale. Pentru a răspunde la aceste întrebări ar trebui să le specificăm, cu alte cuvinte să le facem mai puțin generale, dar totuși, suficient de generale încât să aibă relevanță nu doar pentru un singur individ, pentru un singur județ sau pentru o singură țară, ci pentru mai multe.

Răspunsurile la întrebarea de cercetare (de exemplu, cum putem explica diferențele de performanță academică a studenților?) sunt mai ușor de găsit dacă pornim de la o teorie generală și formulăm / derivăm răspunsurile posibile. Aceste răspunsuri poartă numele de ipoteze, iar ele sunt asumpțiile pe care le facem cu privire la validitatea relațiilor dintre ceea ce vrem să explicăm (diferențele de performanță academică) și ceea ce ne ajută să explicăm această problemă (diferențele de caracteristici personale, cum ar fi genul, nivelul de bunăstare a familiei studentului, nivelul de educație a părinților, nivelul de performanță în educația preuniversitară, numărul de ore de studiu individual, numărul de materiale obligatorii lecturate etc. sau diferențele de caracteristici structurale și organizaționale, cum ar fi nivelul de dotare a universității, numărul de colaborări pe care facultatea le are cu companii private, nivelul de fonduri externe atrase de universitate anul acesta comparativ cu cele de acum un an, sprijinul dat de angajatori educării continue a adulților etc.). Aceste răspunsuri (ipotezele) sunt cele pe care le vom testa.

Ipotezele se referă variabile aflate în relație una cu cealaltă. Deseori vom formula mai multe ipoteze, unele alternative, altele complementare. Rareori realitatea pe care o analizăm (indivizi, grupuri, instituții, comunități, comportamente) este atât de simplă încât să poată primi o explicație printr-o singură ipoteză simplă. De cele mai multe ori ea poate fi explicată din diverse perspective, ținând cont de contexte diferite și de factori explicativi care sunt, cel puțin în aparență, diferiți unii de ceilalți. Prin urmare, ipotezele formulate sunt puse în relație una cu cealaltă astfel încât să identificăm complexul care poate oferi răspunsuri / explicații alternative la întrebarea de la care am pornit cercetarea.

A doua etapă a demersului ştiinţific, extrem de importantă, este cea în care ne asigurăm că putem măsura empiric conceptele (simple sau complexe) pe care vrem să le explicăm. Acest proces se numeşte operaţionalizare şi este procesul prin care transformăm concepte abstracte în concepte măsurabile. Aceste noi concepte primesc caracteristici măsurabile şi observabile. Pe aceste concepte măsurabile le numim variabile. De cele mai multe ori această operaţionalizare are la bază literatura de specialitate, teoriile dezvoltate şi testate de alţi autori. Vom discuta pe larg în secţiunea următoare caracteristicile conceptelor din ştiinţele sociale, diferenţele dintre concepte şi variabile, caracteristicile variabilelor, dar şi modul în care le operaţionalizăm.

A treia etapă este cea în care culegem informaţiile, observaţiile, pe care deseori le numim date, pe care le vom folosi pentru a testa ipotezele cercetării noastre. Pentru a putea produce această testare trebuie să ne asigurăm că datele culese sunt relevante pentru fiecare variabilă din studiul nostru. De exemplu, dacă ipoteza noastră afirmă că performanţa academică ridicată a studenţilor este determinată pozitiv de genul persoanei, de un număr mai mare de ore de studiu individual, de un număr ridicat de prezenţe în clasă, de performanţa academică din clasele de liceu, de gradul de dotare materială a universității, și de numărul de colaborări pe care departamentul le are cu companii private, atunci va trebui să ne asigurăm că informaţiile pe care le culegem de la subiecţii incluşi în studiul nostru sunt măsurători ale acestor variabile incluse în ipoteză.

Colectarea datelor se realizează prin selectarea anumitor cazuri empirice ale căror caracteristici le vom folosi pentru a testa relaţia între variabile. Selecţia

observaţiilor şi a cazurilor se face, de regulă, în studiile cantitative prin eşantionare (sau selecţie, aşa cum o întâlnim şi în studii precum Rotariu et. al (1999) sau Rotariu şi Iluţ (1997)). Eşantionarea ne permite să tragem concluzii despre un întreg (de exemplu un grup de persoane, o comunitate, o societate, instituţii, organizaţii) prin studierea unei mici părţi a acestui întreg (G. King, Keohane, și Verba 2000).

Ulterior colectării informaţiilor empirice încercăm să identificăm tipare din baza de date folosind metode şi tehnici care să distingă variaţiile sistematice de variaţiile ne-sistematice (sau întâmplătoare). Aceste tipare sunt specificate prin modelul explicativ, care la rândul său este specificat matematic. În această etapă a analizei datelor vom încerca să identificăm toate posibilele cauze reale ale efectului pe care dorim să îl explicăm. Vom testa forma, intensitatea şi direcţia relaţiilor dintre variabilele incluse în ipotezele de cercetare. Astfel, vom putea distinge între cauze aparente (studenţii care poară ochelari tind să ia note mai mari, deci să aibă performanţe academice mai ridicate la cursul de Statistică Socială, decât studenţii care nu poartă ochelari), şi cauze reale (studenţii care petrec mai mult timp studiind în afara orelor de clasă, cei care citesc mai mult în bibliotecă, cei care lucrează mai mult la calculator exersând tehnici de analiză cantitativă cu ajutorul unor programe software specializate, tind să ia note mai mari la Statistică Socială). Astfel, de exemplu, putem specifica matematic modelul care rezumă această relaţie: o notă mai mare la statistică = mult studiu individual + multă lectură de specialitate + multe vizite la bibliotecă + multe analize pe baze de date. Desigur, pe bună dreptate, ne-am putea aştepta ca în mod rezonabil ca modelul nostru să nu fie exhaustiv, sau să existe şi alte explicaţii alternative. Aceste probleme le formulăm şi le testăm în această etapă a cercetării ştiinţifice, folosind metode şi tehnici de analiză.

După cum explicam mai sus, rezultatele studiilor ştiinţifice sunt incerte. Concluziile pe care le tragem cu privire la ipotezele noastre sunt la fel de incerte, deşi, cel puţin pentru moment, pe datele culese şi analizate, ele se susţin, și se dovedesc a fi adevărate. Toate aceste etape, însă, sunt raportate întotdeauna la teorie (S. Chelcea 2001; Vlăsceanu 2013). Fără un cadru teoretic care să o susţină, fără o teorie care să însumeze concluziile studiului, cercetarea ştiinţifică este doar o înşiruire de cifre, informaţii și date. În plus, chiar cadrul teoretic şi legăturile teoretice pe care le stabilim între cauze şi efecte (variabilele pe care le studiem), ne ajută să precizăm lanţul cauzal

inferenţial, modelul nostru explicativ, care prezintă logic, plauzibil, relaţia dintre cauze şi efecte, conform argumentelor lui Jon Elster (2013) şi Hubert Blalock (2018).

1.2. Concepte, variabile, ipoteze

Precizam anterior că întrebările noastre de cercetare se referă la probleme reale pe care am dori să le înţelegem mai bine și să le explicăm. Aceste formulări au la bază conceptele. Conceptele reprezintă elementul fundamental al oricărui limbaj ştiinţific comun. Fără aceste concepte nu ne-am înţelege și am discuta fiecare pe „limba" lui având impresia că suntem în consens, deși acesta nu există în realitate. Prin urmare, putem spune ca un concept reprezintă un simbol sau o semnificaţie pe care o dăm unui obiect (masă, individ) sau unei proprietăţi a acestui obiect (lăţime, înălţime), sau unui fenomen comportamental (al indivizilor, cum ar fi, de exemplu, preferinţa unui individ de a vota pentru candidatul A şi nu pentru candidatul B) (Frankfort-Nachmias, Nachmias, și DeWaard 2015, 24–27).

Folosirea și diferențierea conceptelor este învățată de la o vârstă fragedă. Acest proces de învățare a conceptelor permite dezvoltarea unui limbaj comun care să fie ușor înțeles de cei din jurul nostru. Desigur, nu o facem în mod conştient, de dragul diferenţierii ştiinţifice, ci o facem pentru că am fost socializați de părinți și educatori și am învățat de la aceștia cum să folosim limbajul. Un copil cere o cană de apă în mod instinctiv când îi este sete, şi se aşteaptă ca lichidul incolor şi inodor din cană să fie apă care îi reduce senzaţia de sete. În mod identic, părintele care îi dă acea cană cu apă nu stă să evalueze (să definească) conceptul de apă sau pe cel de cană. Ştie deja, pe baza unei socializări şi învăţări pe care le-a avut pe când era copil, ce este aceea o cană şi care sunt diferenţele dintre apă şi, să spunem, lapte, alcool, suc sau orice alt lichid. În mod rezonabil, putem afirma că ambii indivizi (copilul, respectiv părintele) dau aproape instinctiv, aceeaşi definiţie celor trei concepte: cană, apă, sete.

Dar, după cum putem observa, fiecare din aceste concepte are grade diferite de abstractizare şi particularizare. Apa se referă la un lucru mai degrabă puţin abstract (o putem vedea, atinge, mirosi, bea) şi tinde să fie aceeaşi indiferent de locaţia unde

ne-am afla. Ea are, în mare măsură, aceleaşi caracteristici generale care ne determină să afirmăm că acel lichid este apă. Cana, deşi este la rândul său un concept particular, o putem identifica (prin definire) ceva mai greu decât conceptul de apă. O cană poate avea forme diferite, poate fi compusă din elemente diferite, dar, la fel ca şi apa, ea are referenţi empirici destul de clari, obiecte care pot fi văzute, atinse, în general observate. Însă, în ceea ce priveşte setea, acest concept este unul mult mai abstract, poate chiar mai greu de definit. Nu rareori avem o sete care să nu fie satisfăcută de apă, ci de alt lichid, mai mult sau mai puţin dăunător sănătăţii. În plus, setea în sine nu o putem vedea, atinge, observa, sau mirosi. Aşadar, afirmăm despre acest tip de concepte că ele sunt mai abstracte decât altele.

Științele tind mai degrabă să folosească anumite concepte specifice corpului de elemente pe care le studiază. Ne amintim cu toţii de la orele de fizică din şcoala generală sau din liceu, despre discuţiile şi definiţiile, poate chiar şi formulele matematice, pe care le-am folosit pentru a înţelege concepte precum gravitaţia, forţa, masa, viteza, timpul, energia, căldura, puterea etc. În chimie, o altă ştiinţă exactă, folosim concepte precum energie, forţă, electromagnetism, atomi, ioni, reacţie, legătură, termodinamică etc. Alte ştiinţe, cum ar fi psihologia, folosesc mai puţin sau deloc aceste concepte exacte, ci alte concepte, precum personalitatea, devianţa, cogniţia, atenţia, percepţia, conştiinţa. Cercetătorii din sociologie folosesc alte concepte specifice, precum clasa socială, mobilitatea, inegalitatea, sărăcia, acţiunea colectivă, reţeaua, cultura, organizaţia, munca. În demografie, un sub-domeniu foarte important al sociologiei, cercetătorii folosesc concepte specifice, cum ar fi migraţia, fertilitatea, natalitatea, mortalitatea, în vreme ce în ştiinţele politice, se folosesc concepte precum democraţia, regimul politic, instituţiile, partidele, puterea, reprezentarea, justiţia, egalitatea, ca să dăm doar câteva exemple.

Putem observa că, spre deosebire de ştiinţele exacte unde conceptele, deşi sunt abstracte, pot fi mai degrabă măsurate prin ele însele, în ştiinţele sociale conceptele prin care reprezentăm realitatea socială sunt mai degrabă măsurate prin alte concepte. Formulele sau modalitatea de măsurare este standardizată şi universal acceptată. Un chimist din România, care discută cu un chimist din Franţa, unul din Federaţia Rusă şi altul din Coreea de Nord, va avea în mare măsură aceleaşi noţiuni şi definiţii pentru aceste concepte. Într-o situaţie similară va fi un matematician, sau un fizician. Deși

conceptele pot fi mai degrabă imprecise, aceasta nu înseamnă că în științele sociale nu folosim metode științifice de măsurare și analiză a datelor aproape identice cu cele din științe precum medicina sau fizica. Astfel, analizăm informațiile cantitative culese în științele sociale cu ajutorul variabilelor, folosind un raționament inductiv sau deductiv, inferențial, și utilizând tehnici și metode de analiză influențate de sau întru totul preluate din științele exacte cum ar fi matematica. În acest fel putem identifica relațiile dintre variabile și fundamenta logica din modelele explicative.[2]

În ştiinţele sociale, cum sunt cele exemplificate mai sus, conceptele au un grad de abstractizare mare; sunt mai inexacte şi mai puţin universale în ce priveşte definiţia şi consensul privind înţelesul lor. Consensul este mult mai greu de obţinut în aceste ştiinţe, nu doar pentru că avem şcoli diferite de gândire, ci mai ales pentru că realitatea socială, politică şi economică este extrem de diferită între societăţi dezvoltate în mod diferit şi care au urmat căi diferite. Pentru un cercetător american, francez, sau român, concepte precum libertate sau democraţie pot să însemne altceva şi, astfel, să primească definiţii diferite faţă de sensul (definiţia) pe care îl dă acestor concepte un rus, un chinez sau un nord-coreean. Chiar dacă toţi folosesc conceptul de democraţie sau libertate, ne-am aştepta ca, în mod rezonabil, să nu existe aceeaşi înţelegere (definiţie) a acestui concept. Pentru unii, democraţia poată să fie măsurată prin responsabilitatea reprezentanţilor aleşi în faşa cetăţenilor, prin societate civilă dezvoltată, prin competiţia liberă între mai multe partide, sau prin alternanţa la guvernare, în vreme ce pentru alţii, democraţia poate să fie redusă doar la organizarea de alegeri, cu restrângerea drepturilor opoziţiei de a-i contesta pe cei care sunt la putere.

Putem afirma, prin urmare, că în ştiinţele sociale conceptele au mai degrabă un caracter imprecis. În plus, în aceste ştiinţe conceptele nu sunt categorice, precum cele din ştiinţele exacte, ci mai degrabă tentative, având la bază un consens considerabil mai restrâns decât cel obţinut în ştiinţele exacte (Rotariu et al. 1999). Această caracteristică este determinată nu doar de diferenţele regionale sau structurale de înţelegere a conceptelor, ci şi de continua evoluţie a indivizilor, a societăţii, a formelor

[2] Îi mulțumim lui Bogdan Voicu pentru sugestia de a clarifica și detalia aceste argumente.

lor de organizare, de impredictibilitatea unor activităţi şi fenomene sociale, economice sau politice.

Precum cuvintele pe care le folosim pentru a ne exprima, a vorbi, a citi, şi conceptele sunt legate unele de celelalte prin legături logice. Existenţa acestor legături este testată prin referinţe la lucrurile şi fenomenele observabile, pe care le numim şi referenţi empirici. Legăturile dintre concepte sunt determinate de gradul de înţelegere şi consens existent (Frankfort-Nachmias, Nachmias, și DeWaard 2015). Putem exemplifica printr-un grad de înţelegere aproape universală utilizarea unui concept comun, cum ar fi cel de lapte. Când mergem la cumpărături şi cerem vânzătorului un litru de lapte, nu ne aşteptăm să primim un litru de antigel, sau un litru de vopsea albă, nu doar pentru că noi am cerut lapte nu alt produs lichid, ci şi pentru că vânzătorul, la rândul lui, are exact aceeaşi înţelegere ca şi noi, a conceptului de lapte. Prin urmare, ne va oferi o cutie care va conţine un lichid ce reprezintă referentul empiric al conceptului de lapte (definit ca fiind un lichid alb, produs de vacă, oaie, capră sau bivoliţă, pentru hrănirea propriilor pui, şi care este adecvat consumului uman).

În ştiinţe, conceptele sunt definite pe baza unor informaţii extrase din corpul de teorii şi idei produse de studii şi cercetări de specialitate, ele furnizând ceea ce numim literatura de specialitate. Rareori vom analiza probleme care nu au mai fost niciodată, în nici o formă, analizate de alţii; aşadar, rareori vom „reinventa roata”. Dar, aşa cum argumenta Thomas Kuhn (2008, 86), acest lucru nu este imposibil, iar în ştiinţele exacte revoluţia teoriilor ştiinţifice a dus la apariţia unor noi teorii. Totuşi, chiar şi aceste noi teorii se bazează, măcar parţial, pe corpul de idei al teoriilor mai vechi. În ştiinţele sociale, teoriile noi derivă deseori din cele vechi. Regula este mai degrabă inovarea, decât inventarea de noi teorii şi / sau concepte. Să luăm, de exemplu, conceptul de democraţie. Vom constata că deşi aparent pare clar, clasic deja, încercarea de adaptare a acestuia la contexte actuale care variază, a condus la inovarea conceptuală şi la extinderea limitelor conceptului dincolo de ceea ce în mod normal ne-am fi aşteptat să cuprindă. Conceptul a fost întins precum un gumilastic, în funcţie de necesităţile studiilor cercetătorilor care l-au folosit, ajungându-se la o inflaţie de concepte noi ale democraţiei, ce a produs sute de subtipuri ale acestui concept (de exemplu, democraţie populară, democraţie semi-parlamentară, chiar și democraţie iliberală) aşa

cum remarcau David Collier și Steven Levitsky (1997). Elasticitatea conceptuală (Sartori 1970; Collier și Mahon 1993), devine, astfel, o problemă importantă pentru asigurarea unității în utilizarea conceptelor ştiinţifice.

Identificarea limitelor conceptuale se realizează prin definirea conceptului. Ne amintim, însă, că în ştiinţele sociale suntem nevoiţi să transformăm conceptele folosite într-unele speciale, numite variabile, prin procesul numit operaţionalizare. Acest lucru este determinat de studierea unor probleme abstracte și complexe, care sunt extrem de rar neschimbătoare, mult mai des putând fi posibilă observarea unei variații temporale sau transformări ale acestora.

Variabilele conţin în ele însele noţiunile de grad şi diferenţiere, particularitate pe care o urmărim să o obţinem prin operaţionalizarea conceptelor în variabile. Prin urmare, variabilele pot lua valori diferite pentru indivizi / cazuri diferite (Agresti și Finlay 2014). Genul, înălţimea, culoarea părului, preferinţa de vot, multipartidismul, votul secret, libertatea de asociere reprezintă variabile care operaţionalizează concepte, fie prin ele însele, fie prin noi concepte (variabile). Variabilele se referă la cazuri (referenţi empirici) şi la atribute comune ale acestor cazuri. Variabilele păstrează o caracteristică fundamentală pe care am precizat-o anterior ca fiind întâlnită în cazul conceptelor: caracterul incert. Oricând, noi sau alţi cercetători putem oferi o definiţie diferită faţă de cea oferită anterior. În plus, ele pot fi compuse din alte variabile, sau privind din cealaltă perspectivă, mai multe variabile pot forma împreună o nouă variabilă compusă, numită şi variabilă indicator (de exemplu, democraţie, personalitate, devianţă). Nu în ultimul rând, o altă caracteristică importantă pe care o au variabilele este aceea că ele pot lua o valoare pentru fiecare caz sau referent empiric analizat. Pentru fiecare din acești referenți empirici măsurăm caracteristica definită de variabila de analiză. De exemplu, pentru variabila prezența la vot, cazurile noastre analizate vor lua fie valoarea DA, fie valoarea NU, neputând lua concomitent, acelaşi caz, la acelaşi moment de timp, atât valoarea DA, cât şi valoarea NU.

Variabilele se caracterizează prin cuantificare şi măsurare. Cuantificarea reprezintă stabilirea unei cantităţi standard, numerice, unei măsurători a unui lucru, fenomen sau comportament. Astfel, distanţa parcursă de un alergător este măsurată

în *metri*, iar temperatura ambientală este măsurată prin *grade de temperatură*. Alternativ cuantificării, am putea afirma, pentru cele două exemple că alergătorul a parcurs o distanţă *mare*, iar în ce priveşte vremea de afară am putea spune că este *cald*. Cu toate acestea, *distanţă mare* poate să însemne altceva pentru o persoană sedentară, şi cu totul altceva pentru un maratonist. Aceeaşi vreme, cu aceleaşi caracteristici, poate fi *caldă* pentru un locuitor din regiunea Lapland, dar *rece* pentru un locuitor din Sahara. Prin urmare, cuantificarea este superioară alocării de valori nestandardizare unui anumit lucru. De cele mai multe ori, atunci când dorim să avem o exactitate mare, vom prefera să măsurăm vremea prin grade de temperatură, greutatea prin kilograme, distanţa în metri, sau încrederea în guvern printr-o scală numerică. În viaţa de zi cu zi nu vom avea nevoie, în mod necesar, de acest nivel ridicat de exactitate. Totuşi, deşi vom folosi aprecieri ale unor caracteristici sau măsurători inexacte (*cald*, *rece*, *devreme*, *târziu*, *mare*, *mic*) am prefera ca atunci când întrebăm cât este ora să ni se spună ora exactă sau cu aproximaţie de câteva minute, nu aprecieri generale cum ar fi *trecut de prânz*, *aproape seară* etc.

Cuantificarea poate fi discretă sau continuă (Agresti și Finlay 2014). Cuantificarea discretă se realizează prin numărarea unităţilor unui lucru, de exemplu prezenţa la vot (DA, respectiv NU), genul unei persoane (bărbat; femeie). Valoarea pe care o poate lua o asemenea variabilă prin cuantificare este formată din numere întregi, între ele neexistând posibilitatea identificării unei valori intermediare. Termenul provine din limba latină (*discretus*) şi înseamnă separat sau distinct. Cuantificarea continuă permite o variaţie pe un continuum, pe dimensiuni, astfel încât cazul poate poate lua orice valoare pe variabila măsurată. De exemplu, cazurile pot lua orice valoare în intervalul stabilit de variabile precum vârsta, temperatura, nota de absolvire a unui curs, numărul de ani de şcoală, vârsta la absolvirea ultimului ciclu de învăţământ etc. Astfel, din punct de vedere al cuantificării, variabilele pot fi discrete sau continue. Ne vom folosi de această caracteristică a variabilelor când vom discuta despre nivelurile de măsurare. Cuantificarea este mai uşor de obţinut când lucrăm cu variabile şi concepte măsurabile, cum ar fi cele exemplificate din ştiinţe precum fizică sau chimie. Cu toate acestea, în ştiinţele sociale deseori lucrăm cu variabile şi concepte care sunt mai greu de măsurat sau asupra cărora nu există un

consens larg (de exemplu, alienare, stres, securitate personală, fericire, democraţie, productivitate).

Măsurarea este o calitate intrinsecă variabilelor şi discuţiilor analitice, prin urmare o vom întâlni în variabilele care operaţionalizează conceptele nemăsurabile sau greu măsurabile prin ele însele. Măsurarea este exprimată deseori în ştiinţele sociale prin intensitatea unei opinii, de exemplu atunci când folosim numere pentru a măsura un anumit fenomen sau comportament pe care îl măsurăm pe o scală de la 1 la 10 sau de la -10 la +10 (de exemplu performanţa academică măsurată de la 1 la 10, sau gradul de satisfacţie cu salariul obţinut luna trecută, de la -10 la +10). Măsurarea ne ajută să identificăm mai uşor intensitatea caracteristicilor unei variabile, dar şi să comparăm cazuri diferite (de exemplu, elevii din zona urbană obţin, în general, note mai bune decât cei din zona rurală).

În funcţie de poziţia lor în cadrul logicii inferenţiale prin care identificăm cauze ale efectului pe care dorim să îl explicăm, variabilele pot fi independente, intermediare, sau dependente. **Variabilele dependente** mai sunt numite şi variabile de explicat, variabile răspuns, variabile efect sau variabile *outcome*. Variabilele dependente sunt acele variabile despre care vrem să aflăm mai mult, pe ele vrem să le explicăm pe baza raţionamentului inferenţial. Acestea sunt variabilele asupra cărora vrem să vedem dacă o schimbare de un anumit tip şi de un anumit nivel în variabila cauză produce modificări. În exemplul folosit anterior în acest capitol variabila dependentă este performanţa academică a studenţilor. Aceste variabile dependente le regăsim în forma directă, măsurabilă, sau conceptuală, nemăsurabilă, în întrebarea de cercetare. Prin urmare, variabilele dependente se vor regăsi şi în formularea ipotezelor noastre. De regulă, vom avea în fiecare ipoteză câte o singură variabilă dependentă. Chiar dacă studiul nostru vrea să explice mai mult decât un singur lucru sau comportament, (de exemplu, vrem să explicăm diferenţele de performanţă academică între studenţi dar şi modul în care aceasta va produce efecte ulterioare asupra performanţei de la locul de muncă), va trebui să explicăm pe rând fiecare din aceste variabile dependente.

Variabilele independente sunt numite şi variabile explicative, determinanţi, variabile cauză, variabile factor sau variabile predictor. Acestea reprezintă ceea ce

presupunem în ipotezele noastre ca fiind caracteristicile ale căror variaţii determină o schimbare în variabila dependentă. De exemplu, în ipoteza noastră enunțată mai sus, afirmam că **studenţii obţin o performanţă academică mai bună** (măsurată prin notele obţinute sau prin media notelor) atunci când beneficiază de un *nivel de bunăstare mai ridicat al familiei* lor (poate de exemplu, pentru că nu mai simt presiunea asigurării resurselor financiare pentru chirie, taxe, costuri de viaţă); de un *nivel mai ridicat de educaţie al părinţilor* (e posibil ca părinţii mai educaţi să înţeleagă mai bine importanţa educaţiei, să creeze o atmosferă familială care încurajează cititul şi studiul, să îşi încurajeze copiii să înveţe cât mai bine); de un *nivel mai ridicat de performanţă în educaţia preuniversitară* (ne aşteptăm ca elevii foarte buni să fi deprins acele competenţe de organizare şi de învăţare care să le fie utile şi în activitatea educaţională universitară, în plus, să fi acumulat cunoştinţe sumative şi complementare care să îi ajute să adauge mai uşor şi mai eficient alte competenţe pe parcursul studiilor universitare); un *număr mai ridicat de ore de studiu individual*; un *număr mai mare de materiale obligatorii lecturate*.

Diferenţele de performanţă academică nu pot fi explicate doar de caracteristicile individuale ale studenţilor. Ar fi rezonabil şi plauzibil să considerăm că nu doar aceste caracteristici sunt importante pentru o performanţă academică mai ridicată. Prin urmare, consultând literatura de specialitate şi făcând apel la teorie, putem observa că această variabilă dependentă este explicată şi de caracteristici instituţionale şi organizaţionale (Wollersheim et al. 2015). Prin urmare, putem formula şi o a doua explicaţie a diferenţelor de performanţă academică, formulând un model explicativ în care estimăm **performanţa mai ridicată a studenţilor** ca fiind explicată de câteva variabile independente de la nivel organizaţional şi structural cum ar fi un *nivel mai ridicat de dotare a universităţii* (de exemplu gradul de digitalizare al laboratoarelor); un *număr mai mare de colaborări pe care departamentul le are cu companii private* (măsurată prin numărul de acorduri de practică active în ultimul an universitar); un *sprijin mai ridicat dat de angajatori educării continue a adulţilor* (de exemplu măsurată prin număr de zile de concediu cu plată acordate angajaţilor care urmează cursuri universitare).

Variabilele intermediare sunt dependente în raport cu variabilele independente şi independente în raport cu variabila dependentă. Ele sunt acea variabilă care se interpune între variabilele independente şi variabila dependentă în lanţul logic cauzal prin care dorim să explicăm variabila dependentă. În exemplul

nostru de mai sus, între variabilele independente de la nivel structural şi variabila dependentă, „performanţa academică", am putea interpune o variabilă intermediară cum ar fi participarea studenţilor la un număr mare de activităţi extra-curriculare organizate de facultăţi (de exemplu activităţi într-un club de dezbateri şi oratorie) sau angajarea unor experţi practicieni care să predea studenţilor. Această variabilă poate determina creşteri ale performanţei academice a studenţilor care obţin competenţe suplimentare faţă de cele din sala de curs, şi poate intermedia efectele resurselor financiare şi parteneriatelor formale cu mediul privat. Altfel, aceste din urmă variabile e posibil să nu influenţeze în mod direct activitatea individuală a studenţilor ci intermediat prin influenţarea organizaţiei şi a profesorilor.

În cele mai multe studii empirice cantitative vom întâlni un tip de variabile numite **variabile de control**. Earl Babbie le numeşte şi variabile test (2010). Variabilele de control sunt în sine variabile independente (teoretic, ele pot constitui o posibilă cauză pentru efectul studiat) şi sunt folosite pentru a ne asigura că relaţiile dintre variabila dependentă şi variabilele independente din modelul nostru explicativ sunt valabile şi atunci când controlăm această relaţie pentru variabila de control (sau test). În exemplul de mai sus putem transforma variabila gen în variabilă de control şi putem împărţi subiecţii noştri în bărbaţi şi femei, subgrupuri pentru care vom calcula modelele de relaţie dintre variabilele independente ce măsoară caracteristici individuale şi variabila dependentă performanţa academică. Astfel, putem compara efectele la nivel de subgrupuri cu cele la nivelul eşantionului. La fel putem proceda şi pentru al doilea model, care include variabile structurale şi instituţionale.

Menționam anterior că ipotezele prezintă relaţia între două sau mai multe variabile, una fiind dependentă, iar celelalte independente. Ipotezele ne ajută să organizăm şi să structurăm cercetarea ştiinţifică. De asemenea, ele ne ajută să facem presupuneri despre modul în care variabilele independente le influenţează pe cele dependente. Prin ipoteză exprimăm modul în care variaţia variabilei dependente poate fi explicată prin variaţia variabilelor independente. Folosindu-ne de acelaşi exemplu de mai sus am putea afirma că variaţia mediilor obţinute de studenţi poate fi explicată prin variaţia caracteristicilor de la nivel individual şi organizaţional, acestea din urmă fiind, aşadar, variabile explicative. Desigur, aceste caracteristici nu vor reuşi să explice toată variaţia variabilei dependente. Fiecare model explică o parte

din acea variaţie. În mod rezonabil, ne putem aştepta ca, dată fiind complexitatea problemelor sociale analizate (de exemplu, performanţa academică), dar şi eterogenitatea şi gradul ridicat de contexte specifice, atât la nivel individual, cât şi la nivel organizaţional, o parte din variaţia variabilei dependente rămâne neexplicată. Aceasta este ceea ce numim eroare sau termenul de eroare; este partea rămasă neexplicată din variaţia variabilei dependente. Vom folosi în celelalte capitole această noţiune de eroare în diverse forme şi etape ale analizelor cantitative. Pentru o explicaţie detaliată a tipurilor de eroare recomandăm studiul profesorilor Traian Rotariu şi Petru Iluţ (1997). În testarea ipotezelor noastre sau în testele de semnificaţie statistică apelăm la un tip specific de afirmaţie, numită ipoteză nulă. Termenul a fost dezvoltat de statisticianul Ronald Fisher (1971, 15–17). Aceasta reprezintă afirmaţia negativă a răspunsului pe care îl dăm întrebării de cercetare. Aşadar, dacă ipoteza noastră asumă o legătură între variabila dependentă şi cele independente, ipoteza nulă afirmă lipsa acestei relaţii. Uneori testăm ipoteza nulă anterior testării modelului nostru explicativ. Alteori, putem să testăm validitatea modelul afirmat de ipoteza noastră de cercetare (numită ipoteză alternativă ipotezei nule) în raport cu ipoteza nulă, astfel încât respingem ipoteza nulă în momentul în care confirmăm ipoteza alternativă.

Ipotezele au în general câteva caracteristici care ne vor folosi pentru a înţelege cum le putem formula în aşa fel încât să ne ajute în explicarea realităţii empirice. În primul rând **ipotezele sunt empirice, nu normative**; deci formulăm o afirmaţie despre ce este în realitate, cum este această realitate, nu cum ne-am dori să fie. De aceea ipoteza din exemplul de mai sus nu a fost formulată de forma: ar fi bine ca toţi studenţii să aibă performanţă ridicată. **Ipotezele sunt testabile**, adică, ţinând cont de caracteristica de falsificabilitate a teoriilor şi enunţurilor teoretice, putem afirma că şi testarea acestor teorii prin intermediul ipotezelor beneficiază de aceeaşi asumpţie a falsificabilităţii. **Ipotezele trebuie să fie generale**, nu despre un singur caz; prin urmare nu vom formula o ipoteză despre performanţa academică a studentului Popescu, ci despre toţi studenţii sau măcar despre grupuri mari de studenţi (poate dintr-un centru universitar, dintr-un consorţiu universitar, sau dintr-o universitate). În mod similar, cum vom formula ipoteze despre partide în general, despre votanţi în general, despre economii în general etc. **Ipotezele trebuie să fie plauzibile**.

Plauzibilitatea poate fi determinată în mai multe feluri. Cel mai eficient o determinăm făcând apel la teorie: literatura de specialitate pentru acea temă ne va arăta care sunt, de regulă, tipurile de explicaţii pe care le putem da problemei studiate, cu atât mai mult cu cât nu suntem primii şi foarte probabil nu vom fi nici ultimii care vor studia acea problemă. Nu în ultimul rând, facem apel la experienţa noastră, la cunoştinţele noastre despre acea problemă, şi la cunoştinţele de logică pe care le-am acumulat în şcoală. O ipoteză implauzibilă este ilogică. O ipoteză de forma „o performanţă academică mai ridicată este determinată de culoarea părului (persoanele cu părul albastru au performanţă mai ridicată), de înălţimea mai mare a studenţilor, şi de cantitatea de apă consumată înainte de examene" e evident implauzibilă, între variabila dependentă şi cele independente neexistând în mod logic legături cauză-efect.

O altă caracteristică importantă a ipotezelor corect formulate şi utile în demersul de cercetare este **precizarea relaţiei dintre variabilele dependente şi cele independente**: sensul, forma şi intensitatea acestei relaţii. În acest fel, respectăm atât principiul falsificabilităţii (dacă nu indicăm sensul, forma şi intensitatea relaţiei, dacă acesta există în orice grad sau nu există, dacă este pozitivă sau negativă, ipoteza ar fi confirmată, ceea ce, evident, ar produce contradicţii logice), cât și cel al caracterului empiric și al plauzibilității.

Formulate în acest fel ipotezele sunt mai clare şi mai specifice:

„(Ţinând cont de elementele teoretice exprimate cadrul teoretic), presupunem că un sistem economic stabil şi funcţional conduce la dezvoltarea societăţi civile."

„Membrii în sindicate tind să participe mai mult în activităţi politice neconvenţionale decât persoanele care nu sunt membre în sindicat.

Mai mult decât atât, dintre activităţile politice neconvenţionale, sindicatele au impact pozitiv asupra activităţilor care se bazează pe mobilizare şi solidaritatea dintre membri, cum ar fi grevele şi demonstraţiile."

„Studenții obțin o performanţă academică mai ridicată atunci când au un nivel de bunăstare mai ridicat al familiei lor, părinți mai educați, note mai bune în liceu,

studiază individual mai multe ore şi citesc multe materiale obligatorii, în medie, la cursurile dintr-un semestru."

Dacă, în schimb le vom formula astfel, ipotezele nu sunt clare:

„În această lucrare vom încerca să vedem ce determină dezvoltarea societăţii civile."

„Ştim că sindicatele influenţează participarea politică a indivizilor. Vom încerca să vedem cum o influenţează."

„Performanţa studenţilor este influenţată de caracteristicile lor personale."

O ultimă caracteristică a ipotezelor este că sunt consistente cu datele sau, am putea spune, la fel de corect, că datele pe care le colectăm trebuie să fie consistente cu ipotezele noastre. Cercetarea ştiinţifică necesită nu doar cunoştinţe şi dedicaţie ci şi resurse de timp, financiare şi umane, considerabile. Nu vom putea, prin urmare, avea la dispoziţie oricând resurse pe care să le folosim pentru a culege informaţii noi și originale, cu privire la subiectul cercetat. În plus, există multe date pe care nu le putem culege noi personal oricâte resurse am avea (cum ar fi de exemplu date oficiale, culese de guverne, ministere, INS, EUROSTAT, BNR etc.). Nici datele istorice, la nivel micro (individual) sau macro (agregate la nivel regional sau naţional) nu pot fi culese de noi, altfel decât călătorind în timp. Ele au fost deja culese, iar noi le vom prelua din arhive. Prin urmare, în această situaţie ne folosim de datele culese de alţi cercetători sau de instituţii publice. Folosind aceste date formulăm ipotezele ulterior culegerii datelor, prin urmare ipotezele noastre trebuie să cuprindă acele variabile pentru care avem date (culese deja). O a doua situaţie este aceea în care formulăm ipotezele anterior culegerii datelor. În acest caz, datele sunt consistente cu ipotezele formulate după pretestarea instrumentelor de culegere a informaţiilor empirice.

1.3. Cantitativ sau / și calitativ

În funcţie de tipurile de raţionamente folosite pentru a explora realitatea, şi în funcţie de datele folosite și instrumentele de colectare şi analiză a lor, cercetarea poate

fi cantitativă sau calitativă. În cercetarea cantitativă cercetătorii tind să folosească un număr relativ mare de cazuri şi un număr relativ redus de variabile. În literatura de specialitate nu există o limită privind numărul de cazuri sau de variabile. Un studiu în care evaluăm comparativ şomajul din cele 42 de judeţe ale României şi încercăm să îl explicăm utilizând informaţii socio-economice de la nivel mediu de dezagregare (judeţul), chiar dacă va utiliza date longitudinale (deci ar multiplica numărul de cazuri-timp), va avea relativ puţine cazuri. Cu toate acestea, formularea ipotezelor cauzale, raţionamentul inductiv sau deductiv (derivarea teoriei din date, respectiv testarea teoriei pe baza unor date), tehnicile specifice de analiză cantitativă ce permit măsurarea, testele statistice formale de identificare a reprezentativităţii şi a erorilor statistice, dar şi reproductibilitatea analizelor şi posibilitatea de generalizare a concluziilor, vor identifica acest studiu ca unul cantitativ şi nu calitativ. Studiile în care folosim sondaje de opinie, sau cele în care folosim date colectate de instituţii publice sunt exemple de studii cantitative.

Studiile cantitative cu un număr foarte mare de cazuri sunt un tip aparte de studii. Și ele folosesc logica inductivă și testele statistice, în plus, au un grad ridicat de generalizare. Recensămintele sunt un asemenea exemplu de studii cantitative exhaustive ce au milioane, zeci sau sute de milioane de cazuri studiate (indivizi recenzaţi). Recensămintele impun costuri foarte mari care pot fi acoperite doar de guverne sau agenții mari de colectare a datelor. De aceea, majoritatea recensămintelor pot culege un număr relativ mic de informații și, în consecință, vor dispune de mai puține variabile decât sondajele de opinie. În recensăminte vom identifica mai degrabă informaţii socio-demografice şi economice, şi nu informaţii cu privire la atitudinile, comportamentele, relaţiile cu alţi indivizi etc.

În ultimii ani se discută tot mai multe despre conceptul de *big data*. Big data reprezintă un tip aparte de seturi de date folosite în studii cantitative care colectează informaţii despre un număr foarte mare de cazuri. Multă lume foloseşte cardul bancar pentru a face cumpărături sau păstrează legătura cu prietenii, cunoscuţii sau chiar se informează din reţelele de social media. De exemplu, în anul 2021 în România erau active 19,570,208 carduri care aveau cont bancar ataşat, și au fost efectuate peste 1340

de milioane de tranzacţii cu cardul[3]. Conform altor date statistice, în România, în August 2022 erau aproape 13 milioane de utilizatori ai platformei de socializare Facebook[4], iar conform Eurostat mai mult de 34% dintre români au folosit platforme de social media cel puţin o dată pe săptămână[5]. Există estimări că motorul de căutare al Google monopolizează traficul generat de aceste căutări în proporţie de peste 70% în unele ţări sau chiar peste 90% în alte ţări.[6] Important de reținut este faptul că toate aceste acțiuni pe care noi le realizăm au în spate un număr imens de date care sunt stocate și folosite ulterior în diverse scopuri. Companii precum băncile, procesatorii de plăți cu cardul, furnizorii de platforme social media, sau companiile care gestionează motoarele de căutare pe internet au acces la un nivel extrem de complex de informaţii despre indivizi, pe care acum treizeci de ani nu ni-l puteam imagina ca fiind posibil. Multora ni s-a întâmplat ca după ce am avut o discuţie cu un prieten despre, spre exemplu, unde ne-am dori să ne petrecem vacanţa vara viitoare, să constatăm că pe browserul telefonului primim reclame care au ca obiect tocmai o vacanţă în locul despre care tocmai am discutat cu prietenul nostru. Utilizarea zilnică a telefonului *smart*, a televizoarelor *smart*, a frigiderului sau prizei *smart* produce nenumărate informaţii cantitative care măsoară diverse comportamente, atitudini, preferinţe, dorinţe pe care le manifestăm sau pe care le exteriorizăm. Aceste date complexe sunt folosite de companiile sau guvernele care le colectează pentru a construi profiluri de comportament individual sau de grup, pentru utilizatorii şi clienţii lor, respectiv pentru cetăţenii lor. Analiza acestor seturi foarte mari şi complexe de date presupune o serie de metode statistice de management al datelor şi

---

[3] Conform datelor publicate de Banca Centrală Europeană în seriile de date https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=169.PSS.A.RO.F000.I1A.Z00Z.NT.X0.20.Z0Z.Z, respectiv https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=169.PSS.A.RO.S101.I11.Z00Z.NT.X0.20.Z0Z.Z. Accesate ultima dată la 29 octombrie 2022.

[4] Conform Statista: https://www.statista.com/statistics/1178634/romania-number-of-facebook-users. Accesat ultima dată la 29 octombrie 2022.

[5] Sursa datelor Eurostat: https://ec.europa.eu/eurostat/databrowser/view/ilc_scp13/default/table. Accesat ultima dată la 29 octombrie 2022.

[6] Conform Statista: https://www.statista.com/statistics/220534/googles-share-of-search-market-in-selected-countries. Accesat ultima dată la 29 octombrie 2022.

de analiză a acestora. Datele nu sunt selectate aleatoriu, dar analiza acestora este subsumată principiilor şi standardelor de analiză cantitativă.

Cercetările calitative sunt acele studii în care de cele mai multe ori avem un număr mic de cazuri şi foarte multe variabile. Studiile antropologice, etnografice, sau chiar şi cele în care elemente istorice sunt comparate, pot fi mai degrabă etichetate ca studii calitative. Juxtapunerea față de cercetările cantitative este asumată deseori și de autorii acestor tipuri de studii, la fel cum, de altfel, această separație este transmisă deseori în specializările din domeniul sociologic: ești fie cantitativist, fie calitativist. *Tertium non datur*! Încercarea de a identifica o cale de mijloc prin care studiile comparative primesc o metodologie proprie (Ragin 1987; Rihoux 2003; Schneider și Wagemann 2012; Dușa 2014) numită *qualitative comparative analysis* (QCA), nu a reușit să creeze punțile de legătură între tradiția cantitativă și cea calitativă, creând în schimb ceea ce pare mai degrabă o a treia metodă, separată. Există puncte de vedere conform cărora studiile cantitative și cele calitative au la bază aceleași reguli standard de identificare a cazurilor și de confirmare a concluziilor (G. King, Keohane, și Verba 2000; Brady și Collier 2003). În contrast, alte puncte de vedere contrapun datele cu număr mare de cazuri, raționamentul inferențial și concluziile generalizante ale studiilor cantitative, studiilor calitative caracterizate prin date culese de la un număr relativ mic de cazuri, raționament interpretativ și concluzii valide doar pentru cazurile studiate (Goertz și Mahoney 2012).

Totuşi, vom întâlni şi studii în care cercetătorii folosesc o combinaţie de tehnici și instrumente cantitative și calitative. Acest tip de cercetare combinată o numim cercetare mixtă sau cercetare multi-metodă. De exemplu, într-un studiu pe tema clientelismului electoral din alegerile din România, pe care Aurelian Muntean l-a efectuat în perioada 2013-2016 cu studenţii săi din Facultatea de Ştiinţe Politice din SNSPA (Mares și Muntean 2015; Mares, Muntean, și Young 2016; Mares, Muntean, și Petrova 2017; 2018), după pretestarea pe 1000 de respondenți a unor instrumente specifice (prin focus grupuri, întrebări, chestionare, congruenţă a cazurilor), acesta a cules date empirice în sondaje regionale şi naţionale cu eșantioane cumulate la peste 7000 de respondenţi, 400 de interviuri aprofundate, monitorizare mass-media şi alte date secundare (de exemplu, materiale de campanie electorală). Așadar, acest studiu multianual a combinat mai multe metode, tehnici și instrumente pentru asigurarea

validității empirice și de construct[7] dar și pentru a obține informații variate capabile să permită înțelegerea aprofundată și fundamentarea temei cercetate. Acest studiu al clientelismului din alegerile din România a folosit un design de cercetare mixt, care a combinat cercetarea ce foloseşte instrumente cantitative cu una care foloseşte instrumente calitative.

Dacă studiile calitative includ un număr relativ redus de cazuri, fără să existe un consens în ceea ce priveşte o anumită limită numerică, un statut aparte în studiile calitative pare a fi acordat studiilor de caz. După cum le spune numele, studiile de caz analizează un singur caz, de exemplu o țară, sau o organizaţie. Există însă o perspectivă extrem de interesantă potrivit căreia, dacă scopul este acela al explicării cazului, nu doar al descrierii acestuia, un asemenea demers nu poate fi făcut pe un singur caz. Pentru a înţelege această dilemă ontologică, vom exemplifica folosind un studiu ipotetic care analizează şomajul în România. Culegerea unor informaţii cu privire la rata şomajului în România în anul 2022 nu ne va ajuta să răspundem la întrebarea „cum explicăm şomajul în România?". Nu vom reuși să aflăm răspuns la această întrebare chiar dacă am culege informaţii despre alte caracteristici precum nivelul compensaţiilor de ajutor de şomaj, rata inflaţiei, costurile de creditare, recesiunea economică, atitudinea angajaţilor faţă de păstrarea unui loc de muncă, sau adoptarea unor tehnologii care înlocuiesc forţa de muncă umană. Pentru a putea explica acest fenomen, John Gerring (2006) ne propune o schimbare a modului în care înțelegem studiul de caz. Vom discuta pe larg despre aceste probleme în capitolul 2.

[7] Vom discuta în secțiunea 3.1.1 despre validitatea și fidelitatea instrumentelor de cercetare.

1.4. Etica în cercetarea cu subiecți umani: de la Micul Albert la GDPR

Reproductibilitatea studiilor, cu precădere a celor cantitative, reprezintă un pilon fundamental pentru organizarea și avansul cunoașterii științifice. Reproductibilitatea poate să prevină situații în care concluzii care par a fi fundamentate științific sunt dovedite pe baza activității științifice de reproducere metodologică, ca fiind false, uneori chiar contrafăcute. Un caz devenit celebru este cel al studiului care părea că a identificat o legătură între vaccinul împotriva rujeolei și apariția autismului. Acest studiu care a încurajat opiniile anti-vacciniste, coordonat de medicul britanic Andrew Wakefield (Wakefield et al. 1998), a produs efecte devastatoare pentru campaniile de vaccinare împotriva rujeolei, rubeolei și oreionului, ducând la creșterea numărului de cazuri de rujeolă. Studiul a fost criticat de alți cercetători (Taylor et al. 1999; Farrington, Miller, și Taylor 2001; Madsen et al. 2002; Davidson 2017) dar, deși a fost retras de revista Lancet pe motive metodologice (selecția subiecților) și de etică (lipsa aprobării comitetelor de etică a cercetării), a produs efecte negative pe termen lung pentru eficiența politicilor de sănătate pentru imunizarea populației la diferite boli contagioase.

În România, pe fondul unor opinii publice privind efectele adverse ale vaccinului împotriva rujeolei, a scăzut încrederea în eficiența acestora, dovedită anterior științific, iar numărul de cazuri de rujeolă a crescut de la 7 cazuri raportate în 2015 la 20.204 cazuri raportate în 2020.[8] Aceste opinii, pe care le cunoaștem și sub numele de anti-vaccinism, au fost susținute de studii științifice publicate în reviste recunoscute și care păreau că oferă dovezi științifice foarte solide pentru concluziile

[8] Conform rapoartelor publicate de Centrul Național de Supraveghere și Control al Bolilor Transmisibile, din României, de exemplu aici: https://www.cnscbt.ro/index.php/informari-saptamanale/rujeola-1/1871-situatia-rujeolei-in-romania-la-data-de-17-07-2020/file și de Centrul European pentru Controlul și Prevenția Bolilor https://www.ecdc.europa.eu/sites/default/files/documents/measles-2019-aer.pdf Accesate ultima dată la 1 noiembrie 2022.

lor. Caracterul fals al concluziilor a fost dovedit prin reproducerea analizelor din cadrul acestor studii.

Recent, aceste efecte de diminuare a încrederii în descoperirile ştiinţifice din domeniul imunizării prin vaccinare și de creștere a explicațiilor non-științifice conspiraționiste au fost observate și în cazul campaniilor de vaccinare împotriva COVID-19 (Prieto Curiel și González Ramírez 2021; Stoica și Umbreș 2021). Așadar, publicarea rezultatelor și caracterul reproductibil al analizelor științifice, stau la baza avansării cunoaşterii în toate domeniile ştiinţifice şi pot asigura verificarea şi testarea validităţii concluziilor. În lipsa lor, avansul științific ne-ar conduce pe căi închise (Jussim, Krosnick, și Stevens 2022).

Problemele precum cea exemplificată mai sus sunt tratate prin regulile de etică în cercetare, dar şi de reglementări legale. Dacă problemele de etică ştiinţifică sunt abordate mai degrabă la nivel instituţional, fie de universitatea unde cercetătorul îşi desfăşoară activitatea, fie de organizaţia finanţatoare a cercetării, protecţia indivizilor (de exemplu subiecţi ai culegerii de informaţii / date) este asigurată de legislaţia internaţională sau naţională, cum ar fi Declaraţia Universală a Drepturilor Omului[9], din 1948, Convenţia Europeană a Drepturilor Omului[10], din 1950, sau GDPR (General Data Protection Regulation) a Uniunii Europene[11], din 2016. Această reglementare de la nivelul Uniunii Europene, transpusă în legislaţie naţională creează un cadru de obligaţii pentru organizaţiile care colectează informaţii cu caracter personal (de exemplu, atunci când un spital internează pacienţi sau atunci când include pacienţi în studii clinice, trebuie să asigure nu doar informarea acestora, ci şi protecţia activă a drepturilor acestora) dar eşuează în a extinde protecţia indivizilor la situaţii ce pot avea implicaţii etice şi morale. Problemele etice şi de protecţie a indivizilor subiecţi umani în studii ştiinţifice nu sunt, însă, noi. Probleme etice au fost ridicate în prima jumătate a secolului XX cu privire la efectele negative produse de utilizarea ştiinţei

[9] Disponibilă aici: https://www.un.org/en/about-us/universal-declaration-of-human-rights Accesat ultima dată la 1 noiembrie 2022.

[10] Disponibilă aici: https://www.echr.coe.int/Documents/Convention_ENG.pdf Accesat ultima dată la 1 noiembrie 2022.

[11] Disponibilă aici: https://gdpr.eu/what-is-gdpr/ Accesat ultima dată la 1 noiembrie 2022.

pentru a studia indivizi, mai ales în domenii precum medical, psihologic, sociologic sau cel militar.

Un prim asemenea studiu este cel numit Studiul Micului Albert. Pe scurt, doi psihologi, John Watson şi Rosalie Rainer, de la Universitatea Johns Hopkins, în 1919 au folosit un experiment pe un subiect uman (copil de 9 luni), pentru a testa mai multe ipoteze, cum ar fi aceea că indivizii îşi construiesc fricile fiind condiţionaţi de experienţele pe care le au și că, odată acumulate, aceste artefacte cognitive pot fi transferate ulterior altor evenimente similare. Experimentul a fost documentat video, iar înregistrarea stimulilor folosiţi de cei doi cercetători a supravieţuit într-un film disponibil şi astăzi.[12] Experimentul a fost contestat în comunitatea ştiinţifică pe motive metodologice: a avut un singur subiect experimental şi nici un subiect de control, dar a fost contestat şi pe motive etice, deoarece, subiectul a dobândit fobie la animalele cu blană. Efectele negative asupra subiectului uman al studiului au crescut gradul de conştientizare a necesităţii impunerii unor limite de ordin etic în desfăşurarea cercetărilor (B. Harris 1979).

Un al doilea studiu, important pentru modificările radicale pe care le-a determinat în ceea ce priveşte principiile etice de cercetare pe subiecţi umani, este reprezentat de experimentele medicale din Al Doilea Război Mondial. Mii de prizonieri de război au primit injecţii şi au fost subiecţi ai unor tratamente experimentale medicale inimaginabile în prezent: de exemplu, injecţii cu benzină, viruşi vii/activi, otravă, sau scufundări în apă cu gheaţă. Sub pretextul unor motivaţii aparent ştiinţifice, medici şi cercetători susţinători ai regimului nazist au folosit copii gemeni ca subiecţi umani consideraţi a fi ideali pentru experimente medicale şi psihologice în care subiecţii din grupul de tratament sunt identici cu participanţii din grupul de control. Una dintre justificările invocate de aceşti cercetători, dar şi de finanţatorii lor politici, a fost necesitatea de a efectua cercetări pe subiecţi umani pentru a putea anula avantajul medical al Aliaţilor, cu importante efecte militare, ca urmare a descoperirii penicilinei în 1940 în Marea Britanie şi a utilizării ei cu succes pe soldaţii răniţi pe front (Weindling 2004).

[12] Înregistrarea este disponibilă aici: https://www.youtube.com/watch?v=kGn_EoHiOvc Accesat ultima dată la 1 noiembrie 2022.

Procesul de la Nürnberg, (1945-1946), care a pus sub acuzare nu doar liderii politici şi militari, ci şi medicii, cercetătorii şi asistenţii acestora, a definit într-un mod categoric noi reguli morale, etice şi legale, pe care orice cercetare ştiinţifică care include subiecţi umani ar trebui să le îndeplinească. Astfel, a fost produs un corp de reguli clare, Codul de la Nürnberg, care, împreună cu Declaraţia Universală a Drepturilor Omului, din 1948, au constituit primele reguli de stabilire a unui standard internaţional pentru efectuarea de cercetări pe subiecţi umani. Codul conţine un set de 10 reguli[13] pe care cercetătorii care fac studii folosind subiecţi umani trebuie să le respecte pentru a asigura onestitatea, încrederea şi respectul faţă de indivizi. Acest cod de etică a fost ulterior preluat de comisii de etică instituţională din universităţile, centrele de cercetare şi finanţatorii cercetărilor ştiinţifice, dar şi de codurile de etică ale asociaţiilor profesionale (de exemplu Asociaţia Americană de Psihologie).

---

[13] Oferim aici o traducere rezumată a principalelor precizări din regulile de etică ale Codului de la Nürnberg (1949, 2-The Medical Case:181–82):

„1. Este absolut esențială existența consimțământului voluntar al subiecților umani. Persoana implicată ar trebui să aibă capacitatea juridică de a-și da consimțământul; trebuie să fie într-o situație în care să își poată exercita liberul arbitru, fără intervenția vreunui element de forță, fraudă, înșelăciune, constrângere, sau orice alte forme ulterioare de constrângere sau coerciție; și ar trebui să aibă suficiente cunoștințe și să înțeleagă suficient de bine elementele subiectului experimentului, pentru a-i permite să ia o decizie avizată, pe care înțelege.

2. Experimentul trebuie să producă cunoștințe ce pot fi generalizate, care nu ar putea fi altfel obținute în niciun alt mod și care nu sunt de natură aleatorie și inutilă.

3. Experimentele pe animale ar trebui să preceadă experimentele pe subiecți umani.

4. Trebuie evitate toate suferințele și rănile fizice și mentale inutile.

5. Nu trebuie efectuat niciun experiment dacă există motive să credem că va avea loc decesul sau vătămarea invalidantă.

6. Trebuie să se facă pregătiri corespunzătoare și să se asigure facilități adecvate pentru a proteja subiecții experimentali chiar și împotriva unor posibilități îndepărtate de rănire, invaliditate sau deces.

8. Experimentul trebuie să fie condus numai de persoane calificate din punct de vedere științific. Cel mai înalt grad de competență și de grijă ar trebui să fie necesar în toate etapele experimentului de către persoanele care conduc sau se implică în experiment.

9. Pe parcursul experimentului, subiectul uman ar trebui să aibă libertatea de a se retrage din experiment, dacă a ajuns la o stare fizică sau psihică în care continuarea experimentului i s-ar părea imposibilă.

10. Pe parcursul desfășurării experimentului, omul de știință responsabil trebuie să fie pregătit să încheie experimentul în orice etapă a acestuia, dacă are motive să creadă, în exercitarea bunei credințe, a competenței superioare și a discernământului atent care i se cer, că o continuare a experimentului este susceptibilă a provoca rănirea, invaliditatea sau decesul subiectului experimental.”

Două alte studii, din domeniul ştiinţelor sociale (psihologie), efectuate în 1963 şi 1971, au ridicat alte semne de întrebare cu privire la respectarea de către cercetători a regulilor etice de studiere a indivizilor umani. Stanley Milgram (1963) a efectuat un studiu experimental, în care încerca să identifice explicaţii pentru obedienţa indivizilor la autoritate. În cadrul studiului, acesta a folosit mecanisme de învăţare prin experimente care simulau electrocutarea sub supravegherea unei persoane cu autoritate ştiinţifică. Studiul a fost criticat pentru că a încălcat reguli etice privind informaţiile incomplete date subiecţilor şi lipsa de monitorizare a efectelor adverse post-experimentale asupra subiecţilor (Baumrind 1964). Răspunsul lui Milgram arată importanţa înţelegerii de către cercetător a cerinţelor etice de protecţie a : „*Prin declaraţiile şi acţiunile lor, subiecţii au arătat că au învăţat multe lucruri şi mulţi s-au simţit mulţumiţi că au luat parte la o cercetare ştiinţifică pe care o considerau importantă*" (Milgram 1964). Un alt studiu experimental cu subiecţi umani, realizat de psihologul Philip Zimbardo[14], a urmărit explicarea factorilor care duc la atitudinile agresive ale indivizilor. Beneficiind de finanţare de la Biroul pentru Cercetare Navală a Armatei SUA, Zimbardo a folosit o metodă experimentală într-un mediu controlat, pentru a simula un cadru instituţional represiv. În acest laborator au fost integraţi şi izolaţi 22 de subiecţi umani, timp de o săptămână. Aceste studii experimentale nu au respectat regulile etice privind studiile pe subiecţi umani, dar nici condiţiile metodologice elementare privind designul experimental, pe care le vom discuta în secţiunea dedicată din capitolul următor.

Un alt studiu experimental, determinant pentru configurarea condiţiilor legale de etică în cercetare în secolul XX, este reprezentat de studiul despre sifilis din Tuskegee, Alabama. A fost desfăşurat de United States Public Health Service între 1930 şi 1972. Pe scurt, studiul a inclus 600 de persoane, doar Afro-Americani, împărţiţi în două grupuri: 400 bolnavi de sifilis, 200 fără sifilis. Acest studiu a marcat mai multe probleme etice în desfăşurarea sa: recrutarea s-a făcut fără consimţământul informat al subiecţilor; acestora li s-a spus că este o campanie în care vor primi tratament

[14] O colecție de materiale video și documente din Colecția de Documente Philip Zimbardo, este disponibilă aici: http://www.oac.cdlib.org/findaid/ark:/13030/kt7f59s371/dsc/?dsc.position=2501#aspace_ref68_f5j Accesat ultima dată la 1 noiembrie 2022.

medical în mod gratuit; deşi penicilina a fost folosită cu succes pe subiecţi umani bolnavi de sifilis încă din 1940, experimentul Tuskegee nu a fost oprit iar subiecţii nu au fost informaţi de existenţa noului tratament descoperit (Brandt 1978). În urma unor investigaţii de presă studiul a fost oprit în 1972, determinând înfiinţarea unei comisii de analiză din partea Departamentului de Sănătate al SUA, care a produs recomandări de întărire a reglementarii pentru cercetarea pe subiecţi umani. Revolta opiniei publice şi a unor politicieni faţă de încălcarea eticii în cercetare a condus la adoptarea National Research Act a SUA în 1974 şi la publicarea în 1979 a Raportului Belmont[15] (Reverby 2009). Acesta constituie şi astăzi principala sursă de proceduri legale şi etice pe baza cărora sunt evaluate proiectele de cercetare din universităţile americane în comitetele instituţionale de etică.

Prevederile acestui raport pot constitui o sursă utilă de pentru amendarea Legii nr 206 / 2004 din România, privind buna conduită în cercetarea ştiinţifică, care eşuează în forma actuală în a reglementa problemele etice inerente studiilor ce implică subiecţi umani (de exemplu cele din medicină sau ştiinţe sociale), sau a Ordinului MEC 4402/2005 de constituire a comisiilor de etică din universităţi (care nici măcar nu menţionează vreun principiu de etică a cercetărilor pe subiecţi umani, pe care aceste comisii să le codifice şi să le urmărească). Cercetarea ştiinţifică care include subiecţi umani nu se face doar în universităţi, ci şi ci în alte institute publice sau în

[15] Pe scurt, Raportul Belmont, disponibil aici https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html (Accesat ultima dată la 1 noiembrie 2022), prevede trei principii importante: respectul pentru persoane, beneficierea de pe urma studiilor, justiția. În mod specific aceste principii se traduc în câteva reguli clare pentru etica cercetărilor ce includ subiecți umani. Persoanele trebuie tratate ca indivizi autonomi (ce pot lua singuri decizii), prin urmare, acestea trebuie incluse în studii pe baza consimțământului informat. Persoanele cu autonomie scăzută (copiii sub 12 ani, persoanele cu dizabilități, cu probleme neuropsihice, fetușii umani, femeile însărcinate, nou-născuții, prizonierii, deținuții) au dreptul la protecție suplimentară. Raportul prevede obligativitatea evitării influențării deciziei prin coerciție (amenințare) sau prin plată exagerată (recompensa pentru participarea la studiu). În ce privește beneficierea de pe urma studiilor, acestea din urmă trebuie să asigure un echilibru între costurile pentru subiecți și beneficiile obținute de aceștia sau de societate, în urma participării la studiu. Persoanele nu trebuie rănite fizic sau psihic. Posibilele beneficii ale studiilor trebuie maximizate, iar posibila suferință trebuie minimizată. Nu în cele din urmă, cercetările pe subiecți umani trebuie să respecte criterii de selecție a participanților și a grupurilor. Indivizii sau grupurile de indivizi trebuie tratate corect și echitabil în ce privește costurile și beneficiile pe care ei le suportă. Grupurile nu trebuie incluse în studiu doar pentru ca sunt disponibile, au o poziție de inferioritate, sau sunt vulnerabile, ele trebuie incluse doar pe motive care țin de problema studiată.

departamente de cercetare ale unor companii private. Prin urmare, necesitatea și utilitatea reglementării acestor probleme în legi cu aplicabilitate națională și acoperire generală, este mai mult decât evidentă.[16]

În științele sociale avem circumstanțe variate în care subiecții umani trebuie protejați suplimentar. Pe lângă principiile enunțate mai sus, derivate din evenimente decisive pentru dezvoltare regulilor de etică științifică, studiile cantitative, mai precis cele care realizează anchete sociologice cum ar fi sondajul de opinie, trebuie să asigure protecția respondenților din punct de vedere etic, moral și legal. Institutele europene care colectează asemenea date, cum ar fi de exemplu institutele de sondare din România, se supun, de regulă, standardelor etice și recomandărilor metodologice stabilite de ESOMAR (European Society for Opinion and Market Research).[17]

Consimțământul informat pentru participarea în studiile sociale trebuie să fie obligatoriu, la fel ca în studiile clinice medicale. Prin utilizarea pe scară tot mai largă a designurilor de cercetare experimentale, științele sociale din România vor fi nevoite să își asume adoptarea standardelor etice în cercetările pe subiecți umani. În sondajele de opinie, dar și în tehnicile folosite în studiile calitative sau mixte (interviul aprofundat, focus grupul, studiile etnografice etc.), respondenții trebuie să primească o informare anterior începerii interviului. Prin acest consimțământ informat, respondenții sunt informați nu doar cu privire la modalitatea și motivul selectării lor în studiu, la caracterul anonim al răspunsurilor pe care le vor da, la metodele prin care se asigură anonimitatea, dar și asupra acceptului pentru înregistrarea video, audio sau în scris a răspunsurilor lor. De exemplu, într-un studiu pe angajații dintr-o companie, respondenții nu trebuie selectați de către management, ci doar de către cercetători, altfel, existând riscul ca participanții la studiu să nu poată efectiv refuza participarea la acest studiu de frica conducătorilor ierarhici. De asemenea, respondenții sunt informați și cu privire la modalitățile de anonimizare a participării lor, a răspunsurilor

[16] A se vedea și analiza detaliată a problemelor de etică din spațiul universitar românesc și recomandările pertinente de etică academică, oferite de Emanuel Socaciu (2018) și colaboratorii săi.

[17] Disponibile aici: https://esomar.org/codes-and-guidelines Accesat ultima dată la 1 noiembrie 2022.

pe care le oferă intervievatorilor, a modului în care informaţiile pe care le oferă vor fi analizate şi utilizate ulterior, a modului în care aceste informaţii vor fi stocate.

Chiar dacă, în teorie, subiecții umani nu sunt supuşi unor potenţiale riscuri la adresa sănătăţii sau integrităţii fizice precum cele din studiile clinice medicale, în ştiinţele sociale aceștia pot fi integraţi în studii empirice, cu design observaţional sau experimental, care pot crea riscul unor suferințe fizice sau psihice în timpul desfăşurării studiului sau ulterior. Prin urmare, orice potenţial risc sau suferinţă la adresa subiecţilor umani (respondenţi sau persoane intervievate, sau participanţilor la un experiment) trebuie evitată, iar atunci când aceasta nu poate fi evitată, nici redusă participanţii trebuie informaţi suplimentar în consimţământ.

Respondenţilor trebuie să li se ofere căi de comunicare ulterioară cu cercetătorii, mai ales atunci când studiul poate avea un impact imprevizibil ulterior interviului. De exemplu, un asemenea impact apare atunci când studiul se face pe o temă sensibilă, iar respondenţii ar putea manifesta efecte psihologice ulterioare participării la studiu (de exemplu, după un studiu despre ritualuri funerare, sau despre experienţele de agresiune). Un asemenea efect negativ poate apărea și când răspunsurile lor ar putea fi folosite de autorităţi sau alte persoane pentru a-i trage la răspundere sau a-i ameninţa (de exemplu, după participarea lor la un studiu despre clientelismul electoral). Nu în ultimul rând, participanţilor în studiile din ştiinţele sociale trebuie să li se permită să se retragă din studiu la orice moment după începerea acestuia, fără teama de a fi penalizaţi (de exemplu, fără reducerea recompensei de participare la discuţiile de tip focus grup), sau de a fi pusă presiune din partea altor persoane sau instituţii ierarhice (de exemplu, într-un sondaj sau un interviu aprofundat cu angajaţii dintr-o companie, reprezentanţii managementului companiei nu trebuie să ia parte sau să monitorizeze procesul de culegere a informaţiilor prin chestionar, interviu sau focus grup, deoarece angajaţii participaţi la studiu s-ar putea simţi presaţi să nu se retragă din studiu înainte de finalizarea acestuia).

În ceea ce priveşte raportarea rezultatelor studiilor şi acestea se supun unor reguli etice clare. Să ne referim pentru început la utilizarea datelor rezultate în urma cercetării empirice. Astfel, cercetătorii trebuie să respecte dreptul de proprietate intelectuală asupra informaţiilor empirice pe care alţi cercetători le-au cules. Acest

principiu este integrat în ceea ce numim plagiat. De exemplu atunci când un cercetător face un studiu de teren în care colectează date, cantitative sau calitative, prin sondaje, interviuri, focus grupuri, alte instrumente de observaţie, atât din punct de vedere legal, cât şi moral, deoarece cercetătorul este cel care a cules informaţiile, le-a conceput şi le-a prelucrat în mod sistematic pe baza unei metodologii, rezultatul fiind modalitatea în care apar ele agregate (de exemplu o bază de date) ca produs al demersului teoretic, metodologic şi empiric, le-a dat o formă ce are un caracter original, acesta are dreptul de a permite sau nu altor cercetători sa-i folosească rezultatul cercetării empirice.

Prin bază de date putem înţelege, de exemplu, observaţii, fapte, informaţii, date, sunete, imagini, care sunt prelucrate în mod original pe baza unor principii metodologice, rezultând o configuraţie care nu este predeterminată extern (cum ar fi cazul unei baze de date preexistente care predetermină modul în care aceasta ar putea fi alterată prin transformarea datelor conţinute) ci este rezultatul ideilor şi activităţii cercetătorului. Ea este rezultatul definirii conceptelor şi variabilelor, verificarea validităţii acestora, selectarea respondenţilor, intervievarea acestora, centralizarea răspunsurilor lor în baza de date prestabilită de către cercetător, verificarea calităţii bazelor de date.

Preluarea și manipularea acestor informații fără acceptul sau autorizarea dată de cercetător (de exemplu, prin publicarea acestora, prin acord personal etc.) constituie nu doar o abatere de la dreptul de proprietate intelectuală[18], ci mai ales o abatere de la principiile de etică în cercetare. Mai mult, dacă aceste date empirice sunt prezentate ca și când ar fi fost culese de autorul lucrării publicate, fără să aibă acceptul utilizării lor dat de cercetătorul care le-a cules și le-a agregat / centralizat într-o bază de date sau alte forme, și fără să indice sursa exactă a datelor empirice, desigur, putem afirma că și aceste fapte constituie abateri de la normele etice de cercetare, putând fi încadrate în categoria plagiatului și a încălcării drepturilor de proprietate intelectuală. Investiția cantitativă și calitativă pe care cercetătorii o fac în culegerea datelor empirice trebuie respectată de toată comunitatea științifică, inclusiv prin recunoașterea transparentă a calității acestora de autori ai datelor empirice (baze de date, interviuri

[18] A se vedea prevederile Directivei nr. 96/9/CE, și ale Legii nr. 8/1996.

etc.), publicațiilor și oricăror produse ale demersului științific original. Prin urmare, preluarea unor date, texte sau teorii, culese sau elaborate de alți autori, și prezentarea acestora ca și când ar fi culese, produse sau elaborate de către autorul publicației, constituie plagiat. Sursa acestora și recunoașterea muncii altora trebuie indicată în mod expres, folosind formele acceptabile în comunitatea științifică, care includ, dar nu sunt limitate la, citarea și raportarea în lista de referințe.

Raportarea cercetării empirice care include subiecţi umani (de exemplu în sondaje, studii etnografice, studii antropologice, studii clinice, experimente etc.) trebuie să asigure respectarea demnităţii subiecţilor ale căror informaţii acordate cercetătorilor sunt prezentate în publicaţiile ştiinţifice. Aceste utilizări ale informaţiilor trebuie să asigure protecţia sub forma anonimizării, dacă aceasta este cerută de subiecţi, dacă publicarea lor şi asocierea cu un subiect poate aduce acestuia orice formă de discriminare, repercusiuni sau pedepse sociale, morale sau legale. Cercetătorii sunt, aşadar, principalii responsabili pentru asigurarea acestor condiţii de anonimizare şi confidenţialitate a datelor personale. Protecţia acestor subiecţi nu se încheie odată cu culegerea datelor ci continuă prin securizarea datelor şi respectarea confidenţialităţii acestora, securizarea accesului la date. Editorii publicaţiilor (reviste de specialitate sau cărţi) trebuie să participe şi ei activ în acest proces de protecţie a subiecţilor umani prin asigurarea confidenţialităţii informaţiilor personale obţinute în cercetarea empirică sau obţinerea consimţământului informat pentru informaţiile ce vor fi publicate şi care ar putea permite identificarea subiecţilor; dar şi în procesul de protecţie a drepturilor autorilor datelor empirice incluse în publicaţiile pe care le editează, de depistare a plagiatului şi de asigurare a mecanismelor de verificare a validităţii studiilor empirice (de exemplu prin publicarea datelor ce asigură reproductibilitatea analizelor şi certifică originalitatea acestora).[19] În plus, anonimizarea trebuie să asigure comunitatea ştiinţifică şi publicul că datele nu sunt falsificate şi că designul de cercetare nu a produs efecte negative asupra subiecţilor umani.

[19] A se vedea și prevederile Comitetului pentru Etica Publicațiilor: https://publicationethics.org

2. Design de cercetare

În acest capitol vom discuta despre câteva probleme metodologice care stau la baza proiectării cercetării de teren. Designul de cercetare reprezintă totalitatea elementelor de care ţinem cont pentru a selecta cazurile studiate, a asigura controlul cauzelor alternative pentru efectul pe care dorim să-l explicăm sau a interpreta informaţiile detaliate ale cazurilor studiate. Diferitele tipuri de raţionamente, date, instrumente şi utilizări ale studiilor cantitative şi calitative conduc la strategii diferite privind culegerea datelor şi analizarea lor. În acest capitol vom discuta despre condiţiile de proiectare a culegerii informaţiilor pentru cazurile incluse în studiile observaţionale. Vom discuta despre constrângerile metodologice de selecţie a cazurilor şi tehnicile de reducere sau multiplicare a cazurilor incluse în aceste studii. De asemenea, vom discuta despre studiile experimentale şi condiţiile pe care acestea le îndeplinesc pentru a ne fi de folos în explicarea efectelor (fenomenelor studiate).

2.1. Unități de analiză și unități de observație: selecția cazurilor

În această secţiune vom discuta despre modalitatea în care înţelegem diferenţele între diverse niveluri la care se află lucrurile pe care le studiem, despre diferenţele între unităţile de analiză şi cele de observaţie. Pentru început, să clarificăm ce sunt unităţile de analiză şi cele de observaţie. Ne vom folosi de un exemplu: vrem să studiem modul în care stabilitatea guvernamentală variază între diverse forme de regim politic şi modul în care aceasta este explicată de mărimea coaliţiilor. Întrebarea de cercetare ar putea fi ce determină stabilitatea guvernamentală în democraţii? Pornind de la concluziile altor studii, ne aşteptăm ca un număr mare de partide care intră în componenţa guvernului să reducă stabilitatea acestuia (Krauss 2018). Dacă într-o coaliţie sunt mai multe partide, cresc şansele ca între acestea să apară neînţelegeri pe marginea politicilor publice, a alocării bugetare, a distribuirii funcţiilor

de conducere în instituţiile publice sau în ceea ce priveşte asumarea responsabilităţii în faşa alegătorilor, pentru deciziile luate, mai ales în momente de criză. În guvernele monocolore ne aşteptăm ca aceste neînţelegeri să fie mai uşor tranşate de conducerea partidului care formează singur guvernul.

Pentru a cerceta aceste probleme am putea include în studiu ţări cu regim semi prezidenţial, dar şi cu regim parlamentar însă nu ţări cu regim prezidenţial, unde durata mandatului cabinetului este fixă, egală cu mandatul preşedintelui, unicul cap al executivului. Prin urmare, am putea include cazuri precum România, Polonia, Italia şi Franţa. Observaţiile de care avem nevoie în acest studiu pentru a răspunde la întrebarea de cercetare şi pe care le vom culege prin intermediul variabilelor ar putea fi: numărul guvernelor şi durata de viaţă a fiecăruia; numărul partidelor care negociază formarea unei coaliţii; numărul partidelor care formează coaliţia; numărul partidelor care sprijină în parlament coaliţia guvernamentală fără a avea membri în guvern; tipul de regim; orientarea ideologică a partidelor etc.

Trebuie să observăm că unele dintre aceste observaţii pot părea ca fiind măsurători ale cazurilor pe care le analizăm, în vreme ce altele măsoară unităţi de la un alt nivel de analiză. Prin urmare, trebuie să diferenţiem între unităţile de analiză şi unităţile de observaţie. O discuţie utilă este prezentată şi de Earl Babbie (2010, 146–53). Unitatea de analiză (de exemplu, un caz) este la nivelul la care tragem concluziile, nivelul la care facem inferenţele bazându-ne pe date. Unitatea de analiză este identificată la nivelul indicat de întrebarea de cercetare. O putem identifica prin întrebările: ce studiem? sau cine este studiat? Unitatea de observaţie este la nivelul la care colectam datele, prin urmare, ea poate fi o subunitate sau o caracteristică a unităţii de analiză. De regulă, unitatea de analiză este la un nivel mai ridicat de agregare decât unitatea de observaţie. Identificarea corectă a unităţilor de analiză şi a celor de observaţie, precum şi clarificarea acestora, ne poate salva mult timp şi resurse în momentul colectării datelor, dar mai ales după ce acestea au fost colectate.

Folosindu-ne de o reprezentare grafică tabelară a bazei de date pe care ne-am propune să o colectăm în acest studiu, am putea ilustra, conform Figurilor 2.1 și 2.2 de mai jos, diferenţa dintre unităţile de analiză şi cele de observaţie. Astfel, dat fiind că vom trage concluzii referitoare la ţări, acestea reprezintă nivelul nostru de analiză,

configurate ca variabilă țară-an (din 1990 până în 2022) (Figura 2.1); și este precizat pe rânduri, iar pe coloane avem specificate unitățile de observație, într-o bază de date care ar ilustra modalitatea în care putem colecta și organiza datele. Cu toate acestea, putem trage concluzii la nivelul de analiză al guvernelor, prin urmare, în baza noastră de date pe rânduri vom avea unitatea de analiză guvern-țară (Figura 2.2), iar pe coloane unitățile de observație precizate mai sus, în vreme ce fiecare celulă reprezintă o observație.

Figura 2.1 Unitate de analiză țară-an

tara-an	guvern	durata_de_viata	nr_partide_negociere	nr_partide_coalitie	regim
romania 1990					
romania 1991					
romania 1992					
polonia 1990					
polonia 1991					
polonia 1992					
italia 1990					
italia 1991					
italia 1992					
franta 1990					
franta 1991					
franta 1992					

Figura 2.2 Unitate de analiză guvern-țară

guvern-tara	an	durata_de_viata	nr_partide_negociere	nr_partide_coalitie	regim
romania guvern 1					
romania guvern 2					
romania guvern 3					
polonia guvern 1					
polonia guvern 2					
polonia guvern 3					
italia guvern 1					
italia guvern 2					
italia guvern 3					
franta guvern 1					
franta guvern 2					
franta guvern 3					

În cazul în care dorim să analizăm performanţa academică unităţile de observare ar putea fi studenţii, iar unităţile de analiză grupele de studiu, în cazul în care grupele de studiu sunt comparate. În cazul în care avem mai multe observaţii pentru fiecare student, aceştia pot fi unităţile de analiză iar notele la fiecare materie vor reprezenta unitatea de observaţie. Alternativ, studenţii pot fi atât unităţile de observaţie, cât şi unităţile de analiză, dacă ei sunt cei comparaţi şi tragem concluzii la acest nivel de dezagregare.

În situația în care dorim să explicăm comportamentul consumatorilor pe piața de carte din România, eventual comparativ cu toate celelalte țări din UE, în funcție de nivelul la care vom trage concluziile unitatea noastră de analiză poate fi reprezentată de cititori, sau de edituri, sau de sector economic, sau chiar țările. Unitatea de observaţie poate fi reprezentată de numărul de cărţi vândute, numărul de cărţi citite, tipul / genul de cărţi citite, momentul din ziuă în care este citită cartea, suma cheltuită luna trecută pentru achiziţia de cărţi, suma cheltuită pentru cărţi electronice, suma cheltuită pentru cărţi pe suport de hârtie, cât de des cumpără cărţi. În concluzie, unitatea de observaţie şi unitatea de analiză *pot fi identice*, dar nu e neapărată nevoie să fie aceleaşi; depinde de designul de cercetare. Designul nostru de cercetare poate să includă unităţi de la niveluri diferite de analiză. Putem analiza indivizi (studenţii,

cititorii de cărți), dar și grupuri de indivizi (grupă de studenți, cluburi ale cititorilor), organizații (facultăți, universități, edituri), structuri non-organizatorice (învățământ, sectorul producției și vânzării de carte), țări.

Putem observa că fiecare din aceste unități de analiză are un nivel diferit de individualizare și agregare: unele dintre ele sunt mai generale decât altele, iar unele cuprind în cadrul lor mai multe unități de analiză inferioare. De aceea, este util să diferențiem între diferite grade de agregare a unităților de analiză. Astfel, putem selecta unitățile de analiză la nivelul cel mai dezagregat, sau nivelul micro de analiză (indivizii). La acest nivel vom avea acele unități de analiză pe care cu greu le mai putem sparge în sub-unități componente, cel puțin nu în științele sociale, ci poate în științele medicale. Acest nivel de analiză este extrem de important pentru că permite cunoașterea la nivel granular a unor probleme din societate. Indivizii sunt sursa ultimă dar și cea mai importantă de informații pentru cercetătorii din științele sociale.

Al doilea nivel important de analiză pe care putem să ne plasăm unitățile analizate este nivelul mediu dezagregat, sau nivelul mezo / mediu de analiză. La acest nivel de analiză avem unități precum grupuri, organizații, structuri sociale, unități organice precum o stradă sau un cartier, unități administrative locale sau chiar companii. Acest nivel de analiză este important deoarece ne asigură furnizarea unor informații care, deși nu sunt atât de dezagregate, oferă informații despre caracteristici comune ale unui număr relativ mic de indivizi. Aceste unități de analiză pot fi analizate atât unitar, cât și dezagregate, folosind sub-unități de analiză, care se află la primul nivel de analiză, cel mai dezagregat.

Al treilea nivel de analiză este nivelul agregat, numit și nivelul macro de analiză. La acest nivel unitățile noastre de analiză sunt foarte generale. De exemplu, țările sunt unități de analiză agregate. Utilitatea acestor unități de analiză rezidă în capacitatea acestora de a cumula caracteristici medii ale sub-unităților lor de analiză (unități dezagregate mediu sau micro). Astfel, ele simplifică realitatea pe care dorim să o explicăm și, în același timp, ne permit să comparăm multe asemenea cazuri pentru a găsi răspuns la probleme generale, structurale. Analiza la nivel macro este utilă atunci când, de exemplu, dorim să aflăm răspunsuri la întrebări precum: de ce unele state au obținut stabilitate democratică, în vreme ce altele nu?, de ce unele state au

economii mai performante în vreme ce alte state nu reușesc să își folosească resursele pentru dezvoltarea economică? Unitățile de analiză aflate la nivel foarte agregat pot fi la rândul lor dezagregate, la nivel mediu sau macro.

Dată fiind complexitatea realității sociale pe care o studiem, de cele mai multe ori nu putem explica o unitate de analiză de la un nivel, fără să studiem şi să înţelegem unităţi de analiză de la un alt nivel. De exemplu, nu putem explica performanţa academică a studenţilor fără a înţelege diferenţele dintre specializări, facultăţi, centre universitare, sau chiar sisteme universitare naționale. Această combinație a celor trei niveluri de analiză impune însă o atenționare privind inferențele și concluziile inter-nivel (G. King, Keohane, și Verba 2000; Babbie 2010).

Aceste inferenţe se numesc eroare ecologică şi reprezintă utilizarea incorectă a unor date agregate pentru a formula inferenţe despre indivizi, descriindu-i pe aceştia şi caracteristicile lor fără să avem informaţii despre ei, ci doar informaţii despre grupuri agregate din care indivizii fac parte (G. King, Keohane, și Verba 2000; Babbie 2010; Glynn și Wakefield 2010). Eroarea este identificată atunci când se fac deducţii pornind de la date agregate iar concluziile se aplică unor subgrupuri sau indivizilor. De exemplu, facem această eroare atunci când având doar date agregate despre nivelul de poluare, tragem concluzii la nivelul indivizilor sau al grupurilor de indivizi, despre care, însă, nu avem informații dezagregate. Astfel, vedem în mod frecvent informaţii în mass-media din România, cu privire la puseurile de poluare existente în zona Bucureşti-Ilfov.[20] Multe dintre explicaţiile date acestor valuri de poluare observate doar prin date agregate sau mediu agregate, sunt explicate prin comportamentul individual (utilizarea autoturismului propriu în mod exagerat, încălzirea locuinţei cu combustibili fosili, sau chiar producerea unor arderi neautorizate) , iar concluziile sunt formulate cu referire la unități de analiză aflate la nivelul individual. Cu toate acestea, în lipsa unor date dezagregate de la nivel individual (de exemplu al autoturismelor, sau al locuințelor care folosesc combustibil fosil), aceste concluzii sunt eronate.

