UNIVERSITATEA DE STAT DIN MOLDOVA
Facultatea de Psihologie și Științe ale Educației, Sociologie și Asistență Socială
Departamentul Sociologie și Asistență Socială

Oleg BULGARU

APLICAȚII STATISTICE
în cercetarea sociologică

Suport de curs

Aprobat
de Consiliul Calității al USM

CEP USM
Chişinău – 2018

CZU 311:303(075.8)
B 91

Recomandat de Consiliul Facultății de Psihologie și Științe ale Educației, Sociologie și Asistență Socială

Autor: ***Oleg BULGARU, doctor, conferențiar universitar***

Recenzent: **Svetlana TOLSTAIA,** ***doctor, conferențiar universitar***

Lucrarea reprezintă un suport de curs ce conține teme din domeniul cercetării sociologice cantitative, în care sunt utilizate prelucrări statistice ale datelor. Temele abordate pot fi întâlnite atât în cursul general *Metodologia cercetării sociologice*, cât și în cursurile specializate (*Aplicații statistice în cercetarea socială*, *Metode avansate în cercetarea socială, Managementul datelor*, *Măsurarea în științele sociale* etc.).

Destinatarii prezentei lucrări sunt studenţii de la facultăţile de ştiinţe socioumane, specialiştii în domeniul sociologiei, psihologiei, demografiei, asistenţei sociale, medicinii, economiei etc., toți cei interesați de utilizarea aplicațiilor statistice în cercetarea sociologică cantitativă.

Descrierea CIP a Camerei Naţionale a Cărţii
Bulgaru, Oleg
Aplicații statistice în cercetarea sociologică: Suport de curs / Oleg Bulgaru; Univ. de Stat din Moldova, Fac. de Psihologie și Științe ale Educației, Sociologie și Asistență Socială, Dep. Sociologie şi Asistenţă Socială. – Chişinău: CEP USM, 2018. – 150 p. : fig., tab.
Bibliogr.: p. 149 (12 tit.). – 50 ex.

ISBN 978-9975-142-17-5

311:303(075.8)
B 91

ISBN 978-9975-142-17-5

Cuprins

Prefață

„...gândirea statistică va deveni într-o zi la fel de necesară pentru un cetăţean eficient, la fel ca şi abilitatea de a citi şi a scrie”.

***H.G. Wells* (1866-1946)**

Mulți studenți vin la specialitățile socioumane (sociologie, psihologie, pedagogie, asistență socială etc.) pentru a scăpa de numere, în genere – de matematică, și pentru că le place să manipuleze cu cuvinte. Probabil că așa se întâmplă și cu dumneavoastră. Ați ales una din specialitățile socioumane pentru că sunteți fascinați de oameni, de comportamentul lor, de relațiile dintre ei, de viață lor interioară, chiar și de viața voastră proprie. Vă zicem un „Bine ați venit!” la acest curs de aplicații statistice în cercetarea sociologică și vă asigurăm că el este un curs care vă va lărgi orizontul, vă va pune la dispoziție un șir de metode și tehnici noi, bazate pe legități matematice, de studiere a acelorași oameni, ba chiar și a unor populații întregi.

Termenul „statistica” precum și primele conturări ale conceptului de statistică au pătruns în literatura de specialitate abia în secolul al XVIII-lea. Elementele concrete de evidență statistică își au însă originea în cele mai vechi timpuri. Statistica, ca știință, derivă din numeroase surse, unele chiar inedite. Ideea de bază de a aduna date provine de la necesitățile celor care guvernau (pentru a stabili taxele), dar și din timpuri mai vechi, când marinarii își calculau costurile echipării corăbiilor (bazându-se pe probabilitatea de a fi atacate de pirați sau de a naufragia).

La nivelul cercetării de orice tip, inclusiv cea sociologică, elementele de statistică au devenit indispensabile. Revistele științifice internaționale publică în majoritatea situațiilor date empirice, care să susțină eventualele construcții teoretice, iar standardele acestora în ceea ce privește prelucrarea datelor cantitative sunt foarte ridicate. Astfel că orice autor sau cercetător serios are nevoie de cunoștințe avansate de prelucrare a datelor empirice cantitative.

Prin această lucrare veți lua o primă cunoștință de programul SPSS (*Statistical Package for the Social Sciences*), care este unul dintre cele mai utilizate în prelucrarea și analiza statistică a datelor. De la prima versiune, apărută în anul 1968, programul a evoluat până la versiunea 25 şi aria de aplicabilitate s-a extins de la versiune la versiune, odată cu modul de operare și cu facilitățile oferite. Programul este utilizat astăzi nu numai în prelucrarea datelor sondajelor sociologice, dar și în cercetarea experimentală, în economie, marketing, educație, medicină, sănătate etc.

Lucrarea de față reprezintă un suport de curs și conține un număr de teme din domeniul cercetării sociologice cantitative, în care sunt utilizate prelucrări statistice ale datelor. Temele abordate pot fi întâlnite atât în cursul general *Metodologia cercetării sociologice*, cât și în cursurile specializate, cum ar fi *Metode statistice în cercetarea socială, Metode avansate în cercetarea socială, Managementul datelor*, *Măsurarea în științele sociale* ș.a. din domeniul general de studii – Științe sociale şi comportamentale. Fiecare temă a suportului de curs este urmată de exerciții și întrebări de control, care vor permite verificarea și aprofundarea cunoștințelor. Se recomandă ca rezolvarea problemelor, ce necesită calcule, să fie realizată în mediul programului Excel.

Suportul de curs *Aplicații statistice în cercetarea sociologică* este destinat studenților în programul de master „Sondaje de opinie, marketing și publicitate" de la Departamentul Sociologie și Asistență Socială din cadrul Universității de Stat din Moldova, care conține disciplina *Metode statistice în cercetarea socială* cu un volum de 150 de ore (15 ore curs, 30 de ore laborator și 105 ore lucru individual). Cursul respectiv își propune să prezinte, în modul cel mai serios și aprofundat, arsenalul metodelor și tehnicilor de culegere a informației, precum și al procedurilor avansate de prelucrare și interpretare a datelor obținute din cercetările empirice. Accentul este pus pe metodologiile cantitative, orientându-se pe aspectele cele mai moderne utilizate în cercetările recente. Predarea cursului este orientată spre însușirea tehnologiilor oferite de mijloacele electronice de calcul, dar nu doar în sensul aplicării mecanice de proceduri, ci insistându-se pe analiza critică a acestora. În sfârșit, se va completa arsenalul metodologic specific sociologiei cu cunoștințe generale privind cercetarea științifică, argumentarea și comunicarea rezultatelor obținute.

Acest suport de curs va contribui la cunoașterea elementelor de statistică descriptivă și la interpretarea lor, va permite tuturor absolvenților de profil să finalizeze studii și cercetări sociologice specifice organizațiilor în care vor activa. Mai mult, deprinderile de operare cu softuri specializate (SPSS) sau cu aplicația Excel din pachetul Microsoft Office vor contribui la proiectarea și realizarea bazelor de date pentru cercetările sociologice de profil, respectiv la operarea cu date de cercetare, prelucrarea acestora și analiza rezultatelor.

Autorul

Obiective

Obiectivele de bază ale cursului urmăresc:

- ***Să identifice noțiunile statistice de bază ale cercetării sociologice cantitative.***
- ***Să construiască eșantioane reprezentative.***
- ***Să elaboreze baza de date a cercetării.***
- ***Să utilizeze procedurile de verificare-corectare a datelor.***
- ***Să aplice metodele statistice la prelucrarea datelor și analiza rezultatelor cercetării.***
- ***Să utilizeze în comun programele SPSS și Excel pentru reprezentarea cât mai reușită a rezultatelor prelucrării datelor.***
- ***Să aplice metodele de prelucrare statistică a datelor în activitatea profesională.***

Competențe

Lucrarea va contribui la formarea următoarelor competențe profesionale:

- ***Determinarea componentelor de bază ale cercetării sociologice cantitative***
- ***Elaborarea și managementul bazei de date.***
- ***Prelucrarea statistică a datelor.***
- ***Gestionarea cazurilor și a variabilelor.***
- ***Utilizarea în comun a programelor SPSS și Excel pentru reprezentarea cât mai reușită a rezultatelor prelucrării datelor.***
- ***Dezvoltarea capacității de aplicare și transfer a cunoștințelor în vederea utilizării metodelor de prelucrare a datelor în activitatea profesională.***

Tema 1

Noţiuni statistice elementare. Baze de date statistice

Cuvântul „statistică” provine din limba italiană *statista,* ce desemna, în trecut, persoana care se ocupa de afacerile statului: număra populaţia sau alte elemente ce ajutau statul să gestioneze mai bine politica de taxe sau costurile războaielor. Acest termen este introdus în anul 1746 de către Gottfried Achenwall pentru a desemna *„ştiinţa de descriere a statului”*.

Def. 1.1. ***Statistica*** este disciplina care se ocupă cu culegerea, înregistrarea, gruparea, analiza şi interpretarea datelor referitoare la un anumit fenomen, precum şi cu formularea unor previziuni privind comportarea viitoare a acestuia.

Obiectul de studiu al statisticii îl constituie fenomenele şi procesele care prezintă următoarele particularităţi:

- se produc într-un număr mare de cazuri (sunt fenomene de masă);
- variază de la un element la altul, de la un caz la altul;
- sunt forme individuale de manifestare în timp, în spaţiu şi ca formă organizatorică.

Pentru rezolvarea problemelor, care fac obiectul său de studiu, statistica, ca orice ştiinţă, şi-a elaborat procedee şi metode speciale de cercetare, cum sunt cele ale observării de masă, ale centralizării şi grupării, procedee şi modele de analiză şi interpretare statistică. Putem spune că metoda statisticii este constituită din „totalitatea operaţiilor, tehnicilor, procedeelor şi metodelor de investigare statistică a fenomenelor ce aparţin unor procese de tip stocastic*”.

* Stocastic – întâmplător.

Complexitatea şi amploarea cercetării statistice fac imperios necesară perfecţionarea continuă a metodelor de observare, prelucrare, analiză. În acelaşi timp, dezvoltarea metodelor statisticii este strâns legată de progresele înregistrate de teoria probabilităţilor şi statistica matematică, precum şi de cele din domeniul informaticii.

Definiția 1.1 evidențiază două laturi ale statisticii, care poartă denumirile de *statistică descriptivă* și *statistică inferențială*.

Scopul principal şi specific ***statisticii descriptive*** este acela de a sintetiza şi structura, într-o manieră cât mai directă şi mai intuitivă, datele de observaţie şi informaţia conţinută de acestea. În atare sens utilizează, de regulă, tabele, grafice, indicatori statistici etc., prin care se obține descrierea fenomenului cercetat.

Statisticii inferenţiale îi revine rolul de a extinde rezultatele obţinute pe baza datelor din eşantion (o parte a populației cercetate) la nivelul populaţiei generale şi de a confirma sau invalida ipotezele emise *a priori* sau formulate după faza exploratorie.

Noţiunile statistice elementare sunt cele de *individ* (statistic) şi *populaţie* (statistică).

Def. 1.2. ***Indivizii*** sau ***unităţile statistice*** sunt nişte entităţi elementare, purtătoare de însuşiri (proprietăţi, caracteristici, calități).

Dintre toate însuşirile indivizilor se pot evidenţia una sau câteva *comune,* care exprimă natura însăşi a entităţilor respective, fiind atributul cu ajutorul căruia aceste entităţi sunt şi desemnate ca atare (oameni, ţări, mărfuri, plante etc.).

Celelalte însuşiri sunt *variabile,* diferă de la un individ la altul. Aşa, de exemplu, oamenii pot avea înălţimi diferite, opinii diferite, cunoştinţe diferite etc. Anume aceste însuşiri se studiază cu ajutorul instrumentelor statistice.

Unităţile statistice pot fi *simple* sau *complexe. Unităţile complexe* sunt rezultate ale organizării sociale ori economice a colectivităţii statistice (de exemplu, familia, colectivul întreprinderii).

Def. 1.3. Mulţimea indivizilor de aceeaşi natură formează ***populaţia statistică*** sau ***colectivitatea statistică***.

Numărul indivizilor ce formează populaţia statistică poate fi foarte diferit. Tehnicile, metodele statistice funcționează, de preferinţă, cu populaţiile mari.

Studierea populaţiilor mari pune probleme practice destul de dificile în culegerea şi prelucrarea informaţiei. Una dintre cele mai importante particularităţi ale statisticii este cea de cercetare a unei submulţimi (subpopulaţii, eşantion), foarte mici în comparaţie cu întreaga populaţie, şi generalizarea rezultatelor pentru întreaga populaţie. Astfel, se poate vorbi despre două tipuri de cercetări statistice: *cercetări exhaustive* (cercetări care cuprind populaţia în întregime sau *recensăminte*) şi *cercetări selective* (cercetări ale unei părți a populației special selectată, numită eşantion, sau *sondaje*).

Def. 1.4. ***Eşantion*** se numeşte acea parte a populaţiei asupra căreia se efectuează un studiu statistic (sau subset de elemente selectate dintr-o colectivitate statistică).

Def. 1.5. Prin ***reprezentativitate*** (a eşantionului) se înţelege proprietatea eşantionului de a reprezenta fidel populaţia.

Def. 1.6. Se numeşte ***variabilă statistică*** sau ***caracteristică*** proprietatea în funcţie de care se cercetează o populaţie statistică şi care, în general, poate fi măsurată, căpătând valori diferite de la un individ la altul.

Def. 1.7. ***Valoarea*** (starea, realizarea) reprezintă forma concretă de manifestare a unei variabile statistice pentru un individ.

Def. 1.8. Se numeşte ***scală*** totalitatea valorilor diferite ale unei caracteristici sau intervalul care le conţine (domeniul de valori al variabilei);

De exemplu, fie dată populaţia unei localităţi. Numărul de locuitori ai acestei localităţi reprezintă volumul populaţiei. În calitate de eşantion ar putea fi luaţi locuitorii de pe o stradă oarecare din localitate sau dintr-un bloc locativ. Locuitorii sunt acei indivizi care pot fi studiaţi prin metode statistice, culegându-se de la ei valori ale diferitelor caracteristici, cum ar fi sexul, nivelul studiilor, vârsta, opiniile faţă de o problemă sau de un eveniment etc. Valorile acestor caracteristici, desigur, vor fi diferite de la un individ la altul, dar se vor încadra în nişte limite − domenii de valori (sexul poate fi feminin sau masculin, vârsta poate fi între 0 şi 200 de ani, de exemplu, etc.).

Valorile caracteristicilor studiate, culese de la indivizii din populaţie, reprezintă nişte date statistice (mărimi concrete, determinate prin numărare, măsurare, intervievare etc.), care se grupează în aşa-numitele baze de date.

Def. 1.9. Se numeşte ***bază de date*** un set structurat de date pentru a le putea regăsi cât mai rapid şi mai eficient.

Sistemul de structurare a datelor care se utilizează cel mai frecvent este *tabelul*, iar în cazul volumelor şi diversităţii mari de date, vorbim despre *baze de date*, formate din tabele legate între ele.

Datele culese dintr-o populaţie prin metoda anchetei sociologice se structurează, de regulă, tot sub formă de tabel, coloanele căruia corespund caracteristicilor, iar liniile − indivizilor. Astfel, acest tabel reprezintă baza de date a cercetării (a se vedea Tabelul 1.1).

Tabelul 1.1

Structura tabelului – bază de date a cercetării prin metoda anchetei sociologice

	Caracteristica 0 *(nume individ)*	**Caracteristica 1**	**Caracteristica 2**	**Caracteristica 3**	**…**
1	Individul 1	Valoarea 11	Valoarea 21	Valoarea 31	…
2	Individul 2	Valoarea 12	Valoarea 22	Valoarea 32	…
3	Individul 3	Valoarea 13	Valoarea 23	Valoarea 33	…
...	…	…	…	…	…

Să presupunem că se studiază, de exemplu, participarea populaţiei dintr-o localitate la ultimele alegeri parlamentare în funcție de sexul, studiile și vârsta indivizilor. Datele pot fi culese de la indivizi prin intermediul următoarelor întrebări:

A1. Dvs. aţi participat la alegerile parlamentare?

1. Da
2. Nu

D1. Sexul individului

1. Feminin
2. Masculin

D2. Ce studii aveţi Dvs.?

1. Fără studii
2. Primare
3. Medii
4. Superioare

D3. Indicați vârsta Dvs. _____ ani.

Răspunsurile indivizilor (fie *Ion, Vasile, Ana* etc. − numele câtorva dintre ei) pot fi introduse într-un tabel de forma Tabelului 1.1, coloanele căruia corespund caracteristicilor studiate (*nume, votat, sex, studii* etc.), determinate de sensul întrebărilor, iar liniile − indivizilor (Tabelul 1.2).

Tabelul 1.2

Exemplu de bază de date completată

	nume	**votat**	**sex**	**studii**	**varsta**	**…**
1	Ion	da	masculin	medii	22	…
2	Vasile	nu	masculin	superioare	30	…
3	Ana	da	feminin	superioare	28	…
...	…	…	…	…		…

Însă, pentru prelucrarea statistică a datelor, e mai comod de utilizat valori numerice ale variabilelor, şi nu valori textuale. În acest scop, valorile înregistrate ale variabilelor, dacă acestea nu sunt numerice, se codifică, de regulă, cu numere întregi. Cele din urmă, deseori, sunt nu altceva decât numerele de ordine ale variantelor de răspuns din întrebări. Deoarece în majoritatea cazurilor răspunsurile la întrebări sunt anonime (numele indivizilor nu se înregistrează), prima coloană a bazei de date se foloseşte pentru numerele de ordine ale indivizilor intervievați (ale respondenților).

Presupunând, de exemplu, că numerele de ordine ale indivizilor din exemplul de mai sus sunt 17 (Ion), 29 (Vasile), 103 (Ana) etc., baza de date din Tabelul 2.2 va primi forma prezentată de Tabelul 1.3, în care deja figurează numai numere: coduri ale variantelor de răspuns sau valori ale caracteristicilor numerice.

Tabelul 1.3

Bază cu date codificate

	nume	**votat**	**sex**	**studii**	**varsta**	**…**
1	17	1	2	2	22	…
2	29	2	2	3	30	…
3	103	1	1	3	28	…
...	…	…	…	…		…

În continuare, datele din tabele, astfel construite şi completate, pot fi prelucrate, utilizând diferite metode statistice, ca rezultat obţinându-se răspunsuri la un şir de întrebări, cum ar fi: *„Ce parte din populaţia cercetată a participat la alegeri?”, „Cum au participat la alegeri femeile şi bărbaţii?”, „Cum sunt repartizaţi după nivelul de studii indivizii din localitate?”* şi multe, multe altele.

Un astfel de studiu, care urmăreşte obţinerea şi prelucrarea informaţiilor dintr-o populaţie, reprezintă scopul cercetării sociologice cantitative, despre care se va vorbi în continuare.

Exerciții, întrebări de control

1. Care va fi populația necesară de sondat pentru a prezice viitoarea structură a Parlamentului țării?
2. În cursa pentru fotoliul de primar al capitalei au fost acceptate 9 persoane. Definiți populația ce trebuie sondată pentru a prezice șansele candidaților la câștig?
3. Definiți populația ce trebuie sondată pentru a cerceta situația copiilor de vârstă timpurie (0-7 ani). Cine vor fi respondenții într-un astfel de sondaj?
4. Care va fi populația cercetată pentru a determina mărcile automobilelor implicate în accidente rutiere?
5. Într-o instituție de învățământ superior a fost formulată problema de a cerceta absenteismul de la ore al studenților de la diferiți ani de studii (licență și masterat). Care va fi populația cercetată și ce caracteristici vor fi culese de la respondenți?

6. Indicați care din următoarele grupuri de indivizi formează o populație sau un eșantion:
   a) studenții Universității de Stat din Moldova (USM);
   b) studenții Facultății de Drept de la USM;
   c) un grup de persoane din or. Chișinău;
   d) primarii localităților din r-nul Ialoveni;
   e) o lingură de fasole luate din cratița în care ele se fierb;
   f) 10 nuci dintr-un sac cu nuci;
   g) o alee de copaci din parc.

   Pentru populații – dați exemple de eșantioane, iar pentru eșantioane – numiți populațiile din care au fost extrase.
7. Elaborați structura bazei de date în scopul de a studia opțiunile electorale ale bărbaților și femeilor, ale cetățenilor din mediul rural și cel urban, ale tinerilor, adulților și persoanelor în vârstă, ale diferitelor etnii din țară pentru alegerea Parlamentului.
8. O bază de date conține următoarele caracteristici ale indivizilor: vârsta (în ani), mediul de reședință (sat, oraș), opinia față de diferite canale TV (preferat, indiferent, nepreferat). Care din următoarele informații pot fi determinate din această bază de date:
   a) repartizarea respondenților după vârstă;
   b) repartizarea respondenților după culoarea ochilor;
   c) procentul pensionarilor din populația cercetată;
   d) numărul de canale TV preferate de fiecare respondent;
   e) care canal TV e cel mai preferat de respondenții de la sat;
   f) atitudinea ucrainenilor față de canalele TV.

Tema 2

Măsurarea în științele sociale. Scalarea

În tema precedentă s-a vorbit despre determinarea valorilor variabilelor statistice pentru diferiți indivizi, înlocuirea unora din ele, exprimate prin cuvinte, cu valori numerice etc. În continuare vom analiza mai pe larg aceste lucruri.

Def. 2.1. ***Măsurarea*** reprezintă o exprimare simbolică, numerică sau nenumerică, a gradului în care un obiect sau fenomen posedă o anumită caracteristică sau proprietate. Aceasta expresie simbolică permite să se compare obiecte şi fenomene concrete între ele.

Ca exemple de măsurare, utilizate frecvent, servesc: măsurarea greutăţii, vitezei, lungimii, temperaturii, dar şi, de exemplu, nivelului de cunoaştere a unui obiect sau eveniment, nivelului de încredere într-un politician, stării civile şi sexului unei persoane etc. Instrumentul cu ajutorul căruia se realizează măsurarea se numeşte *scală* (de exemplu, scala metrică sau metrul în cazul măsurării lungimii), iar activitatea de construire a scalelor − *scalare*.

Def. 2.2. ***Scalarea*** (engl. *scaling*, rus. *шкалирование*) − activitate de construire a scalelor. Cu alte cuvinte, scalarea cuprinde totalitatea metodelor, procedurilor, modalităților de construire a scalelor de diferite tipuri, de modificare a lor.

Deoarece termenul *scalar* are sensul de *valoare numerică,* prin scalare se mai înțelege și atribuirea de numere sau de alte constructe matematice obiectelor. Scala, așadar, reprezintă regula unei astfel de atribuiri. Numerele obținute ca rezultat al scalării se mai numesc valori scalare.

Suplimentar, scalarea urmărește obiectivele *infra*:

- metoda propusă să fie atât de simplă, încât datele obţinute prin măsurare să fie adecvate condiţiilor existente;
- metoda să corespundă unui nivel cât mai înalt de măsurare (despre nivelurile de măsurare se va vorbi în continuare), în aşa fel ca la prelucrarea datelor să se poată folosi metodele numerice tradiţionale (în special, atunci când datele se organizează în baze de date statistice);
- metoda să fie funcţională, astfel încât rezultatele obţinute în baza ei pe eşantion să poată fi transferate pentru întreaga populaţie.

O scală de calitate asigură o măsurare de calitate. Pentru a asigura calitatea măsurării, este necesar ca la elaborarea scalei să fie îndeplinite două condiţii: a) ea *să fie înţeleasă* de către subiecţii de la care se culeg informaţiile; b) ea *să diferenţieze* nivelurile de intensitate ale proprietăţilor fenomenului cercetat, adică să cuprindă toate variantele posibile de situaţii.

Pentru măsurarea proprietăţilor fenomenelor în cercetările sociologice, de marketing etc., se utilizează patru tipuri de scale: *nominală, ordinală, de interval* şi *de raport*, corespunzătoare celor patru niveluri de măsurare: nominal, ordinal, de interval și de raport.

Scala nominală este cea mai simplă din punctul de vedere al capacităţii de măsurare, fiind şi cea mai puţin restrictivă din perspectiva instrumentului statistico-matematic. Respectiv, *nivelul nominal* de masurare este cel mai inferior dintre toate nivelurile de măsurare.

Scala nominală permite clasificarea subiecţilor cercetaţi în grupe (două sau mai multe), ai căror membri diferă în funcţie de proprietatea ce a fost scalată. Scala nominală nu permite însă ordonarea acestor subiecţi în funcţie de intensitatea proprietăţilor fenomenului cercetat sau de măsurarea distanţelor care îi separă (acestea nici nu pot fi definite!). Practic, *toate componentele unei grupe vor primi acelaşi simbol numeric, de regulă − un număr întreg, indicând apartenenţa unei componente la o anumită grupă*. Pot fi aduse numeroase exemple de proprietăți ce se măsoară cu scala nominală: sexul, culoarea ochilor, starea civilă, specialitatea, naționalitatea etc. În construirea unei anumite scale nominale, se va urmări ca, în clasificarea propusă, să fie prevăzute toate grupele posibile, recurgând, în unele cazuri, chiar şi la variante de tipul „altul”, „alta”, „altceva” etc. În acelaşi timp, este necesar ca grupele să se excludă reciproc din punctul de vedere al proprietăţii scalate.

În calitate de exemplu de scalare a unei caracteristici nominale poate servi procedura de prelucrare a răspunsurilor la întrebările care presupun obţinerea de valori nenumerice ce nu pot fi ordonate (de exemplu: *În ce domeniu activați Dvs.?*). De regulă, toate răspunsurile obţinute de la respondenţi se clasifică, fiecare clasă reprezentând un item al scalei nominale (în exemplul nostru scala nominală ar putea avea următorii itemi: 1 – *transport,* 2 – *construcții,* 3 – *industrie,* 4 – *învățământ,* 5 – *știință,* 6 – *alimentație publică, 7 – agricultură, 8 – altul*).

Scala ordinală, la fel ca scala nominală, clasifică diverse situaţii, evenimente, obiecte sau fenomene, însă între subiecţii din diferite grupe este introdusă o relaţie suplimentară, de ordine. Respectiv, *nivelul ordinal* de măsurare este superior celui nominal.

Scala ordinală permite ordonarea subiecţilor cercetaţi în funcţie de o anumită preferinţă, de un anumit criteriu, *folosindu-se pentru codificare, de data aceasta, şiruri ordonate de numere întregi*, nepermiţând însă evaluarea distanţelor dintre variante. În scopul prelucrării ulterioare a datelor cu ajutorul metodelor numerice, anume pentru astfel de scale au fost elaborate un şir de metode de scalare, care transformă valorile calitative ale caracteristicilor în valori numerice.

Vom prezenta în continuare câteva metode de scalare pentru astfel de scale, frecvent utilizate, care au succes datorită uşurinţei în aplicare şi calităţii informaţionale obţinute.

Metoda *diferenţialei semantice*, creată de Charles E. Osgood şi dezvoltată ulterior de alţi cercetători, porneşte de la identificarea acelor cuvinte opuse (perechi de adjective bipolare, antonime) care pot descrie subiectul cercetat. Ele vor fi plasate pe scala ce poate avea un număr impar de trepte − 3, 5 sau 7. De exemplu, respondentului i se propune să aprecieze calitatea unui produs pe o scală cu 5 trepte, marcând cu semnul **X** segmentul care corespunde opiniei sale:

Foarte joasă	_____	_____	_____	__**X**__	_____	**Foarte înaltă**

Sau, pentru o prelucrare cantitativă ulterioară, scala se propune având variantele codificate:

Foarte joasă	**1**	**2**	**3**	**4**	**5**	**Foarte înaltă**

După ce fiecare persoană investigată a încercuit numărul care reprezintă opinia sa, cercetătorul are posibilitatea să facă o medie a tuturor opiniilor, stabilind un punct final pe scală, sintetizând imaginea eşantionului cercetat. Această medie poate fi comparată apoi cu mediile obţinute la alte produse sau servicii, cu mediile altor eşantioane sau cu media aceluiaşi eşantion, obţinută într-o altă perioadă de timp.

Scala lui Stapel reprezintă o variantă, asemănătoare cu diferenţiala semantică. Ea are 10 niveluri: 5 cu semnul „+” şi 5 cu semnul „–”, iar între aceste zone se inserează atributul ce urmează a fi evaluat (un nivel mediu sau de mijloc nu există!):

-5	**-4**	**-3**	**-2**	**-1**	**Nivelul înţelegerii unei teme de curs**	**+1**	**+2**	**+3**	**+4**	**+5**

Subiecţii investigaţi încercuiesc numărul care reprezintă opinia lor. Prelucrarea datelor este asemănătoare cu cea specifică diferenţialei semantice, ambele conducând la informaţii specifice scalelor de tip interval.

Scala Likert, la fel, reprezintă o scală ordinală, care se foloseşte pentru a aprecia mai multe afirmaţii cu calificative cuprinse între un *„acord total”* până la un *„dezacord total”*. Aceste afirmaţii se compun pentru a descrie diferite laturi (aspecte) ale unui fenomen (obiect), pentru ca în consecință să se găsească o medie pentru descrierea fenomenului (obiectului) în întregime. Numărul de trepte ale scalei este unul şi acelaşi pentru toate afirmaţiile despre fenomen (obiect).

Etapele de lucru cu scalele Likert sunt următoarele:

- se alcătuieşte un set de propoziţii care reprezintă afirmaţii cu caracter favorabil sau nefavorabil despre fenomenul (obiectul) investigat;

- propoziţiile sunt prezentate subiecţiilor, care trebuie să-şi dea acordul sau dezacordul încercuind una din gradaţiile scalei (de exemplu, cu cinci trepte):

Acord total	**+2**	**+1**	**0**	**-1**	**-2**	**Dezacord total**

- scorul realizat de un subiect se calculează făcând suma algebrică a valorilor.

Prin *metoda comparaţiilor perechi* respondentul trebuie să indice care din cele două obiecte din perechea evaluată are o poziţie mai bună în ceea ce priveşte atributele care stau la baza comparaţiei. De exemplu, se testează $n=4$ variante A, B, C, D, deci este posibil să se realizeze $n(n-1)/2$ comparaţii sau se pot forma şi compara 6 perechi (A-B, A-C, A-D, B-C, B-D, C-D). Datele obţinute pot fi analizate şi interpretate cu ajutorul metodelor specifice scalelor ordinale.

În continuare, vom exemplifica această metodă, pe o scală ordinală, în cazul a patru probleme din societate (*corupția, migrația, sărăcia, șomajul*), prin prelucrarea răspunsurilor tuturor indivizilor dintr-un eșantion de 200 de persoane, solicitate să indice în cazul fiecărei perechi problema ce-i îngrijorează mai mult; indecișii (non-răspunsurile) nu vor intra în calcul la comparațiile respective.

Rezultatele sunt prezentate în Tabelul 2.1. Fiecare celulă a tabelului indică numărul de persoane care consideră că problema din coloana respectivă (j) e mai îngrijorătoare decât cea din rândul respectiv (i). Deoarece problemele nu se compară cu ele însele, diagonala principală nu conține date.

Pentru a putea interpreta datele Tabelului 2.1, distribuția de frecvente absolute se poate exprima sub formă de proporții, redate în paranteze, tot în Tabelul 2.1.

Pentru a stabili ordinea problemelor ce-i îngrijorează cel mai mult pe respondenți, în baza datelor din Tabelul 2.1 se elaborează un nou tabel (a se vedea Tabelul 2.2), în care în toate celulele cu proporții mai mari de 0,50 se trece cifra „1”, iar in celulele cu proporții mai mici sau egale cu 0,50 – cifra „0”.

Tabelul 2.1

Numărul (proporția) persoanelor îngrijorate mai mult de problema din coloana „j”, in comparație cu cea din rândul „i”

Problema (rândul „i”)	**Problema (coloana „j”)**			
	corupția	*migrația*	*sărăcia*	*șomajul*
corupția	-	80 (0,40)	70 (0,35)	50 (0,25)
migrația	120 (0,60)	-	140 (0,70)	90 (0,45)
sărăcia	130 (0,65)	60 (0,30)	-	50 (0,25)
șomajul	150 (0,75)	110 (0,55)	150 (0,75)	-

Cele doua cifre au următoarele semnificații:

1 – problema respectivă îngrijorează mai mult în perechea considerată;

0 – problema respectivă îngrijorează mai puțin în perechea considerată.