[20] O simplă căutare pe internet va identifica peste opt mii de articole mass-media dedicate poluării sporadice din zona București-Ilfov. Câteva exemple de raportare în mass-media sunt oferite de articolele Catiușei Ivanov, pentru Hotnews (Ivanov 2020; 2021; 2022).

În asemenea rapoarte privind poluarea nu avem informaţii privind caracteristicile individuale, pentru că nu măsurăm comportamentul individual, ci ne rezumăm culegerea de informaţii şi date de la un nivel agregat. Unitatea noastră de analiză şi de observaţie nu sunt indivizii, ci cartierele, localităţile, regiunile. Prin urmare, nu putem trage concluzii inferenţiale, deductive, afirmând că un comportament al indivizilor (dar pe care nu îl măsurăm) ar determina un efect la nivelul de analiză agregat. În realitatea, ar trebui să ne folosim de informații agregate (sau dezagregate) pentru a trage concluzii la nivel agregat (de localitate sau regiune), sau de informații dezagregate pentru a trage concluzii la nivel individual.

Dacă facem această eroare de logică inferenţială, vom trage concluzii greşite: nu avem de unde să ştim că un anumit comportament al indivizilor este cel care produce efectul estimat, în acest caz „poluarea”. Ar trebui ca datele culese, pe baza cărora am putea trage asemenea concluzii, să fie la nivel individual, dezagregat, deci să instalăm foarte mulţi senzori de poluare, capabili să identifice poluarea produsă de un eşantion reprezentativ de autoturisme care folosesc un anumit carburant, de sisteme de încălzire care folosesc combustibili fosili, de persoane sau firme care gestionează deşeuri etc. Orice altă conjectură privind comportamentul individual, susținută exclusiv pe baza unor date de la nivel agregat, nu poate conduce la concluzii valide ştiinţific şi, deci, la politici publice bazate pe date obiective, ştiinţifice.

Concluziile inferenţiale despre indivizi pe baza unor date agregate, cum ar fi informaţii politice (ele pot fi și economice sau sociale), pot constitui exemple de eroare ecologică. Un alt exemplu, destul de întâlnit, de raţionament în spaţiul public politic, şi care produce această eroare ecologică este interpretarea rezultatelor alegerilor şi tragerea unor concluzii cu privire la comportamentul unor indivizi specifici, pe baza informaţiilor de la nivel agregat (de exemplu, secţii de votare sau circumscripţii electorale). Rezultatele alegerilor din circumscripţia electorală X arată că partidul C a câştigat cele mai multe voturi. În circumscripţia X se constată că a existat un număr de votanţi pe liste electorale suplimentare, mai mare decât la ultimele alegeri. Concluzia trasă deseori în aceste situaţii este că votanţii de pe listele suplimentare au susţinut partidul C care astfel a câştigat circumscripţia X (poate pentru că s-a folosit de turismul electoral aducând votanţi din alte localităţi şi care au votat, prin urmare pe listele suplimentare). Aceste concluzii sunt însă false. Deşi avem informaţii şi date

despre procentul de votanţi ai partidului C din circumscripţia X şi informaţii despre numărul de votanţi pe listele permanente şi pe listele suplimentare, totuşi, nu avem date dezagregate despre preferinţa de vot a votanţilor de pe listele suplimentare sau permanente. Toate buletinele de vot, de pe ambele liste, se introduc în aceeaşi urnă, iar voturile sunt numărate şi procesate indiferent de lista pe care a fost înscris alegătorul. Dacă, însă, am face un sondaj de tip exit-poll în acea circumscripţie, pe un eşantion reprezentativ de alegători, şi am colecta informaţii de la aceştia, despre votul lor, statutul lor rezidenţial, atunci am putea trage concluzii la nivelul individual, iar unităţile de analiză pot fi indivizii şi nu circumscripţia electorală. În concluzie, această problemă şi importanţa înţelegerii erorii ecologice este relevantă nu doar în ştiinţele politice sau sociologie, ci şi în ştiinţele care deseori utilizează indivizii ca unitate de analiză: psihologie, ştiinţele educaţiei, criminologie (Walker 2021) etc.

Prin urmare, putem concluziona că, pentru a întări raționamentul științific, este întotdeauna preferabilă colectarea datelor la cel mai dezagregat nivel cu putință, desigur, ținând cont de resursele umane, financiare și de timp disponibile studiului empiric. Datele culese la nivel individual pot fi, ulterior, agregate la un nivel superior (de exemplu județul), pentru a trage concluzii despre acest nivel, mai ales atunci când avem eșantion reprezentativ pentru sub-grupuri, dar chiar și atunci când avem date nereprezentative pentru subgrupuri. Însă, odată ce am cules informațiile la un nivel agregat (de exemplu județul) nu mai putem să le dezagregăm la nivel individual, deci nu putem trage concluzii la nivelul individual bazându-ne pe aceste date agregate, fără să culegem noi date, activitate pentru care trebuie să investim noi resurse.

O altă problemă importantă, de design de cercetare, pe care o putem întâmpina atunci când ne selectăm cazurile de analiză, o reprezintă selecţia acestora în funcţie de variabilele incluse în analiză (în ipoteză) (G. King, Keohane, și Verba 2000). Selecția cazurilor pe variabila dependentă, sau de explicat, este o eroare importantă pe care chiar și cercetători experimentați o pot face. Deoarece deseori ne dorim atât de mult să explicăm o problemă riscăm să îngustăm prea mult câmpul de analiză şi să nu mai vedem pădurea din cauza copacilor. Prin formularea ipotezei de regulă ne aşteptăm să obţinem o anumită variaţie a variabilei dependente, la care ne vom raporta chiar şi atunci când ne selectăm cazurile de analizat. Această selecţie este însă mai degrabă potrivită pentru analize descriptive, exploratorii, pe baza cărora am putea mai

degrabă să formulăm întrebări de cercetare, nu pentru studii cu caracter analitic, inferenţiale, prin care să testăm ipotezele. Să ne folosim de următorul exemplu: în studiul nostru ne dorim să explicăm performanţa academică de la cursul de Metode de Analiză a Datelor. Deoarece performanţa semnifică mai degrabă un rezultat foarte bun, e posibil să ne concentrăm atenţia asupra studenţilor care au luat note mari, de exemplu 10 sau 9, a căror performanţă ne-am dori să o înţelegem şi, eventual, să o replicăm. Ar fi o eroare dacă am include în studiul nostru doar studenţi care au luat note de 9 şi de 10. Pe baza ipotezei, am putea trage concluzia că performanţa academică este determinată de caracteristici pe care le observăm ca fiind prevalente la studenţii de 9 şi 10 pe care i-am inclus în studiul nostru (de exemplu, toţi aceşti studenţi sunt femei, toate poartă ochelari de vedere, au avut medii de 9 sau 10 la matematică în liceu, şi au participat la toate întâlnirile cursului de Metode de Analiză a Datelor).

Cu toate acestea, trebuie să ne fie clar că procedând astfel am pierde din vedere posibilitatea ca aceste caracteristici să existe în această formă şi la studenţii care nu obţin performanţă academică. Vom putea trage concluzii cu privire la relaţia dintre performanţa academică şi acele caracteristici individuale ale studenţilor, doar dacă cazurile alese variază (şi) din punctul de vedere al variabilei dependente. De aceea, este recomandabil să permitem cazurilor noastre să aibă o variaţie maximă pe variabila de explicat. Doar în acest fel ele ne vor putea ajuta să respingem ipoteza nulă, conform căreia variaţia variabilei dependente nu este determinată (cauzată) de variaţia variabilei independente. Astfel, vom putea discerne dacă avem cauzalitate sau doar corelaţie.

În final, vom da un exemplu foarte ilustrativ, care ne poate descreţi frunţile, oferit de Ronald King pentru a ilustra modul în care producem erori de selecție a cazurilor folosind informaţiile despre ceea ce vrem să explicăm: un bătrân pe patul de moarte îi spune soţiei sale „ai fost lângă mine în toate momentele noastre grele, atunci când mi-am pierdut slujba, când am avut accidentul de maşină, când au plecat copiii, şi când m-am îmbolnăvit. Pleacă, căci îmi aduci ghinion!” (R. F. King 2005, 207). Prin urmare, nu vom face aceeaşi eroare precum acest bătrân: nu vom selecta cazurile pornind de la situaţiile în care apare efectul formulat în ipoteză, ci pornind de la

cauzele pe care le-am inclus în formularea ipotezei, diversificând şi maximizând efectele acestora.

O altă eroare de selecţie a cazurilor este selecţia acestora pe baza unei variabile independente. Folosindu-ne de exemplul de mai sus, în care dorim să explicăm performanţa academică a studenţilor, ar fi greşit dacă ne-am alege cazurile de studiat pe baza unei variabile independente, cum ar fi, de exemplu genul studentului sau participarea sa la cursuri. Dacă am alege în acest fel, am include fie doar femei în studiu (una dintre conjecturi formula legătură între gen şi performanţa academică), fie doar studenţi care au participat la toate întâlnirile cursului (o altă conjectură enunţată mai sus). În această situaţie am putea trage concluzii cauzale (deci raţionamentul inferenţial nu este compromis), însă acestea ar putea fi generalizate doar la femei sau la studenţii care au participat la cursuri, nu la orice student (caz) din populaţia totală, pe care am dori să o explicăm (deci puterea de generalizare a studiului este compromisă) (G. King, Keohane, și Verba 2000). În concluzie, trebuie să permitem cazurilor potențial a fi incluse în studiul nostru o variație mare și pentru variabilele independente. Ronald King (2005, 209) aprecia că această variabilitate maximă permisă poate fi raportată la variabilitatea din populație, aceasta putând fi obţinută cel mai uşor prin selecţia aleatorie a cazurilor.

Teoretic, însă ne putem confrunta cu două dileme, deşi ne putem propune să permitem variabilitate maximă a cazurilor: prima este aceea a situației în care avem prea multe cazuri; a doua este cea în care avem prea puține cazuri în populația totală, pe care le-am putea include în studiu. Despre prima situație vom discuta în continuare în această secțiune a capitolului, urmând ca în următorul subcapitol să discutăm despre situația în care avem prea puține cazuri și trebuie să creștem numărul acestora prin tehnici specifice.

De regulă, în studiile cantitative, aşa cum menţionam în capitolul anterior, avem un număr mare sau foarte mare de cazuri, la fel şi în populaţia totală pe care dorim să o studiem. Acest număr mare de cazuri ce ar putea fi teoretic incluse în studiul nostru poate fi redus cel mai simplu prin selecţie aleatorie: dând cu banul, aruncăm cu zarul, extrăgând bile dintr-o urnă, sau extrăgând automat cu ajutorul unui program statistic numere dintr-o listă. În studiile cantitative numim eşantionare

această reducere a numărului mare de cazuri din populația totală. Ea reprezintă selectarea unei părți din populația totală, parte care reproduce la scară anumite caracteristici ale populației totale, astfel încât cunoscând anumite caracteristici ale acestei părți să putem estima aceste caracteristici și pentru populația totală. În studiile cantitative pornim, de regulă, de la o populație totală (indivizi, grupuri, instituții, organizații, țări) care urmează să fie studiată, și pe baza căreia vrem să testăm validitatea ipotezelor formulate.

Pentru a înțelege utilitatea selecției aleatorii a unui număr redus de cazuri dintr-o populație totală mai mare, pentru a trage concluzii cu grad ridicat de generalizare la nivelul întregii populații, folosim câteva caracteristici matematice ale selecției aleatorii probabiliste. Probabilitatea unui eveniment reprezintă „proporția numărului de cazuri în care evenimentul ar avea loc, dacă am măsura în mod repetat acest eveniment" (Agresti și Finlay 2014, 73). Fiind o proporție, sau raport a două mărimi (partea și întregul), probabilitatea poate orice valoare între 0 și 1 sau echivalent, între 0% și 100%. O întâlnim exprimată, de exemplu, în probabilitatea (50%) de a nimeri una dintre fețele banului, atunci când dăm cu banul, probabilitatea (2,04%) de a se extrage un număr la loto 6 din 49.

Atunci când dintr-o mulțime (de cazuri, de indivizi etc.) extragem aleatoriu un număr redus de cazuri, dacă pentru aceste cazuri calculăm un indicator cum ar fi media pentru o anumită caracteristică, dacă eșantionul va fi suficient de mare media acestuia va aproxima, cu o anumită eroare, media caracteristicii populației din care eșantionul a fost extras. Dacă extragem mai multe asemenea eșantioane, media mediilor acestora va fi din ce în ce mai apropiată de media populației din care aceste eșantioane au fost extrase. Această caracteristică poartă numele de teorema limitei centrale, și ea ne ajută să înțelegem modul în care un distribuția dintr-un eșantion aleatoriu poate aproxima distribuția din populația totală, pe o caracteristică dată. Astfel, extrăgând un eșantion aleatoriu dintr-o populație, indiferent de distribuția acestei populații pe o caracteristică de interes (variabilă continuă, cantitativă), calculându-i media pe această caracteristică și adăugând-o într-o distribuție a mediilor unor eșantioane extrase în mod repetat din aceeași populație, vom putea observa cum distribuția mediilor urmează o distribuție normală, de tip Gaussian (Agresti și Finlay 2014, 93–95).

Distribuţia normală, pe care o întâlnim nu doar în cazul acestor eşantioane extrase aleator, în mod repetat, ci şi în cazul unor variabile (de exemplu înălţimea, greutatea, notele şcolare), are câteva caracteristici pe care ne vom baza atunci când calculăm probabilitatea de a greşi când estimăm (deoarece nu avem informaţii anterioare despre) valoarea pe populaţia totală a unui parametru calculat pe un eşantion extras din acea populaţie. Ea este simetrică faţă de medie şi prezintă o omogenitate pe care o putem calcula cu ajutorul abaterii standard de la medie, un indicator pe care îl vom discuta în capitolul următor. Aceste caracteristici ale distribuţiei normale ne sunt utile pentru a estima că pentru o caracteristică măsurată pe o scală continuă, a unor cazuri (de exemplu, indivizi), un individ extras aleator din populaţie are 68% şanse să aibă o valoare pe acea caracteristică într-un interval cuprins între media acelei caracteristici plus sau minus o abatere standard. Altfel spus, dacă extragem aleatoriu 100 de eșantioane din populație, 68 dintre aceste eșantioane vor avea media în intervalul cuprins între +/- o abatere standard în jurul mediei. În plus, pe acelaşi tip de distribuţie normală, un individ extras aleator din populaţie are 95% şanse să aibă o valoare pe acea caracteristică într-un interval cuprins între media acelei caracteristici plus sau minus două abateri standard. Deci din 100 de eșantioane extrase aleatoriu 95% vor avea media acelei caracteristici de interes, în intervalul cuprins între +/- două abateri standard în jurul mediei. Putem utiliza aceşti doi parametri (media şi abaterea standard) pentru a standardiza unitățile de măsură ale caracteristicii evaluate și a o putea compara cu alte caracteristici ale aceleiași populații. În acest fel, obținem o distribuție normală standard, în care media ia valoarea 0, iar abaterea standard valoarea 1. Caracteristicile enunțate mai sus rămân valabile (Agresti și Finlay 2014, 78–85).

Utilitatea eşantionului constă în capacitatea lui de a putea estima caracteristici ale populaţiei totale (de exemplu totalitatea locuitorilor dintr-o ţară sau dintr-o regiune) pe care altfel nu le cunoaştem. O încercare de a măsura aceste caracteristici pentru populaţia totală s-ar lovi de timpul şi resursele exorbitante necesare. De exemplu, recensământul, care realizează exact acest tip de studiu exhaustiv pe populaţia totală (sau aproape totală), este la rândul său limitat la un număr de întrebări sociale, economice şi demografice, fără să măsoare atitudini şi comportamente (care sunt, de altfel, extrem de inconstante în timp). Costul

recensământului din 2011 realizat în România, s-a ridicat la 40 de milioane de euro (Răducanu 2013), în vreme ce recensământul din 2022 a costat aproximativ 329 milioane de lei (aproximativ 67 milioane de euro) (Diochetanu 2022). Studiile ştiinţifice nu vor putea niciodată obţine finanţări de acest nivel pentru a măsura atitudini ale persoanelor. Utilizarea eşantionului permite nu doar un cost mult mai redus decât un studiu pe întreaga populaţie, ci şi o rapiditate în culegerea datelor, analizarea lor, publicarea rezultatelor, dar şi utilizarea unor tehnici şi instrumente inovative (cum ar fi de exemplu experimentele integrate în sondaje, despre care vom discuta la finalul acestui capitol).

Reprezentativitatea eşantionului, definită ca fiind capacitatea acestuia de a reproduce cât mai fidel caracteristicile populaţiei din care a fost extras (Rotariu et al. 1999, 86), este caracteristica fundamentală pentru utilizarea unei părţi din întreg pentru a estima caracteristici ale întregului. Trebuie subliniat faptul că această reprezentativitate a eşantionului este estimată (şi calculată) întotdeauna prin raportare la populaţia totală, mai precis la informaţiile pe care deja le cunoaştem din populaţia totală, de cele mai multe ori cu ajutorul recensământului. Însă, aşa cum am precizat mai sus, în recensământ avem puţine informaţii, acestea fiind mai degrabă unele precum genul, vârsta, mediul de rezidenţă, mărimea localităţii, educaţia, etnia, religia. Nu ne oferă informaţii detaliate despre statusul economic şi social al indivizilor, nici despre comportamentele şi atitudinile acestora, cum ar fi încrederea în ceilalţi, gradul de satisfacţie cu locul de muncă etc. Prin urmare, reprezentativitatea eşantionului se referă nu la orice caracteristică a populaţiei ci ea este particulară, fiind raportată la o anumită caracteristică a acestuia. Reprezentativitatea eşantionului depinde de cât de omogenă este populaţia totală şi nu de cât de mare este această populaţie.

De exemplu, dacă avem o populaţie totală formată doar din femei, este suficient să extragem o singură femeie pentru eşantionul nostru pentru a afirma că acesta este reprezentativ pentru populaţia totală, pe variabila gen. Dacă populaţia totală ar fi formată din femei care locuiesc în mediul rural, toate având 44 de ani şi înălţimea de 1,7 m, atunci ar fi suficient să extragem aleatoriu o singură persoană, pentru ca eşantionul nostru să fie reprezentativ pentru toate cele trei variabile. Am putea asemui situaţia cu aceea a unui bucătar care găteşte un ghiveci de legume. Dacă poate

amesteca în ghiveci pentru a spune cât de cald este, atunci va reuşi să îl omogenizeze suficient de bine încât să îi fie suficientă o mostră luată cu lingura, pentru a spune dacă mâncarea este suficient de caldă sau nu, fără a fi nevoit să mănânce tot ghiveciul pentru a da un răspuns la această întrebare. Dacă, însă, bucătarul nu poate amesteca în ghiveci pentru a-l omogeniza, atunci acesta poate fi foarte eterogen distribuit şi deci ar fi nevoie de mai multe mostre pe care bucătarul să le poată lua din zone diferite ale oalei cu ghiveci, astfel încât să poată trage o concluzie pe baza „mediei temperaturii tuturor mostrelor de ghiveci".[21]

Prin urmare, reprezentativitatea eşantionului este dependentă şi de mărimea acestuia, dar şi de selecţia aleatorie a indivizilor din eşantion. Sporul de reprezentativitate creşte rapid atunci când creştem mărimea eşantionului în grade mici, ajungând rapid în apropierea plafonului P=100 (Rotariu et al. 1999, 88–89). O creştere suplimentară, peste aproximativ 800 de indivizi nu mai aduce un spor notabil de reprezentativitate. Acest spor permite însă ca subcategorii ale unor caracteristici să fie la rândul lor reprezentative pentru subcategoriile similare din populaţia totală (de exemplu un eşantion de 2000 de respondenţi poate fi reprezentativ şi pentru subcategoria bărbaţi. Din acest motiv, eşantioanele mari sunt utile pentru analizele pe subeşantioane reprezentative (Rotariu și Iluț 1997; Rotariu et al. 1999).

Reprezentativitatea eşantionului probabilist se măsoară întotdeauna prin două mărimi: eroarea maximă sau mărimea maximă a erorii de eşantionare aleatorie (***d***) (care reprezintă diferenţa cea mai mare pe care o acceptăm atunci când estimăm o caracteristică în populaţia totală pe baza măsurării acestei caracteristici în eşantionul aleator extras din acea populaţie totală; şi este calculată conform formulei de mai jos); respectiv probabilitatea (***P***) ce reprezintă şansele ca eroarea de estimare a caracteristicii din populaţie să se situeze în limitele determinate de eroarea (***d***). În practică, deseori folosim inversul lui (***P***), adică (***p***), calculat ca ***1-P***, ce reprezintă nivelul de încredere statistică sau nivel de semnificaţie statistică. Practica din ştiinţele sociale foloseşte un nivel maxim acceptat pentru eroarea d≤3%, un nivel de încredere de maxim p≤0,05, iar o probabilitate P≥0,95 (Rotariu și Iluț 1997; Rotariu et al. 1999).

[21] Exemplul este inspirat de Traian Rotariu și Petru Iluț (1997, 122).

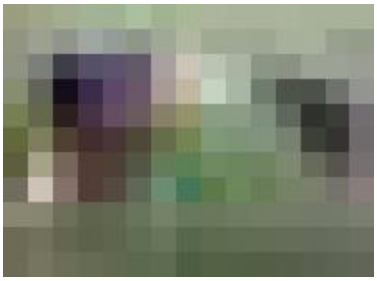
$$e = \mathcal{Z} \times \frac{\sigma}{\sqrt{n}} \quad (1)$$

unde σ reprezintă abaterea standard a caracteristicii populației (fiind rareori cunoscută aceasta este înlocuită cu abaterea standard σ' a eșantionului dacă eșantionul extras aleator este mare), n reprezintă mărimea eșantionului, iar $\mathcal{Z}$ reprezintă scorul $\mathcal{Z}$ al intervalului de încredere (1,96, pentru un nivel al P=95%, sau 2,58 pentru un nivel al P=99%).[22]

Pentru a înțelege modul în care un eșantion este capabil să reproducă populația totală putem folosi o analogie cu fotografia digitală la o rezoluție mai mică pe care o utilizăm pentru a surprinde realitatea, pentru a identifica suficient de multe detalii încât să putem spune care sunt caracteristicile generale ale realității fotografiate, fără să fim nevoiți să spargem pușculița pentru a cumpăra cel mai performant aparat de fotografiat digital (a se vedea ilustrația Foto 2.1.1). Crescând numărul de pixeli putem obține o imagine mai bună a populației totale pe caracteristicile pe care dorim să le explorăm; această creștere ne-ar permite să identificăm cele mai mici detalii, precum firul de iarbă din fotografia de mai jos. Cu toate acestea, pentru a putea trage concluzii cu privire la imaginea de ansamblu, ne-am putea folosi de o fotografie (eșantion) suficient de redusă ca dimensiune, pentru a estima cu un anumit grad de eroare valori medii pe unele caracteristici din realitate (populația totală). Pentru comparație, în fotografiile de mai jos fiecare pixel reprezintă o medie a pixelilor din imediata sa apropiere, cu care are caracteristici comune, astfel încât împreună, pixelii să fie capabili să reproducă cât mai bine unele caracteristici ale imaginii complete (populația totală a pixelilor).

[22] Manualele de statistică prezintă în detaliu modul de calcul al scorului Z, prin urmare, nu vom relua această prezentare, recomandând însă manuale de statistică devenite clasice în mediul academic, cum ar fi Rotariu et al. în limba română (1999) sau Agresti și Finlay în limba engleză (2014).

63 de cazuri (pixeli)



165 de cazuri (pixeli)

300 de cazuri (pixeli)

800 de cazuri (pixeli)

1530 de cazuri (pixeli)

2700 de cazuri (pixeli)

Toată populaţia = 16,084,992 cazuri (pixeli)

Foto 2.1: Pic (2007 - 2022) şi Alf (2011-)

În studiile cantitative în care folosim ancheta sociologică şi sondajul de opinie, selecţia aleatorie este principalul instrument de selectare a cazurilor analizate. Literatura de specialitate, extrem de bogată (Kish 1995; Rotariu și Iluț 1997; Lohr 2010; Agresti și Finlay 2014; Fowler 2014), care tratează sondajele de opinie și eșantionarea, analizează în detaliu diverse metode de eșantionare probabilistă (aleatorie – când fiecare individ din populație are o șansă diferită de zero de a fi selectat în eșantion (Rotariu și Iluț 1997, 130) și ne-probabilistă (ne-aleatorie). Din prima categorie, amintim eșantioanele simple aleatorii (pe baza de date de angajați raportați la ITM putem folosi un program statistic pentru a extrage aleatoriu o listă de angajați); stratificate (populația este formată din grupuri sau straturi de indivizi, de exemplu în funcție de aria culturală, mediul de rezidență, tipul de localitate, gen sau etnie și putem extrage aleatoriu din fiecare strat un număr de respondenţi proporţional cu

mărimea stratului[23]); cluster (indivizii sunt parte dintr-un grup, care la rândul său face parte dintr-un grup mai mare şi aşa mai departe, prin urmare putem folosi această grupare naturală a indivizilor extrăgând aleatoriu asemenea grupuri de indivizi pe care le putem utiliza în întregime în eşantion, de exemplu o grupă de studenţi). Aceste eşantionări sunt caracterizate de asumpţiile distribuţiilor normale şi ale limitei centrale, astfel încât vor permite calcularea erorii de eşantionare şi estimarea parametrilor din populaţia totală.

Dintre eşantionările non-aleatorii putem menţiona eşantionarea pe cote, cea mai des folosită metodă de selecţie non-probabilistă a indivizilor în sondaje de opinie. Această metodă este utilizată de regulă cu mulţi (minim 3 sau mai mulţi, de ex. gen, vârstă, educaţie, venit, etnie, ocupaţie) parametri pe care îi combinăm astfel încât operatorul de teren să poată alege doar persoanele care corespund strict combinaţiei de parametri şi cotelor alocate, iar eşantionul astfel obţinut să fie cat mai apropiat de caracteristicile populaţiei totale. Deoarece selecţia respondenţilor este aproape cu totul la îndemâna operatorului de teren, fără o selecţie aleatorie a subiecţilor, eşantioanele non-probabiliste nu se supun aceloraşi asumpţii privind calculul probabilităţii de selecţie a indivizilor în cazul extragerii unor eşantioane repetate, precum cele probabiliste, prin urmare calcularea erorii de eşantionare nu poate fi făcută după aceleaşi principii.

2.2. Studiul de caz

Dacă în cercetările cantitative studiem de cele mai multe ori o populaţie mare de cazuri, fiind astfel, deseori, obligaţi să reducem numărul acestora, strategia folosită fiind aceea a eşantionării, în studiile calitative sau comparative adeseori numărul total de cazuri pe care le putem analiza pentru a răspunde la întrebarea de cercetare, este limitat. La extremă acest număr poate fi egal cu 1. Numărul redus al cazurilor poate fi

[23] Un exemplu de eșantionare stratificată și multistadială utilizată pe scară largă în sondajele din România, și care a fost în mare parte dezvoltată și materializată în lucrările lui Dumitru Sandu (1996; 1999).

determinat nu doar de limitarea naturală a acestora, ci și de scopul cercetării, acela de a analiza în cel mai mic detaliu o unitate de analiză și de observație, apelând deci la cazuri unice, reprezentative. Prin urmare, în literatura de specialitate, studiul de caz este deseori definit (Yin 2005) ca o metodă de analiză calitativă a unui număr de cazuri egal cu unu (pentru a trage concluzii generalizabile pentru alte cazuri asemănătoare), un studiu etnografic sau de tip *process-tracing* (identificarea posibilelor cauze ale unui efect observabil într-un caz, prin analizarea informațiilor relevante în conformitate cu predicțiile teoretice despre acel caz) (George și Bennett 2005), sau un studiu în care analizăm o singură problemă; prin urmare este mai degrabă un mod de definire a cazurilor, nu o metodă specială de analiză a unui singur caz (Gerring 2004).

Un exemplu de studiu de tip *process-tracing* este oferit de un studiu (Adăscăliței și Muntean 2019) ce a încercat să explice modalitatea în care două grupuri de interese pot dezvolta strategii diferite pentru a obține același tip de beneficii: condiții mai bune de muncă, creșteri salariale, consultări bipartite. Folosind analiză de documente, interviuri și date cantitative agregate, studiul documentează transformarea mișcării sindicale din sectoarele sănătății și al educației, și adaptarea acestora la schimbările politice, economice și sociale. Preferințele individuale, deciziile de grup, alianțele politice, mecanismele de consultare a membrilor, sunt documentate pentru a ilustra adaptarea combinației de strategii pe care sindicate reprezentative din cele două sectoare le-au folosit în negocierile cu reprezentații guvernului.

Pentru a înțelege mai bine aceste diferențe, ne vom folosi de tipologia propusă de John Gerring pentru a diferenția între mai multe tipuri de studiu de caz, tipologie bazată pe delimitarea spațială și temporală a unui caz (fenomen studiat) (Gerring 2006). Utilitatea studiului de caz, este derivată din înțelegerea acestuia ca fiind „un studiu intensiv al unei singure unități de analiză cu scopul de a înțelege o clasă mai mare de unități similare" (Gerring 2004, 342). Folosind aceste distincții putem identifica modul în care studiul de caz poate fi utilizat pentru a permite explicarea sau înțelegerea fenomenului studiat. Pentru a face aceasta folosim logica inferențială. În studiile comparative utilizarea acestei logici a studiului de caz ne poate ajuta să înțelegem metodele prin care putem multiplica numărul de cazuri studiate (Lijphart 1971), situație opusă aceleia în care suntem în cazul studiilor cantitative, unde folosim eșantionarea pentru a reduce numărul de cazuri.

Să presupunem că dorim să studiem şomajul din România. Informaţia pe care o primim este aceea că şomajul este de 5%. În plus, am putea culege multe alte informaţii în conformitate cu teoriile din literatura de specialitate care explică şomajul prin intermediul unor variabile precum inflaţia, nivelul investiţiilor străine directe, rata de creştere a PIB etc. Având toate aceste informaţii putem produce ceea ce în aparenţă pare a fi un studiu de caz (o unitate de analiză, ţara, România). Cu toate acestea, raţionamentul inferenţial ne arată că toate aceste informaţii despre unitatea noastră de analiză sunt detalii ale unui instantaneu. Utilitatea sa poate fi una descriptivă sau cel mult exploratorie (Yin 2005), însă nu ne va ajuta să explicăm, nici să înţelegem cazul şi problema pe care o studiem.

Pentru a putea trage asemenea concluzii avem nevoie de o abordare diferită a studiului de caz: trebuie să îi permitem o variaţie temporală şi / sau una spaţială. Dacă permitem o variaţie temporală, atunci vom avea informaţii despre şomaj la momente de diferite de timp ($T_0, T_{-1}, T_{-2}, \dots, T_{-n}$) care ne vor permite să comparăm unitatea noastră de analiză (cazul) cu ea însăşi. Având acelaşi tip de informaţii cu variaţie temporală, despre determinanţii şomajului, putem mai uşor înţelege şi explica fenomenul analizat. În plus, avem acelaşi tip de cercetare: studiul de caz. John Gerring (2004) îl numeşte studiu diacronic.

Dacă permitem o variaţie spaţială a unităţii de analiză, astfel încât cazul nostru să poată fi analizat observând sub-unităţile sale componente ($Ti_0, Tj_0, Tk_0, \dots, Tz_0$) putem colecta informaţii de la acelaşi moment în timp dar din regiuni diferite (de exemplu judeţe sau arii culturale), producând un studiu de caz de tip sincronic. Dacă analizăm fiecare sub-unitate de analiză la momente de timp diferite atunci vom avea un studiu de caz de tip diacronic şi sincronic (Gerring 2004), care permite o explicaţie mai bună a fenomenului studiat şi, în acelaşi timp, ne permite să studiem un singur caz. Acelaşi raţionament ce foloseşte variaţia temporală şi spaţială poate fi folosit şi în studiile ce includ mai multe unităţi de analiză, acestea fiind de exemplu analizele comparative, analizele de serii de timp ale unor cazuri diferite, modelele ierarhice cu mai multe niveluri de analiză sau seriile de timp ierarhice.

În studiile de caz informaţiile pot provin de la un număr mare de variabile, nu doar din observaţia directă a fenomenului analizat, ci şi din analiza unor documente

relevante pentru acel caz sau produse de acel caz, documente de arhivă, obiecte şi alte elemente materiale şi imateriale, dar şi observaţii din participarea directă a cercetătorului, şi, nu în ultimul rând, din interviuri aprofundate. Studiile de caz pot deveni studii la fel de costisitoare ca şi cele cantitative sau comparative, oferind însă o libertate mai mare în alegerea unităţii de analiză. Atunci când studiul de caz este folosit pentru compararea mai multor cazuri sunt îndeobşte preferate strategii metodologice capabile să asigure un grad cât mai ridicat de comparabilitate din punct de vedere teoretic, de regulă prin tehnici precum analiza de congruenţă sau *pattern matching* (Nielsen 2016), care implică folosirea unor algoritmi, dezvoltaţi pe baza teoriei explicative, pentru a asigura în fiecare caz potenţial, cea mai bună compoziţie a variabilelor relevante.

Complexitatea datelor culese în studiile de caz se pretează atât la analize cantitative, cât şi la analize calitative. Bazându-ne studiul de caz pe raţionamentul inferenţial şi covariant, aşa cum am detaliat mai sus, instrumentele pe care le putem folosi sunt cele de analiză descriptivă, univariată, dar şi cele specifice modelelor de analiză multivariată, pe care le vom detalia în capitolul următor. Informaţiile calitative din studiile de caz pot fi culese şi analizate folosind instrumente specifice, cum ar fi analiza de conţinut, analiza de discurs, focus grupul şi interviul aprofundat. Interpretarea concluziilor trebuie însă să ţină cont de limitările studiului de caz, cum ar fi puterea redusă de generalizare a concluziilor la alte cazuri, subiectivitatea cercetătorului în selecţia şi utilizarea unor unităţi formale şi informale (indirecte) de analiză care pot distorsiona concluziile (Gerring 2004), lipsa de putere explicativă determinată de limitarea numărului (sub)unităţilor de analiză, sau erori în validitatea internă a studiului, despre care vom discuta în capitolul următor.

2.3. Cele mai asemănătoare cazuri / cele mai diferite cazuri

În secţiunile anterioare am discutat despre metodele pe care le putem folosi pentru a specifica unităţile de analiză în studii cantitative şi calitative. În studiile comparative, în care numărul de cazuri tinde să fie mai mic decât în cele cantitative, dar mai mare decât în studiile calitative, ne confruntăm cu câteva probleme metodologice specifice: cum putem multiplica numărul de cazuri, cum putem reduce numărul acestora, cum selectăm cazurile comparabile şi cum reducem numărul de variabile şi, deci, de observaţii. Dacă multiplicarea numărului de cazuri poate fi făcută în mod elegant prin utilizarea variaţiei diacronice şi sincronice, iar dacă reducerea numărului de cazuri poate fi obţinută prin eşantionare, identificarea celor mai potrivite cazuri pentru comparaţii poate ridica probleme metodologice chiar şi celor mai experimentaţi cercetători. Motivaţia renunţării la un caz, sau a introducerii unui caz într-un studiu, trebuie să permită nu doar variaţia maximă ci şi capacitatea acestor cazuri de a ne ajuta să tragem concluzii inferenţiale pe baza unui lanţ cauzal plauzibil.

Pornind de la diferenţele sau asemănările efectelor pe care dorim să le analizăm în cazurile comparate, putem selecta în studiu acele cazuri a căror selecţie poate fi fundamentată pe metodele de diferenţiere şi acord, dezvoltate de John Stuart Mill (1882). Astfel, putem selecta cazurile cele mai asemănătoare, sau cazurile cele mai diferite.

În designul ce foloseşte cele mai asemănătoare cazuri pornim de la asumpţia că avem efecte diferite care apar în cazurile studiate (de exemplu, de ce în unele state democraţia se dezvoltă, iar în altele nu). Pentru a permite identificarea corectă a cauzelor acestor efecte diferite ar trebui să urmărim selecţia acelor cazuri care sunt asemănătoare sau chiar similare pe cele mai multe caracteristici, sau măcar pe cele mai relevante din punct de vedere inferenţial. Totuşi, aceste cazuri trebuie să fie diferite pentru, de regulă, o singură caracteristică, sau, oricum, pentru un număr foarte mic de caracteristici relevante din perspectiva teoriei pe care ne bazăm argumentul. Astfel, diferă variabila dependentă, de explicat, şi o variabilă independentă, cea pe care o

considerăm a fi cauză pentru efectul de explicat, aşa cum ilustrăm în tabelul 2.1. În acest fel putem ţine constante explicaţiile alternative.

Tabel 2.1 Metoda celor mai asemănătoare cazuri

caracteristici	Cazul 1	Cazul 2
Variabila independentă 1	**A**	**A′**
Variabila independentă 2	**B**	**B′**
Variabila independentă 3	**C**	**C′**
Variabila independentă 3	**X**	**Y**
Variabila dependentă	**Efect 0**	**Efect 1**

În designul care foloseşte cele mai diferite cazuri pornim de la asumpţia că avem efecte asemănătoare care apar în cazurile studiate (de exemplu dorim să explicăm succesul consolidării democraţiei în două state diferite). Pentru a permite identificarea corectă a cauzelor acestor efecte diferite ar trebui să urmărim selecţia acelor cazuri care sunt diferite pe cele mai multe caracteristici, sau măcar pe cele mai relevante din punct de vedere inferenţial. Totuşi, aceste cazuri trebuie să fie asemănătoare pentru, de regulă, o singură caracteristică, sau, oricum, pentru un număr foarte mic de caracteristici relevante din perspectiva teoriei pe care ne bazăm argumentul. Astfel, sunt identice variabila dependentă, de explicat, şi o variabilă independentă, cea care considerăm că este cauză pentru acest efect, aşa cum ilustrăm în tabelul 2.2. În acest fel putem ţine constante explicaţiile alternative.

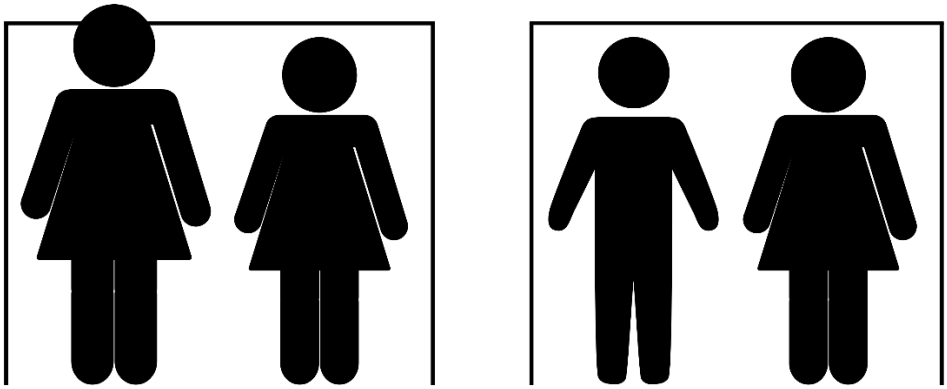
Tabel 2.2 Metoda celor mai diferite cazuri

caracteristici	Cazul 1	Cazul 2
Variabila independentă 1	**A**	**B′**
Variabila independentă 2	**C**	**D′**
Variabila independentă 3	**E**	**F′**
Variabila independentă 3	**X**	**X′**
Variabila dependentă	**Efect 1**	**Efect 1**

Ca să ilustrăm grafic diferenţele dintre aceste două metode şi să înţelegem cele două strategii metodologice prin care am putea explica un fenomen, să ne imaginăm

că vrem să explicăm ce caracteristici determină capacitatea a două persoane de a trece pe sub un prag (a se vedea pentru ilustrare figura 2.3). Astfel, prin metoda celor mai asemănătoare cazuri vom selecta două persoane, femei, surori gemene, care au comune toate caracteristicile mai puţin înălţimea. Rezultatul pe care vrem să îl explicăm (a trece sau nu pe sub un prag) este determinat tocmai de caracteristica pe care cele două persoane sunt diferite: înălţimea. Prin metoda celor mai diferite cazuri vom selecta două persoane care au o singură caracteristică comună – înălţimea, din lista caracteristicilor relevante din punct de vedere teoretic pentru explicarea capacităţii de a trece pe sub un prag. Toate celelalte caracteristici sunt diferite. Efectul pe care dorim să îl explicăm este determinat, de această dată, de caracteristica pe care cele două persoane o au în comun: înălţimea.

Figura 2.3 Cele mai asemănătoare (stânga) și cele mai diferite (dreapta) cazuri



2.4. Experimente: considerații generale

Spre deosebire de metodele observaționale pe care le-am discutat în capitolele și secțiunile anterioare, și care sunt deseori preferate în științele sociale, studiile experimentale se bazează pe metoda comparativă și pe o planificare și organizare atentă pentru condițiile în care o cauză produce un efect (Freedman, Pisani, și Purves

2007). Această cauză este tratamentul, iar cercetătorul încearcă să ţină sub control orice potenţială cauză alternativă a efectului pe care doreşte să îl explice. De aceea, experimentele sunt designul de cercetare preferat în ştiinţele naturii, medicale, fizică, chimie, putând asigura condiţii de cercetare controlate, de regulă în laborator, pentru explicarea unor fenomene. Controlul perfect pe care cercetătorii din aceste domenii îl pot realiza în laborator este mai greu a fi replicat în studierea în laborator a unor comportamente sociale comparabile cu cele din viaţa reală. Experimentele de teren trebuie să suporte mai multe constrângeri pentru a putea obţine un nivel ridicat de explicaţie a unui efect. Cu toate acestea, în ultimii ani, experimentele sunt din ce în ce mai utilizate şi în studiile din ştiinţele sociale care au ca subiecţi un număr mare de cazuri şi fenomene sociale complexe. Deoarece problemele cercetate în ştiinţele sociale se desfăşoară în mod natural în realitate, fiind influenţate sau denaturate de un control în laborator, designul experimental este integrat adeseori în tehnici şi instrumente cantitative observaţionale, cum este sondajul de opinie. Vom discuta despre acest design de cercetare în subcapitolul următor.

Pe scurt, experimentele reprezintă un tip de design de cercetare care permite cercetătorilor să aloce subiecţii unui tratament cu o probabilitate calculabilă între 0 și 1 (Gerber și Green 2012). Subiecții care iau parte la experiment sunt alocați fie unui grup de tratament, fie unui grup de control. Un design experimental poate avea mai multe grupuri de tratament şi de control, în funcţie de numărul de categorii definite de variabila cauză.

În toate tipurile de design experimental selecţia aleatorie a unităţilor / subiecţilor (de exemplu localităţi), respectiv alocarea aleatorie a subiecţilor (de exemplu indivizii) asigură independenţa efectelor tratamentului. Prin urmare, alocarea complet aleatorie a tratamentului subiecţilor incluşi în experiment aduce beneficiul controlării erorii cauzate de variabilele neobservate omise, permiţând astfel măsurarea validă a efectului mediu al tratamentului, „orice corelaţie între tratament şi alţi factori cauzali fiind una pur întâmplătoare” (Gerber și Green 2012, 95). De asemenea, designul experimental permite utilizarea straturilor sau blocurilor de subiecţi care sunt împărţiţi (clasificaţi) în funcţie de una sau mai multe variabile explicative, astfel încât să permită eficientizarea observării efectelor tratamentului, dar şi posibilitatea de analiză a subgrupurilor (Gerber și Green 2012, 71–77).

De exemplu, să presupunem că vrem să explicăm efectul cauzal al transportului public gratuit asupra abandonului şcolar al elevilor dintr-o şcoală. Putem aloca aleatoriu jumătate din elevi grupului de tratament în care toţi elevii primesc transport gratuit, iar cealaltă jumătate de elevi îi vom aloca grupului de control, care nu primesc nimic. După o perioadă, de exemplu după un an şcolar, putem măsura rata de abandon şcolar în cele două grupuri. Acesta este un exemplu simplu de design experimental cu alocare complet aleatorie a subiecţilor studiaţi. Dat fiind însă că e posibil ca în realitate să existe şi constrângeri în selecţia subiecţilor, cum ar fi un număr mic de subiecţi (de exemplu, şcoala inclusă în studiu este în mediul rural şi are aproximativ 40 de copii) şi condiţii specifice (cum ar fi aceea că unii elevi au familii care pot să asigure transportul sau locuiesc foarte aproape de şcoală). De aceea, dacă selectăm pur aleatoriu, e posibil ca aceşti din urmă elevi să fie alocaţi prin hazard în grupul de tratament, în vreme ce ceilalţi să fie alocaţi grupului de control. În această situaţie ar fi posibil ca elevii care nu au nici un mijloc personal de transport sau locuiesc departe de şcoală să abandoneze, iar cei care au mijloace personale de transport sau locuiesc lângă şcoală să nu abandoneze, iar transportul public gratuit aplicat ca tratament acestui din urmă grup să nu producă nici un efect asupra ratei abandonului şcolar. Soluţia este folosirea alocării aleatorii în blocuri (subgrupuri sau straturi) de indivizi, chiar dacă aceste straturi nu sunt egale ca mărime. Astfel, ne putem baza pe împărţirea naturală a indivizilor în subgrupuri sau straturi care sunt relevante pentru dimensiuni de analiză (de exemplu distanţa faţă de şcoală, sau posibilitatea familiei de a asigura transportul elevului), şi putem aloca aleatoriu indivizii din fiecare din aceste straturi (blocuri), jumătate în grupul de tratament şi jumătate în grupul de control. În cazul acestui tip de alocare aleatorie a tratamentului există riscul de apariţie a unei contagiuni între subiecţii aceluiaşi grup care primesc tratamente diferite sau sunt alocaţi grupului de tratament şi cel de control. Elevii interacţionează între ei şi e posibil ca tratamentul aplicat unui subiect să influenţeze comportamentul altui subiect din acelaşi grup, dar care este în grupul de control. De aceea, putem decide să includem toţi subiecţii din acelaşi grup într-un singur grup (de tratament sau de control), după cum vom exemplifica mai jos.

Desigur, aceste tipuri de design experimental prezentate mai sus se bazează pe alocarea perfect aleatorie a indivizilor în grupul de tratament sau grupul de control,

dar şi pe estimarea rezultatului tratamentului la acelaşi nivel, individual. Însă, deseori suntem interesaţi de explicarea unor probleme care nu ne permit alocarea individuală a subiecţilor, ci alocarea tratamentului unui grup de subiecţi în întregimea sa, deşi noi estimăm rezultatul tratamentului la nivelul indivizilor. Acesta reprezintă designul experimental cu alocare cluster a subiecţilor. De exemplu, dorim să explicăm abandonul şcolar în localităţile din mediul rural dintr-o regiune de dezvoltare din România. Pentru a estima acest lucru, putem alege aleatoriu un număr de 100 de comune. Pe acestea le grupăm în funcţie de mărimea acestora alocând aleatoriu grupului de tratament jumătate din comunele care au aceeaşi caracteristică de mărime. În acest fel asigurăm controlul şi reducerea interferenţei altor posibile variabile cauză asupra efectului tratamentului (în acest exemplu ipotetic mărimea comunei poate influenţa abandonul deoarece în localităţile mai dispersate distanţa faţă de şcoală este mai mare decât în localităţile mai condensate). În fiecare categorie selectăm aleatoriu jumătate din comune în care asigurăm transportul gratuit timp de un an tuturor elevilor din satele comunei pentru a se deplasa la şcoală, iar pentru restul comunelor nu asigurăm transportul gratuit. La finalul anului putem evalua efectul tratamentului pentru fiecare comună din clusterele incluse în designul experimental.

Experimentele sunt, în general, de două tipuri: experimente controlate sau de laborator, respectiv experimente naturale. Experimentele de laborator, numite şi experimente randomizate controlate, sunt folosite în ştiinţele care permit aducerea în laborator a subiecţilor (animale, oameni) şi controlul unul număr cât mai mare de cauze alternative şi caracteristici ale subiecţilor. Alocarea subiecţilor în grupul de tratament şi cel de control se face aleatoriu, ceea ce permite ca aceste două grupuri să fie aproape similare şi comparabile. Diferenţele dintre aceste grupuri pot fi măsurate pe diverse dimensiuni relevante pentru analiză, de exemplu, gen, vârstă, venit, etnie, stare de sănătate etc. O altă metodă folosită pentru a asigura comparabilitatea subiecţilor este *matching*, prin care sunt măsurate caracteristici ale subiecţilor (sau grupurilor) pentru stabilirea gradului de convergenţă a acestora pe dimensiunile incluse în analiză (Lee 2016). Această tehnică este utilă şi în analiza unui număr foarte mic de subiecţi în care avem o alocare în perechi a acestora (câte unul în grupul de control şi grupul experimental) având astfel „o serie de mini-experimente, câte unul

pentru fiecare bloc" (Gerber și Green 2012, 71) sau strat folosit în alocarea aleatorie în funcţie de o caracteristică specifică, de exemplu, genul, vârsta, ocupaţia. Alocarea aleatorie în blocuri permite folosirea variabilelor explicative pentru a întări precizia cu care putem identifica efectul mediu al tratamentului, adică estimarea cu privire la puterea de explicare a efectului prin intermediul variabilei de tratament (cauză).

Pentru a putea calcula efectul mediu al tratamentului într-un design experimental calculăm mai întâi rezultatul potenţial folosind formula demonstrată în detaliu de Gerber și Green (2012, 23–25):

$$\tau_i = Y_i(1) - Y_i(0) \tag{2}$$

unde τ_i este efectul cauzal al tratamentului, $Y_i(1)$ este efectul din comuna i în care se aplică tratamentul, iar $Y_i(0)$ este efectul din comuna i unde nu se aplică tratamentul.

Pe baza acestei formule putem estima şi efectul mediu al tratamentului:

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}\tau_i = \mu_{Y(1)} - \mu_{Y(0)} \tag{3}$$

unde N este numărul de subiecţi, $\mu_{Y(1)}$ este valoarea medie a efectului tratamentului $Y_i(1)$, iar $\mu_{Y(0)}$ este valoarea medie a efectului lipsei tratamentului $Y_i(0)$.

Experimentele, atât cele de laborator, cât şi cele naturale au fost folosite în special în studiile medicale. Un tip aparte de experiment, cel natural, este deseori folosit pentru a studia probleme la nivel structural, agregat, care au loc meteoric, şi nu pot fi replicate în laborator.

Experimentul natural folosit de medicul englez John Snow a fost unul dintre primele studii care au folosit designul experimental pentru a identifica cauza pandemiei de holeră care afecta cea mai mare parte din Europa la mijlocul secolului al XIX-lea. Snow a cules date observaţionale din zone ale Londrei unde holera era activă şi din zone unde rata de mortalitate era mult mai redusă. A construit una dintre primele proiectări geografice a datelor culese, în prezent cunoscută sub numele

proiectare GIS, pentru a identifica posibilele cauze ale epidemiei. Dacă până atunci explicaţia ce obţinuse consens în rândul cercetătorilor şi al opiniei publice era că holera este cauzată fie de calitatea aerului (miasme), fie de anumite grupuri etnice, Snow s-a folosit de un design experimental natural, în care cercetătorul nu aplică tratamentul asupra unor subiecţi, nu alocă subiecţii în grupuri, iar aceştia nu au informaţii despre experiment, nici nu pot alege dacă sunt supuşi tratamentului sau nu, pentru a descoperi adevărata cauză a epidemiei: consumul de apă contaminată (Vinten-Johansen 2003; Dunning 2012).

Fără să depisteze agentul patogen care stă la baza infecţiei cu holeră, Snow a putut oferi o explicaţie inferenţială validă, chiar dacă designul experimental natural pe care l-a folosit nu i-a permis să se asigure că nu există alte cauze ale apariţiei efectului explica, folosind variabile de tip *confounder*[24] cum ar fi caracteristici medicale individuale sau de familie care ar fi putut favoriza infectarea sau decesul, consumul de apă sau alte lichide din alte zone decât gospodăria proprie, ceea ce ar fi putut duce la infectarea din alte surse, sau contactul interpersonal dat de mobilitatea indivizilor între regiuni, ce ar fi produs infectarea în afara gospodăriei sau cartierului. Însă, după cum explică Thad Dunning (2012), în designul experimental natural, unde cercetătorul nu poate asigura controlul asupra unui număr mare de variabile ce pot fi cauze alternative, putem folosi controlul observaţional în modelarea prin analizele de regresie, astfel încât să putem identifica posibile explicaţii alternative şi efectul produs de aceste variabile *confounder*.

Principala caracteristică a experimentelor naturale, pe care ne bazăm pentru a explica fenomenele este aceea a alocării subiecţilor ca şi când ar fi făcută aleatoriu: fără ca subiecţii să ştie şi fără ca ei să poată alege participarea sau nu în experiment. Experimentele naturale nu sunt create de cercetător, ci, de regulă, de o autoritate publică sau de un fenomen natural, de aceea influenţa preferinţelor subiective ale

[24] Putem aproxima în limba română prin termenul variabile omise; termenul de variabilă de control folosit deseori în studiile observaționale acoperă parțial înțelesul variabilei *confounder* în designul experimental. Aceste variabile sunt cele pe care uneori le pierdem din vedere, sau nu le includem în modelul explicativ dar care creează sau pot crea un efect în variabila dependentă pe care nu îl putem explica. Prin comparație, variabilele de control sunt acele variabile pe care nu le schimbăm, le ținem constante și nu influențează explicația pe care o dăm variabilei dependentă.

cercetătorului asupra problemei de studiat nu apare în aceste experimente ce apar și se dezvoltă într-un context real, natural, nu controlat.

De exemplu, o politică publică care este obligatorie pentru toți indivizii, dar produce efecte doar pentru unii. Într-un articol publicat recent, Simon Ress și Florian Spohr (2022) se folosesc de un experiment natural, implementarea în 2015 a legislației privind salariul minim în Germania. Opoziția partenerilor sociali, sindicate și patronate, la introducerea unui nivel minim al salariului a fost în mare parte determinată de potențiala subminare a procesului prin care partenerii sociali luau deciziile în mod consensual și creșterea rolului guvernului în probleme de relații de muncă pe care partenerii le gestionau în mod tradițional, dar și potențialul efect asupra scăderii numărului de membri de sindicat. Folosindu-se de această schimbare de politici publice, cei doi autori folosesc deci un design experimental natural și un sondaj de opinie panel cules de Agenția Federală a Muncii în perioada 2012-2017, cu 14362 respondenți, care permite deci realizarea unei măsurători pre-test și post-test atât pentru grupul de tratament, cât și pentru cel de control. Astfel angajații au fost alocați în acest studiu cu design experimental natural „aproape aleatoriu" în grupul de control (cei care aveau nivel salarial peste salariul minim, înainte și după adoptarea politicii publice) și cel de tratament (cei care aveau înainte de 2015 un salariu sub nivelul minim determinat de adoptarea politicii publice), în acest studiu fiind comparați membrii de sindicat cu cei care s-au retras și cei care intrat în sindicat.

Prin analiza de diferență-în-diferențe, putem măsura efectul de explicat nu ca o diferență între două grupuri ci ca o măsură a schimbării de la momentul pre-test la cel post-test. Potrivit lui Gerber și Green (2012, 98–102) estimatorul (metoda prin care măsurăm efectul tratamentului) diferență-în-diferențe produce o variantă mai mică decât diferența de medii, care nu folosește informații de tip pre-test. Analiza lui Ress și Spohr (2022) a putut releva inexistența vreunui efect al introducerii salariului minim asupra predispoziției de a ieși din sindicat, observând chiar un efect pozitiv în ce privește intrarea unor noi membri în sindicat.

2.5. Experimente integrate în sondaje – Tehnica de numărare a itemilor

În sondaje (față în față, autoaplicate, telefonice sau pe internet) respondenţii vor fi precauţi când trebuie să dea detalii personale sau să răspundă la întrebări sensibile. De exemplu, dacă facem un studiu în care încercăm să măsurăm comportamentul de evitare a plății impozitelor şi taxelor, sau un studiu despre consumul de droguri, ambele comportamente fiind incriminate legale, ne putem aştepta ca respondenţii să nu fie deschişi şi onești. Prin urmare, dacă vom folosi o întrebare directă în care le cerem să ne spună dacă au evitat în ultimii doi ani plata taxelor şi impozitelor, numărul celor care vor răspunde afirmativ poate fi mai scăzut decât ar fi în realitate. Acest rezultat distorsionat este foarte probabil să apară chiar şi atunci când asigurăm subiecţii că răspunsurile lor sunt întru totul anonime, spre exemplu prin autocompletarea chestionarului şi introducerea lui într-o urnă sigilată, ce conţine alte asemenea chestionare anonime. Cu toate acestea, studierea problemelor sensibile este important a fi făcută cu metode ştiinţifice.

Cum putem studia, totuşi, aceste probleme sensibile? Dacă chestionarul cu întrebări directe poate avea erori mari de dezirabilitate socială, de frica pedepsirii comportamentului ilegal, nici interviul aprofundat nu ne poate ajuta foarte mult pentru că nu ne permite să estimăm cât de extins este acest comportament pentru probleme sensibile, chiar dacă anonimitatea va putea fi percepută în mod mai puternic de subiecţi prin durata mai mare a interacţiunii cu intervievatorul şi încrederea investită în acesta odată cu acceptarea realizării unui interviu de durată. Soluţia o reprezintă metoda experimentală, ce foloseşte alocarea aleatorie a subiecţilor în grup experimental, sau de tratament, respectiv grup de control. Asupra grupului experimental aplicăm tratamentul, în vreme ce grupul de control nu primeşte acel tratament. Alocarea aleatorie a subiecţilor în cele două grupuri permite un grad ridicat de comparabilitate a acestora, subiecţii celor două grupuri fiind la fel de omogeni sau eterogeni.

Însă, designul experimentul de laborator nu oferă același grad de generalizare a concluziilor, atunci când evaluăm comportamente sociale, precum sondajul de opinie efectuat pe eșantion reprezentativ. Motivul este acela că în laborator rareori putem aduce un număr mare (peste 800) de subiecți umani, selectați aleatoriu. Costurile ar fi considerabil mai mari decât atunci când subiecții sunt chestionați prin intermediul unui sondaj de opinie, față în față sau telefonic. Din acest motiv, pentru a păstra avantajele metodologice de identificare a efectului produs de o cauză independentă, pe care experimentele sunt capabile să le ofere cu ajutorul alocării aleatorii a subiecților, integrăm experimentul în chestionar și în sondaj, cu ajutorul unor liste de itemi, vignete, stimuli vizuali. Acest tip de design de cercetare se numește experiment integrat în sondaje.

În mod particular, vom detalia mai jos o tehnică specifică de integrare a experimentului în sondaj, numit experiment cu liste, sau tehnica de numărare a itemilor (în literatura de limbă engleză poartă numele de *item count technique* sau *unmatched count technique*). Există mai multe tehnici experimentale incluse în sondaje, tipurile acestora fiind în continuă dezvoltare: *vignete factoriale* (Auspurg și Hintz 2015), *conjoint* (Hainmueller, Hopkins, și Yamamoto 2014), *priming* (Valentino et al. 2008), *endorsement* (Bullock, Imai, și Shapiro 2011), *person count technique* (Wolter 2019), *crosswise model* (Jann, Jerke, și Krumpal 2012), *randomized response technique* (J. A. Fox și Tracy 1986), *item sum technique* (Trappmann et al. 2014), *item count technique* (Kuklinski, Cobb, și Gilens 1997; Droitcour et al. 2004; Imai 2011). Ne vom referi doar la această ultimă tehnică, eficiența ei fiind evaluată în literatura de specialitate ca fiind superioară multora dintre celelalte metode experimentale înglobate în sondaj pentru evaluarea problemelor sensibile (Holbrook și Krosnick 2010; Wolter și Laier 2014).

Tehnica de numărare a itemilor reprezintă o tehnică specifică de experiment integrat în sondaje (Corstange 2008; Blair și Imai 2012; Blair, Coppock, și Moor 2020), ce permite măsurarea unor elemente sensibile, cu ajutorul unui design relativ simplu și ușor de pus în practică. Acest design permite atât utilizarea diferențelor dintre subiecți (un subiect este alocat aleatoriu doar unui singur grup, de control sau cel de tratament), cât și a diferențelor în interiorul subiecților (un subiect poate fi în același timp, dar pe problemă măsurată diferită, atât în grupul de control, cât și în grupul de tratament). Acest din urmă design permite utilizarea unui număr mai mic de subiecți

în eșantionul total, pentru măsurarea mai multor probleme sensibile prin metoda experimentală.

Respondenții din eșantionul sondajului sunt alocați aleatoriu grupului de tratament sau celui de control pentru fiecare din problemele sensibile pe care ne propunem să le măsurăm în sondaj. De exemplu, dacă măsurăm trei probleme sensibile, vom aloca aleatoriu cei 1000 de respondenți din eșantion, în grupul de tratament (500) și grupul de control (ceilalți 500).

Tehnica de numărare a itemilor se bazează pe dorința subiecților de a avea o anonimitate maximă posibilă și de a nu lăsa urme prin răspunsul oferit. Astfel, vom asigura această condiție prin alocarea aleatorie a subiecților, prin furnizarea unei liste de afirmații, asupra cărora subiecții nu trebuie să se pronunțe în mod direct (de exemplu dacă a făcut sau nu acel lucru) ci raportându-se la alte afirmații, prin anonimizarea completă a răspunsului la întrebare și la chestionar (de exemplu, pentru un surplus de anonimitate putem oferi respondenților posibilitatea de a introduce chestionarul completat într-o urnă mobilă[25]).

Configurarea listei de itemi poate fi precum în lista de mai jos. Poziția itemilor în listă este aleatorie (aceștia sunt rotiți) pentru fiecare respondent, astfel încât să controlăm pentru distorsiunea indusă de poziția în listă a fiecărui item. În plus, atunci când avem mai multe tratamente (probleme sensibile pe care dorim să le măsurăm) poziția listelor cu itemi din chestionarul alocat fiecărui respondent este aleatoare. Prin urmare, putem folosi un design experimental *within subjects* (Charness, Gneezy, și Kuhn 2012), în care fiecare subiect este în grupul de tratament pentru un subiect sensibil, respectiv în grupul de control pentru un alt subiect sensibil. De aceea, numărul total de respondenți necesari pentru a implementa un design experimental pentru mai multe subiecte sensibile poate fi mult mai mic decât atunci când folosim un design *between subjects*, în care fiecare subiect nu poate fi inclus decât într-un singur grup.

[25] Aceste metode au fost implementare de studiul pe clientelism electoral efectuat de Aurelian Muntean în România între 2013-2016. Urna mobilă a fost folosită în valul de cercetare din 2013, așa cum se observă în fotografia de pe coperta acestei cărți.

În tehnica de numărare a itemilor, lista de itemi are următoarea configuraţie:

- un item cu prevalenţă ridicată
- un item cu prevalenţă ridicată, opus primului
- un item cu prevalenţă scăzută (improbabil)
- un item ce conţine tratamentul (fenomenul sensibil)

Întrebarea ce conţine doar itemii cu prevalenţă ridicată şi cel cu prevalenţă scăzută se aplică subiecţilor din grupul de control. Întrebarea ce conţine toţi itemii plus itemul cu tratament se aplică doar subiecţilor alocaţi aleatoriu în grupul de tratament: ***Câte (nu care) din aceste afirmaţii sunt adevărate în cazul dumneavoastră?***

În cazul în care am folosi doar itemi cu prevalenţă ridicată într-o listă pentru tehnica de numărare a itemilor, ar fi posibil ca respondenţii să fie obligaţi să răspundă afirmativ. Astfel, am avea prea multe erori cauzate de aceste efecte de plafonare a răspunsurilor, numite efecte *ceiling*. Corolarul este situaţia în care am folosi doar itemi cu prevalenţă redusă, caz în care respondenţii ar fi obligaţi să răspundă negativ. În această situaţie efectele ar fi opuse, ar conduce răspunsurile la cota zero: efecte de tip *floor*. În ambele situaţii anonimitatea furnizată de lista de itemi ar fi compromisă. De aceea, soluţia recomandată este de a folosi în liste itemi negativ corelaţi (Glynn 2013). Media pe răspunsurile din grupul de control ar trebui să fie 1 sau foarte aproape de 1 pentru a creşte precizia listei. Pretestarea listelor (fără itemul ce conţine tratamentul) ne permite să ne asigurăm că afirmaţiile sunt corect înţelese de respondenţi, şi să identificăm acele perechi de afirmaţii cu prevalenţă ridicată care au o corelaţie negativă.

Analiza răspunsurilor din grupul de control şi cel de tratament ne permite să calculăm efectul mediu al tratamentului (ATE), într-un design experimental cu alocarea aleatorie completă a subiecţilor, prin diferenţa mediilor calculate pentru fiecare listă.

Tabel 2.3 Calcul al diferenţelor dintre grupul de tratament şi cel de control

Răspunsurile subiecţilor din grupul de tratament pentru lista 1	Răspunsurile subiecţilor din grupul de control pentru lista 1
1	1
2	1
1	1
2	1
1	1
1	1
2	1
2	1
2	1
Media=1,55	Media=1

Calculând diferența mediilor din distribuția ipotetică de mai sus vom obține ATE = 1,55 - 1 = 0,55. Prin urmare, vom concluziona că 55% dintre subiecții din grupul de tratament au răspuns pozitiv la acesta, deci au experimentat subiectul sensibil. Pentru a testa dacă sunt diferențe semnificative statistic între mediile calculate pe grupul experimental și cel de control vom raporta testul *t* de comparație a mediilor pentru eșantioane independente, la un nivel de semnificație $p \leq 0.05$ și limitele intervalului de încredere de 95%. Desigur, spre deosebire de exemplul ipotetic din tabelul 2.3, în realitate, este posibil ca unii dintre respondenții din grupul de control să răspundă 0 (adică nici una dintre afirmații nu sunt adevărate), la fel cum în grupul de tratament să avem subiecți care vor răspunde că 3 afirmații sunt corecte. Însă, dacă designul experimentului pe liste asigură controlul distorsiunilor și erorilor de design, de listă de item, numărul de răspunsuri aberante va fi foarte mic, deci impactul asupra calității estimării efectului va fi redus. Pe lângă controlul efectelor *ceiling* și *floor*, prezentate mai sus, în următoarele paragrafe vom discuta despre alte mecanisme de control al erorilor și vom oferi un exemplu de design experimental pe liste.

După cum vom explica mai jos, aceste erori pot fi reduse și ținute sub control prin pretestare, controlarea efectelor cauzate de formulare, testarea efectelor de design, dar și prin randomizarea listelor și itemilor. Efectele acestor erori, deși ținute

sub control, pot fi exacerbate de zgomotul de fond (*noise* în literatura de specialitate de limbă engleză) (Gerber și Green 2012). Atunci când problema pe care dorim să o explicăm are o variație mică (de exemplu studierea efectului unui vaccin împotriva unei boli rare care are o rata redusă de apariție, sau utilizarea amenințării cu arme de foc ca strategie de clientelism electoral în alegerile din România) zgomotul de fond este redus, pentru că nivelul de bază este constant, deci chiar și un efect mic cauzat de tratament poate fi identificat prin experiment, atunci când controlăm pentru factori alternativi. Dacă problema pe care o analizăm are o variație mare (de exemplu efectul unor calmante asupra durerii musculare, care are o rată mare de apariție în populație, sau acordarea de favoruri administrative de către funcționarii publici care îl susțin pe primarul aflat în funcție, ca formă de clientelism electoral în alegerile din România), atunci zgomotul de fond este mai mare, deoarece există mulți factori și caracteristici (de exemplu, predispoziții de sănătate, diferențe între grupuri, respectiv diferențe la nivel de votant, susținător, partid politic etc.) care pot influența aceste efecte, zgomotul de fond este mare iar o schimbare redusă a efectului, cauzată de tratamentul din grupul experimental va fi greu de depistat.

De asemenea, vom testa diferenţele dintre grupul experimental şi cel de control. pe variabile de control, socio-demografice, cum ar fi genul, vârsta, etnicitatea, clasa socială, pentru a evalua potenţialele dezechilibre de distribuţie între cele două grupuri. Pentru a testa validitatea listelor experimentale, putem folosi testul dezvoltat de identificare a unor posibile efecte de design al listelor (Blair și Imai 2012). Pentru a calcula aceste efecte vom utiliza pachetul R *list* (Blair și Imai 2010). Putem astfel să identificăm în ce măsură includerea itemului sensibil (de tratament) în lista experimentală schimbă răspunsul subiecţilor pentru itemii de control; şi dacă diferenţa dintre răspunsurile din lista de control şi cea de tratament sunt mai mari decât 1. Acest test permite identificarea efectelor de design şi în cazul unor diferenţe reduse între lista de tratament şi cea de control.

În continuare vom prezenta pe scurt un exemplu de utilizare a tehnicii de numărare a itemilor, folosită în studiul pe clientelismul electoral din Romania, efectuat în perioada 2013-2016 de Aurelian Muntean împreună cu studenţi din SNSPA. În sondajul efectuat în zona rurală din Buzău şi Teleorman, au fost incluşi

1497 de respondenţi din 86 de comune, cu un eşantion multistratificat, cu selecţia probabilistă a subiecţilor.

Pentru un respondent formularea listelor de itemi a urmat exemplul de mai jos, în care ilustrăm pentru două subiecte sensibile, deci pentru două experimente diferite. Aşadar, acest respondent a fost alocat aleatoriu grupului de tratament pentru prima lista experimentală, şi grupului de control pentru a doua listă. Astfel, acest respondent a primit următoarele întrebări:

Vă voi citi câteva afirmații legate de alegerile prezidențiale recente, din luna Noiembrie 2014. Pentru fiecare afirmație aș dori să îmi spuneți câte din aceste evenimente s-au întâmplat și aici, în comuna dumneavoastră. Nu trebuie să ne spuneți care din ele au avut loc, ci doar câte dintre ele.

[2 EXEMPLE] Ca să înțelegem mai bine vă rog să-mi spuneți câte din următoarele 3 lucruri sunt adevărate?

- *Acum este vară.*
- *Avem biserică în comuna noastră.*
- *În comuna noastră avem magazin.*

###

- *Unii oameni din comună se ocupă cu agricultura.*
- *Traian Băsescu s-a născut în comuna noastră.*
- *Comuna noastră este situată într-o zonă de munte.*

Lista 1) Vă rog să vă gândiți la ultimele alegeri prezidențiale din Noiembrie 2014. Câte din aceste 4 lucruri vi s-au întâmplat?

- *Când am mers să votez, am stat la coadă.*
- *Am văzut oameni care erau beți când au mers la vot.*
- *Mi-a fost teamă că mi se vor tăia ajutoarele de la primărie dacă nu votez cu cine trebuie.*
- *La noi în comună campania s-a desfășurat fără violențe.*

0 – Nici unul 1 – Unul 2 – Două 3 – Trei 4 – Patru 99 - Nu știu

Lista 2) Vă rog să vă gândiți la ultimele alegeri prezidențiale din Noiembrie 2014. Câte din aceste 3 lucruri vi s-au întâmplat?

- *La secția de vot, am văzut observatori din Germania care au venit să vadă cum se desfășoară alegerile.*

• *Când am ajuns la secția de votare, secția era deschisă.*

• *Traian Băsescu a vizitat comuna noastră în ziua alegerilor.*

0 – Nici unul 1 – Unul 2 – Două 3 – Trei 99 - Nu știu

Un al doilea respondent a fost alocat aleatoriu grupului de control pentru prima listă, respectiv grupului de tratament pentru a doua listă. Pentru grupul de tratament formularea listelor de itemi a urmat exemplul de mai jos:

Vă voi citi câteva afirmații legate de alegerile prezidențiale recente, din luna Noiembrie 2014. Pentru fiecare afirmație aș dori să îmi spuneți câte din aceste evenimente s-au întâmplat și aici, în comuna dumneavoastră. Nu trebuie să ne spuneți care din ele au avut loc, ci doar câte dintre ele.

[2 EXEMPLE] Ca să înțelegem mai bine vă rog să-mi spuneți câte din următoarele 3 lucruri sunt adevărate?

• *Acum este vară.*

• *Avem biserică în comuna noastră.*

• *În comuna noastră avem magazin.*

###

• *Unii oameni din comună se ocupă cu agricultura.*

• *Traian Băsescu s-a născut în comuna noastră.*

• *Comuna noastră este situată într-o zonă de munte.*

L2) Vă rog să vă gândiți la ultimele alegeri prezidențiale din Noiembrie 2014. Câte din aceste 3 lucruri vi s-au întâmplat?

• *Când am mers să votez am stat la coadă.*

• *Am văzut oameni care erau beți când au mers la vot.*

• *La noi în comună campania s-a desfășurat fără violențe.*

0 – Nici unul 1 – Unul 2 – Două 3 – Trei 99 - Nu știu

L3) Vă rog să vă gândiți la ultimele alegeri prezidențiale din Noiembrie 2014. Câte din aceste 4 lucruri vi s-au întâmplat?

• *La secția de vot, am văzut observatori din Germania care au venit să vadă cum se desfășoară alegerile.*

• *Când am ajuns la secția de votare, secția era deschisă.*

• *Traian Băsescu a vizitat comuna noastră în ziua alegerilor.*

• *Cineva de la primărie mi-a spus cum ar fi bine să votez ca să mă servească și în viitor.*

0 – Nici unul 1 – Unul 2 – Două 3 – Trei 4 – Patru 99 - Nu știu

Pentru aceste două liste efectul mediu al tratamentului a fost 0,11, respectiv 0,07, prin urmare, aproximativ 11% dintre respondenţi au afirmat că au experimentat presiuni cu privire la vot din perspectiva pierderii ajutorului social, în vreme ce 7% dintre respondenţi au raportat presiuni la vot din partea unor funcţionari publici pentru a beneficia de servicii administrative corecte.

Pentru a testa diferenţele dintre cele două eşantioane principale folosite (cele două tipuri de chestionar) în ce priveşte subiecţii, putem calcula mediile unor variabile socio-demografice, care, în studiul indicat mai sus arată astfel:

Tabel 2.4 Test de distribuire a respondenţilor pe grupuri

	Versiunea 1	Versiunea 2	Diferență	Nivel de semnificație
Vârstă	50.11	50.46	0.03	0.69
Femei (dummy)	0.50	0.50	0.00	0.93
Etnie Romă (dummy)	0.20	0.20	0.00	0.74
Săraci (3 categorii)	0.57	0.60	0.03	0.19

Notă: Cifrele din fiecare celulă reprezintă proporții

Pentru testul Blair-Imai (2012) de identificare a efectelor de design instalăm și folosim pachetul *list* şi pachetele pe care acesta le solicită în mod automat pentru a permite analizele solicitate. Putem rula comanda de mai jos pentru a obţine rezultatul testului pentru fiecare listă alocată grupului de tratament şi celui de control.

```
library(list)

deftest.socasist <- ict.test(DATEBZTR$control, DATEBZTR$tratament, J = 3,
alpha = 0.05, gms = TRUE, pi.table=TRUE)

print(deftest.socasist)
```

```
      Estimated population proportions
                          est.   s.e.
      pi(y = 0, t = 1)  0.0000 0.0000
      pi(y = 1, t = 1) -0.0488 0.0341
      pi(y = 2, t = 1)  0.0000 0.0000
      pi(y = 3, t = 1)  0.0000 0.0000
      pi(y = 0, t = 0)  0.0000 0.0000
      pi(y = 1, t = 0)  1.0000 0.0000
      pi(y = 2, t = 0)  0.0488 0.0341
      pi(y = 3, t = 0)  0.0000 0.0000

      Bonferroni-corrected    p-
 value

      If this value is below alpha, you reject the null
 hypothesis of no design effect. If it is above alpha, you
 fail to reject the null.

      Sensitive Item 1 Sensitive Item 2
          3.168142e-11     1.598469e-01
```

Pentru a putea explica tipul de comportament sensibil măsurat prin designul experimental şi a identifica modalitatea în care variază efectele tratamentului, putem efectua analize de regresie, prin includerea rezultatelor variabilei tratament (efect) în analize de regresie logistică. De asemenea, putem include şi factori de interacţiune pentru a identifica modul în care atribute ale subiecţilor în funcţie de tratament, permiţând astfel, aşa cum precizează Gerber şi Green (2012) să observăm designul experimental la nivel de subgrupuri (de exemplu introducând în modelele noastre explicative variabile cauză suplimentare, precum tratamentXetnicitate;

tratamentXdeținătorul-funcției; tratamentXpartid; tratamentXgen; tratamentXvârstă; tratamentXclasa-socială etc.). Nu în ultimul rând, putem adăuga în analiză modele care să identifice diferențele nu doar la nivel de individ, dar și la nivel de localitate, utilizând modele ierarhice de analiză , adăugând în baza noastră de date rezultată în urma sondajului cu liste experimentale, informații de dezvoltare economică și socială sau de competiție politică, de la nivel mediu agregat (de exemplu localitate).

3. Metode de analiză cantitativă

3.1. Datele: surse ale datelor, tipuri de date

Sistemele sociale în care trăim sunt din ce în ce mai greu de înţeles fără o cantitate imensă de date. În consecinţă, organizaţiile din diferite domenii au înţeles necesitatea colectării datelor în procesul de elaborare a strategiilor de afaceri, în procesul de elaborare a politicilor publice, în formarea deciziilor privind viitoarea politică monetară, în cercetarea ştiinţifică, dar şi în strategii militare, tehnologice şi strategii de campanie electorală.

Astfel, colectarea datelor reprezintă o parte crucială a procesului de procesare şi analiză al datelor. Colectarea eficientă poate oferi o serie de informaţii esenţiale pentru a răspunde la întrebări de cercetare pentru a analiza performanţa afacerii, pentru a decide necesitatea elaborării unei politici publice, dar şi pentru a prezice tendinţele, acţiunile şi scenariile viitoare.

Procesul de colectare a datelor diferă de la domeniu la domeniu. Spre exemplu, mediul privat colectează date prin mai multe surse: prin sistemele software cu care lucrează colectează în mod regulat date despre clienţi, angajaţi, vânzări şi alte aspecte ale operaţiunilor comerciale; prin diseminarea de sondaje la nivelul publicului larg; dar şi prin utilizarea reţelelor sociale pentru a obţine feedback de la clienţi. Pe de altă parte, în mediul academic şi de cercetare, colectarea datelor este adesea un proces mai specializat, în care cercetătorii creează şi implementează instrumente pentru a colecta seturi specifice de date. Cu toate acestea, atât în contextul mediului public sau privat, cât şi în cel al cercetării, datele colectate trebuie să fie exacte pentru a se asigura validitatea rezultatelor.

În studiile din ştiinţele sociale folosim date extrem de diverse, nu doar din punct de vedere al complexităţii realităţii pe care acestea o măsoară, ci şi din punct de vedere al surselor producerii datelor. Nu întotdeauna vom culege propriile date, prin urmare demersul nostru ştiinţific nu va mai fi unul deductiv (pe baza teoriei

formulăm ipotezele şi le vom testa pe datele pe care le culegem şi le codificăm în modul predeterminat de construcţia ipotezelor), ci inductiv (pe baza datelor culese şi pe baza analizei acestora vom produce generalizări teoretice). Deseori vom încerca să combinăm aceste două abordări, să folosim date deja culese, la care să adăugăm noi informaţii relevante pentru subiectul cercetat.

Comunicarea şi transmiterea digitală a informaţiilor facilitează culegerea unui număr din ce în ce mai mare de date de către diverşi actori şi instituţii (guverne, agenţii regionale sau internaţionale, universităţi şi institute de cercetare, companii, organizaţii non-guvernamentale etc.). Prin urmare, formele în care putem găsi aceste date sunt extrem de diverse. Uneori vom identifica date relevante în surse diferite, deci va trebui să prelucrăm suplimentar aceste informaţii. Prelucrarea datelor prin curăţarea lor (filtrarea informaţiilor / variabilelor relevante pentru analiză, filtrarea observaţiilor lipsă, corectarea erorilor de codificare) anticipează analiza propriu zisă. Datele trebuie transformate pentru a fi aduse în formatul necesar analizei în conformitate cu ipotezele noastre. Astfel, putem utiliza proceduri specifice de transformare a datelor, cum ar fi simple recodificări, de exemplu transformarea codurilor aberante care sunt cauzate de eroarea de operator. Aceste erori pot fi reduse prin utilizarea procedurilor automate de colectare a răspunsurilor, cum ar fi CATI (sondaj realizat telefonic), CAPI (sondaj față în față folosind o tableta) sau chestionarele online auto-administrate. Putem apela și la recodificări complexe, cum ar fi transformările factoriale ale variabilelor, scalarea multidimensională a variabilelor sau transformarea valorilor lipsă (non-răspunsurilor din baza de date) prin imputare multiplă, care permite înlocuirea acestora cu valori care aproximează distribuţia datelor valide.[26]

Uneori datele pe care le preluăm din alte surse sunt prezentate într-un format care nu este în mod necesar pregătit pentru a fi prelucrat prin programul statistic pe care îl folosim. Deşi multe programe statistice au dezvoltat opţiunile de importare a datelor din formate proprietare ale altor programe sau companii, uneori importarea poate produce modificări ale numelor variabilelor, etichetelor acestora sau ale

[26] Tehnicile de imputare și analiză a valorilor lipsă s-au dezvoltat foarte mult în ultimii ani; putem recomanda cartea lui Roderick Little și Donald Rubin (2020).

valorilor. De aceea este recomandabilă verificarea și testarea datelor prin analize simple de frecvenţe sau de tabele de contingenţă. Despre aceste tehnici vom discuta în acest capitol. Pentru transformarea datelor din formatul în care acestea au fost publicate în formatul specific programului în care dorim să lucrăm, putem folosi și programe dedicate, printre care nominalizăm StatConverter (un program gratuit) sau StatTransfer (un program contra cost).[27] Pentru programele R și Stata, pe care le vom folosi pentru a ilustra analiza datelor reale, există pachete speciale care permit importarea directă a unor baze de date în formate diferite, și chiar salvarea exportarea lor în alte formate decât cele specifice programului. Pentru operaţiunile de curăţare și pregătire a datelor putem folosi fie programe cum ar fi R, Stata, sau Excel, fie programe specializate în minarea datelor, organizarea și transformarea acestora, precum RapidMiner. Pentru acest din urmă program, recomandăm manualul publicat de Mircea Comșa (2022), care prezintă în detaliu modul în care putem folosi RapidMiner pentru pregătirea datelor în vederea analizelor statistice.

Datele pot avea mai multe proprietăţi și în funcţie de aceste proprietăţi pot fi aplicate tehnici de analiză specifice. Vom lua în considerare trei niveluri de date: microdate (la nivel dezagregat, de exemplu indivizi), macrodate și metadate (la nivel agregat). Microdatele reprezintă date privind unităţile statistice individuale (persoane fizice, gospodării, agenţi economici). Macrodatele, în schimb, descriu obiecte agregate. Ele sunt produse prin combinarea scorurilor unităţilor unui set de unităţi pentru a obţine un singur scor pentru fiecare set. Macrodatele sunt astfel frecvenţe, sume, medii etc. Multe tehnici statistice au ca scop agregarea seturilor de observaţii din motive descriptive sau rezumative. Macrodatele sunt supuse unei analize ulterioare. Metadatele sunt date ce permit clasificarea, organizarea și stocarea altor date (de obicei în format digital). Metadatele sunt definite ca fiind date despre date sau informaţii despre informaţii. Prin metadate sunt furnizate informaţii descriptive cu privire la producătorul, conţinutul, calitatea sau starea unor anumite obiecte. Aceste obiecte pot fi imagini, cărţi sau seturi de date digitale, etc. Metadatele ne pot ajuta să decidem ce fel de întrebări de cercetare pot fi puse în mod legitim.

[27] StatConverter poate fi descărcat de aici: https://roda.github.io/StatConverter. StatTransfer poate fi descărcat de aici: https://stattransfer.com.

Mai mult, datele pot fi primare sau secundare. Datele primare sunt acele date care sunt colectate prin aplicarea unor chestionare, a unor experimente, interviuri și alte metode de colectare de date. Datele secundare sunt acele date colectate din studii, sondaje sau experimente care au fost conduse de alte persoane sau pentru alte cercetări. De regulă, colectarea de date primare este foarte costisitoare și consumă foarte mult timp. De aceea, de foarte multe ori vom apela la surse de date secundare pentru a ne îndeplini obiectivele de cercetare.