Tabelul 2.2

Distribuția nivelurilor de îngrijorare pentru cele patru probleme analizate

Problema	**Problema**			
	corupția	*migrația*	*sărăcia*	*șomajul*
corupția	-	0	0	0
migrația	1	-	1	0
sărăcia	1	0	-	0
șomajul	1	1	1	-
Suma frecvențelor	3	1	2	0

Suma frecvențelor reflectă locul ocupat de fiecare problemă după nivelul de îngrijorare al respondenților, respectiv: *corupția* – locul întâi, *sărăcia* – locul al doilea, *migrația* – locul al treilea și *șomajul* – locul al patrulea.

Metoda comparațiilor perechi este avantajoasă pentru un număr mic de variante, datorita faptului că permite:

- comparare directă și expunerea unei comparații deschise a preferințelor;

- urmărirea într-un timp foarte scurt a reacțiilor comparative ale respondenților.

Pe de alta parte, rezultatele comparațiilor pot fi neconcludente, iar metoda devine anevoioasă pentru un număr mare de variante.

Faptul că o variantă este preferată alteia, nu înseamnă că în mod absolut aceasta este și dorită sau plăcută. Varianta comparată poate fi apreciată cu mai puține aspecte negative decât celelalte și numai din acest punct de vedere apare a fi preferată.

Menționăm că metoda comparațiilor perechi este potrivită pentru cercetarea de marketing, deoarece permite colectarea de date privind preferințele față de produse, servicii etc.

Prin *metoda ordonării rangurilor,* subiectului i se cere să considere toate alternativele odată, să le compare, apoi să le ordoneze în funcție de o anumită caracteristică. Ea se aplică cu ușurință atunci când numărul obiectelor sau fenomenelor este mare, este mai economică, conduce la rezultate mai precise, iar pentru interpretarea datelor se pot folosi metode statistice caracteristice scalei ordinale.

Metoda ordonării rangurilor este apreciată de specialiști ca fiind deosebit de eficienta, ea prezentând următoarele avantaje față de metoda comparațiilor perechi:

- evită erorile de tranzitivitate, posibile în cazul metodei comparațiilor perechi (de exemplu, s-ar putea să se aprecieze că A este preferat față de B și B este preferat față de C, pentru ca apoi, eronat, să se aprecieze că C este preferat față de A);

- poate fi utilizată cu ușurință și dacă numărul variantelor este mai mare, fiind totodată mai economică și mai simplu de gestionat, conducând și la rezultate mai precise și mai puțin distorsionate de erorile de răspuns.

Utilizarea acestei metode presupune evaluarea concomitentă a tuturor variantelor de comparat și solicitarea respondenților de a le ordona în funcție de un anumit atribut. Scala utilizată pentru prelucrarea datelor este de tip ordinal.

Vom exemplifica, în continuare, utilizarea metodei ordonării rangurilor într-o cercetare, având ca obiectiv evaluarea a cinci valori din punctul de vedere al importanței acestora și prelucrarea răspunsurilor înregistrate pe un eșantion format din 8 respondenți. Respondenții au ordonat valorile propuse spre a fi evaluate, atribuindu-le locuri de la 1 (cea mai importantă) până la 5 (cea mai puțin importantă). Rezultatul este prezentat în Tabelul 2.3:

Tabelul 2.3

Evaluarea importanței valorilor (locul atribuit)

Numărul respondentului	*Familia*	*Serviciul*	*Prietenii*	*Studiile*	*Timpul liber*
1	1	5	3	4	2
2	4	5	1	3	2
3	2	1	3	4	5
4	1	4	2	3	5
5	1	3	4	2	5
6	2	4	3	1	5
7	5	1	4	2	3
8	1	3	2	5	4

În continuare numărul respondenților care plasează valorile pe locurile 1…5 se centralizează într-un alt tabel sub aspectul importanței (a se vedea Tabelul 2.4):

Tabelul 2.4

Numărul de locuri diferite atribuite valorilor

	Locuri 1 (5 pct)	Locuri 2 (4 pct)	Locuri 3 (3 pct)	Locuri 4 (2 pct)	Locuri 5 (1 pct)	Total
Familia	4	2	0	1	1	8
Serviciul	2	0	2	2	2	8
Prietenii	1	2	3	2	0	8
Studiile	1	2	2	2	1	8
Timpul liber	0	2	1	1	4	8

Celor cinci locuri li se acordă punctaje conform următoarei reguli: locului 1 – 5 puncte, locului 2 – 4 puncte, locului 3 – 3 puncte, locului 4 – 2 puncte și locului 5 – 1 punct.

Ierarhia fiecărei valori se va determina prin ponderarea ei cu punctajul acordat locului pe care a fost plasată, astfel:

Familia = 4*5+2*4+0*3+1*2+1*1 = 31

Serviciul = 2*5+0*4+2*3+2*2+2*1 = 21

Prietenii = 1*5+2*4+3*3+2*2+0*1 = 26

Studiile = 1*5+2*4+2*3+2*2+1*1 = 24

Timpul liber = 0*5+2*4+1*3+1*2+4*1 = 17

Ierarhia finala a celor cinci valori, din punctul de vedere al aprecierii celor 8 respondenți investigați este:

Familia (31) > *Prietenii* (26) > *Studiile* (24) >
> *Serviciul* (21) >*Timpul liber* (17)

O variantă simplificată a metodei ordonării rangurilor se folosește, atunci când respondenților li se cere să ordoneze după importanță numai o parte (de exemplu, 3) din valorile propuse. Atunci, pentru prelucrarea răspunsurilor, la fel se construiesc tabele de tipul 2.3 și 2.4, în care valorilor neordonate nu li se atribuie locuri, celulele respective rămânând goale/necompletate.

Scala cu suma constantă impune subiectul să împartă o sumă constantă (10 sau 100) între două sau mai multe variante de apreciere. Informația este de calitate mai ridicată, deoarece este măsurată cu ajutorul unei scale numerice.

Vom demonstra aplicarea scalei cu sumă constantă pentru aprecierea calităților unui profesor de facultate în viziunea studenților. În acest un grup de 8 studenți au fost rugați să împartă 100 de puncte la următoarele calități: *cunoașterea materialului, modalitatea de predare, atitudinea față de studenți*. Rezultatele sunt reflectate în Tabelul 2.5:

Tabelul 2.5

Aprecierea calităților profesorului de către studenți

Numărul studentului	**Puncte acordate pentru calitate**			**Total puncte**
	cunoaștere	*predare*	*atitudine*	
1	20	50	30	100
2	40	30	30	100
3	15	45	40	100
4	10	40	50	100
5	25	65	10	100
6	60	30	10	100
7	35	40	25	100
8	45	30	25	100
Total:	**250**	**330**	**220**	**800**

Importanța calității profesorului se determină după punctajul mediu obținut (sumele pe colane în tabel împărțite la numărul studenților intervievați):

cunoașterea materialului: 250/8=31,25;

modalitatea de predare: 330/8=41,25;

atitudinea față de studenți: 220/8=27,5.

Analizând comparativ punctajele obținute, remarcăm faptul că *modalitatea de predare* a profesorului reprezintă calitatea cea mai importantă în viziunea studenților intervievați, urmată de *cunoașterea materialului* de către profesor și *atitudinea lui față de studenți*.

În practică se mai folosesc şi alte scale. De exemplu, *scala Guttman*, care permite ca prin răspunsuri de tip „Da” – „Nu” la anumite întrebări să fie evaluate atitudinile şi satisfacţiile respondenţilor, sau *scala Thurstone*, utilă în cercetări referitoare la atitudini, intenţii, preferinţe sau comportament ale respondenţilor.

Scalele numerice (de interval şi de raport) se utilizează pentru măsurări ale caracteristicilor cantitative şi, de regulă, nu se construiesc, având la bază etaloane respective sau convenţii.

Nivelul de interval **(**numit şi *nivelul cardinal***)** este foarte util pentru că permite determinarea distanţelor şi diferenţelor dintre variante. Pentru acest nivel este caracteristic faptul că originea este marcată de un zero convenţional.

Scala de interval se bazează pe utilizarea unor unităţi de măsură egale, făcând posibilă stabilirea atât a ordinii variantelor analizate, cât şi a distanţelor dintre acestea. Se stabileşte un nivel de pornire zero, de la care se creează trepte sau grade plasate la distanţe egale unele de altele. Semnificaţia punctului zero (original), cât şi mărimea unităţii de măsură vor fi stabilite de cercetători. Exemple sunt scalele de măsurare a temperaturii Celsius și Fahrenheit în care, după cum se cunoaşte, zero reprezintă la prima scală punctul de înghețare a apei, iar la a doua punctul de înghețare a unui amestec de clorură de amoniu.

În cercetările de marketing, variabilele de tip interval sunt foarte des utilizate. Scalele de atitudine sunt, în general, considerate ca fiind scale de interval. Pe aceste scale se consideră că fiecare interval are aceeași lungime, în acest caz diferenţele dintre atitudini având sens.

Nivelul proporţional (de proporţii sau *de raport)* este acela care asigură măsurarea cea mai riguroasă, utilizând unităţi de măsură reale (cifra de afaceri, profit, volum, greutate ş.a.). Punctul de plecare este, de asemenea, zero, dar acesta este zero natural, care semnifică absenţa proprietăţii respective la fenomenul cercetat. De la acest zero se creează trepte plasate la distanţe egale una faţă de cealaltă.

Scala proporţională este cea mai complexă dintre toate tipurile de scale. Ca şi cea anterioară, este împărţită în intervale egale, fiecăruia dintre acestea corespunzându-i un anumit număr. O asemenea scală are însă un „zero” unic, acest „zero” indicând „absenţa”, respectiv – o cantitate nulă, o viteză nulă. Diferitele unităţi de măsură pentru exprimarea lungimii, vânzărilor, vitezei sunt exemple semnificative de scale proporţionale. Ele oferă posibilitatea efectuării tuturor operaţiunilor admise de celelalte tipuri de scală prezentate, inclusiv multiplicarea şi divizarea unui număr de pe scală la altul.

Exerciții, întrebări de control

1. Explicați deosebirile dintre noțiunile *măsurare, scală* și *scalare.*
2. Propuneți trei exemple de scale nominale.
3. Ce au comun și prin ce se deosebesc scalele ordinale de cele nominale.
4. Construiți un exemplu de măsurare cu scala diferențialei semantice a unei caracteristici pentru 10 indivizi. Determinați valoarea „medie” a acestei caracteristici pentru grupul de indivizi.
5. Construiți un exemplu de măsurare cu scala lui Stapel a unei caracteristici pentru 10 indivizi. Determinați valoarea „medie” a acestei caracteristici pentru grupul de indivizi.
6. Alcătuiți cinci propoziții-afirmații și propuneți o scală Likert de apreciere a lor.
7. Propuneți trei probleme și examinați importanța lor, în cadrul grupei academice, prin metoda comparațiilor perechi.
8. Identificați patru obiecte importante pentru viața omului. Prin metoda ordonării rangurilor, evaluați în cadrul grupei academice importanța acestor obiecte.
9. Apreciați în cadrul grupei academice blocul de studii în care vă aflați prin metoda scalei cu sumă constantă după următoarele calități: amplasare geografică, înfățișare exterioară, condiții de instruire, dotare cu echipamente.
10. Propuneți câte trei exemple de scale de interval și scale de raport.

Tema 3

Variabile statistice: clasificare, indicatori

Reamintim, pentru început, că ***variabilă statistică*** este o proprietate (caracteristică, însuşire), în funcţie de care se cercetează o populaţie statistică şi care, în general, poate fi măsurată, căpătând valori diferite de la un individ la altul. Prin *valoare (stare, realizare)* se înţelege forma concretă de manifestare a unei variabile statistice pentru un individ, iar *scala* reprezintă instrumentul cu care se măsoară această valoare. Pe de altă parte, se poate spune că scala reprezintă totalitatea valorilor diferite ale unei caracteristici sau intervalul care le conţine (numit și domeniu de valori ale variabilei).

Procesul prin care se obţin valorile variabilelor sau atribuirea de valori caracteristicilor indivizilor potrivit unor reguli, după cum s-a menționat în tema precedentă, se numeşte ***măsurare***.

Def. 3.1. Prin ***cercetare*** vom înţelege studiul variabilelor şi al relaţiilor dintre ele.

În continuare vom nota variabilele în felul următor:

<*nume*> = {<*domeniu de valori*>}.

De exemplu, variabila ce caracterizează sexul individului se va scrie: *sex={feminin, masculin},* variabile ce caracterizează nivelul de studii − *studii={fără studii, primare, medii, superioare, altele}*, variabila ce caracterizează vârsta − *varsta={[18, 24]},* unde [18, 24] sunt vârste exprimate în ani din intervalul 18-24 de ani inclusiv etc.

Variabilele statistice se clasifică după un şir de criterii. Astfel:

a) După modul de exprimare se deosebesc *variabile calitative* şi *variabile cantitative:*

- ***variabile calitative*** − variabile ale căror valori sunt exprimate prin cuvinte care desemnează apartenenţa individului la una din categoriile scalei (exemple: sexul, calificativul, profesia, starea civilă etc.).

Variabilele calitative sunt de două tipuri: *nominale* şi *ordinale*.

- ***variabile cantitative*** − variabile ale căror valori se exprimă numeric (exemple: vârsta, salariul, înălţimea etc.).

Variabilele cantitative, la fel, sunt de două tipuri: *de interval* şi *de raport*.

b) După numărul de valori ale variabilelor calitative, se cunosc *variabile dihotomice* şi *variabile categoriale*:

- ***variabile dihotomice*** (binare, alternative) − variabile calitative a căror scală e compusă din două valori antonime (*da – nu, prezent – absent, aprins – stins* etc.).

Noţiunea de „variabilă binară” provine de la codificarea valorilor acestora cu 0 şi 1. Codificarea prin 0/1 permite utilizarea acestor variabile în proceduri dedicate nivelurilor mai înalte de măsurare (ordinal, de interval).

- ***variabile categoriale*** (nealternative) − celelalte variabile calitative ce nu posedă proprietăţi ale variabilelor dihotomice.

c) După modul de obţinere variabilele se clasifică în *primare* şi *derivate*:

- ***variabile primare*** − variabile obţinute în etapa de culegere a datelor (exemplu: *vârsta* înregistrată în ani, *notele* primite de student la examenele din sesiune etc.);
- ***variabile derivate*** (auxiliare) − variabile obţinute în urma procesului de prelucrare a variabilelor primare (exemplu: *vârsta* pe grupe de vârstă, *nota medie* la sesiune etc.).

d) După natura variaţiei caracteristicii numerice deosebim *variabile continue* şi *variabile discrete*:

- ***variabilele continue*** sun acele variabile cantitative, care pot lua orice valoare din domeniul lor de variaţie (exemple: *înălţimea, greutatea, cifra de afaceri* etc.);

- ***variabilele discrete*** sunt acele variabile cantitative, care nu pot lua decât anumite valori din domeniul lor de variaţie, de regulă − numere întregi (exemple: *numărul de membri ai familiei, numărul de copii din familie, numărul de localități din raion* etc.).

Cea din urmă clasificare se mai poate completa prin următoarea explicaţie: *datele discrete* sunt răspunsuri numerice care apar în urma unui proces de numărare, în timp ce *datele continue* sunt răspunsuri numerice care apar în urma unui proces de măsurare.

Vom analiza, în continuare, **tipurile variabilelor calitative şi cantitative**.

Def. 3.2. Variabilele calitative, care pot lua un număr finit de valori neordonate sau variabile ce permit doar clasificarea observaţiilor, se numesc ***variabile nominale***.

Observăm că nivelul de măsurare a acestor variabile este cel nominal, iar scala – nominală.

În calitate de exemple de variabile de acest tip pot fi aduse *sexul, profesia, culoarea ochilor* individului etc.

În vederea prelucrării, valorile variabilelor nominale se codifică, de regulă, cu numere naturale. În acest caz, nivelul de măsurare (tipul variabilei) nu se modifică prin utilizarea unei astfel de codificări.

Def. 3.3. Variabilele calitative, ale căror valori sunt ordonate, dar nu este definită (sau nu se poate defini) distanţa dintre oricare două valori, se numesc ***variabile ordinale***.

Nivelul de măsurare al a acestor variabile este cel ordinal, iar scalele de măsură – ordinale.

Exemple de variabile de acest tip pot fi: *aprecierea de către individ a unui film* (cu valorile: „foarte bun”, „bun”, „rău”, „foarte rău”, „nu l-am privit”), *opinia individului faţă de temperatura dintr-o încăpere* (valori posibile: „foarte cald”, „cald”, „normal”, „rece”, „foarte rece”, „nu pot aprecia”) etc.

La codificarea valorilor (ordonate) ale variabilelor ordinale se folosesc numai şiruri ordonate de numere naturale.

Def. 3.4. Variabilele cantitative (numerice) care utilizează o valoare 0 convenţională se numesc ***variabile de interval***.

Nivelul de măsurare a variabilelor de interval este cel de interval, iar scala – scala interval.

La compararea valorilor acestor variabile, găsim răspuns la întrebări de tipul: „Cu cât e mai mare?” sau „Cu cât e mai mică?” (de exemplu, *temperatura mediului înconjurător* măsurată în diferite zile).

Def. 3.5. Variabilele cantitative (numerice) care utilizează o valoare 0 naturală se numesc ***variabile de raport.***

Acestor variabile le corespunde nivelul proporțional de măsurare, iar scala respectivă este cea proporțională.

La compararea valorilor variabilelor de raport găsim răspuns şi la întrebări de tipul: „De câte ori e mai mare?” sau „De câte ori e mai mică?” (de exemplu, *greutatea individului* sau *înălţimea individului*). Este important a observa că valoarea 0 indică inexistenţa variabilei.

Observaţia 3.1. Valorile variabilelor numerice nu se codifică: în calitate de „cod” în baza de date se introduce chiar valoarea variabilei.

Observaţia 3.2. Variabilele de interval şi cele de raport practic nu se deosebesc în procesele de prelucrare; în continuare ele vor fi examinate împreună şi numite, pur şi simplu, *numerice* (sau *cantitative*).

Observăm, astfel, o corespondență biunivocă între tipurile de variabile și nivelurile de măsurare. Respectiv, tipurile de variabile pot fi definite prin modalitatea de măsurare: cele ce se măsoară cu scala nominală se numesc variabile nominale, cu scala ordinală – variabile ordinale etc.

Pentru continuarea expunerii materialului, vom introduce următoarele notări:

- n − numărul indivizilor cercetaţi (volumul populaţiei);
- X − o caracteristică studiată;

- x_1, x_2,…,x_m – valorile posibile ale caracteristicii X (scala de valori a caracteristicii X);
- n_1, n_2,...,n_m – numărul de indivizi corespunzător valorilor caracteristicii (sau care posedă valoarea respectivă a caracteristicii).

Atunci:

Def. 3.6. Se numeşte ***frecvenţă absolută*** a unei valori x_i a caracteristicii X numărul de unităţi ale populaţiei n_i corespunzătoare acestei valori.

Def. 3.7. Se numeşte ***frecvenţă relativă*** a unei valori x_i a caracteristicii X raportul dintre frecvenţa absolută n_i a valorii x_i şi numărul total al indivizilor n.

Def. 3.8. Frecvenţele relative, exprimate în procente, se mai numesc ***frecvenţe procentuale***. Ele se calculează după formula:

$$f_i = \frac{n_i}{n} \times 100\%$$

Def. 3.9. Se numeşte ***frecvenţă cumulată*** procentul de indivizi ce se găsesc până la sau sub o treaptă (valoare) a scalei. Ea se calculează după formula:

$$F_i = \frac{n_1 + n_2 + n_3 + ... + n_i}{n} \times 100\% = f_1 + f_2 + f_3 + ... + f_i$$

Observaţia 3.3. Frecvenţele cumulate au sens numai pentru variabilele ordinale şi cele cantitative (numerice).

Def. 3.10. Un tabel de forma:

X	x_1	x_2	x_3	...	x_m
F	f_1	f_2	f_3	...	f_m

poartă denumirea de *distribuţie de frecvenţe.*

Distribuţiile de frecvenţe pot fi reprezentate şi grafic, sub formă de diagrame cu bare, diagrame circulare („plăcintă”, pie) etc. Vom demonstra acest lucru pentru un exemplu concret de distribuţii de frecvenţe (a se vedea Tabelul 3.1)

Tabelul 3.1.

Distribuţia a 20 de figuri geometrice de patru forme diferite

Figură	□	△	○	◇
Frecvenţe absolute	8	4	6	2
Frecvenţe relative[*]	40%	20%	30%	10%

Diagramele corespunzătoare acestei distribuţii de frecvenţe sunt prezentate în Figurile 3.1 şi 3.2.

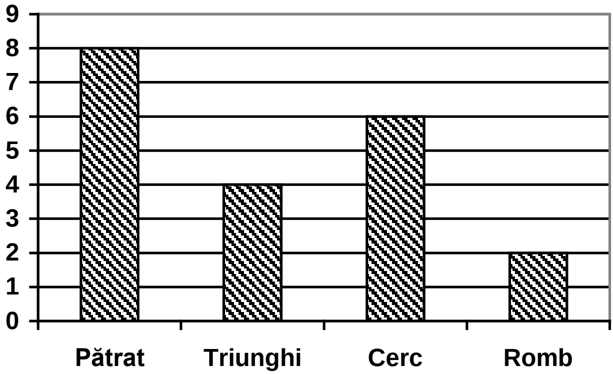


Fig. 3.1. Reprezentarea distribuţiei de frecvenţe sub formă de diagramă cu bare

[*] În cercetările sociologice frecvenţele procentuale se calculează şi se prezintă pentru populaţii cu volumul ce depăşeşte 100 de indivizi.

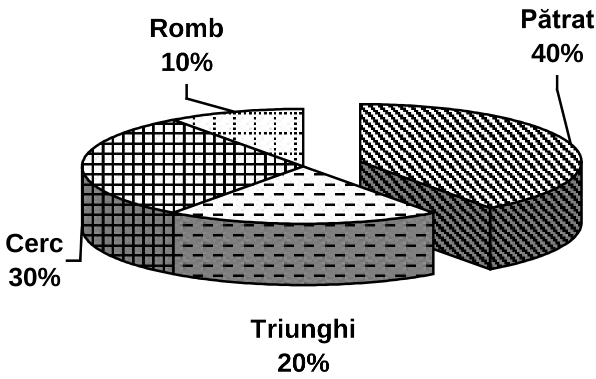


Fig. 3.2. Reprezentarea distribuţiei de frecvenţe sub formă de diagramă circulară

Astfel, distribuția de frecvențe poate fi privită ca o proprietate a întregii populații, determinată de variabila pentru care a fost construită. Însă, având două populații de aceeași natură, pentru care au fost construite distribuțiile de frecvențe ale uneia și aceleași variabile, va face dificilă compararea acestora. Pentru a compara totuși populațiile, ne vin în ajutor așa-numiții indicatori statistici. În continuare, vom examina această noțiune.

Def. 3.11. Vom numi ***indicator statistic*** o mărime cantitativă sau calitativă, care descrie populaţia în întregime, determinată sau calculată din valorile variabilelor statistice.

Astfel, indicatorul statistic poate fi considerat drept o caracteristică a întregii populaţii.

Condiţiile ce trebuie îndeplinite de către indicatorii statistici:

- să fie definiţi în mod obiectiv, independent de dorinţa utilizatorului;
- să depindă de toate valorile individuale înregistrate ale caracteristicilor (variabilelor);
- să aibă o semnificaţie concretă, uşor de înţeles chiar şi de nespecialişti;

- să fie simplu şi uşor de calculat;
- să fie puţin sensibil la fluctuaţiile de selecţie a eşantioanelor.

Se deosebesc două tipuri de indicatori statistici: ai tendinței centrale și de dispersie (împrăștiere).

Def. 3.12. Indicatorul tendinţei centrale este un indicator ce caracterizează valoarea medie a unei variabile din populaţie sau valoarea ei cea mai frecventă întâlnită în populaţie (modul, mediană, medie – exemple de astfel de indicatori).

Def. 3.13. Indicatorul de dispersie este un indicator ce caracterizează împrăştierea valorilor variabilei faţă de valoarea medie sau uniformitatea/neuniformitatea distribuţiei valorilor acesteia (de exemplu: IVC, amplitudine, dispersie, abatere standard etc.).

Un exemplu simplu: dacă în calitate de variabilă se ia nota la un examen a studenţilor dintr-o grupă academică, atunci unul din indicatorii statistici ar putea fi nota medie a grupei la acest examen, altul – diferența dintre nota maximală și cea minimală primite la examen de către studenți, diferență ce caracterizează împrăștierea notelor (amplitudinea).

În calitate de indicatori statistici ai **`variabilelor nominale`** se utilizează:

− indicatorul tendinţei centrale:

- *modul* (**Mo**) − categoria cu cea mai mare frecvenţă;

− indicatorul de dispersie:

- *Indicele variaţiei calitative* (**IVC**) − raportul dintre variaţia distribuţiei observate şi variaţia distribuţiei uniforme. (*Distribuția uniformă* este o așa distribuție pentru care categoriile scalei conțin unul și același număr sau același procent de indivizi, determinat, de exemplu, de raportul *n/m*.)

Remarcăm următoarea proprietate a **IVC**: cu cât valoarea lui este mai apropiată de 100%, cu atât valorile observate ale caracteristicii sunt repartizate mai uniform sau distribuția observată se apropie de cea uniformă. Valoarea IVC este egală cu 0, atunci când toți indivizii din populație au aceeași valoare a variabilei sau toți se găsesc în una și aceeași categorie a scalei.

Vom determina aceşti indicatori pentru distribuţia figurilor geometrice din exemplul anterior (a se vedea Tabelul 3.1):

- **Mo** = „Pătrat" (este categoria cu cea mai mare frecvenţă);
- pentru determinarea **IVC** se utilizează distribuţia observată {8, 4, 6, 2} şi cea uniformă {5, 5, 5, 5}, pentru care frecvențele tuturor valorilor variabilei cercetate coincid. Atunci:

$$IVC = \frac{8\cdot(4+6+2)+4\cdot(6+2)+6\cdot 2}{5\cdot(5+5+5)+5\cdot(5+5)+5\cdot 5}\cdot 100\% \approx 93.3\%$$

Indicatorii **variabilelor ordinale** sunt:

− indicatori ai tendinţei centrale:

- *modul* (**Mo**);
- *mediana* **(Me)** − valoarea din mijloc a şirului ordonat (în creştere sau descreştere) de valori ale caracteristicii. În cazul unui număr par de valori (numerice!), mediana se calculează că media aritmetică a celor două valori din mijlocul șirului ordonat.

− indicator de dispersie:

- *Indicele variaţiei calitative* (**IVC**).

Suplimentar, în cazul variabilelor ordinale se poate vorbi şi despre *forma distribuţiei de frecvenţe*, care poate fi *simetrică* sau *nesimetrică*.

Vom demonstra calcularea indicatorilor variabilelor ordinale printr-un exemplu. Fie că la întrebarea *„În ce măsură sunteţi mulţumit de calitatea deservirii la cantină?"*, având variantele de răspuns: *foarte nemulţumit* (FN), *nemulţumit* (NM), *indiferent* (I), *mulţumit* (M), *foarte mulţumit* (FM), 15 studenţi au răspuns în felul următor:

I, FN, M, I, NM, FM, I, NM, M, NM, I, FN, M I, FM

Distribuţia de frecvenţe ale acestei caracteristici şi distribuţia uniformă, necesară pentru calcularea IVC, au formele prezentate în Tabelul 3.2.

Tabelul 3.2

Nivelul de mulţumire a studenţilor faţă de deservirea în cantină

	Foarte nemulţumit	*Nemulţumit*	*Indiferent*	*Mulţumit*	*Foarte mulţumit*
Distribuţia observată	2	3	5	3	2
Distribuţia uniformă	3	3	3	3	3

Din Tabelul 3.2 determinăm modul caracteristicii studiate (varianta de răspuns care se întâlneşte cel mai des, de 5 ori):

Mo = „indiferent”

Pentru determinarea medianei, aranjăm cele 15 variante de răspuns în ordine crescătoare (de la *foarte nemulţumit*, la *foarte mulţumit*):

FN, FN, NM, NM, NM, I, I, **I**, I, I, M, M, M, FM, FM

Valoarea caracteristicii din mijlocul acestui şir (ea este evidenţiată), conform definiţiei, este mediana:

Me = „indiferent”

Indicele variaţiei calitative se calculează în felul următor (folosim datele din Tabelul 3.2):

$$IVC = \frac{2\cdot(3+5+3+2)+3\cdot(5+3+2)+5\cdot(3+2)+3\cdot 2}{3\cdot(3+3+3+3)+3\cdot(3+3+3)+3\cdot(3+3)+3\cdot 3}\cdot 100\% \approx 96.7\%$$

Distribuţia de frecvenţe studiată este simetrică. Acest lucru se observă atât din Tabelul 3.2, cât şi din diagrama prezentată *infra* (a se vedea Figura 3.3).

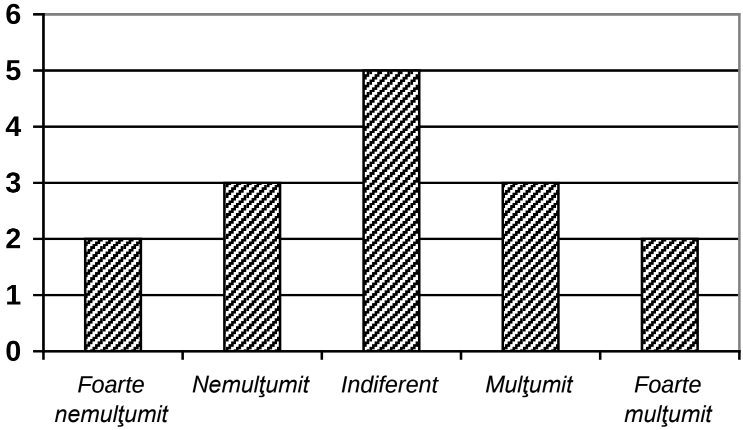


Fig. 3.3. Nivelul de mulţumire faţă de deservirea la cantină

În sfârşit, vom analiza indicatorii statistici ai **variabilelor numerice:**

− Indicatorii tendinţei centrale pentru acest tip de variabile sunt:

- **Mo** − *modul* (el se determină în cazul variabilelor discrete, dacă numărul valorilor observate depăşeşte cu mult numărul categoriilor din scala de valori ale variabilei, sau după transformarea variabilei numerice continue într-o variabilă ordinală);
- **Me** − *mediana* (se determină conform definiţiei pentru un număr impar de valori ale caracteristicii; în cazul unui număr par de valori − ca medie aritmetică a celor două valori situate în mijlocul şirului ordonat de valori ale caracteristicii);
- **M** − *media* − se determină ca media aritmetică a valorilor caracteristicii:

$$M = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

– printre indicatorii de dispersie menţionăm:

- **A** – *amplitudinea,* definită ca diferenţa dintre valorile maximală şi minimală observate ale caracteristicii:

$$A = x_{\max} - x_{\min}$$

- **σ** – *abaterea standard*, calculată după formula:

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

Observaţia 3.4. De rând cu abaterea standard în calitate de indicator al împrăștierii se folosește și $\boldsymbol{\sigma^2}$ – *dispersia***.**

Observaţia 3.5. Cu cât sunt mai mari valorile indicatorilor **A** şi **σ,** cu atât sunt mai împrăştiate valorile observate ale caracteristicii. Menţionăm însă că nivelul de împrăştiere a valorilor caracteristicii e descris cu mult mai bine de abaterea standard (sau de dispersie) decât de amplitudine.

Vom finaliza expunerea temei cu un exemplu de calculare a indicatorilor variabilelor numerice. Astfel, fie date notele obţinute de 20 de studenţi la un examen:

7, 5, 7, 8, 4, 6, 8, 2, 7, 1, 8, 10, 9, 7, 9, 6, 4, 2, 3, 7

Trebuie determinaţi indicatorii statistici ai acestei caracteristici.

Începem cu construirea distribuţiei de frecvenţe (a se vedea Tabelul 3.3).