Printre cele mai cunoscute surse prin care putem obține cantități mari de date în funcție de domeniu putem identifica câteva din cele de mai jos. În anexa de la finalul acestei cărți vom furniza o listă exhaustivă de surse de date.

Pentru cercetarea fenomenelor sociale, politice și economice:

- **World Values Survey (WVS)** și **European Values Study (EVS)** sunt două proiecte de cercetare care colectează date despre valorile și convingerile oamenilor, cum se schimbă acestea în timp și impactul social și politic al acestor convingeri. EVS a fost lansat în 1981 iar WVS a fost lansat ca parte din acest prim val de studiere comparată a cetățenilor Uniunii Europene (Voicu și Voicu 2002). EVS și WVS măsoară, monitorizează și analizează: sprijinul pentru democrație, toleranța față de străini și minorități etnice, sprijinul pentru egalitatea de gen, rolul religiei și schimbarea nivelurilor de religiozitate, impactul globalizării, atitudinile față de mediu, muncă, familie, politică, identitate națională, cultură, diversitate, insecuritate și bunăstare subiectivă. Datele colectate prin intermediul sondajelor EVS și WVS oferă informații pentru factorii de decizie politică și pentru instituții democratice. Aceste date comparative au fost folosite, de exemplu, pentru a înțelege mai bine motivațiile din spatele unor evenimente precum Primăvara Arabă, tulburările civile din Franța din 2005, genocidul din Rwanda din 1994 și războaiele și revoltele politice din Iugoslavia din anul 1990. Datele colectate în European Values Study pot fi accesate și descărcate gratuit de aici: https://europeanvaluesstudy.eu/about-evs/, iar datele colectate în cadrul World Values Survey pot fi accesate și descărcate gratuit la următorul link: https://www.worldvaluessurvey.org/wvs.jsp.

- **European Social Survey (ESS)** este un proiect european de cercetare care colectează date despre atitudinile, credințele și modelele de comportament ale diferitelor populații din Europa. Datele colectate pot fi accesate și descărcate gratuit de la adresa https://www.europeansocialsurvey.org/.
- **European Elections Studies (EES)** oferă date despre comportamentul electoral. Acoperă subiecte precum evoluția unei comunități politice în Uniunea Europeană, percepțiile și preferințele cetățenilor cu privire la regimul politic al UE și evaluarea performanței politice a UE. Oferă oportunități generoase de analize comparative și longitudinale în statele membre ale UE. Datele colectate pot fi accesate și descărcate gratuit de la https://www.gesis.org/en/services/finding-and-accessing-data/international-survey-programs/european-election-studies.
- **Sondajele Eurobarometru** efectuate la nivelul Uniunii Europene începând cu anul 1973. Aceste sondaje se realizează de mai multe ori pe an, în varianta extinsă primăvara și toamna, pentru măsurarea comparativă (longitudinală și națională) a unor probleme de interes general sau specific pentru țările membre ale UE, sau în varianta scurtă, pentru probleme ad-hoc, realizată telefonic (numite Flash Eurobarometer). Eurobarometrul reprezintă o sursă importantă de informații la nivel individual. Poate fi descărcat de aici: https://europa.eu/eurobarometer/screen/home.
- **Institutul Național de Statistică al României (INS)**. INS oferă prin serviciul gratuit Tempo Online, o serie de date statistice la nivel național, regional, județean, cu serii de timp: http://statistici.insse.ro:8077/tempo-online/

Pentru cercetarea indicatorilor macroeconomici și microeconomici putem menționa:

- **World Bank Open Data**: https://data.worldbank.org/
- **International Monetary Fund (IMF)**: https://www.imf.org/en/Data.
- **OECD Data**: https://data.oecd.org/
- **Thomson Reuters** – contra cost

Mai mult, o serie de date pot fi descărcate pentru replicare accesând paginile oficiale ale depozitelor de date care sprijină conceptul de *open science*. Aceasta este o

inițiativă care își propune să elimine barierele pentru partajarea oricărui tip de rezultate, resurse, metode sau instrumente, în orice etapă a procesului de cercetare. Astfel, datele şi analizele folosite în procesul de cercetare au fost publicate gratuit de o serie de organizaţii în aşa numitele depozite de date (în engleză *data repository*). Prin aceste depozite de date, pe lângă setul de date folosit în analiză, cercetătorii oferă spre descărcare şi codul analizei statistice folosite. Accesul la astfel de date permite aprofundarea unor cunoştinţe prin replicarea cercetării.

Câteva exemple de depozite de date sunt:

- **Harvard Dataverse**: https://dataverse.harvard.edu/dataverse/harvard/
- **Mendeley Repository**: https://data.mendeley.com/
- **OFS Repository**: https://osf.io/
- **Github**: https://github.com/

În capitolul 3 vom prezenta tipuri de analiză a datelor cantitative, cum ar fi analiza univariată, bivariată și multivariată. Astfel, vom ilustra modul în care putem prelucra datele și explica distribuția observațiilor, relațiile între variabile și vizualizarea grafică a datelor și a analizelor inferențiale. În acest capitol vom exemplifica aceste analize folosind programul de analiză statistică R, sub implementarea și interfața RStudio. Acest capitol va oferi o introducere cât mai simplă și prietenoasă în folosirea acestui program, astfel încât inițierea în limbajul de programare R să fie mai ușor de înțeles. Întrucât unii dintre practicieni și studenți nu folosesc R, în capitolul 4 vom exemplifica în programul statistic Stata aceleași analize prezentate în capitolul 3, folosind același set de date și modele explicative. În capitolul 5 vom ilustra aceste modele de analiză folosind programul Excel. Deoarece concluziile și modelele pe care le identificăm în analizele noastre statistice sunt mai ușor de înțeles cu ajutorul imaginilor, prin reprezentări grafice, capitolul 6 este dedicat aprofundării vizualizării grafice a datelor folosind R și pachetul ggplot2.

3.1.1. Validitate și fidelitate. Niveluri de măsurare

Opţiunile metodologice şi de design de cercetare, dar şi utilizarea şi operaţionalizarea variabilelor şi a unităţilor de analiză conduc la culegerea unor tipuri specifice de informaţii empirice. Indiferent de abordare (cantitativă, calitativă, comparativă) datele empirice culese pot suferi de pe urma erorii de măsurare. Această eroare se referă la problemele pe care cercetătorul le poate avea atunci când încearcă să măsoare concepte nu prin ele însele (de exemplu greutate, venit), ci prin concepte indirecte (de exemplu, măsurarea indirectă a conceptului de democraţie reprezentativă).

Instrumentele folosite pentru culegerea şi analizarea datelor empirice trebuie să îndeplinească două condiţii pe baza cărora ne asigurăm că datele culese sunt utile pentru a studia problema de cercetare: validitatea şi fidelitatea. Validitatea se referă la capacitatea instrumentului folosit de a măsura ceea ce ne dorim să măsurăm. Astfel, prin evaluarea validităţii unui instrument putem să identificăm dacă operaţionalizarea conceptului sub forma unei variabile (sau mai multor variabile care măsoară părţi separate din acel concept) este capabilă să măsoare conceptul pe care dorim să îl măsurăm (Frankfort-Nachmias, Nachmias, și DeWaard 2015, 131). Acest tip de validitate mai este cunoscută sub numele de validitate de conţinut. O putem evalua pentru instrumentul / întrebarea folosită, prin raportarea la alte instrumente teoretice, de exemplu cele folosite de alţi cercetători, pe care le putem identifica în studiile publicate, sau în bazele de date publicate de alţi cercetători. Deseori, aceste baze de date sunt însoţite de un manual (*codebook*) ce explică codurile folosite, indicând modul în care a fost formulată fiecare întrebare ce a condus la culegerea acelei informaţii empirice, dar și modul în care fiecare valoare a variabilei se regăsește în baza de date, cum ar fi codurile numerice sau etichetele acestor valori.

O altă formă a validităţii este cea empirică şi se referă la capacitatea instrumentului folosit de a măsura efectul empiric pe care dorim să îl măsurăm. O putem verifica prin calcularea unor coeficienţi de corelaţie între două instrumente diferite, de exemplu, calculate pe eşantioane diferite de indivizi, sau compararea unui instrument folosit pentru a măsura o caracteristica a unor indivizi dintr-un eşantion

cu criterii de la nivelul populaţiei, numite criterii externe (Frankfort-Nachmias, Nachmias, și DeWaard 2015, 132). Validitatea de construct este o a treia faţetă a validităţii şi se referă la calitatea operaţionalizării conceptului nostru. O putem evalua identificând capacitatea instrumentului nostru de a fi în concordanţă cu utilizarea unor concepte similare în cadrul teoretic relevant pentru acea problemă studiată.

Fidelitatea este o altă trăsătură importantă a măsurătorii pe care o facem folosind instrumente de cercetare. Ne dorim ca, odată ce am operaţionalizat instrumentele noastre şi am cules datele folosind aceste instrumente, acestea să poată fi replicate, fie de noi, fie de alţi cercetători. În plus, uneori validitatea instrumentelor este greu de stabilit, de aceea, spun Frakfort-Nachmias et al. (2015, 135) ne putem fundamenta analizele şi instrumentele folosit prin evaluarea fidelităţii acestora. În acest fel putem să obţinem recunoaşterea validităţii concluziilor şi a demersului analitic. Capacitatea instrumentelor folosite în culegerea informaţiilor de a produce în mod constant rezultate identice pentru situaţii / măsurători identice, poartă numele de fidelitate a instrumentului de cercetare. Fidelitatea reprezintă, deci, capacitatea instrumentului de a produce cât mai puţină eroare de măsurare. Aceasta poate fi testată în mai multe feluri. De exemplu, putem aplica aceeaşi întrebare la momente de timp diferite, prin tehnica testare-retestare (Frankfort-Nachmias, Nachmias, și DeWaard 2015, 136). Deşi memoria respondenţilor poate fi foarte bună, deci poate să intervină în măsurătoarea noastră o eroare de respondent în care ne sunt furnizate aceleaşi răspunsuri false la momente de timp diferite, dacă răspunsul la întrebare este similar probabil instrumentul nostru este fidel. Din motive care ţin de eroarea cauzată de respondent, nu putem asigura o fidelitate perfectă prin această tehnică.

Testarea fidelităţii instrumentelor se poate realiza şi prin repetarea întrebărilor sub forme uşor diferite, în acelaşi chestionar la momente diferite. În acest fel, putem să comparăm răspunsurile subiecţilor la cele două întrebări pentru a estima gradul de fidelitate a instrumentului. De asemenea, mai putem sparge variabilele complexe în mai multe variabile simple fiecare măsurând segmente diferite ale aceluiaşi concept, astfel încât să reducem eroarea de măsurare determinată de variabilă. Nu în ultimul rând am putea utiliza operatori diferiţi, însă această strategie este recomandabilă doar în etapa de pretestare sau atunci când vrem să comparăm instrumentele pe subiecţi

diferiţi şi vrem să controlăm pentru eroarea cauzată de operator în măsurarea unei variabile.

Caracteristicile intrinseci ale variabilelor, pe care dorim să le obţinem prin operaţionalizare (a se revedea capitolul 1) sunt măsurarea şi cuantificarea. Ne amintim că variabilele pot fi cuantificate discret sau continuu. Diferenţele sunt date de modul în care categoriile de valori pe care le poate lua un referent empiric pe variabila analizată permit sau nu valori intermediare. Este important să diferenţiem între variabilele calitative sau categoriale (numite și categorice de unii autori români) şi variabilele cantitative sau numerice, deoarece tipurile de analize în care putem folosi aceste variabile depind de caracteristica variabilelor. Putem privi aceste două tipuri de variabile şi din perspectiva unor niveluri de diferenţiere în interiorul fiecărui tip, în funcţie de caracteristicile generale ale categoriilor determinate de variabilă, diferenţe pe care le numim **niveluri de măsurare**.

Aşadar, putem constata că unele variabile calitative produc categorii distincte de valori (referenţi empirici), fără ca între aceste valori / categorii să existe o ordine. Aceste variabile se rezumă la a nominaliza fiecare categorie în parte, de aceea le numim variabile de **nivel de măsurare nominal**, sau pur şi simplu **variabile nominale**. Aceste variabile sunt caracterizate prin categorii sau clase mutual exclusive (fiecare clasă conţine caracteristici unice, care o diferenţiază întru totul de celelalte clase). Chava Frankfort-Nachmias şi colaboratorii ei (2015) recomandă o condiţie suplimentară pentru acest nivel de măsurare, respectiv exhaustivitatea claselor (nici o caracteristică nu rămâne fără clasa sa specifică). Epuizarea referenţilor empirici prin clasele furnizate de variabila nominală este o condiţie *sine qua non* pentru asigurarea validităţii variabilei. Cu toate acestea, considerăm că în cazul variabilelor nominale nu rareori vom fi în situaţia în care nu vom putea include exhaustiv clasele, mai ales atunci când operaţionalizarea variabilei ne conduce la un număr exagerat de clase ci vom apela la clase umbrelă, cum este de exemplu categoria „Alţii" sau „Restul". Această categorie umbrelă ne ajută să epuizăm, teoretic, clasele. De exemplu, atunci când colectăm informaţii despre ocupaţia unei persoane, nu ne vom permite, din motive de cost şi timp, să integrăm în chestionar toate ocupaţiile posibile, astfel încât să asigurăm epuizarea tuturor posibililor respondenţi în aceste clase. De cele mai multe ori vom folosi o variantă extrem de redusă a listei posibilelor ocupaţii ale

respondenţilor şi, în plus, vom permite codificarea unei categorii umbrelă „Altă ocupaţie, precizaţi care”. Precizăm că orice codificări numerice am da acestor valori, ele nu au nici un rol matematic, nici nu pot fi folosite pentru a face operaţii matematice logice. De exemplu, variabila gen produce categoriile bărbat, respectiv femeie, între aceste categorii neexistând nici o relaţie de ordonare, iar orice cod numeric am da acestor categorii este folosit doar pentru a facilita integrarea informaţiei într-o bază de date. Prin urmare, aceste coduri numerice (de exemplu 1 pentru bărbat, 2 pentru femeie, dar la fel de bine putem codifica 0, respectiv 1, sau 35 respectiv 40) nu au valoarea cantitativă; ele nu pot fi folosite pentru a spune că valoarea „femeie” este de două ori valoarea „bărbat”, nici nu putem calcula indicatori cum ar fi media. Mai mult, putem permuta categoriile acestor variabile fără să schimbăm sensul variabilei sau rezultatul măsurătorii produse. Un caz nu poate lua două valori pe aceeaşi variabilă, în acelaşi timp. Alte exemple de variabile măsurate pe nivelul nominal pot fi votul pentru preşedinte (votezi cu candidatul A, sau candidatul B sau candidatul C, nu poţi vota cu candidatul ½A sau cu candidatul A,5); votul pentru partide (votezi cu partidul F, sau partidul H, prin urmare pe această variabilă un respondent nu poate lua valoarea H,7 sau 0,F); numele localităţii de rezidenţă (de exemplu, poate lua valori precum Aiud, Cluj-Napoca, Bucureşti, etc.); etnie; religie; surse de informare; zona de rezidență (de exemplu urban, respectiv rural), tipul de sistem electoral, ocupaţia etc.

Un al doilea tip de variabilă calitativă este reprezentat de variabilele care produc categorii mutual exclusive, dar care, în plus, indică existenţa intrinsecă şi acceptată a unei ordini sau ierarhii între aceste categorii. Aceste variabile se numesc variabile de **nivel de măsurare ordinal**, sau, simplu, **variabile ordinale**. Relaţiile între categoriile determinate de acest tip de variabile respectă câteva proprietăţi ale relaţiilor de echivalenţă: ireflexivitate (un caz A nu poate fi diferit de el însuşi), asimetria (dacă A este mai mare decât B, atunci B nu poate fi mai mare decât A) şi tranzitivitatea (dacă A este mai mic decât B, iar B este mai mic decât C, atunci A nu poate fi mai mare decât C) (Frankfort-Nachmias, Nachmias, şi DeWaard 2015, 126). Nu putem permuta aceste categorii, ordinea lor fiind predefinită şi, în general, unanim acceptată. Deoarece aceste variabile au mai multe caracteristici pentru categoriile pe care le definesc, deci sunt mai uşor cuantificabile, putem afirma că aceste variabile sunt superioare variabilelor nominale. De exemplu, educaţia măsurată în categorii:

primară, gimnazială, liceală, universitară. Şi pentru valorile acestor variabile putem folosi coduri numerice în vederea integrării lor într-o bază de date, însă, la fel ca în cazul variabilelor nominale, aceste coduri nu au nici un sens matematic. În exemplul de mai sus, dacă codificăm cu valori de la 1 la 4 categoriile educaţionale, nu înseamnă că aceste coduri produc sens matematic, altul decât ordinea. La fel de bine am putea codifica această variabilă cu valori precum 10, 20, 30, 40 sau 11, 12, 13, 14, oricare dintre ele indicând succesiunea categoriilor de educaţie formală. În plus, cu aceste valori nu putem face nici o operaţie matematică, deci nu am putea spune că nivelul universitar, care a primit codul 4, este de două ori mai mare decât nivelul gimnazial, care a primit codul 2. Deoarece nu cunoaştem mărimea diferenţelor dintre categorii, nu putem că diferenţa dintre codul 1 (nivel primar) şi 2 (nivel gimnazial) este egală cu diferenţa dintre codul 3 (nivelul liceal) şi codul 4 (nivelul universitar). Nici nu putem spune că un respondent care a absolvit doar ciclul de învăţământ liceal şi în variabila noastră a primit codul 3, este cu 1 punct sau cu 50% mai educat decât respondentul care a absolvit doar învăţământul gimnazial şi care este codificat cu cifra 2. Aceste coduri numerice pot fi privite, mai degrabă, ca fiind ranguri şi, prin urmare, putem discuta de diferenţe între ranguri, nu de diferenţe între valori. Alte exemple de variabile ordinale pot fi variabilele de atitudine sau de opinie, cum ar fi nivelul de încredere în guvern (măsurat prin valori precum deloc, puţină, aşa şi aşa, multă, foarte multă), nivelul de satisfacţie cu calitatea vieţii (deloc, puţin, nici puţin nici mult, mult, cu totul), venitul în categorii (mic, mediu, mare), vârsta în categorii (copil, tânăr, adult, bătrân), tip de localitate (comună, oraş mic, oraş mediu, municipiu, capitală).

Un prim tip de variabile cantitative, sau continue, este acela al variabilelor care pe lângă caracteristicile variabilelor ordinale (categorii mutual exclusive şi o ordine sau ierarhie între aceste categorii) mai sunt caracterizate de existenta unei valori zero acceptată, dar şi de existenţa unor diferenţe (intervale) clare şi măsurabile între categorii. Acestea se numesc variabile de **nivel de măsurare de interval**, sau **variabile de interval**. De exemplu, temperatura în grade Celsius, Fahrenheit, sau Réaumur, este o variabilă de interval. Valorile folosite pentru a codifica şi măsura acest tip de variabile au un sens matematic însă pot fi folosite doar pentru operaţii de scădere şi adunare. Operaţiile prin care calculăm rapoarte nu produc sens matematic. De exemplu, nu putem spune că este de zece ori mai cald la temperatura de 10 grade

Celsius (echivalentul a 50 de grade Fahrenheit, sau 8 grade Réaumur) decât la temperatura de 0 grade Celsius (32 grade Fahrenheit, sau 0 grade Réaumur).

Un al doilea tip de variabilă cantitativă este cel al variabilelor aflate la **nivel de măsurare de rapoarte**, numite simplu **variabile de rapoarte**. Aceste variabile au toate caracteristicile variabilelor de interval, dar în plus au un zero absolut. Codurile numerice alocate valorilor pe care respondenţii le pot lua pe aceste variabile au un sens matematic, deci pot fi folosite pentru operaţii matematice. Temperatura măsurată în grade Kelvin, care oferă un nivel zero absolut, este un exemplu de variabilă de rapoarte. Alte variabile care sunt măsurate pe nivelul de rapoarte sunt: pulsul, tensiunea arterială, distanţa între localităţi, numărul de votanţi, numărul de studenţi dintr-o clasă, număr de ani petrecuţi în sistemul de educaţie formală etc.

Putem spune că între nivelurile de măsurare există o relaţie ierarhică, iar ceea ce e valabil pentru un nivel inferior este valabil şi pentru un nivel superior. În forma lor brută, netransformată, de exemplu, în variabilă dihotomică, valorile numerice ale variabilelor nominale pot fi folosite doar pentru analize de frecvenţe, fără a putea calcula mediane, medii sau alţi indicatori ce necesită operaţii precum scăderea sau rapoartele. Variabilele ordinale sunt superioare pentru că valorile lor numerice pot fi folosite pentru calcularea unor frecvenţe şi a unor indicatori precum mediana, dar nu pentru calcularea mediilor în operaţii precum scăderea sau fracţiile. Variabilele de interval sunt superioare celor nominale şi ordinale pentru că valorile lor numerice pot fi folosite pentru calcularea frecvenţelor, a unor indicatori precum mediana, media sau abaterea standard, pentru operaţii matematice precum adunarea şi scăderea, dar nu pentru operaţii matematice de rapoarte. Variabilele de rapoarte pot fi folosite pentru orice tip de operaţii matematice.

Variabilele măsurate pe un nivel de măsurare superior (cantitativ) pot fi transformate într-o variabilă măsurată pe un nivel inferior (calitativ), însă odată operaţionalizată o variabilă pe un nivel de măsurare inferior, şi odată culese datele empirice la acel nivel, este extrem de greu, dacă nu imposibil, să le mai transformăm ulterior în variabile superioare. De exemplu, venitul unei persoane, măsurat în suma de bani obţinuţi luna trecută, este o variabilă de rapoarte şi poate fi transformată, prin operaţiune de recodificare, într-o variabilă ordinală (de exemplu în venit mic, mediu,

mare), sau în variabilă nominală (venit suficient sau venit insuficient pentru traiul zilnic). Însă odată colectate informaţiile empirice pentru această variabilă la nivel nominal, care limitează foarte mult diferenţele fine dintre categorii, nu mai putem să o transformăm într-o variabilă superioară (cantitativă). Am putea oferi însă un exemplu de aproximare a unei variabile cantitative pornind de la informaţii empirice culese pe o variabilă calitativă: educaţia măsurată la nivel ordinal, poate fi transformată în educaţie la nivel de măsurare de rapoarte, recodificând-o în număr de ani petrecuţi în şcoală. Pentru a face această transformare putem estima durata medie a fiecărui ciclu de învăţământ formal, folosind suma acestor ani ca valoare pentru fiecare individ care a absolvit acel nivel specific de educaţie. Într-un eşantion este puţin probabil să avem un număr apreciabil de persoane repetente care au petrecut mai mulţi ani într-un ciclu de învăţământ, la fel cum e puţin probabil să avem un număr mare de persoane care au studiat în sistem intensiv (doi ani într-unul), pentru a avea distorsiuni în măsurarea acestei variabile.

Variabilele dihotomice sunt un tip specific de variabile, care formal sunt variabile pe nivel de măsurare nominal, fiind variabile categorice. Ele poartă şi denumirea de variabile *dummy* sau variabile binare. Sunt formate din două categorii de tipul da / nu, atributiv / non-atributiv, A / non-A, prezenţa / absenţa caracteristicii, 1 / 0. Din acest motiv, aceste variabile pot fi obţinute prin transformarea oricărui tip de variabilă, calitativă sau cantitativă. Astfel, pornind de la o variabilă ce are patru categorii (de exemplu etnia), putem produce trei (4 minus 1) variabile dihotomice, a patra fiind dedusă din celelalte trei.

3.1.2. Analiza practică – ce vom testa în exemplele din carte

Aplicaţiile practice exemplificate în capitolele următoare îşi propun să testeze empiric diferiţi determinanţi ai încrederii oamenilor în scena politică, precum şi impactul locului de muncă asupra nivelului de încredere. În vederea formulării ipotezelor de cercetare, am stabilit o serie de obiective fundamentate teoretic ca urmare a realizării analizei literaturii de specialitate. Analiza literaturii de specialitate

evidenţiază faptul că încrederea în instituţiile guvernamentale şi politice, precum şi satisfacţia cetăţenilor cu privire la o serie de indicatori (de exemplu, politici, economici, democratici, stil de viaţă, fericire etc.) au un rol important în funcţionarea democraţiei (Armingeon și Ceka 2014). Mai mult, nivelul de încredere şi de satisfacţie a cetăţenilor în scena politică reflectă nu doar buna funcţionarea a democraţiei, dar şi efectul asupra comportamentului politic al cetăţenilor (E. U. Weber 2016). Cercetătorii adoptă de obicei una dintre cele trei perspective în ceea ce priveşte relaţia cauzală între satisfacţie şi încredere: (1) satisfacţia duce la încredere, (2) încrederea duce la satisfacţie, sau (3) ambele construcţii se consolidează reciproc (E. U. Weber 2016). În baza informaţiilor obţinute în urma realizării analizei literaturii de specialitate, ne propunem ca în capitolele următoare, folosind diverse modele statistice, (1) să observăm dacă nivelul de satisfacţie al cetăţenilor cu privire la o serie de factori influenţează nivelul acestora de încredere în scena politică; (2) să identificăm dacă prezenţa la vot a cetăţenilor este influenţată de încrederea cetăţenilor în scena politică şi (3) să observăm dacă locul de muncă influenţează nivelul de satisfacţie a cetăţenilor cu propria viaţă şi cu situaţia economică.

Astfel, plecând de la obiectivele de cercetare menţionate mai sus am formulat o serie de ipoteze pe care le vom testa în capitolele următoare prin aplicarea modelul statistic potrivit fiecăreia dintre ele. Am formulat trei seturi de ipoteze care îşi propun să răspundă obiectivelor de cercetare formulate mai sus. Un prim set de ipoteze testează relaţia dintre nivelul de încredere şi satisfacţia cetăţenilor cu privire la propria situaţia economică, cu privire la viaţa personală, la situaţia democraţiei şi cu privire la gestionarea pandemiei de COVID-19 de către Guvern, dar şi nivelul de fericire al acestora. Altfel spus, vrem să investigăm dacă nivelul de încredere în politicieni şi partidele politice este influenţat de nivelul de satisfacţie al cetăţenilor şi de nivelul de fericire perceput de aceştia. În acest context am formulat următoarele ipoteze:

I1a. Indivizii cu un nivel ridicat de fericire au mai multă încredere în politicieni.

I1c. Un nivel de satisfacţie ridicat cu propria situaţie economică creşte nivelul de încredere în politicieni.

I1d. Un nivel de satisfacţie ridicat cu propria viaţă creşte nivelul de încredere în politicieni.

I1e. Un nivel de satisfacţie ridicat cu situaţia democratică creşte nivelul de încredere în politicieni.

I1f. Un nivel de satisfacţie ridicat cu modul în care guvernul a gestionat pandemia de COVID-19 creşte nivelul de încredere în politicieni.

Similar ipotezelor care testează impactul nivelului de satisfacţie a cetăţenilor asupra încrederii acestora în politicieni, am formulat o serie de ipoteze care testează impactul nivelului de satisfacţie asupra încrederii în partidele politice în general. Mai mult, am considerat că pe lângă nivelul de satisfacţie, încrederea în partidele politice este influenţată şi de ataşamentul faţă de un partid politic (Koleva și Rip 2009). Astfel, ipotezele formulate sunt următoarele:

I2b. Cu cât nivelul de fericire al indivizilor este mai mare, cu atât încrederea acestora în partidele politice creşte.

I2c. Cu cât interesul faţă de politică este mai mare cu atât creşte încrederea în partidele politice.

I2d. Cu cât satisfacţia faţă de propria situaţia economică este mai mare cu atât nivelul de încredere în partidele politice creşte.

Impactul nivelului de încredere al cetăţenilor a fost testat şi în relaţia cu participarea la vot. În baza literaturii de specialitate, considerăm că participarea la vot este influenţată pozitiv de nivelul de încredere al cetăţenilor în parlament (Grönlund și Setälä 2007).

I3a. Probabilitatea ca un individ să fi votat la ultimele alegeri este mai mare dacă acesta are un nivel ridicat de încredere în parlament.

De asemenea, conform literaturii de specialitate, ne propunem să testăm dacă participarea la vot este influenţată pe lângă nivelul de încredere şi de nivelul de educaţie al cetăţenilor. Este subliniat în studiile consacrate care analizează comportamentul alegătorilor faptul că un nivel de educaţie ridicat este asociat cu o rată de participare la vot crescută. Această relaţie se datorează faptului că un nivel de educaţie ridicat presupune şi dobândirea unor abilităţi civice şi cunoştinţe politice

care dezvoltă simțul de implicare politică al cetățenilor și implicit participarea la vot (Lazarsfeld, Berelson, și Gaudet 2021).

I3b. Probabilitatea ca un individ să fi votat la ultimele alegeri este mai mare dacă acesta are un nivel de educație ridicat.

Ultimul set de ipoteze își propune să testeze impactul nivelului de satisfacție cu propria viață și cu situația economică asupra tipului de loc de muncă al indivizilor. Astfel, formulăm următoarele ipoteze:

I4a. Antreprenorii au un nivel de satisfacție cu propria viață ridicat.

I4b. Indivizii care lucrează pentru mediul privat au un nivel ridicat de satisfacție privind situația economică.

Ipotezele formulate mai sus vor fi testate pe un eșantion de 6,000 de respondenți cu vârsta peste 18 ani care au mediul de rezidență în țările din Estul Europei. Datele utilizate în analizele noastre au fost colectate de European Social Survey (ESS) în anul 2020. În vederea testării ipotezelor formulate vom aplica o serie de analize bazate în principal pe metoda regresiei.

3.1.3. Citirea (Importarea) setului de date în RStudio

Prima etapă a procesului de analiză statistică a datelor constă în importarea (citirea) setului de date în programul pe care vrem să-l utilizăm. Pentru a descărca setul de date utilizat în următoarele capitole accesăm pagina oficială a ESS (https://ess-search.nsd.no/en/study/172ac431-2a06-41df-9dab-c1fd8f3877e7). Setul de date poate fi descărcat cu extensii diferite în funcție de programul statistic folosit (.sav sau .zsav pentru SPSS, .dta pentru Stata și RStudio, .csv moștenit din programul Fortran, din anii 1970 și acum folosit pentru majoritatea programelor statistice). Noile versiuni ale unor programe precum Stata, R (RStudio) sau SPSS pot citi, direct sau prin intermediul

unor pachete gratuite, dezvoltate independent, toate aceste tipuri e date. Pentru realizarea aplicațiilor practice în Stata și RStudio descărcăm fie fișierul .dta fie .csv.[28]

Importarea (citirea) fișierelor de tip .csv[29] în programul RStudio se realizează prin introducerea în consola programului a următoarei funcții:

```
### Citirea fișierelor care au extensia .csv ###
read.csv ("ESS10.csv")
```

RStudio permite, de asemenea, importarea fișierelor care au extensia .dta. Pentru a importa fișiere cu extensia .dta este necesară instalarea pachetului **'haven'**[30] (Wickham, Miller, și Smith 2022) și introducerea următoarei funcții în consolă:

```
### Citirea fișierelor care au extensia .dta ###
library(haven)
read_dta ("ESS10.dta")
```

În ecosistemul de programare R, există mai multe pachete ce permit utilizatorului importarea și citirea bazelor de date în format .dta, în afara pachetului arhicunoscut **haven**, sau a pachetului dedicat importării datelor Stata **readstata13** (Garbuszus și Jeworutzki 2022).

De exemplu, cu ajutorul pachetului **DDIwR** (Dușa 2022a) care permite conversia bazei de date din și în programul statistic Stata prin transformarea automatizată a valorilor lipsă (în engleză *missing values*). Pachetul folosește standardul internațional Data Documentation Initiative (DDI) care atribuie variabilelor tipul de

[28] Pentru informații referitoarea la instalarea și deschiderea sesiunii de lucru în RStudio și Stata parcurgem pașii descriși de Adrian Dușa și colaboratorii (2015).

[29] Se recomandă importul fișierelor de tip .csv ca alternativă la a celor Excel.

[30] Un pachet este o colecție de funcții R, date și cod compilat într-un format bine definit. Pachetele sunt stocate în directorul numit bibliotecă (library). R vine cu un set standard de pachete. Cu ajutorul comenzii search(), putem găsi toată lista pachetelor disponibile care sunt instalate în sistemul nostru de operare R. Altele sunt disponibile pentru descărcare și instalare. Odată instalate, trebuie să le încărcăm în sesiune utilizând funcția library() pentru a le utiliza.

nivel de măsurare adecvat: "nominal", "ordinal", "de interval", "de rapoarte", "procente", și "altul". Astfel, în RStudio putem aplica următoare comandă:

```
### Citirea fișierelor cu ajutorul pachetului DDIwR și a funcției
convert() ###

library(DDIwR)
mydata_dta <- convert("ESS10.dta")
```

Seturile de date importate în RStudio pot fi stocate ca *obiect,* care poate fi salvat și accesat ulterior din Global Environment. Stocarea fișierelor în RStudio poate fi realizată prin următoarea linie de cod:

```
### Atribuirea unui nume setului de date importat (citit) ###
mydata_csv <- read.csv("ESS10.csv")
mydata_dta <- read_dta("ESS10.dta")
```

Pentru exemplificare am atribuit numele mydata urmat de extensia specifică fiecărui set de date. Când lucrăm pe propriul set de date nu este necesară importarea sau numirea tuturor extensiilor. De regulă, este recomandat să lucrăm cu extensia specifică programului statistic în care realizăm analiza. În această carte vom lucra cu extensia .dta.

În urma creării obiectului mydata și rulării liniei de cod respective prin apăsarea comenzii Run sau a scurtăturii Ctrl + Enter, în chenarul Global Environment va apărea setul de date importat salvat cu noul nume (mydata). Se recomandă utilizarea unor nume simple, formate dintr-un singur cuvânt. Dacă numele este compus din două sau mai multe cuvinte, se recomandă ca acestea să fie despărțite de o bară jos (_).

Tabelul 3.1 ne oferă o privire de ansamblu exemplificând funcțiile și câteva pachete care pot fi utilizate pentru citirea/importarea seturilor de date în programul RStudio.

Tabel 3.1 Funcții pentru citirea și scrierea datelor

Pachet	Format fișier	Formulă citire
utilis (preinstalat)	.csv	read.csv(''nume_file.csv'')
utilis (preinstalat)	.rds	readRDS("name_file.rds")
haven (trebuie instalat)	.dta (citește și .sav)	read_dta("name_file.dta")
readxl (trebuie instalat	.xlsx	read_excel("name_file.xlsx")

3.1.4. Familiarizarea cu variabilele din setul de date

Familiarizarea cu setul de date importat reprezintă a doua etapă a procesului de analiză statistică. În cadrul acestei etape vom selecta și transforma variabilele de interes pentru analiza pe care o efectuăm conform obiectivelor de cercetare formulate în subcapitolul 3.1.2.

În RStudio există mai multe comenzi care sunt deosebit de utile pentru etapa de familiarizare cu setul de date:

- **summary()** - furnizează statistici rezumative, cum ar fi medii, abaterea standard etc.
- **glimpse()** și **str()**

Utilizarea comenzilor menționate necesită instalarea și încărcarea pachetului dyplyr (Wickham et al. 2022)[31], folosind funcția library().

```
library(dyplyr)
summary(mydata_dta)
glimpse(mydata_dta)
str(mydata_dta)
```

[31] Pentru a vedea ce poate face un pachet și care sunt funcțiile de care dispune, scriem în consolă ?numepachet (de exemplu, ?haven) și click enter.

Din rezultatele rezumate, se pot vizualiza numele variabilei, eticheta variabilei, distribuția frecvenței, valoarea maximă și minimă etc. Pentru un set de date foarte mare, cum ar fi cel de la ESS, informațiile furnizate pot fi copleșitoare, având în vedere numărul mare de variabile. O soluție este să limităm rezultatele prin listarea numai a variabilelor pe care dorim să le explorăm și, ulterior, să creăm un nou set de date care să conțină doar variabilele dorite. Pentru a face acest lucru, primul pas este să afișăm numele variabilelor din setul de date cu comanda.[32]

```
names(mydata_dta)
```

Al doilea pas este să obținem informații despre variabilele de interes. Plecând de la ipoteza 2c formulată în subcapitolul 3.1.2, să presupunem că suntem interesați de variabila independentă *polintr* care are la bază întrebarea "How interested are you in politics?" („Cât de interesat sunteți de politică?").

```
summary(mydata_dta$polintr)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    2.00    3.00    2.78    3.00    4.00      42
```

Înainte de a trece la următoarea etapă, este recomandat să ne familiarizăm foarte bine cu setul de date și cu variabilele dependente și independente pentru a facilita procesul de transformare a variabilelor finale în funcție de scopul analizei.[33]

---

[32] Pentru o vizualizare mai clară a variabilelor existente în setul de date ESS, odată cu descărcarea setului de date descărcăm din același loc, de la secțiunea documente (https://ess-search.nsd.no/en/study/172ac431-2a06-41df-9dab-c1fd8f3877e7), fișierul ESS10 Appendix A7 Codebook ed. 1.0.

[33] Este recomandat ca înainte de a trece mai departe să identificăm în setul de date, în baza fișierului oferit de ESS cu descrierea variabilelor, toate variabilele care alcătuiesc ipotezele de cercetare pe care vom lucra și să aplicăm funcția summary().

3.1.5. Manipularea variabilelor din setul de date

Familiarizarea cu variabilele de interes din setul de date permite trecerea la a treia etapă, și anume manipularea datelor. Manipularea datelor presupune transformarea variabilelor de interes într-un mod organizat și adaptat. În această secțiune vom exemplifica ce presupune manipularea datelor utilizând mai multe tipuri de comenzi, plecând de la variabilele din setul nostru de date.[34]

Accesarea rapidă a distribuției de frecvențe a răspunsurilor și a categoriilor de răspuns ale unei variabile este extrem de utilă atunci când utilizăm seturi mari de date. În RStudio, utilizarea comenzilor **count(name_file$nume_variabile)** și **list(nume_file$nume_variabila)** permite accesarea distribuției de frecvențe a răspunsurilor și a categoriilor de răspuns. În baza ipotezei 2c formulată în subcapitolul 3.1.2. suntem interesați să aflăm categoriile de răspuns și distribuția acestora pentru variabila independentă *polintr*:

```
### Afișarea distribuției de frecvențe pentru fiecare categorie ###

count(mydata_dta, polintr)

## # A tibble: 5 × 2
##   polintr                           n
##   <dbl+lbl>                     <int>
## 1     1 [Very interested]        1319
## 2     2 [Quite interested]       5469
## 3     3 [Hardly interested]      7093
## 4     4 [Not at all interested]  4137
## 5 NA(b) [Don't know]               42
```

[34] Manipularea datelor se realizează prin aplicarea mai multor comenzi din RStudio. Accesarea acestor comenzi se realizează prin instalarea și încărcarea pachetului **dplyr** (Wickham et al. 2022).

```
### Afișarea categoriilor de răspuns ale unei variabile ###

list(mydata_dta$polintr)

1 [Very interested]
2 [Quite interested]
3 [Hardly interested]
4 [Not at all interested]
5 NA(b) [Don't know]
```

Alternativ funcției count() putem obține distribuția de frecvențe a unei variabile prin utilizarea pachetului **DDIwR** (Dușa 2022a). Acest pachet include o funcție dedicată din pachetul **declared** (Dușa 2022b) pentru obținerea și afișarea distribuției de frecvențe. Astfel, putem aplica:

```
### Distribuția de frecvențe pentru variabila polintr utilizând funcția
w_table() ###
w_table(mydata_dta$polintr)

##                          fre    rel   per   vld   cpd
##                          ------------------------------
## Very interested         1319  0.073   7.3   7.3   7.3
## Quite interested        5469  0.303  30.3  30.4  37.7
## Hardly interested       7093  0.393  39.3  39.4  77.0
## Not at all interested   4137  0.229  22.9  23.0 100.0
##                          ------
##           No answer       11  0.001   0.1
##           Don't know      23  0.001   0.1
##              Refusal       8  0.000   0.0
##                          ------------------------------

##                           18060  1.000 100.0
```

Pentru a vizualiza etichetele valorilor unei variabile putem folosi comanda labels():

```
### Vizualizarea etichetelor unei variabile prin utilizarea funcției
labels() ###
labels(mydata_dta$polintr)
Very interested      Quite interested     Hardly interested Not at all
interested
1                     2                     3                     4
                Refusal            Don't know             No answer
                  -91                   -92                   -93
```

Seturile mari de date nu sunt doar dificil de accesat, dar conțin și o serie de informații care nu sunt de interes în cazul unei analize particulare. Astfel, este necesar să curățăm setul de date și să selectăm doar acele informații care sunt necesare analizei pe care o realizăm. Conform precizărilor din subcapitolul 3.1.2. suntem interesați să testăm ipotezele de cercetare doar pe respondenții din statele din Estul Europei. Astfel, din setul de date ESS dorim să rămânem cu un set de date care să conțină doar statele de interes.[35] În RStudio, crearea unui nou set de date dintr-un set de date existent se realizează prin comanda **filter():**

```
sample_state_estice <- filter(mydata_dta, cntry %in% c("BG", "CZ","HR",
"HU","SI", "SK"))
```

De asemenea, un set mare de date conține și o gamă variată de categorii de variabile care nu sunt de interes pentru analiza pe care o efectuăm. Astfel, pentru a utiliza eficient setul de date, putem crea un set de date secundar care să conțină doar variabilele de interes. Să presupunem că din setul de date sample_state_estice, vrem să păstrăm doar variabilele de interes: încredere în partide politice, în politicieni, participarea la vot la ultimele alegeri, orientare ideologică, satisfacție cu viața, economia, guvernul, educația și sănătate, nivelul de fericire, etc. și variabile socio -

[35] Statele incluse în analizele statistice din această carte sunt Bulgaria, Cehia, Croația, Ungaria, Slovacia și Slovenia.

demografice precum gen, vârstă și educație. În RStudio vom folosi comanda **select()** pentru a selecta într-un nou set de date doar variabilele de interes:

```
sample_scurt <- subset(sample_state_estice, select = c(idno, cntry,
polintr, vote, clsprty, trstsci, trstlgl,trstprt, trstplt, trstprl,
lrscale, stflife, stfeco, stfgov, stfedu, stfhlth, stfdem, happy, health,
eisced, mbtru, stfmjob, c19whome, gvhanc19, respc19, agea, gndr))
```

În general, variabilele necesită o serie de transformări astfel încât să corespundă tipului de analiză statistică utilizată. Un pas extrem de important care trebuie realizat înainte de transformarea variabilelor este curățarea setului de date de valori nule (de exemplu, respondenții care nu au vrut să răspundă, nu au știut să răspundă sau pentru care operatorul nu a consemnat nici un răspuns). Putem fie să renunțăm pur și simplu la valorile lipsă, dar dacă sunt prea multe aceasta poate duce la scăderea mărimii eșantionului (bazei de date) și la efecte asupra rezultatelor analizei. De aceea, deseori este recomandabil să folosim tehnici specifice imputare, prin care înlocuim valorile lipsă cu o altă valoare, estimată pe baza distribuției datelor valide, de exemplu, valoarea medie a întregului eșantion. Să luăm ca exemplu variabila încrederea în partidele politice *(trstprt).*

Să presupunem că *trstprt* trebuie să fie transformată dintr-o variabilă categorică ordonată de 11 nivele, într-o variabilă categorică ordonată pe trei nivele după cum urmează: de la 0 la 3 neîncredere totală, de la 4 la 6 încredere medie, iar de la 7 la 10 încredere totală. Odată ce am transformat toate variabilele de interes, vom crea un nou set de date cu toate variabilele transformate. Pentru exemplificare vom transforma o singură variabilă. Primul pas în transformarea variabilei *trstprt* este introducerea valorii medii în locul valorilor nule. În RStudio vom aplica următoarea linie de cod:

```
### trstprt ###
#mean trstprt###
summarise(filter(sample_scurt, trstprt<=10), avginc = mean(trstprt,
na.rm=TRUE)) ## 3.05
```

```
### înlocuim NA din trstprt cu valoarea medie 3 ###
sample_scurt %>%
replace_na(list(trstprt = 3))
```

Al doilea pas este transformarea propriu – zisă a scalei variabilei *trstprt* conform precizărilor de mai sus. În RStudio, transformarea variabilelor se face folosind comanda **mutate()**, iar pentru transformarea variabilei *trstprt* vom folosi următoarea linie de cod:[36]

```
trstprt_recodare <- select(sample_scurt, idno, cntry, trstprt) %>%
  mutate(trstprt = case_when(trstprt == 0 ~ 1,
trstprt == 1 ~ 1,
trstprt == 2 ~ 1,
trstprt == 3 ~ 1,
trstprt == 4 ~ 2,
trstprt == 5 ~ 2,
trstprt == 6 ~ 2,
trstprt == 7 ~ 3,
trstprt == 8 ~ 3,
trstprt == 9 ~ 3,
trstprt == 10 ~ 3,))
```

Noul set de date pe care vom lucra conține toate variabilele transformate după modelul de mai sus.

```
dataset_exemplu <- select(joined_final, idno, cntry, vote, polintr,
clsprty, mbtru, trstprt, trstplt, trstprl, trstsci, trstlgl, lrscale,
stflife, stfeco, stfgov, stfedu, stfdem, stfhlth, stfmjob, happy, health,
eisced, c19whome, gvhanc19, respc19, agea, gndr)
```

[36] Vom reveni la această secțiune de cod, ori de câte ori vrem să transformăm categoriile unei variabile. Să presupunem că vrem ca în loc de 1, 2 și 3 să apără neîncredere, neutru și încredere: trstprt_recodare <- select(sample_scurt, idno, cntry, trstprt) %>% mutate(trstprt = case_when(trstprt == 0 ~ 1, trstprt == 1 ~ 1, trstprt == 2 ~ 1, trstprt == 3 ~ 1, trstprt == 4 ~ 2, trstprt == 5 ~ 2, trstprt == 6 ~ 2, trstprt == 7 ~ 3, trstprt == 8 ~ 3, trstprt == 9 ~ 3, trstprt == 10 ~ 3,)) %>% mutate(trstprt = case_when(trstprt == 1 ~ 'neincredere', trstprt == 2 ~ 'neutru', trstprt == 3 ~ 'incredere',)).

În RStudio putem elimina răspunsurile nule (lipsă) din setul de date prin comanda **na.omit().**

```
dataset_exemplu %>%
  na.omit()
```

Finalizarea procesului de transformare a variabilelor și curățarea setului de date de valori nule permite demararea procesului de analiză a datelor și testarea ipotezelor formulate în subcapitolul 3.1.2. În continuare vom discuta despre analiză univariată, bivariată și multivariată a datelor.

3.2. Analiza univariată

În capitolele și secțiunile anterioare am discutat despre modul în care culegem datele, cum proiectăm cercetarea, ce standarde științifice și metodologice urmăm pentru a ne asigura în designul studiului nostru că vom putea produce concluzii valide din punct de vedere științific. Culegerea datelor și pregătirea lor în baza de date sunt etape fundamentale. Aceste etape consumă, în general, cele mai multe resurse într-un studiu științific. De calitatea datelor și a modului în care acestea au fost culese depinde calitatea analizei și a concluziilor. Despre metodele prin care analizăm datele vom discuta în secțiunile următoare.

Primul și cel mai simplu mod de a înțelege datele culese este de a analiza, pe rând, fiecare variabilă din punct de vedere a distribuției observațiilor. Această analiză este cunoscută sub numele de analiză univariată, deoarece nu vom analiza relațiile dintre variabile, ci doar modul în care sunt distribuite observațiile (valorile) în interiorul variabilelor. Acest tip de analiză este foarte util în procesul de curățare și pregătire a setului de date pentru analiza relațiilor dintre variabile. În multe rapoarte publice ale unor sondaje de opinie publică pe teme politice sunt prezentate cu

precădere informații de analiză univariată. În analiza univariată nu urmărim explicarea cauzelor sau a relațiilor dintre variabile, ci descrierea acestora și evidențierea tiparelor din aceste observații (de exemplu, distribuția subiecților din punct de vedere al genului, a notelor obținute la un curs, a venitului, a ocupației, a prezenței la vor, a atitudinilor cu privire la anumite fenomene etc.).

Indiferent de complexitatea modelului statistic ales, realizarea unei analize exploratorii a datelor este recomandată pentru procesul de înțelegere și validare a datelor din două motive. În primul rând, indicatorii analizei univariate avertizează cu privire la existența unor probleme ale setului de date. Astfel că, dacă există erori în modalitatea de codificare a datelor sau există date lipsă, indicatorii analizei univariate ajută la identificarea unor astfel de probleme. În al doilea rând, distribuția valorilor (cât de răspândite sunt datele sau unde tind să se grupeze în raport cu valoarea medie) pe variabile cheie ajută la identificarea relațiilor dintre variabile. Întrucât aceste tendințe pot fi greu de identificat când sunt utilizate statistici mai avansate, este recomandat să demarăm procesul de analiză al datelor cu analiza univariată a acestora.

Analiza univariată are ca scop descrierea și prezentarea informațiilor fiecărei variabile de interes. Astfel, statistica univariată este diferită de statistica inferențială care permite testarea și validarea unor ipoteze de cercetare. O analiză de tip univariat se poate realiza fie prin prezentarea distribuției de frecvențe sau procentuale a datelor, prin prezentarea numerică sau grafică a acestor informații; fie prin prezentarea unor indicatori capabili să ne ofere o imagine despre distribuia generală a valorilor fiecărei variabile. Distribuția de frecvențe a datelor poate fi reprezentată prin utilizarea tabelelor de distribuție a frecvenței, prin histograme, diagrame circulare și diagrame cu bare. Acești indicatori sunt numiți statistici descriptive. Ne vom folosi de tipologia oferită de Traian Rotariu (1999, 42) prin care diferențiem între indicatori de poziție (ai tendinței centrale) și indicatori ai dispersiei (de omogenitate, de variație). Principalii indicatori prin care putem măsura tendința centrală a datelor sunt media, mediana și valoarea modală, iar indicatorii specifici dispersiei datelor sunt amplitudinea, varianța și abaterea standard (Rotariu et al. 1999; Reisz 2017). Prin acești indicatori putem obține două tipuri de informații: care este individul tipic, cum putem reduce la

un singur număr distribuția generală a valorilor, respectiv cât de omogenă sau eterogenă este populația măsurată.

Acest capitol își propune să exemplifice utilizarea tabelelor de frecvențe simple, a diagramelor cu bare și a histogramelor pentru a sublinia modul în care sunt distribuite valorile unei variabile. În secțiunile ce urmează vom explica câteva tipuri de indicatori pe care îi calculăm prin analiza univariată și vom ilustra cu exemple practice modalitatea în care putem să obținem aceste măsurători. Vom prezenta, de asemenea, modalitatea de calcul a indicatorilor tendinței centrale și a indicatorilor dispersiei datelor. Secțiunile următoare prezintă aplicat comenzile în RStudio prin care pot fi realizate analize univariate a datelor prin calcularea distribuțiilor de frecvențe, a tendinței centrale și a dispersiei datelor. În capitolele 4 și 5 vom exemplifica aceste analize prin programele Stata și Excel.

3.2.1. Distribuţia de frecvenţe a datelor

Distribuția de frecvențe este cea mai elementară informație pe care o putem obține despre modul în care variază răspunsurile (valorile) pe fiecare variabil în parte. Ea indică frecvența (numărul de apariții) cazurilor dintr-o categorie sau, cu alte cuvinte, observațiile pentru fiecare valoare în parte pe care o poate lua variabila analizată. Distribuțiile pot fi și de procente, și reprezintă proporția observațiilor pe care le identificăm pentru fiecare valoare a variabilei și pentru totalul valorilor. Nu în ultimul rând, distribuțiile pot fi valide (adică nu includ non-răspunsurile), sau cumulate (însumează frecvențele valorilor anterioare, de ex pentru variabile ordinale sau cantitative).

De exemplu, distribuția de frecvențe ne poate arăta câte persoane au mers la vot și câte nu, dintr-un număr total de persoane. Astfel, distribuția de frecvențe este un indicator al analizei univariate folosit pentru a evalua proprietățile distribuției scorurilor dintr-un set de date. Putem reprezenta distribuția de frecvențe fie numeric, prin tabele de frecvențe, fie grafic, prin diagrame cu bare (în engleză *bar charts*) sau histograme, în funcție de nivelul de măsurare a variabilei.

3.2.1.1. Tabele de frecvențe

Un tabel de frecvențe este un indicator al analizei univariate prin intermediul căruia sunt indicate numărul de apariții sau procentul de apariție a categoriilor variabilelor de interes (Stockemer 2019). De regulă, tabelul de frecvențe este utilizat pentru variabile categorice ordinale, variabilele continue (cantitative), având o distribuție variată cu foarte multe valori diferite pe care le pot lua unitățile de analiză studiate, nu se pretează la raportarea tabelelor de distribuție, deoarece au prea multe categorii (rânduri).

Să presupunem că vrem să aflăm distribuția de frecvențe pentru variabila *trstprt* (încrederea în partidele politice) din setul nostru de date. În RStudio vom folosi funcția **table()** pentru a construi tabelul de frecvențe:

```
#### Realizarea unei tabel de frecvențe ####

table(dataset_exemplu$trstprt)

##
##   1   2   3
## 879 443 112
```

Rezultatul obținut în RStudio poate fi transpus într-un tabel de frecvențe organizat și care poate fi editat în Microsoft Office.

Tabel 3.2 Distribuția frecvențelor absolute pentru variabila trstprt

Categoriile variabilei *trstprt*	Frecvența absolută
Neîncredere (1)	879
Încredere medie (2)	443
Încredere totală (3)	112

Odată obținut tabelul de frecvențe, este important să înțelegem cum poate fi acesta interpretat. Tehnic, observăm că tabelul este împărțit în 2 coloane și 4 rânduri. Prima coloană include numele categoriilor variabilei de interes[37], în timp ce coloana a doua indică frecvența fiecărei categorii exprimată în termeni absoluți. Din punct de vedere al informațiilor pe care tabelul le evidențiază observăm că 879 de respondenți fac parte din categoria respondenților care nu au încredere în partidele politice, în timp ce 443 au o încredere medie și doar 112 au declarat o încredere totală în partidele politice. Astfel, conform informațiilor oferite de tabelul de frecvențe putem aprecia că există un nivel de încredere scăzut în partidele politice.

Informațiile prezentate în Tabelul 3.2 pot fi transmise nu doar în termeni absoluți (*f*), ci și în termeni relativi (%). Frecvențele relative sunt preferate frecvențelor absolute, deoarece cele absolute au dezavantajul de a fi dificil de interpretat (cel puțin fără realizarea unor calcule matematice mentale). Să luăm ca exemplu categoria „neîncredere" unde 879 de respondenți au declarat că nu au încredere în partidele politice. În comparație cu numărul de răspunsuri pentru celelalte categorii, putem afirma că 879 este un răspuns semnificativ pentru categoria celor care nu a încredere. Cu toate acestea, în funcție de mărimea eșantionului la care ne raportăm, 879 poate fi un răspuns semnificativ sau nesemnificativ. Astfel, trebuie să ne punem întrebarea: 879 de răspunsuri din câte? Cu siguranță, această frecvență de 879 este interpretată diferit în cazul unei eșantion de 4.000 de respondenți față de un eșantion de 10.000 de respondenți. Astfel, pentru a aprecia dacă 879 de răspunsuri indică o cantitate semnificativă este necesar să ne raportăm la mărimea eșantionului utilizat și pe baza acestuia să calculăm frecvențele relative (%).[38] Cum putem afla aceste valori într-un mod rapid? Să nu uităm că RStudio poate fi folosit ca un calculator supra puternic care permite, pe lângă altele, adunarea frecvențelor din toate categoriile variabilei *trstprt*

[37] Numele categoriilor poate fi schimbat în momentul în care recodificăm variabila de interes. Modalitatea prin care putem face acest lucru este să redenumim numele categoriilor cu funcția **mutate()**.

[38] Frecvențele relative sunt calculate raportând frecvența absolută la totalul distribuției.

pentru a afla dimensiunea totală a eșantionului.[39] Comenzile RStudio pentru determinarea mărimii eșantionului setului de date sunt:

```
### Adunăm frecvențele fiecărei categorii și le salvăm într-un obiect nou
numit „esantion_marime”, ###
esantion_marime <- 879 + 443 + 112
###Deschidem valoarea obiectului "esantion_marime"###
esantion_marime

## [1] 1434
```

În baza identificării dimensiunii eșantionului putem aprecia faptul că 879 din 1.434 de respondenți au declarat un nivel de neîncredere în partidele politice. Următorul pas este să ne gândim dacă, în raport cu dimensiunea eșantionului, 879 este un răspuns semnificativ sau nu. Pentru a afla acest lucru trebui să calculăm frecvența relativă. Frecvența relativă se calculează prin împărțirea numărului de respondenți dintr-o categorie la dimensiunea totală a eșantionului (de exemplu, 879/1434). În aceste condiții, frecvențele relative sunt de obicei exprimate prin raportare la întreg sau la unitate. În RStudio, comenzile pentru determinarea frecvenței relative sunt următoarele:

```
###Calculăm procentul pentru categoria "neîncredere"###
(879/esantion_marime)*100

## [1] 61.29707

### Calculăm proporția pentru categoria "neîncredere"###
879/esantion_marime

## [1] 0.6129707
```

[39] O altă modalitate de a determina mărimea eșantionului este identificarea acestei informații din fereastra Global Environment din programul RStudio.

Observăm că aproximativ 61% (sau 0.612)[40] din răspunsuri corespund respondenților care nu au încredere în partidele politice. În acest caz, 61,29% (sau .612) înseamnă că mai mult de jumătate din toate răspunsurile sunt în categoria celor care nu au încredere. Importanța substanțială a frecvențelor relative depinde de numărul de categorii din variabilă. În cazul de față, avem trei categorii de răspunsuri. Dacă răspunsurile au fost distribuite aleatoriu pentru celelalte două categorii, ne așteptăm să avem aproximativ 20% în fiecare categorie. Știind asta, rezultatul de 61% sugerează că aceasta este o categorie de răspuns destul de populară.

RStudio permite, de asemenea, calcularea proporțiilor fiecărei categorii a unei variabile prin utilizarea funcției **prop.table()**. Pentru a aplica această funcție trebuie să salvăm rezultatele frecvențelor relative într-un obiect nou și apoi funcția **prop.table()** va folosi acel obiect pentru a calcula proporțiile. Reținem că trebuie să folosim extensia „.tbl" când denumim noul obiect. Acest lucru servește ca un memento pentru RStudio că acest obiect special este un tabel. Când executăm comenzi ca aceasta, ar trebui să observăm că noul obiect apare în fereastra Global Environment. Comenzile pentru calcularea frecvențelor relative în RStudio sunt următoarele:

```
### Salvăm tabelul de frecvențe într-un obiect nou numit "poorAid.tbl" ###
trstprt.tbl <- table(dataset_exemplu$trstprt)

### Creăm un tabel de proporții folosind conținutul tabelului de
frecvențe###
prop.table(trstprt.tbl)

##         1         2         3
## 0.61297071 0.30892608 0.07810321
```

[40] Atenție, de cele mai multe ori programele statistice au setate în mod automat standarde folosite în litratura anglo-saxonă, în care separatorul zecimal este punctul „ . ", și nu virgula „ , ". Putem modifica aceste setări din opțiunile programelor, astfel încât virgula să fie folosită ca separator zecimal, iar punctul separator al sutelor, dar, în acest caz trebuie să folosim și să raportăm în mod constant valorile în același mod. Dacă, însă, publicăm rezultatele în jurnale sau cărți în limba engleză, separatorul zecimal este punctul iar cel de grupare a sutelor este virgula.

O metodă mai rapidă prin care putem obține tabelul frecvențelor relative în RStudio este folosirea funcției **freq()**, furnizată de pachetul **descr** (Aquino et al. 2021). Acest pachet oferă mai multe instrumente pentru efectuarea analizei descriptive. Pentru obținerea tabelului de frecvențe relative în RStudio trebuie să aplicăm următoarea linie de cod:

```
###Încărcăm pachetul necesar obținerii frecvențelor relative###
library(descr)
###Furnizăm un tabel de frecvențe, dar nu un grafic###
freq(dataset_exemplu$trstprt, plot=F)
## dataset_exemplu$trstprt
##       Frequency Percent
## 1           879   61.30
## 2           443   30.89
## 3           112    7.81
## Total      1434  100.00
```

Observăm că prin această linie de cod am obținut toate informațiile furnizate în tabelele anterioare, plus o serie de informații suplimentare. De asemenea, tabelul generat prin funcția **freq()** transpune rezultatele într-o manieră mult mai organizată. Astfel, Tabelul 3.2 de mai sus poate fi completat prin adăugarea unei coloane cu frecvențele relative după cum urmează:

Tabel 3.3 Distribuția frecvențelor absolute și relative pentru variabila trstprt

Categoriile variabilei trstprt	Frecvența absolută (f)	Frecvența relativă (%)
Neîncredere (1)	879	61.30%
Încredere medie (2)	443	30.89%
Încredere totală (3)	112	7.81
Total	1434	100.00%

Utilitatea informațiilor obținute prin intermediul tabelor de frecvențe poate fi maximizată prin prezentarea unor grafice univariate (de exemplu, diagrame cu bare și histograme). Graficele univariate nu doar completează informațiile obținute prin tabelele de frecvențe, dar oferă și o vizualizare a informațiilor ceea ce facilitează interpretarea informațiilor. Vizualizarea grafică a datelor ajută la contextualizarea rezultatelor, oferind consumatorului cercetării o perspectivă suplimentară asupra constatărilor statistice. În baza frecvențelor prezentate mai sus pentru variabila *trstprt*, vom reprezenta informațiile sub forma unor diagrame cu bare, a unor histograme și a unor *boxplots*.

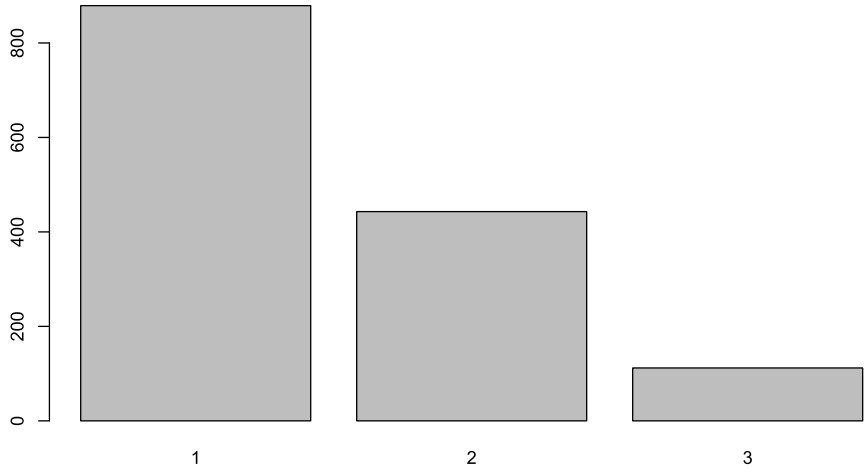
3.2.1.2. Diagrame cu bare

Diagramele cu bare (în engleză *bar charts*) sunt grafice simple care rezumă distribuția valorilor variabilelor categorice. Numele categoriilor pot fi identificate pe axa orizontală, chiar sub barele verticale; numerele de pe axa verticală indică numărul (sau procentele) din cazuri (indivizi); iar înălțimea fiecărei bare reprezintă numărul (sau procentul) de cazuri pentru fiecare categorie. Este important să precizăm că axa orizontală reprezintă diferențe categoriale, nu distanțe cantitative între categorii.

În RStudio, codul de mai jos este folosit pentru a genera diagrama cu bare pentru variabila *trstprt*, variabilă care măsoară nivelul de încredere în partidele politice. Important de reținut este faptul că funcția **barplot()** folosește tabelul de frecvențe salvat anterior cu denumirea de poorAid.tbl, și nu numele variabilei așa cum îl avem în setul de date. Astfel, diagramele cu bare sunt reprezentarea grafică a datelor brute dintr-un tabel de frecvențe.

```
###Plotăm distribuția de frecvențe pentru variabila trstprt###
barplot(trstprt.tbl)
```

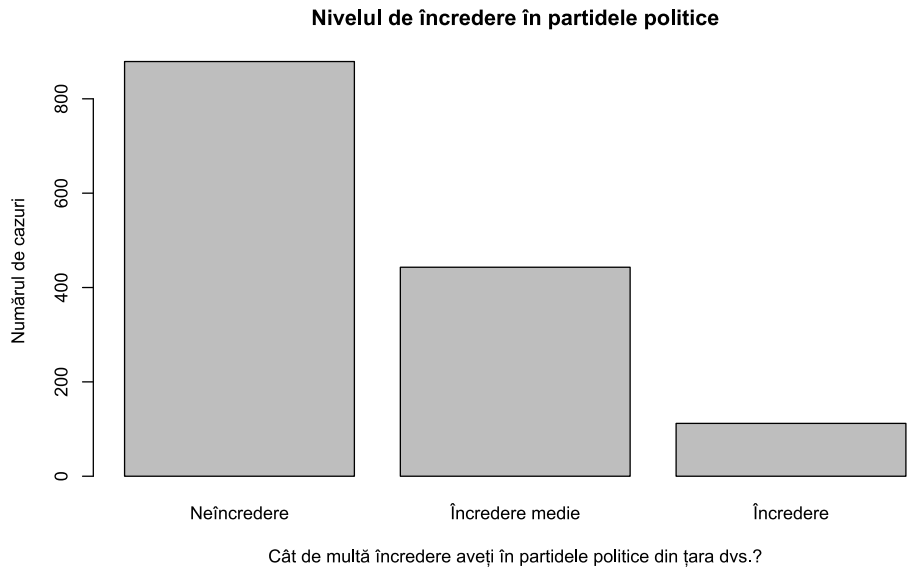
Figura 3.1 Distribuția de frecvențe pentru variabila trstprt (încrederea în partide)



Realizarea vizualizărilor grafice este un proces care consumă timp, întrucât informațiile transmise prin intermediul instrumentelor grafice trebuie prezentate într-o manieră ușor de înțeles pentru publicul larg. Astfel, plecând de la graficul din Figura 3.1 am realizat o serie de modificări prin intermediul cărora informațiile prezentate de diagrama cu bare pot fi cu ușurință înțelese de cei care citesc rapoartele cercetării noastre.

```
###Plotăm distribuția de frecvențe pentru variabila trstprt dar cu
etichetele modificate pentru claritate ###
barplot(trstprt.tbl,
names.arg=c("Neîncredere", "Încredere medie", "Încredere"),
xlab= "Cât de multă încredere aveți în partidele politice din țara dvs.?",
ylab= "Numărul de cazuri",
main= "Nivelul de încredere în partidele politice")
```

Figura 3.2 Distribuția de frecvențe pentru variabila trstprt (încrederea în partide) cu etichete

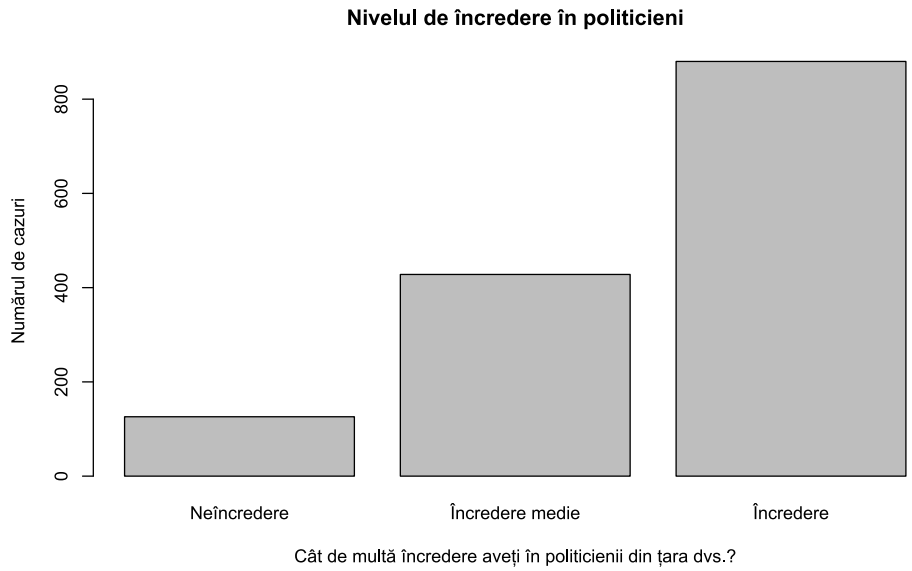


Să presupunem că vrem să comparăm nivelul de încredere al cetățenilor în partidele politice (*trstprt*), cu nivelul de încredere în politicieni (*trstplt*). Deoarece nu am salvat conținutul tabelului de frecvențe original pentru această variabilă într-un obiect nou, putem insera precizarea **table(dataset_exemplu$trstplt)** în comanda **barplot()**:

```
###Plotăm distribuția de frecvențe pentru variabila trstplt###

barplot(table(dataset_exemplu$trstplt),
names.arg=c("Neîncredere", "Încredere medie", "Încredere"),
xlab="Cât de multă încredere aveți în politicienii din țara dvs.?",
ylab="Numărul de cazuri",
main="Nivelul de încredere în politicieni")
```

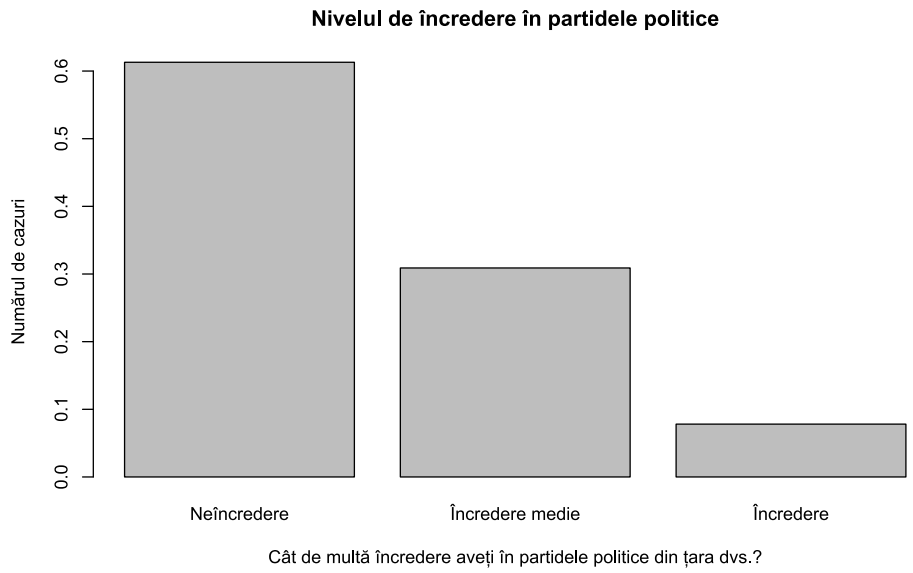
Figura 3.3 Distribuția de frecvențe absolută pentru variabila trstplt (încrederea în politicieni)



De asemenea, putem reprezenta grafic nu doar frecvențele absolute, ci și frecvențele relative. Astfel, pentru reprezentarea frecvențelor relative putem indica programului RStudio să folosească tabelul de proporții ca intrare și să schimbe în consecință eticheta axei y:

```
###Folosim "prop.table" ca input###
barplot(prop.table(trstprt.tbl),
names.arg=c("Neîncredere", "Încredere medie", "Încredere"),
xlab="Cât de multă încredere aveți în partidele politice din țara dvs.?",
ylab="Numărul de cazuri",
main="Nivelul de încredere în partidele politice")
```

Figura 3.4 Distribuția de frecvențe relativă pentru variabila trstprt (încrederea în partide)



Observăm că informațiile vizualizate prin intermediul diagramelor cu bare respectă informațiile furnizate de tabele de frecvențe realizate la subcapitolul 3.2.1.1. Important de menționat este că atunci când comparăm grafic nivelul de încredere în partidele politice (Figura 3.2 și 3.4) cu nivelul de încredere în politicieni (Figura 3.3) observăm că încrederea în politicieni este mai mare ca încrederea în partidele politice în termeni de frecvențe absolute.

3.2.1.3. Histograme

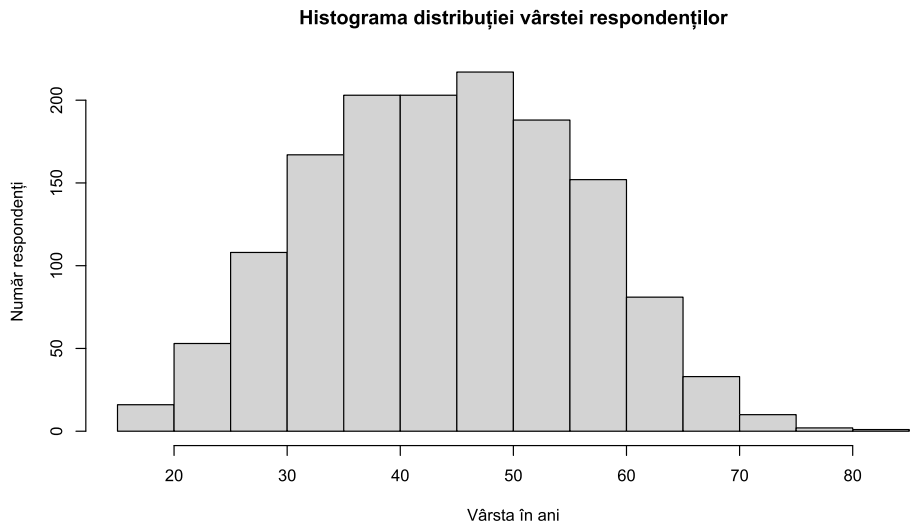
Histogramele sunt utilizate pentru a rezuma vizual distribuția de frecvențe a variabilelor cantitative (de interval și de rapoarte). Ele oferă o interpretare vizuală a datelor numerice, arătând numărul de puncte de date care se încadrează într-un interval de valori. Histograma este similară unei diagrame cu bare verticale. Cu toate acestea, o histogramă, spre deosebire de diagrama cu bare verticale, nu dispune de spații între bare, ci ea reflectă caracterul continuu al informațiilor furnizate de variabilele cantitative. Mediana, media și dispersia datelor pot fi, de asemenea,

ilustrate printr-o histogramă. În plus, histograma poate afișa valorile aflate exterior tiparului de distribuție (în engleză *outliers*).

Să presupunem că vrem să aflăm distribuția vârstei respondenților din eșantionul studiat și să o reprezentăm grafic. Astfel, în RStudio vom utiliza următoarea linie de cod:

```
###Histograma vârstelor###
hist(dataset_exemplu$agea,
xlab="Vârsta în ani",
ylab="Număr respondenți",
main="Histograma distribuției vârstei respondenților")
```

Figura 3.5 Histograma vârstelor



Intuitiv, observăm că cei mai mulți respondenți au o vârstă cuprinsă între 30 și 60 de ani, iar dintre aceștia cea mai mare distribuție se află la cei cu vârsta cuprinsă între 40 și 50 de ani.

Pentru a observa mai bine distribuția datelor putem folosi și o diagramă a densității datelor. Astfel, prin generarea unei linii care uniformizează denivelările

histogramei putem observa forma distribuției. Diagramele de densitate pot fi utilizate împreună cu histogramele sau independent acestora.

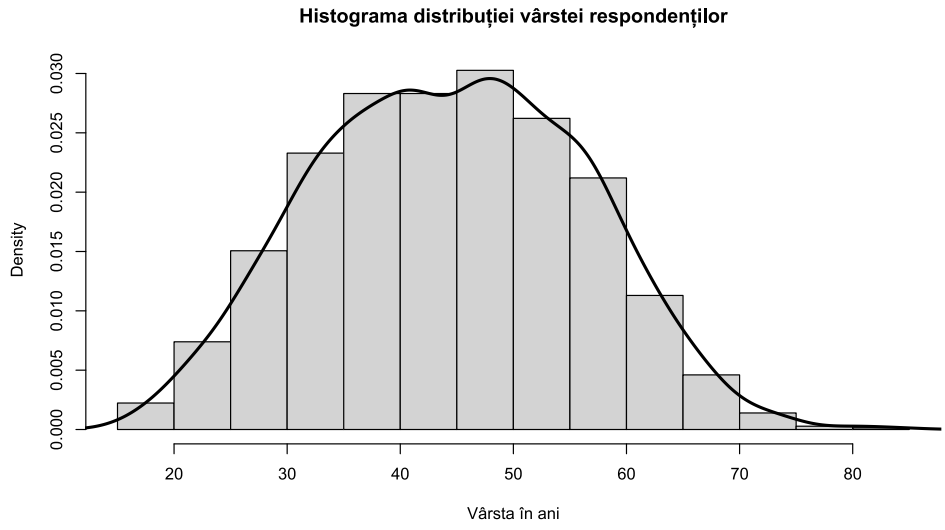
Adăugarea de diagrame de densitate la histograme este destul de simplă, folosind funcția **lines()**. Această funcție poate fi utilizată pentru a adăuga multe tipuri diferite de linii la graficele existente. Pentru variabila noastră, ar arăta astfel:

```
####Histograma vârstelor###

hist(dataset_exemplu$agea,
xlab="Vârsta în ani",
main="Histograma distribuției vârstei respondenților",
prob=T) ###Utilizăm densitățile de probabilitate pe axa verticală###


###Suprapunem o diagramă de densitate pe histogramă###
lines(density(dataset_exemplu$agea), lwd=3) #modificăm grosimea liniei
```

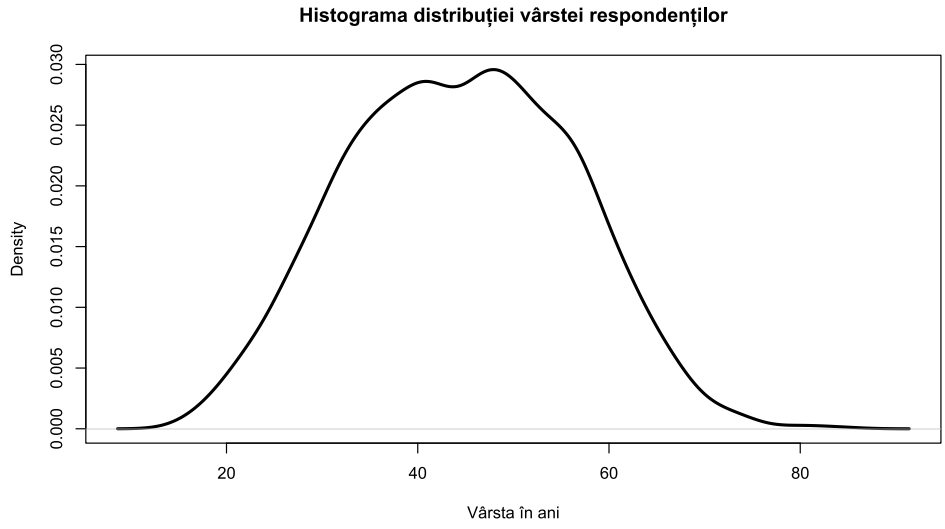
Figura 3.6 Histograma vârstelor și linia densității



De asemenea, putem vizualiza graficul densității separat de histogramă, folosind funcția **plot():**

```
###Generăm o diagramă de densitate fără histogramă###
plot(density(dataset_exemplu$agea),
xlab="Vârsta în ani",
main="Histograma distribuției vârstei respondenților",
lwd=3)
```

Figura 3.7 Diagramă de densitate fără histogramă



Aspectul digramelor cu bare și a histogramelor simple poate fi transformat printr-o serie de funcții de bază din RStudio, precum și prin utilizarea unui pachet numit **ggplot2** (Wickham 2016). Funcțiile pachetului **ggplot2** permit realizarea unor grafice mai complexe pe care le vom exemplifica în Capitolul 6. Însă, la acest moment, putem aplica la funcțiile **plot(), hist()** și **barplot()** trei argumente esențiale care vor oferi îmbunătățiri vizuale semnificative graficelor noastre și anume:

col = “” este argumentul folosit pentru a alege culoarea barelor. Gri este culoarea implicită, dar putem alege altă culoare. Pentru a vedea o listă cu toate culorile disponibile în R, tastăm **colors()** la promptul din fereastra consolei.

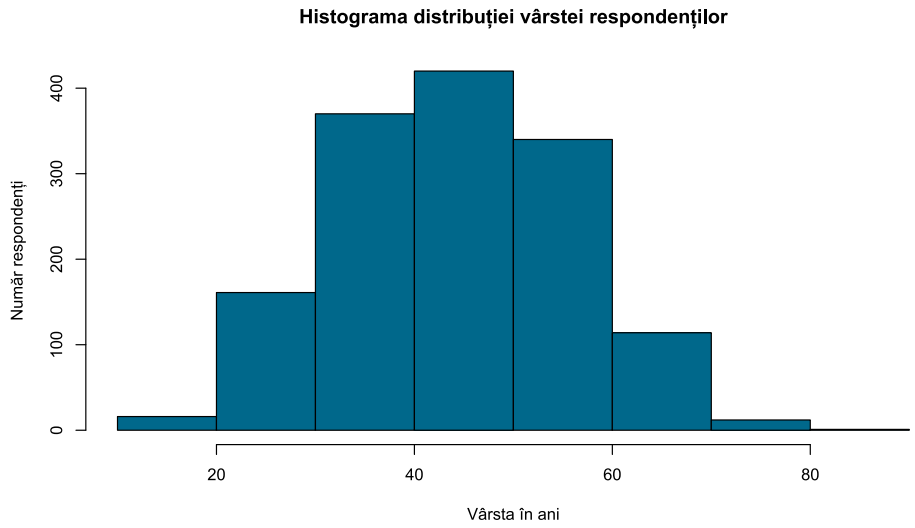
horiz = T este argumentul folosit dacă dorim să întoarcem o diagramă cu bare, astfel încât barele să se transforme în bare orizontale.

breaks = este argumentul utilizat într-o histogramă pentru a modifica numărul de bare utilizate pentru afișarea datelor.

Graficele de mai jos adaugă o parte din aceste informații la graficele pe care le-am realizat mai sus.

```
###Histograma vârstelor###
hist(dataset_exemplu$agea,
xlab="Vârsta în ani",
ylab="Număr respondenți",
main="Histograma distribuției vârstei respondenților",
col = "deepskyblue4", #exemplificăm folosind culoarea albastru#
breaks = 5) #folosim doar 5 categorii
```

Figura 3.8 Histograma vârstelor (schimbarea culorii)

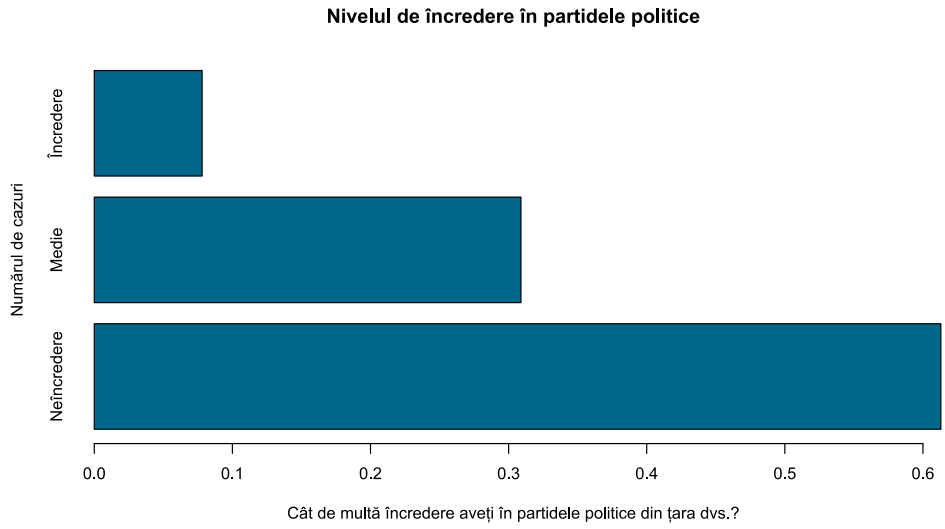


Pentru a inversa dispunerea barelor din verticală în orizontală vom aplica următoarea funcție:

```
###Inversăm barele din vertical în orizontal###
```

```
barplot(prop.table(trstprt.tbl),
names.arg=c("Neîncredere", "Medie", "Încredere"),
xlab="Cât de multă încredere aveți în partidele politice din țara dvs.?",
ylab="Numărul de cazuri",
main="Nivelul de încredere în partidele politice",
horiz=T, #Plotăm barele pe orizontală
col="deepskyblue4") #exemplificăm folosind culoarea albastru
```

Figura 3.9 Distribuția de frecvențe pentru variabila trstprt (inversarea dispunerii barelor)



Tabelele de frecvențe simple, diagramele cu bare și histogramele pot oferi o serie de informații esențiale referitoare la distribuția datelor. Uneori, aceste informații pot indica potențiale erori în codificarea sau colectarea datelor, dar în mare parte sunt utile procesului de familiarizare cu datele. Realizarea acestei analize preliminare prin aplicarea instrumentelor grafice oferă cercetătorilor un nivel ridicat de înțelegere asupra datelor care este esențial pentru realizarea unor analize statistice și grafice complexe.

3.2.2. Tendința centrală a datelor

În continuare vom discuta despre tendința centrală a datelor (în engleză *central tendency*). Pe lângă distribuția de frecvențe a datelor, putem produce o analiză univariată cu ajutorul unor parametri calculați pe baza de date, prin care încercăm să surprindem distribuția informațiilor dintr-o variabilă printr-un singur număr sau observație. Un asemenea tip de informație este cel oferit de indicatorii tendinței centrale, prin care putem reduce întreaga distribuție a valorilor pe care le iau indivizii (unitatea de analiză) pe o variabilă la o singură valoare numerică capabilă să rezume și să ne arate care este, în general, distribuția tuturor valorilor acelei variabile. Media ia valoarea care se obține împărțind suma valorilor tuturor indivizilor din populație la numărul acestora. Ea ilustrează nivelul general al valorilor unei caracteristici și se calculează conform formulei de mai jos.

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

(4)

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. În cazul în care media trebuie calculată pe baza unui tabel de frecvențe, putem folosi formula alternativă, prin care multiplicăm valoarea categoriei cu frecvența de apariție a categoriei.

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

(5)

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.

Cu toții am întâlnit un asemenea indicator atunci când în școală ne calculam la sfârșitul fiecărui semestru sau trimestru media pentru fiecare materie studiată, apoi

media generală anuală. Așadar, **media** (aritmetică) este un asemenea indicator pe care îl putem calcula pentru variabile continue precizate pe niveluri de măsurare cantitative (de interval sau de rapoarte). Ne amintim, de asemenea, că o notă discordantă în raport cu notele tipice pe care le obțineam la o materie, putea distorsiona media, în sus sau în jos, pozitiv sau negativ. Astfel, o notă de 4 ne scădea mult această medie, deși mai aveam două-trei note de 8, 9 sau 10, în vreme ce o notă de 10 ne putea ajuta să ne creștem semnificativ media, deși mai aveam două note foarte proaste, cum ar fi 4. Un alt exemplu sugestiv este acela în care într-o firmă avem un angajat cu un salariu lunar de 10.000 RON, iar alți cinci angajați cu salariul minim pe economie (3.000 RON, la nivelul din ianuarie 2023). Media ne va oferi o imagine distorsionată privind distribuția generală a salariilor. Am putea în mod fals să concluzionăm că, în medie, angajații din această companie au un salariu cu 38% mai mare decât salariul minim pe economie. Putem, așadar, concluziona că media este influențată și distorsionată de valorile extreme, ceea ce nu ne ajută să înțelegem individul tipic sau distribuția generală a valorilor variabilei.

În RStudio comanda pentru a calcula media aritmetică este următoarea:

```
mean(dataset_exemplu$trstprt)
## [1] 1.465132
```

Observăm, astfel, că media aritmetică pentru variabila *trstprt* este de 1.46 fapt care ne indică că majoritatea respondenților se află între categoria de neîncredere și încredere medie.

Pentru a scăpa de distorsiunile produse de cazurile extreme (*outliers*), putem calcula **mediana** (în engleză *median*). Ea este o măsură a tendinței centrale folosită atunci când avem variabile ordinale, de interval sau de rapoarte, neputând fi calculată pentru variabile aflate la nivel de măsurare nominal. Mediana reprezintă valoarea pe care o ia individul de mijloc (cel care are în stânga lui tot atâtea unități câte are și în dreapta sa). Mediana este foarte utilă pentru a ilustra tendința centrală a datelor,

deoarece ea nu este influențată de valorile extreme. O putem calcula folosind formula simplă de mai jos.

$$Me = \frac{n}{2} + 1 \quad (6)$$

Atunci când avem intervale, putem folosi formula de mai jos.

$$Me = l + \frac{\frac{N}{2} - nc}{n} \times L \quad (7)$$

unde: 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 ce conține mediana (adică numărul de observații care iau valori mai mici decât l), n este frecvența intervalului care conține mediana, iar L este lărgimea sau mărimea intervalului care conține mediana.

În RStudio putem afla mediana valorilor unei variabile folosind următoarea linie de cod:

```
median(dataset_exemplu$trstprt)
## [1] 1
```

Pentru variabilele aflate pe nivelul de măsurare nominal nu putem calcula nici unul dintre acești indicatori ai tendinței centrale, nici mediana, nici media, deoarece codurile numerice folosite pentru a identifica fiecare categorie definită de variabila nominală nu au nici un sens matematic. În plus, aceste categorii pot fi permutate, ceea ce, ipotetic, ar conduce la schimbarea valorii unui asemenea indicator al tendinței centrale, în realitate datele oferindu-ne aceleași informații, neschimbate. De aceea, singurul indicator al tendinței centrale pe care îl putem calcula și analiza pentru variabilele nominale este **valoarea modală** (în engleză *mode*). Ea este valoarea ce caracterizează individul tipic, deci o întâlnim cel mai frecvent în baza de date. Cele

mai multe cazuri, situație numită și pluralitatea cazurilor, iau această valoare. Pluralitatea cazurilor este, de exemplu, folosită în cazul alegerilor pentru primarii din România, din 2016 până în prezent, unde formula electorală folosită pentru a identifica numărul minim necesar de voturi pentru a câștiga alegerile pentru funcția de primar este pluralitatea $n_i + 1$, unde n_i reprezintă numărul de voturi obținut de al doilea candidat în ordinea voturilor. Într-un mod similar putem identifica frecvența minimă pe care o valoare a unei variabile trebuie să o ia pentru a afirma că ea este valoarea modală sau tipică a acelei variabile de interes. Valoarea modală poate fi identificată, prin urmare, ca fiind valoarea cea mai frecventă în rândul indivizilor. Este posibil ca unele variabile să fie bimodale sau plurimodale, atunci când mai multe categorii au un număr egal de indivizi (cazuri).

În cazul în care valoarea modală se calculează pentru valori grupate, putem folosi formula:

$$Mo = l_i + h_i \frac{f_i + f_1}{(f_i - f_{i-1}) + (f_i - f_{i-2})} \quad (8)$$

unde l_i reprezintă limita inferioară a intervalului modal, h_i este mărimea acestui interval, f_i este frecvența intervalului modal, f_{i-1} este frecvența intervalului care precedă intervalul modal, iar f_{i-2} este frecvența intervalului care succedă intervalul modal.

Individul tipic poate fi observat într-o diagramă cu bare, deoarece va fi cea mai înaltă bară din grafic. Pentru graficul care arată nivelul de încredere în exemplul aplicat din Figura 3.2 putem afirma că modulul datelor este categoria neîncredere, adică valoarea 1 așa cum avem definit numele categoriei în setul de date și în Figura 3.1.

3.2.2.1. Reprezentarea grafică a tendinței centrale cu ajutorul boxplots

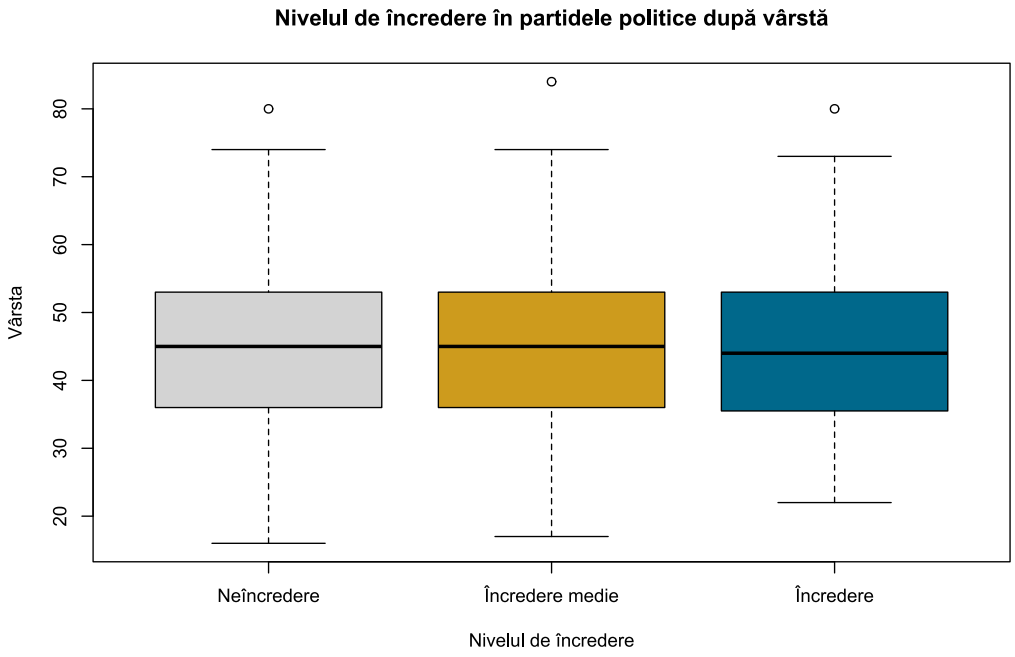
Similar diagramelor cu bare și a histogramelor sunt și graficele de tip *boxplot*. Acest tip de grafic este foarte util în reprezentarea grafică a indicatorilor care descriu tendința centrală a datelor, dar și atunci când vrem să comparăm două sau mai multe grupuri. Să presupunem că vrem să reprezentăm grafic indicatorii tendinței centrale pentru variabilele *trstprt* (încrederea în partidele politice), *vote* (participarea și ultimele alegeri) și *stfdem* (nivelul de satisfacție cu democrația) în funcție de variabila care descrie vârsta respondenților (*agea).* În termeni grafici, vrem să reprezentăm pe axa X variabila de interes (*trstprt, vote, stfdem)*, iar pe axa Y variabila în funcție de care sunt raportate variabilele de pe axa X, în cazul nostru *agea.* Comanda în RStudio pentru a crea un grafic de tip *boxplot* este următoarea:

```
### Boxplot care indică nivelul de încredere în funcție de vârstă ###
boxplot(agea~trstprt,data=dataset_exemplu,
names = c("Neîncredere", "Încredere medie", "Încredere"),
col = c("lightgray", "goldenrod3", "deepskyblue4"),
xlab = "Nivelul de încredere",
ylab = "Vârsta",
main = "Nivelul de încredere în partidele politice după vârstă")
```

Figura 3.10 ilustrează graficul rezultatului liniei de cod de mai sus. Cum interpretăm acest tip de reprezentare grafică? Înainte de toate trebuie să înțelegem ce semnifică fiecare element al graficului. Linia îngroșată de pe fiecare culoare a categoriilor variabilei noastre reprezintă mediana sau punctul de mijloc. Casetele colorate reprezintă intervalul intercuartilic (IQR) care include mijlocul a 50% din date. Cuartilele împart distribuția în 4 grupuri egale (25% din total, fiecare). Mediana reprezintă cuartila a doua sau bara de mijloc a casetei colorate. Celelalte două linii subțiri din partea de sus și de jos denotă intervalul sau amplitudinea distribuției valorilor (nivelul minim și nivelul maxim), iar valorile care ies din acest interval

(*outliers)* sunt reprezentate de un cerculeț aflat deasupra sau dedesubtul liniilor care indică valoarea maximă și minimă a datelor noastre.

Figura 3.10 Nivelul de încredere în partidele politice după vârstă

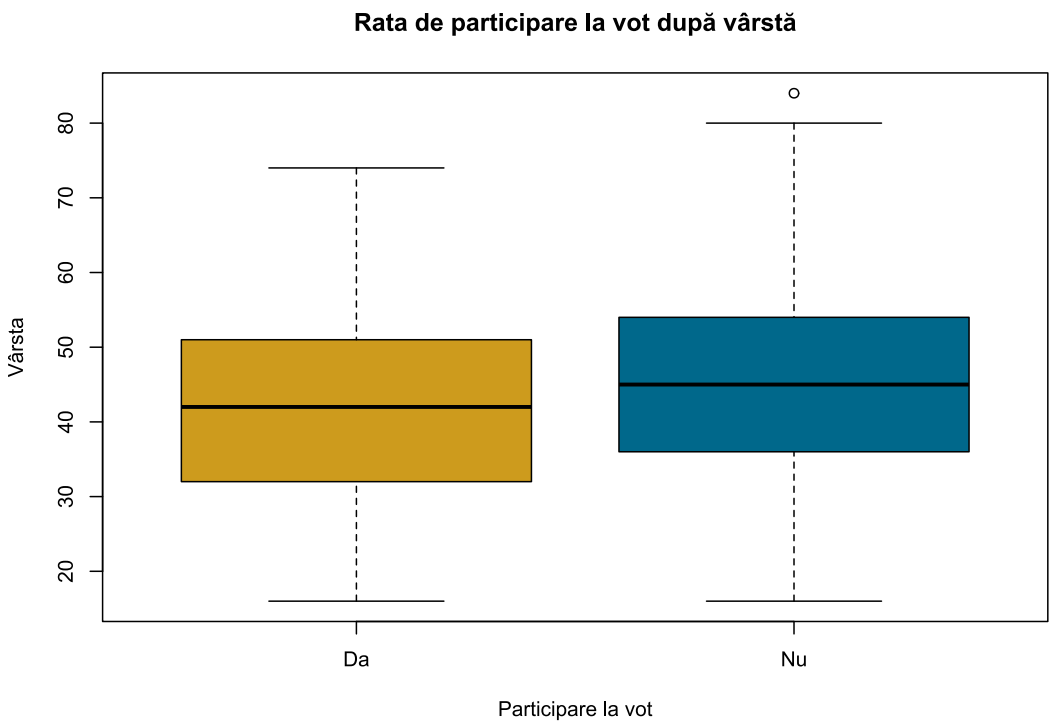


Acum că am înțeles ce reprezintă fiecare element al graficului, să încercăm să interpretăm informațiile oferite. Observăm că datele sunt prezentate pe fiecare categorie în parte a variabilei de interes, iar vârsta medie a fiecărei categorii este 45 de ani. Din grafic putem aprecia că respondenții care nu au încredere în partidele politice se află în categoria de vârstă cuprinsă între aproximativ 38 și 53 de ani, în timp ce cei care au încredere au o vârstă cuprinsă între 40 și 50 de ani. În continuare, vom prezenta comenzile pentru celelalte două variabile pentru care vrem să obținem o reprezentare grafică a tendinței centrale, și anume variabilele *vote* și *stfdem*.

```
###Boxplot care indică participarea la alegeri în funcție de vârstă###
boxplot(agea~vote,data=dataset_exemplu,
names = c("Da", "Nu"),
col = c("goldenrod3", "deepskyblue4"),
```

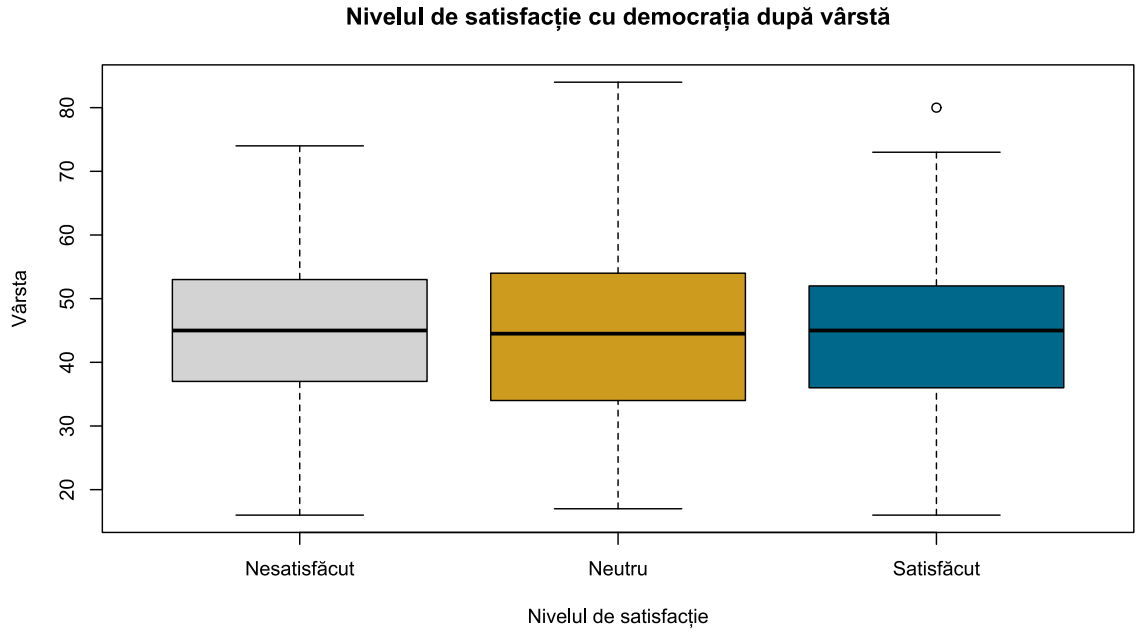
```
xlab = "Participare la vot",
ylab = "Vârsta",
main = "Rata de participare la vot după vârstă")
```

Figura 3.11 Rata de participare la vot




```
###Boxplot care indică nivelul de satisfacție în funcție de vârstă###
boxplot(agea~stfdem,data=dataset_exemplu,
names = c("Nesatisfăcut", "Neutru", "Satisfăcut"),
col = c("lightgray", "goldenrod3", "deepskyblue4"),
xlab = "Nivelul de satisfacție",
ylab = "Vârsta",
main = "Nivelul de satisfacție cu democrația după vârstă")
```

Figura 3.12 Nivelul de satisfacție cu democrația



3.2.3 Dispersia datelor

Analiza preliminară a datelor constă și în determinarea dispersiei datelor analizate. Această dispersie poate fi examinată doar pentru variabilele aflate pe nivel de măsurare de interval sau de rapoarte, nu și pentru variabilele calitative. Pentru a face acest tip de analiză vom folosi indicatori sau măsuri de dispersie. Aceștia mai sunt numiți și indicatori de variație sau de împrăștiere, și ne ajută să îmbogățim informația pe care ne-o oferă indicatorii tendinței centrale. Indicatorii de dispersie măsoară gradul de împrăștiere a indivizilor (cazurilor) în cadrul seriei de valori pe care aceștia le iau. Acești indicatori sunt extrem de utili, deoarece ne oferă o imagine asupra caracterului omogen sau eterogen al populației / eșantionului studiat (Rotariu et al. 1999, 49). Printre cele mai des folosite metode pentru a calcula dispersia datelor amintim amplitudinea sau intervalul valorilor (în engleză *range*), abaterea intercuartilă (*interquartile range*), abaterea medie de la medie, varianța (*variance*), abaterea standard (în engleză *standard deviation*), coeficientul de variație, și doi

indicatori ai formei distribuției, cel de asimetrie sau alungire (*skewness*) și cel de aplatizare sau boltire (*kurtosis*).

Amplitudinea (sau mărimea intervalului) este indicatorul care arată diferența dintre valoarea maximă și valoarea minimă a variabile analizate. Notele pe care un elev le poate primi în sistemul de învățământ românesc au o amplitudine maximă de 9 (10-1).

Înainte de a calcula amplitudinea în RStudio, trebuie să parcurgem doi pași: (1) să calculăm cea mai mică valoarea existentă în variabila analizată și (2) să calculăm cea mai mare valoare existentă în variabila de interes. Ulterior, pentru a determina valoarea intervalului de amplitudine vom face diferența dintre valoarea maximă și valoarea minimă.

```
### calculăm cea mai mică valoare ###
min(dataset_exemplu$trstprt)
## [1] 1

### calculăm cea mai mare valoare ###
max(dataset_exemplu$trstprt)
## [1] 3

### calculăm intervalul valorilor ###
max(dataset_exemplu$trstprt)-min(dataset_exemplu$trstprt)
## [1] 2
```

Deoarece amplitudinea este influențată de valorile extreme, putem apela la un alt indicator care are la bază ordonarea cazurilor în funcție de valorile luate de acestea. Aceasta este **abaterea intercuartilă** și reprezintă diferența dintre cuartila a treia (Q3), ce reprezintă nivelul până la sunt 75% dintre cazuri, și cuartila întâi (Q1), ce reprezintă nivelul până la care avem 25% din distribuția cazurilor ce iau valori pe acea variabilă. Ne amintim că mediana este egală cu a doua cuartilă (Q2), ce reprezintă 50% din cazuri, în vreme ce cuartila a patra (Q4) reprezintă totalul cazurilor. Abaterea intercuartilă are avantajul că nu este influențată de valorile extreme.

Putem avansa în evaluarea gradului de dispersie a valorilor variabilei și putem încerca să identificăm modul în care valoarea luată pe variabilă de fiecare individ este comparabilă cu o altă valoare (de exemplu valoarea individului tipic), calculând astfel o abatere medie de la o anumită valoare. De exemplu, putem raporta valoarea fiecărui individ (observația) la valoarea medie a distribuției. Acest indicator poartă numele de **abatere de la medie** și reprezintă diferența dintre valoarea pe care o ia respectiva observație și media variabilei.

$$A = x_i - \bar{x} \tag{9}$$

unde x_i reprezintă valoarea individului i din distribuția variabilei, iar $\bar{x}$ reprezintă media acelei distribuții.

Putem calcula media acestor abateri de la medie:

$$A = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} \tag{10}$$

Însă, putem constata că suma tuturor abaterilor individuale de la medie este egală cu 0:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{11}$$

De aceea, folosim suma valorilor absolute (în modul) ale abaterilor individuale de la medie (iar indicatorul obținut poartă numele de abaterea medie absolută) sau folosim suma pătratelor acestor abateri (un indicator pe care îl numim varianță). **Varianța** este, prin urmare, pătratul abaterii de la medie sau abaterea medie pătratică de la media grupului (sau eșantionului) pe care îl analizăm. Ea ne arată cât de departe se află în medie fiecare valoare din distribuție față de media distribuției. Dat fiind că variabilele pot măsura caracteristici extrem de diverse, fiecare având propria unitate de măsură de exemplu, vârsta măsurată în ani, greutatea măsurată în kilograme, venitul măsurat în ROL, venitul măsurat în RON), este dificil să interpretăm varianța,

raportându-ne la distribuția valorilor unei variabile și la eșantion sau populație. Varianța nu are o valoare standard comparabilă, ci variază în unități de măsură (Rotariu et al. 1999, 53–58).

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(12)

În cazul în care o calculăm pentru un eșantion, nu pentru o populație, varianța poate fi calculată cu o corecție *n-1*, unde *n* este mărimea eșantionului (Agresti și Finlay 2014, 48, 118–19), pe care o vom folosi și în cazul abaterii standard (conform formulei 14 de mai jos).

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(13)

Dacă extragem rădăcina pătrată din valoarea variației, obținem un alt indicator important, numit **abaterea standard** sau **deviația standard**. Aceasta măsoară modul în care valorile sunt distribuite (se abat) în jurul mediei și mărimea acestor abateri de la medie. Într-o distribuție normală, cum este cea a mediilor eșantioanelor aleatoare extrase succesiv dintr-o populație, aproximativ 68% dintre valori se vor afla la plus / minus o abatere standard de la medie, iar aproximativ 95% dintre valori se vor afla la plus / minus două abateri standard de la medie, iar aproximativ 99% dintre valori se vor afla la plus / minus trei abateri standard de la medie (Sandu 1992, 205). Abaterea standard crește pe măsură ce valorile pe care le iau observațiile se abat mai mult de la medie. De asemenea, abaterea standard se caracterizează prin constanță atunci când transformăm datele pentru care ea este calculată folosind operațiuni de înmulțire sau împărțire la un număr constant. Nu în ultimul rând, abaterea standard nu este afectată de adunarea sau scăderea unui număr constant la valorile pe care le iau observațiile din datele pentru care calculăm indicatorul abaterii standard (Rotariu et al. 1999, 56).

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(14)

Un alt indicator important pentru estimarea omogenității sau eterogenității distribuției datelor este **coeficientul de variație**. Acesta reprezintă raportul dintre abaterea standard și media variabilei. Este util în compararea variației a două variabile măsurate în același eșantion sau populație și ne va permite să afirmăm că o populație este mai omogenă pentru caracteristica A decât pentru caracteristica B. Cu ajutorul variației sau al abaterii standard asemenea comparații nu ar fi posibile deoarece ambii indicatori sunt măsuri dimensionale și preiau unitatea de măsurată a caracteristicii pentru care au fost calculați. Am compara mere cu pere. De aceea, folosim coeficientul de variație, calculat astfel:

$$CV = \frac{\sigma}{\bar{x}}$$

(15)

În strânsă legătura cu gradul de omogenitate sau de eterogenitate al observațiilor pe variabilă, este foarte important să observăm și să măsurăm forma distribuției observațiilor. Aceasta poate fi făcută cu ajutorul a doi indicatori: indicatorul de asimetrie sau alungire, uneori numit și oblicitate (*skewness*) și cel de boltire (*kurtosis*).

Alungirea sau **asimetria** se referă la gradul în care distribuția observațiilor față de medie este simetrică. Aceasta poate fi pozitivă, atunci când distribuția este alungită la dreapta, sau negativă, atunci când alungirea distribuției este spre stânga. Indicatori mai mari de +1 și mai mici de -1 indică mai degrabă o distribuție alungită. Asimetria ne oferă și o imagine asupra poziției celor trei indicatori ai tendinței centrale. Atunci când o distribuție este alungită la dreapta media este mai mare decât mediana care , la rândul ei, este mai mare decât valoarea modală. Într-o distribuție alungită la stânga media este mai mică decât mediana, care este mai mică decât valoarea modală (Rotariu et al. 1999, 60–61).

$$asimetria(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3}$$

(16)

Aplatizarea sau **boltirea** unei distribuții ne oferă informații privind gradul în care distribuția variabilei analizate prezintă cazuri deviante (*outliers*). Boltirea poate fi redusă (de regulă cu valori mai mici decât -1), numită platicurtică (dacă avem puține cazuri deviante), sau ridicată (de regulă cu valori mai mari de +1), numită și lepticurtică (dacă avem mai degrabă multe cazuri deviante).

$$boltirea(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3$$

(17)

În RStudio putem calcula abaterea standard cu ajutorul comenzii:

```
sd(dataset_exemplu$trstprt)
## [1] 0.6366108
```

Pentru a ne putea folosi de indicatorii tendinței centrale și ai dispersiei unei distribuții în înțelegerea în ansamblu a datelor noastre, este bine ca aceste rezultate să fie integrate într-un tabel ce raportează indicatori ai statisticii descriptive. Multe dintre articolele și cărțile care raportează analize empirice includ aceste tabele. Pentru a obține un tabel cu date descriptive în RStudio vom folosi următoarea funcție din pachetul **psych** (Revelle 2022), care trebuie instalat și încărcat în sesiunea de lucru deschisă.

```
library(psych)
```

Ulterior, vom aplica codul de mai jos pentru a calcula doar indicatori descriptivi pentru anumite variabile din setul nostru de date:

```
###tabel de statistică descriptivă doar pentru categoriile unei variabile###
describe(dataset_exemplu)
describeBy(dataset_exemplu, group=dataset_exemplu$trstprt)
```

De asemenea, putem alege să calculăm doar statisticile descriptive pentru anumite variabile din setul nostru de date:

```
describe(dataset_exemplu[ , c('trstprt', stfeco, 'agea')], fast=TRUE)
##         vars    n  mean   sd min  max range   se
## trstprt    1 1434  1.47  0.64   1    3     2 0.02
## stfeco     2 1434  1.81  0.73   1    3     2   NA
## agea       3 1434  44.48 11.73  16  84    68   NA
```

O altă funcție pe care o putem folosi pentru a integra rezultatele statisticilor descriptive este funcția **summary()**.

```
summary(dataset_exemplu)
```

Tabelul de ieșire al rezultatelor statistice poate fi salvat folosind pachetul stargazer (Hlavac 2022).

```
library(stargazer)
statistica_descriptiva <- describe(dataset_exemplu[ , c('trstprt', stfeco,
'agea')], fast=TRUE)
###generare tabel statistică descriptivă în html###
stargazer(statistica_descriptiva, type="html", out="tabel_descriptiv.htm")
```

Tabelul generat de funcția **stargazer()** în format "html" poate fi editat și transpus în Microsoft Word sau alte programe de editare conținut după cum urmează:

Tabel 3.4 Statistică descriptivă a variabilelor analizate

Variable	N	Mean	St. Dev.	Min	Max	Range	SE.
trstprt	1434	1.47	0.64	1	3	2	0.02
stfeco	1434	1.81	0.73	1	3	2	0.02
agea	1434	44.48	11.73	16	84	68	0.31

În primele două secțiuni ale capitolului 3 am explorat setul de date, variabilele și valorile acestora. În continuare, vom discuta despre diferitele modele statistice (bivariate sau multivariate) pe care le putem utiliza pentru a testa ipoteze și a face inferențe pe baza rezultatelor statistice.

3.3. Analiza bivariată

În capitolul anterior am discutat despre modul în care putem identifica și ilustra caracteristici parametrice ale datelor noastre, analizând variabilă cu variabilă. Această analiză este însă una descriptivă, deci ne ajută doar la descrierea fiecărei variabile în parte. De exemplu, dacă aflăm din datele pe care le analizăm că 35% dintre respondenți nu au încredere în guvern, sau că rata șomajului în România a fost de 5,2% în septembrie 2022, iar în alte țări a fost de minim 2,6% și maxim 12,7%, putem ilustra numeric și grafic aceste informații. Totuși, ele rămân una sau mai multe fotografii ale realității studiate. Este foarte posibil să ne punem întrebări mai complexe decât „Câți respondenți au încredere în guvern?" sau „Care este șomajul în România comparativ cu alte state membre ale UE?". Astfel, am putea dori să înțelegem ce determină această variație a șomajului în Uniunea Europeană, ce determină volatilitatea șomajului în fiecare țară, sau de ce nu au încredere românii în guvern. Înainte, însă, de a răspunde la aceste întrebări, o etapă importantă a analizei este aceea prin care identificăm modele în baza de date, și dorim să vedem dacă există o relație

între variabile, luate două câte două, care este intensitatea acestei relații și care este direcția ei (pozitivă sau negativă).

De exemplu, am putea evalua relația dintre rata șomajului și rata de creștere a PIB, cea dintre încrederea în guvern și încrederea în parlament, cea dintre temperatura exterioară și vânzările de înghețată, relația dintre vârstă și gradul de educație, sau relația dintre participarea politică și gradul de informare. În acest capitol vom discuta despre aceste relații dintre două variabile, cunoscute sub numele de relații bivariate. În capitolele următoare vom discuta despre relațiile multivariate și diverse modele statistice care se aplică în analizarea acestora.

Pentru analizarea relației dintre două variabile (să le numim de exemplu variabila X, respectiv Y) folosim date pe care le numim date bivariate. Ele ne oferă valori sau măsurători a două caracteristici diferite pentru fiecare individ (unitate de analiză) din eșantion sau populație. Atunci când analizăm relațiile dintre variabile calitative (dihotomice, nominale sau ordinale), relația este numită relație de asociere; când însă analizăm relațiile dintre două variabile cantitative (de interval și raportare) aceasta se numește corelație. Existența unei asemenea relații perfecte (fie asociere, fie corelație) ne permite să estimăm variația unei variabile (indiferent care din cele două, să spunem că aici este vorba despre variabila Y) pe baza informațiilor pe care le avem despre variația celeilalte variabile (pe care să o numim variabila X). Cu alte cuvinte, dacă avem o relație puternică între aceste două variabile, cunoscând valorile pe care indivizii (cazurile despre care culegem informații) le iau pe variabila X putem reduce erorile de estimare a valorilor pe care aceștia le iau pe cealaltă variabilă (Y).

Acest tip de estimare este un pas foarte important pe care îl facem pentru a înțelege modul în care putem explica relațiile cauzale dintre două sau mai multe variabile. Însă, la acest moment al evaluării relației dintre cele două variabile nu discutăm despre o relație de tip cauză-efect, ci doar despre măsura în care două variabile variază concomitent. De fapt, această relație cauză-efect este una mai degrabă teoretică, decât statistică, după cum vom discuta în secțiunea 3.4. Asocierea sau corelația dintre două caracteristici diferite ale indivizilor (unităților noastre de observație) poate să fie pur întâmplătoare. De exemplu, dacă pe baza unor observații constatăm că studenții care iau note mai mari la statistică tind să aibă părul mai lung,

am putea spune că există un model de relație între lungimea părului și nota de la statistică, relație pe care o putem reprezenta grafic sau o putem estima prin indicatori ai asocierii. Totuși, nu ar fi nici plauzibil, nici logic, să argumentăm că lungimea părului este o cauză pentru mărimea notei de la statistică, sau că mărimea părului este determinată de mărimea notei de la statistică. Aceste tipuri de relații, uneori extrem de puternice din punct de vedere statistic, sunt numite relații aparente sau eronate (în engleză *spurious relationship*). Chiar dacă statistic sunt relații puternice, ele pot fi însă pur întâmplătoare. Întotdeauna trebuie să raportăm aceste analize și concluziile statistice pe teorie. Dacă nu există un suport teoretic pentru a explica rațional relația dintre cele două variabile, atunci este foarte probabil ca aceasta să fie una aparentă.

Reprezentarea datelor bivariate poate fi făcută numeric, cu ajutorul unor indicatori capabili să exprime într-un singur număr gradul în care există această relație, intensitatea sa și direcția relației; sau vizual, folosind fie tabele de asociere sau de contingență, fie grafice, cu reprezentări vizuale capabile să prezinte perechea de valori $[X_i; Y_i]$ pe care fiecare caz le ia pentru cele două variabile, X, respectiv Y. O asemenea reprezentare grafică este cea prin care folosim un grafic cu două dimensiuni, în care fiecare punct din grafic reprezintă perechea de valori pe care individul le ia pentru fiecare din cele două variabile reprezentate de cele două dimensiuni (abscisa, axa OX, sau axa orizontală, respectiv ordonată, axa OY, sau axa verticală). Această reprezentare produce o diagramă cu puncte, uneori numită nor de puncte sau grafic de dispersie (în limba engleză *scatterplot*).

Nici tabelele de asociere, nici analiza grafică nu pot însă să ne spună cât de intensă este relația dintre cele două variabile analizate. Tabelele de asociere sau de contingență pot fi deseori extrem de complexe, de exemplu, atunci când variabilele incluse în analiză au multe categorii sau sunt continue, deci numărul de coloane și rânduri este mare, și când există o distribuție mai degrabă eterogenă pe celulele tabelului. Identificarea în aceste tabele a unui model sau a intensității relației de asociere devine aproape imposibilă folosind doar analiza vizuală. Uneori, atunci când relația nu este una foarte puternică, reprezentarea grafică nu ne va edifica nici în ce privește direcția asocierii. Pentru aceste informații (intensitatea și direcția asocierii) avem nevoie de indicatori care, printr-un singur număr, să ne poată indica ambele

informații. Astfel, calculăm și interpretăm indicatori ai asocierii sau corelației pe care îi vom prezenta în secțiunile de mai jos.

În secțiunea 3.3.1 vom exemplifica modul de utilizare a programului RStudio pentru a determina puterea relației de asociere dintre două variabile nominale și/sau ordinale, iar în secțiunea 3.3.2 vom exemplifica analiza de corelație dintre două variabile măsurate pe nivel de interval sau de rapoarte.

3.3.1. Asocierea

Asocierea reprezintă relația dintre două variabile calitative (nominale sau ordinale). Ea este prezentă atunci când există o legătura între valorile unei variabile și valorile celeilalte variabile, atunci când aceste variabile variază concomitent într-un sens pe care îl putem identifica din punct de vedere statistic. De exemplu, când constatăm că există o legătură între mediul de rezidență (urban sau rural) și preferința de vot pentru un partid. Această asociere poate fi reprezentată printr-un tabel de contingență în care să prezentăm distribuția valorilor celor două variabile. Observând această distribuție am putea estima dacă există asociere între aceste două variabile.

Pentru început putem produce un tabel de contingență precum tabelul 3.5., care conține date ipotetice pentru două variabile calitative (dihotomice, de data aceasta, prin urmare este un tabel cu două rânduri și două coloane). Distribuțiile de frecvențe din acest tabel ne arată că avem un număr egal de votanți ai partidului A care locuiesc în mediul urban și rural. Aceeași distribuție este valabilă și pentru votanții partidului B. În această distribuție putem afirma că nu există nici o legătura între distribuția votului pentru partid și distribuția indivizilor pe mediul de rezidență. Numim această relație ca fiind una de independență statistică. Cunoscând mediul de rezidență al unui subiect nu putem estima care este preferința sa de vot: sunt șanse egale de a vota cu partidul A sau cu partidul B. Cunoscând preferința de vot a unui subiect, sunt șanse egale de a estima că locuiește în mediul urban sau în mediul rural.

Tabel 3.5 Tabel de asociere – independență statistică

Rezidență / Preferință vot partid	Partidul A	Partidul B
Urban	200 (50%)	200 (50%)
Rural	200 (50%)	200 (50%)

Un tabel de contingență ne poate oferi informații și despre intensitatea sau direcția asocierii. Transformând datele ipotetice din tabelul de mai sus, putem ilustra în tabelul 3.6 situația în care identificăm o relație de asociere între cele două variabile analizate. Astfel, putem observa în distribuția datelor că toate persoanele care votează cu partidul B sunt din mediul rural, în vreme ce doar un sfert din votanții partidului A locuiesc în mediul rural. Astfel, cunoscând că o persoană locuiește în mediul rural, putem spune că sunt 75% șase să voteze cu partidul B. Cunoscând că o persoană locuiește în mediul urban, putem spune că sunt 100% șanse să voteze cu partidul A. Putem concluziona că există o relație de asociere puternică între cele două variabile. Subliniem, însă, încă o dată că această relație nu este una cauzală. Identificarea relației de asociere nu ne ajută să răspundem la întrebarea în ce măsură X este o cauză pentru Y, sau în ce măsură X îl determină pe Y. Acest informații pot fi obținute folosind analiza de regresie, despre care vom discuta în secțiunile 3.4 și 3.5.

Tabel 3.6 Tabel de asociere – relație puternică

Rezidență / Preferință vot partid	Partidul A	Partidul B
Urban	300 (75%)	0 (0%)
Rural	100 (25%)	400 (100%)

Un tabel de asociere este cel mai edificator pentru analiza relației dintre două variabile cu un număr redus de categorii atunci când avem o relație perfectă de asociere pentru variabilele analizate. De exemplu, în tabelul 3.7 avem ilustrarea unei relații de asociere perfectă, în care distribuția indivizilor pe una dintre variabile este estimată perfect prin distribuția lor pe cealaltă variabilă.

Tabel 3.7 Tabel de asociere – relație perfectă

Rezidență / Preferință vot partid	Partidul A	Partidul B
Urban	400 (75%)	0 (0%)
Rural	0 (0%)	400 (100%)

Subliniem încă o dată că în lipsa unei explicații fundamentate teoretic, orice relație între variabile, chiar dacă e perfectă, nu ne arată că avem neapărat o relație de cauzalitate sau determinare între variabile. Orice analiză statistică trebuie să se bazeze pe teorie (Rotariu et al. 1999, 126).

Asocierea poate fi măsurată și prin coeficienți de asociere. Aceștia reprezintă exprimarea printr-un indicator statistic a calității asocierii dintre două variabile. Pe lângă prezența sau absența unei asocieri, coeficienții ne oferă și informații cu privire la intensitatea acesteia. Aceștia variază între 0 și 1 (unde 0 înseamnă că nu există nici o relație, iar 1 înseamnă ca exista relație de asociere perfectă) sau între -1 și +1 (unde 0 înseamnă că nu există nici o relație; -1 înseamnă că există o relație negativă perfectă, iar +1 înseamnă că există o relație pozitivă perfectă).

Coeficienții de asociere pot fi simetrici sau asimetrici (Rotariu et al. 1999, 137). Cei simetrici tratează variabilele *X* și *Y* simetric fără a diferenția între cele două variabile, în termeni de utilitate a unei variabile pentru a prezice valorile celeilalte variabile (fără a implica o relație de cauzalitate, putem reduce această diferență ca fiind una între variabila independentă și variabila dependentă). Coeficienții asimetrici

diferențiază între variabile din perspectiva prezicerii valorilor unei variabile în funcție de valorile celeilalte variabile.

Pentru variabilele măsurate pe **nivel nominal**, unde, ne reamintim, valorile pot fi permutate în ceea ce privește ordinea lor sau pot să primească orice codificare numerică fără nici un sens matematic, nici măcar cel de ordine, putem să calculăm și să interpretăm doar coeficienți de asociere care identifică prezența sau absența, respectiv intensitatea asocierii dintre variabile. Asocierea dintre două variabile nominale poate fi măsurată prin aplicarea unor coeficienți de măsurare specifici variabilelor nominale:

- **Pearson C:** este un coeficient simetric care poate lua valori între 0 și 1. Reprezintă procentul variației maxime posibile a celor două variabile. Este o simplificare la nivel nominal a coeficientului de corelație Pearson r folosit pentru variabilele cantitative, despre care vom discuta în secțiunea următoare.
- **Cramer V:** este un coeficient simetric care poate fi interpretat ca procent al maximei variații posibile a celor două variabile nominale din analiză. Variază între 0 și 1.
- **Lambda:** este un coeficient asimetric. Variază între 0 și 1. Reprezintă procentul cu care se reduc erorile de predicție a valorilor variabilei dependente atunci când cunoaștem valorile variabilei independente. Dacă lambda este egal cu 0, atunci cunoașterea valorilor variabilei independente nu ne ajută să estimăm valorile variabilei dependente și putem, astfel, concluziona absența relației între variabile.
- **Testul Chi Pătrat (Hi Pătrat):** este un test prin care măsurăm calitatea asocierii. Acesta se măsoară pentru a testa ipoteza nulă (de independență statistică), potrivit căreia nu avem nici o asociere a variabilelor (a coloanelor și a rândurilor în tabelele de contingență). Chi pătrat este utilizat și pentru a testa dacă două variabile sunt sau nu asociate. În funcție de tipul variabilelor (nominale, ordinale etc) interpretarea coeficienților se face diferit. Testul este folosit atunci când eșantionul este mare și numărul valorilor din fiecare celulă a tabelului de asociere este mare (de ordinul zecilor sau sutelor).

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

(18)

unde: O_i reprezintă frecvențele observate (numite și frecvențe empirice), iar A_i reprezintă frecvențele așteptate (numite și frecvențe teoretice).

În continuare vom calcula intensitatea asocierii dintre variabilele ipotezei 1a care testează dacă cei care se declară fericiți au mai multă încredere în politicieni. Prima dată vom aplica testul Chi pătrat pentru a testa existența unei situații de independență statistică între cele două variabile. În RStudio trebuie să încărcăm pachetul **MASS** (Venables și Ripley 2002), să închidem pachetul **dplyr** (Wickham et al. 2022) și să aplicăm următoarele comenzi pentru aplicarea testului Chi pătrat.

```
### Detașăm pachetul dplyr și încărcăm pachetul MASS. ###
detach("package:dplyr")
library("MASS")
### Creăm un dataframe din setul de date principal.###
chi.data <- data.frame(dataset_exemplu$trstplt, dataset_exemplu$happy)

### Creăm un tabel cu variabilele necesare.###
chi.data = table(dataset_exemplu$trstplt, dataset_exemplu$happy)
print(chi.data)
##                   fericit nefericit neutru
##   incredere           105         2     19
##   incredere medie     312        19     97
##   neincredere         623        44    213

### Calculăm testul Chi-Square.###
print(chisq.test(chi.data)
##  Pearson's Chi-squared test
##
## data:  chi.data
## X-squared = 9.2655, df = 4, p-value = 0.0548
```

Testul Chi-pătrat are scopul de a testa probabilitatea ca o asociere să fie întâmplătoare. Se mai numește și statistică *"goodness of fit"*, deoarece măsoară cât de bine se potrivește distribuția observată a datelor cu distribuția așteptată dacă variabilele sunt independente. Cum interpretăm rezultatul testului Chi-pătrat?

Observăm că obținem un chi-pătrat de 9,2655 și o valoare a lui *p* aproape de nivelul de semnificație de 0,05. În aceste condiții respingem ipoteza nulă și concluzionăm că între cele două variabile (*trstplt* și *happy*) există o relație.

Al doilea pas este realizarea unui tabel de contingență care să ne arate dacă există o asociere semnificativă între cele două variabile.

```
### Încărcăm din nou pachetul dplyr###
library(dplyr)
### Creăm un tabel de contingență între trstplt și happy###
dataset_exemplu %>%
  group_by(trstplt, happy) %>%
  tally() %>%
  spread(trstplt, n)

## # A tibble: 3 × 4
##   happy     incredere `incredere medie` neincredere
##   <chr>         <int>             <int>       <int>
## 1 fericit         105               312         623
## 2 nefericit         2                19          44
## 3 neutru           19                97         213
```

O modalitate avansată de a realiza un tabel de contingență este prin utilizarea pachetului **sjPlot** (Lüdecke 2022). Această modalitate oferă și rezultatele testului Cramer's V care măsoară mărimea efectului testului chi-pătrat. Altfel spus, măsoară cât de puternic sunt asociate două variabile calitative categoriale.

```
### Încărcăm pachetului sjPlot###
library(sjPlot)
### Creăm un tabel de contingență folosind funcția tab_xtab a pachetului s
jPlot###
sjPlot::tab_xtab(var.row = dataset_exemplu$happy, var.col = dataset_exempl
u$trstplt, title = "Table Title", show.row.prc = TRUE)
```

Tabel 3.8 Tabel de contingență pentru variabilele happy și trstplt

happy	trstplt			
	Neîncredere	Încredere medie	Încredere	**Total**
Fericit	623	312	105	1040
	59.9%	30%	10.1%	100%
Nefericit	44	19	2	65
	67.7%	29.2%	3.1%	100%
Neutru	213	97	19	329
	64.7%	29.5%	5.8%	100%
Total	880	428	126	1434
	61.4%	29.8%	8.8%	100%

$X^2 = 9,265$ df = 4 Cramer's V = 0.057 p = 0.055

Astfel, din cele două modalități se poate observa faptul că nivelul de încredere în politicieni este mai mare pentru respondenții cu un nivel raportat de fericire mare. Respondenții nefericiți tind să aibă mai puțină încredere în politicieni. Rezultatele indică faptul că relația este semnificativă statistic la p<0,005.

Asocierea dintre două variabile nominale se poate realiza și prin aplicarea testului Goodman Kruskal Lambda. În RStudio vom aplica următoarele comenzi pentru a aplica testul menționat:

```
### Creăm un tabel de contingență între trstplt și happy###
tab <- table(dataset_exemplu$trstplt, dataset_exemplu$happy)
tab

### Încărcăm pachetul DescTools ###
library(DescTools)

### Aplicăm testul de asociere specific variabielor nominal, Goodman Krusk
al Lambda###
Lambda(tab)
Lambda(tab, conf.level=0.95)

Lambda(tab, direction="row")
Lambda(tab, direction="column")

> Lambda(tab)
[1] 0.01961679
> Lambda(tab, conf.level=0.95)
      lambda       lwr.ci       upr.ci
```

```
0.0196167883 0.0001194114 0.0391141653
>
> Lambda(tab, direction="row")
[1] 0.02451839
> Lambda(tab, direction="column")
[1] 0.01428571
```

Coeficientul Lambda variază de la 0 la 1. Un rezultat de 0 nu reflectă nicio asociere între variabile, iar un rezultat de 1 indică o asociere perfectă între variabilele de interes. În cazul nostru, 0.014 indică o asociere redusă între cele două variabile.

În această secțiune am exemplificat noțiuni precum tabelele de contingență (încrucișate), măsuri de asociere și testul Chi pătrat. Toate aceste instrumente evidențiază informații importante despre variabilele noastre. Cu toate acestea, este important de reținut faptul că modalitatea de recodificare a variabilelor are o mare influență asupra rezultatelor. Astfel, este important să verificăm modalitatea de recodificare înainte de a evalua relații dintre variabile. În următoarele secțiuni vom discuta despre relația de asociere dintre două variabile ordinale, dar și despre relația de corelație dintre două variabile cantitative.

Atunci când analizăm relația dintre două **variabile ordinale** tabelul de asociere ne ajută să identificăm nu doar existența relației, ci și direcția acesteia. O concentrare mare a frecvenței cazurilor, precizate în fiecare celulă a tabelului, pe una dintre diagonalele tabelului de contingență înseamnă existența unei asocieri între cele două variabile. Dacă concentrarea este pe diagonala "/" atunci asocierea este pozitivă, iar dacă concentrarea este pe diagonala "\" atunci asocierea este negativă. Desigur, această interpretare trebuie să țină cont de modul în care sunt codificate cele două variabile: care sunt valorile ridicate și care sunt valorile scăzute. Atunci când avem variabile ordinale vom folosi caracteristica lor de ordonare a valorilor și vom putea calcula ranguri (indivizii sunt raportați unii la alții din perspectiva valorilor ridicate sau scăzute pe care le iau pe variabila ordonată) ale indivizilor în funcție de valorile luate pentru fiecare variabilă. De exemplu, evaluarea guvernului și partidelor de către respondenții unui sondaj duce la evaluări diferite, și anume la ranguri pe care le primesc cele două instituții în funcție de cele două evaluări. Astfel, unii respondenți vor acorda evaluări bune (ridicate) guvernului dar slabe partidelor, iar alții vor acorda

evaluări bune sau slabe atât guvernului, cât și partidelor. Primii vor avea ranguri diferite, în vreme ce ultimii vor avea ranguri identice. Putem concluziona, așadar, că vom avea perechi concordante atunci când individul are un rang ridicat pe ambele variabile, respectiv perechi discordante când un individ are un rang ridicat pe o variabilă și scăzut pe cealaltă variabilă.

Pe baza acestei caracteristici importante vom constata că relațiile dintre variabilele ordinale au nu doar intensitate (de la 0 la 1), ci și direcție (de la -1 la +1) pozitivă sau negativă. Prin urmare, coeficienții calculați pot lua valori între -1 și +1. Printre cei mai importanți coeficienți prin care putem măsura asocierea dintre două variabile ordinale sunt:

- **Gama**: este un coeficient simetric și reprezintă proporția cu care este redusă eroarea de predicție a rangurilor (nu valorilor) variabilei dependente, cunoscând variabila independentă. Dacă indicatorul ia valoarea 0,25 putem concluziona că identificând rangurile pe care individul le ia pe variabila independentă reducem erorile de predicție a variabilei dependente cu 25%. Formula de calcul a coeficientului gamma este:

$$\gamma = \frac{N_c - N_d}{N_c - N_d} \tag{19}$$

unde, N_c este numărul total de perechi concordante (ranguri identice), iar N_d este numărul total de perechi discordante (ranguri diferite).

- **Kendall tau-b** si **Kendall tau-c**: Kendall's Tau este o măsură non-parametrică a relațiilor dintre coloanele de date ordonate. Coeficientul de corelație Tau returnează o valoare între -1 și +1, unde -1 indică o asociere negativă, 0 indică lipsa asocierii, iar +1 indică o asociere perfectă între cele două variabile. Coeficientul tau-b este calculat pentru variabilele ordinale ce au același număr valori, în vreme ce coeficientul tau-c este raportat pentru variabilele ordinale ce au un număr diferit de categorii.

- **Somer's d**: are o interpretare similară coeficientul Gama, dar este un coeficient asimetric. Poate lua valori între -1 și +1.

În concluzie, avem o asociere pozitivă între variabile ordinale 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 ranguri mici și pentru variabila Y. Avem o asociere negativă atunci când indivizii cu rang mare pentru variabila X tind să aibă ranguri mici pentru variabila Y și invers. Dacă ia valoarea 0 atunci cele două variabile sunt independente. Cu cât relația dintre două variabile ordinale va fi mai puternică, cu atât măsura asocierii (coeficientul) va fi mai aproape de 1 (cu + sau -).

În continuarea vom testa asocierea dintre două variabile de tip ordinal și anume satisfacția cu economia (*stfeco)* și satisfacția cu locul de muncă (*stfmjob*) din setul de date ESS. Inițial vom crea un tabel de contingență pentru cele două variabile de interes, iar ulterior vom folosi tabelul creat în comanda de calcul a testelor de asociere pentru variabilele ordinale. În RStudio, testarea asocierii se realizează prin aplicarea următoarelor comenzi:

```
### Creăm un tabel de contingență între stfeco și stfmjob###
table <- table(dataset_exemplu_final$stfeco, dataset_exemplu_final$stfmjob
)
table

### Încărcăm pachetul DescTools ###
library(DescTools)

### Aplicăm testul de asociere specific variabielor ordinale, Gama###
GoodmanKruskalGamma(table, conf.level=0.95)

> GoodmanKruskalGamma(table, conf.level=0.95)
     gamma     lwr.ci     upr.ci
0.07555492 0.02815254 0.12295730
```

Reamintim că rezultatele pentru coeficientul Gama pot lua valori între -1 și 1, unde 1 înseamnă o asociere pozitivă perfectă (dacă satisfacția cu economia crește, crește și satisfacția cu locul de muncă), iar -1 evidențiază o relație negativă perfectă (dacă satisfacția cu economia crește, satisfacția cu locul de muncă scade). 0 indică faptul că nu există o asociere între cele două variabile. Cu cât ne apropiem mai mult de 1 (sau -1), cu atât relația este mai puternică. În cazul nostru, am obținut un coeficient Gama în valoare de 0.075 ceea ce indică o relație de asociere pozitivă, dar nu perfectă.

Asocierea dintre două variabile ordinale mai poate fi testată și prin aplicarea testelor Kendall tau-b sau Kendall tau-c. În RStudio, testul Kendall tau-b se aplică similar testului Gama. Astfel, vom aplica următoarele comenzi.

```
### Creăm un tabel de contingență între stfeco și stfmjob###
table <- table(dataset_exemplu_final$stfeco, dataset_exemplu_final$stfmjob
)
table

### Încărcăm pachetul DescTools ###
library(DescTools)

### Aplicăm testul de asociere specific variabielor ordinale, Kendall tau-
b###
KendallTauB(table, conf.level=0.95)

> KendallTauB(table, conf.level=0.95)
     tau_b     lwr.ci     upr.ci
0.06474785 0.02406914 0.10542657
```

În cazul nostru, rezultatul de 0.064 indică o asociere pozitivă, dar slabă. Similar testelor Gama și Kendall (tau-b sau tau-c) poate fi folosit și testul Somers' d. În RStudio, vom folosi următoarele comenzi pentru aplicarea acestui test:

```
### Creăm un tabel de contingență între stfeco și stfmjob###
table <- table(dataset_exemplu_final$stfeco, dataset_exemplu_final$stfmjob
)
table

### Încărcăm pachetul DescTools ###
library(DescTools)

### Aplicăm testul de asociere specific variabielor ordinale, Somers'd###
SomersDelta(table, direction="column", conf.level=0.95)
# Somers' D R|C
SomersDelta(table, direction="row", conf.level=0.95)

> SomersDelta(table, direction="column", conf.level=0.95)
    somers     lwr.ci     upr.ci
0.06289434 0.02335701 0.10243168

> # Somers' D R|C
> SomersDelta(table, direction="row", conf.level=0.95)
    somers     lwr.ci     upr.ci
```

```
0.06665599 0.02464148 0.10867049
```

Ca mod de interpretare, similar testului Kendall (tau-b sau tau-c) și a testului Gama, coeficientul delta al lui Somers ia valori între -1 și 1. În cazul nostru, valoarea de 0.066 a testului Somers' d indică o relație pozitivă, deși slabă, între cele două variabile de interes. Putem concluziona că valorile ridicate pe care indivizii le iau pentru variabila satisfacția cu economia, ne ajută să reducem cu 6% erorile de predicție a valorilor pe care aceștia le iau pentru variabila satisfacția cu locul de muncă. Observăm că aceste teste tind să producă rezultate convergente.

3.3.2. Covarianța și corelația

Pentru a observa modul în care variabilele cantitative se află în relație una cu cealaltă vom discuta despre covarianță și corelație. Ne amintim despre problemele discutate la secțiunea 3.2.3 în care observam modul în care variabilele cantitative tind să fie mai omogene sau mai eterogene, putând, cu ajutorul reprezentărilor grafice și al unor indicatori statistici, să identificăm variația fiecărei variabile raportată la un indicator al tendinței centrale, precum media. Această variație a fiecărei caracteristici din distribuția datelor noastre poate fi observată și măsurată pentru mai multe asemenea variabile (caracteristici ale unității noastre de observație). Atunci când observăm că două variabile tind să varieze în același mod, sumarizăm această relație prin termenul de covarianță (Frankfort-Nachmias, Nachmias, și DeWaard 2015, 85, 334). Aceasta descrie cât de variată e distribuția valorilor variabilei față de media acesteia și care este relația dintre valorile celor două variabile.

$$cov_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

(20)

Unde: x şi y sunt variabilele a căror variație o analizăm, x_i este valoarea luată de fiecare individ pe variabila x, $\bar{x}$ este media variabilei x, y_i este valoarea luată de

fiecare individ pe variabila y, $\bar{y}$ este media variabilei y, iar n este numărul de indivizi statistici (totalul cazurilor din baza de date).

Prin urmare, prin covarianță putem observa în ce măsură valorile mari pe care indivizii le iau pe o variabilă (de exemplu, variabila x) se află în relație cu valorile mari sau mici pe care aceștia le iau pe cealaltă variabilă (de exemplu, variabila y). Unitățile de observație care iau valori mari pe o variabilă, iau în medie valori mici sau mari pe cealaltă variabilă. Covarianța este însă un produs a două unități de măsură care pot fi diferite: cea a variabilei x și cea a variabilei y. Ea se modifică la orice schimbare cu o constantă a valorilor variabilelor, prin urmare fiind greu de utilizat în comparații. Teoretic, valorile pe care covarianța le poate lua pot varia de la $-\infty$ la $+\infty$. Standardizarea valorilor variabilelor din analiză prin împărțirea la abaterea standard (așa cum observăm în formula *(22)* de mai jos) transformă scala de măsurare în intervalul -1 și +1, conducându-ne la o măsură specifică a covarianței, pe care o numim corelație. Aceasta nu are unitate de măsură și, prin urmare, poate fi folosită mai facil în comparații (Agresti și Finlay 2014; Rotariu et al. 1999).

Corelația este o măsură a intensității asocierii dintre două variabile cantitative, fără a diferenția între variabila cauză și variabila efect (Agresti și Finlay 2014). Coeficienţii prin care măsurăm gradul de corelaţie dintre aceste variabile ne vor ajuta:

- să identificăm dacă există relaţie între variabilele cantitative (de exemplu: în ce măsură putem anticipa rata şomajului dintr-o ţară cunoscând valoarea PIB/cap de locuitor);
- să identificăm care este intensitatea relaţiei (cât de bine putem prezice valorile variabilei dependente atunci când cunoaştem valorile variabilei independente);
- să identificăm care este forma relaţiei între cele două variabile (dacă avem relaţii liniare, valorile unei variabile tind să se modifice cu o constantă atunci când valorile celeilalte variabile se modifică cu o constantă);
- să identificăm care este forma relației dintre cele două variabile, mai precis în ce măsură putem reduce norul de puncte din diagrama cu

puncte la o linie dreaptă care să îl aproximeze și prin care să ilustrăm relația dintre cele două variabile.

Linia dreaptă aproximează relații liniare dintre cele două variabile, care semnifică faptul că valorile unei variabile tind să se modifice cu o constantă, atunci când valorile celeilalte variabile se modifică cu o constantă.

$$r = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov_{x,y}}{n\sigma_x\sigma_y}$$

(21)

unde r este coeficientul de corelație, x şi y sunt variabilele a căror variație o analizăm, x_i este valoarea luată de fiecare individ pe variabila x, $\bar{x}$ este media variabilei x, y_i este valoarea luată de fiecare individ pe variabila y, $\bar{y}$ este media variabilei y, iar n este numărul de indivizi statistici.

Analiza de corelație pe un eșantion presupune câteva asumpții ale datelor analizate: (1) variabilele sunt cantitative, (2) sunt distribuite normal (verificarea poate fi efectuată prin testul Shapiro-Wilk de interpretare a ipotezei nule cu privire la asumpția că eșantionul provine dintr-o populație cu distribuție normală), (3) relația dintre cele două variabile analizate este liniară, (4) nu există sau sunt foarte puține cazurile deviante *outliers* (verificarea o facem prin intermediul coeficientului de corelație Spearman rho care indică prezența acestor cazuri deviante atunci când produce rezultate puternic diferite de rezultatele coeficientului de corelație Pearson r), (5) homoscedasticitatea (cuvânt compus preluat din limba greacă – *skedastikós*, însemnând aceeași dispersie sau aceeași varianță) se referă la distribuției datelor sau omogenitatea varianței (varianța datelor este aceeași indiferent de punctul pe care ne situăm pe linia de aproximație a relației dintre cele două variabile); opusul ei fiind heteroscedasticitatea (dispersie diferită pe valori diferite ale celor două variabile).

Principalii coeficienți pe care îi folosim pentru estimarea intensității corelației sunt Pearson r, coeficientul de determinare r pătrat și Spearman rho.

- **Pearson r**, numit și coeficientul de corelație al produsului momentelor de ordinul întâi (al abaterilor de la medie) (Rotariu et al. 1999, 172) măsoară cât de dispersat sau de concentrat este norul de puncte

(diagrama cu puncte) față de o linie dreaptă și ia valori între -1 și +1, unde -1 indică o relație negativă perfectă, +1 o relație pozitivă perfectă, iar 0 indică lipsa relației între cele două variabile. Acest coeficient nu ne arată o direcție a relației de cauzalitate dintre cele două variabile. Aceasta trebuie stabilită de cercetător pe baza teoriei și a analizei de regresie, despre care vom discuta în secțiunea următoare.

- **Coeficientul de determinare r^2** reprezintă pătratul coeficientului r, fiind interpretat ca procent din variația variabilei dependente explicat de variabila independentă. Acesta ia valori între 0 și 1.
- **Spearman rho**, numit și coeficient de corelație a rangurilor, este un test non-parametric, prin urmare este utilizat pentru a măsura relația dintre două variabile ordinale (Rotariu et al. 1999, 175). Așa cum spuneam mai sus, este utilizat și pentru identificarea cazurilor deviante în relația dintre două variabile cantitative. Ia valori între -1 și +1, unde -1 indică o relație negativă perfectă, +1 o relație pozitivă perfectă, iar 0 indică lipsa relației între cele două variabile.

Analiza de corelație nu ne arată direcția cauzalității, înclinația liniei drepte care estimează (aproximează) dispersia norului de puncte, nici dacă alte variabile independente influențează variabila dependentă (prin urmare, covariația va fi atribuită unei singure variabile independente, ceea ce este nesatisfăcător din punct de vedere teoretic, deoarece în studierea problemelor sociale niciodată un fenomen pe care dorim să îl explicăm nu are o singură cauză).

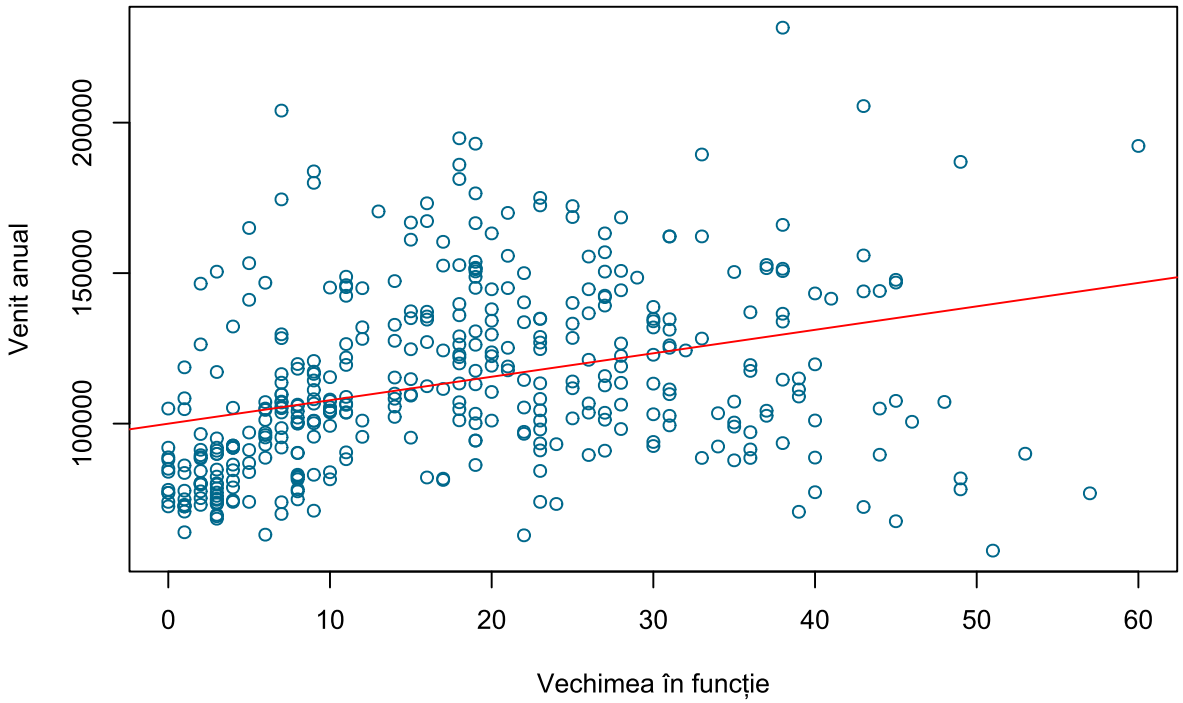
În acest context, pentru a evalua relația dintre două variabile continue (de interval sau de rapoarte) folosim tehnici statistice diferite față de tehnicile folosite pentru a evalua asocierea dintre două variabile categorice. Să presupunem că vrem să aflăm corelație dintre venitul anual și vechimea în muncă. Pentru a testa această corelație vom utiliza setul de date **Salaries** inclus în pachetul **carData**. Setul de date conține informații referitoare la genul, salariul, vechimea în muncă, disciplina predată, perioada de la finalizarea doctoratului. Inițial, putem produce o diagramă cu puncte pentru a observa vizual dacă pare să existe sau nu o relație liniară între

vechimea în muncă și venitul anual. În RStudio realizăm acest lucru prin următoarea linie de cod:

```
###Atașăm setul de date Salaries ###
attach(Salaries)

###Generăm un grafic de dispersie (diagramă cu puncte) pentru "venitul anu
al" după "vechimea în funcție la locul de muncă"###
plot(Salaries$yrs.service,Salaries$salary, col = 'deepskyblue4',
xlab="Vechimea în funcție",
ylab="Venitul anual")
```

Figura 3.13 Diagramă cu puncte pentru corelație



Graficele de dispersie sunt similare tabelelor de contingență prin faptul că afișează rezultate comune pentru ambele variabile, dar arată diferit datorită naturii datelor. Pe baza graficului de mai sus putem observa că există o corelație între vechimea în muncă și venitul anual în sensul în care pe măsură ce crește vechimea, crește și venitul.

Măsura corelației dintre cele două variabile poate fi aflată și prin calcularea coeficientului de corelație Pearson r. Pentru a estima această valoare vom folosi comanda **pwcorr()**, și vom adăuga comanda **sig()** din cadrul pachetului **DAMisc**

(Armstrong 2022) astfel încât rezultatul să includă nivelul de semnificație statistică (p) pentru relație. Vom aplica și greutatea eșantionului cu această comandă. În RStudio vom aplica următoarele linii de cod:

```
### Încărcăm pachetului DAMisc###
library(DAMisc)
### Calculăm coeficientul de corelație Pearson r###
pwCorrMat(~salary + yrs.service, data=Salaries)
## Pairwise Correlations
##         salary service
## salary
## service 0.335*
```

În codul de mai sus, identificăm variabilele care trebuie corelate specificându-le în partea dreaptă a unei formule separate prin plusuri (+). În interpretarea rezultatelor vom începe cu coeficientul de corelație. Rezultatele indică o relație pozitivă între vechimea în muncă și venit (0.335): la creșterea vechimii în muncă, crește și venitul.

În continuare, vom calcula coeficienții de corelație în RStudio:

```
### Calculăm coeficientul de corelație Pearson###
pearson <- cor.test(Salaries$salary, Salaries$ yrs.service, method="pearso
n")
pearson

##  Pearson's product-moment correlation
##data:  Salaries$salary and Salaries$yrs.service
##t = 7.0602, df = 395, p-value = 7.529e-12
##alternative hypothesis: true correlation is not equal to 0
##95 percent confidence interval:
## 0.2443740 0.4193506
##sample estimates:
      cor
## 0.3347447
```

Având în vedere că valoarea nivelului de semnificație statistică $p \leq 0.001$ este mai mică decât nivelul de semnificație alfa = 0,05, putem concluziona că vechimea și venitul sunt corelate pozitiv, coeficientul de corelație r fiind 0.33. Pentru identificarea,

altfel decât grafică a existenței cazurilor deviante, putem folosi coeficientul de corelație a rangurilor, Spearman rho, conform exemplului de mai jos.

```
### Calculăm coeficientul de corelație Spearman###
spearman <- cor.test(Salaries$salary, Salaries$yrs.service,method="spearma
n")
spearman

##   Spearman's rank correlation rho
## data:  Salaries$salary and Salaries$yrs.service
## S = 5996863, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
      rho
## 0.4249486
```

3.4. Analiza de regresie liniară

Analiza de regresie ilustrează relația dintre două variabile cantitative, la fel ca și analiza de corelație despre care am discutat în secțiunea anterioară. Regresia ne ajută însă să înaintăm în procesul de explicare a corelației dintre două variabile și să stabilim sensul / direcția acestei relații. După cum ne amintim, corelația este simetrică: corelația lui x cu y este identică corelației lui y cu x; regresia, în schimb, nu este simetrică, astfel diferențiem din punct de vedere cauzal între variabila dependentă (aceasta este variabila pe care dorim să o explicăm) și cea independentă (de această variabilă ne vom folosi pentru a explica variația variabilei dependente). Norul de puncte ce reprezintă valorile luate de indivizi pe ambele variabile poate fi aproximat printr-o dreaptă (curbă) care aproximează forma și direcția norului de puncte. Aceasta este dreapta de regresie (curba de regresie). Această dreaptă este identificată prin ecuația:

$$y = a + bx \tag{22}$$

unde y reprezintă variabila dependentă; x este variabila independentă; b reprezintă panta de regresie (cu câte unități proprii se modifică în medie variabila dependentă atunci când variabila independentă se modifică cu o unitate proprie. Semnul + semnifică faptul că valori scăzute ale lui y corespund valorilor scăzute ale lui x. Semnul – semnifică faptul că relația este negativă, deci valorile scăzute ale lui y corespund valorilor ridicate ale lui x. b ia valoarea 0 atunci când nu putem aproxima norul de puncte printr-o dreapta), b este numit și **coeficient de regresie** (numit și pantă) și se calculează înmulțind coeficientul de corelație r cu raportul abaterilor standard ale celor două variabile din regresie; iar a reprezintă constanta (reprezintă valoarea pe care o are variabila dependentă atunci când variabila independentă ia valoarea zero). De exemplu, într-o regresie prin care încercăm să explicăm venitul profesorilor în funcție de vechimea în muncă, a reprezintă nivelul salarial de început, atunci când profesorul nu are nici o vechime.

Estimarea dreptei de regresie are la bază metoda liniară a celor mai mici pătrate ce produce o singură soluție optimă pentru ecuația de regresie precizată mai sus. Acestea minimizează erorile de estimare a distribuției cazurilor din norul de puncte produs de cele două variabile prin intermediul liniei de regresie care aproximează această distribuție, prin urmare reduce la minim valorile reziduale, adică diferențele dintre valorile observate și cele estimate de panta de regresie. Cu cât estimăm mai bine prin dreapta de regresie distribuția cazurilor din modelul estimat, cu atât valorile reziduale vor fi mai mici. Astfel, putem estima parametrii de regresie (coeficienții pantei de regresie și a valorii constante a) și putem prezice valorile variabilei dependente y.

Pentru analiza de regresie ne bazăm pe următoarele asumpții de configurare a datelor de analiză: (1) variabilele sunt cantitative, (2) ele sunt distribuite normal (verificarea poate fi efectuată prin testul Shapiro-Wilk de interpretare a ipotezei nule cu privire la asumpția că eșantionul provine dintr-o populație cu distribuție normală), (3) relația dintre ele este liniară (se observă grafic cu ajutorul crPlots(fit) ce ilustrează valorile reziduale, adică diferențele dintre valorile observate și cele estimate de panta de regresie), (4) nu prezintă cazuri deviante semnificative, (5) homoscedasticitatea distribuției datelor sau omogenitatea variantei (varianța erorilor trebuie să fie constantă pentru toată distribuția cazurilor pe cele două variabile, iar distribuția

cazurilor trebuie să fie similară, uniformă), (6) lipsa multicoliniarității, variabilele independente nu sunt corelate (dacă avem multicoliniaritate atunci variabilele independente sunt în corelație puternică una cu cealaltă, prin urmare, putem prezice într-un grad ridicat valorile unei variabile independente printr-o variabilă sau mai multe variabile independente introduse în modelul explicativ), (7) independența observațiilor statistice sau lipsa autocorelație (dacă există autocorelație observată prin testul Durbin Watson pe care îl vom exemplifica mai jos, înseamnă că am omis din modelul nostru explicativ una sau mai multe variabile independente importante).

Estimarea calității modelului de regresie se face cu ajutorul **coeficientului de determinare** notat R^2. Acesta se calculează ca raport între varianța neexplicată (pătratul valorilor reziduale pe care le măsurăm față de dreapta de regresie) și varianța totală (pătratul valorilor reziduale pe care le măsurăm față de media variabilei dependente y), pe care îl scădem din 1. Coeficientul de determinare R^2 reprezintă proporția din variația variabilei dependentă explicată de variația variabilei independente (Rotariu et al. 1999, 184). R^2 ia valori între 0 și 1, unde 0 ne arată că variabila independentă nu explică nimic din variația variabilei dependente, iar 1 semnifică faptul că toată variația variabilei dependente poate fi explicat de variația variabilei independente (toate cazurile sunt pe dreapta de regresie). Așadar, cu cât valorile lui R^2 sunt mai apropiate de 1, cu atât relația dintre variabila dependentă și cea independentă este mai intensă iar reprezentarea grafică a distribuției este mai apropiată de o dreaptă. Atunci când R^2 ia valoarea 1, coeficientul de regresie b ia una din valorile extreme (-1 sau +1). Dacă R^2 ia valoarea 0, atunci coeficientul de regresie b ia valoarea 0.

Analiza de regresie simplă liniară ne este utilă pentru a explica variația variabilei dependente prin intermediul unei singure variabile independente. Dacă analizăm efectul mai multor variabile independente asupra unei variabile dependente atunci regresia se numește multiplă sau multiliniară. Dat fiind că cele mai multe, dacă nu toate, fenomenele sociale pe care le explicăm au multiple cauze, putem afirma că regresia multiplă este cea pe care o vom folosi pentru a testa ipotezele ce formulează legătura dintre aceste multiple cauze și efectul de explicat. Analiza de regresie multiplă analizează relația dintre trei sau mai multe variabile, dintre care una este

variabila dependentă, iar celelalte sunt variabilele independente. Modelul de regresie multiliniară poate fi estimat prin următoarea ecuație de regresie multiplă:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + \varepsilon \quad (23)$$

unde b_i este panta de regresie pentru variabila x_i, si reprezintă numărul de unități proprii cu care se modifică în medie variabila y atunci când variabila independentă x_i, crește cu o unitate proprie iar celelalte variabile independente x_j, sunt constante, iar ε este partea rămasă neexplicată din variația variabilei y, ε este numit și termen de eroare.

Deoarece coeficienții b_i au unități de măsură care sunt produs al variabilei dependente y și al variabilei independente x_i, compararea valorilor absolute ale mai multor asemenea coeficienți b este dificilă (comparăm mere cu pere). Acești coeficienți sunt, prin urmare, nestandardizați. De aceea pentru comparația acestor coeficienți îi putem standardiza cu ajutorul abaterii standard a variabilelor dependentă y și independente x_i astfel încât să obținem coeficienți de regresie standardizați (β). Pe scurt, standardizarea variabilei se face prin calcularea diferenței fiecărei valori, măsurate pe variabilă, față de media variabilei, și împărțirea acesteia la abaterea standard a variabilei. Acești coeficienți de regresie standardizați permit ierarhizarea variabilelor independente în funcție de mărimea coeficientului standardizat, de importanța pe care o are fiecare în explicarea variației variabilei dependente printr-o relație directă. Interpretarea acestora este identică cu cea a coeficienților b, doar că nu mai discutăm despre unități proprii cu care se modifică variabila dependentă și cea independentă, ci de abateri standard. În cazul analizei de regresie multiplă, coeficientul r pătrat R^2 este numit coeficient de determinație multiplă și se interpretează la fel ca R^2 din analiza de regresie simplă liniară. În cazul regresiilor multiliniare, în general, adăugarea unor variabile independente suplimentare în model (de preferat pe baza ipotezelor fundamentate teoretic, nu prin ajustarea modelului explicativ de regresie prin extragerea sau adăugarea treptată a variabilelor predictor) conduce la creșterea valorii coeficientului de determinare multiplă R^2 de aceea, se calculează și este de preferat să raportăm coeficientul R^2 ajustat (Rotariu et al. 1999, 193; Agresti și Finlay 2014, 446–47).

În secțiunile următoare, folosind setul de date utilizat în capitolele anterioare, vom exemplifica utilizarea în practică a modelelor de regresie simplă liniară (metoda celor mai mici pătrate – în limba engleză Ordinary Least Squares - OLS), regresie multiliniară și regresie logistică.

3.4.1. Exemplu de analiză cu regresia simplă liniară (OLS - metoda celor mai mici pătrate)

În cazul modelului de regresie simplă o singură variabilă independentă prezice variabila de rezultat (dependentă). Plecând de la setul de date ESS, să presupunem că vrem să investigăm, conform ipotezei 2b, dacă variabila independentă *happy* (nivelul de fericire declarat de respondenți) influențează nivelul de încredere în partidele politice (*trstprt*). Variabila independentă, *happy*, măsoară nivelul de fericire al respondenților pe o scală de la 0 la 10. Variabila dependentă, *trstprt*, măsoară nivelul de încredere în partidele politice pe o scală de la 0 la 10.

În RStudio, modelul de regresie OLS se realizează prin utilizarea funcției **lm()**, iar rezultatele modelului se pot afișa cu ajutorul funcției **summary()** după cum urmează:

```
regresie_simpla <- lm(trstprt ~ happy, data = dataset_exemplu)

summary(regresie_simpla)

## Call:
## lm(formula = trstprt ~ happy, data = dataset_exemplu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##-3.2544 -2.0509 -0.0509  1.7456  7.2543
##
## Coefficients:
## (Intercept)  2.23707    0.26772   8.356  < 2e-16 ***
##happy        0.10173    0.03558   2.859  0.00432 **
```

```
##---
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

##Residual standard error: 2.314 on 1174 degrees of freedom
##Multiple R-squared:  0.006917,    Adjusted R-squared:  0.006071
##F-statistic: 8.177 on 1 and 1174 DF,  p-value: 0.004318
```

Rezultatele obținute ne arată o relație liniară pozitivă între cel două variabile. Coeficienții modelului sunt semnificativi statistic la un nivelul $p \leq 0.001$, și sugerează că o creștere a nivelului de fericire determină o creștere cu 0,10 unități a încrederii în partidele politice. Totuși, putem observa că modelul, fiind unul redus la extrem, explică doar 0,6% din variația variabilei dependente.

În continuare, vom extinde modelul, astfel încât să analizăm evoluția unei variabile Y în raport cu doi factori de influență. Plecând de la setul de date ESS, să presupunem că vrem să estimăm dacă interesul față de politică (*polintr*) și nivelul de fericire (*happy*) influențează încrederea în partidele politice (*trstprt*) conform ipotezelor 2b și 2c. De asemenea, intenționăm să evaluăm gradul în care nivelul de satisfacție cu situația economică (*stfeco*) și nivelul de fericire (*happy*) influențează nivelul de încredere (*trstprt*) conform ipotezelor 2b și 2d formulate.

Vom testa prima regresie formulată mai sus în RStudio folosind funcția **lm()**. Funcția **lm()** estimează valoarea așteptată a variabilei dependente ca o funcție a variabilei independente. Este important să scriem funcția în ordinea corectă. Variabila dependentă este specificată în partea stângă a formulei, iar variabila independentă în partea dreaptă.

În cazul nostru, conform ipotezelor 2b și 2d, variabila dependentă este nivelul de încredere în partidele politice (*trstprt*), iar cele două variabile independente sunt nivelul de fericire (*happy*) și interesul politic (*polintr*). Toate cele trei variabile incluse în model, atât dependenta, cât și cele două variabile independente, sunt măsurate pe o scală de la 0 la 10[41]. Primul lucru pe care trebuie să-l facem este să ne asigurăm că variabilele sunt codificate corect. Putem face acest lucru folosind funcția **class()**.

[41] Similar transformării variabilei *happy* pentru aplicarea primei regresii din această secțiune, trebuie să transformăm și variabila *stfeco*.

```
### Verificăm clasa variabilelor###
class(dataset_exemplu$polintr)
## [1] "numeric"
class(dataset_exemplu$happy2)
## [1] "numeric"
```

Vom aplica funcțiile **lm()** și **summary()** pentru a estima și vizualiza primul model de regresie:

```
regresie_1 <- lm(trstprt ~ polintr + happy, data = dataset_exemplu)
summary(regresie_1)
## Call:
## lm(formula = trstprt ~ polintr + happy, data = dataset_exemplu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##-4.1253 -1.9105 -0.2051  1.6146  7.6950
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept)  3.66657    0.33375  10.986  < 2e-16 ***
##polintr     -0.54018    0.07789  -6.936 6.67e-12 ***
##happy        0.09989    0.03488   2.864  0.00426 **
##---
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
##Residual standard error: 2.269 on 1173 degrees of freedom
##Multiple R-squared:  0.04604,     Adjusted R-squared:  0.04441
##F-statistic:  28.3 on 2 and 1173 DF,  p-value: 9.895e-13
```

Cum putem interpreta rezultatele regresiei bifactoriale de mai sus? În baza rezultatelor coeficienților de regresie de mai sus putem estima că nivelul de fericire și interesul exprimat față de politică au un efect semnificativ asupra nivelului de încredere în partidele politice. Creșterea cu un nivel a interesului față de politică determină o scădere cu 0,54 unități în nivelul de încredere în partidele politice atunci când comparăm cu categoria de referință. De asemenea, observăm că o creștere a

nivelului de fericire determină o creștere cu 0,10 unități a nivelului de încredere în partidele politice.

În continuare vom folosi analiza de regresie pentru a explica variația unei variabile dependente cu variabile independente de tip interval. Vom continua să folosim funcția **lm()** introdusă în secțiunea precedentă și vom testa a doua regresie prin testarea ipotezelor 2b și 2d. Astfel, suntem interesați să testăm dacă nivelul de satisfacție cu situația economică (*stfeco*) și nivelul de fericire (*happy2*) influențează nivelul de încredere în partidele politice (*trstprt*). În RStudio vom aplica funcția **lm()** și comanda **summary()**:

```
regresie_2 <- lm(trstprt ~ stfeco + happy, data = dataset_exemplu)
summary(regresie_2)

## Call:
## lm(formula = trstprt ~ stfeco + happy, data = dataset_exemplu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##-5.4538 -1.4278 -0.1693  1.2877  7.1154
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept)  1.31590    0.24610   5.347 1.07e-07 ***
##stfeco       0.45692    0.02695  16.956  < 2e-16 ***
##happy       -0.04313    0.03302  -1.306    0.192
##---
##Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
##Residual standard error: 2.075 on 1173 degrees of freedom
##Multiple R-squared:  0.2024,      Adjusted R-squared:  0.201
##F-statistic: 148.8 on 2 and 1173 DF,  p-value: < 2.2e-16
```

Conform rezultatelor analizei de regresie, în care putem observa nivelul de încredere $p \leq 0.001$, doar nivelul de satisfacție cu situația economică are o influență semnificativă statistic asupra nivelului de încredere în partidele politice. O creștere a nivelului de satisfacție cu situația economică determină o creștere de 0,45 unități a nivelului de încredere în partidele politice, în timp ce nivelul de fericire arată o relație pozitivă, dar nesemnificativă statistic.