Tabelul 3.3

Distribuţia notelor de la examen

Nota	1	2	3	4	5	6	7	8	9	10
Frecvenţa absolută	1	2	1	2	1	2	5	3	2	1

În primul rând, observăm că cea mai frecvent întâlnită este nota 7. Astfel:

$$\mathbf{Mo} = 7$$

Notele aranjate în creștere formează următorul șir:

1, 2, 2, 3, 4, 4, 5, 6, 6, **7, 7**, 7, 7, 7, 8, 8, 8, 9, 9, 10

Valorile din mijlocul acestui şir (ele sunt evidenţiate) permit a calcula mediana:

$$\mathbf{Me} = (7+7)/2 = 7$$

Media se calculează ca medie aritmetică a tuturor notelor obţinute de studenţi la examen:

$$\mathbf{M} = (7+5+7+8+4+6+8+2+7+1+8+10+9+7+9+6+4+2+3+7)/20 = 6$$

Amplitudinea:

$$\mathbf{A} = 10 - 1 = 9$$

Abaterea standard se calculează după formula adusă *supra*:

$$\sigma = \sqrt{63} \approx 8$$

Recapitulăm cele expuse cu Tabelul 3.4, în care sunt prezentate tipurile de variabile şi indicatorii ce pot fi determinaţi pentru acestea:

Tabelul 3.4

Indicatorii statistici ai diferitelor tipuri de variabile

Tip de variabilă	*Indicatori ai tendinţei centrale*			*Indicatori ai împrăştierii*		
	Mo	**Me**	**M**	**IVC**	**A**	**σ**
Nominală	**X**			**X**		
Ordinală	**X**	**X**		**X**		
Numerică	**X**	**X**	**X**		**X**	**X**

Exerciții, întrebări de control

1. Propuneți un exemplu de populație. Pentru această populație formulați câte trei exemple de:

a) variabile nominale;
b) variabile ordinale;
c) variabile numerice;
d) variabile dihotomice;
e) variabile discrete;
f) variabile continue.

2. Construiţi distribuţiile de frecvenţe absolute, relative, valide şi cumulate pentru datele despre numărul de copii din familiile unui bloc de locuit din Chişinău (n/s − nu se ştie):

   0 2 1 0 2 2 0 1 1 0 3 2 0 n/s 1 2 0 4 1 0 2 2 0 1 1 2 2 3 n/s 1 2 1 0 0 1 0 1 2 4 2 1 1 1 0 2 0 2 2 4 1 0 1 n/s 0 1 0 0 1 2 4 2 n/s 1 2 0 1 1 2 2 1 0 n/s 2 2 3 0 1 0 2 1

3. Să se compare trei parcuri după distribuţia speciilor de copaci (U – ulm, P – pin, S – stejar, B – brad, A – arţar):

   Parcul 1: P P S B B B P P A P P S S A A B B B P P

   Parcul 2: U U U P S B A B S P P P P P U U S P P U

   Parcul 3: U U P P P S B A P P P S U U P U U U P P

4. Să se compare următoarele sectoare ale mun. Chişinău după nivelul de poluare a aerului (F – foarte înalt, I – înalt, M – mediu, J – jos, N – practic nu e poluat), conform opiniilor a 20 de experţi:

   Ciocana: I I M M F F J I I I M M M I M M I I I F

   Botanica: F I F I M J M M F F I I I M F F I I I F

   Centru: F N M M M I I I I F F F I M M I F F I I

5. Să se compare următoarele trei grupe după notele obţinute la examen:

   Grupa 1: 8 8 6 6 6 5 7 5 7 5 7 7 5 5 5 8 5 5 7 7

   Grupa 2: 5 9 10 9 10 10 10 9 8 7 8 7 8 9 9 9 10 10 9 9

   Grupa 3: 6 7 6 7 8 9 6 7 6 7 8 10 6 7 6 7 7 7 6 7

6. Ca rezultat al cercetării unei păduri după soiurile de copaci ce cresc în ea, a fost determinat indicatorul statistic *modul*, care s-a dovedit a fi **pinul**. Aceasta înseamnă că (alegeți variantele corecte):

   a) În pădure cel mai des se întâlneşte pinul.

   b) În pădure cresc numai pini.

   c) În medie în pădure se întâlneşte soiul de pin.

d) Cei mai înalţi copaci din pădure sunt pinii.
e) În pădure este necesar a mări numărul pinilor.
f) În pădure pinii sunt repartizaţi uniform.

7. În urma cercetării unei păduri după soiurile de copaci ce cresc în ea, a fost determinat indicatorul statistic *Indicele variaţiei calitative (IVC)*, care s-a dovedit a fi egal cu 98,5%. Care din următoarele afirmaţii este falsă:
   a) În pădure este acelaşi număr de copaci de fiecare soi.
   b) Distribuţia soiurilor de copaci este aproape uniformă.
   c) Distribuţia soiurilor de copaci este neuniformă.
   d) În pădure cresc mai multe soiuri de copaci, unul din care se întâlnește cel mai frecvent.
8. După susţinerea unui examen de către o grupă de studenţi, a fost determinată nota medie a grupei – 8. Ţinând cont de faptul că studenţii au fost notaţi în sistemul zecimal (note de la 1 la 10), determinaţi dacă următoarele afirmaţii pot fi adevărate:
   a) Grupa este compusă din 8 studenţi care au susţinut examenul cu note diferite.
   b) Toţi studenţii au primit note de 8.
   c) Niciun student n-a primit nota 8.
   d) Nota 8 a fost cea mai frecventă în grupă.
   e) Atât nota maximală, cât şi cea minimală în grupă a fost nota 8.
   f) Un student a primit nota 10, unul – nota 6, iar restul studenţilor − note de 8.

Tema 4

Ancheta sociologică − furnizor de date statistice

Cercetarea sociologică, în sens larg, semnifică obţinerea şi prelucrarea informaţiilor obiectiv verificate, în vederea construirii explicaţiilor ştiinţifice ale faptelor, fenomenelor, proceselor sociale. Cercetarea sociologică se efectuează prin diferite metode şi tehnici, pentru fiecare din ele utilizându-se instrumente corespunzătoare de culegere şi înregistrare a datelor.

Def. 4.1. ***Metoda*** reprezintă o modalitate generală, strategică de abordare, studiere a realităţii.

Def. 4.2. ***Tehnicile*** sunt forme concrete pe care le îmbracă metodele (există posibilitatea ca una şi aceeaşi metodă să se realizeze cu tehnici diferite).

Def. 4.3. ***Instrumentul*** reprezintă mijlocul cu ajutorul căruia se realizează „captarea” informaţiei ştiinţifice, a datelor, este cel care se interpune între cercetător şi realitatea studiată.

Metodele de cercetare sociologică se clasifică în *metode cantitative* şi *metode calitative*.

Def. 4.4. ***Metodele cantitative*** sunt cele mai frecvente şi cele mai cunoscute modalităţi de obţinere a unor volume mari de date din mediul social pentru o ulterioară prelucrare şi analiză statistică.

Metoda cantitativă de bază este *ancheta sociologică,* iar instrumentul principal − *chestionarul*. Toate instrumentele de studiu sunt administrate şi aplicate on-line, iar culegerea şi centralizarea datelor se face automat şi securizat.

Sondajul, recensământul reprezintă tehnici ale anchetei.

Def. 4.5. ***Metodele calitative*** sunt folosite pentru a obţine date mai bogate în conţinut şi mai în profunzime.

Menționăm că cercetarea calitativă nu se bazează pe măsurări numerice, urmărind descrierea comprehensivă a unui eveniment sau a unei unități sociale.

Cercetarea calitativă dispune de metode, tehnici şi instrumente de studiu, adaptate la specificul problemei studiate.

Metodele calitative pot fi clasificate după cum urmează:

- *experimentul* − provocarea variaţiei unuia sau mai multor fenomene într-o situaţie controlată pentru determinarea legăturilor cauzale, confirmarea sau respingerea ipotezelor de cercetare;
- *observaţia* − culegerea on-line a informaţiei despre evenimente, fenomene, obiecte, persoane etc.;
- *analiza documentelor* − culegerea informaţiei despre evenimente, fenomene trecute, despre urmările lor;
- *interviul* (individual sau de grup) – discuţie ce presupune folosirea unui ghid de interviu, nestructurat sau semistructurat, aplicat indivizilor, cu posibilităţi de manevrare.

În continuare, vor fi examinate în exclusivitate metodele cantitative, reprezentanta principală a cărora este *ancheta sociologică.*

Particularităţile specifice ale anchetei sociologice sunt următoarele:

1. Tehnicile de realizare a anchetei au un evident caracter standardizat (nu se permit abateri de la schema de realizare stabilită anterior).
2. Ancheta foloseşte, prin definiţie, un chestionar în calitate de instrument de cercetare.
3. Ancheta urmăreşte să satisfacă cerinţa de reprezentativitate a eşantionului în raport cu populaţia incomparabil mai mare (în sens statistic).

4. Pentru asigurarea reprezentativităţii, ancheta se realizează pe eşantioane mari.

5. Ancheta urmăreşte colectarea unor informaţii relativ simple (datorită numărului mare de indivizi cercetat).

6. Prelucrarea datelor culese prin metoda anchetei presupune folosirea procedurilor statistice standard.

7. Ancheta, prin definiţie, se realizează culegând informaţii de la persoane în mod individual (spre deosebire de interviu, care poate fi şi de grup).

8. Ancheta se realizează, de regulă, cu personal auxiliar (operatori de anchetă sau interviu), nu numaidecât calificat, dar instruit respectiv.

Etapele unei cercetări sociologice prin metoda anchetei pot fi divizate în trei grupe:

I) ***etape de pregătire:***

- Formularea temei, determinarea scopului şi obiectivelor cercetării.
- Construirea eşantionului sau determinarea populaţiei spre a fi cercetată.
- Evaluarea costurilor fiecărei operaţii, elaborarea bugetului.
- Elaborarea instrumentelor (chestionar, fişă de observaţie etc.).
- Realizarea cercetării-pilot şi definitivarea instrumentelor (după necesitate).
- Stabilirea calendarului şi asigurarea măsurilor de respectare a lui.
- Asigurarea tuturor mijloacelor şi instrumentelor necesare pentru deplasarea pe teren.
- Rezolvarea problemelor, pe care le-ar putea întâmpina operatorii pe teren.

- Asigurarea condiţiilor de cazare, masă şi transport pentru operatori, personal.
- Stabilirea modului şi mijloacelor de verificare şi control al lucrului operatorilor în teren.
- Selectarea şi instruirea operatorilor.

II) ***lucrul în teren:***

- Intervievarea (completarea chestionarelor).
- Verificarea chestionarelor (de către operatori şi şefii de echipe).

III) ***etape finale*** (presupun utilizarea unor aplicaţii statistice pe calculator, cum ar fi, de exemplu, programul SPSS):

- Elaborarea structurii bazei de date.
- Codificarea răspunsurilor (după necesitate).
- Introducerea datelor.
- Verificarea datelor şi corectarea greşelilor de introducere.
- Prelucrarea primară a datelor şi analiza preliminară a rezultatelor.
- Introducerea de corecţii (după necesitate).
- Elaborarea rezultatelor şi a raportului final.
- Prezentarea rezultatelor cercetării.

Def. 4.6. Principalul instrument de culegere a datelor prin metoda anchetei sociologice este *chestionarul*.

Construirea unui chestionar nu este un proces chiar atât de simplu pe cât se crede de obicei. Adecvarea lui la tema de cercetat presupune, în primul rând, **operaționalizarea** obiectului de studiu, adică găsirea unor **indicatori** pertinenți pentru ceea ce vrem să măsurăm. Opiniile indivizilor despre un anumit fapt nu pot fi măsurate unidimensional decât simplificând la extrem ceva care este prin natura lui multidimensional. A operaționaliza înseamnă a împărți obiectul de cercetat pe dimensiunile și subdimensiunile care îl caracterizează (adică a-l defini), apoi a selecta dintre acestea pe cele pe care le considerăm cele mai relevante pentru ceea ce vrem să cunoaştem și, într-un ultim stadiu, să construim indicatorii care să estimeze cât mai exact cu putință dimensiunile obiectului de cercetat. Aceşti indicatori sunt reprezentați în chestionar prin întrebări.

Formularea întrebărilor se recomandă să respecte un șir de condiții:

- Fiecare întrebare trebuie să fie logică și individuală.
- În întrebări este interzisă utilizarea cuvintelor rar întâlnite, neînțelese și a termenilor tehnici/speciali.
- Întrebările trebuie să fie cât mai scurte.
- Dacă este necesar, întrebarea poate fi însoțită de o explicație, însă formularea ei trebuie să fie concisă/laconică.
- Întrebările trebuie să fie concrete, nu abstracte.
- Întrebările nu trebuie să conțină un indiciu. În cazul în care indiciul face referire la răspunsurile posibile, lista acestora trebuie să fie completă.
- Modul de formulare a întrebării trebuie să evite obținerea de răspunsuri stereotipe.
- Întrebarea nu trebuie să oblige respondenții la răspunsuri inacceptabile pentru ei.
- Limbajul întrebărilor nu trebuie să provoace dezgust (de exemplu, să fie prea expresiv).
- Nu se admit întrebări sugestive, care ar inspira răspunsul.

Ordinea întrebărilor în chestionar, tipul întrebărilor (daca sunt cu răspuns deschis sau cu răspunsuri prestabilite), forma grafică a chestionarului influențează semnificativ răspunsurile obținute de la cei chestionați și de aceea construirea chestionarului trebuie realizată cu foarte mare grijă, respectând o serie întreaga de reguli.

În plus, orice chestionar, pentru a deveni un instrument valid de măsurare, trebuie în prealabil pretestat, deși în practica curentă se trece adesea peste această etapă, îndeosebi din lipsă de timp și pentru că se folosesc întrebări considerate standard.

Numărul de întrebări din chestionar depinde de problema cercetată şi poate fi determinat, de exemplu, în modul următor:

- problema principală A se descompune în k dimensiuni A_1, A_2,...,A_k, care generează în medie câte m întrebări;

- se introduc r factori complecşi B, C, D,..., necesari pentru explicarea lui A, sau care prezintă un alt interes pentru cercetare, fiecare având s dimensiuni a câte t indicatori;

- se adaugă v întrebări de identificare (variabile personale): sex, vârstă, naţionalitate, ocupaţie, zonă de reşedinţă etc.

Astfel, numărul de întrebări $n = k \times m + r \times s \times t + v$.

De exemplu: pentru $k \approx m \approx r \approx s \approx t \approx 5$ şi $v \approx 10$ primim $n \approx 160$ (cantitate normală pentru o cercetare serioasă).

Tipurile de întrebări folosite în chestionar se clasifică în funcţie de conţinut, de înregistrare a răspunsurilor şi de numărul de variabile pe care le generează.

Tipuri de întrebări în funcţie de conţinut:

- factuale (elemente de comportament al indivizilor, calităţile lor fizice, situații obiective și verificabile prin alte mijloace etc.);

- de cunoştinţe (despre cunoştinţele indivizilor cu privire la ceva sau cineva; astfel de întrebări nu se folosesc pentru obținerea de informații, dar pentru a caracteriza persoanele intervievate, pot fi utilizate și în calitate de întrebări de control etc.);

- de opinie (vizează aspecte ce ţin de universul interior al individului: păreri, atitudini, opinii, aşteptări, evaluări, ataşamente, explicaţii, justificări, motivaţii etc.);

- întrebări-filtru (pentru trecerea, condiționată de varianta de răspuns, la unele sau altele întrebări din chestionar).

Tipuri de întrebări în funcţie de înregistrarea răspunsurilor:

- închise: ele oferă toate variantele posibile de răspuns, dintre care individul întrebat le alege pe cele potrivite;

- semideschise: aceste întrebări se aseamănă cu întrebările închise, având o variantă de răspuns de tipul *altceva, alta* etc. care acoperă toate variantele de răspuns posibile;

- deschise: ele nu conţin variante de răspuns, răspunsul fiind lăsat la discreţia individului. Prelucrarea acestor întrebări (gruparea şi codificarea răspunsurilor) se face de către cercetător după completarea tuturor chestionarelor.

Tipuri de întrebări după numărul de răspunsuri solicitate *(respectiv − după numărul de variabile din baza de date pe care le generează):*

- întrebări cu o singură variantă de răspuns (generează o singură variabilă în baza de date);

- întrebări cu un număr specificat de variante de răspuns (generează numărul respectiv, specificat, de variabile);

- întrebări cu orice număr de variante de răspuns (generează atâtea variabile, câte variante de răspuns se propun).