3.4.2 Exemplu de analiză de regresie multiplă (multiliniară)

Înțelegerea sistemelor sociale complexe necesită analiza unor serii de potențiali factori care pot avea efect asupra unui fenomen. În exemplul nostru de regresie cu două variabile independente, am observat că un nivel de satisfacție ridicat crește nivelul de încredere în partidele politice, demonstrând o relație pozitiv semnificativă. Apare întrebarea dacă pot exista mai mulți factori care să aibă un efect asupra nivelului de încredere al oamenilor. Poate nu doar nivelul de satisfacție față de economie influențează nivelul de încredere, ci și nivelul de satisfacție cu guvernul, interesul față de politică, nivelul de atașament față de un partid politic și altele. Prin urmare, folosind regresia multiplă putem obține estimări mai bune ale efectului cumulat al variabilelor independente asupra rezultatului de interes. Analiza de regresie multiplă este concepută pentru a clarifica efectul explicativ a două sau mai multe variabile independente întrucât aceasta va estima efectul fiecărei variabile independente asupra variabilei dependente, controlând pentru efectele tuturor celorlalte variabile independente din model. Pentru aceeași variabilă dependentă (încrederea în partidele politice) vom testa efectul mai multor variabile independente după cum urmează:

```
regresie_multivariata <- lm(trstprt~ happy2 + polintr + clsprty + stfeco +
                                  stfgov + stfdem + stfmjob  + agea +
eisced + gndr, data = dataset_exemplu)
summary(regresie_multivariata)

## Call:
## lm(formula = trstprt ~ happy + polintr + clsprty + stfeco +
##  stfgov + stfdem + stfmjob + agea + eisced + gndr, data = dataset_exemp
lu_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##-6.4143 -1.1850 -0.1078  1.1659  7.3167
##
##Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
##(Intercept)  2.305242   0.479549   4.807 1.73e-06 ***
##happy       -0.061230   0.031216  -1.962  0.05006 .
##polintr     -0.383285   0.068144  -5.625 2.33e-08 ***
##clsprty      0.223529   0.114397   1.954  0.05094 .
##stfeco       0.089539   0.031964   2.801  0.00518 **
##stfgov       0.244335   0.029198   8.368  < 2e-16 ***
##stfdem       0.276065   0.031819   8.676  < 2e-16 ***
##stfmjob      0.043365   0.028617   1.515  0.12996
##agea        -0.005997   0.004730  -1.268  0.20506
##eisced      -0.197433   0.082995  -2.379  0.01753 *
##gndrM       -0.250260   0.108947  -2.297  0.02179 *
##---
##Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##Residual standard error: 1.824 on 1165 degrees of freedom
##Multiple R-squared:  0.3877,      Adjusted R-squared:  0.3825
##F-statistic: 73.77 on 10 and 1165 DF,  p-value: < 2.2e-16
```

Rezultatele regresiei multiple susțin ipoteza inițială că a fi mulțumit de situația economică crește nivelul de încredere în partidele politice, dar subliniază faptul că nivelul de satisfacție cu activitatea guvernului și cu democrația în general are un efect semnificativ mai mare asupra nivelului de încredere. O creștere în nivelul de satisfacție cu activitatea guvernului (*stfgov*) determină o creștere de 0,24 unități a nivelului de încredere, în timp ce satisfacția cu democrația (*stfdem*) determină o creștere de 0,27 unități a nivelului de încredere în partidele politice. Putem observa că o relație negativă este observată între nivelul de educație (*eisced*) și încredere: cu cât respondenții au un nivel de educație mai mare, cu atât scade încrederea acestora în partidele politice, semn că aceștia sunt mai greu de convins. De asemenea, bărbații (*gndrM*) tind să aibă mai puțină încredere în partidele politice decât categoria de referință. Astfel, luând toate variabilele independente (factorii explicativi) în considerare, putem explica aproximativ 38% (valoarea R^2 ajustat) din variația nivelului de încredere în partide.

Rezultatul regresiei multiple obținut în RStudio poate fi transpus în Microsoft Word astfel încât acesta să poată fi mai ușor lecturat.

Tabel 3.9 Model regresie multiplă

	Încredere în partide
Nivelul de fericire (happy2)	-0.061
	(0.031)
Interesul pentru politică (polintr1)	**-0.383 *** **
	(0.068)
Atașamentul față de un partid (clsprty)	0.223
	(0.114)
Satisfacția cu situația economică (stfeco)	**0.089****
	(0.031)
Satisfacție cu guvernul (stfgov)	**0.244*****
	(0.029)
Satisfacția cu democrația (stfdem)	**0.276*****
	(0.031)
Satisfacția cu locul de muncă (stfjob)	0.043
	(0.028)
Vârsta (agea)	-0.005
	(0.004)
Nivelul de educație (eisced)	**-0.197***
	(0.08)
Genul (gndrM)	**-0.250***
	(0.108)
Valoarea R^2 ajustat	0.382
Număr de observații (N)	1,434

Notă: Nivelul de semnificație statistică este *p<0.1; **p<0.05; ***p<0.001

3.4.2.1. Testarea ipotezelor referitoare la erorile din model

Regresia nu poate oferi rezultate valide dacă setul de date nu îndeplinește asumpțiile pe care le-am enumerat mai sus. Reamintim condițiile care trebuie să fie îndeplinite de datele care intră în modelul de regresie și pe care le vom testa RStudio fie prin generarea unor grafice, fie prin aplicarea unor teste statistice

1. Relație liniară (verificăm grafic).
2. Independența observațiilor statistice sau autocorelația (test Durbin Watson din pachetul car).

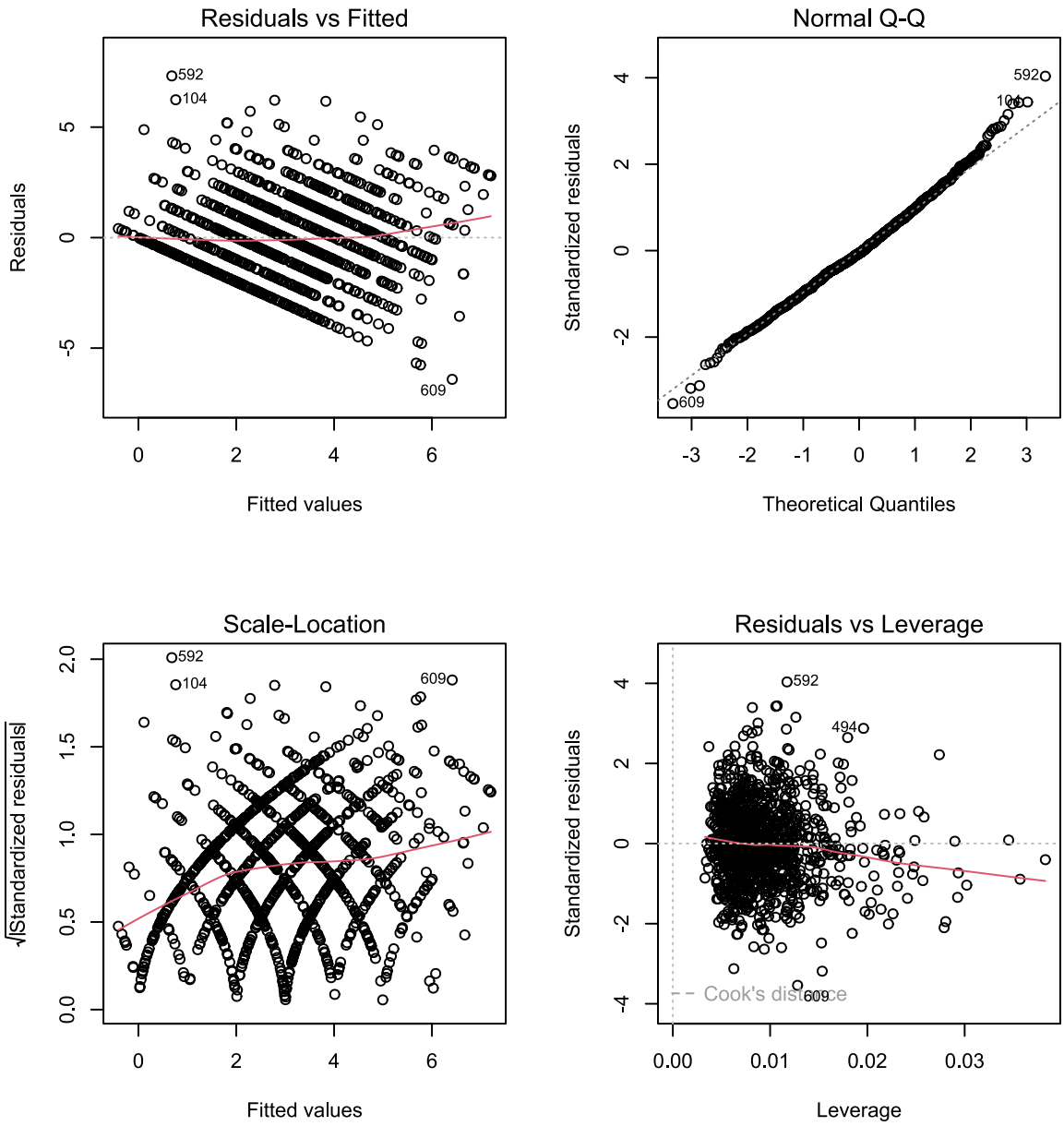
3. Homoscedasticitatea. Reziduurile au varianță constantă la fiecare nivel al lui x (verificăm grafic).
4. Distribuție normală (verificăm grafic).
5. Nu există multicoliniaritate (test VIF – factor de inflație a variației - cu cât sunt mai mari cu atât este mai mare coliniaritatea).
6. Nu sunt outlieri semnificativi (verificăm grafic).

```
### Toate graficele în același timp###
### definește model###
regresie_multivariata <- lm(trstprt~ happy + polintr + clsprty + stfeco +
                                     stfgov + stfdem + stfmjob  + agea +
                                     eisced + gndr, data = dataset_exemplu)

#### Modificăm aspectul la 2 x 2 (pentru a vizualiza simultan toate cele 4
grafice)###
par(mfrow = c(2, 2))

### Utilizarea funcției plot() pentru a crea graficele de testare###
plot(regresie_multivariata)
```

Figura 3.14 Reprezentarea grafică a condițiilor datelor în regresia de tip OLS



Conform celor 4 grafice de mai sus putem observa că modelul ales respectă condițiile impuse de modelul de regresie.În continuare, vom aplica o serie de teste pentru a verifica dacă datele noastre respectă condițiile impuse de utilizarea regresiei de tip OLS.

3.4.2.2. Normalitatea distribuției erorilor

Unul dintre testele folosite pentru identificarea normalității erorilor este testul Jarque-Bera. Acest test folosește momentele necentrate ale reziduurilor pentru a estima conformitatea distribuția erorilor. Testul Jarque-Bera nu se află în pachetul standard oferit de RStudio, dar se găsește în pachetul **tseries** (Trapletti și Hornik 2022)**:**

```
library(tseries)
```

În RStudio, aplicarea testului Jarque-Bera se realizează prin aplicarea funcției **jarque.bera.test()**. Observăm mai jos rezultatele obținute:

```
jarque.bera.test(summary(regresie_multivariata)$residuals)
##
##  Jarque Bera Test
##
## data:  summary(regresie_multivariata)$residuals
## X-squared = 22.19, df = 2, p-value < 1.519e-05
```

Interpretarea este: dacă se respinge ipoteza nulă, conform căreia erorile sunt normal distribuite, probabilitatea de a greși este mică decât 0,001 ($p \leq 0.001$). În consecință, ipoteza nulă este respinsă.

Alte teste de normalitate a distribuției erorilor sunt testul Pearson și testul Shapiro-Francia care pot fi aplicate în RStudio prin încărcarea pachetului **nortest** (Gross și Ligges 2015) după cum urmează:

```
library(nortest)

# Testul c2 - Pearson
pearson.test(summary(regresie_multivariata)$residuals)
```

```
##
##  Pearson chi-square normality test
##
## data:  summary(regresie_multivariata)$residuals
## P = 38.874, p-value < 0.2888
### Testul Shapiro-Francia###
sf.test(summary(regresie_multivariata)$residuals)
##
##  Shapiro-Francia normality test
##
## data:  summary(regresie_multivariata)$residuals
## W = 0.99482, p-value = 0.000639
```

Toate testele indică normalitatea distribuției erorilor la pragul standard de semnificație $p \leq 0.05$.

3.4.2.3. Autocorelarea erorilor

Autocorelarea erorilor presupune existența unei covarianțe nenule între erorile din ecuația de regresie. Autocorelarea erorilor apare în special în modelele construite pentru seriile de timp. Testul Durbin – Watson (Durbin și Watson, 1950, 1951) este cea mai cunoscută procedură utilizată pentru identificarea autocorelării de ordinul întâi a erorilor din modelele de regresie liniară.

Testul Durbin-Watson se găsește în pachetul **lmtest** (Zeileis și Hothorn 2022) și se folosește apelând funcția **dwtest():**

```
library(lmtest)
dwtest(regresie_multivariata)
##
##  Durbin-Watson test
##
## data:  regresie_multivariata
## DW = 1.8899, p-value = 0.02852
## alternative hypothesis: true autocorrelation is greater than 0
```

Pe baza rezultatului testului Durbin-Watson putem afirma că nu există autocorelație între valorile reziduale. În general, dacă Durbin-Watson este mai mic de 1.5 sau mai mare de 2.5, atunci există autocorelația datelor. În caz contrar, dacă Durbin-Watson este între 1.5 și 2.5, atunci autocorelația nu este probabil un motiv de îngrijorare.

3.4.2.4. Multicoliniaritatea

Una dintre asumpțiile de fundamentare a modelului de regresie liniară multiplă afirmă faptul că nu există nici o relație liniară între două sau mai multe variabile explicative (avem deci absența coliniarității). Pentru a testa existența multicoliniarității vom aplica mai multe teste. Prima modalitate de a testa multicoliniaritate în RStudio este prin aplicarea testului VIF, dar și prin a ne uita la valoarea de toleranță a datelor. Toleranța măsoară procentul de variație a variabilei independente care nu poate fi explicat de celelalte variabile. Prin urmare, dacă toleranța este scăzută, atunci celelalte variabile independente sunt capabile să explice o creștere sau o scădere a valorii variabilei independente. Dacă rezultă că variabilele sunt corelate atunci putem afirma că există multicoliniaritate în datele noastre și ar trebuie să revizuim modalitatea de transformare a variabilelor , prin agregarea lor, sau chiar să renunțăm la unele dintre variabilele independente.

În RStudio, putem utiliza funcția **ols_vif_tol()** din pachetul **olsrr** (Hebbali 2020) pentru a calcula valorile toleranței și ale factorului de inflație al variației. Funcția **ols_vif_tol()** returnează un tabel care conține numele variabilei, gradul de toleranță și VIF. De asemenea, VIF mai poate fi calculat și prin utilizarea pachetului **car** (J. Fox și Weisberg 2019). Vom exemplifica mai jos ambele variante.

```
library(car)
library(olsrr)

### testăm multicoliniaritatea 1 prin calcularea valorii VIF###
vif(regresie_multivariata)
```

```
##happy  polintr  clsprty   stfeco   stfgov   stfdem  stfmjob     agea   e
##isced     gndr
##1.239142 1.184619 1.145740 1.950959 2.116214 2.178232 1.180445 1.070362
##1.111655 1.044655
```

Un VIF egal cu 1 indică absența multicoliniarității întrucât coeficientul de regresie nu este mărit de prezența celorlalți predictori.

```
### testăm multicoliniaritatea 2 prin calcularea gradului de toleranță și
a valorii VIF ###
ols_vif_tol(regresie_multivariata)
##   Variables Tolerance      VIF
##1      happy 0.8070102 1.239142
##2    polintr 0.8441532 1.184619
##3    clsprty 0.8727986 1.145740
##4     stfeco 0.5125685 1.950959
##5     stfgov 0.4725419 2.116214
##6     stfdem 0.4590879 2.178232
##7    stfmjob 0.8471383 1.180445
##8       agea 0.9342631 1.070362
##9     eisced 0.8995593 1.111655
##10     gndrM 0.9572543 1.044655
```

Așa cum poate fi observat din tabel, toate variabilele au o toleranță mai mare de 0.1 și o valoare VIF mai mică de 2. Ca o măsură generală, o toleranță <0,1 ar putea indica multicoliniaritate, iar un VIF care depășește 5 necesită investigații suplimentare, în timp ce VIF de peste 10 indică în mod clar multicoliniaritate. În mod ideal, factorii de inflație ale variației sunt sub 3.

3.5. Analiza de regresie logistică

În capitolele anterioare, în care a fost prezentată regresia liniară, obiectivul principal era acela de a modela o variabilă dependentă (y) continuă (cantitativă). Uneori, în măsurătoarea fenomenelor pe care dorim să le explicăm, variabilele nu sunt continue ci categoriale sau dihotomice. Ne amintim din secțiunea 3.1.1 că variabilele

dihotomice sunt formate din două categorii de tipul prezența sau absența caracteristicii, 1 sau 0. Din acest motiv, aceste variabile pot fi obținute prin transformarea oricărui tip de variabilă, calitativă sau cantitativă. În acest capitol vom prezenta o derivație a modelelor de regresie, numită regresia logistică, în care obiectivul este modelarea unei variabile dependente de tip dihotomic.

Variabilele dihotomice (numite și variabile *dummy* sau binare) pot fi folosite în modelele de regresie nu doar ca variabile explicative, ci și ca variabile de explicat. Totuși, acest tip de variabile nu respectă asumpțiile precizate în secțiunile anterioare necesare pentru a putea folosi analiza de regresie pentru explicarea unei variabile dependente, cum ar fi normalitatea, homoscedasticitatea, liniaritatea relației sau nivelul de măsurare cantitativ (de interval sau de rapoarte) al variabilelor. Analiza de regresie logistică rezolvă această limitare folosind relația dintre categoriile pe care le putem crea atunci când avem variabile non-cantitative (categoriale). De exemplu, atunci când vrem să explicăm preferința de vot pentru diverse partide, participarea sau neparticiparea la un eveniment; acestea prezentând o distribuție binomială (de tip 1 sau 0) (Agresti și Finlay 2014, 483).

Astfel, din aceste variabile putem construi *n-1* variabile dihotomice, unde *n* este numărul de categorii iar suma categoriilor este egală cu 1. (Rotariu et al. 1999). Așadar, modelul de regresie logistică poate fi înțeles ca sumă de proporții de răspunsuri pentru categoriile variabilei dependente. Acest model explică raportul de șanse (în limba engleză *odds ratio*) între probabilitatea ca evenimentul (sau comportamentul) să se producă ($p(y = 1)$) (Scenariul 1) și probabilitatea ca evenimentul să nu se producă ($1 - p(y = 1)$) (Scenariul 0)[42], în funcție de variabilele independente (x). Folosind logaritmul natural, modelul are, prin urmare, forma generală de estimare a curbei de regresie logistică:

$$log\left[\frac{p(y = 1)}{1 - p(y = 1)}\right] = \alpha + \beta x + \varepsilon \quad (24)$$

[42] Îi mulțumim lui Andrei Gheorghiță pentru această sugestie a simplificării explicației scenariilor de succes și eșec al probabilităților evenimentului.

Folosind transformata logit (Agresti și Finlay 2014, 484) putem prescurta modelul la:

$$logit[p(y = 1)] = \alpha + \beta x + \varepsilon$$

(25)

unde β reprezintă coeficientul de regresie logistică, interpretarea fiind aceea potrivit căreia o creștere cu o unitate a variabilei independente x conduce la o creștere a șanselor de succes pentru Scenariul 1 în raport cu Scenariul 0 atunci când β are valori pozitive, SAU o creștere cu o unitate a variabilei independente x ce conduce la o scădere a șanselor Scenariului 1 în raport cu Scenariul 0 atunci când β are valori negative atunci, luând valoarea 0 atunci când șansele Scenariului 1 și ale Scenariului 0 sunt egale, chiar dacă variabila independentă x se modifică. α este constanta. Putem transforma raționamentul și explicația coeficienților β din terminologia de evaluare a șanselor, care poate fi uneori mai greu de evaluat, în probabilități, folosind funcția exponențială $p(y = 1) = \frac{exp^{\beta}}{1+exp^{\beta}}$ care transformă șansele în probabilități.

Desigur, la fel ca în cazul regresiei multivariate ce permite modelarea unei explicații cu mai mult de o variabilă independentă, şi modelul de regresie logistică permite o asemenea modelare, caz în care, pe lângă prima variabilă independentă x_1 mai putem adăuga şi celelalte variabile x_2, x_3, ..., x_n formula fiind ajustată în consecință (Rotariu et al. 1999, 229; Agresti și Finlay 2014, 488). În această situație, în interpretarea coeficienților β pentru relația dintre variabila dependentă și fiecare variabilă independentă asumăm că celelalte variabile predictor (independente) sunt constante.

Așadar, regresia logistică este un model probabilistic de analiză statistică. Rezultatele regresiei logistice subliniază probabilitatea cu care variabila rezultat (dependentă) înregistrează una dintre categoriile de răspuns posibile, estimarea parametrilor ecuației de regresie respectând criteriul verosimilității maxime. Există situații în care variabila dependentă poate înregistra două sau mai multe categorii de răspuns și în funcție de categoriile dependentei aplicăm modelul de regresie logistică potrivit după cum urmează:

- când variabila dependentă înregistrează două categorii de răspuns, adică este dihotomică (de exemplu, variabila sex poate înregistra două valori: masculin și feminin), folosim regresia logistică binomială;
- când variabila dependentă înregistrează mai mult de două categorii de răspuns folosim regresia logistică multinomială (de exemplu, variabila orientare politică poate înregistra mai multe categorii: de stânga, de dreapta sau de centru);
- când variabila dependentă are categoriile ordonate, adică este ordinală, folosim regresia logistică ordinală (de exemplu, variabila nivel de educație înregistrează categorii ordonate precum educație primară, secundară și terțiară. Acestea sunt ordonate întrucât între cele trei categorii există, de regulă același interval de timp petrecut în școală).[43]

Regresia logistică poate avea una sau mai multe variabile independente (x_i), denumite și predictori, sau variabile explicative. În modelele de regresie logistică, variabilele explicative pot fi continue și/sau categoriale.

În secțiunea următoare vom prezenta parametrii regresiei logistice care ne ajută să interpretăm rezultatele modelului de regresie, iar apoi vom exemplifica trei tipuri de regresiei logistică: binară, multinomială și ordinală.

3.5.1 Evaluarea modelului de regresie logistică

Pentru regresia logistică estimarea parametrilor se bazează pe maximizarea probabilității de realizare cu succes a unui eveniment. Modelul de regresie logistică prezice probabilitatea ca un eveniment să se producă pentru o anumită observație ($P(Y_i)$), astfel că pentru o observație evenimentul Y poate fi fie 0 (evenimentul nu s-a

[43] Modelele de regresie pentru care variabila dependentă este calitativă pot fi de tip probit sau logit, fiind diferite în ceea ce privește specificarea distribuției erorilor. Dacă distribuția cumulată a erorilor este o funcție logistică rezultă un model de tip logit. Dacă distribuția cumulată a erorilor urmează o distribuție normală, rezultă un model de tip probit.

întâmplat), fie 1 (evenimentul s-a întâmplat), iar rezultatul prezis $P(Y)$ va avea o valoare între 0 (nu există nicio șansă ca rezultatul să apară) și 1 (rezultatul va avea loc cu siguranță).

Am văzut în regresia liniară multiplă că, dacă dorim să evaluăm gradul în care modelul estimează datele, putem compara valorile observate și cele estimate ale rezultatului (folosind R^2). Similar, pentru regresia logistică putem utiliza valorile observate și estimate pentru a evalua gradul în care modelul estimează datele. Măsura pe care o folosim pentru a calcula coeficienții de regresie este verosimilitatea (în engleză *log-likelihood*). Se consideră că un model este mai mult sau mai puțin verosimil atunci când, folosind variabilele independente din model, se pot estima corect valorile variabilei dependente y. Conceptul de verosimilitate poate fi utilizat în calcularea coeficienților de regresie pe baza unui algoritm iterativ, denumit metoda verosimilității maxime (în engleză *Maximum-Likelihood Estimation - MLE*) care se bazează pe transformarea variabilei dependente într-o variabilă de tip logit (logaritmul natural al șansei ca evenimentul să se producă sau nu).

Coeficientul de determinare multiplă R^2 al regresiei de tip OLS este un instrument care indică măsura în care modelul regresiei estimează datele. De asemenea, coeficientul *log-likelihood* discutat anterior este similar prin faptul că se bazează pe nivelul de corespondență dintre valorile estimate (prognozate) și cele observate (reale) ale rezultatului. Cu toate acestea, putem calcula și o versiune a corelației multiple în regresia logistică cunoscută sub numele de R-statistic. Aceasta este corelația parțială dintre variabila rezultat și fiecare dintre variabilele predictor și poate varia între -1 și 1. O valoare pozitivă indică faptul că, pe măsură ce variabila predictor crește, la fel și probabilitatea de apariție a evenimentului. O valoare negativă, implică faptul că, pe măsură ce variabila predictor crește, probabilitatea ca rezultatul să apară scade. Dacă o variabilă are o valoare mică a lui R, atunci contribuie doar într-o mică măsură la model.

Evaluarea modelului unei regresii logistice se poate realiza și prin compararea indicatorilor Akaike Information Criterion (AIC) (Akaike 1974) și Bayesian Information Criterion (BIC) (Schwarz 1978) pentru a evalua potrivirea modelului. AIC și BIC sunt folosiți atunci când vrem să excludem nesiguranța pe care o ridică valoarea

R^2 și anume că de fiecare dată când adăugăm o variabilă la model, R^2 crește. Cei doi indicatori de informare ne ajută să decidem care este modelul de regresie mai bun, atunci când comparăm mai multe modele cu variabile independente similare. De regulă un indicator AIC sau BIC mai scăzut indică un model de regresie mai bun.

Modelul de regresie logistică este legat direct de noțiunea de șanse de succes (*odds*, în limba engleză), și reprezintă raportul dintre probabilitatea de succes și probabilitatea de eșec a evenimentului. Șansa de succes este exponențialul lui B (adică eB sau exp(B)) și este un indicator al modificării șanselor de succes care rezultă dintr-o schimbare cu o unitate a valorii predictorului.

3.5.2 Regresia logistică binară

Există mai multe tipuri de regresie logistică printre care cele mai cunoscute sunt regresia logistică binară, regresia logistică multinomială și regresia logistică ordinală. Pentru început, vom considera un model de regresie logistică binară, în care variabila dependentă (y) este dihotomică (binară) și în care există o singură variabilă independentă (x) de tip cantitativ.

În baza setului nostru de date ESS să presupunem că suntem interesați să aflăm probabilitatea cu care un respondent a votat la ultimele alegeri (*vote*) dacă are încredere în parlament (*trstprl*) și dacă are un nivel de educație ridicat (*eisced*).

În RStudio, funcția de bază utilizată pentru modelarea fenomenelor care au ca răspuns variabile categorice este funcția **glm()** (*generalized linear models* în engleză) și are următoarea structură:

```
regresie_logit <- glm(vote~trstprl + eisced, data = dataset_exemplu, famil
y = "binomial")
summary(regresie_logit)

## Call:
```

```
## glm(formula = vote ~ trstprl + eisced, family = "binomial", data = data
set_exemplu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0957   0.4857   0.6535   0.7237   1.0392
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.53768    0.26798  -2.006 0.044811 *
## trstprl      0.32131    0.09246   3.475 0.000511 ***
## eisced       0.55060    0.09283   5.931 3.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1515.0  on 1433  degrees of freedom
## Residual deviance: 1468.1  on 1431  degrees of freedom
## AIC: 1474.1
##
## Number of Fisher Scoring iterations: 4
```

Similar regresiei liniare, atenția se îndreaptă, în special asupra coeficientului de regresie b1 care exprimă modificarea y atunci când *x* crește cu o unitate. Uneori, este mai ușor de interpretat eb1 care semnifică efectul generat de coeficientul de regresie b1 asupra șanselor de succes. Așadar, eb1 reprezintă raportul de șanse care arată ce se întâmplă atunci când *x* se modifică cu o unitate. În cazul nostru, o creștere a nivelului de încredere în Parlament va rezulta într-o creștere de 32% a șanselor ca respondentul să fi votat la ultimele alegeri.

Odată ce am estimat coeficienții de regresie este important să evaluăm modelul. Evaluarea modelului de regresie logistică constă în parcurgerea a două etape:

- **Prima etapă** determină dacă există variabile independente care nu au o influență semnificativă asupra dependentei;
- **A doua etapă** evaluează prin intermediul unor măsuri stabilite convențional dacă modelul este adecvat (în engleză *goodness-of-fit*); această etapă presupune și evaluarea capacității de predicție a modelului.

Pentru prima etapă de evaluare a modelului de regresie se verifică dacă variabilele independente (încrederea în Parlament și nivelul de educație) au sau nu o influență semnificativă asupra variabilei dependente (participarea la vot). Pentru aceasta, se verifică dacă coeficienții de regresie estimați (b1 = 0.32131 și b2 = 0.55060) sunt semnificativ statistic. Aplicând în RStudio funcția **glm()** setului de date considerat, aceasta returnează următorul rezultat:

```
prima_etapa1 <- glm(formula = vote~trstprl, family = "binomial",
                    data =  dataset_exemplu)
summary(prima_etapa1)


## Call:
## glm(formula = vote ~ trstprl, family = "binomial", data = dataset_exemp
lu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9259   0.5835   0.6708   0.7680   0.7680
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.76286    0.15671   4.868 1.13e-06 ***
## trstprl      0.30718    0.09111   3.372 0.000747 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1515.0  on 1433  degrees of freedom
## Residual deviance: 1503.2  on 1432  degrees of freedom
## AIC: 1507.2
##
## Number of Fisher Scoring iterations: 4
```

Mai sus am verificat influența variabilei *trstprl* asupra variabilei dependente, iar mai jos vom testa influența variabilei *eisced* din modelul nostru.

```
prima_etapa2 <- glm(formula = vote~eisced, family = "binomial",
                    data =  dataset_exemplu)
summary(prima_etapa2)
```

```
## Call:
## glm(formula = vote ~ eisced, family = "binomial", data = dataset_exempl
u)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9010   0.5989   0.5989   0.7628   0.9568
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.002146   0.217930   0.010    0.992
## eisced      0.541795   0.092280   5.871 4.33e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1515.0  on 1433  degrees of freedom
## Residual deviance: 1480.6  on 1432  degrees of freedom
## AIC: 1484.6
##
## Number of Fisher Scoring iterations: 4
```

Conform rezultatului obținut, coeficientul este semnificativ statistic (semnificativ diferit de zero, conform ipotezei nule), cu un nivel de semnificație $p \leq 0.001$. În această situație, se poate afirma că încrederea în Parlament influențează probabilitatea ca individul să fi votat la ultimele alegeri, astfel încât prin creșterea cu o unitate a încrederii sale, șansele ca el să fi votat cresc, în medie, la 30%.

Calculul limitelor intervalelor de încredere cu care au fost estimați parametrii ecuației de regresie se realizează cu funcția **confint():**

```
confint(prima_etapa1, level  = 0.95)

##                 2.5 %    97.5 %
## (Intercept) 0.4564981 1.0711588
## trstprl     0.1307668 0.4882054

confint(prima_etapa2, level  = 0.95)

##                  2.5 %    97.5 %
## (Intercept) -0.4228707 0.4321982
## eisced       0.3611238 0.7231401
```

Așadar, intervalul de încredere pentru b1, pentru un nivel de semnificație de 5% este [0.3611238 0.7231401]. Similar se procedează pentru toți predictorii din modelul de regresie logistică.

A doua etapă de evaluare constă în testarea modelului de regresie logistică și se aplică testul hi pătrat care este inclus în RStudio prin funcția **anova().**

```
anova(regresie_logit, test = "Chisq")

     ## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vote
##
## Terms added sequentially (first to last)
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     1433     1515.0
## trstprl  1   11.836      1432     1503.2 0.0005808 ***
## eisced   1   35.122      1431     1468.0 3.097e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mai mult, în loc să estimăm coeficientul R^2 ca în regresia liniară, pentru regresia logistică putem calcula o valoare cunoscută sub numele de coeficientul pseudo R^2 al lui McFadden, care variază de la 0 la 1. Valorile apropiate de 0 indică faptul că modelul nu are putere de predicție. În practică, valorile peste 0,3 sau 0,4 indică faptul că un model estimează bine datele.

Putem calcula coeficientul pseudo R^2 lui McFadden pentru modelul nostru folosind funcția **pR2()** din pachetul **pscl** (Jackman 2020):

```
library(pscl)

pscl::pR2(regresie_logit)["McFadden"]
```

```
## fitting null model for pseudo-r2

##   McFadden
## 0.03099525
```

O valoare de 0,03099525, precum cea obținută în analiza de mai sus, indică faptul că modelul nostru estimează datele în proporție de doar 3%. Introducerea mai multor predictori în modelul de regresie va mări puterea predictivă.

De asemenea, putem calcula importanța fiecărei variabile predictor din model folosind funcția **varImp()** din pachetul **caret** (M. Kuhn 2022):

```
library(caret)

caret::varImp(regresie_logit)

##          Overall
## trstprl 3.475128
## eisced  5.931199
```

Valorile mai mari indică o importanță mai mare. Nivelul de educație este cea mai importantă variabilă predictivă. De asemenea, putem calcula valorile VIF ale fiecărei variabile din model pentru a vedea dacă multicoliniaritatea este o problemă:

```
car::vif(regresie_logit)

##  trstprl   eisced
## 1.001798 1.001798
```

Ca regulă generală, valorile VIF de peste 5 indică multicoliniaritate severă. Deoarece toate variabilele predictor din modelele noastre au un VIF puțin mai mare de 1, putem presupune că multicoliniaritatea nu este o problemă.

Putem folosi următorul cod pentru a calcula probabilitatea de a fi votat pentru fiecare individ din setul nostru de date:

```
predicted <- predict(regresie_logit, dataset_exemplu, type="response")
```

În ansamblu, putem aprecia pe baza rezultatelor modelului de regresie că probabilitatea ca un respondent să fi votat la ultimele alegeri depinde de nivelul de încredere (cu cât este mai mare încrederea cu atât probabilitatea este mai mare) și de nivelul de educație (un nivel de educație ridicat indică o probabilitate mare de participare la vot).

3.5.3 Regresia logistică multinomială

Modelul de regresie logistică multinomială este o generalizare a regresiei logistice binomiale; în acest caz, variabila dependentă de tip categoric are mai mult de două categorii de răspuns. În setul nostru de date de la ESS am ales ca variabilă dependentă cu mai multe categorii, variabila *tporgwk* care indică ocupația respondenților. Astfel, suntem interesați să observăm dacă nivelul de încredere și satisfacția respondenților diferă în funcție de ocupația acestora.

Pentru a realiza acest lucru am creat din setul de date original ESS un set de date secundar dataset_multinomial.dta încărcat în RStudio.

Descrierea variabilelor:

- *tporgwk*: variabila dependentă care indică ocupația are următoarele categorii de răspuns: sectorul public, sectorul public (educația/sănătate), compania de stat, sectorul privat și antreprenor. Avem astfel o variabilă dependentă categorică cu 5 categorii. Categoria *companie de stat* este categoria de referință.
- *trstprt*, *trstplt*, *stfeco*, *stflife*, *polintr* și *vote* vor fi variabilele independente sau factoriale ale modelului de regresie multinomială.

Calculăm coeficienții de regresie pe baza funcției multinom(), furnizată în pachetul de date nnet (Venables și Ripley 2002). De asemenea, vom mai avea nevoie

de pachetul foreign (R Core Team 2022a). Primul pas este să transformăm clasa variabilei dependente în factor:

```
table(dataset_multinomial$tporgwk)

##
##                                     Antreprenor
##                                             192
##                                Companie de stat
##                                             130
##                                 Sectorul privat
##                                             763
## Sectorul public (inclusiv educatie si sanatate)
##                                             345

dataset_multinomial$tporgwk <- as.factor(dataset_multinomial$tporgwk)
```

Următorul pas este să selectăm categoria de referință:

```
dataset_multinomial$tporgwk2 <- relevel(dataset_multinomial$tporgwk, ref =
"Companie de stat")
```

Al treilea pas este să calculăm coeficienții de regresie pe baza funcției multinom() din pachetul nnet (Venables și Ripley 2002):

```
library(nnet)
library(stargazer)

### Rulăm modelul multilogistic folosind funcția multinom().###
regresie_multinomiala = multinom(tporgwk2 ~ trstprt + trstplt + stflife +
stfeco +stfdem + stfgov + polintr + clsprty,
                    data=dataset_multinomial, Hess = TRUE)

summary(regresie_multinomiala)

## Call:
## multinom(formula = tporgwk2 ~ trstprt + trstplt + stflife + stfeco +
##     stfdem + stfgov + polintr + clsprty, data = dataset_multinomial,
##     Hess = TRUE)
```

```
##
## Coefficients:
##                                                 (Intercept)     trstprt
## Antreprenor                                      -0.8009336 -0.18256058
## Sectorul privat                                   1.3131993 -0.07645457
## Sectorul public (inclusiv educatie si sanatate)   0.1032041 -0.08874988
##                                                     trstplt   stflife
## Antreprenor                                      0.1205722660 0.6176875
## Sectorul privat                                  0.0009916111 0.2567398
## Sectorul public (inclusiv educatie si sanatate) 0.0794491927 0.3905884
##                                                     stfeco     stfdem
## Antreprenor                                      -0.1051122 -0.2146175
## Sectorul privat                                   0.3907265 -0.2057645
## Sectorul public (inclusiv educatie si sanatate)   0.1331892 -0.1683603
##                                                     stfgov     polintr
## Antreprenor                                      -0.0588526  0.23068993
## Sectorul privat                                  -0.2280032 -0.31960210
## Sectorul public (inclusiv educatie si sanatate) -0.1417826  0.09304631
##                                                   clsprty
## Antreprenor                                      0.5275289
## Sectorul privat                                  0.3234944
## Sectorul public (inclusiv educatie si sanatate) 0.3844861
##
## Std. Errors:
##                                                 (Intercept)   trstprt   trstplt
## Antreprenor                                       0.5557444 0.3040389 0.3015470
## Sectorul privat                                   0.4266971 0.2498885 0.2499016
## Sectorul public (inclusiv educatie si sanatate)   0.4748761 0.2710892 0.2706862
##                                                     stflife    stfeco    stfdem
## Antreprenor                                      0.1945090 0.2091997 0.2124620
## Sectorul privat                                  0.1516257 0.1756531 0.1783800
## Sectorul public (inclusiv educatie si sanatate) 0.1680226 0.1893020 0.1927291
##                                                     stfgov   polintr   clsprty
## Antreprenor                                      0.2005025 0.2428959 0.2449776
## Sectorul privat                                  0.1698347 0.2021457 0.2059429
## Sectorul public (inclusiv educatie si sanatate) 0.1830616 0.2186161 0.2219061
##
## Residual Deviance: 3279.762
## AIC: 3333.762
```

În baza rezultatelor regresiei multinomiale putem sublinia o serie de observații. Astfel, pentru o creștere cu o unitate a scorului referitor la nivelul de satisfacție față propria viață șansele multinomiale pentru a fi antreprenor și nu angajat al unei companii de stat cresc cu 0,618 unități, menținând toate celelalte variabile din model constante. Pentru angajații din sectorul privat și sectorul public satisfacția cu propria

viață are un impact mai mic. Conform așteptărilor, o creștere cu o unitate a satisfacției cu situația economică crește șansele multinomiale pentru a lucra în sectorul privat și nu la o companie de stat cu 0,391 unități.

Pentru a înțelege mai bine modelul, estimăm probabilitățile asociate fiecărei categorii de răspuns, pentru variabilele independente, utilizând funcția **fitted():**

```
pp <- fitted(regresie_multinomiala)
head(pp)

##   Companie de stat Antreprenor Sectorul privat
## 1       0.09820361  0.19696700       0.4003757
## 2       0.08401873  0.17823973       0.4654975
## 3       0.06024854  0.09997542       0.6353140
## 4       0.10198023  0.09261871       0.6016695
## 5       0.07723166  0.14236123       0.5509940
## 6       0.09706297  0.07007044       0.6445141
##   Sectorul public (inclusiv educatie si sanatate)
## 1                                       0.3044537
## 2                                       0.2722440
## 3                                       0.2044621
## 4                                       0.2037315
## 5                                       0.2294131
## 6                                       0.1883525
```

De asemenea, pentru e evalua modelul de regresie multinomială vom calcula valoare *log likelihood:*

```
### Calculăm log likelihood###
logLik(regresie_multinomiala)

## 'log Lik.' -1639.881 (df=27)
```

Pentru a accesa într-o manieră organizată rezultatele regresiei vom crea cu ajutorul funcției **stargazer()** un tabel de ieșire de tip html. Tabelul html a fost salvat în destinația de lucru curentă.

```
### Generăm tabel de ieșire HTML###
stargazer(regresie_multinomiala, type="html", out="my_multinomial.htm")
```

De asemenea, pentru simplificarea lecturii rezultatelor regresiei multinomiale putem reproduce în Microsoft Word un tabel de rezultat similar celui de mai jos:

Tabel 3.10 Model de regresie logistică multinomială

	Antreprenor	**Sectorul privat**	**Sectorul public (inclusiv educatie și sănătate)**
	(1)	(2)	(3)
Încrederea în partidele politice (trstprt)	-0.183	-0.076	-0.089
	(0.304)	(0.25)	(0.271)
Încrederea în politicieni (trstplt)	0.121	0.001	0.079
	(0.302)	(0.25)	(0.271)
Satisfacția cu propria viață (stflife)	**0.618*****	**0.257***	**0.391****
	(0.195)	(0.152)	(0.168)
Satisfacția cu situația economică (stfeco)	-0.105	**0.391****	0.133
	(0.209)	(0.176)	(0.189)
Satisfacția cu democrația (stfdem)	-0.215	-0.206	-0.168
	(0.212)	(0.178)	(0.193)
Satisfacția cu guvernul (stfgov)	-0.059	-0.228	-0.142
	(0.201)	(0.17)	(0.183)
Interesul pentru politică (polintr1)	0.231	-0.32	0.093
	(0.243)	(0.202)	(0.219)
Atașamentul față de un partid (clsprty)	**0.528****	0.323	**0.384***
	(0.245)	(0.206)	(0.222)
Constant	-0.801	**1.313*****	0.103
	(0.556)	(0.427)	(0.475)
Akaike Inf. Crit. (AIC)	3,333.76	3,333.76	3,333.76
LogLik		-1639.881	

Notă: Nivelul de semnificație statistică este *p<0.1; **p<0.05; ***p<0.001

3.5.4 Regresia logistică ordinală

Regresia logistică ordinală este o extensie a modelului de regresie logistică simplă. În regresia logistică simplă, variabila dependentă este categorică și urmează o distribuție Bernoulli. În timp ce, în regresia logistică ordinală, variabila dependentă este ordinală, adică există o ordonare a categoriilor de răspuns.

Să presupunem că vrem să testăm, conform ipotezelor formulate la începutul acestui capitol, impactul nivelului de satisfacție cu economia, propria viață, democrația, dar și cu modul în care guvernul a gestionat pandemia de COVID-19, asupra nivelului de încredere în politicieni. Astfel, din setul de date ESS vom folosi ca variabilă dependentă *trstplt* care este transformată într-o variabilă categorică ordinală cu 3 nivele: 1 încredere mică, 2, încredere medie și 3, încredere mare.

În RStudio calculăm coeficienții de regresie pe baza funcției **polr(),** furnizată în pachetul de date **MASS** (Venables și Ripley 2002).

```
library(MASS)

regresie_ordinala <- polr(as.factor(trstplt) ~ stflife + stfeco + stfdem +
polintr + clsprty, data = dataset_exemplu, Hess=TRUE)
summary(regresie_ordinala)

## Call:
## polr(formula = as.factor(trstplt) ~ stflife + stfeco + stfdem +
##     polintr + clsprty, data = dataset_exemplu, Hess = TRUE)
##
## Coefficients:
##           Value Std. Error t value
## stflife  0.2854    0.10751   2.654
## stfeco  -0.6729    0.09610  -7.002
## stfdem  -1.3243    0.09998 -13.245
## polintr -0.3672    0.12375  -2.968
## clsprty -0.2090    0.12408  -1.685
##
## Intercepts:
##                             Value    Std. Error t value
## incredere|incredere medie    -6.1784   0.3425   -18.0370
```

```
## incredere medie|neincredere  -3.7641   0.3062   -12.2915
##
## Residual Deviance: 2045.186
## AIC: 2059.186
```

Pentru a putea stabili puterea de predicție a modelului următorul pas este să rulăm un model de comparație în care includem doar valorile constante ale categoriilor din variabila dependentă.

```
### Rulăm un model „numai valoarea constantă a categoriilor variabilei dep
endente”###
OIM <- polr(as.factor(trstplt) ~ 1, data = dataset_exemplu)
summary(OIM)


## Call:
## polr(formula = as.factor(trstplt) ~ 1, data = dataset_exemplu)
##
## No coefficients
##
## Intercepts:
##                              Value    Std. Error t value
## incredere|incredere medie    -2.3400   0.0933   -25.0857
## incredere medie|neincredere  -0.4628   0.0542    -8.5325
##
## Residual Deviance: 2507.249
## AIC: 2511.249
```

Acest model va fi folosit cu modelul de bază al regresiei noastre prin aplicarea funcției **anova()** după cum urmează:

```
### Comparăm modelul nostru cu modelul test „OIM”.###
anova(OIM,regresie_ordinala)

## Likelihood ratio tests of ordinal regression models
##
## Response: as.factor(trstplt)
##                                          Model Resid. df Resid. Dev   Test
## 1                                            1      1432   2507.249
## 2 stflife + stfeco + stfdem + polintr + clsprty      1427   2045.186 1 vs 2
##      Df LR stat. Pr(Chi)
```

```
## 1
## 2     5 462.0622       0
```

În continuare vom calcula valorile de predicție pentru categoriile variabilei de interes *trstplt*:

```
### Verificăm probabilitatea prezisă pentru fiecare categorie###
p <- regresie_ordinala$fitted.values
head(p)

##     incredere incredere medie neincredere
## 1 0.140332608      0.50570956   0.3539578
## 2 0.015132146      0.13147431   0.8533935
## 3 0.012562476      0.11197106   0.8754665
## 4 0.013959647      0.12270074   0.8633396
## 5 0.006449275      0.06121717   0.9323336
## 6 0.059818986      0.35586123   0.5843198
```

Rezultatul predicțiilor poate fi obținut în RStudio și prin utilizarea funcției **predict():**

```
### Putem obține rezultatul predicțiilor folosind funcția predict()###
predicted_value <- predict(regresie_ordinala)
head(predicted_value)

## [1] incredere medie neincredere     neincredere     neincredere
## [5] neincredere     neincredere
## Levels: incredere incredere medie neincredere
```

În cele din urmă vom realiza testul pentru estimarea modelului folosind funcția **chisq.test()**, precum și tabelul de confuzie și eroarea de clasificare:

```
### Testăm pentru goodness of fit###
chisq.test(dataset_exemplu$trstplt,predict(regresie_ordinala))
```

```
##  Pearson's Chi-squared test
## data:  dataset_exemplu$trstplt and predict(regresie_ordinala)
## X-squared = 300.97, df = 4, p-value < 2.2e-16

### Calculăm confusion table și ssmisclassification error###
predictOLR <-  predict(regresie_ordinala,dataset_exemplu)
cTab <- table(dataset_exemplu$trstplt, predictOLR)
mean(as.character(dataset_exemplu$trstplt) != as.character(predictOLR))

## [1] 0.3361227

### Calculăm classification rate###

(CCR <- sum(diag(cTab)) / sum(cTab))

## [1] 0.6638773
```

Pentru regresia logistică ordinală există mai multe statistici asemănătoare R^2 care pot fi utilizate pentru a măsura puterea asocierii dintre variabila dependentă și variabilele predictor. În RStudio vom încărca pachetul DescTools (Signorell 2022) și vom folosi funcția PseudoR2():

```
### Încărcăm pachetul DescTools package pentru a calcula R2###
library("DescTools")

### Calculăm R2###
PseudoR2(regresie_ordinala, which = c("CoxSnell","Nagelkerke","McFadden"))

##   CoxSnell Nagelkerke   McFadden
##  0.2754606  0.3335064  0.1842906

### Folosim funcția coef() pentru a verifica estimările parametrilor###
(OLRestimates <- coef(summary(regresie_ordinala)))

##                                 Value Std. Error    t value
## stflife                     0.2853811 0.10751150   2.654424
## stfeco                     -0.6729184 0.09609989  -7.002280
## stfdem                     -1.3242685 0.09998433 -13.244760
## polintr                    -0.3672214 0.12374643  -2.967531
## clsprty                    -0.2090287 0.12408371  -1.684578
## incredere|incredere medie  -6.1783609 0.34253901 -18.036956
## incredere medie|neincredere -3.7641437 0.30623852 -12.291543
```

```
### Adăugăm valoarea p la tabelul de estimare a parametrilor###
p <- pnorm(abs(OLRestimates[, "t value"]), lower.tail = FALSE) * 2
(OLRestimates_P <- cbind(OLRestimates, "p value" = p))

##                                Value Std. Error    t value      p value
## stflife                     0.2853811 0.10751150   2.654424 7.944391e-03
## stfeco                     -0.6729184 0.09609989  -7.002280 2.518296e-12
## stfdem                     -1.3242685 0.09998433 -13.244760 4.838242e-40
## polintr                    -0.3672214 0.12374643  -2.967531 3.002018e-03
## clsprty                    -0.2090287 0.12408371  -1.684578 9.207002e-02
## incredere|incredere medie  -6.1783609 0.34253901 -18.036956 9.989787e-73
## incredere medie|neincredere -3.7641437 0.30623852 -12.291543 1.005685e-34
```

Înainte de a interpreta rezultatul regresiei logistice ordonate este indicat să calculăm efectele marginale. Efectele marginale arată modificarea probabilității atunci când variabila independentă crește cu o unitate. În RStudio, putem calcula efectele marginale prin aplicarea următoarei sintaxe cu ajutorul pachetului **margins** (Leeper et al. 2021):

```
### Încărcăm pachetul margins pentru a calcula efectele marginale###
library("margins")

### Calculăm efectele marginale###
efecte_marginale_logit_ordered <- margins(regresie_ordinala)

summary(efecte_marginale_logit_ordered)

##factor     AME     SE       z      p   lower   upper
##clsprty  0.0146 0.0093  1.5649 0.1176 -0.0037  0.0329
##polintr  0.0256 0.0105  2.4296 0.0151  0.0050  0.0463
##stfdem   0.0924 0.0213  4.3359 0.0000  0.0506  0.1342
##stfeco   0.0470 0.0124  3.7815 0.0002  0.0226  0.0713
##stflife -0.0199 0.0058 -3.4098 0.0007 -0.0314 -0.0085
```

Pentru a accesa într-o manieră organizată rezultatele regresiei logistice ordonate și pentru a ușura interpretarea coeficienților, în tabelul de ieșire vom trece valorile pentru efectele marginale (AME) și în paranteză eroarea standard (SE) a acestora după cum urmează:

Tabel 3.11 Model regresie logistică ordinală

	Nivelul de încredere
Satisfacție cu viața (stflife)	0.285
	(0.107)
Interesul pentru politică (polintr1)	**-0.367**
	(0.123)
Atașamentul față de un partid (clsprty)	-0.209
	(0.124)
Satisfacția cu situația economică (stfeco)	**-0.672**
	(0.096)
Satisfacția cu democrația (stfdem)	**-1.324**
	(0.099)
AIC	2511.249
R^2 al lui Nagelkerke	0.333
Număr de observații (N)	1,434

Pentru modelul nostru, variabilele independente explică 33,3% din variația variabilei dependente, conform coeficientului R^2 al lui Nagelkerke. În baza rezultatelor de mai sus putem conclude faptul că nivelul de satisfacție al oamenilor are un impact pozitiv semnificativ asupra nivelului de încredere în politicieni. O creștere cu o unitate a nivelului de satisfacție cu economia determină o creștere de 4,7 puncte procentuale a nivelului de încredere în politicieni. De asemenea, o creștere cu o unitate a interesului pentru politică, determină o creștere de 2,5 puncte procentuale a încrederii în politicieni.

4. Aplicații practice în Stata

Similar aplicațiilor practice exemplificate în RStudio, în capitolul 4 vom prezinta modalitatea de realizare a acestor analize în programul Stata. Versiunea pe care o vom folosi în aplicarea analizelor este Stata SE 16.0.

4.1. Încărcarea setului de date în Stata

Setul de date pe care vom lucra pentru exemplificarea în Stata este cel de la European Social Survey numit ESS10.dta. Încărcarea și deschiderea setului de date în programul Stata se poate realiza fie prin folosirea meniului și a comenzilor acestuia, fie direct prin scrierea următoarei comenzi în chenarul Command:[44]

```
use ESS10.dta
```

4.2. Familiarizarea cu variabilele din setul de date în Stata

Odată deschis setul de date este important să ne familiarizăm cu variabilele acestuia.[45] Primul pas este acela de a realiza o analiză exploratorie preliminară a variabilelor. Există mai multe comenzi care sunt deosebit de utile în procesul de familiarizare cu datele. Una dintre acestea este comanda **list** care oferă o listă a tuturor variabilelor incluse în setul de date:

[44] În caseta Command nu este necesar să adăugăm și extensia .dta întrucât aceasta este citită implicit.

[45] De asemenea, este recomandabil să examinăm documentația tehnică și codurile disponibile pentru descărcare împreună cu setul de date. Acesta din urmă ne oferă o listă a întrebărilor adresate în timpul sondajului, numele variabilei corespunzătoare și categoriile de răspunsuri.

```
list ESS10
```

O altă comanda extrem de utilă atunci când vrem să ne familiarizăm cu variabilele dintr-un set mare de date este comanda **codebook().** Această comandă oferă o descriere a variabilelor, iar comenzile **inspect** și **summarize**, afișează un rezumat al unei variabile, inclusiv o mică histogramă, respectiv statistici rezumative, cum ar fi medii și abateri standard.

```
codebook
inspect
summarize
```

Stata permite aplicarea acestor comenzi și pentru variabile particulare. Să presupunem că vrem să aplicăm comenzile **inspect** și **summarize** pentru variabile *stfeco:*[46]

```
inspect stfeco
summarize stfeco
```

Tabele de ieșire pentru cele două comenzi aplicate la variabila de interes *stfeco* ne prezintă următoarele informații:

[46] Utilizăm funcția Stata Keyword Search pentru a afla mai multe informații despre modurile în care aceste comenzi pot fi modificate și despre diferitele opțiuni disponibile pentru fiecare. Pentru a căuta rapid informații despre anumite comenzi, putem tasta **help** urmat de cuvântul cheie pe care dorim să îl căutăm în fereastra Command sau DO – file.

Figura 4.1 Tabelul de ieșire pentru comanda inspect

```
stfeco:  How satisfied with present stat        Number of Observations
------------------------------------------------------------------------------
                                                      Total      Integers   Nonintegers
          #                     Negative                  -             -             -
          #   #                 Zero                  1,472         1,472             -
      #   #   #                 Positive             16,169        16,169             -
      #   #   #                                  ----------    ----------    ----------
  #   #   #   #   #             Total                17,641        17,641             -
  #   #   #   #   #             Missing                 419
|-----------------------|
0                      10                           18,060
  (11 unique values)
```

Foarte important la comanda **inspect** este faptul că indică și numărul da valori lipsă. În cazul nostru, pentru variabila *stfeco* un număr de 419 respondenți dintr-un total de 17,4641 nu au răspuns la această întrebare.

Figura 4.2 Tabelul de ieșire pentru comanda summarize

```
. summarize stfeco
```

Variable	Obs	Mean	Std. Dev.	Min	Max
stfeco	17,641	4.654725	2.488362	0	10

Pentru un set de date foarte mare, cum ar fi ESS, informațiile furnizate pot fi copleșitoare, având în vedere numărul mare de variabile. O soluție este să limităm rezultatele prin listarea numai a variabilelor pe care intenționăm să le explorăm. Pentru a identifica variabilele de interes după cuvinte cheie vom utiliza comanda **lookfor**. Spre exemplu, dacă aplicăm comanda **lookfor** urmată de cuvântul cheie *vote*, Stata va genera toate variabilele care includ cuvântul „vot” în numele sau eticheta variabilei:[47]

[47] Cel mai rapid mod de a scana variabilele unui set mare de date și de a alege acele variabile de interes pentru analiză este să petrecem un timp citind codebook-ul disponibil spre descărcare o dată cu setul de date.

```
lookfor vote
```

Obținem un tabel de ieșire precum în Figura 4.3.

Figura 4.3 Tabel de ieșire pentru comanda lookfor

vote	byte	%1.0g	vote	Voted last national election
prtvtebg	byte	%2.0g	prtvtebg	Party voted for in last national election, Bulgaria
prtvtbhr	byte	%2.0g	prtvtbhr	Party voted for in last national election, Croatia
prtvtecz	byte	%2.0g	prtvtecz	Party voted for in last national election, Czechia
prtvthee	byte	%2.0g	prtvthee	Party voted for in last national election, Estonia
prtvtefi	byte	%2.0g	prtvtefi	Party voted for in last national election, Finland
prtvtefr	byte	%2.0g	prtvtefr	Party voted for in last national election, France (ballot 1)
prtvtghu	byte	%2.0g	prtvtghu	Party voted for in last national election, Hungary
prtvclt1	byte	%2.0g	prtvclt1	Party voted for in last national election 1, Lithuania (first vote, party)
prtvclt2	byte	%2.0g	prtvclt2	Party voted for in last national election 2, Lithuania (second vote, party)
prtvclt3	byte	%2.0g	prtvclt3	Party voted for in last national election 3, Lithuania (third vote, party)
prtvtfsi	byte	%2.0g	prtvtfsi	Party voted for in last national election, Slovenia
prtvtesk	byte	%2.0g	prtvtesk	Party voted for in last national election, Slovakia
vteurmmb	byte	%2.0g	vteurmmb	Would vote for [country] to remain member of European Union or leave
vteubcmb	int	%1.0g	vteubcmb	Would vote for [country] to become member of European Union or remain outside
votedir	byte	%2.0g	votedir	Citizens have the final say on political issues by voting directly in referendum
votedirc	byte	%2.0g	votedirc	In country citizens have the final say on political issues by voting directly in
iplylfr	byte	%1.0g	iplylfr	Important to be loyal to friends and devote to people close

Odată ce am identificat variabilele de interes putem să obținem informații despre acestea, folosind comanda **codebook**:

```
codebook stfeco vote
```

4.3. Crearea și recodificarea variabilelor în Stata

Recodificarea variabilelor înainte de a fi utilizate în modele statistice complexe este o caracteristică comună a procesului de analiză a datelor. Recodificarea variabilelor este recomandată a se realiza prin generarea unei noi variabile și ulterior să aplicăm transformările. Transformarea variabilelor originale dintr-un set de date nu este recomandată din nouă motive: (1) generarea unei noi variabile permite să ne întoarcem oricând la versiunea veche a acelei variabile și a verifica dacă transformarea

s-a produs corect și (2) menținerea variabilei originale ne permite să realizăm orice nouă modificare a variabilei dacă este necesar în analize ulterioare, pornind de la datele originale.

4.3.1. Crearea unei noi variabile în Stata

Comanda **generate (gen)** este utilizată atunci când dorim generarea unei noi variabile. Selectăm un nou nume de variabilă pentru noua variabilă și apoi indicăm programului să genereze o nouă variabilă din variabila existentă. Alternativ, putem clona o variabilă existentă, folosind comanda **clonevar vote2 = vote**, ce preia nu doar valorile variabilei vechi, ci și etichetele acestora; iar apoi vom prelucra noua variabilă ce reprezintă o clonă a variabilei originale care este de preferat să o păstrăm în baza de date, pentru referință. Să aplicăm comanda **generate** pe variabila *vote* din setul de date ESS prin generarea unei noi variabile numită vote2.

```
gen vote2 = vote
```

Variabila nou generată poate fi comparată cu variabila inițială prin realizarea unui tabel de contingență. Realizarea unui astfel de tabel pentru verificarea recodificării este recomandată doar atunci când lucrăm cu variabile nominale sau ordinale. În cazul variabilelor de tip interval/raport, folosim o distribuție de frecvențe. În Stata pentru a genera un tabel de contingență vom adăuga numele celor două variabile după comanda **tab**:

```
tab vote vote2
```

Figura 4.4 Tabel de contingență pentru noua variabilă vote2

```
. tab vote vote2
```

Voted last national election	vote2			Total
	1	2	3	
Yes	12,037	0	0	12,037
No	0	4,684	0	4,684
Not eligible to vote	0	0	1,155	1,155
Total	12,037	4,684	1,155	17,876

4.3.2. Recodificarea unei variabile în Stata

Finalizarea etapei de generare a variabilelor conform obiectivelor analizei permite avansarea în realizarea transformărilor necesare pentru aplicarea tehnicilor statistice complexe. Să presupunem că odată cu generarea variabilei *vote2* am generat și variabila *stfeco2* și acum dorim să transformăm variabila *stfeco* în sensul în care vrem să eliminam cazurile lipsă. Primul pas este să transformăm categoriile .a, .b și .c care denotă respondenții care nu au răspuns la întrebare în "." (. reprezintă modalitatea de recunoaștere a cazurilor lipsă în Stata):

```
recode stfeco2 = .a=. .b=. .c=.
```

Variabila nou recodificată trebuie comparată cu variabila originală pentru a ne asigura că nu s-a produs nici o eroare. Vom compara prin realizarea unui tabel de contingență. Pentru a include cazurile lipsă (acele cazuri pe care le-am setat la „.") adăugăm opțiunea lipsă (*missing)* la comanda **tab**, după cum urmează:

```
tab stfeco stfeco2, missing
```

Rezultatul este următorul:

Figura 4.5 Tabelul de ieșire pentru comanda tab

```
. tab stfeco stfeco2, missing
```

How satisfied with present state of economy in country	stfeco2												Total
	0	1	2	3	4	5	6	7	8	9	10	.	
Extremely dissatisfie	1,472	0	0	0	0	0	0	0	0	0	0	0	1,472
1	0	703	0	0	0	0	0	0	0	0	0	0	703
2	0	0	1,520	0	0	0	0	0	0	0	0	0	1,520
3	0	0	0	2,069	0	0	0	0	0	0	0	0	2,069
4	0	0	0	0	1,889	0	0	0	0	0	0	0	1,889
5	0	0	0	0	0	3,117	0	0	0	0	0	0	3,117
6	0	0	0	0	0	0	2,317	0	0	0	0	0	2,317
7	0	0	0	0	0	0	0	2,354	0	0	0	0	2,354
8	0	0	0	0	0	0	0	0	1,458	0	0	0	1,458
9	0	0	0	0	0	0	0	0	0	441	0	0	441
Extremely satisfied	0	0	0	0	0	0	0	0	0	0	301	0	301
.	0	0	0	0	0	0	0	0	0	0	0	419	419
Total	1,472	703	1,520	2,069	1,889	3,117	2,317	2,354	1,458	441	301	419	18,060

.

4.4. Grafice cu bare, diagrame circulare și histograme în Stata

Programul Stata permite realizarea unor reprezentări grafice variate care pot fi extrem de utile atunci când vrem să prezentăm rezultatele analizei noastre publicului larg. În această secțiune vom descrie pașii pentru a realiza grafice și diagrame care pot fi utilizate atunci când vrem să prezentăm analize univariate. Spre exemplu, să raportăm distribuția de frecvențe a variabilei recodificate *stfeco2*, din baza de date ESS, într-o diagramă cu bare și o diagramă circulară. Vom aplica următoarele sintaxe:

```
graph bar, over(stfeco2) title(Nivelul de satisfacție cu situația economică)
```

Figura 4.6 Analiză grafică de tip bar chart în Stata

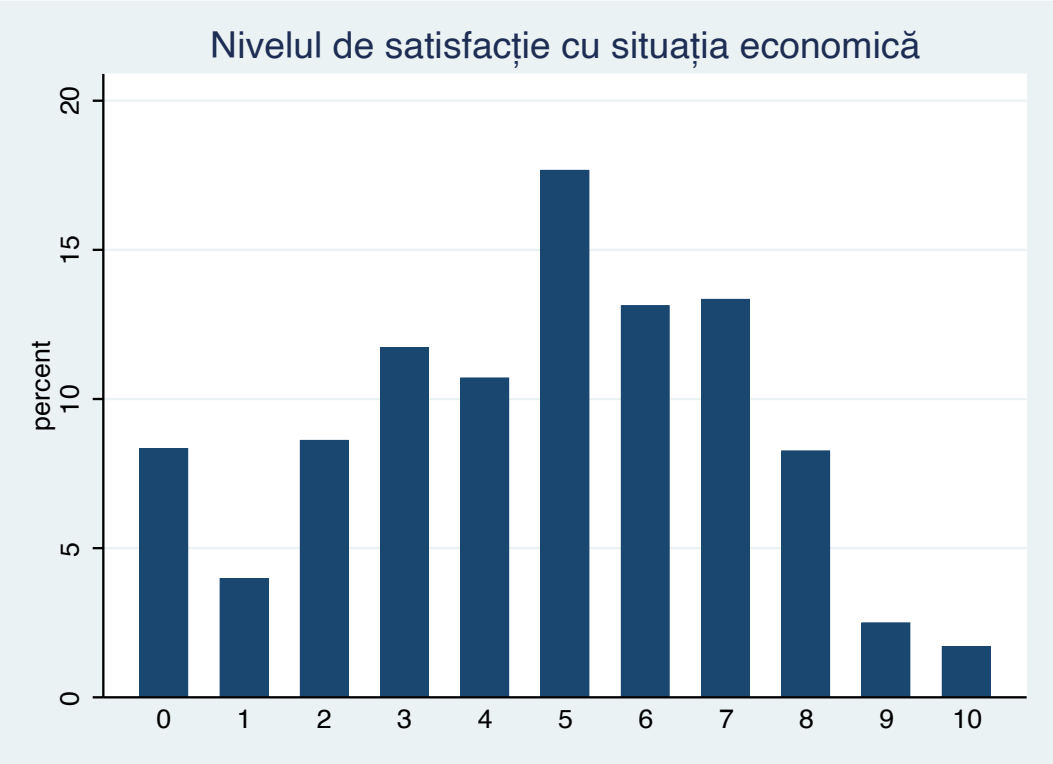
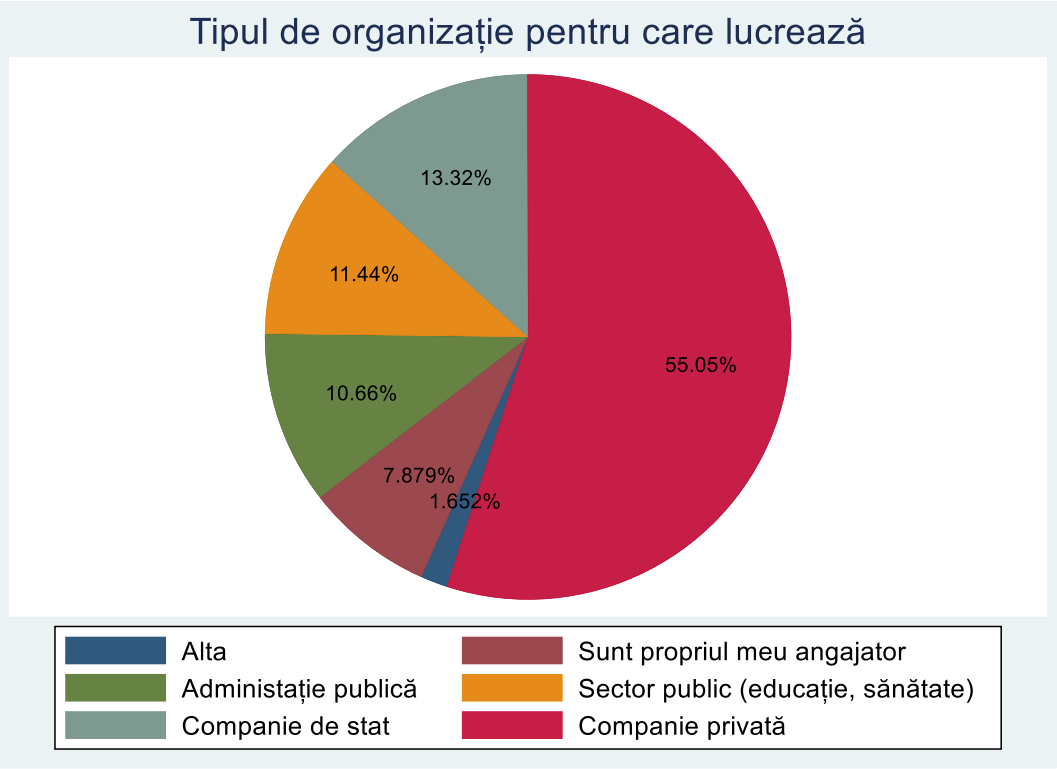


Diagrama circulară este deseori folosită pentru ilustrarea grafică a distribuției unor variabile calitative, măsurate la nivel nominal sau ordinal. Variabilele cantitative au, de regulă, mult prea multe categorii pentru a folosi diagrama circulară. Pentru Acest tip de vizualizare a datelor este indicat a fi folosit pentru variabile cu puține sau relativ puține categorii. Cu cât sunt mai multe categoriile variabilei, și cu cât categoriile au frecvențe de apariție mai redusă, cu atât crește dificultatea de înțelegere a informației transmisă prin acest tip de diagramă. Stata permite includerea unor informații suplimentare în graficele generate prin adăugarea unor instrucțiuni suplimentare în sintaxă. Spre exemplu să adăugăm procentul eșantionului din fiecare categorie la diagrama circulară. Vom folosi următoarea sintaxă pentru ilustrare, folosind variabila *tporgwk* din baza de date ESS:

```
graph pie, over(tporgwk) title (Tipul de organizație pentru care lucrează)
sort angle(252) plabel(_all percent)
```

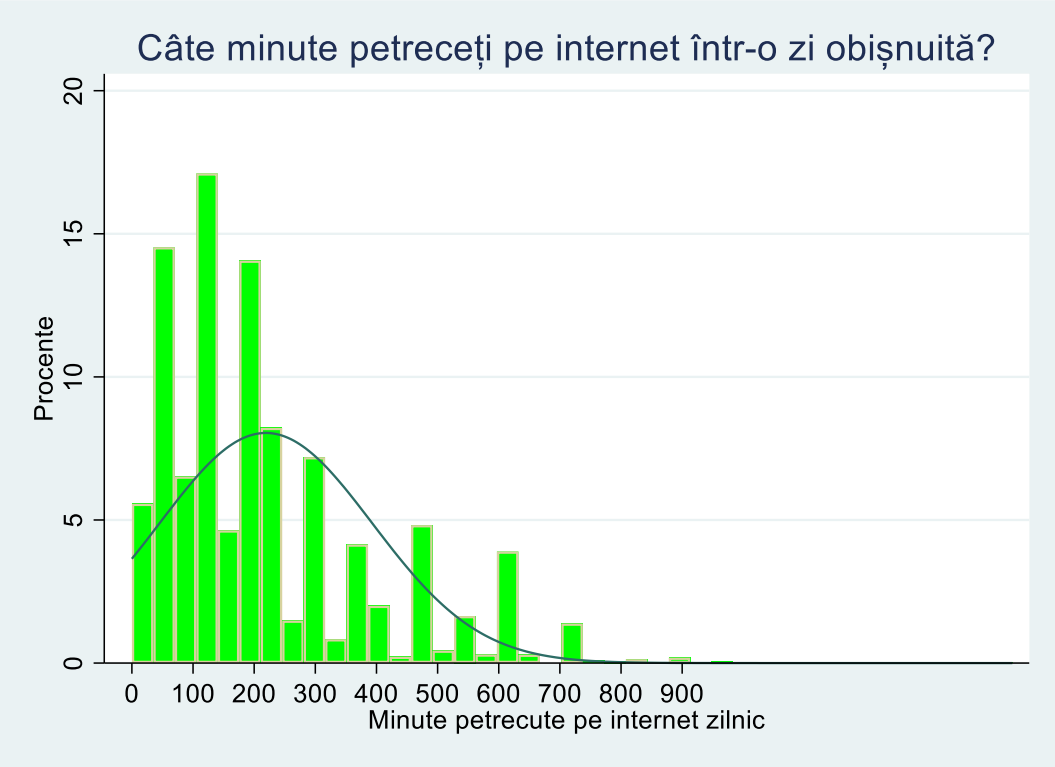
Figura 4.7 Analiză grafică de tip pie chart, în Stata



Variabilele cantitative pot fi reprezentate grafic folosind diagrame de tip histogramă, așa cum am discutat și în secțiunea 3.2. Putem modifica atât etichetele celor două axe, cât și raportarea fiecărei valori (de exemplu la frecvențe absolute sau relative), dar putem adăuga și linia de densitate (curba de normalitate). Exemplificăm producerea unei histograme în Stata, folosind variabila *netustm*, din baza de date ESS.

```
histogram netustm, percent fcolor(lime) normal ytitle(Procente)
xtitle(Minute petrecute pe internet zilnic) title(Câte minute petreceți
pe internet într-o zi obișnuită?)
```

Figura 4.8 Analiză grafică de histogramă, în Stata



4.5. Compararea a două eșantioane independente în Stata

Această secțiune își propune să prezinte modalitatea prin care putem compara mediile a două grupuri independente pentru a observa dacă acestea diferă. Pentru a calcula diferențele de medii dintre două grupuri putem folosi **testul t student** (Agresti și Finlay 2014, 197). Aplicarea testului t necesită existența unei variabile cu două grupuri (de exemplu, genul) și o variabilă de tip interval (încrederea în partidele politice – *trstprt*).

Raportat la setul de date analizat să presupunem că vrem să comparăm nivelul de încredere în partidele politice în funcție de gen. Vom folosi următoarele sintaxe pentru a găsi, examina, recodifica și verifica variabilele pe care le folosim în această analiză:

Utilizăm comanda **lookfor** pentru a identifica variabila gen:

```
lookfor gender
```

1. Analizăm categoriile și distribuția variabilei originale:

```
des gndr
label list gndr
tab gndr
```

2. Generăm, recodificăm și etichetăm noua variabilă:

```
gen gender = gndr
recode gender .a=.
lab var gender "Variabila binara"
lab define gender 1 "M" 2"F"
lab values gender gender
```

3. Verificăm distribuțiile celor două variabile (gndr și gender) utilizând o tabulare încrucișată:

```
tab gndr gender
```

4. Utilizăm comanda **lookfor** pentru a găsi variabila încredere în partidele politice:

```
lookfor trust in political parties
```

5. Analizăm categoriile și distribuția variabilei originale:

```
des trstprt
label list trstprt
tab trstprt
```

6. Generăm, recodificăm și etichetăm noua variabilă:

```
gen trust_parties = trstprt
```

```
recode trust_parties .a=. .b=. .c=.
lab var trust_parties "Increderea in partidele politice"
lab values trust_parties trust_parties
```

7. Verificăm distribuțiile celor două variabile (gndr și gender) utilizând o tabulare încrucișată:

```
tab1 trstprt trust_parties
```

În final, putem utiliza comanda **ttest** pentru a compara încrederea în partidele politice a celor două grupuri. Să reținem că opțiunea *by* indică variabila de utilizat pentru cele două grupuri; în cazul nostru: bărbați și femei. Rezultatul este prezentat în Figura 4.9.

```
ttest trust_parties, by(gender)
```

Figura 4.9 Tabel ieșire test t, în Stata

```
Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       M |   8,005    3.338788    .0276897    2.477414    3.284509    3.393067
       F |   9,789    3.415671    .0249974    2.473225    3.366671    3.464671
---------+--------------------------------------------------------------------
combined |  17,794    3.381084    .0185566    2.475336    3.344711    3.417456
---------+--------------------------------------------------------------------
    diff |           -.0768824    .0372977               -.1499894   -.0037754
------------------------------------------------------------------------------
    diff = mean(M) - mean(F)                                      t =  -2.0613
Ho: diff = 0                     degrees of freedom =    17792

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0196         Pr(|T| > |t|) = 0.0393          Pr(T > t) = 0.9804

.
```

4.6. Analiza univariată. Examinarea distribuțiilor de frecvențe și a statisticilor univariate în Stata

În această secțiune discutăm despre modul în care putem să examinăm distribuțiile de frecvențe și statisticile sumative pentru fiecare variabilă din setul de date. În Stata, statisticele univariate pot fi generate în mai multe moduri. O primă modalitate este utilizarea comenzii **tabulate**[48] care generează un tabel de frecvențe a variabilei de interes. În cazul nostru, vom aplica comanda **tabulate** pentru variabila *vote* după cum urmează:

```
tabulate vote
```

Rezultatul obținut este prezentat în Figura 4.10. Observăm astfel că un număr de 12,037 de respondenți au votat la ultimele alegeri în timp ce 4,684 nu au participat la vot și doar 1,155 de respondenți din eșantionul nostru nu sunt eligibili să voteze. De asemenea, observăm că tabelul generat de Stata ne indică și dimensiunea eșantionului (17,876), precum și frecvențele relative (%) nu doar frecvențele absolute (n).

[48] Comenziile în Stata pot fi prescurtate. Comanda **tabulate** poate fi prescurtată în **tab.**

Figura 4.10 Tabel de frecvențe pentru variabila vote

```
. tabulate vote
```

Voted last national election	Freq.	Percent	Cum.
Yes	12,037	67.34	67.34
No	4,684	26.20	93.54
Not eligible to vote	1,155	6.46	100.00
Total	17,876	100.00	

```
.
```

În Stata, putem, de asemenea, genera un tabel de frecvențe pentru mai multe variabile prin aplicarea unei singure comenzi (**tabulate**) la care adăugăm variabilele de interes:

```
tabulate vote stfeco stfdem
```

Similar programului RStudio, Stata generează și tabele pentru statistica descriptivă prin aplicarea următoarei sintaxe:

```
tabstat vote stfeco, statistics(mean n max min range sd)
columns(statistics)
```

Vom obține un rezultat similar cu cel din Figura 4.11. Observăm astfel că variabila *vote* are o valoare medie de 1.39 și o deviație standard de .606, în timp ce variabila *stfeco* are o valoare medie de 4.65 și o deviație standard de 2.488. Aceste valori ne oferă o imagine despre gradul de dispersie a fiecărei variabile analizată, însă nu ne permit o comparație, deoarece aceste variabile au „unități de măsură diferite". Pentru a putea trage concluzii comparative cu privire la gradul lor de omogenitate sau

eterogenitate va trebui să calculăm coeficientul de variație, împărțind valoarea deviației standard la media variabilei. Vom obține o valoare de 0.44 pentru variabila *vote*, respectiv 0.55 pentru variabila *stfeco.* Această a doua variabilă este prin urmare mai dispersată decât variabila *vote*. Putem aprecia, astfel, că are un grad mai ridicat de eterogenitate.

Figura 4.11 Statistică descriptivă

```
. tabstat vote stfeco, statistics(mean n max min range sd) columns(statistics)
```

variable	mean	N	max	min	range	sd
vote	1.391251	17876	3	1	2	.6061499
stfeco	4.654725	17641	10	0	10	2.488362

```
.
```

4.7. Analiza bivariată în Stata

4.7.1. Asocierea. Examinarea relațiilor bivariate pentru variabile nominale și/sau ordinale în Stata

Examinarea relației dintre două variabile nominale și/sau ordinale se realizează prin crearea unui tabel încrucișat, numit și tabel de contingență, prin calcularea coeficientului de asociere, precum și prin aplicarea statisticii inferențiale potrivite. Stata permite aplicarea acestor pași prin utilizarea unei singure linii de sintaxă.

Să presupunem că vrem să testăm, conform ipotezelor enunțate în capitolul 3, dacă respondenții cu un nivel mai mare de educație au mai puțină încredere în partidele politice. Primul pas constă în recodificarea variabilelor de interes. Astfel,

aplicat acestei ipoteze, vom recodifica variabila dependentă (DV), încrederea în partidele politice, într-o variabilă pe trei nivele, variind de la încredere mică la încredere mare, după cum urmează:

1. Utilizăm comanda **lookfor** pentru a identifica variabila *trstprt* (încrederea în partidele politice).

```
lookfor trust in political parties
```

2. Analizăm categoriile și distribuția variabilei originale:

```
des trstprt
label list trstprt
tab trstprt
```

3. Generăm, recodificăm și etichetăm noua variabilă:

```
gen trust_parties = trstprt
recode trust_parties .a=. .b=. .c=.
lab var trust_parties "Increderea in partidele politice"
lab values trust_parties trust_parties
```

4. Verificăm distribuțiile celor două variabile (*trstprt* și *trust_parties*) prin generarea unei tabel de contingență:

```
tab trstprt trust_parties, missing
```

Transformarea variabilei *trstprt* într-o variabilă pe 3 nivele se realizează prin generarea unui tabel de frecvențe pentru noua variabilă. Prin intermediul coloanei de frecvență cumulativă vom determina valorile pentru aproximativ 33%, 66% și 99% din eșantion[49]:

1. Generăm o distribuție de frecvențe pentru noua variabilă:

```
tab trstprt trust_parties
```

[49] O modalitate alternativă de recodificare a acestei variabile ar fi să setăm respondenții care aleg 0-4 ca încredere scăzută, 5 ca punct de mijloc și toate răspunsurile peste 5 ca încredere ridicată. Optăm să folosim punctele limită raportate în procentul cumulat pentru a produce trei grupuri de dimensiuni aproximativ egale.

2. Setăm cele trei nivele:

```
gen trust_parties_nivele = trust_parties
recode trust_parties_nivele 0/4=1 5=2 6/10 =3
lab define trust_parties_nivele 1 " Incredere scazuta" 2 "Incredere
medie" 3 "Incredere ridicata" , replace
lab values trust_parties_nivele trust_parties_nivele
```

3. Verificăm distribuțiile celor două variabile *(trust_parties, trust_parties_nivele*) prin realizarea unui tabel de contingență:

```
tab1 trust_parties trust_parties_nivele, missing
```

În continuare, vom recodifica variabila noastră independentă, *eisced* (nivelul de educație) pe trei nivele (scăzut, mediu și ridicat).

4. Utilizăm comanda **lookfor** pentru a identifica variabila *eisced:*

```
lookfor education
```

5. Analizăm categoriile și distribuția variabilei originale:

```
des eisced
label list eisced
tab eisced
```

6. Generăm, recodificăm și etichetăm noua variabilă:

```
gen education = eisced
recode education .a=. .b=. .c=.
lab var education "Nivelul de educatie"
lab values education education
```

7. Verificăm distribuțiile celor două variabile (*eisced* și *education*) prin generarea unui tabel de contingență:

```
tab eisced education, missing
```

Pentru a transforma această variabilă într-o variabilă pe 3 nivele, generăm o distribuție de frecvențe pentru noua variabilă și folosim coloana de frecvență cumulativă pentru a determina valorile pentru aproximativ 33%, 66% și 99% din eșantion[50]:

1. Generăm o distribuție de frecvențe pentru noua variabilă:

```
tab eisced education
```

2. Setăm cele trei nivele:

```
gen education_nivele = education
recode education_nivele 0/4=1 5=2 6/7=3
lab define education_nivele 1 "Primara" 2 "Secundara" 3 "Universitara"
, replace
lab values education_nivele education_nivele
```

3. Verificăm distribuțiile celor două variabile (*education, education_nivele*) prin generarea unui tabel de contingență:

```
tab2 education education_nivele, missing
```

Finalizarea etapei de generare și transformare a variabilelor de interes permite testarea relației dintre cele două prin aplicarea următoarei sintaxa în Stata:

```
tab trust_partiee_nivele education_nivele, col all
```

[50] O modalitate alternativă de recodificare a acestei variabile ar fi să setăm respondenții în funcție de numărul de ani petrecuți în sistemul formal de educație, de exemplu: 0 ani (fără școală), 4 ani (școala primară), 8 ani (gimnaziu), 12 ani (liceu), 16 ani (facultate 3 sau 4 ani). Pentru a face această transformare putem estima durata medie a fiecărui ciclu de învățământ formal, folosind suma acestor ani ca valoare pentru fiecare individ care a absolvit acel nivel specific de educație. Într-un eșantion este puțin probabil să avem un număr apreciabil de persoane repetente care au petrecut mai mulți ani într-un ciclu de învățământ, la fel cum e puțin probabil să avem un număr mare de persoane care au studiat în sistem intensiv (doi ani într-unul), pentru a avea distorsiuni în măsurarea acestei variabile.

4.7.2. Corelația. Examinarea relațiilor bivariate între variabilele de interval/rapoarte în Stata

În această secțiune vom evalua relația dintre două variabile continue (de tip interval/raport) pentru care folosim tehnici statistice diferite față de evaluarea relației dintre două variabile nominale și/sau ordinale.

Spre exemplu, pe baza setului nostru de date, am putea asuma faptul că respondenții mai în vârstă sunt mai puțini satisfăcuți de democrație decât respondenții mai tineri. Pentru a testa acest lucru, mai întâi recodificăm variabilele pentru analiză. Pentru a face acest lucru, aplicăm următoarea sintaxă:

1. Utilizăm comanda **lookfor** pentru a identifica variabila (*stfdem*) satisfacția față de democrație.

```
lookfor satisfaction democracy
```

2. Analizăm categoriile și distribuția variabilei originale:

```
des stfdem
label list stfdem
tab stfdem
```

3. Generăm, recodificăm și etichetăm noua variabilă:

```
gen democratie_satisfactie = stfdem
recode democratie_satisfactie .a=. .b=. .c=.
lab var democratie_satisfactie "Satisfactie cu democratia"
lab values democratie_satisfactie democratie_satisfactie
```

4. Verificăm distribuțiile celor două variabile (*trstprt* și *trust_parties*) prin generarea unui tabel de contingență:

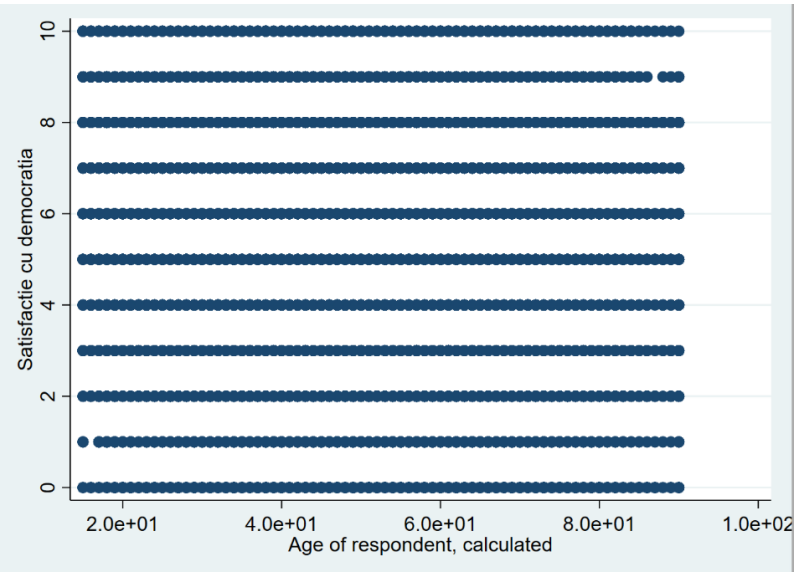
```
tab stfdem democratie_satisfactie, missing
```

Variabila care indică vârsta nu necesită modificare. Finalizarea procesului de generare și transformare a variabilelor permite generarea unei diagrame cu puncte (de dispersie) prin care putem inspecta vizual dacă pare să existe sau nu o relație liniară

între vârstă și satisfacția față de democrație. În Stata vom aplica următoarea sintaxă în baza căreia vom obține graficul din Figura 4.12.

```
twoway (scatter democratie_satisfactie agea)
```

Figura 4.12 Grafic de dispersie



Corelația dintre cele două variabile poate fi calculată prin calcularea coeficientului Pearson. În Stata, estimarea coeficientului se realizează prin utilizarea comenzii **pwcorr** la care se adaugă comanda **sig** astfel încât rezultatul să includă și nivelul de semnificație statistică a relației:

```
pwcorr democratie_satisfactie agea, sig
```

Figura 4.13 Corelația dintre stfedm și agea

```
. pwcorr democratie_satisfactie agea, sig

             | democr~e     agea
-------------+------------------
democratie~e |   1.0000
             |
             |
        agea |  -0.0528   1.0000
             |   0.0000
             |
```

Cum am putea interpreta rezultatele corelației de mai sus? Rezultatul include corelația dintre fiecare variabilă cu ea însăși (o valoare de 1 indică o corelare perfectă) și măsura asocierii dintre vârstă și nivelul de satisfacție cu democrație. Rezultatele indică o relație negativă, între vârstă și nivelul de satisfacție cu democrația (-0.0528): pe măsură ce vârsta crește, nivelul de satisfacție cu democrație scade.

În continuare, vom apela la statistica inferențială pentru a observa dacă această relație negativă este și semnificativă statistic. Valoarea 0.0000 aflată sub coeficientul de corelație indică probabilitatea inexistenței unei astfel de relații în eșantionul nostru. În acest caz, relația este semnificativă statistic la p<0,001.

4.7.3. Regresia simplă liniară în Stata

În timp ce informațiile obținute prin calcularea asocierii/corelației și a puterii relației dintre două variabile sunt informative, regresia liniară de bază oferă informații despre variabilele independente care pot fi utilizate pentru a face predicții despre variabila dependentă. În Stata, modele de regresie pot fi generate prin utilizarea comenzii **regress (reg)**.

În continuarea exemplului de la analiza asocierii și a corelației, vom estima modul în care nivelul de satisfacție cu democrația (*stfdem*) se schimbă pentru fiecare an în care crește vârsta:

```
reg democratie_satisfactie agea
```

Figura 4.14 Regresia liniară simplă

```
. reg democratie_satisfactie agea
```

Source	SS	df	MS	Number of obs	=	17,485
				F(1, 17483)	=	48.90
Model	351.520279	1	351.520279	Prob > F	=	0.0000
Residual	125673.647	17,483	7.18833422	R-squared	=	0.0028
				Adj R-squared	=	0.0027
Total	126025.168	17,484	7.20802834	Root MSE	=	2.6811

democratie~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agea	-.0077342	.001106	-6.99	0.000	-.0099021	-.0055663
_cons	5.306101	.0597804	88.76	0.000	5.188926	5.423277

4.7.4. Regresia multiliniară în Stata

Sintaxa pentru regresia multiliniară în Stata este aceeași cu cea pentru regresia liniară simplă, doar că adăugăm, în conformitate cu ipotezele studiului nostru, variabilele suplimentare independente și/sau de control în modelul de estimare a datelor. De exemplu, pe lângă vârstă, putem evalua modul în care încrederea în partidele politice, interesul pentru politică, nivelul de fericire influențează nivelul de satisfacție cu democrație:

```
reg democratie_satisfactie agea trstprt polintr clsprty happy
```

Figura 4.15 Regresia liniară multiplă

```
. reg democratie_satisfactie agea trstprt polintr clsprty happy
```

Source	SS	df	MS	Number of obs	=	16,902
				F(5, 16896)	=	1935.14
Model	44309.8655	5	8861.97311	Prob > F	=	0.0000
Residual	77375.1637	16,896	4.57949596	R-squared	=	0.3641
				Adj R-squared	=	0.3639
Total	121685.029	16,901	7.19987156	Root MSE	=	2.14

democratie~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agea	-.0016917	.0009398	-1.80	0.072	-.0035337	.0001504
trstprt	.5931107	.0069637	85.17	0.000	.5794612	.6067602
polintr	.0161467	.0206575	0.78	0.434	-.024344	.0566375
clsprty	.0389879	.0361349	1.08	0.281	-.0318404	.1098161
happy	.2203106	.0083565	26.36	0.000	.203931	.2366902
_cons	1.295981	.1199459	10.80	0.000	1.060875	1.531088

.

4.7.5. Regresia logistică în Stata

Să presupunem că suntem interesați să investigăm dacă participarea la vot este influențată de interesul în politică, apropierea față de un partid, încrederea în partidele politice, încrederea în politicieni, nivelul de satisfacție cu economia și nivelul de fericire raportat. Astfel, variabila dependentă este participarea la vot (*vote*) care este o variabilă binară unde 1 este Da și 0 este Nu.

În primul rând vom transforma variabila dependentă după cum urmează:

1. Analizăm categoriile și distribuția variabilei originale:

```
des vote
label list vote
tab vote
```

2. Generăm, recodificăm și etichetăm noua variabilă:

```
gen vot = vote
recode vot .a=. .b=. .c=. 3=. 2=0
lab var vot "Participarea la vot"
lab values vot vot
```

3. Verificăm distribuțiile celor două variabile (*trstprt* și *trust_parties*) prin generarea unui tabel de contingență:

```
tab vote vot, missing
```

În al doilea rând, pentru a realiza o regresie logistică în Stata, vom aplica următoarea sintaxă:

```
logit vot polintr clsprty trstprt trstplt stfeco happy
```

Figura 4.16 Regresia logistică

```
. logit vot i.polintr clsprty trstprt trstplt stfeco happy

Iteration 0:   log likelihood = -9373.3723
Iteration 1:   log likelihood = -8023.4829
Iteration 2:   log likelihood =  -7947.189
Iteration 3:   log likelihood =  -7946.952
Iteration 4:   log likelihood =  -7946.952

Logistic regression                             Number of obs     =     15,855
                                                LR chi2(8)        =    2852.84
                                                Prob > chi2       =     0.0000
Log likelihood =  -7946.952                     Pseudo R2         =     0.1522

------------------------------------------------------------------------------
         vot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      polintr |
Quite interested |  -.0345951   .1012438    -0.34   0.733    -.2330293    .1638392
Hardly interested |  -.5940424   .0974746    -6.09   0.000     -.785089   -.4029957
Not at all interested |  -1.369806   .1005157   -13.63   0.000    -1.566814   -1.172799
             |
      clsprty |  -1.435817   .0469881   -30.56   0.000    -1.527912   -1.343722
      trstprt |   .0286345   .0169434     1.69   0.091     -.004574     .061843
      trstplt |     .03741   .0167589     2.23   0.026     .0045632    .0702568
       stfeco |  -.0076339   .0096801    -0.79   0.430    -.0266065    .0113386
        happy |   .0836362   .0096523     8.66   0.000     .0647181    .1025543
        _cons |   3.188179    .134718    23.67   0.000     2.924137    3.452221
------------------------------------------------------------------------------
```

Putem, de asemenea, să calculăm valoarea exponențială a coeficienților și să îi interpretăm ca probabilități *odds ratio*. În Stata putem obține coeficienții sub formă de *odds ratio* prin aplicarea următoarei sintaxe:

```
logit , or
```

Figura 4.17 Regresia logistică exprimată în odds ratio

```
. logit , or

Logistic regression                             Number of obs     =     15,855
                                                LR chi2(8)        =    2852.84
                                                Prob > chi2       =     0.0000
Log likelihood =  -7946.952                     Pseudo R2         =     0.1522
```

vot	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
polintr						
Quite interested	.9659965	.0978012	-0.34	0.733	.7921303	1.178025
Hardly interested	.552091	.0538148	-6.09	0.000	.4560791	.6683149
Not at all interested	.2541562	.0255467	-13.63	0.000	.2087092	.3094994
clsprty	.2379209	.0111795	-30.56	0.000	.2169882	.2608729
trstprt	1.029048	.0174356	1.69	0.091	.9954364	1.063795
trstplt	1.038119	.0173977	2.23	0.026	1.004574	1.072784
stfeco	.9923951	.0096064	-0.79	0.430	.9737443	1.011403
happy	1.087233	.0104943	8.66	0.000	1.066858	1.107997
_cons	24.24424	3.266135	23.67	0.000	18.61814	31.57044

```
Note: _cons estimates baseline odds.
```

În Stata putem calcula probabilitățile prezise pentru a înțelege modelul, folosind comanda **margins**.

```
margins polintr, atmeans
```

Figura 4.18 Regresia logistică cu probabilități prezise

```
. margins polintr, atmeans

Adjusted predictions                            Number of obs     =     15,855
Model VCE    : OIM

Expression   : Pr(vot), predict()
at           : 1.polintr       =    .0773888 (mean)
               2.polintr       =    .3176916 (mean)
               3.polintr       =     .396468 (mean)
               4.polintr       =    .2084516 (mean)
               clsprty         =      1.5655 (mean)
               trstprt         =    3.373258 (mean)
               trstplt         =    3.437023 (mean)
               stfeco          =    4.620624 (mean)
               happy           =    7.128098 (mean)
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
polintr						
Very interested	.8489605	.0118105	71.88	0.000	.8258124	.8721086
Quite interested	.8444707	.0054965	153.64	0.000	.8336979	.8552436
Hardly interested	.7562869	.0057175	132.28	0.000	.7450807	.767493
Not at all interested	.5882329	.0095425	61.64	0.000	.56953	.6069358

De asemenea, s-ar putea să dorim să observăm cât de bine estimează modelul nostru datele. Acest lucru poate fi deosebit de util atunci când comparăm mai multe modele. Comanda **fitstat** (Long și Freese 2000) produce o varietate de statistici de estimare[51]:

```
fitstat
```

[51] Comanda fitstat trebuie instalată prin tastarea comenzii ssc install fitstat în câmpul de comandă din Stata.

Figura 4.19 Regresia logistică – estimarea modelului

```
.  search fitstat

. fitstat

Measures of Fit for logit of vot

Log-Lik Intercept Only:    -9373.372     Log-Lik Full Model:        -7969.009
D(15848):                  15938.018     LR(6):                      2808.727
                                         Prob > LR:                     0.000
McFadden's R2:                 0.150     McFadden's Adj R2:             0.149
Maximum Likelihood R2:         0.162     Cragg & Uhler's R2:            0.234
McKelvey and Zavoina's R2:     0.267     Efron's R2:                    0.173
Variance of y*:                4.486     Variance of error:             3.290
Count R2:                      0.754     Adj Count R2:                  0.115
AIC:                           1.006     AIC*n:                     15952.018
BIC:                    -137331.797      BIC':                      -2750.700
```

4.7.6. Regresia multinomială în Stata

Regresia logistică multinomială este utilizată pentru a modela variabilele dependente măsurate pe nivel nominal. Să presupunem că vrem să observăm dacă opțiunea pentru locul de muncă este influențată de nivelul de educație și nivelul de educație al tatălui. Astfel, primul pas este să transformăm variabilele de interes în variabile dihotomice:

1. Analizăm categoriile și distribuția variabilei originale:

```
des tporgwk
label list tporgwk
tab tporgwk
des eiscedf
label list eiscedf
tab eiscedf
des eisced
label list eisced
```

```
tab eisced
```

2. Generăm, recodificăm și etichetăm noua variabilă:

```
gen ocupatie = tporgwk
recode ocupatie .a=. .b=. .c=. 3=. .d=. 6=.
recode ocupatie 1/2=1 3/4=2 5=3
lab define ocupatie 1 "Sector public" 2 "Sector privat" 3 "Antreprenor"
, replace
lab var ocupatie "Ocupatie"
lab values ocupatie ocupatie
```

```
gen educatie_tata = eiscedf
recode educatie_tata 0/4=1 5=2 6/7=3
lab define educatie_tata 1 "Primara" 2 "Secundara" 3 "Universitara" ,
replace
lab values educatie_tata educatie_tata
```

```
gen educatie_respondent = eisced
recode educatie_respondent 0/4=1 5=2 6/7=3
lab define educatie_respondent 1 "Primara" 2 "Secundara" 3
"Universitara" , replace
lab values educatie_respondent educatie_respondent
```

3. Verificăm distribuțiile celor două variabilele prin generarea unui tabel de contingență:

```
tab tporgwk ocupatie, missing
tab eiscedf educatie_tata, missing
tab eisced educatie_respondent, missing
```

Al doilea pas constă în obținerea unor statistici descriptive ale variabilelor de interes:

```
tab ocupatie educatie_respondent, chi2
tab ocupatie educatie_tata, chi2
table ocupatie, con(mean educatie_respondent sd educatie_respondent)
table ocupatie, con(mean educatie_tata sd educatie_tata)
```

Mai jos folosim comanda **mlogit** pentru a estima un model de regresie logistică multinomială. Am folosit, opțiunea „base" pentru a indica categoria pe care am dori să o folosim pentru grupul de comparație de referință. În modelul de mai jos, am ales să folosim ocupația *antreprenor* drept categorie de referință.

```
mlogit ocupatie educatie_tata educatie_respondent, base(3)
```

Figura 4.20 Regresia logistică multinomială

```
. mlogit ocupatie educatie_tata educatie_respondent, base(3)

Iteration 0:   log likelihood = -11109.273
Iteration 1:   log likelihood = -11029.571
Iteration 2:   log likelihood = -11021.877
Iteration 3:   log likelihood = -11021.744
Iteration 4:   log likelihood = -11021.744

Multinomial logistic regression                 Number of obs     =     12,972
                                                LR chi2(4)        =     175.06
                                                Prob > chi2       =     0.0000
Log likelihood = -11021.744                     Pseudo R2         =     0.0079

------------------------------------------------------------------------------
    ocupatie |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Sector_public |
      educatie_tata |  -.0097515   .0123258    -0.79   0.429    -.0339096    .0144066
educatie_respondent |   .0644669   .0294274     2.19   0.028     .0067903    .1221435
              _cons |   .9476853   .0631585    15.00   0.000      .823897    1.071474
-------------+----------------------------------------------------------------
Sector_privat |
      educatie_tata |  -.0153587   .0114172    -1.35   0.179     -.037736    .0070186
educatie_respondent |  -.1925801   .0300757    -6.40   0.000    -.2515273   -.1336328
              _cons |   2.286565   .0616254    37.10   0.000     2.165781    2.407348
-------------+----------------------------------------------------------------
Antreprenor  |  (base outcome)
------------------------------------------------------------------------------
```

Raportul dintre probabilitatea de a alege o categorie de rezultat și probabilitatea de a alege categoria de referință este adesea numită risc relativ. Riscul relativ poate fi obținut prin exponențierea ecuațiilor liniare de mai sus, obținând coeficienți de regresie care sunt rapoarte de risc relative pentru o modificare unitară a variabilei

predictoare. Putem folosi opțiunea **rrr** pentru comanda **mlogit** pentru a afișa rezultatele regresiei în termeni de raporturi de risc relative.

```
mlogit, rrr
```

Figura 4.21 Rezultatele regresiei multinomiale în termeni de raport de risc relativ

```
. mlogit, rrr

Multinomial logistic regression                 Number of obs     =     12,972
                                                LR chi2(4)        =     175.06
                                                Prob > chi2       =     0.0000
Log likelihood = -11021.744                     Pseudo R2         =     0.0079
```

ocupatie	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Sector_public						
educatie_tata	.9902959	.0122062	-0.79	0.429	.9666589	1.014511
educatie_respondent	1.06659	.0313869	2.19	0.028	1.006813	1.129916
_cons	2.579732	.1629319	15.00	0.000	2.279365	2.919679
Sector_privat						
educatie_tata	.9847586	.0112432	-1.35	0.179	.9629672	1.007043
educatie_respondent	.8248283	.0248073	-6.40	0.000	.7776122	.8749113
_cons	9.841073	.6064603	37.10	0.000	8.721412	11.10448
Antreprenor	(base outcome)					

```
Note: _cons estimates baseline relative risk for each outcome.
```

De asemenea, s-ar putea să dorim să observăm cât de bine sunt estimate datele prin modelul nostru. Acest lucru poate fi deosebit de util atunci când comparăm mai multe modele. La fel ca mai sus, vom folosi comanda **fitstat** pe care am instalat-o deja în Stata prin comanda ssc install fitstat:

```
fitstat
```

Figura 4.22 Estimarea modelului

```
Note: _cons estimates baseline relative risk for each outcome.

. fitstat

Measures of Fit for mlogit of ocupatie

Log-Lik Intercept Only:   -11109.273     Log-Lik Full Model:       -11021.744
D(12963):                  22043.487     LR(4):                      175.058
                                         Prob > LR:                    0.000
McFadden's R2:                 0.008     McFadden's Adj R2:            0.007
Maximum Likelihood R2:         0.013     Cragg & Uhler's R2:           0.016
Count R2:                      0.642     Adj Count R2:                -0.001
AIC:                           1.701     AIC*n:                   22061.487
BIC:                    -100723.232      BIC':                     -137.176
```

5. Aplicații practice în Excel

Similar aplicațiilor practice exemplificate în RStudio și Stata, în acest capitol vom prezinta modalitatea de realizare a acestor aplicații în Microsoft Office Excel. Excel este un program statistic foarte puternic. Ca să citățm o glumă care circulă între statisticieni: „Microsoft pare că nu știe cât de puternic din punct de vedere statistic este programul său, Excel, de aceea nu îl dezvoltă". Ca dovadă, acest program este folosit de foarte mulți angajați din companii private sau instituții ale administrației publice, de la cele mai elementare gestionări și analize de date, la vizualizări grafice, și până la cele mai complicate formule.

5.1. Funcții de bază și curățarea setului de date în Excel

Setul de date folosit este ESS10.csv pe care l-am descărcat inițial de pe pagina European Social Survey (ESS) - https://ess-search.nsd.no/. Ulterior, am transformat extensia .csv în .xlsx și am păstrat din ESS10.xls doar acele variabile de interes pentru analiza noastră pe care le-am salvat într-un nou fișier .xlsx numit dataset_exemplu.xlsx.[52]

În continuare, vom exemplifica cele mai utile funcții disponibile în Excel care vor ușura procesul de familiarizare cu setul nostru de date. Aceasta nu reprezintă o listă exhaustivă a funcțiilor Excel, ci o listă a funcțiilor care sunt cel mai des folosite de analiști.

Una dintre cele mai importante beneficii pentru care Excel este folosit este capacitatea acestuia de a opera ca un calculator foarte performant. În Excel putem realiza o serie de calcule, fie simple fie foarte complexe, într-un timp foarte scurt. În

[52] Variabilele care fac parte din setul nostru de date numit dataset_exemplu.xlsx sunt următoarele: idno, cntry, vote, polintr, clsprty, mbtru, trstprt, trstplt, trstprl, trstsci, trstlgl, lrscale, stflife, stfeco, stfgov, stfedu, stfdem, stfhlth, stfmjob, happy, health, eisced, c19whome, gvhanc19, respc19, agea, gndr.

Tabelele 1 și 2 am expus câteva dintre cele mai importante funcții de calcul matematic utile în administrarea și analiza datelor în Excel.

Tabel 5.1 Funcții care operează adunări

Funcție	Folosită pentru...	Structura funcției	Exemplu practic
SUM	A aduna valorile dintr-o serie de celule	=SUM(seria de celule)	=SUM(A2:A37)
SUMIF	A aduna valorile dintr-o serie doar dacă celulele îndeplinesc o anumită condiție	=SUMIF(seria de celule, condiția de îndeplinit, seria de celule adăugată)	=SUMIF(D2:D37, "1", A2:A37) Adună toate celulele din coloana A, dacă corespondentul din coloana D conține valoarea 1.
SUMIFS	A aduna valorile dintr-o serie doar dacă celule îndeplinesc mai multe condiții	=SUMIFS(seria de celule adunate, seria de celule pentru prima condiție, prima condiție, seria celulelor pentru a doua condiție, a doua condiție de îndeplinit)	=SUMIFS(A2:A37, D2:D37, "1", I2:I37, ">2") Adună toate celulele din coloana A, dacă coloana D conține valoarea 1 și coloana I are valori mai mari de 2.

Tabel 5.2 Funcții care operează numărări

Funcție	Folosită pentru...	Structura funcției	Exemplu teoretic	Exemplu practic
COUNT	A număra celulele numerice	=COUNT(seria de celule de numărat)	Vrem să vedem câte celule sunt pentru variabila *vote* între	=COUNT(C2:C37)

Funcție	Folosită pentru...	Structura funcției	Exemplu teoretic	Exemplu practic
			respondentul 2 și respondentul 37.	
COUNTIF	A număra celulele	=COUNTIF(seria de celule, valoarea pe care vrem să o numărăm din acea serie de celule)	Vrem să vedem într-o serie de celule de câte ori apare țara BG.	=COUNTIF(C2:C1435, "BG")
COUNTBLANK	A număra celulele care nu au valori (în engleză *empty*)	=COUNTBLANK(seria de celule)	Să presupunem că vrem să aflăm dacă există celule goale pentru variabila *polintr* (coloana D) între respondenții de la rândurile 1 până la 400.	=COUNT(D1:D400)
COUNTIFS	A număra valorile dintr-o serie doar dacă celule îndeplinesc mai multe condiții	=COUNTIFS(seria de celule numărate, seria de celule pentru prima condiție, prima condiție, seria celulelor pentru a doua condiție, a doua condiție de îndeplinit)	Adună toate celulele din coloana A, dacă coloana D conține valoarea 1 și coloana I are valori mai mari de 2.	=COUNTIFS(A2:A37, D2:D37, "1", I2:I37, ">2")

De asemenea, cu ajutorul programului Excel putem opera o serie de funcții logice și de căutare care sunt extrem de utile atunci când operăm cu volume mari de date. În tabelele de mai jos prezentăm funcțiile de bază care ne permit manevrarea unor seturi mari de date. Fără aceste funcții este dificil de operat cu funcții complexe.

Tabel 5.3 Funcții logice

Funcție	Folosită pentru...	Structura funcției	Exemplu teoretic	Exemplu practic
AND	A testa dacă mai mult de o condiție este adevărată. Rezultatul este TRUE doar dacă toate condițiile sunt îndeplinite	=AND(condiția de testat 1, condiția de testat 2,...)	Să presupunem că vrem să observăm care sunt acei respondenți care au votat la ultimele alegeri (vote =1) și au și un interes în politică (polintr = 1). Pentru a calcula pentru întreaga coloană trebuie să dăm dublu click în colțul din dreapta jos al primei celule în care am inserat funcția.	=AND(C2=1, D2=1)
OR	A testa dacă o condiție este adevărată. Rezultatul este TRUE dacă oricare dintre condiții este îndeplinită	=OR(condiția de testat 1, condiția de testat 2,...)		=OR(C2=1, D2=1)

Tabel 5.4 Funcții de căutare

Funcție	Folosită pentru...	Structura funcției	Exemplu practic
VLOOKUP	Caută o valoare într-un tabel și scoate datele căutate într-o coloană	=VLOOKUP(valoare căutată, tabel în care să caute, numărul coloanei din tabel în care să scoată informația, potrivire aproximată sau exactă)	=VLOOKUP(E2,C2:D300,2,FALSE)
HLOOKUP	Caută o valoare într-un tabel și scoate datele căutate într-un rând	=HLOOKUP(valoare căutată, tabel în care să caute, rândul din tabel în care să scoată informația, potrivire aproximată sau exactă)	=HLOOKUP(E2,C2:F30,2,FALSE)
XLOOKUP	Caută într-o serie sau într-o matrice, apoi returnează elementul corespunzător primei potriviri pe care o găsește. Dacă nu există nicio potrivire, atunci XLOOKUP poate returna cea mai apropiată potrivire (aproximativă).	=XLOOKUP(valoare căutată, matrice de căutare, matricea de rezultat, [if_not_found], [match_mode], [search_mode])	=XLOOKUP(B2,B2:B1435, C2:C1435)

5.1.1. Pregătirea și curățarea setului de date în Excel

Realizarea unor analize complexe în Excel necesită ca valorile nule din setul de date utilizat să fie eliminate. Aplicând funcția **filter** vom observa că, pe lângă categoriile variabilelor (de obicei de la 0 la 10), avem și valori de 55, 66, 77, 88 și 99. De regulă, aceste valori reprezintă fie respondenți care au trecut peste întrebare, care au răspuns cu *nu știu* sau au răspuns cu *nu răspund*. Rămase în setul de date și introduse în analize, aceste valori pot schimba semnificativ rezultatele noastre.

5.1.2. Numirea unei variabile în Excel

Numirea unei variabile în Excel, precum și schimbarea numelui prestabilit acesteia de program (coloana A) se realizează prin aplicarea următorilor pași. Să presupunem că pentru coloana A din setul nostru de date dorim să schimbăm numele variabilei din A în numele *idno:*

1. Vom selecta celula A1 și vom apăsa simultan **Ctrl + Shitf + Arrow Down** pentru a selecta toate celulele din acea coloană;
2. Vom da click dreapta și vom selecta **Define Name**, iar la opțiunea **Name** vom scrie numele dorit, în cazul nostru *idno;*
3. Click **OK**

5.1.3. Mărimea eșantionului în Excel

Identificarea dimensiunii eșantionului cu care lucrăm este un pas extrem de important indiferent de tipul de analiză pe care vrem să-l aplicăm. În Excel putem calcula mărimea eșantionului cu formula **=count** pe care o vom aplica variabilei *idno* care indică numărul de identificare a fiecărui respondent din eșantion.

```
=COUNT(idno)
```

5.1.4. Tabel pivot în Excel

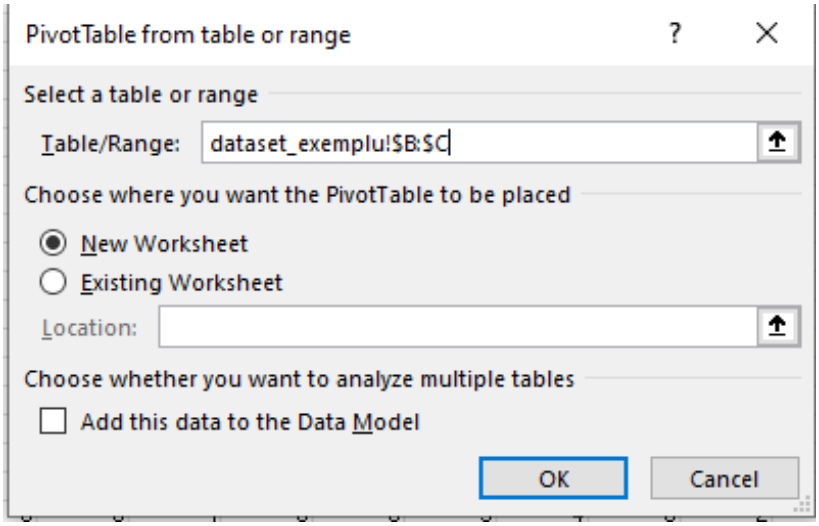
Tabelul pivot reprezintă acea metodă prin care putem obține în Excel un tabel de contingență sau putem să reorganizăm baza de date. Astfel, un tabel pivot este un instrument puternic pentru a calcula, rezuma și analiza date, a realiza comparații, modele și tendințe ale datelor.

Să presupunem că vrem să aflăm câți respondenți au participat la vot în fiecare țară. Cele două variabile de interes sunt *cntry* (variabila care indică țara

respondenților) și *vote* (variabilă cu categorii 1 și 0, unde 1 indică dacă respondentul a votat sau nu). Pentru a realiza un tabel pivot în Excel vom parcurge următorii pași:

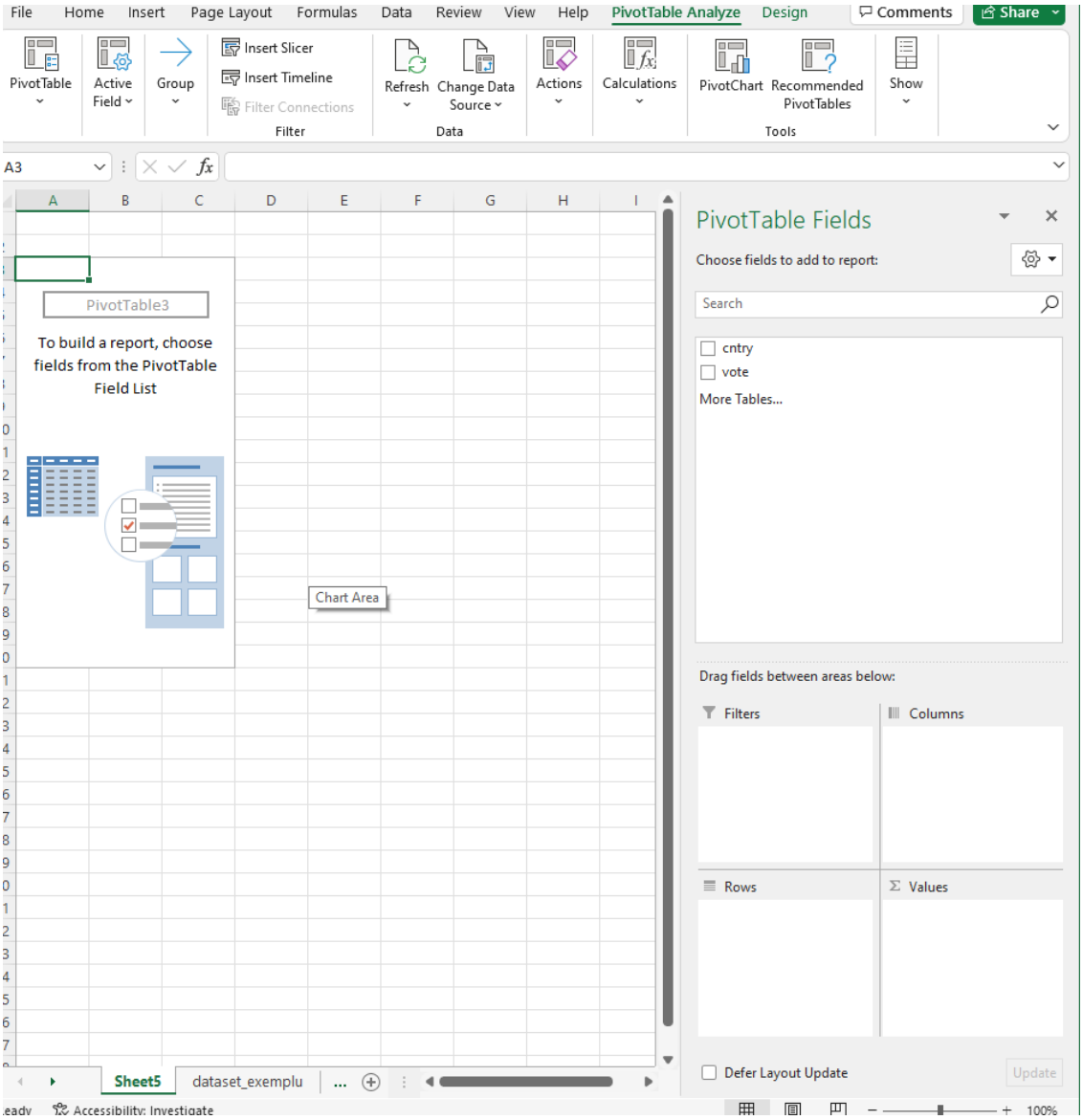
1. Selectăm coloanele de interes.
2. Selectăm Insert -> Pivot Table
3. În caseta Pivot Table introducem la Table/Range coloanele de interes (cntry și *agea* în cazul nostru)
4. Bifăm New Worksheet

Figura 5.1 Tabel pivot selecție



5. Click OK
6. Rezultatul este structura necompletată a tabelului pivot pe o nouă foaie de lucru.

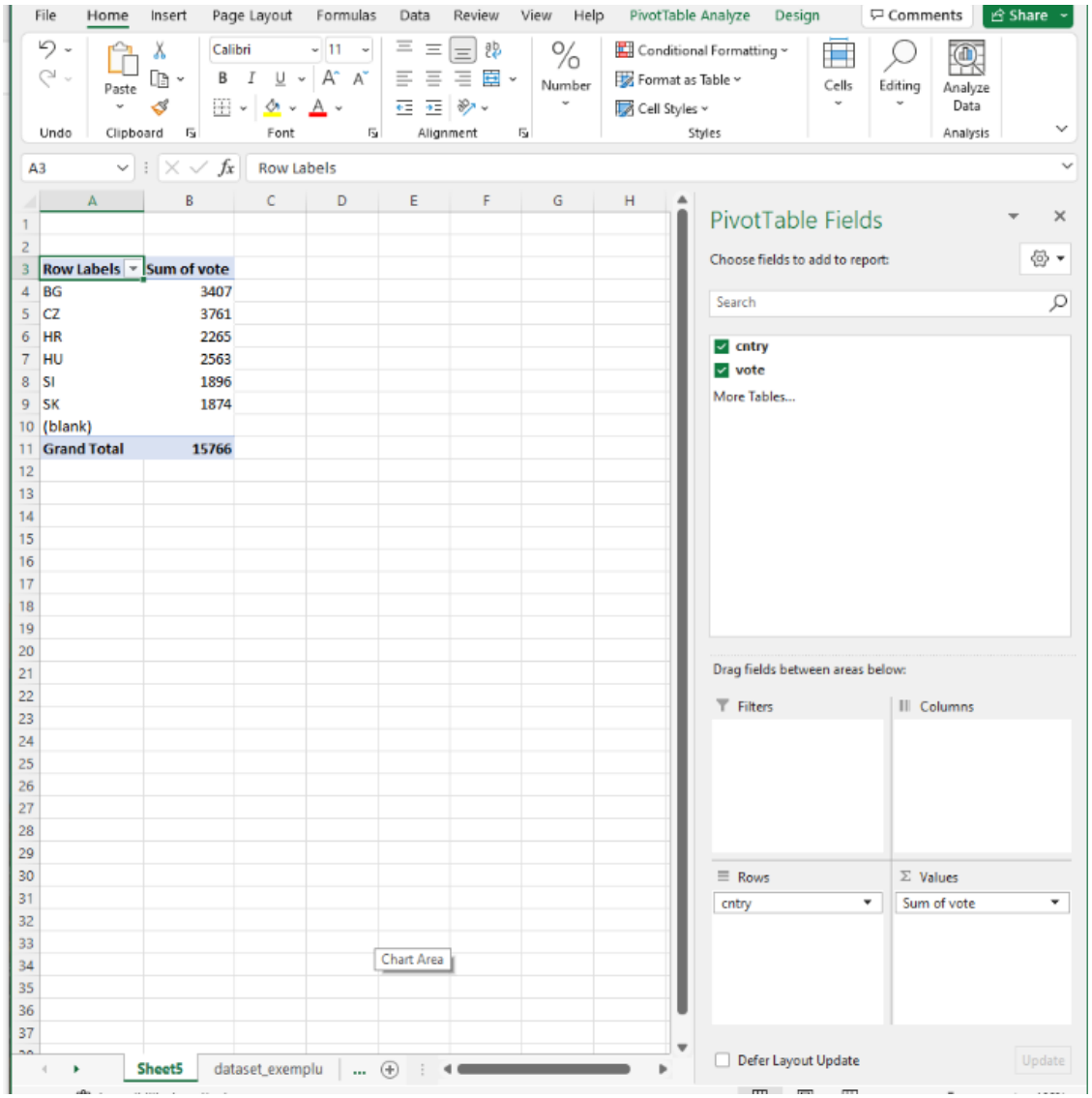
Figura 5.2 Tabel pivot câmpuri



7. Pentru a completa tabelul pivot, selectăm un câmp din Pivot Table Fields și îl introducem în caseta corespunzătoare de mai jos. În cazul nostru am selectat *cntry* pentru a fi dispus pe rânduri și *vote* am plasat-o în chenarul Values.

Rezultatul este următorul:

Figura 5.3 Tabel pivot specificare câmpuri de analiză de frecvențe



În urma inserării variabilelor de interes în chenarul corespunzător al tabelului pivot vom obține un rezultat similar cu cel de mai jos:

Figura 5.4 Tabel pivot frecvențe

Row Labels	Sum of vote
BG	3407
CZ	3761
HR	2265
HU	2563
SI	1896
SK	1874
(blank)	
Grand Total	**15766**

Prin urmare, importanța tabelelor pivot constă în faptul că permite o sintetizare rapidă a datelor de interes. Prin intermediul tabelelor pivot putem observa relații între date și putem efectua analize care altfel ar putea rămâne neidentificate dacă analizăm doar baza de date brută.

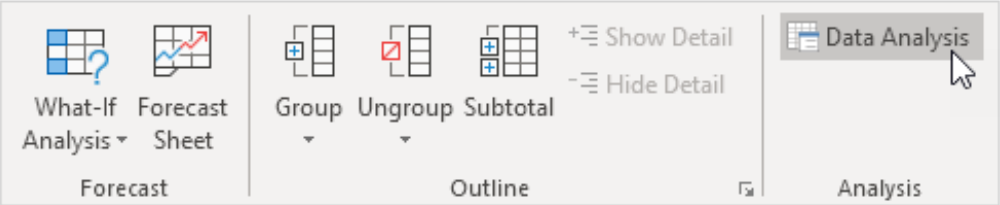
5.2. Analiza univariată în Excel

Excel are posibilitatea de a adăuga o serie de extensii extrem de utile în procesul de analiză a datelor. Una dintre aceste extensii este **Analysis Toolpak** care ne permite să generăm statistici descriptive pentru datele noastre. Să presupunem că vrem să generăm statistici descriptive pentru variabilele *trstprt*, *stfeco* și *happy*. O serie de pași trebuie făcuți pentru a obține statisticile descriptive pentru aceste variabile:

1. La **Data *tab***, în grupul **Analysis** și selectăm **Data Analysis**[53]

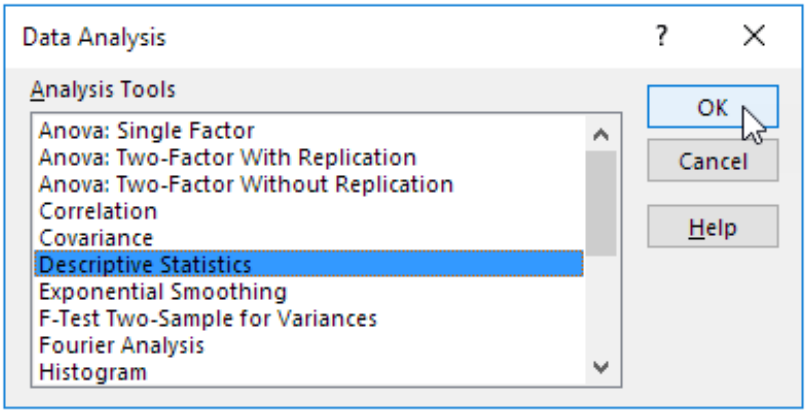
[53] Dacă Data Analysis nu apare în lista din meniu, trebuie să instalăm extensia. Mergem la File -> Options -> Add – ins -> Manage: Excel Add-ins -> selectăm Go -> în chenar selectăm Analysis ToolPak -> OK.

Figura 5.5 Tabel pivot data analysis



2. Selectăm **Descriptive Statistics**

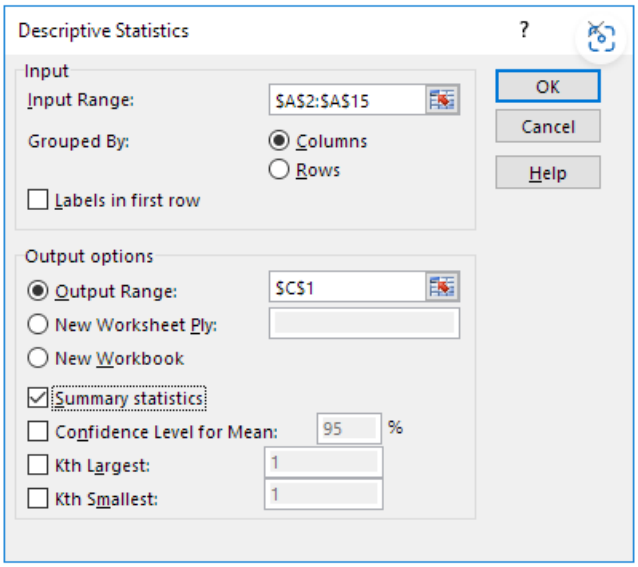
Figura 5.6 Tabel pivot descriptive statistics



3. Selectăm seria celulelor pentru care vrem să se calculeze statistica descriptivă, selectăm o celulă goală unde dorim să fie inserat rezultatul și bifăm **Summary Statistics**[54]

[54] Pentru a obține statistica descriptivă pentru mai multe variabile în același timp selectăm seria celulelor pentru toate variabile de interes.

Figura 5.7 Tabel pivot summary statistics



4. Selectăm OK

În urma efectuării pașilor prezentați mai sus vom obține următorul tabel:

Tabel 5.5 Statistică descriptivă pentru variabila trstprt (încredere în partide)

	trstprt
Mean	3.098439485
Standard Error	0.024046515
Median	3
Mode	0
Standard Deviation	2.52454053
Sample Variance	6.373304887
Kurtosis	-0.542749594
Skewness	0.49954449
Range	10
Minimum	0
Maximum	10
Sum	34151
Count	11022

5.2.1. Tendința centrală a datelor în Excel

Această secțiune tratează modul în care putem utiliza Excel pentru analiza descriptivă a datelor. Vom exemplifica pe setul nostru de date cum putem calcula media, abaterea standard, eroarea standard a mediei și altele. Toți acești indicatori ai analizei descriptive sunt parte din prima etapă în realizarea analizelor statistice complexe. Analiza descriptivă oferă, în acest context, o înțelegere preliminară a datelor noastre și a ceea ce ne permit acestea să realizăm. Așa cum am discutat în capitolul 3, analiza descriptivă stă la baza analizei inferențiale care explică relațiile dintre date și testează ipoteze bazate pe relația dintre cauze și efect.

5.2.1.1. Valoarea modală în Excel

Alături de valoarea medie și valoarea mediană, valoarea modală este o altă măsură prin care putem calcula tendința centrală a datelor. Ne reamintim că valoarea modală reprezintă individul tipic, are valoarea care apare cel mai frecvent într-o serie de date. În Excel, putem calcula valoarea modală a răspunsurilor pentru variabila *trstprt* prin inserarea în aceeași foaie de calcul în care am calculat celelalte valori centrale a următoarei formule:

```
=MODE(dataset_exemplu!J2:J11023)
```

5.2.1.2. Valoarea mediană în Excel

Mediana reprezintă valoarea de mijloc a distribuției datelor noastre. În Excel, putem calcula valoarea mediană a răspunsurilor pentru variabila *trstprt* prin inserarea în aceeași foaie de calcul în care am calculat valoarea medie a următoarei formule:

```
=MEDIAN(dataset_exemplu!J2:J11023)
```

5.2.1.3. Valoarea medie în Excel

În baza setului nostru de date, să presupunem că ne raportăm la variabile *trstprt* care măsoară nivelul de încredere al respondenților în partidele politice pe o scală de la 0 la 10 unde 0 înseamnă neîncredere totală, iar 10 înseamnă încredere totală. Pentru a afla valoarea medie a răspunsurilor pentru variabila *trstprt* vom aplica următoarea formulă în Excel: într-o foaie de calcul (sheet) nouă scriem într-o celulă următoarea formulă:[55]

```
=AVERAGE(dataset_exemplu!J2:J11023)
```

unde **average** reprezintă funcția dorită (valoarea de mijloc), **dataset_exemplu!** reprezintă foaie de calcul de unde dorim ca Excel să își selecteze valorile, iar **N2:N1435** reprezintă seria de celule care compun coloana N care reprezintă variabila *trstprt*[56].

5.2.2. Analiza de dispersie a datelor în Excel: abaterea standard și amplitudinea

Deviația standard (sau abaterea standard) ne indică distanța valorilor față de medie. Dacă abaterea standard este mică înseamnă că valorile sunt grupate aproape de medie, iar dacă este mare înseamnă că valorile sunt depărtate de medie. Formula pentru abaterea standard în Excel este:

```
=STDEV(dataset_exemplu!J2:J11023)
```

[55] Putem ajunge la funcțiile statistice ale Excel selectând **Formulas -> More Functions -> Statistical.**

[56] Motivul pentru care rezultatele nu coincid cu cele din Analiza RStudio și Stata este că în Excel am folosit o altă metodă de a înlocui valorile nule care a dus la obținerea unei eșantion mai mare ca în celelalte programe statistice.

În Excel, pentru a afla intervalul unei serii de date vom aplica următoarea formulă:

```
=RANGE(dataset_exemplu!J2:J11023)
```

5.3. Analiza bivariată în Excel

5.3.1. Corelația în Excel

După cum am discutat în secțiunea 3.3, corelația indică relația dintre două variabile, X și Y. Ne reamintim că rezultatul analizei de corelație este reprezentat de un indicator (Pearson r) ce variază între -1 și +1. Astfel, o corelație poate fi pozitivă sau negativă. O corelație pozitivă înseamnă că pe măsură ce X crește, Y crește. O corelație negativă înseamnă că pe măsură ce X crește, Y scade. Corelația ne indică, de asemenea, intensitatea relației dintre X și Y. Pe măsură ce indicatorul de corelație se apropie mai mult de +1, spunem că relația este puternică și pozitivă. Pe măsură ce indicatorul de corelație se apropie mai mult de -1, spunem că relația este puternică și negativă. Un indicator de corelație egal cu zero ne arată că este probabil să nu existe nicio relație între X și Y.

Să presupunem că vrem să analizăm relația dintre nivelul de încredere în partidele politice și (*trstprt*) și satisfacția respondenților cu situația economică (*stfeco*). Astfel, vom parcurge în Excel următorii pași:

1. Vom crea o nouă foaie de calcul în care vom include următoarele informații: valorile celor două variabile, mărimea eșantionului, valoarea medie și valoarea deviației standard.
2. Calculăm valoarea pentru media, mărimea eșantionului și abaterea standard pentru fiecare variabilă în parte.
3. Pentru a afla valoarea corelației vom aplica următoarea formulă:

```
=correl(A2:A11023,B2:B11023)
```

Rezultatul corelației este +.517, ceea ce înseamnă ca avem o corelație pozitivă între cele două variabile. Acest rezultat indică faptul că există o relație puternică între nivelul de încredere și nivelul de satisfacție. În concluzie, creșterea nivelului de satisfacție tindă să fie asociată cu o creșterea a nivelului de încredere în partidele politice.

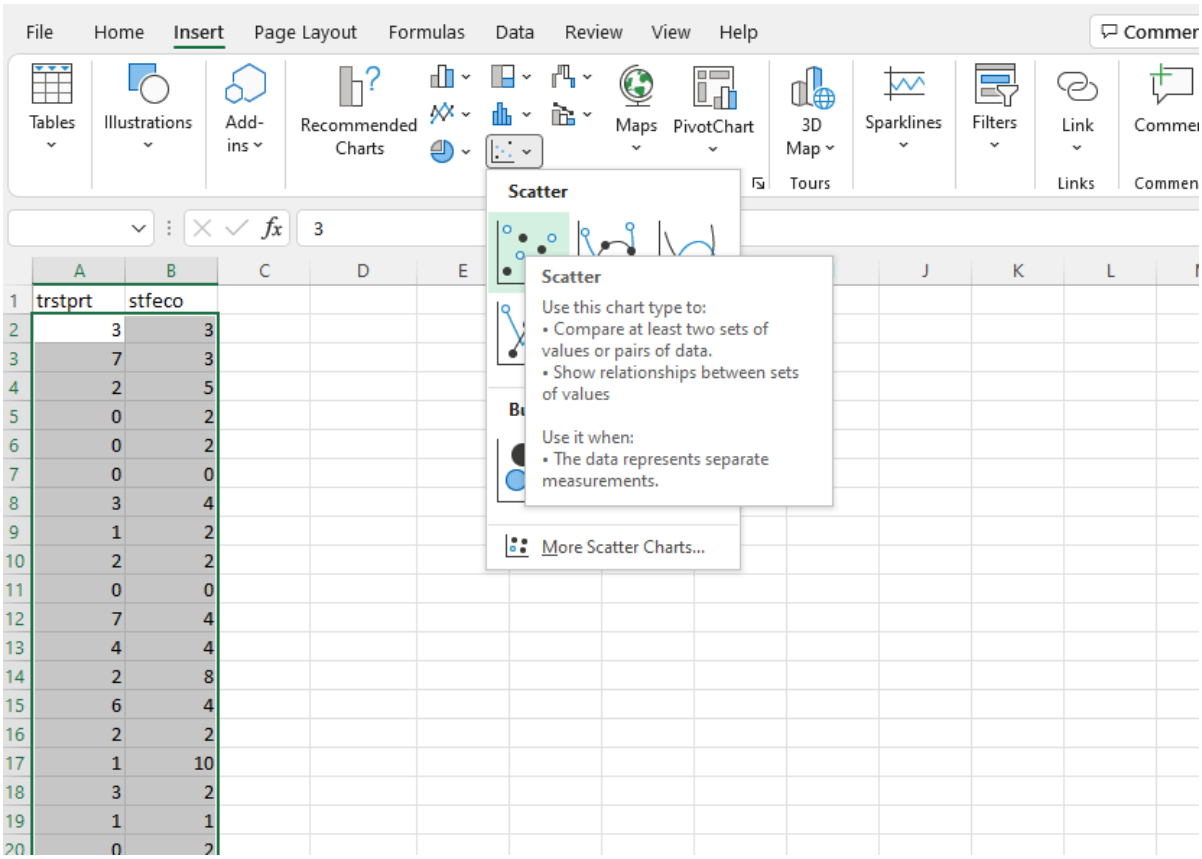
5.3.2. Regresia simplă liniară în Excel

Această secțiune oferă un exemplu de aplicare a analizei de regresie liniară în Excel. Ne reamintim din secțiunea 3.4 că utilizarea modelului de regresie liniară simplă înseamnă că estimăm modul în care o variabilă independentă prezice valorile unei variabile dependente.

Să presupunem că am vrea să folosim nivelul de satisfacție (*stfeco*) ca variabilă independentă și nivelul de încredere ca variabilă dependentă. Având în vedere că rezultatul corelației dintre aceste variabile este +.517, putem asuma că există o relație pozitivă și că nivelul de satisfacție ar putea fi un predictor bun pentru nivelul de încredere.

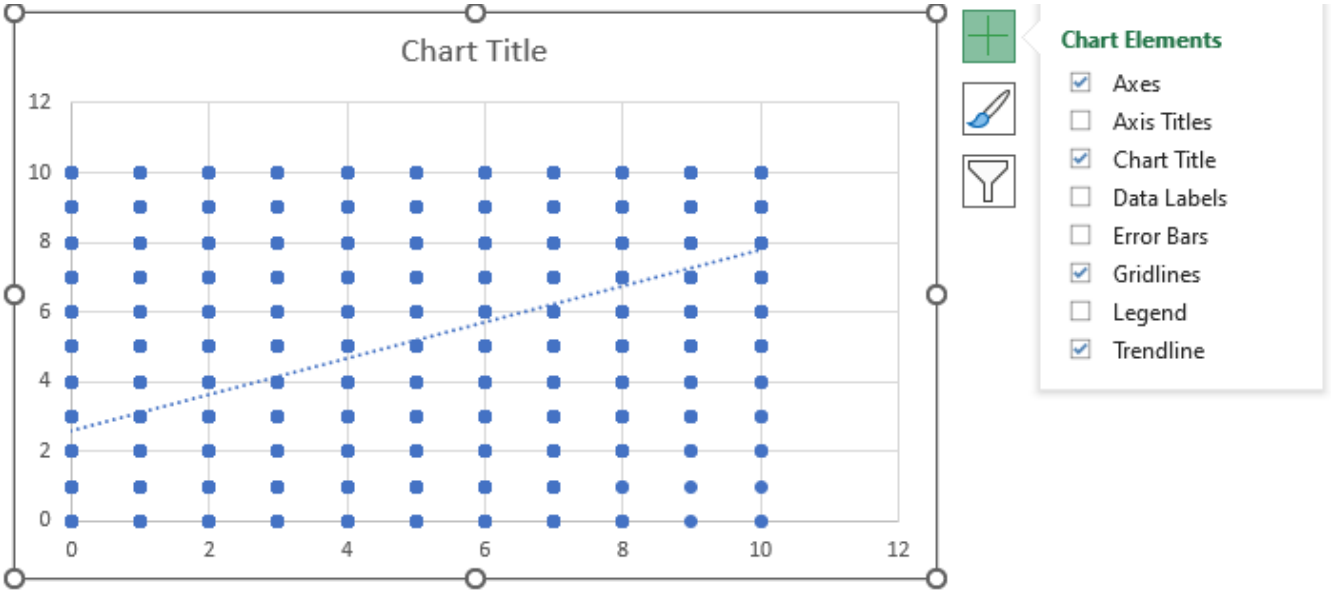
1. În foia de calcul folosită anterior pentru a calcula corelația, selectăm ambele variabile fără rândul 1 (eticheta variabilei).
2. Selectăm **Insert** și apoi iconița Scatter chart

Figura 5.8 Pregătirea diagramei cu puncte



3. Vom insera dreapta de regresie prin a da click pe iconița + din dreapta sus a graficului generat și vom selecta Trendline.

Figura 5.9 Diagrama cu puncte



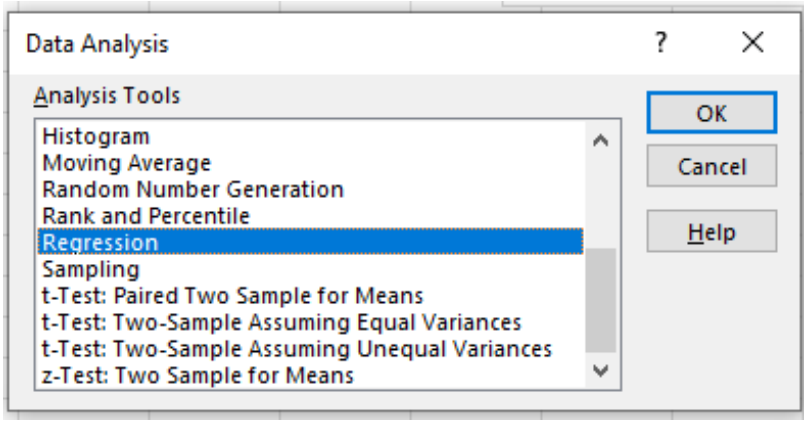
Reprezentând grafic relația dintre X și Y putem estima vizual dacă există o relație puternică între cele două variabile. Putem identifica, astfel, dacă ecuația de regresie care rezumă relația dintre X și Y poate fi folosită pentru a prezice Y pentru o valoare dată a lui X.

Următorul pas după realizarea graficului de tip Scatterplot este să aplicăm ecuația de regresie pe cele două variabile. Pentru a aplica această ecuație în Excel trebuie să instalăm **Data Analysis ToolPak**.

Pentru a aplica ecuația de regresie cu ajutorul **Data Analysis ToolPak** aplicăm pașii următori:

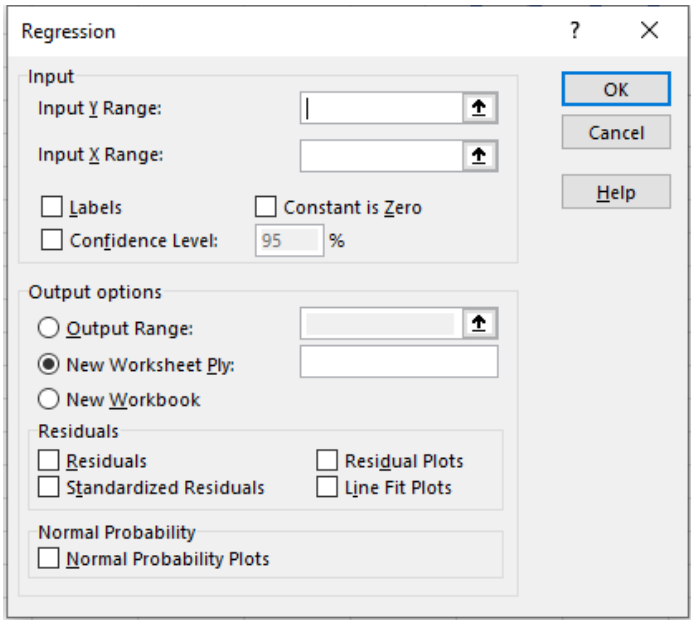
1. Selectăm **Data**, apoi Data Analysis

Figura 5.10 Regresia simplă liniară în Excel pasul 1



2. Selectăm *Regression*

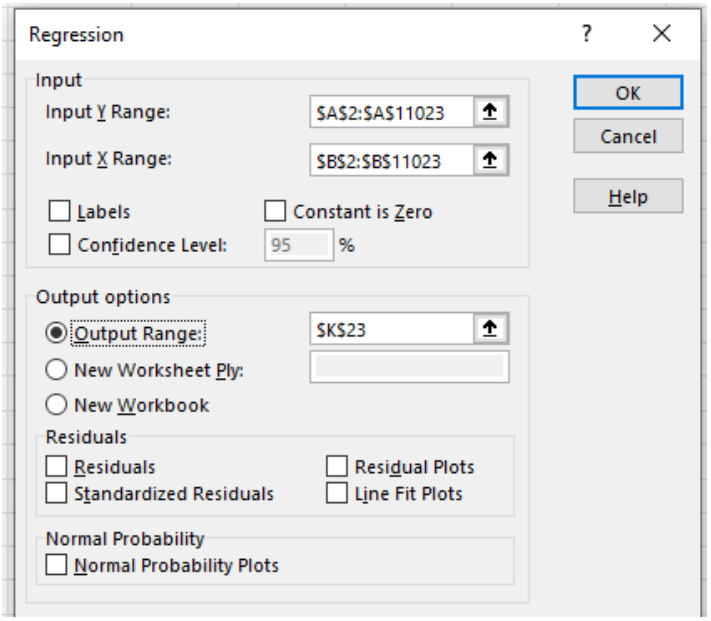
Figura 5.11 Regresia simplă liniară în Excel pasul 2



3. Click OK
4. În chenarul din partea de sus, introducem la Input Y Range variabila dependentă *trstprt*, iar la Input X Range variabila independentă *stfeco.*

5. La căsuța Output Range selectăm celula în care dorim generarea rezultatului.

Figura 5.12 Regresia simplă liniară în Excel specificarea modelului



6. Click OK

În urma pașilor făcuți în Excel, vom genera un tabel de ieșire similar celui din Figura 5.13, de mai jos.

Figura 5.13 Rezultatele regresiei simple liniare în Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.517349001
R Square	0.267649989
Adjusted R Square	0.267583533
Standard Error	2.160535455
Observations	11022

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18799.78693	18799.78693	4027.449764	0
Residual	11020	51440.40623	4.667913451		
Total	11021	70240.19316			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.917497986	0.040056591	22.90504396	1.8016E-113	0.838979886	0.996016087	0.838979886	0.996016087
X Variable 1	0.518078777	0.008163582	63.46219161	0.0000000000	0.502076694	0.534080861	0.502076694	0.534080861

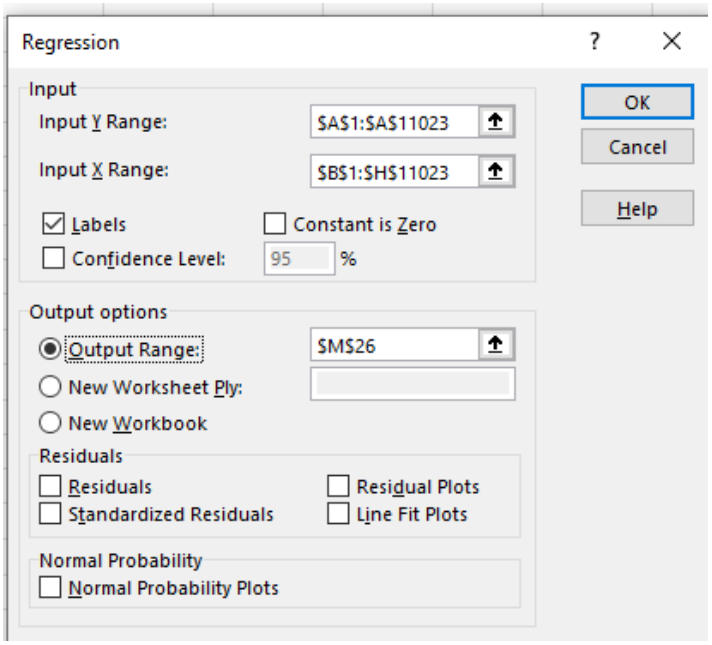
5.3.3. Analiza de regresie multiliniară în Excel

Să presupunem că pe baza setului nostru de date vrem să investigăm relația dintre nivelul de încredere în partidele politice (*trstprt*) și nivelul de satisfacție cu situația economică (*stfeco*) cu propria viață (*stflife*), cu democrația (*stfdemo*), cu nivelul de fericire (*happy*), cu interesul în politică (*polintr*), cu participarea la vot la ultimele alegeri (*vote*) și cu apropierea față de un partid politic (*clsprty*). În Excel vom aplica următorii pași pentru a specifica modelul de regresie multiliniară:

1. Creăm o nouă foaie de calcul numită Regresie Multiplă.
2. În noua foaie de calcul vom introduce următoarele informații: valorile pentru variabilele de interes[57]
3. Selectăm tab-ul **Data**, apoi **Data Analysis**
4. În caseta **Data Analysis** selectăm **Regression**
5. Click OK
6. În chenarul regresiei vom introduce la Input Y Range seria valorile variabilei dependente (*trstprt)*
7. La Input X Range vom introduce seria celulelor cu valorile pentru variabilele dependente
8. Bifăm caseta Labels
9. La Output range introducem celula în care vrem să fie generat rezultatul.

[57] Când utilizăm mai mulți predictori în regresia multiplă este esențial ca în tabelul cu variabile variabila dependentă să fie în partea stângă, iar toți predictorii să fie la dreapta variabilei dependente, astfel încât să identificăm cu ușurință variabila dependentă și predictorii.

Figura 5.14 Regresie multiliniară în Excel specificarea modelului



Tabelul de ieșire în urma realizării pașilor în Excel este similar celui din Figura 5.15, de mai jos.

Figura 5.15 Rezultatele regresiei multiliniare în Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.596367903
R Square	0.355654676
Adjusted R Square	0.355245159
Standard Error	2.027120908
Observations	11022

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	24981.25314	3568.750449	868.4741055	0
Residual	11014	45258.94002	4.109219177		
Total	11021	70240.19316			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.75396584	0.096642303	18.14904837	1.48306E-72	1.564529589	1.943402092	1.564529589	1.943402092
stfeco	0.242798389	0.01038939	23.36983971	6.2345E-118	0.22243332	0.263163458	0.22243332	0.263163458
stflife	-0.063103631	0.012186459	-5.178176255	2.28019E-07	-0.086991276	-0.039215986	-0.086991276	-0.039215986
stfdem	0.381126658	0.009601367	39.695041	0	0.362306256	0.39994706	0.362306256	0.39994706
happy	0.028851619	0.012379122	2.330667558	0.019788855	0.004586318	0.05311692	0.004586318	0.05311692
clsprty	-0.059323036	0.017589455	-3.372647688	0.000747055	-0.093801524	-0.024844548	-0.093801524	-0.024844548
polintr	-0.314747212	0.021919807	-14.35903223	2.4482E-46	-0.357713966	-0.271780459	-0.357713966	-0.271780459
vote	-0.097037703	0.023580745	-4.115124569	3.89788E-05	-0.143260194	-0.050815213	-0.143260194	-0.050815213

5.4. Funcții și comenzi rapide adiționale în Excel

5.4.1. Funcția TRIM

Această funcție este utilă în pregătirea datelor în Excel, pentru eliminarea spațiilor suplimentare dintr-o celulă, fie la început, la final sau oriunde între cuvinte. Adesea, atunci când extragem date dintr-o bază de date, este posibil să observăm că unele spații suplimentare sunt puse în spatele sau în fața datelor legitime. Acest lucru poate crea discrepanțe uriașe dacă încercăm să comparăm date folosind funcții precum IF sau VLOOKUP. Pentru a elimina aceste spații vom folosi următoarea funcție:

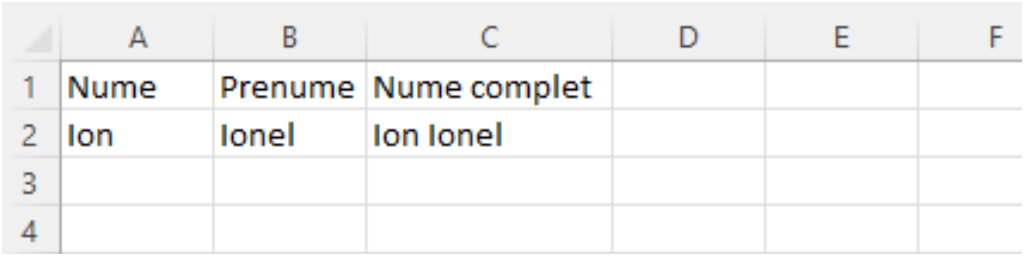
```
=TRIM(dataset_exemplu!J2:J11023)
```

5.4.2. Funcția CONCATENATE

CONCATENATE este o funcție simplă, dar foarte eficientă și utilă pentru multe operațiuni de pregătire a datelor. Este folosită pentru unirea valorilor din diferite celule. De multe ori obținem o bază de date în care este necesar să combinăm 2 coloane sau 2 câmpuri. De exemplu, numele și prenumele respondenților sunt scrise în coloane separate. Pentru a le scrie împreună într-o coloană vom aplica funcția CONCATENATE după cum urmează:

```
=CONCATENATE(A2," ",B2)
```

Figura 5.16 Concatenare valori în Excel



	A	B	C	D	E	F
1	Nume	Prenume	Nume complet			
2	Ion	Ionel	Ion Ionel			
3						
4						

5.4.3. Sinteză comenzi utile în Excel

În această secțiune arătăm câteva **comenzi rapide și generale** pe care le putem realiza folosind doar tastatura pentru manipularea mai rapidă a registrelor de lucru.

Tabel 5.6 Comenzi rapide în Excel

Comandă	Definiție
Ctrl+N	Creează un nou fișier excel
Ctrl+O	Deschide un fișier excel
Ctrl+S	Salvează un fișier excel
F12	Deschide fereastra Save As
Ctrl+W	Închide un fișier excel
F4	Repetăm ultima comandă sau acțiune. De exemplu, dacă ultimul lucru pe care l-am introdus într-o celulă este „bună ziua” sau dacă schimbăm culoarea fontului, făcând clic pe altă celulă și apăsând F4 repetă acțiunea în noua celulă.
Shift+F11	Creează o nouă foaie de calcul
Ctrl+Z	Anulăm o acțiune
Ctrl+Y	Refacem o acțiune
Ctrl+F2	Comutăm la Print Preview
F1	Deschidem panou Help
Alt+Q	Accesăm caseta “Tell me what you want to do”
F7	Verificăm ortografia
F9	Calculăm toate foile de lucru din toate registrele de lucru deschise
Shift+F9	Calculăm foile de lucru active
F11	Creăm o diagramă cu bare pe baza datelor selectate (pe o foaie separată)
Alt+F1	Creăm o diagramă cu bare încorporată pe baza datelor selectate (aceeași foaie)
Ctrl+F	Căutăm într-o foaie de calcul sau utilizăm „Găsiți și înlocuiți”

Comandă	Definiție
Ctrl+Tab	Comutăm între registrele de lucru deschise
Shift+F3	Inserăm o funcție
Alt+F8	Creăm, rulăm, editățm sau ștergem o comandă

O serie de **comenzi pentru navigarea rapidă** pot fi utilizate în interiorul foii de lucru, într-o celulă sau în întregul registru de lucru.

Tabel 5.7 Comenzi rapide ale foii de calcul

Comandă	Definiție
Left/Right Arrow	Mutăm o celulă la stânga sau la dreapta
Ctrl+Left/Right Arrow	Ne deplasăm la cea mai îndepărtată celulă din stânga sau dreapta
Up/Down Arrow	Mutăm o celulă în sus sau în jos
Ctrl+Up/Down Arrow	Ne deplasăm în celula cea mai de sus / jos care este folosită
Tab	Mergem la următoarea celulă
Shift+Tab	Mergem la celula anterioară
Ctrl+End	Accesăm celula folosită cea mai jos din dreapta
F5	Accesăm orice celulă apăsând F5 și tastând coordonatele celulei sau numele celulei.
Ctrl+Home	Trecem la începutul unei foi de lucru
Page Up/Down	Mutăm un ecran în sus sau în jos într-o foaie de lucru
Alt+Page Up/Down	Mutăm un ecran la dreapta sau la stânga într-o foaie de lucru
Ctrl+Page Up/Down	Trecem la foaia de lucru anterioară sau următoare

Există și alte câteva comenzi rapide combinate pentru accelerarea selecției celulelor din tabelul Excel. Utilizarea tastei Shift pentru a modifica tastele săgeți ne permite să extindem celulele selectate.

Tabel 5.8 Comenzi utile de selecție a celulelor

Comandă	Definiție
Shift+Left/Right Arrow	Extindem selecția celulelor la stânga sau la dreapta
Shift+Space	Selectăm tot rândul
Ctrl+Space	Selectăm întreaga coloană
Ctrl+Shift+Space	Selectăm întreaga foaie de lucru

6. Tehnici de vizualizare grafică a datelor

6.1. Introducere în vizualizarea datelor

Vizualizarea datelor este reprezentarea grafică a informațiilor și a datelor. Folosind elemente vizuale precum diagrame, grafice și hărți, oferim publicului o modalitate accesibilă de a vedea și înțelege tendințele, valorile și modelele descrise de date. Astfel, instrumentele și tehnologiile de vizualizare a datelor sunt esențiale pentru a analiza cantități masive de informații și pentru a lua decizii bazate pe date.

În era Big Data în care ne aflăm, vizualizarea datelor presupune nu doar grafice simple de tipul celor cu bare și a diagramelor circulare, ci există o serie întreagă de metode de vizualizare avansate de prezentare a datelor într-o manieră eficientă.

Vizualizarea datelor a devenit în ultimele decenii o artă în sine, cu atât mai mult cu cât instrumentele de generare a acestor vizualizări și cele de diseminare sunt facilitate de digitalizarea informațiilor. Această dezvoltare a fost acompaniată de propuneri de reguli și recomandări specifice vizualizării datelor dezvoltate într-o literatură abundentă (R. L. Harris 1996; Cairo 2013; Murray 2013; Wilke 2019). De exemplu, atunci când prezentăm un grafic cu bare, este recomandabil ca distanța dintre bare să fie mai îngustă decât lățimea barelor (așa cum am ilustrat în Figura 4.6). Niciodată nu este recomandabilă utilizarea graficelor statice cu efect tridimensional, cum ar fi, de exemplu, graficul de bare sau diagramele cu secțiuni (*pie chart*) cu efect 3D. Deși în aparență dovedesc o dexteritate și competență în utilizarea programului Excel (principalul program care este folosit pentru aceste grafice 3D), aceste grafice nu ajută beneficiarul sau cititorul să înțeleagă mai bine informația pe care dorim să i-o transmitem prin vizualizări grafice. Acest efect 3D, inutil de altfel din punct de vedere vizual, mai ales în rapoartele imprimate pe hârtie, poate induce foarte ușor în eroare cititorul deoarece nu mai permite identificarea exactă a valorii fiecărei bare pe axa frecvențelor sau a procentelor, efectul 3D mutând punctul de referință al barei față de axă. În plus, efectul 3D nu mai permite identificarea diferențelor vizuale, de mărime,

dintre bare sau feliile diagramelor. În cazul diagramelor cu secțiuni (*pie charts*) este recomandabil să aranjăm feliile descrescător, cea mai mare felie și următoarea fiind așezate în dreapta, respectiv în stânga axei verticale a graficului (ora 12:00) (așa cum am ilustrat în Figura 4.7). Nu este recomandabilă extragerea unor felii. Deși în aparență pare că subliniază o categorie anume, a atrage atenția asupra acestei categorii (felii) din grafic, distrage în același timp atenția de la celelalte informații cuprinse în grafic. Nu în ultimul rând, este recomandabil ca în acest tip de grafic să includem și valorile lipsă (non-răspuns) mai ales atunci când acestea nu reprezintă mai mult de 10% din cazuri. Dacă acestea abundă, trebuie să ne punem semne de întrebare cu privire la calitatea operaționalizării variabilei, a exhaustivității categoriilor (a se vedea și discuția de la secțiunea 3.1.1) sau chiar a erorilor de codificare cauzate de operator sau de sistemul de centralizare a datelor.

Scopul acestui capitol este de a exemplifica practic o serie de instrumente de vizualizare. Printre cele mai cunoscute tipuri de instrumente de vizualizare pe care le vom exemplifica în această secțiune a cărții sunt cele din tabelul de mai jos.

Tabel 6.1 Instrumente generale de reprezentare vizuală a datelor

Instrument	Descriere
Diagramă (în engleză Chart)	Informații prezentate sub formă grafică, cu datele afișate de-a lungul a două axe (grafic cu bare sau histogramă), lini sau secțiuni (diagrame cu secțiuni pie chart).
Tabel (în engleză Table)	Un set de cifre afișate în rânduri și coloane
Grafic (în engleză Graph)	O diagramă cu puncte, linii, segmente sau curbe care reprezintă anumite variabile în comparație, de obicei de-a lungul a două axe în unghi drept.
Hartă (în engleză Geospatial)	O vizualizare care arată datele sub formă de hartă, pe baza unei localizări spațiale, folosind diferite forme și culori pentru a arăta relația dintre date și locații specifice.
Dashboards	O colecție de vizualizări și date afișate într-un singur loc (de exemplu pe o pagină de internet), pentru a ajuta la analiza și prezentarea datelor

Programele în care pot fi aplicate instrumentele de vizualizare a datelor variază de la simple la complexe, de la intuitive la obtuze. Nu orice instrument este potrivit pentru fiecare persoană care dorește să învețe tehnici de vizualizare și nu orice instrument se poate adapta la scopuri industriale sau academice. Programul pe care îl vom folosi pentru a exemplifica instrumentele de vizualizare menționate mai sus este RStudio cu pachetul specific ggplot2.

6.2. Vizualizarea datelor în RStudio (ggplot2)

6.2.1. Înțelegerea sintaxei ggplot2

Aplicațiile practice elaborate în secțiunile următoare se bazează pe seturile de date oferite de RStudio. Cu ajutorul pachetului **ggplot2** (Wickham 2016) și a funcțiilor acestora vom realiza o serie de elemente grafice des întâlnite în procesul de analiză a datelor.

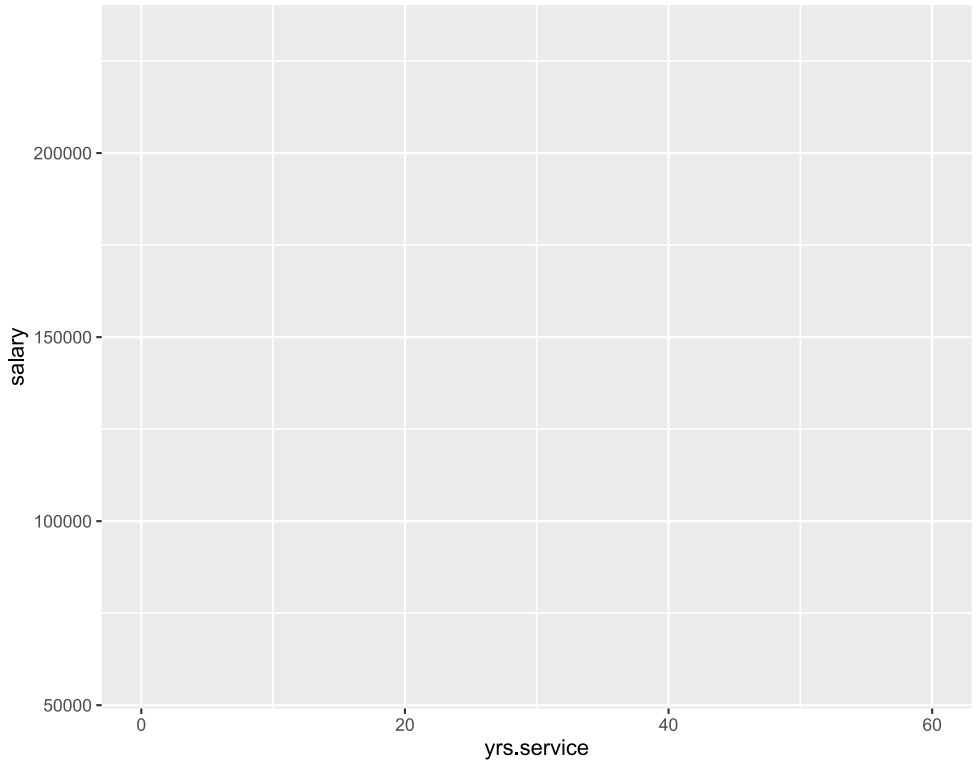
Înainte de a discuta structura funcției **ggplot2** este important să înțelegem în linii mari ce reprezintă acest pachet. **ggplot2** este un pachet de vizualizare a datelor creat pentru limbajul de programare statistic R. Acest pachet poate îmbunătăți considerabil calitatea și estetica graficii și permite realizarea elementelor grafice într-un mod mult mai eficient. Funcțiile din pachetul **ggplot2** construiesc un grafic în straturi, astfel că graficele realizate cu **ggplot2** pot fi îmbunătățite constant prin adăugarea mai multor teme la o diagramă existentă. Pentru realizarea graficelor din acest capitol vom avea nevoie de încărcarea pachetului **ggplot2** (Wickham 2016):

```
### pachete necesare ###
library(ggplot2)
library(car)
```

Primul element grafic pe care îl vom realiza cu ajutorului pachetului **ggplot2** este diagrama cu puncte (în engleză *scatterplot*). Înainte să realizăm diagrama cu puncte pe baza setului de date oferit de RStudio (Salaries) din pachetul **car** (J. Fox și Weisberg 2019)vom inițializa un ggplot de bază.

```
### Inițiem graficul ggplot de bază și alegem două variabile ale setului d
e date ###
ggplot(Salaries, aes(x= yrs.service, y= salary))
```

Figura 6.1 Vizualizare în ggplot2



Observăm că rezultatul sintaxei de mai sus este un grafic ggplot gol. Chiar dacă x și y sunt specificate, nu există puncte sau linii în el. Acest lucru se datorează faptului că nu am indicat funcției ggplot tipul de grafic pe care dorim să-l realizăm, ci am indicat doar ce set de date să folosească și ce coloane ar trebui folosite pentru axa X și Y. Nu i-am cerut în mod explicit să tragă niciun punct. De asemenea, reținem că

funcția **aes()** este utilizată pentru a specifica axele X și Y. Astfel, când folosim funcția ggplot variabile de interes trebuie să fie specificate în cadrul funcției **aes()**.

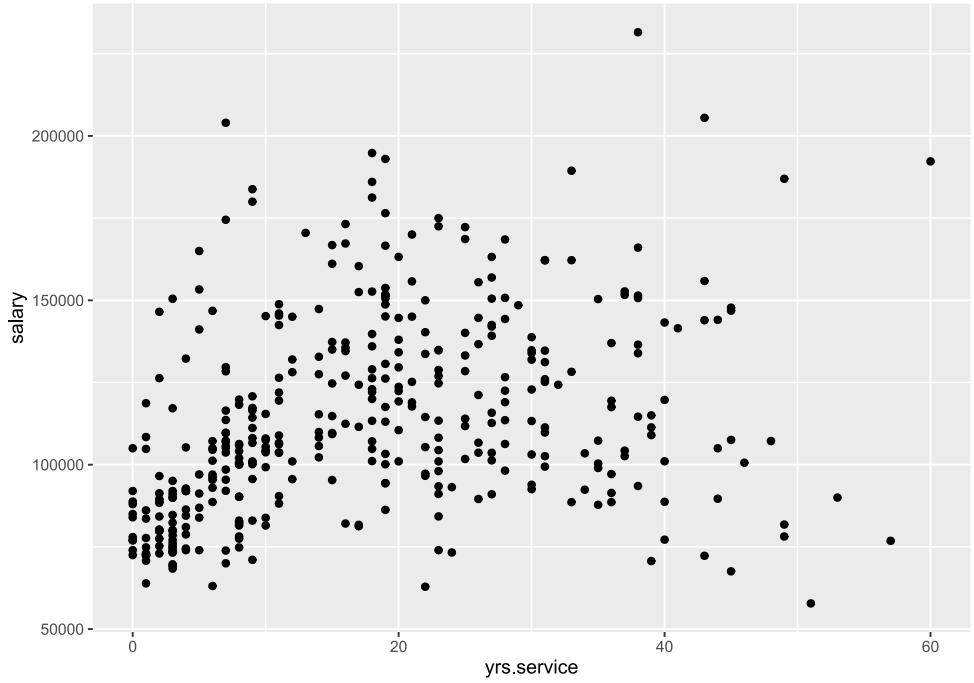
Plecând de la graficul ggplot de bază realizat anterior, vom elabora elementele vizuale din următoarele secțiuni.