Structura chestionarului respectă următoarele reguli:

- la începutul chestionarului se formulează întrebări ce favorizează comunicarea şi stimulează cooperarea individului;

- nu se recomandă aşezarea întrebărilor într-o formă logică, unde următoarele întrebări sunt o consecinţă a precedentelor (individul, astfel, este direcţionat spre un răspuns aşteptat ce nu prezintă opinia lui);

- întrebările factuale (sociodemografice) se aşază, de regulă, la sfârşit;

- pentru verificarea sincerităţii sau acurateţei răspunsurilor, pentru depistarea fraudelor, unele întrebări pot să se repete într-o formulare schimbată (ele se mai numesc întrebări de control).

În sfârşit, este important *designul chestionarului,* care se supune următoarelor reguli:

- întrebările şi variantele de răspuns se situează pe aceeaşi pagină a chestionarului;

- enunţul întrebărilor este evidenţiat (sau cu litere grase, sau cu litere mai mari decât cele din variantele de răspuns);
- întrebările din chestionar se numerotează, această numerotare putând conţine şi litere, în felul acesta evidenţiindu-se diferite compartimente ale chestionarului (de exemplu, A1, A2,…, B1, B2,…, D1, D2,… etc.), dar și permițând posibilitatea de a facilita denumirea variabilelor din baza de date (acest lucru va fi discutat în una din temele următoare);
- variantele de răspuns se numerotează cu cifre arabe, ele fiind şi codurile răspunsurilor (astfel, devine comodă introducerea de către operatori a datelor în calculator, fără o codificare suplimentară a răspunsurilor);
- este de dorit ca variantele de răspuns să se situeze într-o singură coloană, astfel facilitând atât completarea chestionarului de către respondenţi, cât şi introducerea datelor.

Așadar, cercetarea sociologică prin metoda anchetei, având chestionarul în calitate de instrument de culegere a datelor, devine o sursă de date ce pot fi organizate sub formă de tabel (bază de date). Datele, astfel organizate, în continuare se prelucrează la calculator prin utilizarea diferitelor programe de prelucrare statistică. Un reprezentant al acestora, utilizat pe larg în lume, este programul SPSS, care va fi analizat pe larg în temele următoare.

Exerciții, întrebări de control

1. Numiți și argumentați trei deosebiri esențiale dintre interviu și ancheta sociologică.
2. Care sunt asemănările dintre interviu și ancheta sociologică? Dar dintre ancheta sociologică și metoda observației?
3. Există asemănări între analiza de conținut și ancheta sociologică? Argumentați răspunsul.
4. Formulați scopul și obiectivele sondajului sociologic „Studentul USM”. Să se planifice cercetarea următoarelor aspecte: studiile, timpul liber, condițiile de alimentare și de trai, în general, și în funcție de facultate, an de studii și sex.

5. Elaborați, în funcție de obiectivele formulate în pct.4, chestionarul cercetării, care să conțină cel puțin câte trei întrebări deschise, semideschise, închise, factuale, de cunoștințe, de opinie și de control.
6. Pentru cercetarea din pct.4, formulați câte trei întrebări cu o variantă de răspuns, cu trei variante de răspuns și cu orice număr de variante de răspuns.
7. Indicaţi care din următoarele obiecte nu reprezintă instrument de culegere a datelor în cercetarea sociologică prin metoda anchetei:
   a) termometrul;
   b) ceasornicul;
   c) pixul;
   d) chestionarul;
   e) cântarul;
   f) ruleta.

Tema 5

Sondajul statistic.
Volumul eșantionului

După cum s-a menționat în Tema 1, ***eșantion*** se numește acea parte a populației asupra căreia se efectuează un studiu statistic (sau subset de elemente selectate dintr-o colectivitate statistică). Prin ***reprezentativitate*** (a eșantionului) se înțelege proprietatea eșantionului de a reprezenta fidel populația.

Def. 5.1. Cercetarea al cărei scop este ca, pe baza rezultatelor prelucrării datelor obținute pe eșantion, să se estimeze, folosind principiile teoriei probabilităților, parametrii corespunzători ai colectivității totale, poartă denumirea de ***sondaj statistic***. Sondajul statistic reprezintă o tehnică a metodei anchetei sociologice, el realizându-se prin parcurgerea tuturor etapelor cercetării sociologice prin metoda anchetei.

Cercetarea prin sondaj se desfășoară în două faze:

- la prima fază se culeg și se prelucrează date statistice de la unitățile colectivității generale, incluse în eșantion, din care rezultă indicatori derivați care descriu statistic eșantionul folosit (etapa descriptivă);
- la a doua fază indicatorii obținuți prin prelucrarea datelor din eșantion se extind, cu o anumită probabilitate, asupra întregii colectivități în scopul caracterizării acesteia din punct de vedere statistic (etapa inferențială).

Originile sondajului sunt legate de psihologul și sociologul american George Gallup. Acesta și-a susținut, în 1928, teza de doctor prezentând „O metodă obiectivă pentru determinarea intereselor cititorilor față de textele unui ziar”. Ideile susținute în această teză au fost puse în practică cu ocazia alegerilor generale din 1934 (ideea studierii opiniilor pe grupuri reprezentative prin intermediul chestionării directe a publicului). Tot el a înființat, în 1935, și primul institut de studiere a opiniei publice, care îi poartă numele, și care astăzi este cel mai cunoscut și mai prestigios institut de gen din lume. De fapt, prestigiul acestui institut vine încă din 1936, când a prevăzut victoria în alegeri a lui Franklin D. Roosevelt contrar opiniei observatorilor politici.

Dintre avantajele pe care le prezintă cercetarea prin sondaj, le enunțăm pe cele mai semnificative:

- când colectivitatea totală este foarte mare, cercetarea ei exhaustivă necesită un volum mare de cheltuieli materiale şi umane, deci este avantajos să se recurgă la sondaj, care este mai operativ şi mai ieftin;
- partea supusă înregistrării fiind mult mai mică decât cea totală, erorile de înregistrare sunt mai puţin numeroase şi mai uşor de înlăturat în faza de verificare a datelor;
- cercetarea prin sondaj este singura posibilă, atunci când prin cercetarea exhaustivă s-ar ajunge la distrugerea produselor (de exemplu, controlul calităţii unui produs);
- sondajul permite verificarea programului unei observări totale şi a ipotezelor statistice.

Cei care apelează la sondaj ca metodă de culegere de date primare trebuie să fie conştienţi de dezavantajele sale:

- cercetarea se bazează doar pe declaraţiile respondenţilor, ceea ce poate genera o serie de erori sistematice;
- respondenţii pot să denatureze, în mod inconştient sau deliberat, informaţiile ce descriu realitatea;
- pot surveni o serie de erori sistematice în ce priveşte, de exemplu: eşantionarea, formularea întrebărilor, culegerea datelor de către operatori, prelucrarea datelor şi analiza informaţiilor.

În esenţă, sondajul este o metodă de culegere de date primare, pe baza unui chestionar administrat unui eşantion reprezentativ de respondenţi. Includerea sondajului în categoria metodelor de obţinere a datelor primare se întemeiază pe faptul că permite culegerea de date în mod special pentru abordarea unei anumite probleme decizionale, a unui anumit proiect de cercetare.

Calitatea rezultatelor sondajelor statistice și posibilitatea de a le generaliza (cu o exactitate și probabilitate oarecare) pentru întreaga populație depind într-o foarte mare măsură atât de volumul eșantionului, cât și de reprezentativitatea lui. În continuare, vom examina aceste noțiuni.

După cum s-a menționat anterior, prin reprezentativitatea eșantionului se înțelege proprietatea lui de a reprezenta fidel populația. Cu alte cuvinte, eșantionul reprezentativ trebuie sa respecte întocmai structura populației cercetate, dacă, desigur, această structură este cunoscută. De exemplu, dacă se cunoaște repartizarea populației din țară după mediul de reședință, atunci același raport rural/urban trebuie sa-l respecte și eșantionul unui sondaj național.

Volumul eșantionului reprezentativ poate fi determinat prin câteva formule (W.G. Cochran, Taro Yamane, P. Mureșan). Una din variantele utilizate frecvent, dedusă de P. Mureșan, ține cont de volumul populației și are forma:

$$n = \frac{t^2 \times p \times (1-p)}{\frac{N-1}{N} \times d^2 + \frac{t^2 \times p \times (1-p)}{N}} \;, \qquad (5.1)$$

în care:

n – volumul eșantionului;

N – volumul populației din care s-a extras eșantionul;

d – marja de eroare sau eroare maximă (în %);

t – parametru, ce depinde de probabilitatea de estimare sau nivelul de încredere *P* a rezultatelor pentru întreaga populație (t = 1,96 pentru P = 95%, t = 2,33 pentru P = 99% etc.);

p – incidența fenomenului cercetat (probabilitatea, că fenomenul va avea loc) și, respectiv, *(1 – p)* – probabilitatea lipsei fenomenului ($0 \le p \le 100\%$).

Deoarece estimarea lui p este dificilă de realizat, în formula (5.1) se folosește valoarea maximală a produsului $p(1 - p)$, care se obține pentru p = 50%. Astfel formula respectivă, pentru populații mici, capătă forma:

$$n = \frac{t^2 \times 50 \times 50}{\frac{N-1}{N} \times d^2 + \frac{t^2 \times 50 \times 50}{N}} \tag{5.2}$$

Pentru populații mari ($N \to \infty$), formula (5.1) se simplifică, nu depinde de volumul populației N, și primește forma propusă de W.G. Cochran:

$$n = \frac{t^2 \times 50 \times 50}{d^2} \ . \tag{5.3}$$

Formula lui Taro Yamane nu depinde nici de incidența fenomenului cercetat, nici de probabilitatea de estimare (din start ea se presupune egală cu 95%), ci numai de volumul populației:

$$n = \frac{N}{1 + N * \Delta^2}, \tag{5.4}$$

unde marja de eroare Δ este exprimată în părți ale unității ($\Delta = d/100$). Observăm că această formulă este o consecință a formulei lui P. Mureșan pentru P = 95% ($t = 1{,}96 \approx 2$) și $d = 100 \cdot \Delta$.

În calitate de exemplu, vom calcula volumele eșantioanelor reprezentative cu marja de eroare d = 3% și probabilitatea de estimare P = 95% (t = 1,96) pentru diferite mărimi ale populațiilor, folosind formulele (5.2), (5.3) și (5.4) (a se vedea Tabelul 5.1). Observăm că pentru populații mai mari de 500.000 de indivizi volumul eșantionului reprezentativ practic nu depinde de mărimea populației și poate fi calculat cu un grad mare de aproximație după una din formulele (5.3) sau (5.4) (volumul eșantionului calculat după formula lui Taro Yamane este puțin mai mare din cauza aproximării lui t cu 2).

Mărimea marjei de eroare a sondajului, pentru populații mari, se obține din formula (5.3), dacă se cunoaște volumul eșantionului reprezentativ:

$$d = \frac{50t}{\sqrt{n}}. \tag{5.5}$$

Tabelul 5.1

Volumul eșantionului reprezentativ în funcție de volumul populației și formula de calcul

Volum populație (*N*)	**Volum eșantion (*n*), calculat după formula 5.2 (P.Mureșan)**	**Volum eșantion (*n*), calculat după formula 5.3 (W.G.Cochran)**	**Volum eșantion (*n*), calculat după formula (5.4) (Taro Yamane)**
1	1	–	1
10	10	–	10
100	92	–	92
1.000	516	–	526
5000	880	1.067	909
10.000	964	1.067	1.000
50.000	1.045	1.067	1.087
100.000	1.056	1.067	1.099
500.000	1.065	1.067	1.109
1.000.000	1.066	1.067	1.110

În particular, pentru probabilitatea de estimare de 95%, care se aplică în majoritatea sondajelor sociologice, marja de eroare va fi:

$$d = \frac{50 \times 1{,}96}{\sqrt{n}} \approx \frac{100}{\sqrt{n}} \tag{5.6}$$

De exemplu, un sondaj sociologic cu probabilitatea de estimare P = 95% pe un eșantion reprezentativ de n = 1600 de indivizi are marja de eroare $d \approx 2{,}5\%$.

Din cele expuse *supra,* observăm că, vorbind despre reprezentativitatea eșantioanelor în cazul populațiilor mari apelăm permanent la două mărimi, ce-l caracterizează: marja de eroare d și probabilitatea de estimare *P*. Sensul acestora este următorul: dacă în urma unui sondaj am determinat o mărime oarecare *m* în eșantion (indicator statistic, frecvență etc.), atunci cu probabilitatea *P* valoarea respectivă μ din populație se va găsi în intervalul, determinat de relația:

$$\frac{|\mu - m|}{\mu} \cdot 100 < d \tag{5.7}$$

sau în intervalul:

$$\frac{m}{1 + \frac{d}{100}} < \mu < \frac{m}{1 - \frac{d}{100}}\ , \tag{5.8}$$

care, ținând cont de formula (5.6), se transformă în relația ce depinde numai de volumul eșantionului:

$$\frac{m}{1 + \frac{1}{\sqrt{n}}} < \mu < \frac{m}{1 - \frac{1}{\sqrt{n}}} \tag{5.9}$$

Spre exemplu, dacă drept rezultat al unui sondaj național, realizat pe un eșantion reprezentativ de 1.200 de indivizi (marja de eroare d ≈ 2,9%), s-a obținut că 35% dintre indivizi la alegerile parlamentare ar vota pentru partidul X, atunci se poate concluziona că, din toată populația, cu probabilitatea de estimare *P* = 95%, pentru partidul X ar vota între 34 și 36% de indivizi din întreaga populație.

Exerciții, întrebări de control

1. Calculați volumul eșantioanelor reprezentative pentru sondaje naționale în următoarele țări: Republica Moldova, România, Ciad, Andora, Italia, San-Marino, Monaco și Vatican. Să se ia în calcule o marjă de eroare de 3% și o probabilitate de estimare de 95%.
2. Să se compare volumele eșantioanelor reprezentative pentru un sondaj național în Republica Moldova, dacă se planifică marje de eroare de 2%, 3%, 4% sau 5% cu o probabilitate de estimare de 99%.

3. În Republica Moldova, România și Ucraina au fost realizate sondaje naționale pe eșantioane reprezentative cu volumele de 1.200, 1.400 și, respectiv, 1.600 de indivizi. Care sunt marjele de eroare ale acestor sondaje, presupunând probabilitatea de estimare egală cu 95%.
4. La ciclul Licență la USM studiază 15.000 de studenți: 7.000 – la anul I, 5.000 – la anul II și 3.000 – la anul III. Care va fi volumul eșantionului reprezentativ pe ani de studii al sondajului din USM în condițiile marjei de eroare de 3% și probabilității de estimare de 95%.
5. La una de la facultățile USM a fost realizat un sondaj pe un eșantion reprezentativ de 101 studenți. Ținând cont de faptul că la facultatea menționată studiază 800 de studenți, să se calculeze marja de eroare a sondajului cu condiția că probabilitatea de estimare este de 95%.
6. În condițiile pct.5, care ar fi trebuit să fie volumul eșantionului reprezentativ pentru ca marja de eroare să nu depășească 3%?
7. În condițiile pct.5, câți studenți trebuie să studieze la facultate pentru ca eșantionul cu volumul de 101 indivizi să fie reprezentativ pentru facultate cu marja de eroare de 3% și probabilitatea de estimare de 95%?

Tema 6

Sondajul statistic. Metode de eșantionare

Deseori întreaga populație nu poate fi cercetată din mai multe cauze: volumul mare al populației și, respectiv, cheltuielile de cercetare enorme; inaccesibilitatea unor elemente ale populației; metodele de investigație, care ar conduce la distrugerea elementelor, perioada mare de timp pentru culegerea datelor și, respectiv, „învechirea” lor până a fi prelucrate, creșterea numărului de erori etc. Soluția salvatoare este de a cerceta o parte, special selectată, a populației, numită eșantion, ca apoi rezultatele obținute să se extindă, cu o anumită probabilitate și exactitate, asupra întregii populații.

După cum a fost menționat în Tema 1, *eșantion* se numește acea parte a populației asupra căreia se efectuează un studiu statistic (sau subset de elemente selectate dintr-o colectivitate statistică). Prin *reprezentativitate* (a eșantionului) se înțelege proprietatea eșantionului de a reprezenta fidel populația. Mai mult, materialul expus în Tema 5 ne conduce la concluzia că gradul de reprezentativitate al eșantionului depinde de două mărimi: marja de eroare d și probabilitatea de estimare P. Aceste două mărimi sunt legate între ele (a se vedea, de exemplu, formula 5.3): pentru unul și același volum al eșantionului, creșterea probabilității de estimare P conduce la creșterea marjei de eroare d, și invers.