6.2.2. Diagrama cu puncte (*scatterplot*)

În această secțiunea vom continua să lucrăm pe graficul generat în secțiunea anterioară. Vom realiza un grafic de tip diagramă cu puncte (în engleză *scatterplot*) adăugând puncte prin folosirea argumentului **geom_point**. Dar la ce ne ajută o diagramă cu puncte? Diagramele cu puncte sunt folosite de regulă atunci când vrem să identificăm gradul de corelație dintre două variabile cantitative. Folosind setul de date **Salaries** oferit în mod gratuit de pachetul **car** (J. Fox și Weisberg 2019), să presupunem că suntem interesați să aflăm dacă există o corelație între variabila care măsoară vechimea în muncă (*yrs.service*) și cea care măsoară venitul anul (*salary)* al unor indivizi. Putem vizualiza această relație dintre cele două variabile, cu ajutorul unei diagrame cu puncte, în care fiecare punct reprezintă un individ (respondent), ce ia două valori, o valoare pentru variabila *yrs.service* și o altă valoare pentru variabila *salary*. Pentru a genera o diagramă cu puncte cu ajutorul funcției ggplot() vom executa în RStudio următoarea comandă:

```
###Indicăm ggplot să adauge puncte pentru generarea diagramei cu puncte###
ggplot(Salaries, aes(x=yrs.service, y=salary)) + geom_point()
```

Figura 6.2. Diagrama cu puncte în ggplot2



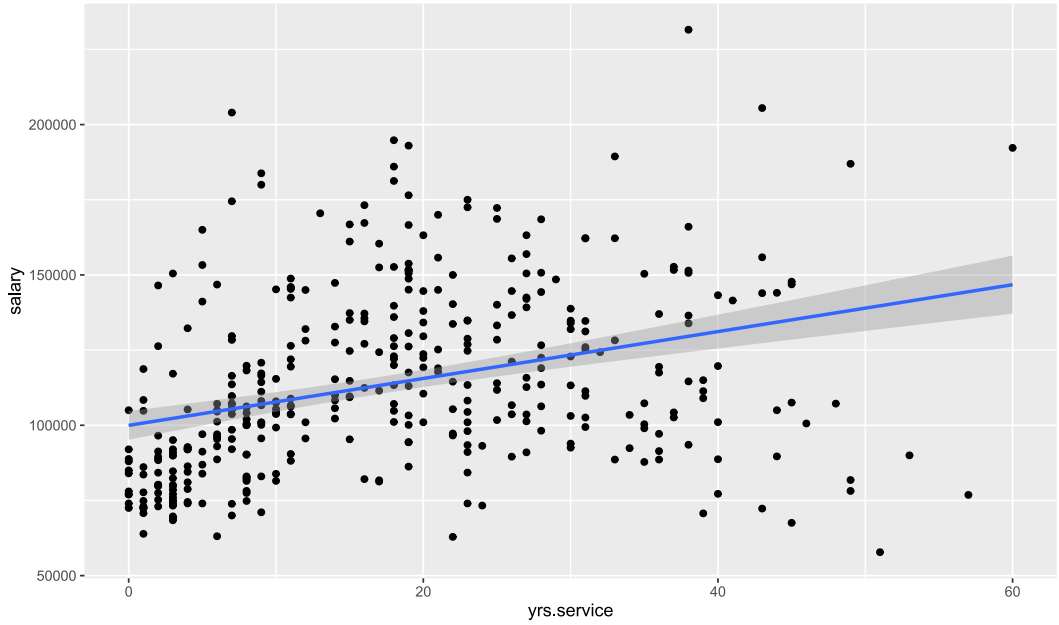
Graficul de mai sus reprezintă un grafic de dispersie de bază, în care fiecare punct reprezintă un respondent. Cu toate acestea, graficului generat îi lipsesc unele componente de bază, cum ar fi titlul, etichetele semnificative ale axelor etc. Vom remedia aceste aspecte în pașii următori.

Similar **geom_point()**, există multe astfel de straturi geom pe care le vom aplica în următoarele secțiuni. În continuare, vom interpola o linie de regresie prin adăugarea **geom_smooth(method='lm')**. Deoarece metoda este setată ca **lm** (prescurtare pentru model liniar), ea trasează linia de cea mai bună potrivire.

```
### Inserăm linia de regresie pentru a indica direcția corelației###

scatterplot <- ggplot(Salaries, aes(x=yrs.service, y=salary)) + geom_point
() + geom_smooth(method="lm")
```

Figura 6.3 Diagrama cu puncte cu linia de regresie în ggplot2



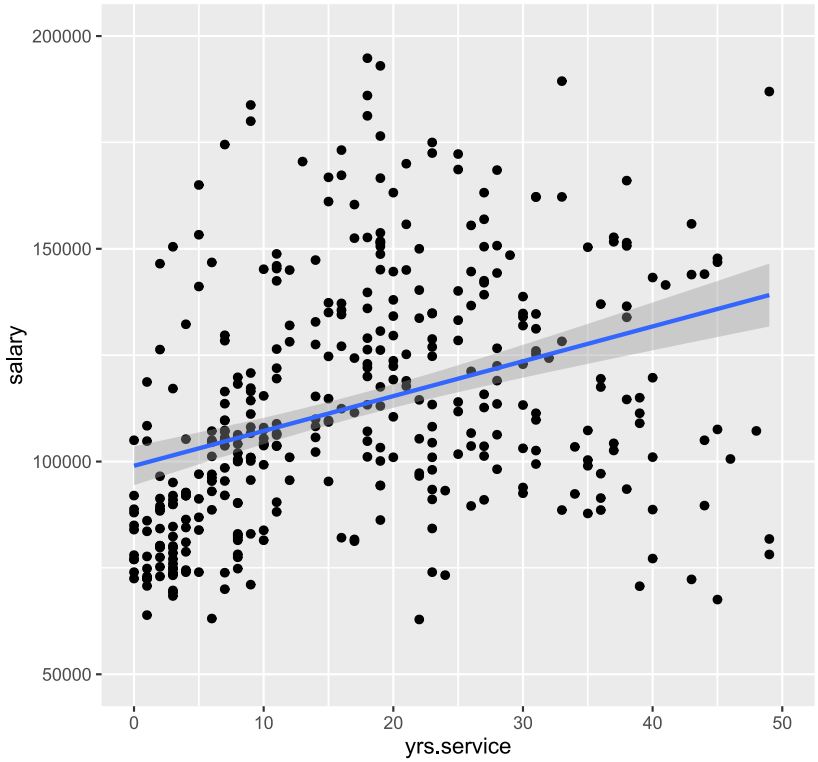
6.2.2.1. Ajustarea limitelor axelor X și Y

În graficul din secțiunea anterioară observăm că majoritatea punctelor sunt concentrate în partea de jos a graficului. Pentru a dispersa punctele este necesar să ajustăm limitele axelor X și Y.

O metodă prin care putem realiza acest lucru este să ștergem punctele aflate în afara limitelor axelor. Acest lucru va schimba liniile de cea mai bună potrivire sau liniile de netezire în comparație cu datele originale. Vom utiliza argumentele **xlim()** și **ylim()**.

```
### Ștergem punctele aflata în afara limitelor###
scatterplot + xlim(c(0, 50)) + ylim(c(50000,200000))
```

Figura 6.4 Diagrama cu puncte cu linia de regresie, ajustată, în ggplot2



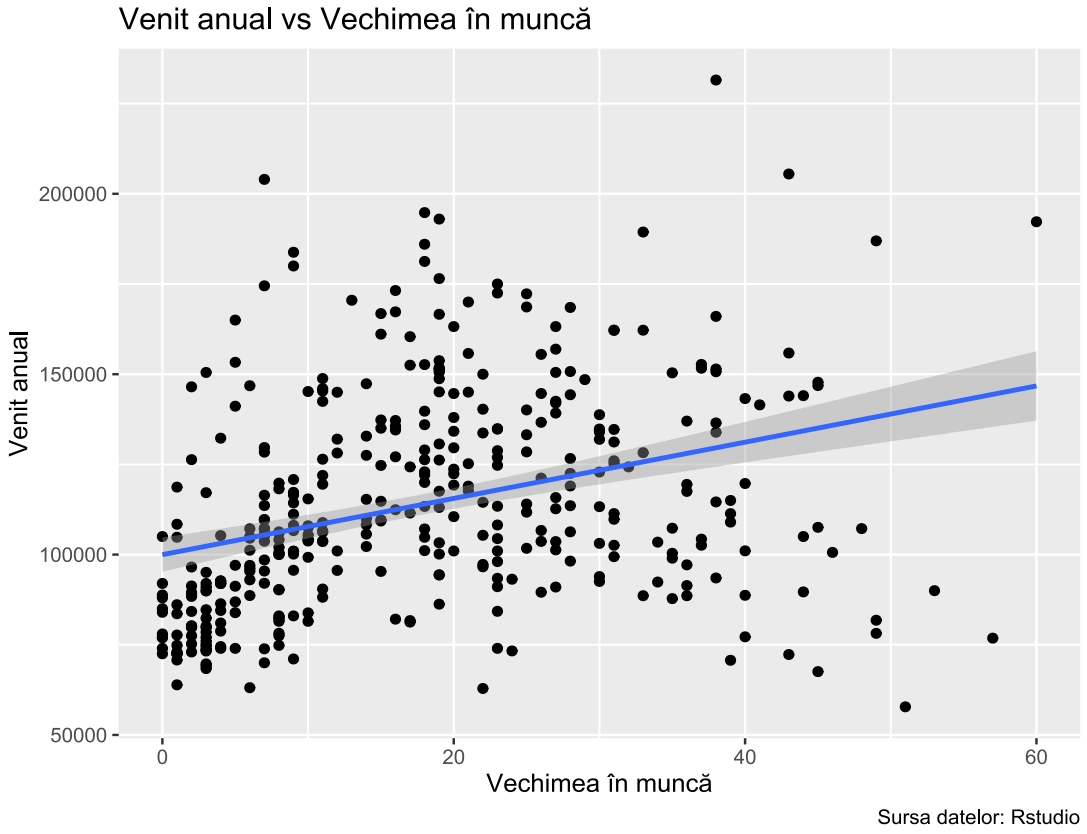
În acest caz, diagrama nu a fost construită de la zero, ci mai degrabă a fost construită în continuarea obiectului creat și denumit "g". Acest lucru se datorează faptului că, graficul anterior a fost stocat ca g, un obiect ggplot, care atunci când este apelat va reproduce graficul original. Folosind ggplot2, putem adăuga mai multe straturi, teme și alte setări peste graficul inițial.

6.2.2.2. Schimbarea numelui titlului și a axelor

La graficul **g** creat și salvat în secțiunea anterioară vom adăuga un titlu și nume pentru axele X și Y. Vom folosi funcția **labs()** cu argumente title, x și y. O altă opțiune este să folosim **ggtitle()**, **xlab()** și **ylab()**.

```
### Agăugăm tilu și nume axelor###
g + labs(title="Venitul anual vs vechimea în muncă", y="Venit anual", x="V
echimea în muncă", caption="Sursa datelor: RStudio")
```

Figura 6.5 Diagrama cu puncte cu linia de regresie, în ggplot2, adăugare titlu




```
### sau###
g1 + ggtitle("Venitul anual vs vechimea în muncă") + xlab("Vechimea în muncă")
+ ylab("Venit anual")
```

La acest moment sintaxa completă pentru a genera diagrama cu puncte în RStudio este:

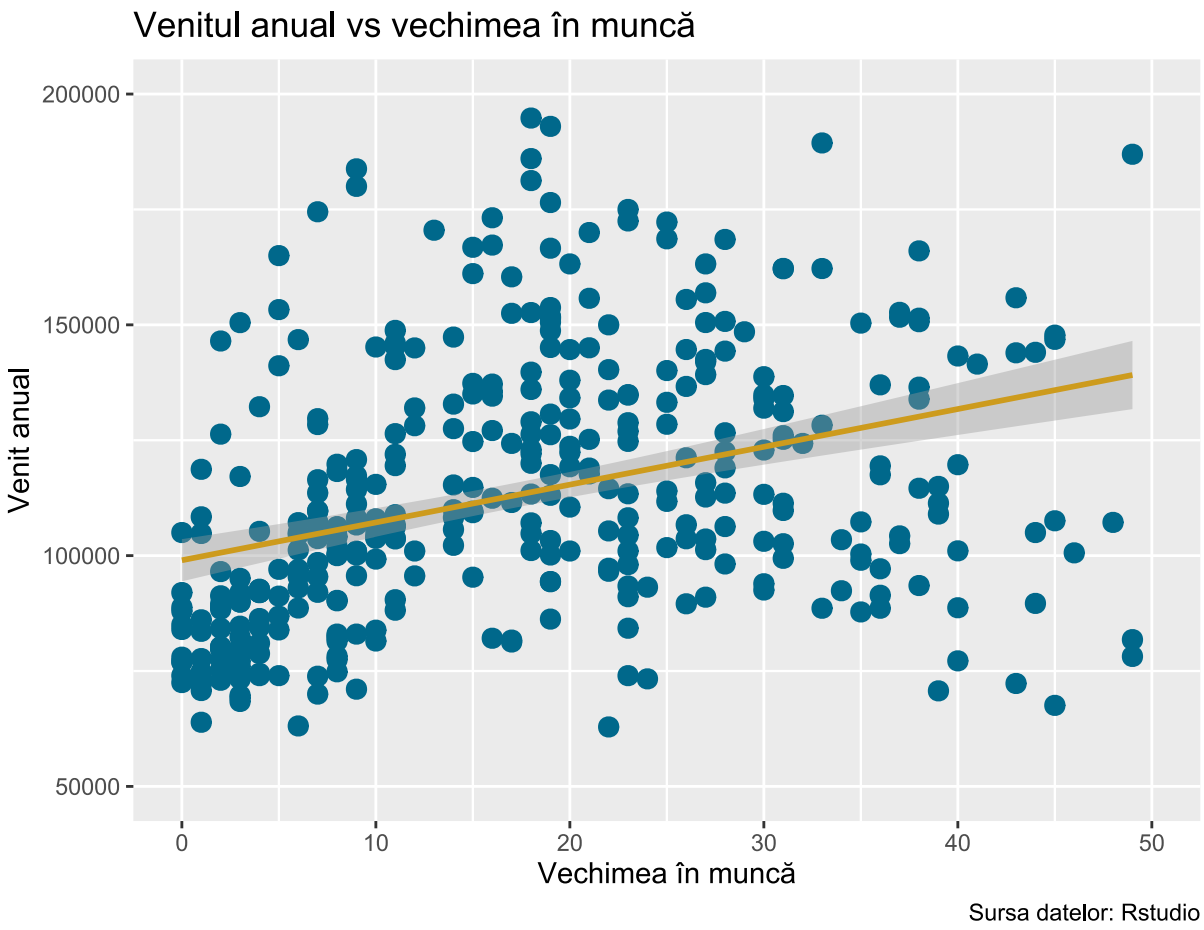
```
### Întreaga sintaxă pentru generarea diagramei cu puncte###
scatterplot <- ggplot(Salaries, aes(x=yrs.service, y=salary)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlim(c(0,50)) + ylimc(50000, 200000)) +
  labs(title="Venitul anual vs vechimea în muncă", y="Venit anual",    x="
Vechimea în muncă", caption="Sursa datelor: RStudio")
```

6.2.2.3. Schimbarea culorilor și a dimensiunii punctelor

Putem schimba estetica unui strat de geom modificând geomurile respective. În continuare vom schimba culoarea punctelor și a liniei la o valoare statică.

```
ggplot(Salaries, aes(x=yrs.service, y=Salary)) +
  geom_point(col="deepskyblue4", size=3) +   ### Setăm culoarea și dimensi
unea statică pentru puncte###
  geom_smooth(method="lm", col="goldenrod3") +  ### Schimbăm culoarea lini
ei###
  xlim(c(0,50)) + ylim(c(50000, 200000)) +
  labs(title="Venitul anual vs vechimea în muncă", y="Venit anual", x="Vec
himea în muncă", caption="Sursa datelor: RStudio")
```

Figura 6.6 Diagrama cu puncte colorată în ggplot2



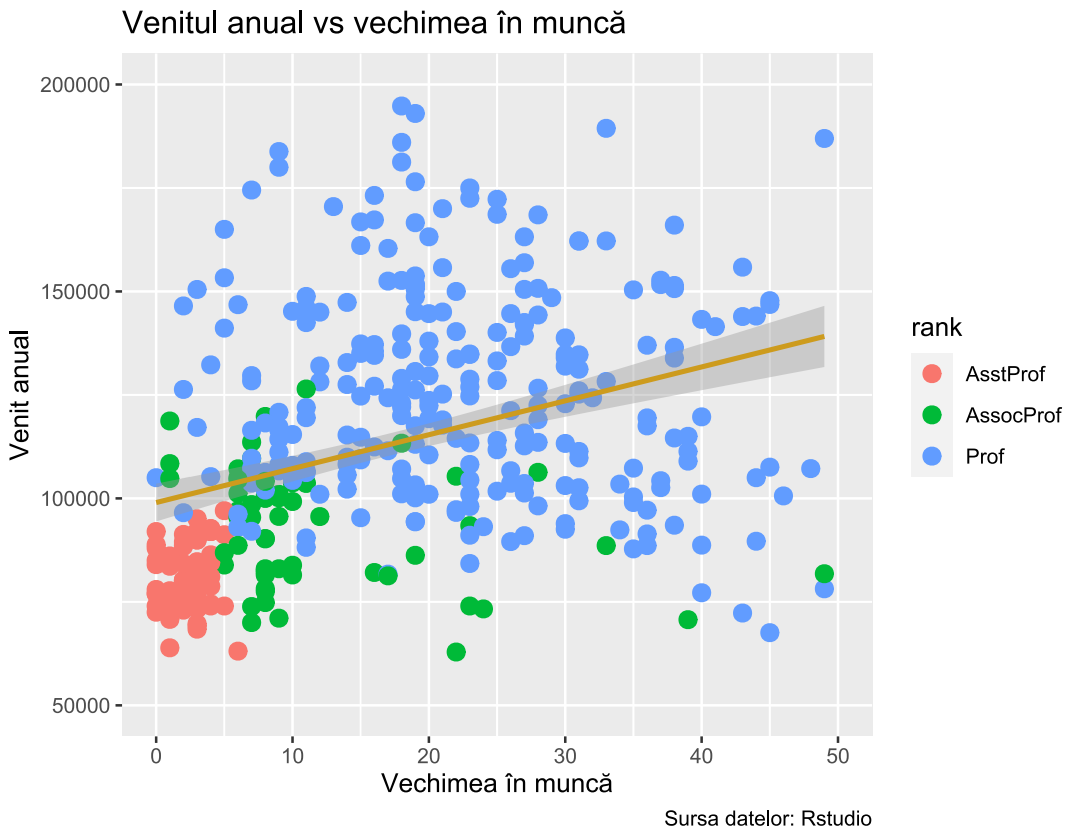
6.2.2.4. Schimbarea culorilor pentru a reflecta categoriile într-o altă coloană

Să presupunem că dorim să schimbăm culoarea pentru grupuri diferite, în funcție de categoria (valoarea) pe care o ia respondentul pe o altă variabilă din setul de date **Salaries**. Pentru a realiza acest lucru trebuie să specificăm în interiorul funcției **aes()**.

```
gg <- ggplot(Salaries, aes(x=yrs.service, y=salary)) +
  geom_point(aes(col=rank), size=3) +  ### Setăm culoarea să varieze în fu
```

```
ncție de poziția respondentului.###
  geom_smooth(method="lm", col="firebrick", size=2) +
  xlim(c(0,50)) + ylim(c(50000, 200000)) +
  labs(title="Venitul anual vs vechimea în muncă", y="Venit anual", x="Vec
himea în muncă", caption="Sursa datelor: RStudio")
plot(gg)
```

Figura 6.7 Diagrama cu puncte, pe grupe, în ggplot2



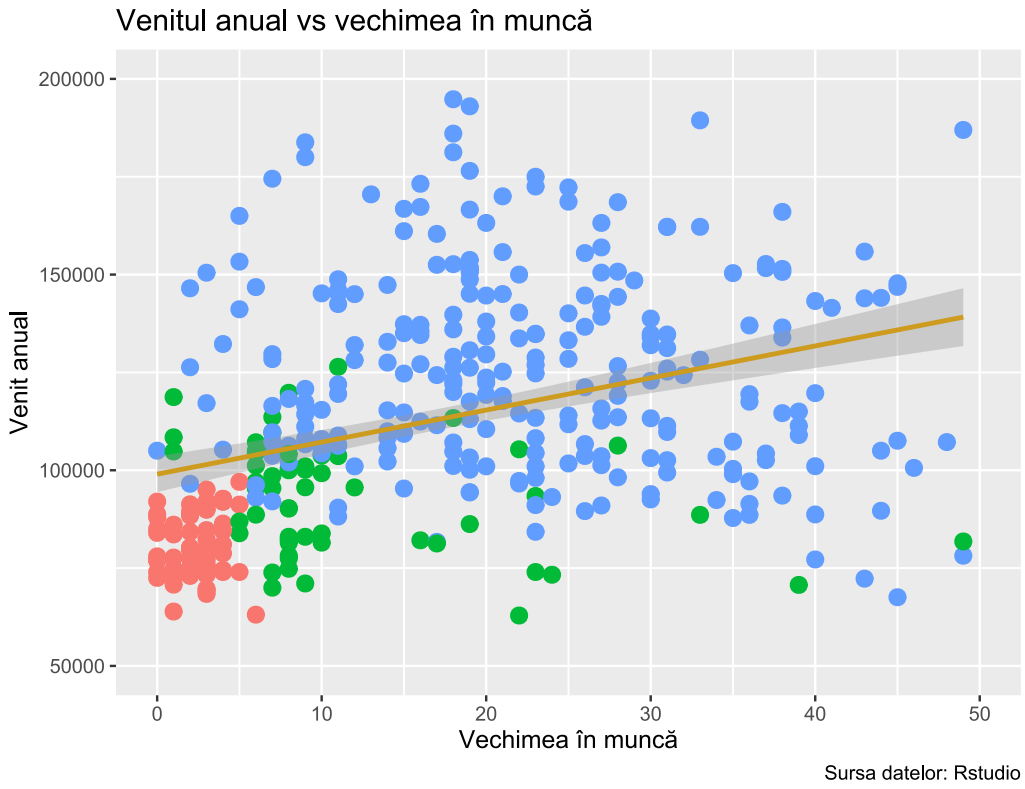
Acum fiecare punct este colorat în funcție de poziția căruia îi aparține folosind **aes(col=state)**. Nu doar culoarea, ci și dimensiunea, forma, conturul (grosimea limitei) și umplerea (culoarea umplerii) pot fi folosite pentru a diferenția grupările.

Un avantaj al acestei împărțiri în funcție de o caracteristică particulară este adăugarea automată a legendei. Dacă este necesar, legenda poate fi eliminată prin specificarea **legend.position** None din cadrul funcției **theme()**.

```
#### Eliminăm legenda###

gg + theme(legend.position="None")
```

Figura 6.8 Diagrama cu puncte, pe grupe, fără legendă, în ggplot2



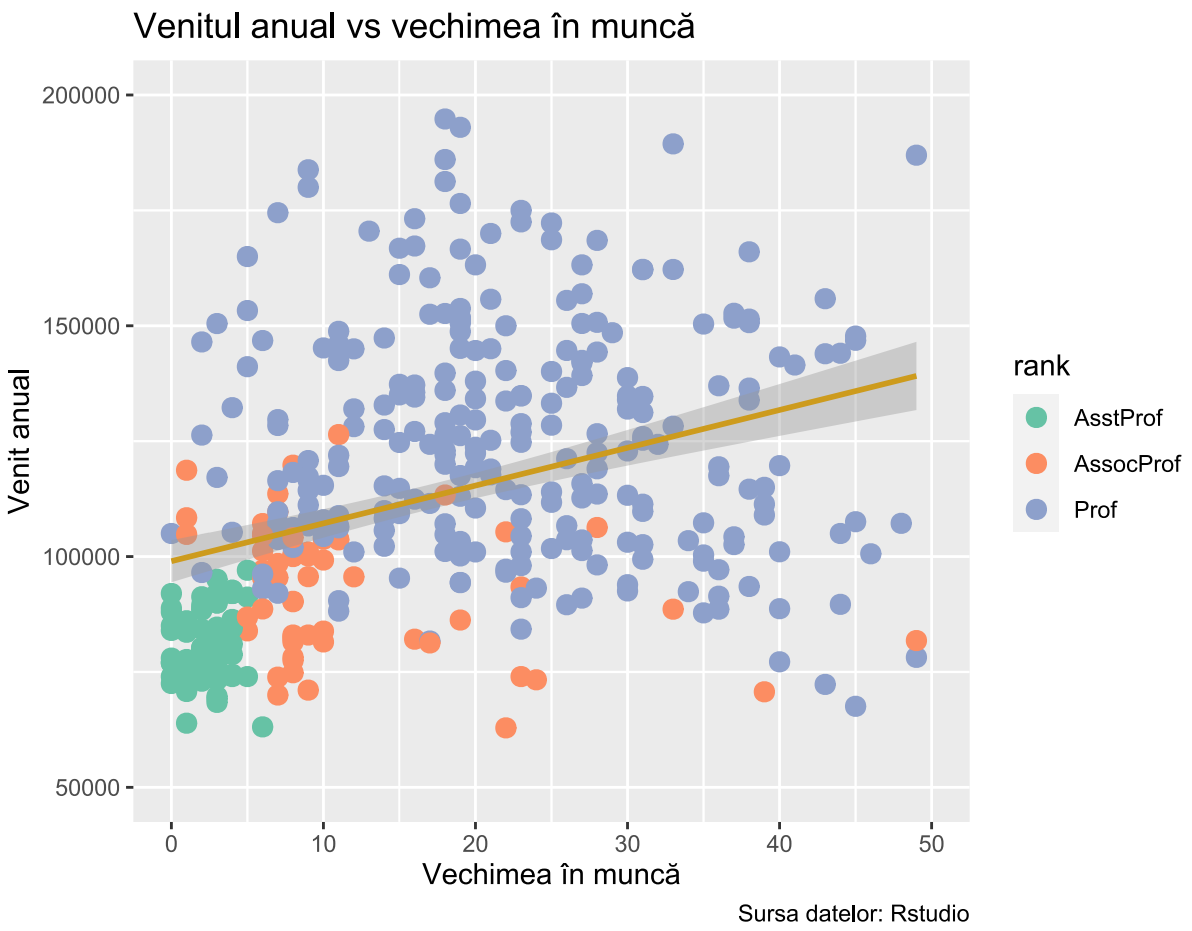
De asemenea, putem schimba complet paleta de culori[58] prin executarea următoarei comenzi:

```
#### schimbăm paleta de culori###

gg + scale_colour_brewer(palette = "Set2")
```

[58] Mai multe astfel de palete pot fi găsite în pachetul **RColorBrewer** (Neuwirth 2022).

Figura 6.9 Diagrama cu puncte, modificare paletă culori, în ggplot2



6.2.2.5. Cum să personalizăm întreaga temă a graficului folosind teme pre-construite?

În cele din urmă, în loc să schimbăm componentele temei individual, putem schimba întreaga temă în sine folosind teme pre-construite. Pagina de ajutor ?theme_bw afișează toate temele încorporate disponibile.

Putem schimba tema graficului prin utilizarea funcției **theme_set()** pentru a seta tema înainte de a desena ggplot-ul. Reținem că această setare va afecta toate graficele viitoare. O altă modalitate de a schimba tema este prin realizarea unui grafic ggplot2 gol și apoi să adăugăm setarea generală a temei (de exemplu, **theme_bw()).**

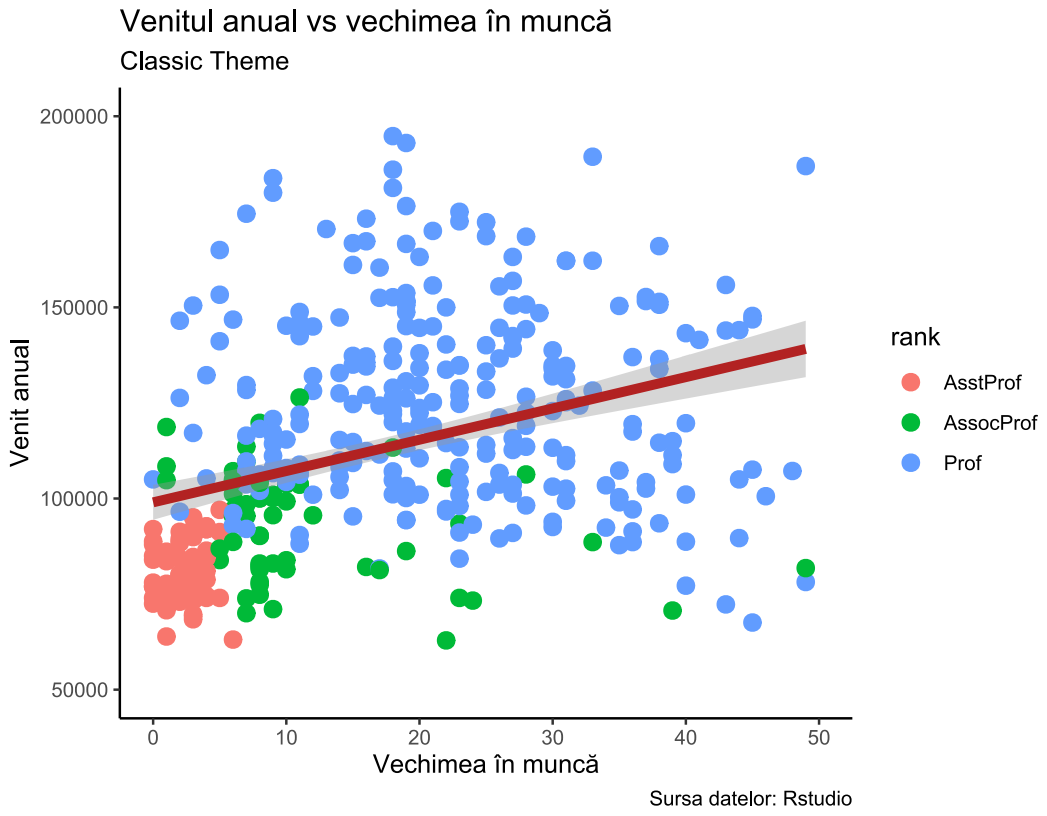
```
#### Grafic de bază###
gg <- ggplot(Salaries, aes(x=yrs.service, y=salary)) +
  geom_point(aes(col=rank), size=3) + ### Setăm culoarea să varieze în fun
cție de poziția respondentului.###
  geom_smooth(method="lm", col="firebrick", size=2) +
  xlim(c(0,50)) + ylim(c(50000, 200000)) +
  labs(title="Venitul anual vs vechimea în muncă", y="Venit anual", x="Vec
himea în muncă", caption="Sursa datelor: RStudio")

### metoda 1: Folosim theme_set()###
theme_set(theme_classic())
gg

### metoda 2: Adăugarea stratului de temă în sine.###
gg + theme_classic() + labs(subtitle="Classic Theme")
```

Pentru mai multe teme personalizate putem accesa pachetul ggthemes (Arnold 2021).

Figura 6.10 Diagrama cu puncte, schimbare temă, ggplot2



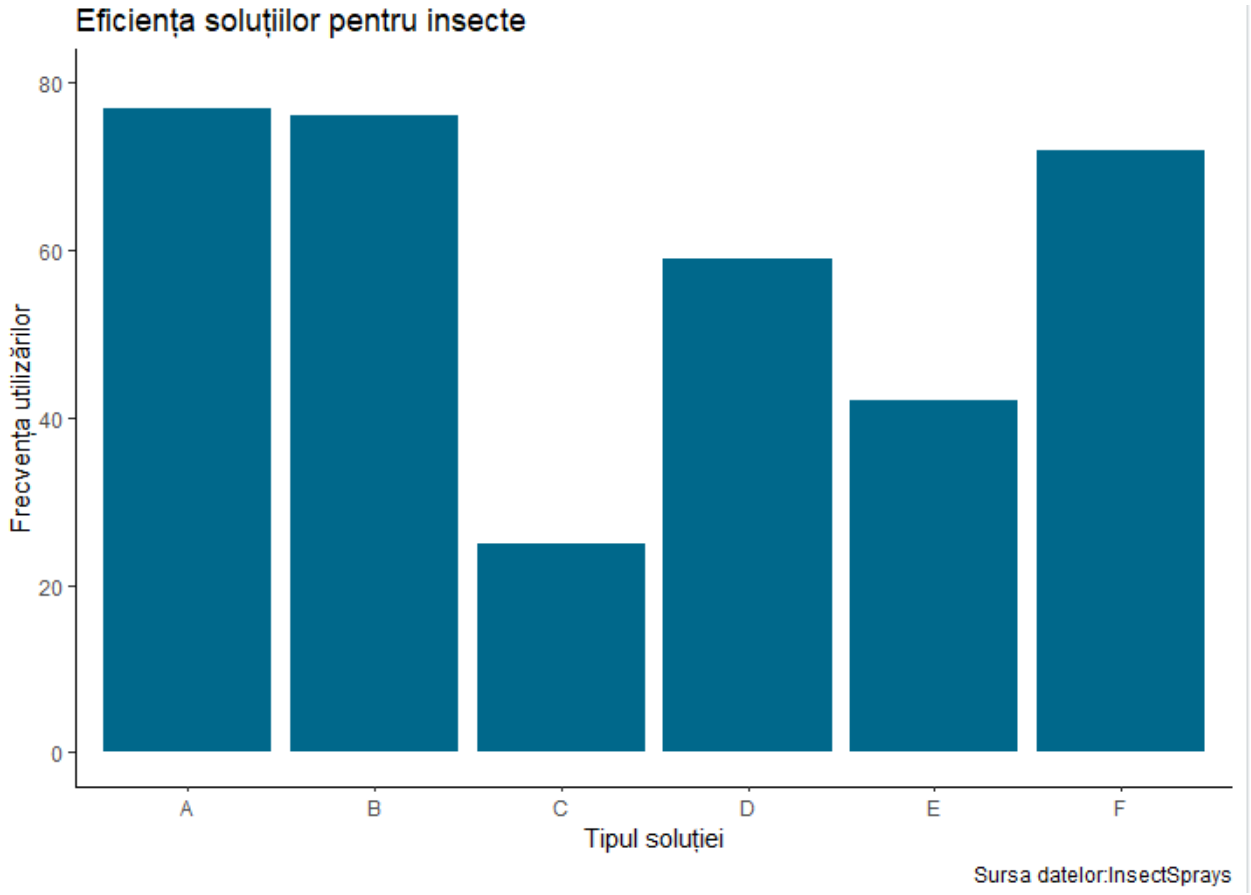
6.2.3. Diagramă cu bare ordonată (*Ordered Bar Chart*)

Diagrama cu bare ordonată (în engleză Ordered Bar Chart) este o diagramă cu bare care este ordonată după variabila de pe axa Y. Doar sortarea cadrului de date după variabila de interes nu este suficientă pentru a ordona diagrama cu bare. Pentru ca diagrama cu bare să păstreze ordinea rândurilor, variabila axei X (adică categoriile) trebuie convertită într-un factor.

Pentru a exemplifica modalitatea de realizare a unei diagrame ordonate folosind ggplot vom utiliza setul de date **InsectSprays.** Să presupunem că suntem interesați să observăm efectivitatea a 6 tipuri de soluții în funcție de numărul de utilizări pentru a se produce efectul dorit. În RStudio vom aplica următoarele comenzi:

```
### Schimbăm dimensiunea barelor###
ggplot(data=InsectSprays, aes(x=spray, y=count)) +
  geom_bar(stat="identity", width=0.5)
### Schimbăm culorile ###
ggplot(data=InsectSprays, aes(x=spray, y=count)) +
  geom_bar(stat="identity", color="deepskyblue4", fill="white")
### Alegem tema și salvăm graficul într-un obiect ggplot numit p###
p <-ggplot(data=InsectSprays, aes(x=spray, y=count)) +
  geom_bar(stat="identity", color="deepskyblue4")
  ylim(c(0, 80)) +
  theme_classic()
p
### Denumim axele și dăm un titlu graficului###
p + labs(title="Eficiența soluțiilor pentru insecte",
         caption="Sursa datelor:InsectSprays",
         y="Frecvența utilizărilor", x="Tipul soluției")
```

Figura 6.11 Diagrama cu bare, în ggplot2



6.2.4. Histogramă

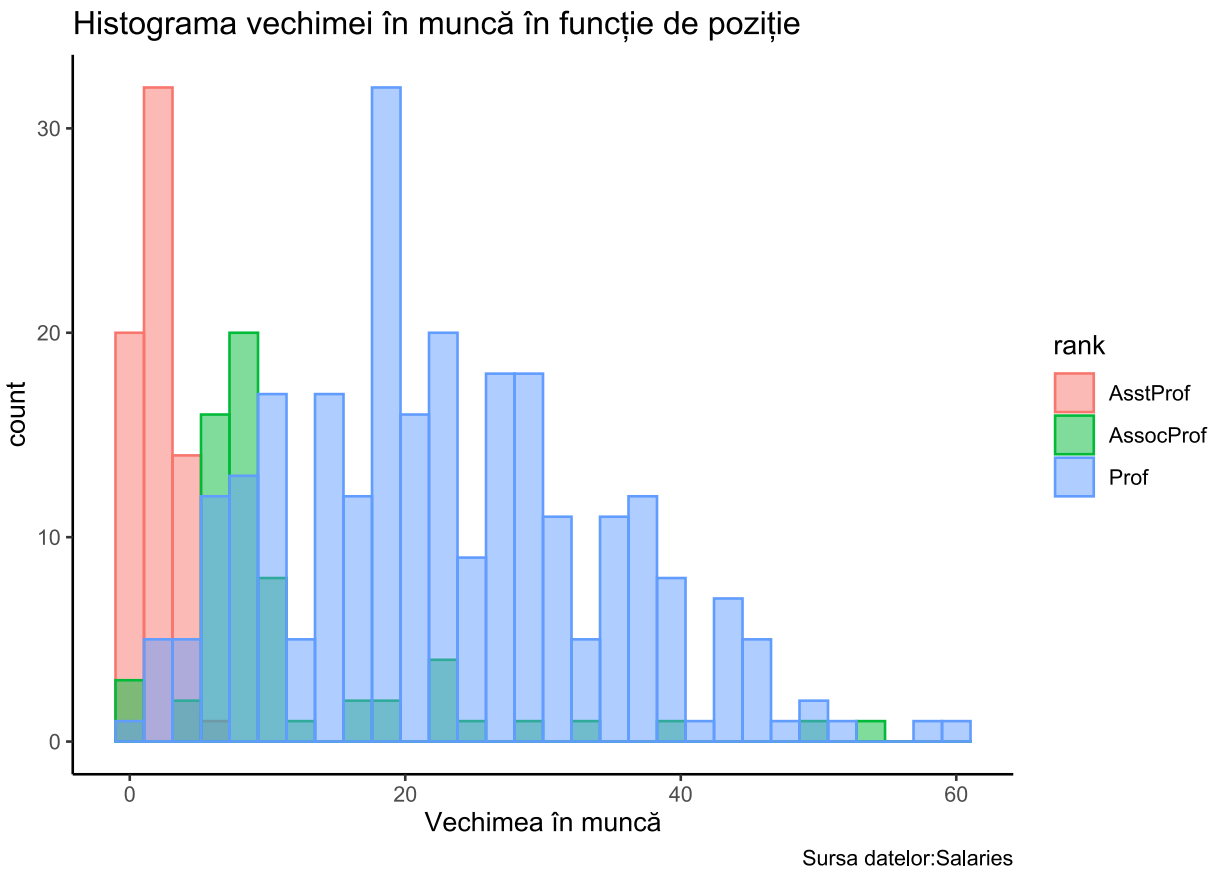
Histograma pe o variabilă continuă poate fi realizată folosind fie **geom_bar()** fie **geom_histogram()**. Vom exemplifica realizarea unei histograme folosind setul de date **Salaries** oferit de pachetul **car** (J. Fox și Weisberg 2019). Să presupunem că dorim să realizăm o histogramă pentru variabila *yrs.service* în funcție de poziția academică. În RStudio vom aplica următoarele comenzi:

```
theme_set(theme_classic())

### Histograma unei variabile continue###
g <- ggplot(Salaries, aes(x = yrs.service, fill = rank, colour = rank)) +
      geom_histogram(alpha = 0.5, bindwidth = .1, position = "identity"),

g +   labs(title="Histograma vechimei în muncă în funcție de poziție",
      caption="Sursa datelor:Salaries",
       x="Vechimea în muncă")
```

Figura 6.12 Histogramă, în ggplot2



6.2.5. Boxplot

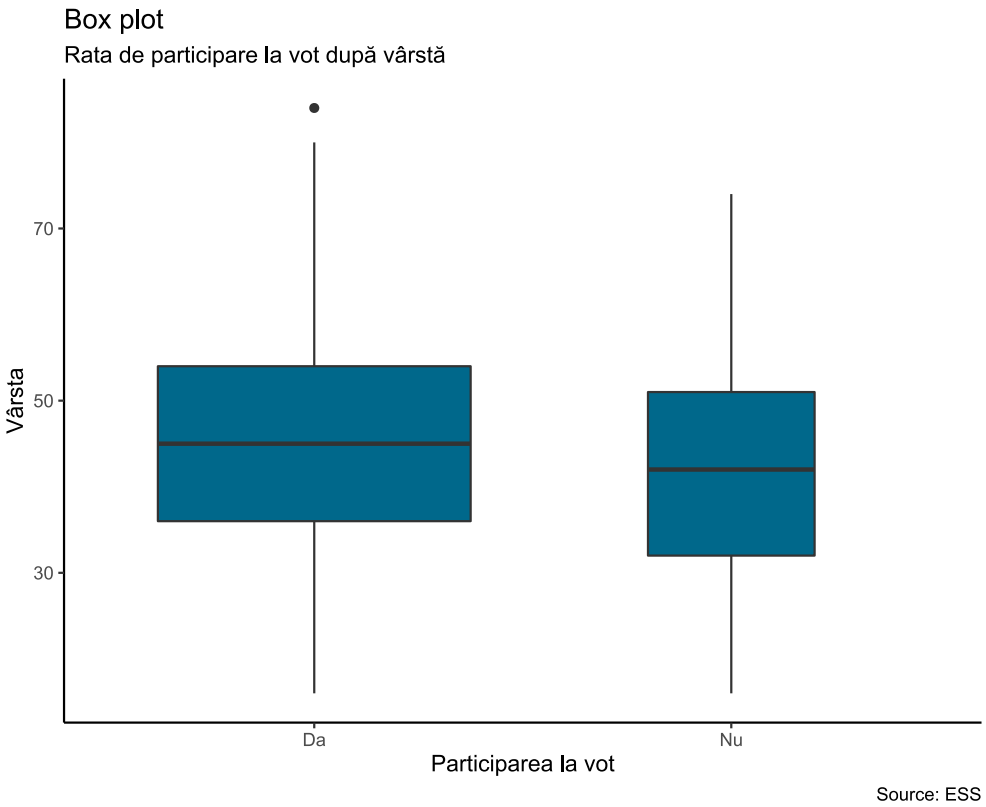
Boxplot este un instrument de vizualizare a datelor foarte utile pentru a studia distribuția acestora. De asemenea, poate afișa distribuțiile în mai multe grupuri, împreună cu mediana, amplitudinea și cazurile deviante, dacă acestea există. Vom exemplifica realizarea unui boxplot folosind setul de date ESS. Să presupunem că suntem interesați în a afla distribuția celor care au votat, grupând indivizii în funcție de variabila vârstă. În RStudio vom aplica următoarele comenzi:

```
data <- read.csv(dataset_exemplu.csv)

theme_set(theme_classic())

### Plot###
g <- ggplot(data, aes(x= vote, y= agea))
g + geom_boxplot(varwidth=T, fill="deepskyblue4") +
    labs(title="Box plot",
         subtitle="Rata de participare la vot după vârstă ",
         caption="Sursa datelor: ESS",
         x="Participarea la vot",
         y="Vârsta")
```

Figura 6.13 Boxplot, în ggplot2



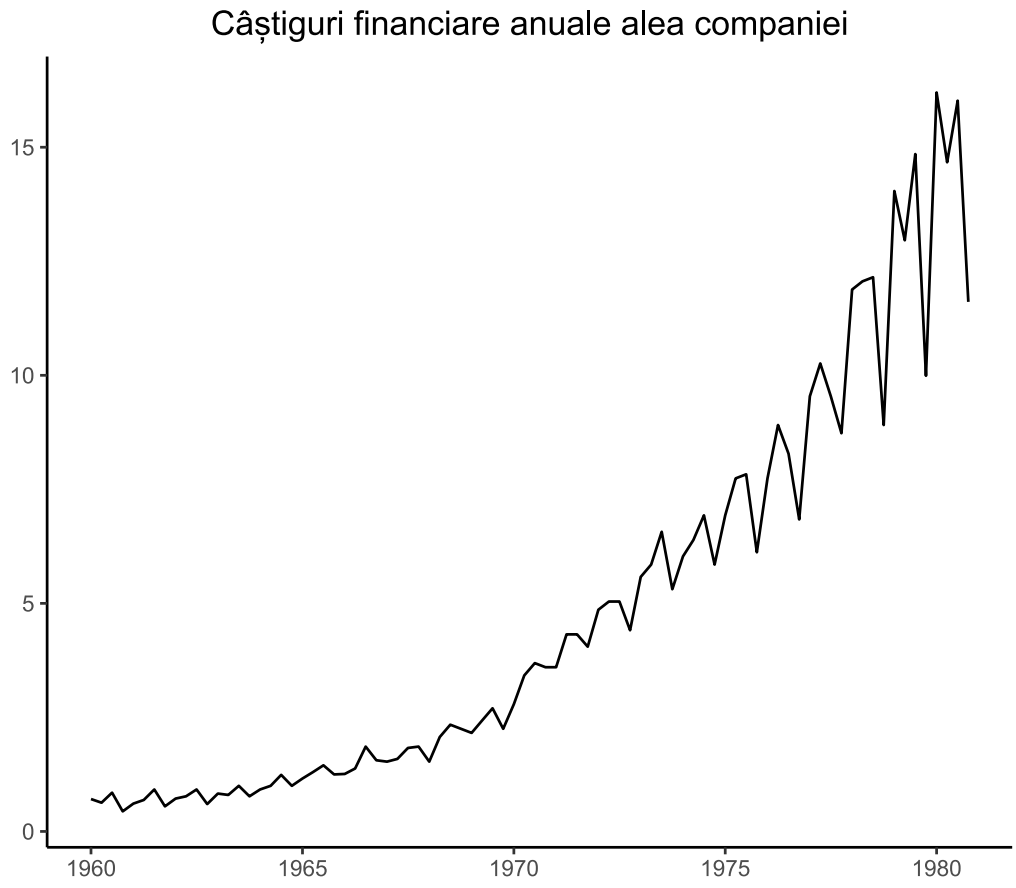
6.2.6. Graficul seriei temporale

Graficele care descriu serii de timp ne ajută să ilustrăm grafic tendințe sau evoluții pe o durată mare de timp într-o manieră organizată. Pachetul ggfortify permite funcției autoplot() să tragă automat direct dintr-un obiect de serie temporală.

```
### From Timeseries object (ts)###
theme_set(theme_classic())

### Plot ###
autoplot(AirPassengers) +
  labs(title="AirPassengers") +
  theme(plot.title = element_text(hjust=0.5))
```

Figura 6.14 Grafic de serie de timp, în ggplot2



7. Metode de analiză calitativă

Datele calitative sunt o categorie aparte de informaţii empirice pe care cercetătorii le culeg pentru a examina realitatea. Aceste date sunt culese şi analizate prin metode specifice de analiză. Metodele calitative folosesc instrumente cum ar fi interviul aprofundat, interviul focus-grup, observaţia, observaţia participativă, analiza de conţinut, analiza de documente video sau audio, naraţiuni auto-etnografice (Leavy 2014). Aceste instrumente sunt folosite în studii care sunt mai degrabă calitative, cum ar fi cele etnografice, studii istorice-comparative, studii de caz, studii *grounded theory* ce au drept scop producerea de noi teorii pornind de la date empirice, studii narative, sau în studii de istorie orală. Totuşi, unele tehnici şi instrumente calitative sunt din ce în ce mai folosite şi în studii cu design şi metode mixte, în care concluziile analizei datelor cantitative primesc substanţă şi sunt întărite de informaţii detaliate oferite de datele calitative (Fearon și Laitin 2009). Metoda calitativă este folosită pentru documentarea despre un caz, o persoană sau un grup restrâns de persoane, despre un context specific identificat într-un cadru temporal şi spaţial restrâns. Această metodă permite descrierea în profunzime a relaţiilor dintre categoriile de observaţii culese.

Putem identifica exemple diverse din punct de vedere al combinării într-un studiu a unor instrumente etnografice variate, în numeroase studii antropologice, dintre care vom menționa câteva pe teme cât se poate de variate, disponibile publicului din România, cum ar fi cele ale lui Liviu Chelcea (2008; 2015), Monica Stroe (2016), Gabriel Jderu (2015) sau Vintilă Mihăilescu (2019). În cartea „Bucureştiul postindustrial" Liviu Chelcea (2008) studiază modul în care transformarea continuă a economiei Bucureștiului secolului XX a produs schimbări în relaţiile urbane între categorii sociale diferite, dar şi modul în care politicile urbane şi de privatizare au condus la reinventări rezidenţiale ale zonelor industriale şi la retrasări ale inegalităţilor sociale. Folosindu-se în principal de analiza documentelor, a arhivelor (fotografii, hărţi), şi documentarea etnografică prin observaţia de teren, acest studiu

oferă un exemplu de analiză de tip *process tracing*. Studiul foloseşte ca o sursă secundară de informaţii şi interviurile semi-structurate. Un alt exemplu de studiu calitativ exhaustiv este cel în care Monica Stroe (2016) documentează modul în care o regiune din Transilvania a fost transformată economic şi cultural de redescoperirea unor tradiţii culinare. Folosind interviurile nestructurate, istoriile personale şi observaţia, analiza documentelor secundare, alături de documentarea proceselor şi transformărilor de la nivelul comunităţii, autoarea explorează efectele pe care aceste redescoperiri gastronomice transcendente diferenţierilor etnice, le-au avut asupra vieţii indivizilor şi a localităţii, dar şi în dezvoltarea economică a regiunii. Un alt studiu calitativ care însă folosește observația participativă, interviurile și istoriile personale, este cel publicat de Gabriel Jderu (2015). În acest studiu, folosindu-se de integrarea personală într-o comunitate relativ închisă, Gabriel Jderu analizează comportamentele motocicliștilor din România, spațiile simbolice și identitățile pe care aceștia și le creează în comunitate.

Studierea unor comunități relativ închise, cunoașterea în detaliu a comportamentului și atitudinilor membrilor unor asemenea comunități, cu greu poate fi obținută în lipsa integrării active a cercetătorului în comunitate. Această strategie de cercetare a unui subiect specific unui grup închis este foarte bine dezvoltată în ceea ce este numită etnografia găștilor (*gang ethnography*), dezvoltată în universitățile din SUA de antropologi și sociologi care au studiat grupurile sociale din orașe precum New York, Boston, Chicago, Atlanta, Los Angeles sau din zona rurală aflată la granița cu Mexic (Thrasher 1963; Brotherton și Barrios 2004; Durán 2018). În acest tip de studii etnografice cercetătorii se integrează în comunitatea studiată, pentru o perioadă lungă, cu scopul de a obține acces la informații sensibile, și pentru a putea observa, documenta și experimenta cele mai mici detalii ale comportamentului individual și de grup. Unele din aceste studii de teren, care folosesc etnografia găștilor, cum ar fi cele al lui Alice Goffman (2014), au fost puternic criticate (Lubet 2015) în special din punct de vedere al respectării standardelor etice de cercetare.

Un exemplu de studiu narativ auto-etnografic ni-l oferă Vintilă Mihăilescu (2019). În cartea sa, bazată pe experiența personală, folosește cu precădere instrumente introspective prin apelul la memoria autobiografică, la istorii comune și observarea fină a celuilalt (medici, asistente, infirmieri, portari, pacienți, aparținători), pentru a

analiza modul în care individul, dominat de relaţii structurale şi instituţionale, dezvoltă mecanisme complexe de apărare, rezistenţă şi supravieţuire. Participarea activă, involuntară dar asumată de cercetător, permite o perspectivă unică asupra subiectului studiat.

Deși focus grupul oferă instrumente calitative puternice, oferind multiple mecanisme de extragere a unor informații utile din discuțiile de grup, după cum vom exemplifica în secțiunea dedicată focus grupului din acest capitol, acesta este mult mai rar raportat (în articole sau cărți) în cercetări științifice efectuate în România în domeniul științelor sociale, comparativ cu interviurile aprofundate. Focus grupul tinde să fie folosit cu precădere în cercetări de piață în scopuri de marketing și comunicare, dar și în studii de evaluare a unor politici publice, cum ar fi politicile de sănătate (Pacurari et al. 2021).

Datele calitative, în comparație cu datele cantitative, se bazează pe informații colectate prin intermediul unei comunicări deschise. Astfel, metodele de cercetare calitativă sunt concepute într-o manieră care încurajează înțelegerea comportamentelor și percepțiilor publicului țintă în ceea ce privește un subiect particular (de exemplu, îngrijirea paliativă). Ambele tipuri de date au un rol esențial în cunoașterea trăsăturilor și caracteristicilor unui anumit set de date. Analiza calitativă a datelor permite cercetătorilor să extragă seturi de date pentru procesul de cuantificare. Prin urmare, putem spune că datele calitative pot constitui o bază pentru colectarea viitoare a datelor cantitative.

Astfel, datele calitative sunt reprezentate de toate tipurile de informații non-numerice care includ documente text, imagini, transcrieri ale interviurilor, note, videoclipuri și înregistrări audio. Ulterior culegerii datelor calitative prin utilizarea acestor instrumente cercetătorul poate demara procesul de analiză a datelor culese. Printre principalele metode de analiză a datelor de tip calitativ amintim: analiza de conținut, analiza narativă și analiza de discurs. În cadrul acestui capitol vom utiliza datele colectate în urma unor discuții de grup și vom aplica analiza de conținut pentru a distinge aspectele principale identificate de participanți. În următoarele secțiuni vom prezenta o serie de instrumente specifice metodelor calitative. Vom discuta despre cel mai utilizat instrument de culegere a informațiilor empirice, interviul

aprofundat. Acesta este preferat în studiile etnografice, dar şi în cele de tip *process-tracing* sau *grounded theory*. În acest din urmă tip de studiu, un alt instrument ce permite colectarea unor informaţii calitative este focus grupul. În ultima secţiune vom discuta despre analiza de conţinut şi analiza de discurs, alte instrumente de culegere a datelor calitative.

7.1. Interviul aprofundat

Culegerea informaţiilor calitative sau cantitative prin interacţiunea activă cu un subiect uman se realizează în cele mai multe cazuri prin interviu (observaţia participativă este o altă tehnică de interacţiune activă) (S. Chelcea 2001). Interviul este folosit nu doar în cercetările calitative, ci şi în cele cantitative, cum ar fi studiile care folosesc sondajul pe eşantioane de subiecţi. Când folosim interviul pentru culegerea unor informaţii perfect structurate cum ar fi cele cuprinse într-un chestionar aplicat unui eşantion de persoane, interviul nu oferă nici o posibilitate intervievatorului sau respondentului, de a devia de la lista şi formatul predeterminat al întrebărilor folosite. Intervievatorul culege informaţiile într-un format predeterminat, consemnând practic răspunsurile în chestionar (pe hârtie, tabletă sau calculator).

În cercetările calitative, prin care ne propunem să culegem informaţii detaliate de la respondenţi interviul este utilizat pentru a cuprinde cât mai multe probleme relevante pentru subiectul cercetat. De aceea, acest tip de interviu este numit interviu aprofundat, interviu comprehensiv. În limba engleză termenii folosiţi sunt *in-depth interview*, sau *long interview* şi ilustrează diferenţa dintre acest tip de interviu, care explorează în detaliu problema studiată, având nevoie de mai mult timp, şi celelalte interviuri de tip unu-la-unu care, de regulă, durează mult mai puţin şi colectează mai puţine informaţii de la subiecţi (McCracken 1988). Informaţiile culese de la respondenţi pot fi şi semi-structurate sau nestructurate. În interviul aprofundat cu ghid semi-structurat cercetătorul foloseşte întrebările ca pe un ghid orientativ, permiţând discuţiei să baleieze în funcţie de subiectul interviului și disponibilitatea conversațională a persoanei intervievate, în anumite limite.

Astfel, atunci când dorim să extindem gradul de detaliere al informaţiilor culese într-o anchetă sociologică, interviul aprofundat poate fi una dintre cele mai bune soluţii. Este instrumentul preferat de cercetători atunci când subiecţii sunt specifici sau, din punct de vedere metodologic, atunci când designul cercetării este unul inductiv (dezvoltarea teoriei se face după culegerea datelor, teoria fiind derivată din date). Interviul aprofundat este folosit cu precădere în studiile etnografice (metodă antropologică de studiere şi cunoaştere a unui grup cultural sau a unei persoane), observaţii (participative sau nu), deoarece este un instrument ce favorizează interacţiunea directă, faţă în faţă, cu persoana şi contextul studiate.

Întrebările folosite în interviuri construiesc împreună un ghid de interviu. În funcţie de gradul de structurare a ghidului de întrebări, interviul poate fi structurat, semi-structurat sau nestructurat. În interviul structurat se folosesc întrebări predeterminate, controlul discuţiei se află în întregime la cercetător. Intervievatorul nu trebuie să schimbe formularea, ordinea sau folosirea în întregime a întrebărilor din ghid. În acest fel, cercetătorul poate controla erorile care pot fi cauzate de comportamentul şi preferinţele intervievatorului din timpul interviului. Avantajul unui interviu structurat constă în gradul ridicat de comparabilitate a informaţiilor obţinute de la subiecţii intervievaţi.

În studiile calitative rareori interviul este structurat. De cele mai multe ori acesta este folosit cu un ghid semi-structurat, care conţine întrebările orientative, folosite în interviu, de cele mai multe ori, aşa cum sunt ele formulate în ghid. Totuşi, ordinea poate să fie adaptată de intervievator ţinând cont de contextul în care se desfăşoară interviul şi de succesiunea informaţiilor pe care le oferă persoana intervievată: dacă discuţia oferă posibilitatea aflării unor informaţii programate ulterior, atunci intervievatorul se va folosi de contextul creat pentru a crea un mod natural de curgere a discuţiei. Astfel, discuţia devine una mult mai liberă şi mai naturală, ceea ce favorizează creşterea încrederii subiectului în intervievator şi permite mai uşor rememorarea unor aspecte trecute relevante pentru discuţie.

Interviul nestructurat se caracterizează printr-o libertate absolută pentru intervievator şi subiect de a conduce discuţia. De regulă, se foloseşte o întrebare sau un context de iniţiere a discuţiei, după care aceasta decurge fără ca intervievatorul să

folosească întrebări predefinite. Acest tip de interviu poate fi folosit atât pentru teme precise, dar pentru care se dorește mai degrabă explorarea decât detalierea și descrierea, cât și pentru teme mai puțin precise, pentru care acest tip de interviu poate oferi o mai bună înțelegere a subiectului. Astfel, în această din urmă situație, interviul are rol mai degrabă de testare a temei, urmând să conducă la restrângerea temei de studiu, la structurarea unor instrumente și la identificarea unor noi subiecți de intervievat.

Organizarea interviului comprehensiv poate cuprinde o serie de etape, pe care Grant McCracken (1988) le-a structurat în colecția devenită clasică, Sage Qualitative Research Methods. Astfel, interviul aprofundat, poate fi folosit nu doar pentru a produce date și a testa teorii, ci și pentru a produce noi teorii.

Pentru o bună fundamentare a interviului McCracken (1988, 29–32) recomandă o documentare analitică a problemei studiate, prin identificarea liniilor principale ale temei, definirea acesteia și identificarea principalelor concluzii, producând astfel, ceea ce în general numim trecere în revistă a literaturii de specialitate, având drept scop identificarea golurilor din teorie, dar și a relevanței subiectului cercetat. Putem opina însă că această etapă, pe care de altfel o parcurge orice tip de studiu, poate ridica însă probleme dacă nu este abordată realist. Încercarea, uneori forțată, de a găsi goluri în teorii, poate să lipsească cercetătorul de obiectivitatea cu care ar trebui să trateze corpul de studii, idei și evidențe produse de alții. Cu alte cuvinte, în încercarea de a fi original, de a publica, un cercetător poate cădea pradă tentației de a critica sau inova (conceptual[59]) de dragul diferențierii analitice, și nu din motive obiective. Pe de altă parte, există riscul ca o problemă mai puțin cercetată să creeze aparența falsă a lipsei de relevanță a temei pe care un cercetător și-ar putea propune să o studieze.

O altă etapă importantă, este auto-examinarea (McCracken 1988, 32–34). În această etapă experiența personală a cercetătorului în ceea ce privește tema studiată îl poate ajuta să construiască întrebări capabile să surprindă mai eficient varietatea

[59] Pentru o discuție aprofundată a inovării conceptuale și a capcanelor produse de aceasta, cum ar fi elasticitatea conceptuală sau inflația de subtipuri conceptuale, recomandăm textele lui Giovanni Sartori (Sartori 1970), David Collier și James Mahon (Collier și Mahon 1993), și David Collier și Steven Levitsky (Collier și Levitsky 1997).

subiectului abordat, dar şi să pregătească mai bine întrebările pentru ghidul de interviu. Astfel, cercetătorul poate identifica relaţiile culturale din interiorul temei şi poate să evidenţieze mai bine propria poziţie faţă de temă. Deşi utilă o asemenea abordare, în lipsa corelării ei cu rezultatele evidenţiate de prima etapă, şi fără o obiectivare a alternativelor teoretice şi empirice, e posibil ca cercetătorul să acorde experienţelor personale un primat epistemologic disproporţionat.

În a treia etapă de organizare a interviului aprofundat construcţia listei de întrebări, numită și ghid de interviu, şi folosirea acestora sunt elementele centrale, conform lui McCracken (1988, 34–41). El recomandă construcţia şi utilizarea unor întrebări care să permită identificarea unor detalii biografice ale persoanelor intervievate, dar şi formularea unor întrebări capabile să surprindă diferite arii din tema studiată. Întrebările nu constituie un cadru fix, ci o jalonare a discuţiilor şi a subiectelor dezvoltate în cadrul acestora. Utilitatea acestei strategii este aceea că ne ajută să evităm anumite subiecte (din motive metodologice sau etice), să evităm distorsiuni deliberate introduse pe parcurs, de respondent sau de context, în desfăşurarea interviului, dar să evităm şi înţelegeri greşite ale unor subiecte.

În ceea ce priveşte implementarea interviului există o serie de reguli elementare pe care le putem folosi. În primul rând crearea atmosferei personalizate fiecărui interviu. Fiecare persoană intervievată este unică, iar informaţiile pe care ni le poate oferi sunt la rândul lor unice, fiind trecute prin propriul filtru de experienţe şi cunoştinţe. În acest fel, putem câştiga încrederea respondentului şi putem obţine informaţii detaliate. Întotdeauna, persoana intervievată trebuie ascultată cu mare atenţie. Astfel, nu doar că îi vom câştiga respectul, dar vom putea surprinde idei şi informaţii relevante pe care le putem dezvolta în interviu. Pe parcursul discuţiei este importantă urmărirea relaţiilor logice dintre termenii cheie şi anticiparea elementelor discuţiei (Ce urmează în mod natural după ce am pus întrebarea aceasta, dincolo de ordinea indicată în ghidul de interviu?). Un alt element important este atenţia acordată tentaţiei persoanei intervievate de a conduce discuţia în direcţia pe care o doreşte ea (pe care noi nu o dorim întotdeauna), dar şi atenţia la tentaţia persoanei intervievate de a se eschiva şi de a nu răspunde la esenţa întrebării (sau la ce ne-am dori noi să răspundă), ci la ce vrea ea (sau la ce înţelege ea din întrebare). Cu toate acestea, persoanei intervievate trebuie să i se lase loc de “manevră”, să i se acorde

libertate în discuție, altfel, interviul se transformă într-un meci de ping-pong la perete (întrebare-răspuns-întrebare-răspuns-...).

O altă recomandare oferită de McCraken pentru implementarea ghidului de interviu este aceea de a juca un rol. În situația unui interviu pe teme sensibile, cercetătorul / intervievatorul poate juca rolul „persoanei neștiutoare" care trebuie "lămurită" de persoana intervievată. Astfel, putem obține mai multe detalii în acel interviu, dar putem identifica legături care altfel ne pot scăpa. De exemplu, în unele dintre cele peste 400 de interviuri calitative aprofundate pe care Aurelian Muntean le-a realizat împreună cu studenții din SNSPA pe teme de clientelism electoral în România, în perioada 2013-2016, a utilizat această strategie pentru a câștiga încrederea persoanelor intervievate și pentru a afla detalii despre derularea competiției electorale din localitate. Fără această abordare temele discutate li s-ar fi putut părea respondenților arhicunoscute, irelevante sau prea sensibile pentru a mai fi menționate.

A patra etapă de diferențiere și evaluare a organizării interviului aprofundat este cea în care determinăm categoriile relevante, relațiile dintre acestea și testăm asumpțiile. Altfel spus, în această etapă vom realiza analiza interviurilor (McCracken 1988, 41–48). Procesul de analizare începe prin transcrierea discuțiilor interviurilor și evaluarea notițelor de teren cu principalele idei pe care intervievatorul le-a preluat la momentul desfășurării interviului. Identificarea legăturilor dintre conceptele principale ale interviului, dintre informațiile oferite și contextele în care acestea au fost plasate de către persoana intervievată, constituie un element important al acestei etape. Concluziile trase în acest fel de cercetător pot fi evaluate din prisma cadrului teoretic evaluat în prima etapă. Nu în ultimul rând, această analiză a interviului poate oferi noi perspective și poate conduce la formularea de noi teorii și ipoteze ce pot fi testate ulterior. Analiza informațiilor interviurilor se poate face atât manual, cât și automatizat, folosind programe specializate, precum Atlas.ti, MAXQDA, sau NVivo. La finalul acestui capitol vom ilustra modalitatea în care putem folosi acest program de analiză, utilizând transcrierea unor interviuri de grup pe teme de politici de sănătate.

Raportarea informaţiilor colectate în interviurile aprofundate trebuie să asigure, de cele mai multe ori, anonimizarea datelor personale, cu atât mai mult cu cât este analizată şi raportată o temă foarte sensibilă sau atunci când persoanele intervievate nu doresc să le fie dezvăluită identitatea. De exemplu, în studiul având tema clientelismul electoral, un broker (intermediar) electoral intervievat faţă în faţă de Aurelian Muntean, într-o localitate de la malul Dunării, în judeţul Teleorman, a relatat modalităţile pe care le foloseşte pentru a se asigura că alegătorii votează cu cine trebuie, adică cu cine le spune el:

> *„L-am schimbat pe fostul primar pentru că mi-a ajuns; a fost un prost și nu știa să fure, efectiv nu știa să fure Domnule!; l-am dus la doctor și i-am spus <<învață-l Domnule să fure, te plătesc, numai dă-i ceva să-l faci fie mai curajos, să fure>>. Nu știa să ne ajute pe noi cei care l-am ajutat să câștige alegerile. [...] Cum mă asigur că votează [notă: alegătorii] cu cine le zic eu și nu altfel? Simplu: au la mine caietul pe datorii; am o echipă de sportivi pe care îi aduc în cantonament la mine, le plătesc tot. Pe unul de aici l-am ținut închis o săptămână. [notă: exprimându-mi nedumerirea și cerând detalii, a clarificat] Da, l-am ținut închis o săptămână în pivniță, ca să își bage mințile în cap. Cel mai sigur mod e însă biblia și banul: pe babe și pe cei în vârstă îi pui să jure pe biblie cu un ban în gură. Le e frică să nu facă ce promit, dacă jură cu banul în gură. Despre unul am aflat că a jurat strâmb pentru că mi-a spus preotul la care s-a dus să se spovedească [...] Eu lucrez și pentru cei de la județ [notă: care candidează în alegerile locale] și pentru cei din parlament [notă: care candidează în alegerile naționale], și pentru cei din localitate, dar și pentru cei din alte localități.”* (Interviu nr. 83, Teleorman, broker, bărbat, interviu realizat la 15 Septembrie 2015) (Mares și Muntean 2015).

Un alt exemplu, este dintr-un studiu de teren în care Aurelian Muntean (2023) a documentat dialogul social şi reprezentarea intereselor angajaţilor din companiile multinaţionale din România. Astfel, pentru a ilustra dificultăţile pe care le întâmpină dialogul social în subsidiarele din România ale acestor companii multinaţionale atunci când managementul local nu se adaptează relaţiilor de muncă din companie, studiul (Muntean 2023) a folosit ilustrarea situaţiei de către un angajat al unei companii multinaţionale din sectorul retail, intervievat prin videoconferinţă:

„Solidaritatea transnațională este mai mare atunci când toate părțile interesate se simt bine. Atunci când unele dintre ele se simt mai slabe, nu există loc pentru solidaritate. Atunci când există o criză, solidaritatea dispare. […] Managerul de țară a venit la filială cu prime mari promise pentru o performanță economică crescută. Astfel, era hotărât să scadă costurile și să maximizeze profiturile cu orice preț. Când a început pandemia COVID-19, a intrat în panică și a judecat greșit implicațiile economice pentru filială. De aceea, a decis să facă reduceri ale costurilor cu forța de muncă. Ulterior, performanța extrem de bună a filialei în timpul pandemiei a subminat raționamentul său. El nu a reușit să administreze în mod corespunzător criza pandemică din companie. […] Această strategie a fost decisivă. I-a arătat directorului general al grupului că liderii sindicali locali sunt persoane rezonabile. El a apreciat faptul că angajații au dorit să rezolve problemele într-un mod privat și calm." (Interviu 4, retail, nivel local, interviu realizat la 09 Aprilie 2021).

De cele mai multe ori interviurile, similar focus grupurilor, sunt înregistrate și ulterior transcrise pentru a putea fi analizate. Informațiile colectate prin interviu pot fi codate și analizate prin metoda analizei de conținut care poate fi realizată fie dintr-o perspectivă cantitativă, fie dintr-o perspectivă calitativă, așa cum vom discuta mai jos. Analiza cantitativă a acestor informații se bazează pe codificarea acestora în conformitate cu temele importante ale studiului identificare de cercetător și se face prin utilizarea unor tehnici precum cele discutate în secțiunea 3.2 a acestui manual. Numărul informațiilor relevante pe care le putem identifica, atât în analiza interviurilor aprofundate, cât și în analiza discuțiilor de tip focus grup, sau în analiza textelor (de conținut sau a discursurilor), depinde de gradul de detaliere a datelor oferite de subiecții sau sursele de informații utilizate în aceste tehnici de colectare și analiză, sau de complexitatea subiectului studiat. De cele mai multe ori vom adăuga mai multe interviuri și mai multe texte în cadrul de analiză pentru a acoperi o varietate cât mai mare de probleme și de opinii. Putem decide să ne oprim din culegerea de noi informații sau din analiza acestora atunci când ajungem la un punct de saturație: constatăm că nu mai obținem informații noi sau diferite de cele obținute deja. Atunci când analizăm informațiile culese ne putem opri atunci când noi relații relevante nu mai sunt posibil a fi identificate cu ajutorul datelor culese. Informațiile redundante

pot fi însă utile pentru a susține și mai mult argumentele deja formulate, sau pentru a asigura validitatea instrumentelor.

7.2. Focus grupul

Studierea problemelor a căror înțelegere poate fi influențată de efectul de grup se realiză prin apelarea la tehnica numită focus grup. Acesta reprezintă o discuţie de grup organizată sub forma unui interviu cu mai multe persoane, sub coordonarea unui moderator, în care încercăm să aflăm opiniile mai multor oameni aflaţi în aceeaşi încăpere, despre un lucru, un comportament sau un fenomen. Prin urmare, o discuţie pe care o avem pe stradă cu mai multe persoane (de exemplu, atunci când documentăm o problemă şi realizăm un interviu cu o persoană în faşa imobilului unde aceasta locuieşte, ca pe parcurs să ni se alăture şi altele), nu este focus grup. Focus grupul se diferenţiază de alte tehnici de intervievare simultană a unui număr relativ mare de persoane, cum ar fi brainstormingul, discuţiile Delphi, interviurile de familie sau observaţia participativă, atât prin numărul de participanţi, cât şi prin moderare, instrumentul folosit sau scopul utilizării sale[60].

În raport cu interviul aprofundat, focus grupul oferă mai puţin timp fiecărui participant, dar are avantajul identificării interacţiunilor de grup între subiecţii participanţi. În comparaţie cu observaţia participativă, focus grupul permite observarea unei interacţiuni între indivizi într-un interval de timp mai scurt, cadrul discuţiilor nefiind însă cel în care subiecţii se află în mod natural. Discuţiile de focus grup sunt organizate sub conducerea unui moderator asistat de una sau două alte persoane. Focus grupul poate permite observarea mecanismului de formare a atitudinilor. Totuşi, focus grupurile prezintă dezavantaje faţă de interviul aprofundat: interacţiunea cu alţi subiecţi poate distorsiona opiniile anumitor participanţi afirmate

[60] Pentru o discuție detaliată a cizelării acestei tehnici de intervievare de către Paul Lazarsfeld și Robert Merton la mijlocul scolului XX recomandăm publicațiile originale ale acestora, dar și ale unor autori precum David Morgan (1996), iar dintre publicațiile în limba română am recomanda lucrările lui Alfred Bulai (2000), sau traducerea cărții lui Richard Krueger și Mary Ann Casey (2005).

în timpul discuţiilor. Acest lucru este valabil îndeosebi pentru subiecţii timizi sau care au în general o atitudine de supunere atunci cânt li se cere opinia. În plus, aceştia vor fi mai reţinuţi în discuţiile de focus grup şi vor interveni rareori, altfel decât la invitaţia expresă venită din partea moderatorului. Interviul aprofundat unu-la-unu rezolvă această problemă oferind tot interviul la dispoziţia unui singur participant. Soluţii la această problemă se pot găsi şi într-un focus grup, dar ele depind aproape în exclusivitate de competenţa şi experienţa moderatorului, de capacitatea acestuia de a observa cu atenţie şi din timp atitudinile şi reticenţele unor participanţi, dar şi de a identifica cele mai bune metode prin care subiecţii timizi pot fi (re)activaţi. Exerciţiile în grupuri restrânse, despre care vom vorbi mai jos în această secţiune, pot oferi asemenea pretexte pentru a rupe timiditatea unor participanţi.

De multe ori focus grupul este folosit în studii mixte pentru a suplini neajunsurile altor tehnici de colectare a informaţiilor de la subiecţi, cum ar fi de exemplu prin sondaj de opinie (Morgan 1996), ale unor date statistice care nu oferă detalii la nivel individual sau ale unor interviuri individuale care nu oferă informaţii directe despre relaţiile dintre indivizi şi dinamica acestora. Focus grupurile sunt folosite şi pentru a identifica probleme specifice temei studiate, astfel, este pot juca rolul unei pretestări a unor instrumente de colectare a informaţiilor.

Participanţii la focus grupuri sunt, de regulă, selectaţi în funcţie de problema studiată. Prin urmare, rareori vom folosi eşantionarea probabilistă şi mai degrabă vom selecta viitorii participanţi la focus grup prin diverse metode nealeatorii, folosind eşantioane intenţionale pe baza unor liste furnizare de instituţii sau companii implicate în studiu, liste de clienţi sau participanţi la alte evenimente, liste de abonaţi, anunţuri publice în mass-media (Krueger și Casey 2005, 101–10) sau direct de pe stradă cu ajutorul unor operatori de teren (Bulai 2000, 39–43).

Aceste distorsiuni pot fi ţinute sub control sau minimizate prin selecţia participanţilor pe baza unor matrici de selecţie care permit variaţia subiecţilor participanţi pe un număr de variabile relevante pentru acel studiu. Obligativitatea identificării unor participanţi care să aibă toate caracteristicile folosite într-o asemenea matrice reduce eroare de selecţie. În tabelul 7.1 de mai jos prezentăm un model de matrice de selecţie a participanţilor, cu subiecți locuitori în zonă urbană, cu venit

mediu. Pe lângă matricea de selecţie, subiecţii sunt selectaţi şi invitaţi la interviul de grup cu ajutorul unui chestionar, redus ca întindere, pe care îl folosim pentru a ne asigura că participanţii pe care îi vom invita respectă criteriile de selecţie precizate în matrice. Aceste criterii se referă la vârstă, gen, ocupaţie, venit, sau educaţie, şi sunt transpuse în întrebări în chestionarul de selecţie a participanţilor. Dacă subiecţii chestionaţi acceptă să participe la discuţiile focus grup la data şi ora prestabilite, este de preferat să li se înmânează o primă invitaţie, scrisă, unde sunt precizate locul şi ora desfăşurării focus grupului. Pentru a putea confirma participarea, subiecţii selectaţi sunt contactaţi, de cele mai multe ori prin telefon, cel puţin cu o zi înainte de data organizării focus grupului. De aceea, este recomandabil să preluăm datele de contact ale subiecţilor selectaţi prin chestionarul de preselecţie: nume, număr de telefon, email. Date fiind regulile stricte de colectare a datelor personale generate de legislaţia GDPR, întotdeauna organizatorii focus grupurilor trebuie să se asigure că respectă aceste reguli atunci când colectează şi procesează datele personale ale participanţilor în vederea contactării acestora sau a recompensării pentru participarea la discuţiile focus grup.

Tabel. 7.1 Matrice de recrutare pentru focus grup

	Sex		Vârstă			Educație		Ocupație			
Respondenți	B	F	18 - 30	31 - 45	45 - 65	Superioară	Liceu	Angajat	Pensionar	Student	Șomer temporar
Respondent 1	+										
Respondent 2	+										
Respondent 3	+										
Respondent 4	+										
Respondent 5	+										
Respondent 6	+										
Respondent 7	+										
Respondent 8		+									

Respondent 9		+									
Respondent 10		+									
Respondent 11		+									
Respondent 12		+									
Respondent 13		+									
Respondent 14		+									
Respondent 15		+									
Total	**7**	**8**	**5**	**5**	**5**	**5**	**10**	**10**	**3**	**1**	**1**

După cum putem observa în matricea de recrutare, numărul de persoane invitate să participe la focus grup este de 15. Dat fiind că ne putem aştepta, în mod rezonabil, ca unele persoane invitate, care au confirmat participarea chiar şi cu o zi înainte de focus grup să nu mai poată participa din diverse motive este de preferat să creştem numărul de invitaţi peste limita numărul de persoane care dorim să participe la discuţiile propriu-zise. Din acest motiv, este obligatorie completarea cotelor folosite în matricea de recrutare astfel încât acestea să fie proporţionale în raport cu matricea de cote stabilite pentru participanţii la focus grup şi, astfel, să ne permită asigurarea unor rezerve de persoane invitate pentru fiecare categorie de participanţi. Pentru a minimiza numărul de persoane recrutate care renunţă să participe, dar şi pentru a compensa timpul pe care participanţii ni-l acordă pentru focus grup, se recomandă recompensarea acestora. Recompensarea poate motiva suplimentar persoanele invitate, mai ales pe cele care sunt mai ocupate. Această recompensare nu trebuie să fie exagerată, dar nici prea mică (Bulai 2000; Krueger şi Casey 2005). Dacă este prea mare, poate distorsiona opiniile participanţilor. Aceştia se pot auto-cenzura în dinamica discuţiilor de grup sau pot să nu fie sinceri în opiniile exprimate. Dacă plata este prea mică, creşte riscul refuzului sau chiar conduce la renunţarea de a participa.

Numărul de participanţi la un focus grup poate varia în funcţie de locaţia unde se desfăşoară întâlnirea. Deşi ne-am dori să avem cele mai potrivite şi mai dotate săli, special amenajate pentru discuţii focus grup, de exemplu cu masă rotundă, oglindă

unidirecţională, antifonare sau sistem video cu circuit închis, existenţa unor camere secundare de lucru pe echipe, vom putea organiza discuţiile în condiţii acceptabile şi în săli ad-hoc amenajate. Numărul participanţilor poate varia şi în funcţie de subiectul discuţiei. Unele teme sunt sensibile (de exemplu, îngrijirea pacienţilor aflaţi în fază terminală) iar numărul persoanelor dispuse să participe la asemenea discuţii de grup poate fi mai degrabă mic. În literatura de specialitate numărul diferă de la minim 5 la maxim 12 (Bulai 2000, 31), între 10 şi 12 sau între 6 şi 8 (Krueger și Casey 2005, 99), în funcţie de tema focus grupului. Deoarece durata totală de desfăşurare a discuţiilor de grup poate ajunge la aproximativ 2 ore, creşte probabilitatea ca unele persoane să nu participe. În plus, alţi factori pot interveni, precum vremea, oboseala acumulată înainte de ora programată a focus grupului, distanţa de acasă sau serviciu la locul de desfăşurare a întâlnirii etc. Prin urmare, rareori un focus grup planificat pentru 12 persoane va reuşi să beneficieze de participarea tuturor acestora. Un alt motiv pentru care este recomandabil să includem în focus grup aproximativ 12 persoane este acela că inevitabil vom avea un grup în care cel puţin 1-2 persoane vor fi foarte tăcute, ceea ce înseamnă că timpul alocat pentru răspunsurile lor va fi extrem de redus. O durată de 2 ore poate fi folosită şi pentru a organiza activităţi în sub-grupuri (cu posibilă alocare aleatorie în grupuri), astfel încât să poată fi trase concluzii ce pot susţine sau contrazice constatările de la nivelul grupului total de participanţi. Aceste activităţi de grup pot ajuta la clarificarea unor probleme sensibile dar importante pentru subiectul studiat, prin urmare, aceste exerciţii nu ar trebui planificate la finalul focus grupului, putând fi folosite pentru a determina discuţii analitice şi problematizări din partea participanţilor în a doua jumătate a focus grupului.

Indiferent de numărul de participanţi, dacă focus grupul nu are drept scop identificarea de probleme de grup în interiorul unei companii sau instituţii, este recomandabil ca participanţii să nu se cunoască între ei sau cel puţin să nu fie rude sau buni prieteni (atunci când populaţia localităţii din care selectăm participanţii este redusă). În cazul focus grupurilor ce au drept scop identificarea unor probleme organizaţionale este important ca, deşi participanţii se cunosc, aceştia să nu fie în relaţii ierarhice În caz contrar, discuţiile pot fi distorsionate de relaţiile apropiate preexistente. Unii participanţi se pot simţi stânjeniţi în prezenţa unor rude apropiate sau prieteni atunci când discută diverse probleme (cum ar fi, de exemplu, probleme

politice), iar gradul de independență în exprimarea propriilor convingeri poate fi mai redus. Prin urmare, într-o asemenea situație este de preferat să avem un grup omogen, nu eterogen de participanți la focus grup. Atunci când tema focus grupului necesită un control ridicat pentru variabile socio-demografice, cum ar fi educaţia sau venitul, este indicat să asigurăm un echilibru al caracteristicilor participanţilor în funcţie de educaţie, venit şi ocupaţie. Astfel, vom evita exacerbarea unor efecte de grup sau de turmă în exprimarea opiniilor. Nu în ultimul rând, pentru a evita efecte precum distorsiuni cauzate de participanţi „profesionişti" la focus grupuri, nu vom invita persoane auto-selectate(invitate), persoane care au mai participat în ultima perioadă la asemenea discuţii sau alte forme de cercetare socială şi de marketing sau persoane care lucrează în domenii precum cercetarea socială şi de marketing.

Tabel. 7.2 Matrice bidimensională de compoziţie omogenă a unui focus grup

SEX \ VÂRSTĂ	18-30	31-45	46-65	**Total**
Bărbați	2	2	2	6
Femei	2	2	2	6
Total	4	4	4	**12**

Similar interviului aprofundat, focus grupul utilizează un instrument specific, numit ghid de focus grup. Acesta este format dintr-o listă de întrebări deschise, dar spre deosebire de interviul aprofundat nestructurat sau semi-structurat, focus grupul foloseşte aproape întotdeauna ghidul structurat. Când trebuie să modereze discuţii cu 10-12 persoane, oricât de experimentat şi de competent ar fi moderatorul, acesta are nevoie de o planificare la minut a discuţiilor şi de proceduri clare de desfăşurare a focus grupului pentru a putea gestiona eficient timpul alocat fiecărui participant, fiecărei întrebări, fiecărui exerciţiu de sub-grup. Prin urmare, ghidul de focus grup nu conţine doar întrebările împărţite pe categorii clare de probleme ale subiectului studiat, ci şi instrucţiuni specifice despre ce trebuie să spună şi să facă moderatorul. Deoarece timpul alocat discuţiilor de grup nu depăşeşte 2 ore, la un număr de maxim 12 participanţi, nu vom putea folosi mai mult de aproximativ 15 întrebări în ghidul de

focus grup. Întrebările folosite sunt mai degrabă simple și scurte, pretestate și ordonate logic, astfel încât să poată fi înțelese de toți participanții. În ANEXA de la final vom exemplifica configurarea unui ghid de focus grup.