Reprezentativitatea eșantionului este o noțiune relativă și depinde de caracteristica din populație studiată. Astfel, un eșantion poate fi reprezentativ pentru o caracteristică și nereprezentativ pentru alta. Pentru una și aceeași caracteristică, un eșantion poate fi mai mult reprezentativ sau mai puțin reprezentativ. Dacă se compară două eșantioane, care au aceeași probabilitate de estimare, mai reprezentativ va fi acel eșantion pentru care marja de eroare este mai mică. Sau dacă ambele eșantioane au aceeași marjă de eroare, mai reprezentativ va fi eșantionul cu o probabilitate de estimare mai înaltă.

Def. 6.1. Procedura de construire a unui eșantion reprezentativ se numește ***eșantionare***.

Prin eșantionare putem demonstra orice, doar că cu o anumită probabilitate și exactitate. Eșantionarea răspunde la întrebările **câți?** (câți indivizi trebuie să fie selectați în eșantion) și **cum?** (cum să fie selectați indivizii, în așa fel ca eșantionul să fie cât mai reprezentativ). Rezultatele obținute în eșantion sunt utilizate pentru a deduce, estima prin inferență statistică rezultatele pe care le-am obține dacă am cerceta întreaga populație.

Se disting doua mari **modalitati de esantionare**:

- eșantionare aleatoare (probabilistică);
- esantionare nealeatoare (empirică, la întâmplare) sau pe bază de rationament.

Procedura fundamentală pentru construirea unui eșantion reprezentativ este **selecția aleatoare** (**randomizarea**). În tehnicile de randomizare toți membrii populației au aceeași șansă de a fi selecționați într-un eșantion și toate posibilele eșantioane au aceeași șansă de a fi selecționate în cercetare. Criteriul de bază este probabilismul. Eșantionarea probabilistă poate fi:

- aleatoare simplă;
- aleatoare sistematică;
- prin stratificare;
- cluster (de grup);
- multistadială (pe trepte).

În continuare, vom trece în revistă tehnicile de eșantionare probabilistice.

1. Eșantionarea aleatoare simplă. Prin această tehnică de eșantionare, fiecare individ din populație are aceeași șansă de a fi selecționat. Această tehnică se aseamănă cu extragerea numărului necesar de bile, toate identice între ele, dintr-o urnă, fiecare bilă corespunzând unui individ din populație.

Pentru construirea practică a eșantionului prin această tehnică, toți indivizii din populație se plasează într-o listă numerotată cu numere naturale de la 1 până la N (N – numărul total de indivizi din populație), după care se generează n numere aleatoare (n – volumul eșantionului) din segmentul $[1, N]$. Indivizii din listă, corespunzători numerelor aleatoare generate, se trec în eșantion.

Pentru generarea șirurilor de numere aleatoare, pot fi folosite tabele cu numere aleatoare (ele se găsesc în literatură) sau mijloacele programului EXCEL: numere aleatoare întregi din segmentul $[1, N]$ pot fi generate cu ajutorul formulei =ROUND(RAND()*(N-1)+1;0).

Observația 6.1. În calitate de variantă alternativă procedurii expuse *supra* pot fi folosite mijloacele programului SPSS de construire a eșantioanelor probabilistice (prin comanda **Analyze** → *Complex Samples* → *Select a Sample…*).

Menționăm că extragerea dintr-o urnă a unui eșantion simplu aleator este, mai degrabă, o procedură teoretică: este greu de imaginat o urnă care să cuprindă milioane de bile, corespunzătoare populațiilor mari. De aceea, în practică sunt folosite celelalte metode de eșantionare probabilistă, care păstrează elemente ale eșantionării simple aleatoare, dar care au caracteristici specifice.

2. Eșantionarea aleatoare sistematică (cu pas). Această tehnică presupune, ca și în cazul eșantionării aleatoare simple, plasarea populației într-o listă numerotată cu numere naturale și trecerea în eșantion a indivizilor din listă, selectați cu un pas, egal cu raportul N/n dintre volumul populației N și cel al eșantionului n.

Astfel, primul individ se selectează aleator din primii N/n din lista populației. Numerele de ordine ale celorlalți indivizi se obțin prin adăugarea pasului N/n la numărul de ordine al individului precedent selectat (începând cu primul). De exemplu, pentru o populație de 9.000 de indivizi și volumul eșantionului de 300 de indivizi, mărimea pasului va fi egală cu 9.000/300=30. Selectând aleator primul individ din primii 30 din lista populației (fie acesta, de exemplu, 24), următorii indivizi trecuți în eșantion vor avea numerele de ordine 24+30=54, 54+30=84, 84+30=114 și așa mai departe.

Observația 6.2. Dacă raportul N/n este fracționar, el se rotunjește până la un număr întreg.

3. Eșantionarea prin stratificare. Tehnica poate fi aplicată dacă populația cercetată poate fi divizată în straturi sau în clase distincte, după anumite caracteristici (de exemplu, divizarea populației după sexul indivizilor o stratifică în femei și bărbați; după mediu de reședință – în rurală și urbană; după categorii de vârstă – în indivizi de 0-4 ani, 5-9 ani, 10-14 ani etc.; după nivelul de studii – fără studii; cu studii primare, cu studii gimnaziale, liceale etc.; după divizarea geografică a teritoriului, cum ar fi în cazul unei cercetări naționale din Republica Moldova, – Nord, Centru, Sud sau r-nul Briceni, r-nul Edineț, … r-nul Ștefan Vodă etc.) Având o așa stratificare a populației, din fiecare strat se vor extrage subeșantioane folosind procedeul eșantionării aleatoare simple sau procedeul eșantionării sistematice.

Pentru a generaliza cele spuse *supra* și a defini câteva noțiuni suplimentare, introducem următoarele notații:

- N – volumul populației;
- m – numărul straturilor populației, în cazul stratificării ei;
- $N_1, N_2, \ldots, N_m$ – volumele straturilor populației;
- n – volumul eșantionului extras din populație;
- $n_1, n_2, \ldots, n_m$ – volumele subeșantioanelor, extrase din fiecare strat al populației.

Def. 6.2. Dacă se respectă egalitățile:

$$n_1/N_1 = n_2/N_2 = \ldots = n_m/N_m = n/N, \tag{6.1}$$

eșantionul se numește ***proporțional***; în caz contrar, dacă:

$$n_1/N_1 \neq n_2/N_2 \neq \ldots \neq n_m/N_m \neq n/N, \tag{6.2}$$

eșantionul se numește ***neproporțional.***

Este clar că un eșantion neproporțional nu este reprezentativ, cel puțin – după caracteristica de stratificare a populației: el nu respectă condițiile (6.1) sau, cu alte cuvinte, nu respectă structura populației din care a fost extras.

Inegalitățile (6.2) pot fi transformate în egalități în felul următor:

$$k_1 \cdot n_1/N_1 = k_2 \cdot n_2/N_2 = \ldots = k_m \cdot n_m/N_m = n/N \qquad (6.3)$$

Def. 6.3. Coeficienții k_1, k_2, …, k_m din egalitățile (6.3) poartă denumirea de ***coeficienți de ponderare.***

Observăm, că valorile coeficienților de ponderare pot fi determinate după formulele:

$$k_i = \frac{n/N}{n_i/N_i} \quad (i = 1, 2, \ldots, m) \qquad (6.4)$$

sau, mai comod, după formulele:

$$k_i == \frac{N_i/N}{n_i/n} \quad (i = 1, 2, \ldots, m) \qquad (6.5)$$

Formulele (6.5) se memorizează ușor: coeficienții de ponderare se determină ca raporturile dintre partea (sau procentul) stratului din populație și partea (sau procentul) stratului din eșantion. Ei sunt mai mari ca *1*, dacă în eșantion au ajuns mai puțini indivizi decât ar fi trebuit să ajungă din stratul respectiv al populației, și mai mici ca *1* – în caz contrar.

Observația 6.3. Coeficienții de ponderare se folosesc pentru „repararea" eșantionului neproporțional, transformându-l în unul proporțional prin ponderarea bazei de date a cercetării. În SPSS această ponderare se execută prin comanda **Data →** *Weight Cases…*

O explicație populară a ponderării bazei de date ar fi că prin acest procedeu se obține „amplificarea" opiniilor indivizilor selectați în eșantion într-un număr mai mic decât cel necesar pentru respectarea reprezentativității și „diminuarea" opiniilor indivizilor selectați întru-un număr mai mare. Se poate spune că aceste ponderi reprezintă o caracteristică suplimentară, atribuită indivizilor cercetați: indivizii din același strat au una și aceeași pondere.

Vom exemplifica cele expuse *supra* printr-un caz concret. Presupunem că populația cercetată constă din 54% femei și 46% bărbați. Normal ar fi fost ca un eșantion reprezentativ format din 1.000 de indivizi să conțină 540 de femei și 460 de bărbați. Însă, ca rezultat al eșantionării a fost obținut un eșantion compus din 480 de femei și 520 de bărbați, în total – 1.000 de indivizi. Coeficienții de ponderare, care vor fi numai doi (după numărul straturilor populației) se vor calcula după formulele (6.5) în felul următor:

- pentru femei k_f = 54% / (480 /1000 · 100%) = 1,125;
- pentru bărbați k_b = 46% / (520 /1000 · 100%) = 0,8846.

4. Eșantionarea cluster (de grup). Atunci când o anumita populație se compune din mai multe grupuri eterogene (clustere), putem considera aceste grupuri ca unități de eșantionare distincte din care urmează să se constituie eșantionul. Astfel, eșantionul se constituie dintr-un număr de grupuri, și nu din indivizi extrași unul câte unul. În schimb, în cadrul grupurilor extrase aleator vor fi intervievați toți indivizii care fac parte din ele.

În calitate de exemplu poate fi adusă cercetarea opiniilor elevilor ce absolvesc liceul din Republica Moldova față de examenele de bacalaureat. În acest caz, în calitate de unitate de eșantionare poate fi luată ultima clasă de liceu. Din mulțimea claselor de absolvire de liceu din toată țara se selectează prin una din tehnicile de eșantionare aleatoare simplă sau sistematică, de exemplu, 40 de clase (pentru a asigura un volum de circa 1.000 de elevi al eșantionului cu condiția că în fiecare clasă învață aproximativ 25 de elevi). În continuare, vor fi intervievați toți elevii din clasele selectate, lucru ce micșorează atât cheltuielile, cât și timpul de colectare a datelor (în fiecare clasă, de exemplu, se aplică concomitent tuturor elevilor chestionare autoadministrate sub supravegherea unui singur operator).

Alte exemple de clustere pot fi blocurile locative dintr-o localitate urbană, gospodăriile dintr-o localitate, străzile dintr-o localitate rurală etc. În toate cazurile unitățile de eșantionare se selectează aleatoriu, ca apoi să fie intervievați toți indivizii ce compun unitatea.

5. Eșantionarea multistadială (pe trepte). Dacă populația cercetată este dispersată geografic, cum ar fi, de exemplu, populația Republicii Moldova, atunci oricare din tehnicile de eșantionare descrise *supra* în cazul sondajelor naționale conduce la cheltuieli financiare destul de mari și la un timp îndelungat de colectare a datelor (în primul rând, ele țin de deplasarea în teritoriu a operatorilor). În astfel de cazuri, pentru constituirea eșantioanelor se folosește tehnica de eșantionare multistadială, care urmărește obținerea rapidă a unor date cu costuri relativ mici. Deși are o reprezentativitate mai redusă în comparație cu eșantionarea aleatorie simplă, această tehnică este intens utilizată în cercetările sociologice din rațiuni de eficiență practică și cost.

Tehnica eșantionării multistadiale presupune parcurgerea unor etape succesive, numite stadii sau trepte, și este indicată pentru populațiile care sunt organizate pe mai multe niveluri. Într-o primă etapă se aleg unitățile din primul nivel de agregare. Aceste unități se numesc unități primare și ele vor constitui baza de sondaj pentru unitățile din al doilea nivel, care se numesc secundare, ș.a.m.d. până la constituirea eșantionului. În această situație, există o dispunere în cascadă a bazelor de sondaj, deoarece unitățile alese într-o etapă formează baza de eșantionare pentru nivelul următor de eșantionare.

De exemplu, în cazul unei cercetări naționale în Republica Moldova, tehnica de eșantionare multistadială poate fi aplicată în felul următor:

- la primul nivel, se iau cele trei zone geografice ale republicii: nord, centru și sud;

- la al doilea nivel, din fiecare zonă geografică se selectează prin tehnica eșantionării simple aleatoare, de exemplu, câte 3 raioane;

- la nivelul al treilea, din fiecare raion prin tehnica eșantionării simple aleatoare se selectează, de exemplu, câte 5 localități, la care se adaugă municipiile Chișinău și Bălți (pentru asigurarea reprezentativității urban-rural);

- la nivelul al patrulea, din fiecare localitate, prin tehnica eșantionării aleatoare sistematice (de exemplu, cu folosirea listelor alegătorilor din localitățile respective), se selectează respondenții, numărul cărora pentru fiecare localitate se stabilește prin tehnica eșantionării prin stratificare (fiecare localitate – un strat al populației selectate la al treilea nivel), proporțional la numărul locuitorilor din fiecare localitate.

Eșantionarea nealeatoare (nonprobabilistă) reprezintă acea tehnică de constituire a eșantionului, care presupune necunoașterea probabilității de includere în eșantion a indivizilor colectivității. Selecția are deci un caracter arbitrar și se bazează, în primul rând, pe judecata personala a cercetătorului, presupunând o „alegere rezonabila”. În anumite situații, o asemenea metodă poate fi utilă pentru scopurile cercetării.

Cele mai utilizate tehnici de eșantionare nealeatoare sunt următoarele:

- eșantionarea pe cote;
- eșantionarea de conveniență (de persoane disponibile);
- eșantionarea prin identificare (tehnica *snowball* – a „bulgărului de zăpadă”);
- eșantionarea prin evaluare (logică, subiectivă).

1. Eșantionarea pe cote este similară cu tehnica stratificată proporțională, cu deosebirea că indivizii nu sunt selectați aleator, ci în funcție de disponibilitatea și accesibilitatea lor, până la constituirea numărului corespunzător. Structura eșantionului este hotărâtă *a priori* (de exemplu, proporție bărbați/femei, proporție rural/urban, procente pe grupe de vârstă etc.), iar în alegerea respondenților intervievatorul are o mai mare influență și libertate (poate căuta persoanele respective în zone unde consideră că este mai probabil să le găsească, nu trebuie sa revină la un domiciliu dacă nu a găsit pe nimeni acasă etc.).

În felul acesta, eșantionarea pe cote este mult mai economicoasă (costuri mai mici de deplasare și, corespunzător, un timp mai scurt de colectare a datelor), ceea ce reprezintă un avantaj. Dezavantajul acestei tehnici constă în faptul că deși structura eșantionului poate fi construită, astfel încât să reproducă populația, nu este nicio garanție că eșantionul este reprezentativ.

2. Prin eșantionarea de conveniență se selectează indivizii apți, disponibili de a participa la sondaj. Este cea mai puțin riguroasă tehnică de eșantionare, deoarece ea presupune alegerea componentelor eșantionului în cel mai simplu mod posibil: prin oprirea și luarea unor interviuri, de obicei scurte, a unor persoane aflate în incinta magazinelor sau pe stradă. Prin această metodă, destul de economă, se realizează un eșantion care nu poate fi reprezentativ pentru o anumită populație sau colectivitate. Concluziile rezultate, desigur, nu se pot generaliza la nivelul populației avute în vedere. Cu toate acestea, o asemenea metodă este utilă în cazul unor cercetări-pilot care, ulterior, vor fi urmate de cercetări ce vor implica eșantioane stabilite probabilistic.

3. Eșantionarea prin identificare (tehnica *snowball* – a „bulgărului de zăpadă”) este o tehnică folosită, atunci când se studiază o populație greu de găsit. Nu există liste ale indivizilor din care s-ar putea selecta eșantionul, însă constituirea lui poate fi bazată pe faptul că astfel de indivizi se cunosc între ei.