Pe lângă clarificările metodologice de selecție a participanților la focus grup și de construcție a ghidului de focus grup, un element important, care diferențiază focus grupul de interviul aprofundat îl constituie specificul metodologic de desfășurare a discuțiilor de grup. Adițional discuțiilor în care fiecare participant la focus grup își exprimă opinia pentru fiecare întrebare, ghidul de focus grup poate include activități interactive, exerciții sau teste de grup, evaluarea unor grafice sau ilustrații. Discuțiile din focus grup sunt de cele mai multe ori înregistrare atât audio (se recomandă utilizarea a cel puțin două reportofoane amplasate în zone diferite, la distanță relativ egală de toți participanții, eventual utilizarea unui al treilea reportofon de rezervă). Înregistrarea video ajută la transcrierea discuțiilor dar permite și vizualizarea și identificarea dinamicilor de grup și sub-grupuri. În funcție de scopul focus grupului, asistenții moderatorului pot lua notițe (de exemplu, referitor la implicarea afectivă a participanților sau coeziunea de grup) în timpul desfășurării discuțiilor, care ulterior vor fi utile în identificarea mai rapidă a unor elemente importante din discuții.

Analiza discuțiilor focus grup se poate face în mare măsură în același fel ca analiza interviului aprofundat. Transcrierea discuțiilor focus grup oferă principalul material de analiză. Pentru creșterea validității de conținut, codificarea inductivă a principalelor subiecte discutate și a opiniilor participanților se poate face în mai multe etape, folosind cercetători diferiți, astfel încât eroarea determinată de preferințele și subiectivismul cercetătorului pentru anumite teme sau construcții semantice să fie cât mai redusă. Codificarea transcrierilor poate astfel să se concluzioneze cu identificarea unor modele tematice principale. Discuțiile din focus grup sunt folosite fie ca elemente citate în publicarea concluziilor cercetării, uneori pentru a întări rezultatele analizelor cantitative, fie sunt prezentate printr-o analiză de frecvențe sau grafică, separată, efectuată de regulă cu ajutorul unor programe speciale, cum ar fi Atlas.ti, MAXQDA, NVivo sau Tropes. Aceste programe permit identificarea mai rapidă a unor coduri (cuvinte sau expresii) a valorilor semantice ale acestora, dar și a relațiilor dintre aceste coduri, a succesiunilor discursive, a relațiilor semantice și morfologice dintre coduri, sau a frecvenței și poziției acestor coduri în ansamblul discuțiilor din focus grup.

La finalul acestui capitol vom ilustra cu ajutorul Nvivo modul în care putem analiza informaţiile culese în discuţiile focus grup. Acest model de analiză nu diferă de analiza pe care o putem face informaţiile culese în interviuri aprofundate, folosind programele software dedicate. În general, aceste programe ne vor oferi instrumente de codificare, analiză cantitativă şi calitativă şi vizualizare, pe care le putem folosi în mod similar pentru marea majoritate a datelor calitative în care informaţia culeasă este de tip text. Deşi extrem de utile, aceste programe nu pot înlocui factorul uman, care este principala sursă a clasificărilor unităţilor semantice incluse în analiză. În plus, cele mai multe dintre aceste programe specializate în analiza textelor, se bazează pe dicţionare configurate şi actualizate în mod constant pentru un număr foarte redus de limbi, de regulă engleză, franceză, spaniolă. Aceste dicţionare permit analize aprofundate, capabile să perceapă legăturile semantice dintre texte, clasele morfologice, sau secvenţele discursive. Prin urmare, analiza textelor publicate în limbi mai puţin utilizate, cum ar fi româna, nu poate beneficia de cele mai avansate instrumente de analiză calitativă, în lipsa integrării unor dicţionare în limba română.

Analiza de tip cantitativ a focus grupului se bazează pe examinarea informațiilor din prisma recurenței lor în discuțiile de grup. Recurența unor anumite elemente (de exemplu, în focus grupul pe care îl vom folosi pentru exemplificarea analizei în NVivo, problemele recurente au fost lipsa accesului la medicamente) sunt ulterior grupate prin intermediul unui mecanism de codificare (de exemplu, bariere în accesul la medicamente) care la rândul lor pot fi grupate în modele tematice. Acest sistem de codificare poate fi cantitativ în sensul în care evidențiază frecvența cu care o temă este discutată mai mult decât alta și astfel putem trage concluzii cantitative de tipul *tema referitoare la accesul la medicamente are o recurență de 67% în comparație cu celelalte bariere identificate în cadrul discuțiilor de grup.* Sistemul de codificare cantitativă a datelor culese prin intermediul discuțiilor de grup este realizat prin parcurgerea unor pași precum:

1. Decizia asupra unității de analiză asupra căreia vrem să tragem concluziile (de exemplu, întregul grup, dinamica de grup, participanții în mod individual etc.);

2. Dezvoltarea unui sistem de codificare propriu. Figura 7.3, de la sfârșitul acestui capitol, oferă un exemplu de sistem de codificare cantitativă;

3. Aplicarea codurilor în mod sistematic asupra tuturor transcrierilor obținute în urma discuțiilor de grup. O exemplificare a modalității de aplicare a codurilor poate fi observată în Figura 7.3;

4. Realizarea unor tabele cu frecvența codurilor.

7.3. Analiza de conținut și analiza de discurs

Analiza de conținut și analiza de discurs reprezintă două tehnici de analiză foarte asemănătoare. Pentru a le diferenția, putem observa că analiza de conținut se poate realiza pe orice tip de text, în vreme ce analiza de discurs se realizează pe un tip specific de texte, cum ar fi discursuri sau poziții publice ale unor persoane publice, de regulă de rang înalt, ale căror opinii produc efecte în societate. Analiza de conținut se realizează cu ajutorul unor instrumente cantitative și calitative în baza unor date calitative compuse din texte predeterminate produse de persoane sau instituții (R. P. Weber 1990; Gee 1999; Jenner și Titscher 2000; Jørgensen și Phillips 2002; Drisko și Maschi 2016; Krippendorff 2018). Spre deosebire de interviul aprofundat și de focus grup, analiza de conținut se bazează pe texte pre-existente și are la bază tehnici mixte de analiză. Sursele principale ale acestor texte sunt reprezentate de mass-media, internet, documente de arhivă, documente oficiale ale unor instituții publice sau actori politici, programe politice, discursuri ale politicienilor.

Aceste texte pot fi analizate folosind instrumente cantitative precum frecvența aparițiilor unor cuvinte sau expresii, variația dihotomică a acestora (pozitive/negative; stânga/dreapta etc.), clasificarea conținutului folosind scheme de codificare (R. P. Weber 1990). Astfel, în analiza de text vom folosi cu precădere instrumente descriptive, precum cele prezentate în capitolul anterior. Prin utilizarea codificării și analizelor statistice, analiza de conținut urmărește maximizarea validității codificării

textului, dar și a rezultatelor analizei. În analiza de conținut, cercetătorii urmăresc testarea validității constructelor teoretice și analitice prin validitatea intuitivă oferită de plauzibilitatea operaționalizării, prin validitatea socială raportată la factori culturali, sociali sau politici, prin validitate empirică evaluată prin raportarea instrumentelor la date empirice noi, alternative (Krippendorff 2018). În ciuda acestor limitări metodologice, analiza de conținut rămâne una dintre cele mai folosite tehnici de analiză calitativă.

Pentru analiza de conţinut cercetătorilor le este recomandat (R. P. Weber 1990) să definească unităţile de codificare (cum ar fi, de exemplu, cuvântul, sensul, propoziţia, paragraful, tema sau textul în întregime atunci când unitatea de analiză este reprezentată de text, iar pentru analiză folosim mai multe texte comparabile), să definească categoriile codificărilor, să verifice validitatea acestora, să pre-testeze codificările şi categoriile, să codifice toate textele incluse în analiză, să verifice acurateţea analizei atât cu ajutorul altor cercetători (prin *peer-review* sau *inter-rater / inter-coder agreement*), cât şi folosind programe software specializate.

Rezultatul acestui tip de analiză de conţinut poate fi constituit din distribuţii numerice a cuvintelor, măsurarea exactă a lungimii frazelor, identificarea poziţiei acestora în pagină şi în corpul textului, evaluarea numărului de propoziţii utilizate în text sau complexitatea vocabularului. Un tip aparte de utilizare a analizei de conţinut este compararea mai multor texte şi identificarea unui grad de similitudine între acestea. Sintetic, acest tip de analiză de conţinut porneşte de la un text de referinţă pe care îl compară cu un număr ridicat de texte care pot fi mai mult sau puţin asemănătoare. Implementarea practică a analizei de conţinut se regăseşte şi în programe specializate, cum ar fi cele de similitudine de text care sunt utilizate pentru a măsura cantitativ elemente de plagiat.

O altă abordare a analizei de conținut și de discurs este cea calitativă și are la bază evaluarea și interpretarea modului în care testele analizate au fost scrise, a mesajului comunicat de ele și a modului de interpretare a sensului acestuia, a motivației sursei care emite sau produce acele texte, a scopului emiterii acelui mesaj, și a destinatarului acestuia. Analiza de conținut de tip calitativ încearcă să ofere informații nu doar despre modul în care indivizii discută despre lume, ci informații și

despre modul în care indivizii trăiesc și simt anumite situații. Astfel, analiza de conținut calitativă suplinește analiza de tip cantitativ prin faptul că urmărește:

1. Să observe viața participanților prin intermediul informațiilor pe care aceștia le oferă în cadrul discuțiilor de grup;
2. Să ofere interpretări ale aspectelor particulare ale contextului studiat (de exemplu, accesul la resurse medicale) din perspectiva participanților;
3. Să prezinte informațiile observate despre anumite fenomene sociale identificate în cadrul discuțiilor de grup prin ilustrarea anumitor fragmente din cadrul discuțiilor de grup (Silverman 2020).

Comparativ cu alte tehnici şi instrumente de analiză, cercetătorul are, desigur, libertate maximă în a stabili modul în care defineşte categoriile şi le analizează, putându-le schimba în funcţie de scopul cercetării. Aceasta fundamentează însă şi una dintre criticile la adresa acestor tipuri de analize: lipsa de fidelitate a instrumentelor, lipsa de structurare a acestora, validitatea internă şi externă incertă şi greu de cuantificat.

Analiza de conţinut este utilă în compararea unor texte diferite. Analiza de conţinut şi de discurs care foloseşte tehnici şi instrumente cantitative permite rezumarea şi concentrarea unui număr mare sau foarte mare de informaţii, comparaţia între texte care în aparenţă sunt diferite. Acest tip de analiză necesită însă un timp mai lung de pregătire a textelor. Acestea ar trebui să poată fi prelucrate OCR (recunoaştere optică a caracterelor) astfel încât şi un text scanat să poată fi integrat în programele de analiză calitativă. Uneori poate exista un volum mare de texte care, teoretic, pot fi analizate pentru a explica tema cercetată. La fel ca în cazul metodelor cantitative, în analiza de conţinut putem folosi eşantionarea pentru a reduce numărul prea mare de cazuri (texte). Metoda de codificare a informaţiei relevante şi metoda de analiză cantitativă a textelor culese pot beneficia de puterea mărită de calcul oferită de calculatoare şi programe specializate, cu ajutorul cărora chiar şi cele mai lungi texte pot fi organizate şi codificate. La fel ca în cazul analizei transcrierilor interviurilor aprofundate sau focus grupurilor şi în cazul analizei de conţinut şi de discurs putem folosi programe specializate cum ar fi NVivo. Acest tip de programe ne ajută să organizăm textele pe baza unor coduri prestabilite (de regulă, de cercetător) şi să identificăm legăturile statistice între aceste coduri.

În ciuda utilizării unor tehnici şi instrumente cantitative care permit reproductibilitatea ridicată a analizelor, analiza de conţinut poate avea un grad ridicat de subiectivitate deoarece cercetătorul este singurul care decide codificarea şi clasificarea informaţiei relevante pe baza acestor coduri (Burnham et al. 2008, 259). Opinii diferite ce conduc la diferenţe chiar minore ale clasificării acestor coduri pot conduce în final la concluzii diferite pentru analiza aceloraşi texte. În plus, modul de selecţie a surselor este influenţat nu doar de preferinţa cercetătorului, ci şi de gradul de accesibilitate a textelor: unele dintre potenţialele texte, mai ales cele din arhive, pot fi inaccesibile cercetătorilor.

Spre deosebire de analiza de conținut, care se pretează la orice tip de text, analiza de discurs foloseşte texte specifice. Una dintre caracteristicile care diferenţiază analiza de discurs de analiza de conţinut este aceea că textele analizate sunt în mod specific comunicări publice ale unor persoane cu rang înalt. De regulă, acestea reprezintă exprimări publice ale opiniilor şi atitudinilor unor persoane care sunt deţinătoarele unor poziţii politice importante, în guvern, parlament, preşedinţie. În analiza de discurs pot fi identificare probleme importante în procesul de luare a deciziilor, idei dominante, contextul în care are loc discursul, reacţia pe care o provoacă, în aceeaşi măsură în care pot fi prezentate caracteristicile principale ale discursului sau neconcordanţele logice. Discursurile sunt importante pentru transmiterea unor mesaje, de exemplu, pentru fidelizarea susţinătorilor, ameninţarea contestatarilor sau semnalizarea puterii de care dispune emiţătorul, dar şi pentru legitimarea unor acţiuni. Analiza de discurs este cel mai des folosită pentru analizarea discursurilor politice, dar uneori este folosită şi pentru analiza discursurilor manageriale (Burnham et al. 2008, 252–53). Analiza de discurs este, deci, o tehnică de cercetare utilizată cu precădere în istorie, ştiinţe politice şi ştiinţele comunicării.

Dacă textul poate fi analizat folosind tehnici cantitative, discursurile sunt rareori analizate în acest mod deoarece ele sunt folosite mai degrabă pentru interpretarea mesajului discursului într-o anumită cheie pe care fie auditoriul o cunoaşte, fie emiţătorul o transmite împreună cu discursul / mesajul său. Există numeroase exemple de discursuri care au avut un impact foarte important asupra desfăşurării unor evenimente. De exemplu, istoricii şi specialiştii în relaţii internaţionale au folosit analiza discursurilor pentru a explica evenimente din

Războiul Rece, cum ar fi criza rachetelor din Cuba (Allison 1971) în care discursurile au fost folosite ca instrumente de analiză pentru a înţelege modul în care au gândit decidenţii politici măsurile pe care le-au luat în timpul crizei.

Discursurile preşedinţilor SUA cum ar fi cele de la ceremonia de inaugurare a mandatului[61], sau discursurile din Congres, privind Starea Uniunii[62], sau mai recentele discursuri ale preşedinţilor Comisiei Europene[63] sunt exemple de colecţii de discursuri ale unor persoane oficiale cu funcţii înalte, pe care cercetătorii le folosesc pentru a înţelege mai bine deciziile pe care acestea le iau în momente importante, dar şi modul în care aceste discursuri pot anticipa cursul unor politici publice. De exemplu, discursul pe atunci senatorului SUA, John F. Kennedy, din 1957, în care susţinea o viziune programatică privind politica externă a SUA, în contextul crizei din Algeria în care era implicată direct Franţa, încheiată în 1962 prin obţinerea independenţei Algeriei. Discursul, apreciat ca fiind unul dintre cele mai importante discursuri anticoloniale din anii 50, a fost interpretat nu doar ca un sprijin pe care Kennedy îl dădea algerienilor, dar şi ca un semnal că SUA şi puterile coloniale trebuie să îşi revizuiască politicile din Africa şi Asia. Importanţa discursului a fost interpretată (Cleva 2022) ca fiind un precursor al politicii externe pe care Kennedy o va duce de la Casa Albă după câştigarea alegerilor din 1960.

Discursurile sunt importante atât pentru cuvintele pe care le conţin şi care pot să dea numele acelui discurs, cât şi pentru contextul şi locul în care au fost prezentate în public. De exemplu, discursul lui John F Kennedy în Berlinul de Vest în iulie 1963,

---

[61] Arhiva este disponibilă aici: https://www.bartleby.com/124/index.html Accesat ultima dată la 4 Noiembrie 2022.

[62] Arhivate în Proiectul Președinției Americane aici: https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/annual-messages-congress-the-state-the-union Accesat ultima dată la 4 Noiembrie 2022.

[63] Discursurile despre Starea Uniunii sunt disponibile aici: https://ec.europa.eu/info/strategy/strategic-planning/state-union-addresses_en, alte discursuri ale înalților oficiali ai EU sunt disponibile aici: https://ec.europa.eu/commission/presscorner/home/en?keywords=&dotyp=4#news-block Accesat ultima dată la 4 Noiembrie 2022.

cunoscut sub numele de discursul *Ich bin ein Berliner*, a fost analizat nu doar pentru mesajul transmis[64], ci şi pentru locul şi contextul politic în care a fost prezentat.

Nu toate discursurile care merită atenţia cercetătorilor sunt publice. Unele, extrem de sensibile la momentul la care sunt pronunţate, precum discursurile secrete ale lui Mao Zedong (MacFarquhar, Cheek, și Wu 1989) în şedinţe restrânse ale PCC cu prilejul unor campanii precum Marele Salt Înainte sau Revoluţia Culturală, sau celebrul Discurs Secret rostit de Nikita Hruşciov la congresul al XIII-lea la PCUS în februarie 1956, care a avut implicaţii importante pentru destalinizarea URSS, dar şi pentru schimbările subsecvente din celelalte ţări din blocul comunist (Taubman 2004), devin publice și pot fi analizate doar după mult timp. Cu toate acestea, utilitatea lor nu este mai scăzută, ci permite o mai bună înțelegere a unor evenimente istorice importante.

7.4. Introducere în NVivo pentru analiza calitativă

În continuare vom exemplifica practic modalitatea de codificare a discuțiilor de grup. Plecând de la o serie de focus grupuri[65], pentru organizarea unor fiind folosit ghidul din ANEXĂ, vom utiliza programul NVivo pentru a realiza analiza de conținut și a identifica temele recurente din cadrul discuțiilor de grup oferind o introducere aplicată în utilizarea programului NVivo (Bazeley și Jackson 2013). NVivo oferă și opțiunea de transcriere a textelor, prin încărcarea documentelor în format audio în NVivo. Acestea sunt transcrise automat în text, permițându-se apoi verificarea manuală a corectitudinii transcrierii.

[64] Discursul fost interpretat ca fiind, pe de o parte, un mesaj de susținere transmis aliaților din blocul occidental și germanilor din zonele Berlinului administrate de SUA, Marea Britanie și Franța și, pe de altă parte, un mesaj de avertisment la adresa URSS în contextul crizei Berlinului inițiată în 1949.

[65] Focus grupurile au fost organizate în cadrul proiectului „Overcoming disparities in access to quality basic palliative care in the community. Partnerships to identify and improve clinical, educational, legal, and economical barriers" finanțat de Programul de Cooperare Elvețiano-Român, implementat de Fundația Hospice Casa Speranței, Fundația MRC-Median Research Centre, și Spitalul Cantonal St.Gallen. Focus grupurile au fost realizate de Aurelian Muntean în decembrie 2013 și ianuarie 2014.

Primul pas în realizarea analizei de conținut în NVivo este introducerea textelor de analizat în program (Figura 7.1). Introducerea textelor în NVivo se realizează prin parcurgerea unor pași similari altor programe de citire a datelor:

```
Import -> Files-> selectăm documentul pe care dorim să îl încărcăm pentru
analiză
```

Programul NVivo nu este limitat la analiza documentelor text, ci oferă și posibilitatea analizării documentelor audio, video sau a imaginilor. Odată citite de programul NVivo, fișierele pot fi analizate în scopul identificării temelor recurente din cadrul discuțiilor de grup. Pentru a identifica temele recurente din cadrul discuțiilor de grup încărcate în NVivo trebuie să parcurgem o serie de etape. Alternativ, putem apela la opțiunea de analiză automată / codificare automată, care poate fi utilizată ca un prim pas, urmând ca apoi analiza să fie dublată de cea manuală, realizată de cercetător.

1. Deschiderea documentului dorit (Figura 7.2):

```
     Data -> Files-> dublu click și selectăm documentul pe care dorim
să-l încărcăm pentru analiză
```

2. Identificarea temelor recurente și crearea codurilor acestora (Figura 7.3):

```
     Data -> Coding-> Codes -> din documentul deschis tragem un
cuvânt/serie de cuvinte care descriu tema identificată către chenarul
din dreapta ecranului și o nouă fereastră va fi deschisă
```

Figura 7.1 Introducerea fișierelor de analizat în NVivo

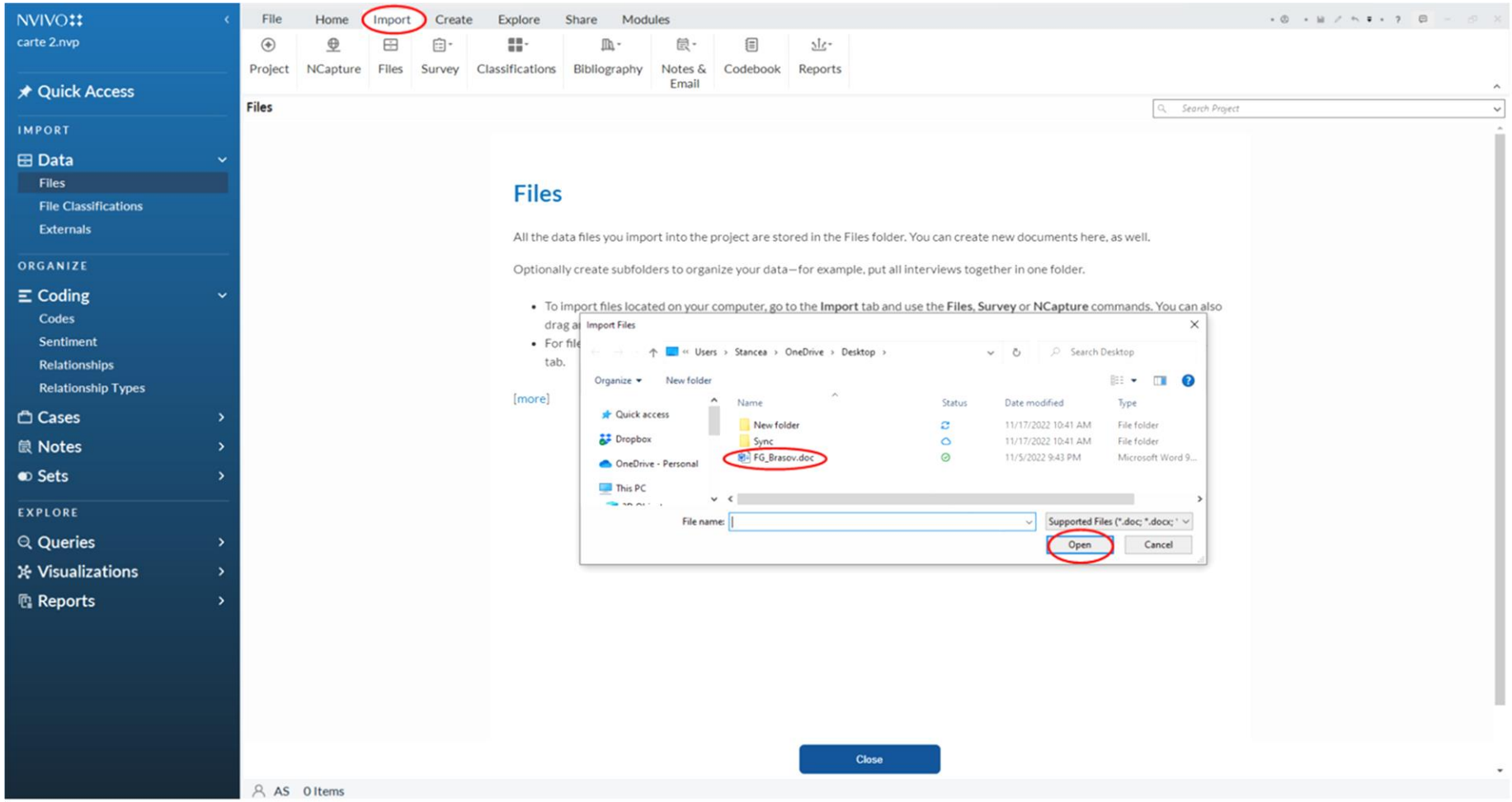


Figura 7.2 Deschiderea documentului în NVivo

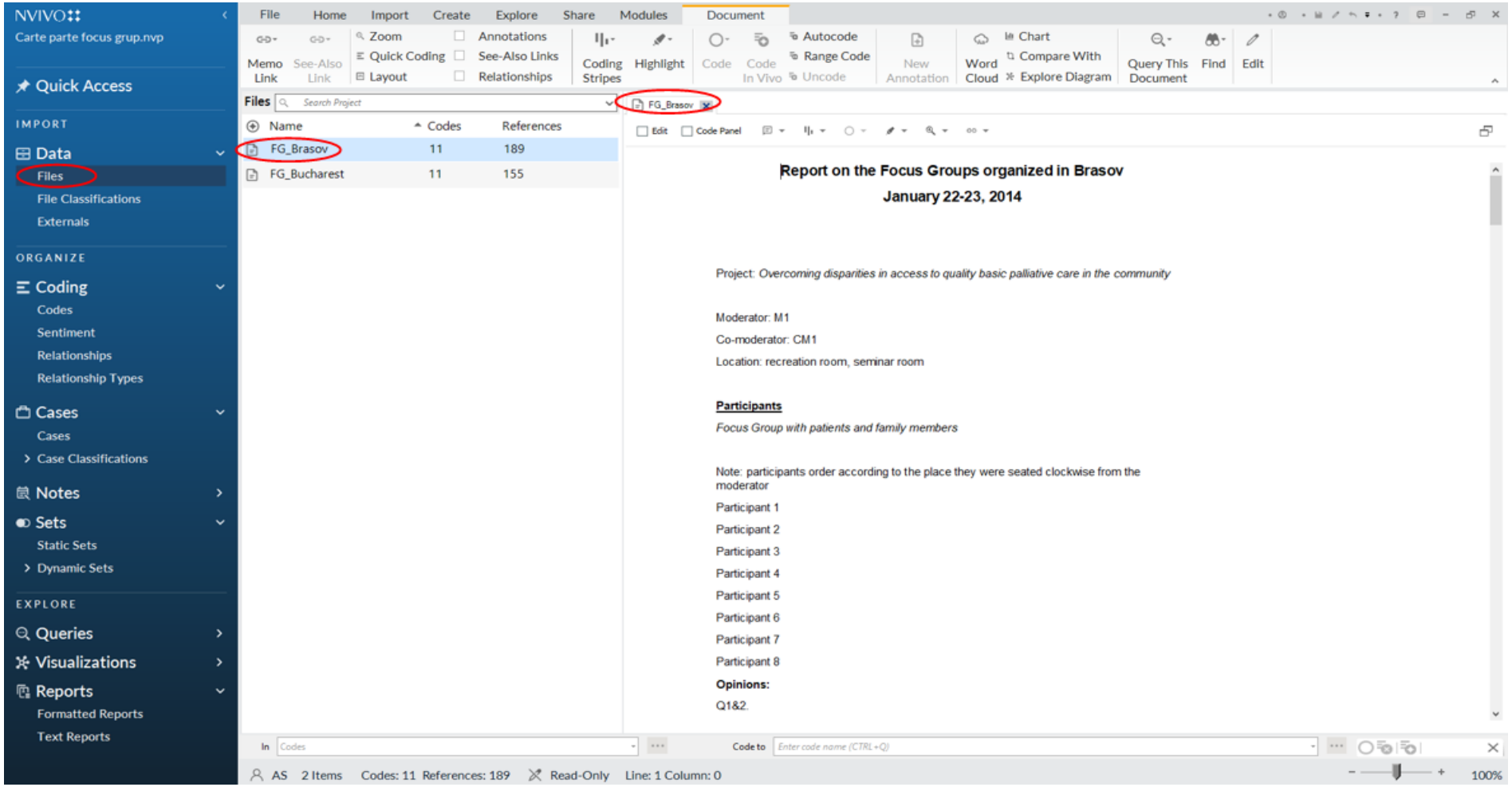
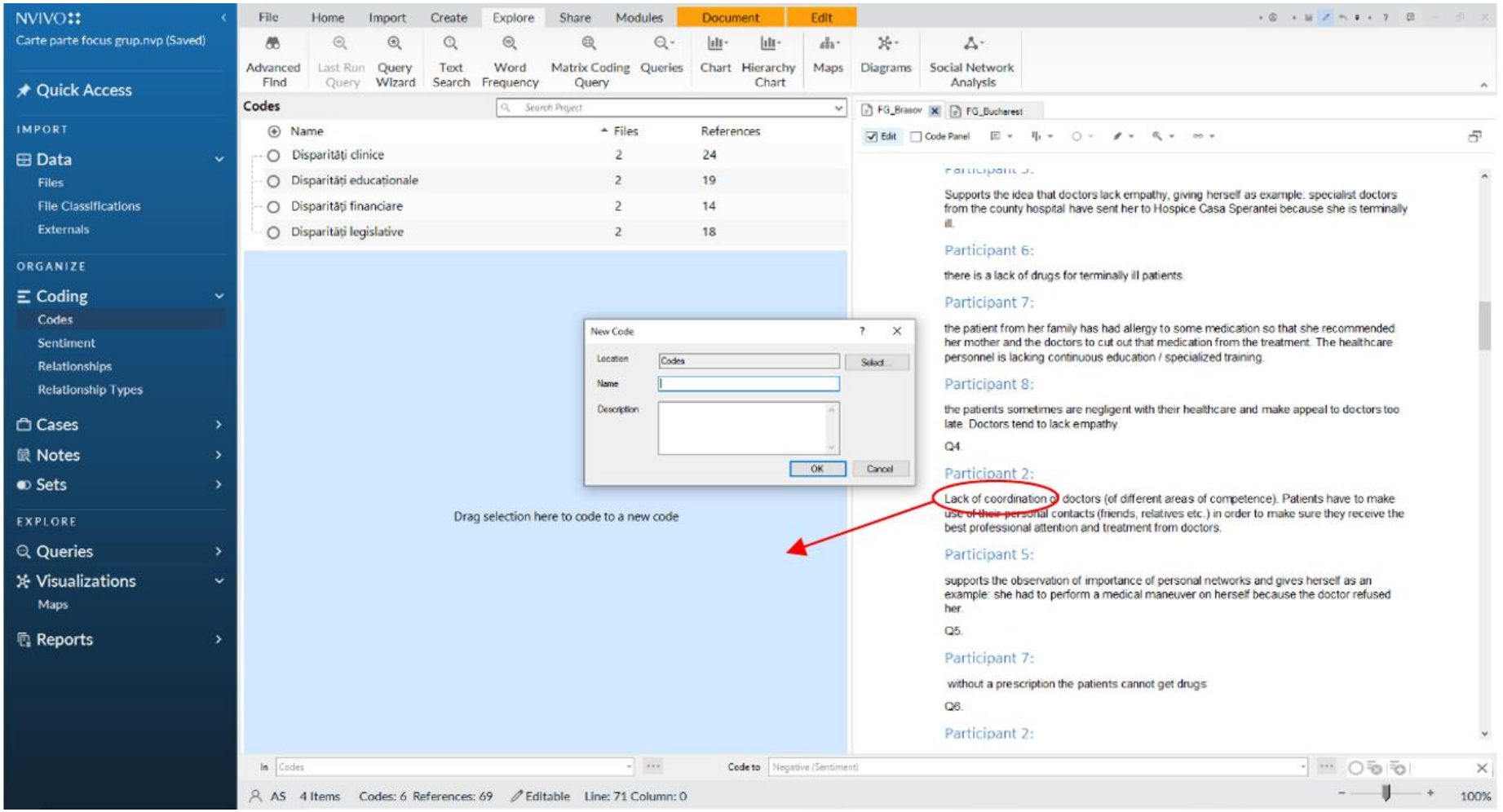


Figura 7.3 Crearea codurilor pentru temele recurente



3. Salvăm tema identificată sub formă de cod:

```
În casetă deschisă la pasul anterior creăm un nume pentru tema
identificată. În cazul discuțiilor de grup oferite ca exemplu în acest
manual temele recurente identificate au fost Disparități clinice,
educaționale, financiare și legislative. După ce am creat un nume pentru
tema identificată selectăm OK și tema va fi salvată în chenarul Codes.
Parcurgem pasul 2 și 3 până când au fost identificate toate temele
recurente din cadrul documentelor supuse analizei în NVivo.
```

Ulterior codării temelor principale din cadrul discuțiilor de grup pot fi accesate o serie de grafice (de exemplu, nor de cuvinte, în engleză *WordCloud*, ilustrat în Figura 7.4) care să permită o înțelegere aprofundată a datelor. De asemenea, pot fi filtrate, în funcție de tema dorită, frazele specifice temei identificate de participanții la discuția de grup. Această filtrare pe teme/coduri permite o vizualizare comparativă a aceleiași teme din perspectiva mai multor subiecți intervievați. Analiza poate fi realizată și pe un subgrup de subiecți, pentru a explora și a scoate în evidență asemănări și diferențe de perspectivă asupra temei, care altfel ar fi dificil de identificat fără utilizarea acestui program. Pentru proiectele de cercetare care includ secțiuni specifice de analiză a teoriei de specialitate (*literature review*), NVivo permite colectarea și extragerea informației calitative pe baza acestor coduri / teme specificate de cercetător. Acest lucru este posibil dacă în prealabil cercetătorul a codificat și aceste texte incluse în analiza literaturii de specialitate. Astfel informația extrasă pe baza codurilor / temelor poate să cuprindă componenta de analiză de literatură de specialitate și de date calitative (de exemplu, interviurile analizate). Aceasta posibilitate tehnică crește eficiența procesului comparativ de analiză calitativă.

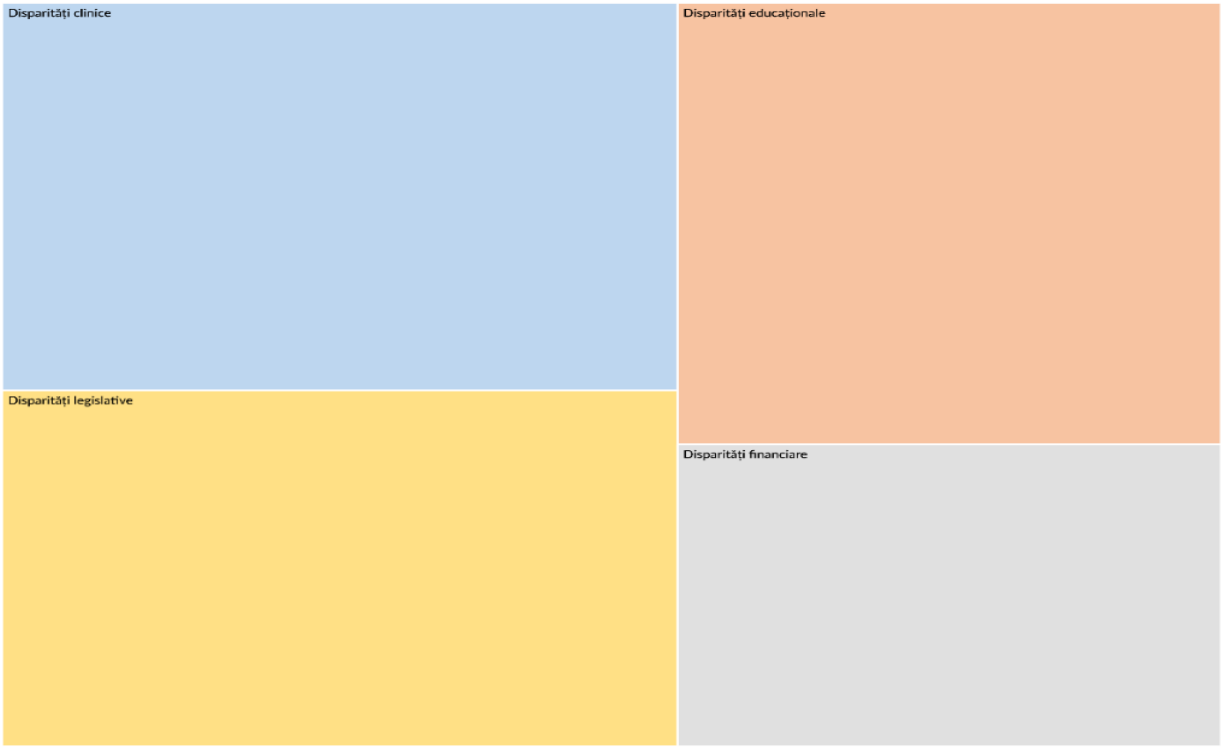
În analiza focus grupurilor utilizate în această exemplificare aplicată pe date reale, am inclus principalele coduri ale disparităților (clinice, educaționale, financiare, legislative). Pe lângă acestea, în analiză adăugăm în codificare problemele specifice incluse în exercițiul integrat în ghidul de focus grup (a se vedea ANEXA), dar și problemele identificate ad-hoc de participanții la discuțiile de grup: comunicare deficitară; tratament medicamentos neadecvat; lipsa asistenței spirituale; lipsa de

empatie a personalului medical; necoordonarea îngrijirii medicale; lipsa banilor; lipsa medicamentelor; lipsa subvențiilor de stat; pregătirea profesională insuficientă a personalului medical; simptome insuficient tratate; lipsa protocoalelor medicale; necunoașterea legislației; salarii nemotivante pentru personalul medical; lipsa serviciilor de îngrijire la domiciliu; lipsa suport familie / apropiați; lipsa secțiilor de îngrijire paliativă în spitale; bariere în accesul la medicamente; lipsa informațiilor privind dieta specială pentru persoane bolnave cronic; lipsa activităților fizice, culturale și sociale pentru persoanele bolnave; lipsa personalului de suport; lipsa materialelor auxiliare pentru tratament; prea multă birocrație.

Figura 7.4 WordCloud cu cele mai întâlnite expresii

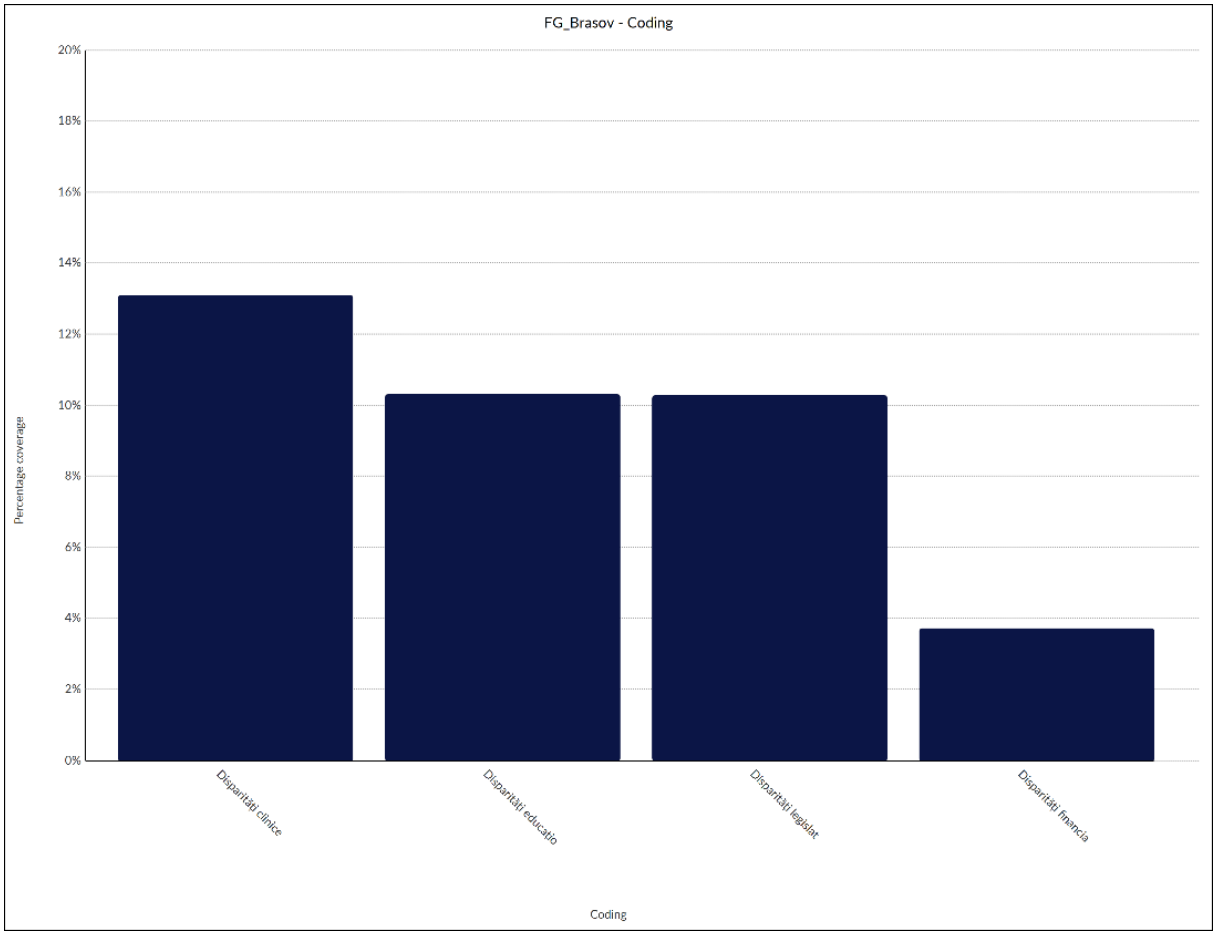
Temele recurente abordate de respondenții la discuțiile de grup pot fi reprezentate vizual prin realizarea unui grafic de tip ierarhic (în engleză *hierarchy chart*), precum cel din Figura 7.5. Acesta permite ilustrarea grafică a diferențelor de prevalență a unor probleme (ulterior codificate) în discuțiile de grup.

Figura 7.5 Hierarchy chart cu temele recurente din cadrul discuțiilor de grup



Aceste diferențe pot fi ilustrate grafic folosind un grafic cu bare (în engleză *bar chart)* precum cel din Figura 7.6, care permite la rândul său compararea frecvențelor temelor recurente dintre în interiorul grupului, dar și între două sau mai multe grupuri.

Figura 7.6 Frecvența temelor recurente într-o discuție de grup



Bibliografie

Adăscăliței, Dragoș, și Aurelian Muntean. 2019. „Trade Union Strategies in the Age of Austerity: The Romanian Public Sector in Comparative Perspective". *European Journal of Industrial Relations* 25 (2): 113–28. https://doi.org/10.1177/0959680118783588.

Agresti, Alan, și Barbara Finlay. 2014. *Statistical Methods for the Social Sciences*. Harlow, Essex: Pearson.

Akaike, Hirotugu. 1974. „A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Allison, Graham T. 1971. *Essence of decision. Explaining the Cuban missile crisis*. Boston: Little, Brown.

Aquino, Jakson, Dirk Enzmann, Marc Schwartz, Nitin Jain, și Stefan Kraft. 2021. „descr: Descriptive Statistics". https://CRAN.R-project.org/package=descr.

Armingeon, Klaus, și Besir Ceka. 2014. „The loss of trust in the European Union during the great recession since 2007: The role of heuristics from the national political system". *European Union Politics* 15 (1): 82–107. https://doi.org/10.1177/1465116513495595.

Armstrong, Dave. 2022. „DAMisc: Dave Armstrong's Miscellaneous Functions". https://CRAN.R-project.org/package=DAMisc.

Arnold, Jeffrey B. 2021. „ggthemes: Extra Themes, Scales and Geoms for «ggplot2»". https://CRAN.R-project.org/package=ggthemes.

Auspurg, Katrin, și Thomas Hintz. 2015. *Factorial Survey Experiments*. http://srmo.sagepub.com/view/factorial-survey-experiments/SAGE.xml.

Babbie, Earl. 2010. *Practica cercetării sociale*. Iași: Polirom.

Baumrind, Diana. 1964. „Some Thoughts on Ethics of Research: After Reading Milgram's «Behavioral Study of Obedience.»" *American Psychologist* 19 (6): 421–23. https://doi.org/10.1037/h0040128.

Bazeley, Patricia, și Kristi Jackson. 2013. *Qualitative Data Analysis with NVivo*. Second edition. Los Angeles [i.e. Thousand Oaks, Calif.] ; London: SAGE Publications.

Blair, Graeme, Alexander Coppock, și Margaret Moor. 2020. „When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments". *American Political Science Review*, august, 1–19. https://doi.org/10.1017/S0003055420000374.

Blair, Graeme, și Kosuke Imai. 2010. „list: Statistical Methods for the Item Count Technique and List Experiment". The Comprehensive R Archive Network (CRAN). http:// CRAN.R-project.org/ package=list.

———. 2012. „Statistical Analysis of List Experiments". *Political Analysis* 20 (1): 47–77.

Blalock, Hubert M. 2018. *Causal Inferences in Nonexperimental Research*. Original edition printed in 1964. Chapel Hill: University of North Carolina Press.

Brady, Henry E., și David Collier. 2003. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Brandt, Allan M. 1978. „Racism and Research: The Case of the Tuskegee Syphilis Study". *The Hastings Center Report* 8 (6): 21. https://doi.org/10.2307/3561468.

Brotherton, David, și Luis Barrios. 2004. *The Almighty Latin King and Queen Nation: street politics and the transformation of a New York City gang*. New York: Columbia University Press.

Bulai, Alfred. 2000. *Focus-grup*. București: Paideia.

Bullock, Will, Kosuke Imai, și Jacob N. Shapiro. 2011. „Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan". *Political Analysis* 19 (4): 363–84. https://doi.org/10.1093/pan/mpr031.

Burnham, Peter, Karin Lutz, Wyn Grant, și Zig Layton-Henry. 2008. *Research methods in politics*. 2nd ed. Political analysis. Basingstoke: Palgrave Macmillan.

Cairo, Alberto. 2013. *The functional art: an introduction to information graphics and visualization*. Berkeley, California: New Riders.

Carney, Dana R., Amy J.C. Cuddy, și Andy J. Yap. 2010. „Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance". *Psychological Science* 21 (10): 1363–68. https://doi.org/10.1177/0956797610383437.

Charness, Gary, Uri Gneezy, și Michael A. Kuhn. 2012. „Experimental Methods: Between-Subject and within-Subject Design". *Journal of Economic Behavior & Organization* 81 (1): 1–8. https://doi.org/10.1016/j.jebo.2011.08.009.

Chelcea, Liviu. 2008. *Bucureştiul postindustrial: Memorie, dezindustrializare şi regenerare urbană*. Iași: Polirom.

Chelcea, Liviu, și Ioana Iancu. 2015. „An Anthropology of Parking: Infrastructures of Automobility, Work, and Circulation". *Anthropology of Work Review* 36 (2): 62–73. https://doi.org/10.1111/awr.12068.

Chelcea, Septimiu. 2001. *Metodologia cercetării sociologice*. București: Editura Economică.

Cleva, Gregory D. 2022. *John F. Kennedy's 1957 Algeria speech: the politics of anticolonialism in the Cold War era*. Lanham: Lexington Books.

Collier, David, și Steven Levitsky. 1997. „Democracy with Adjectives: Conceptual Innovation in Comparative Research". *World Politics* 49 (3): 430–51. https://doi.org/10.1353/wp.1997.0009.

Collier, David, și James E. Mahon. 1993. „Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis". *American Political Science Review* 87 (4): 845–55. https://doi.org/10.2307/2938818.

Comșa, Mircea. 2022. *Data mining pentru Științele Sociale Vol. 1: Pregătirea datelor în RapidMiner Studio*. Cluj-Napoca: Presa Universitară Clujeană.

Corstange, Daniel. 2008. „Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT". *Political Analysis* 17 (1): 45–63.

Davidson, Michael. 2017. „Vaccination as a Cause of Autism—Myths and Controversies". *Dialogues in Clinical Neuroscience* 19 (4): 403–7. https://doi.org/10.31887/DCNS.2017.19.4/mdavidson.

Diochetanu, Andreea. 2022. „Cât a costat Recensământul Populației din 2022 și câți bani au fost cheltuiți pentru promovare". *Main News* (blog). 17 august 2022. https://mainnews.ro/exclusiv-cat-a-costat-recensamantul-populatiei-din-2022-si-cati-bani-au-fost-cheltuiti-pentru-promovare/.

Drisko, James W., și Tina Maschi. 2016. *Content analysis*. Pocket guides to social work research methods. New York: Oxford University Press.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, și Trena M. Ezzati. 2004. „The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study

Application". În *Measurement Errors in Surveys*, 185–210. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118150382.ch11.

Dunning, Thad. 2012. *Natural experiments in the social sciences: a design-based approach*. Strategies for social inquiry. Cambridge ; New York: Cambridge University Press.

Durán, Robert J. 2018. *The gang paradox: inequalities and miracles on the U.S.-Mexico border*. New York: Columbia University Press.

Dușa, Adrian. 2014. *Elemente de analiză comparativă*. București: Tritonic.

———. 2022a. „DDIwR: DDI with R". https://CRAN.R-project.org/package=DDIwR.

———. 2022b. „declared: Functions for Declared Missing Values". https://CRAN.R-project.org/package=declared.

Dușa, Adrian, Bogdan Oancea, Nicoleta Caragea, Ciprian Alexandru, Nicolae Marius Jula, și Ana Maria Dobre. 2015. *R cu aplicații în statistică.* Editura Universității din București.

Elster, Jon. 2013. *Comportamentul social. Fundamentele explicației în științele sociale*. București: Editura All.

Farrington, C. Paddy, Elizabeth Miller, și Brent Taylor. 2001. „MMR and Autism: Further Evidence against a Causal Association". *Vaccine* 19 (27): 3632–35. https://doi.org/10.1016/S0264-410X(01)00097-4.

Fearon, James D., și David D. Laitin. 2009. „Integrating Qualitative and Quantitative Methods". În *The Oxford Handbook of Political Methodology*, ediție de Janet M. Box-Steffensmeier, Henry E. Brady, și David Collier, 1 ed, 756–76. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0033.

Fisher, Ronald Aylmer. 1971. *The Design of Experiments*. Ninth edition. First published 1935. New York: Hafner Press.

Fowler, Floyd J. 2014. *Survey research methods*. Fifth edition. Applied social research methods series. Los Angeles: SAGE.

Fox, James Alan, și Paul E. Tracy. 1986. *Randomized response: a method for sensitive surveys*. Sage Quantitative Applications in the Social Sciences, no. 07-058. Beverly Hills: Sage Publications.

Fox, John, și Sanford Weisberg. 2019. *An {R} Companion to Applied Regression*. Third. Thousand Oaks (CA): Sage.

Frankfort-Nachmias, Chava, David Nachmias, și Jack DeWaard. 2015. *Research methods in the social sciences*. Eighth edition. New York, NY: Worth Publishers, a Macmillan Education Company.

Freedman, David, Robert Pisani, și Roger Purves. 2007. *Statistics*. 4th ed. New York: W.W. Norton & Co.

Garbuszus, Jan Marvin, și Sebastian Jeworutzki. 2022. „readstata13: Import «Stata» Data Files". https://CRAN.R-project.org/package=readstata13.

Gee, James Paul. 1999. *An introduction to discourse analysis: theory and method*. London ; New York: Routledge.

George, Alexander L., și Andrew Bennett. 2005. *Case studies and theory development in the social sciences*. BCSIA studies in international security. Cambridge, Mass: MIT Press.

Gerber, Alan S., și Donald P. Green. 2012. *Field experiments: design, analysis, and interpretation*. 1st ed. New York: W. W. Norton.

Gerring, John. 2004. „What Is a Case Study and What Is It Good for?" *The American Political Science Review* 98 (2): 341–54.

———. 2006. *Case Study Research. Principles and Practices*. Cambridge: Cambridge University Press.

Glynn, Adam N. 2013. „What Can We Learn with Statistical Truth Serum?" *Public Opinion Quarterly* 77 (S1): 159–72. https://doi.org/10.1093/poq/nfs070.

Glynn, Adam N., și Jon Wakefield. 2010. „Ecological Inference in the Social Sciences". *Statistical Methodology* 7 (3): 307–22. https://doi.org/10.1016/j.stamet.2009.09.003.

Goertz, Gary, și James Mahoney. 2012. *A tale of two cultures: qualitative and quantitative research in the social sciences*. Princeton, N.J: Princeton University Press.

Goffman, Alice. 2014. *On the run: fugitive life in an American city*. Fieldwork encounters and discoveries. Chicago ; London: The University of Chicago Press.

Grönlund, Kimmo, și Maija Setälä. 2007. „Political Trust, Satisfaction and Voter Turnout". *Comparative European Politics* 5 (decembrie). https://doi.org/10.1057/palgrave.cep.6110113.

Gross, Juergen, și Uwe Ligges. 2015. „nortest: Tests for Normality". https://CRAN.R-project.org/package=nortest.

Hainmueller, Jens, Daniel J. Hopkins, și Teppei Yamamoto. 2014. „Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". *Political Analysis* 22 (1): 1–30. https://doi.org/10.1093/pan/mpt024.

Harris, Ben. 1979. „Whatever Happened to Little Albert?" *American Psychologist* 34 (2): 151–60. https://doi.org/10.1037/0003-066X.34.2.151.

Harris, Robert L. 1996. *Information graphics: a comprehensive illustrated reference*. Atlanta: Management Graphics.

Hebbali, Aravind. 2020. „olsrr: Tools for Building OLS Regression Models". https://CRAN.R-project.org/package=olsrr.

Hlavac, Marek. 2022. „stargazer: Well-Formatted Regression and Summary Statistics Tables". Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.

Holbrook, Allyson L., și Jon A. Krosnick. 2010. „Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique". *The Public Opinion Quarterly* 74 (1): 37–67.

Imai, Kosuke. 2011. „Multivariate Regression Analysis for the Item Count Technique". *Journal of the American Statistical Association* 106 (494): 407–16. https://doi.org/10.1198/jasa.2011.ap10415.

International Military Tribunal. 1949. *Trials of war criminals before the Nuernberg military tribunals under Control Council law no. 10*. Vol. 2-The Medical Case. Washington D.C.: U.S. Government Printing Office. https://collections.nlm.nih.gov/ext/dw/01130400RX2/PDF/01130400RX2.pdf.

Ivanov, Catiușa. 2020. „Poluare în București și Ilfov sâmbătă noaptea: Senzorii independenți au înregistrat valori spectaculoase pentru poluarea cu praf - peste 1.700% pentru PM2,5 și aproape 1.400% pentru PM10 / Autoritățile de mediu spun că au verificat, dar nu se confirmă". 8 noiembrie 2020. https://www.hotnews.ro/stiri-administratie_locala-24405165-poluare-bucuresti-ilfov-sambata-noaptea-senzorii-independenti-inregistrat-spectaculoase-pentru-poluarea-praf-peste-1700-pentru-pm2-5-aproape-1400-pentru-pm10-autoritatile-mediu-spun-verificat-dar-nu-c.htm.

———. 2021. „Poluarea de weekend din București: Agenția de Mediu dă vina pe trafic și condițiile atmosferice, activiștii de mediu indică arderi de deșeuri și încălzirea rezidențială / În rest, tăcere”. 26 ianuarie 2021. https://www.hotnews.ro/stiri-administratie_locala-24562550-poluarea-weekend-din-bucuresti-agentia-mediu-vina-trafic-conditiile-atmosferice-activistii-mediu-indica-arderi-deseuri-incalzirea-rezidentiala-rest-tacere.htm.

———. 2022. „Poluare extremă sâmbătă seară în București: Valori foarte mari pentru poluarea cu particule PM 2,5 și PM 10 în special la perifieria Capitalei”. 12 noiembrie 2022. https://www.hotnews.ro/stiri-administratie_locala-25899388-poluare-extrema-sambata-seara-bucuresti-valori-25-ori-mai-mari-pentru-2-5.htm.

Jackman, Simon. 2020. „{pscl}: Classes and Methods for {R} Developed in the Political Science Computational Laboratory”. R package version 1.5.5. Sydney, New South Wales, Australia: United States Studies Centre, University of Sydney. https://github.com/atahk/pscl/},.

Jann, Ben, Julia Jerke, și Ivar Krumpal. 2012. „Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism”. *Public Opinion Quarterly* 76 (1): 32–49. https://doi.org/10.1093/poq/nfr036.

Jderu, Gabriel. 2015. *Cultura motocicletelor*. București: Tritonic.

Jenner, Bryan, și Stefan Titscher, ed. 2000. *Methods of Text and Discourse Analysis*. London: SAGE.

Jørgensen, Marianne, și Louise Phillips. 2002. *Discourse Analysis as Theory and Method*. London: Sage Publications.

Jussim, Lee J., Jon A. Krosnick, și Sean T. Stevens, ed. 2022. *Research integrity: best practices for the social and behavioral sciences*. New York, NY: Oxford University Press.

King, Gary, Robert O. Keohane, și Sidney Verba. 2000. *Fundametele cercetării sociale*. Iași: Polirom.

King, Ronald F. 2005. *Strategia cercetării: treisprezece cursuri despre elementele științelor sociale*. Iași: Polirom.

Kish, Leslie. 1995. *Survey Sampling*. A Wiley Interscience Publication. New York: Wiley.

Koleva, Spassena P., și Blanka Rip. 2009. „Attachment Style and Political Ideology: A Review of Contradictory Findings". *Social Justice Research* 22 (2): 241–58. https://doi.org/10.1007/s11211-009-0099-y.

Krauss, Svenja. 2018. „Stability through Control? The Influence of Coalition Agreements on the Stability of Coalition Cabinets". *West European Politics* 41 (6): 1282–1304. https://doi.org/10.1080/01402382.2018.1453596.

Krippendorff, Klaus. 2018. *Content analysis: an introduction to its methodology*. Fourth Edition. Los Angeles: SAGE.

Krueger, Richard A., și Mary Anne Casey. 2005. *Metoda focus grup. Ghid practic pentru cercetarea aplicată*. Iași: Polirom.

Kuhn, Max. 2022. „caret: Classification and Regression Training". https://CRAN.R-project.org/package=caret.

Kuhn, Thomas S. 2008. *Structura revoluțiilor științifice*. București: Humanitas.

Kuklinski, James H., Michael D. Cobb, și Martin Gilens. 1997. „Racial Attitudes and the «New South»". *The Journal of Politics* 59 (2): 323–49. https://doi.org/10.2307/2998167.

Lazarsfeld, Paul F., Bernard Berelson, și Hazel Gaudet. 2021. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign, Legacy Edition*. Columbia University Press.

Leavy, Patricia, ed. 2014. *The Oxford handbook of qualitative research*. Oxford library of psychology. Oxford ; New York: Oxford University Press.

Lee, Myoung-jae. 2016. *Matching, regression discontinuity, difference in differences, and beyond*. New York: Oxford University Press.

Leeper, Thomas J., Jeffrey B. Arnold, Vincent Arel-Bundock, și Jacob A. Long. 2021. „margins: Marginal Effects for Model Objects". https://CRAN.R-project.org/package=margins.

Lijphart, Arend. 1971. „Comparative Politics and the Comparative Method". *The American Political Science Review* 65 (3): 682–93. https://doi.org/10.2307/1955513.

Little, Roderick J. A., și Donald B. Rubin. 2020. *Statistical analysis with missing data*. Third edition. Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Lohr, Sharon L. 2010. *Sampling: design and analysis*. 2nd ed. Boston, Mass: Brooks/Cole.

Long, J. Scott, și Jeremy Freese. 2000. „FITSTAT: Stata module to compute fit statistics for single equation regression models". Statistical Software Components S407201: Boston College Department of Economics.

Lubet, Steven. 2015. „Ethics On The Run - New Rambler Review". New Rambler Review. 2015. https://newramblerreview.com/book-reviews/law/ethics-on-the-run.

Lüdecke, Daniel. 2022. „sjPlot: Data Visualization for Statistics in Social Science". https://CRAN.R-project.org/package=sjPlot.

MacFarquhar, Roderick, Timothy Cheek, și Eugene Wu. 1989. *The Secret Speeches of Chairman Mao: From the Hundred Flowers to the Great Leap Forward*. Harvard Contemporary China Series 6. Cambridge, Mass: Council on East Asian Studies/Harvard University : Distributed by Harvard University Press.

Madsen, Kreesten Meldgaard, Anders Hviid, Mogens Vestergaard, Diana Schendel, Jan Wohlfahrt, Poul Thorsen, Jørn Olsen, și Mads Melbye. 2002. „A Population-Based Study of Measles, Mumps, and Rubella Vaccination and Autism". *New England Journal of Medicine* 347 (19): 1477–82. https://doi.org/10.1056/NEJMoa021134.

Mares, Isabela, și Aurelian Muntean. 2015. „Mayors, ethnic intermediaries and party brokers: explaining variation in clientelistic strategies in rural settings". În *European Political Science Association*. Vienna: European Political Science Association. https://www.academia.edu/download/38062285/Mares_and_Muntean2015.pdf.

Mares, Isabela, Aurelian Muntean, și Tsveta Petrova. 2017. „Pressure, Favours, and Vote-Buying: Experimental Evidence from Romania and Bulgaria". *Europe-Asia Studies* 69 (6): 940–60. https://doi.org/10.1080/09668136.2017.1364351.

———. 2018. „Economic Intimidation in Contemporary Elections: Evidence from Romania and Bulgaria". *Government and Opposition* 53 (03): 486–517. https://doi.org/10.1017/gov.2016.39.

Mares, Isabela, Aurelian Muntean, și Lauren Young. 2016. „Bought or coerced? The nonprogrammatic mobilization of Roma voters in Eastern Europe". În *Council for European Studies*. Philadelphia, PA: Council for European Studies.

https://www.researchgate.net/publication/339618123_Bought_or_coerced_The_nonprogrammatic_mobilization_of_Roma_voters_in_Eastern_Europe.

McCracken, Grant David. 1988. *The long interview*. Qualitative research methods, v. 13. Newbury Park, Calif: Sage Publications.

Microsoft Corporation. 2021. „Microsoft Excel LTSC". Redmond, WA: Microsoft Corporation. https://office.microsoft.com/excel.

Mihăilescu, Vintilă. 2019. *În căutarea corpului regăsit: o ego-analiză a spitalului*. Iași: Polirom.

Milgram, Stanley. 1963. „Behavioral Study of Obedience". *The Journal of Abnormal and Social Psychology* 67 (4): 371–78. https://doi.org/10.1037/h0040525.

———. 1964. „Issues in the Study of Obedience: A Reply to Baumrind". *American Psychologist* 19 (11): 848–52. https://doi.org/10.1037/h0044954.

Mill, John Stuart. 1882. *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methos of scientific investigation*. Eight. New York: Harper & Brothers. https://www.gutenberg.org/files/27942/27942-h/27942-h.html.

Morgan, David L. 1996. „Focus Groups". *Annual Review of Sociology* 22 (1): 129–52. https://doi.org/10.1146/annurev.soc.22.1.129.

Muntean, Aurelian. 2023. „Dependence, Adaptation and Survival: Social Dialogue in Multinational Corporations in Romania". În *Are multinational companies good for trade unions? Evidence from six central and eastern European countries*, ediție de Martin Myant. Brussels: ETUI.

Murray, Scott. 2013. *Interactive data visualization for the web*. 1st ed. Sebastopol, CA: O'Reilly Media.

Neuwirth, Erich. 2022. „RColorBrewer: ColorBrewer Palettes". https://CRAN.R-project.org/package=RColorBrewer.

Nielsen, Richard A. 2016. „Case Selection via Matching". *Sociological Methods & Research* 45 (3): 569–97. https://doi.org/10.1177/0049124114547054.

Pacurari, Nadia, Eva De Clercq, Monica Dragomir, Anca Colita, Tenzin Wangmo, și Bernice S. Elger. 2021. „Challenges of Paediatric Palliative Care in Romania: A Focus Groups Study". *BMC Palliative Care* 20 (1): 178. https://doi.org/10.1186/s12904-021-00871-7.

Popper, Karl R. 1981. *Logica cercetării*. București: Editura Științifică și Enciclopedică.

Prieto Curiel, Rafael, și Humberto González Ramírez. 2021. „Vaccination Strategies against COVID-19 and the Diffusion of Anti-Vaccination Views". *Scientific Reports* 11 (1): 6626. https://doi.org/10.1038/s41598-021-85555-1.

QSR International Pty Ltd. 2020. „NVivo". QSR International Pty Ltd. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.

R Core Team. 2022a. „foreign: Read Data Stored by «Minitab», «S», «SAS», «SPSS», «Stata», «Systat», «Weka», «dBase», ..." https://CRAN.R-project.org/package=foreign.

———. 2022b. „R: A Language and Environment for Statistical Computing". Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Ragin, Charles C. 1987. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ranehill, Eva, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, și Roberto A. Weber. 2015. „Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women". *Psychological Science* 26 (5): 653–56. https://doi.org/10.1177/0956797614553946.

Răducanu, Mara. 2013. „40 de milioane de euro a costat Recensământul Populaţiei şi Locuinţelor din octombrie 2011". 28 iunie 2013. https://adevarul.ro/stiri-interne/societate/40-de-milioane-de-euro-a-costat-recensamantul-1450781.html.

Reisz, Robert D. 2017. *Carte de statistică. Rețete încercate*. Romania: Tritonic.

Ress, Simon, și Florian Spohr. 2022. „Was it worth it? The impact of the German minimum wage on union membership of employees". *Economic and Industrial Democracy* 43 (4): 1699–1723. https://doi.org/10.1177/0143831X211035828.

Revelle, William. 2022. „psych: Procedures for Psychological, Psychometric, and Personality Research". Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Reverby, Susan. 2009. *Examining Tuskegee: the infamous syphilis study and its legacy*. The John Hope Franklin series in African American history and culture. Chapel Hill: University of North Carolina Press.

Rihoux, Benoît. 2003. „Bridging the Gap between the Qualitative and Quantitative Worlds? A Retrospective and Prospective View on Qualitative Comparative Analysis" 15 (4): 351–65. https://doi.org/10.1177/1525822X03257690.

Rosenberg, Joshua M., Marcus Kubsch, Eric-Jan Wagenmakers, și Mine Dogucu. 2022. „Making Sense of Uncertainty in the Science Classroom". *Science & Education*, iunie, 1–24. https://doi.org/10.1007/s11191-022-00341-3.

Rotariu, Traian, Gabriel Bădescu, Irina Culic, Elemer Mezei, și Cornelia Mureșan. 1999. *Metode statistice aplicate în științele sociale*. 1 ed. Iași: Polirom.

Rotariu, Traian, și Petru Iluț. 1997. *Ancheta sociologică și sondajul de opinie. Teorie și practică*. Iași: Polirom.

RStudio Team. 2022. „RStudio: Integrated Development for R". Boston, MA: RStudio, PBC. http://www.rstudio.com.

Sandu, Dumitru. 1992. *Statistică în ştiinţele sociale. Probleme teoretice şi aplicaţii pentru învăţământul universitar*. București: Universitatea București.

———. 1996. *Sociologia tranziției. Valori și tipuri sociale în România*. București: Editura Staff.

———. 1999. *Spațiul social al tranziției*. Iași: Polirom.

Sartori, Giovanni. 1970. „Concept Misformation in Comparative Politics". *The American Political Science Review* 64 (4): 1033–53. https://doi.org/10.2307/1958356.

Schneider, Carsten Q., și Claudius Wagemann. 2012. *Set-theoretic methods for the social sciences: a guide to qualitative comparative analysis*. Strategies for social inquiry. Cambridge: Cambridge University Press.

Schwarz, Gideon. 1978. „Estimating the Dimension of a Model". *The Annals of Statistics* 6 (2). https://doi.org/10.1214/aos/1176344136.

Signorell, Andri. 2022. „{DescTools}: Tools for Descriptive Statistics". https://cran.r-project.org/package=DescTools.

Silverman, David. 2020. *Qualitative Research*. London: SAGE.

Socaciu, Emanuel, Constantin Vică, Emilian Mihailov, Toni Gibea, Valentin Mureșan, și Mihaela Constantinescu. 2018. *Etică și integritate academică*. București: Editura Universității din București.

StataCorp. 2019. „Stata Statistical Software: Release 16". College Station, TX: StataCorp LLC.

Stockemer, Daniel. 2019. *Quantitative Methods for the Social Sciences: A Practical Introduction with Examples in SPSS and Stata*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-99118-4.

Stoica, Cătălin Augustin, și Radu Umbreș. 2021. „Suspicious Minds in Times of Crisis: Determinants of Romanians' Beliefs in COVID-19 Conspiracy Theories". *European Societies* 23 (sup1): S246–61. https://doi.org/10.1080/14616696.2020.1823450.

Stroe, Monica. 2016. *Gustul locului. Productia de peisaje culturale agro-alimentare in sudul Transilvaniei*. București: Tritonic.

Taubman, William. 2004. *Khrushchev: The Man and His Era*. New York: Norton.

Taylor, Brent, Elizabeth Miller, CPaddy Farrington, Maria-Christina Petropoulos, Isabelle Favot-Mayaud, Jun Li, și Pauline A Waight. 1999. „Autism and Measles, Mumps, and Rubella Vaccine: No Epidemiological Evidence for a Causal Association". *The Lancet* 353 (9169): 2026–29. https://doi.org/10.1016/S0140-6736(99)01239-8.

Thrasher, Frederic Milton. 1963. *The Gang: A Study of 1313 gangs of Chicago*. Original Edition 1927. Chicago, IL: The University of Chicago Press.

Trapletti, Adrian, și Kurt Hornik. 2022. „tseries: Time Series Analysis and Computational Finance". https://CRAN.R-project.org/package=tseries.

Trappmann, Mark, Ivar Krumpal, Antje Kirchner, și Ben Jann. 2014. „Item Sum: A New Technique for Asking Quantitative Sensitive Questions". *Journal of Survey Statistics and Methodology* 2 (1): 58–77. https://doi.org/10.1093/jssam/smt019.

Valentino, Nicholas A., Vincent L. Hutchings, Antoine J. Banks, și Anne K. Davis. 2008. „Is a Worried Citizen a Good Citizen? Emotions, Political Information Seeking, and Learning via the Internet". *Political Psychology* 29 (2): 247–73. https://doi.org/10.1111/j.1467-9221.2008.00625.x.

Venables, W.N., și B.D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer.

Vinten-Johansen, Peter, ed. 2003. *Cholera, chloroform, and the science of medicine: a life of John Snow*. Oxford ; New York: Oxford University Press.

Vlăsceanu, Lazăr. 2013. *Introducere în metodologia cercetării sociologice*. Iași: Polirom.

Voicu, Mălina, și Bogdan Voicu. 2002. „Proiectul de cercetare internațională privind studiul valorilor europene". *Calitatea Vieții* XIII (1–4): 1–9.

Wakefield, Andrew J., S. H. Murch, A. Anthony, J. Linnell, D. M. Casson, M. Malik, M. Berelowitz, et al. 1998. „RETRACTED: Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children". *The Lancet* 351 (9103): 637–41. https://doi.org/10.1016/S0140-6736(97)11096-0.

Walker, Jeffery T. 2021. „Ecological Fallacy". În *The Encyclopedia of Research Methods in Criminology and Criminal Justice*, ediție de J.C. Barnes și David R. Forde, 1 ed, 478–82. Wiley. https://doi.org/10.1002/9781119111931.ch98.

Weber, Elke U. 2016. „What Shapes Perceptions of Climate Change? New Research since 2010". *WIREs Climate Change* 7 (1): 125–34. https://doi.org/10.1002/wcc.377.

Weber, Robert Philip. 1990. *Basic content analysis*. 2nd ed. SAGE Quantitative Applications in the Social Sciences, no. 07-049. Newbury Park, Calif: Sage Publications.

Weindling, Paul Julian. 2004. *Nazi Medicine and the Nuremberg Trials. From Medical War Crimes to Informed Consent*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230506053.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wickham, Hadley, Evan Miller, și Danny Smith. 2022. „haven: Import and Export «SPSS», «Stata» and «SAS» Files". https://CRAN.R-project.org/package=haven.

Wickham, Hadley, François Romain, Henry Lionel, și Kirill Müller. 2022. „dplyr: A Grammar of Data Manipulation". https://CRAN.R-project.org/package=dplyr.

Wilke, Claus. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. First edition. Sebastopol, CA: O'Reilly Media.

Wollersheim, Jutta, Annett Lenz, Isabell M. Welpe, și Matthias Spörrle. 2015. „Me, Myself, and My University: A Multilevel Analysis of Individual and Institutional Determinants of Academic Performance". *Journal of Business Economics* 85 (3): 263–91. https://doi.org/10.1007/s11573-014-0735-3.

Wolter, Felix. 2019. „A New Version of the Item Count Technique for Asking Sensitive Questions: Testing the Performance of the Person Count Technique". *Methods* data (ianuarie): 24 Pages. https://doi.org/10.12758/MDA.2018.04.

Wolter, Felix, și Bastian Laier. 2014. „The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency". *Survey Research Methods* Vol 8 (decembrie): 153-168 Pages. https://doi.org/10.18148/SRM/2014.V8I3.5819.

Yin, Robert K. 2005. *Studiul de caz. Designul, analiza și colectarea datelor*. Iași: Polirom.

Zeileis, Achim, și Torsten Hothorn. 2022. „lmtest: Testing Linear Regression Models". https://CRAN.R-project.org/package=lmtest.

Pagini web cu instrumente de analiză a datelor și baze de date

R – https://www.r-project.org – program și limbaj de programare și analiză statistică

RStudio – https://posit.co/products/open-source/rstudio – program și interfață grafică de analiză statistică în R

JASP – https://jasp-stats.org – program și interfață grafică de analiză statistică în R

jamovi – https://www.jamovi.org – program și interfață grafică de analiză statistică în R

BlueSky Statistics – https://www.blueskystatistics.com – program și interfață grafică de analiză statistică în R

Stata – https://www.stata.com – program de analiză statistică

Tableau – https://www.tableau.com – program de analiză statistică și vizualizare

NVivo – https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home – program de analiză calitativă

MAXQDA – https://www.maxqda.com – program de analiză calitativă

ATLAS.ti – https://atlasti.com – program de analiză calitativă

StatConverter – https://roda.github.io/StatConverter – program de conversie a bazelor de date

Zotero – https://www.zotero.org – program gratuit de management al bibliografiei

European Social Survey – https://www.europeansocialsurvey.org/data/ – studiu comparativ european

World Values Survey – https://www.worldvaluessurvey.org/wvs.jsp – studiu comparativ mondial

European Election Study – https://www.gesis.org/en/services/finding-and-accessing-data/international-survey-programs/european-election-studies – studiu comparativ european

Experimente pe liste – http://list.sensitivequestions.org/articles/list-experiment-examples.html

Dataverse – https://dataverse.org – portal agregator de date

Open ICPSR – https://www.openicpsr.org/openicpsr – portal agregator de date

Eurostat – https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes – furnizor al datelor statistice comparative la nivel european

Banca Centrală Europeană – https://sdw.ecb.europa.eu – banca gestionară a zonei euro, furnizează date statistice comparative

OECD – https://stats.oecd.org – bază de date agregate ale țărilor membre OECD

Banca Mondială – https://databank.worldbank.org/home – bază de date agregate ale tuturor țărilor, cu indicatori sociali și economici

Organizația Mondială a Muncii – https://ilostat.ilo.org/data – bază de date agregate despre forța de muncă

Consortium of European Social Science Data Archives – https://datacatalogue.cessda.eu – portal agregator de date

Autoritatea Electorală Permanentă – http://alegeri.roaep.ro

Romanian Data Archive – http://www.roda.ro – portal agregator de date statistice sociale din România

Guvernul României – https://data.gov.ro – portal cu date deschise furnizate de ministere și agenții guvernamentale

Institutul Național de Statistică – http://statistici.insse.ro:8077/tempo-online – furnizor al datelor statistice oficiale din România

Banca Națională a României – https://bnr.ro/Baza-de-date-interactiva-604.aspx - banca gestionară a leului românesc, furnizează date statistice de politică monetară

Activitatea parlamentarilor din România – https://parlament.openpolitics.ro

ANEXĂ

Exemplu de ghid pentru grup focus[66]

Introducere (8-10 minute)

Vă rugăm să pregătiți badge-uri (călărești) cu numele participanților înainte de discuții, sau scrieți badge-uri (călărești) cu un marker înainte de începerea discuțiilor în funcție de numele persoanelor invitate să rămână la discuții, după ce semnează acordul de consimțământ informat. Celelalte persoane care s-au prezentat, dar care nu sunt invitate să rămână, vor primi un chestionar ce va măsura opiniile lor despre aceleași probleme pe care le-am inclus în ghidul de focus grup.

Introducerea de către moderator a obiectivelor și regulilor discuției (vă rugăm să considerați acest text ca o sugestie de introducere și să precizați toate regulile principale ale focus-grupului)

MODERATOR: *Bună ziua. Bine ați venit. Vă mulțumesc că ați acceptat invitația noastră. Ne-am întâlnit aici pentru a discuta disparitățile în accesul la îngrijire paliativă de bază în comunitate pentru bolnavii oncologici. Numele meu este ..... și sunt cercetător la ..... Întâlnirea de față este parte dintr-o cercetare științifică desfășurată de .....*

- *Suntem interesați de opiniile dumneavoastră în ceea ce privește subiectul discuției.*
- *Nu există răspunsuri și opinii bune sau proaste.*
- *Cea mai importantă regulă a acestei discuții este aceea că fiecare participant este liber să vorbească, să ia cuvântul pe parcursul discuției, și să-și prezinte propriile puncte de vedere.*

[66] Acest ghid a fost utilizat în focus grupurile organizate în cadrul proiectului „Overcoming disparities in access to quality basic palliative care in the community. Partnerships to identify and improve clinical, educational, legal, and economical barriers" finanțat de Programul de Cooperare Elvețiano-Român, implementat de Fundația Hospice Casa Speranței, Fundația MRC-Median Research Centre, și Spitalul Cantonal St.Gallen. Focus grupurile au fost realizate de Aurelian Muntean în decembrie 2013 și ianuarie 2014.

- *Dacă nu sunteți de acord cu opinia vreunui participant vă rog să vă exprimați punctul de vedere fără nici o teamă. Vă rog doar să nu vorbiți toți odată, astfel încât să putem înțelege ceea ce spuneți. Dorim să aflăm opinia fiecăruia dintre dumneavoastră.*
- *Această discuție este absolut confidențială și voluntară. Este împotriva regulilor etice după care ne ghidăm activitatea științifică să spunem altor persoane ceea ce o anumită persoană a spus aici.*

Înregistrarea discuțiilor

Vom înregistra audio discuția. Vom face acest lucru doar pentru a ne fi mai ușor să ascultăm opiniile dumneavoastră după această discuție. Vom folosi aceste înregistrări doar în acest scop. Înregistrările nu vor fi făcute publice fără permisiunea dumneavoastră. Vă rog să-mi spuneți dacă toată lumea este de acord cu înregistrarea discuțiilor.

PENTRU MODERATOR: ACEST GHID CONȚINE 4 SUBIECTE:

1. Disparități clinice
2. Disparități educaționale
3. Disparități legislative
4. Disparități financiare

ÎNCERCAȚI SĂ MENȚINEȚI DISCUȚIA PE FIECARE DINTRE ACESTE SUBIECTE.

DACĂ UN SUBIECT ESTE INTRODUS ÎN DISCUȚIE DE UN PARTICIPANT, ÎNAINTE DE A SE AJUNGE LA ACEL SUBIECT, PRECIZAȚI CĂ ACEL SUBIECT VA FI DISCUTAT MAI TÎRZIU.

MODERATOR: *Rog pe fiecare participant să se prezinte.*

MODERATORUL FACE O SCHEMĂ A GRUPULUI CU NUMELE FIECĂRUI PARTICIPANT.

Participanții se prezintă.

MODERATOR: *Acum vom începe discuția noastră.*

Discuția principală (95 de minute)

1. *Ce bariere/greutăți/limitări/dificultăți ați întâmpinat în asigurarea unui control adecvat al durerii sau al altor simptome la domiciliu, pentru pacientul cu cancer?*
2. *Ce temeri aveți în efectuarea unor manevre în îngrijirea pacienților oncologici la domiciliu? Dar piedici concrete ați întâmpinat?*
3. *Ce temeri aveți în informarea pacientului și a familiei asupra diagnosticului de cancer și în facilitarea înțelegerii bolii oncologice (inclusiv a prognosticului) de către aceștia? Dar bariere concrete ați întâmpinat?*
4. *Cine considerați ca ar trebui să coordoneze îngrijirea pacienților oncologici la domiciliu și ce bariere există în realizarea acestei coordonări?*
5. *Ce bariere întâmpinați în colaborarea cu alți specialiști / alte instituții?*
6. *Ce temeri aveți în asistarea la domiciliu a pacienților oncologici în ultimele zile/ore de viață? Dar dificultăți concrete ați întâmpinat?*
7. *Ce lipsuri educaționale v-au împiedicat să faceți un management adecvat al durerii și simptomelor pentru pacienții cu cancer aflați la domiciliu?*
8. *Ce bariere ați identificat în legislația și reglementările actuale, în legătură cu asistarea la domiciliu a acestor pacienți? Vorbiți-ne și despre bariere în accesul la medicamente și echipamente pentru pacienții îngrijiți în comunitate.*
9. *Ce limitări financiare ați întâmpinat în oferirea serviciilor de îngrijiri la domiciliu pentru pacienții cu cancer?*

EXERCIȚIU:

Participanții sunt împărțiți în două grupe, ambele primind câte un set de cartonașe A5 pe care este scrisă câte o problemă. Cartonașele trebuie printate în prealabil pe foi de hârtie similare ca dimensiune (A5). **NU NUMEROTAȚI ACESTE CARTONAŞE.** Problemele ce vor fi imprimate pe aceste cartonașe sunt: Comunicare deficitară; Tratament medicamentos neadecvat; Lipsa asistenței spirituale; Lipsa de empatie a personalului medical; Necoordonarea îngrijirii medicale; Lipsa banilor; Lipsa medicamentelor; Lipsa subvențiilor de stat; Pregătirea profesională insuficientă a personalului medical; Simptome insuficient tratate; Lipsa protocoalelor medicale; Necunoașterea legislației; Salarii nemotivante pentru personalul medical; Lipsa serviciilor de îngrijire la domiciliu; Lipsă suport familie / apropiați.

Grupurile de lucru mai primesc alte trei cartonașe goale pe care vor scrie probleme neexistente în lista celor primite.

INSTRUCȚIUNI PENTRU PARTICIPANȚI:

Vă voi da fiecărui grup câte un set de cartonașe. Pe fiecare cartonaș sunt scrise tipuri de probleme pe care oamenii le întâmpină atunci când apelează la serviciile medicale. Vă rog să vă gândiți care dintre acestea reprezintă disparități ale accesului indivizilor la serviciile de îngrijire paliativă. Folosiți cartonașele goale pentru a adăuga alte asemenea probleme, dacă credeți că sunt și alte probleme și disparități în furnizarea serviciilor medicale în localitatea dvs. / județul dvs.

Vă rugăm să împărțiți cartonașele în trei categorii:

*Lista 1). Include problemele care reprezintă **cele mai importante** forme de disparitate pe care o întâmpină beneficiarii de servicii medicale.*

*Lista 2). Include problemele care reprezintă **cele mai puțin importante** forme de disparitate pe care o întâmpină beneficiarii de servicii medicale.*

Lista 3). Include toate celelalte probleme care se află între celelalte două categorii.

Vă rugăm să lucrați împreună, în grup, iar rezultatul final să reprezinte punctul de vedere al tuturor membrilor. În cazul în care un membru nu este de acord cu acest rezultat, poate să își expună obiecțiile.

Concluzii (5-7 minute)

MODERATOR: *Am vorbit în această întâlnire despre diferite tipuri de disparități în accesul la îngrijirea paliativă în România. Am aflat multe lucruri interesante din aceste discuții. Înainte de a termina această întâlnire doriți să îmi puneți vreo întrebare, sau mai doriți să spuneți ceva?*

Vă mulțumim foarte mult pentru participarea dumneavoastră!

ISBN 978-606-37-1721-5