Procedura se desfășoară în câteva faze. În prima fază cercetătorul identifică o serie de indivizi care îndeplinesc condițiile de includere în eșantionul cercetării. În faza a doua aceștia sunt rugați să caute alți indivizi care îndeplinesc anumite criterii explicite (vârstă, nivel de pregătire, apartenență la anumite grupuri de preocupări etc.). În continuare, operația se repetă cu următorii indivizi intervievați și se aseamănă cu rostogolirea unui bulgăre de zăpadă, având un efect similar – eșantionul devine tot mai mare.

Menționăm că eșantionul obținut nu este unul reprezentativ și poate fi folosit numai pentru studii exploratorii și descriptive.

4. Eșantionarea prin evaluare (logică, subiectivă) se realizează prin includerea cazurilor ca urmare a deciziei subiective a cercetătorului, care alege unitățile de eșantionare în conformitate cu anumite criterii, astfel încât să se asigure ceea ce el consideră că este reprezentativ pentru populația vizată. Reprezentativitatea unui eșantion constituit în acest mod depinde de experiența și intuiția cercetătorului, iar uneori poate funcționa foarte bine.

Eșantionarea este o etapă esențială în cercetare, acuratețea și validitatea rezultatelor investigației depinzând, în mare măsură, de felul în care au fost selectați subiecții și de numărul lor. Argumentele pentru alegerea unei anume metode de eșantionare și a mărimii eșantionului sunt în multe situații pragmatice, depășind considerentele strict teoretice. Resursele financiare și umane, timpul aflat la dispoziție pentru derularea anchetei sociologice și elaborarea raportului de cercetare, structura și mărimea chestionarului, informațiile cu privire la populația investigată (existenta unui cadru de eșantionare) sunt aspecte practice față de care cercetătorul întotdeauna trebuie să țină cont. Aceste constrângeri, împreună cu cele de ordin teoretic aflate în acord cu obiectivele anchetei sociologice, fac dificilă munca cercetătorului, care trebuie să dovedească în etapa de cercetare dedicată eșantionării multă imaginație, o cât mai bună cunoaștere a populației investigate și, evident, a metodologiei de eșantionare.

Exerciții, întrebări de control

1. Fie date facultățile USM cu numărul de studenți la secția zi, licență:

BP – Biologie și Pedologie	400
CTC – Chimie și Tehnologie Chimică	600
D - Drept	2.800
FI – Fizică și Inginerie	1.200
IF – Istorie și Filosofie	400
JSC – Jurnalism și Științe ale Comunicării	600
LLS – Limbi și Literaturi Străine	900
L – Litere	2.200
MI – Matematică și Informatică	800
PSESAS – Psihologie și Științe ale Educației, Sociologie și Asistență Socială	2.900
RISPA – Relații Internaționale, Științe Politice și Administrative	2.700
SE – Științe Economice	2.500

a) În vederea realizării unui sondaj la USM, construiți un eșantion reprezentativ probabilist, proporțional pe facultăți, cu marja de eroare de 3% și probabilitatea de estimare de 95%. Numărul studenților de la fiecare facultate rotunjiți-l până la următorul număr întreg.

b) Pentru realizarea unui sondaj la USM, a fost construit un eșantion reprezentativ cu marja de eroare de 3% și probabilitatea de estimare de 95%, selectându-se aleatoriu de la fiecare facultate același număr de studenți (acesta a fost rotunjit până la următorul număr întreg). Calculați coeficienții de ponderare pentru facultățile USM.

c) Dacă presupunem că în medie la facultățile USM la anul I, Licență studiază 40% din toți studenții facultății, la anul II – 32%, la anul III – 28%, câți studenți de la fiecare an de studii de la fiecare facultate vor fi incluși în eșantionul reprezentativ, proporțional pe facultăți și ani de studii (marja de eroare a eșantionului – 3%, probabilitatea de estimare – 95%). Rezultatul rotunjiți-l până la următorul număr întreg.

3. Folosind metoda eșantionării multistadiale, să se construiască un eșantion național reprezentativ, selectând proporțional respondenții din câte trei raioane din fiecare zonă geografică a Modovei (Nord, Centru, Sud) și municipiile Chișinău și Bălți. Raioanele vor fi selectate aleatoriu, iar în eșantion vor fi incluși respondenți din mediul rural și urban, proporțional numărului de locuitori din raioanele și municipiile selectate. Pentru eșantionul național, să se considere marja de eroare de 2% și probabilitatea de estimare de 95%. Numărul final de respondenți din fiecare subdiviziune administrativă să se rotunjească până la următorul număr întreg. Cunoscând repartizarea populației Republicii Moldova după mediul de reședință, să se determine coeficienții de ponderare a bazei de date după mediul de reședință al respondenților. Informația demografică necesară pentru construirea eșantionului să se preia de pe site-ul www.statistica.md (statistica demografică de la începutul anului calendaristic cel mai recent).

4. Propuneți cercetări în cadrul USM, pentru care eșantionul poate fi construit numai prin metoda „bulgărului de zăpadă”.

Tema 7

Programamul *SPSS*: descriere generală. Definirea variabilelor, introducerea, verificarea şi corectarea datelor

Programul SPSS (*Statistical Package for Social Sciences*) se utilizează pentru prelucrarea statistică a datelor prin:

- elaborarea, prin definirea variabilelor, a bazelor de date din diferite domenii ce studiază populaţii (sociologie, psihologie, medicină, demografie, marketing etc.);
- introducerea, verificarea şi corectarea datelor (de regulă – codificate);
- prelucrarea datelor prin metodele statisticii descriptive (frecvenţe, dependenţe între variabile, indicatori statistici etc.);
- reprezentarea rezultatelor sub formă de tabele şi diagrame;
- analiza datelor şi a rezultatelor prin metode ale statisticii inferenţiale;
- gestiunea variabilelor şi a cazurilor: selectarea cazurilor, sortarea cazurilor, calcularea şi recodificarea variabilelor, adăugarea cazurilor şi a variabilelor, ponderarea datelor, divizarea bazei de date pentru analize comparative etc.

Programul SPSS are diferite versiuni, cele mai recente (versiunile 20-24) fiind elaborate pentru mediile sistemelor de operare Windows XP, Windows 7, Windows 8 și Windows 10. Una dintre cele mai reuşite versiuni pentru mediul sistemului de operare Windows XP este versiunea 11.0.

Interfaţa programului nu se deosebeşte esenţial de cea a programelor din pachetul Microsoft Office, iar structura documentului SPSS e asemănătoare cu cea a registrului şi a foilor de calcul Excel (a se vedea Figurile 7.1 şi 7.2).

Documentul SPSS conţine două foi, numite *Data View* şi *Variable View*. Ambele sunt divizate în linii şi coloane, asemănător foilor de calcul Excel, însă fiecare îşi are destinaţia sa: *Data View* este prevăzută pentru introducerea şi păstrarea datelor, iar *Variable View* − pentru definirea variabilelor şi păstrarea lor. Ambele nu sunt altceva decât nişte baze de date: *Data View* conţine valori ale variabilelor (de regulă – codificate) ce ulterior se prelucrează prin metode statistice, iar *Variable View* − lista variabilelor împreună cu proprietăţile lor.

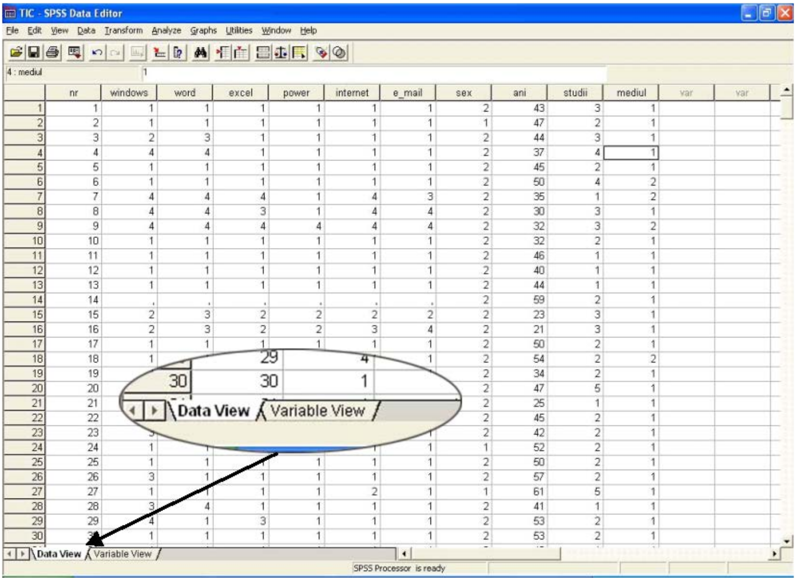


Fig. 7.1. Interfaţa Programului SPSS (foaia *Data View*)

Meniul programului SPSS conţine următoarele unităţi:

- **File** – comenzi de gestiune a întregului document;
- **Edit** − comenzi de redactare;
- **View** − comenzi de vizualizare şi modificare a elementelor interfeţei;
- **Data** − comenzi de gestiune a datelor şi a bazei de date;

- **Transform** – comenzi de calculare a noilor variabile, de recodificare;
- **Analyze** – comenzi de elaborare a rezultatelor, de analiză a variabilelor şi datelor;
- **Graphs** – comenzi de construire şi redactare a diagramelor;
- **Utilities** – utilităţi suplimentare: afişarea informaţiei despre variabile şi baza de date, gruparea variabilelor etc.;
- **Window** – comenzi de gestiune a ferestrelor documentelor;
- **Help** – regim de asistenţă.

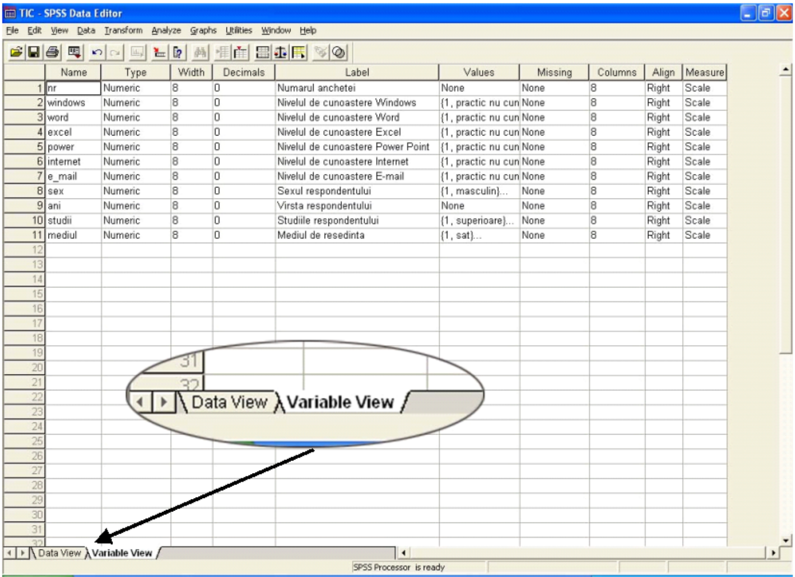


Fig. 7.2. Interfaţa Programului SPSS (foaia *Variable View*)

Definirea variabilelor are ca scop formarea structurii bazei de date a cercetării. Variabilele se definesc completându-se foaia *Variable View*. A defini o variabilă înseamnă a-i atribui următoarele proprietăţi:

- *nume* (Name) − o identifică univoc în mulţimea tuturor variabilelor aferente cercetării;
- *tip* (Type) − stabileşte tipul valorilor variabilei (numeric, text, dată etc.);
- *lungime* (Width) − numărul de poziţii ocupate de valoarea variabilei;
- *număr de zecimale* (Decimals) − exactitatea reprezentării valorilor numerice;
- *etichetă* (Label) − denumirea deplină a variabilei (caracteristicii);
- *valori* (Values) − scala de valori ale variabilei (în cazul variabilelor numerice ea nu se defineşte).

La definirea în SPSS a variabilelor se respectă următoarele condiţii:

- numele variabilei se atribuie de către utilizator, trebuie să fie unic, nu poate să se repete în lista variabilelor;
- numele variabilei poate conţine până la 8 caractere (litere latine, cifre, semnul „_”, punctul în interiorul numelui) şi începe cu o literă;
- dacă variabilele se definesc în baza chestionarului, atunci ordinea lor se recomandă să corespundă ordinii întrebărilor, prima variabilă definită fiind numărul de ordine al chestionarului.

Observaţia 7.1. În procesul definirii variabilelor pot fi utilizate metode de copiere, mutare, corectare a celulelor, asemănătoare cu cele utilizate în Excel.

Observaţia 7.2. În orice moment, structura bazei de date poate fi redactată prin adăugarea sau ştergerea variabilelor şi, respectiv, schimbarea lor cu locul, utilizând aceleaşi procedee de gestiune a celulelor, liniilor şi coloanelor ca şi în Excel.

Prin câteva exemple vom explica relaţiile dintre tipurile întrebărilor din chestionar şi variabilele din baza de date, ele fiind utile pentru înţelegerea metodei de definire a variabilelor.

Exemplul 7.1

Întrebare cu o singură variantă de răspuns:

Care e culoarea din mijloc pe drapelul Moldovei?

1) alb
2) roşu
3) galben
4) verde
5) albastru
6) negru
7) nu ştiu

Variabila:

nume – cul_mijl
tip − numeric
lungime − 8
zecimale − 0
etichetă: Care e culoarea din mijloc pe drapelul Moldovei?
valori:
1 – alb
2 – roşu
3 – galben
4 – verde
5 – albastru
6 – negru
7 – nu ştiu

Exemplul 7.2

Întrebare cu un număr fixat de variante de răspuns:

Care 3 culori se conţin pe drapelul Moldovei?

1) alb
2) roşu
3) galben
4) verde
5) albastru
6) negru
7) alta

Variabile vor fi 3:

3 nume diferite – cul1, cul2, cul3
tip − numeric
lungime − 8
zecimale − 0
3 etichete (ele coincid):
Care culoare se conţine pe drapelul Moldovei?
valorile fiecărei variabile − aceleaşi:
1 – alb
2 – roşu
3 – galben
4 – verde
5 – albastru
6 – negru
7 – alta

Exemplul 7.3

Întrebare cu orice număr de variante de răspuns:

Care din cele enumerate sunt culorile preferate ale Dvs.?

1) alb
2) roşu
3) galben
4) verde
5) albastru
6) negru
7) alta

Variabile vor fi 7
(acelaşi număr ca şi cel al variantelor de răspuns):

7 nume diferite:
alb, rosu, galben, verde, albastru, negru, alta
tip − numeric
lungime − 8
zecimale − 0
7 etichete diferite:
Preferaţi culoarea ... *(se scrie culoarea respectivă)*?
valorile fiecărei variabile − aceleaşi:
0 – nu
1 – da

Exemplul 7.4

Întrebare sub formă de tabel: → **5 întrebări simple:**

În ce măsură preferaţi următoarele culori?

	mult	puţin	deloc
albă	1	2	3
roşie	1	2	3
galbenă	1	2	3
verde	1	2	3
alta	1	2	3

1. În ce măsură preferaţi culoarea albă?
1) mult 2) puţin 3) deloc

2. În ce măsură preferaţi culoarea roşie?
1) mult 2) puţin 3) deloc

3. În ce măsură preferaţi culoarea galbenă?
1) mult 2) puţin 3) deloc

4. În ce măsură preferaţi culoarea verde?
1) mult 2) puţin 3) deloc

5. În ce măsură preferaţi o altă culoare?
1) mult 2) puţin 3) deloc

Observăm că întrebările sub formă de tabel se transformă în atâtea întrebări cu o singură variantă de răspuns, câte linii are tabelul, ca apoi să fie definite variabilele pentru fiecare întrebare în parte.

Practic, procesul de definire a variabilelor şi creării structurii bazei de date (în *Data View* coloanele din tabel automat preiau în calitate de denumiri numele variabilelor definite) se realizează în SPSS în următoarea consecutivitate:

- Se lansează programul SPSS (**Start** → *Programs* → *SPSS for Windows* → *SPSS 11.0);*
- Se trece la foaia *Variable View;*
- Se defineşte prima variabilă, care, de regulă, este numărul chestionarului (de exemplu: *Name* − nr, *Type* − Numeric, *Width* − 8, *Decimals* − 0, *Label* − Numarul chestionarului, *Values* − None). Documentul SPSS primeşte, astfel, conţinutul demonstrat în Figura 7.3;

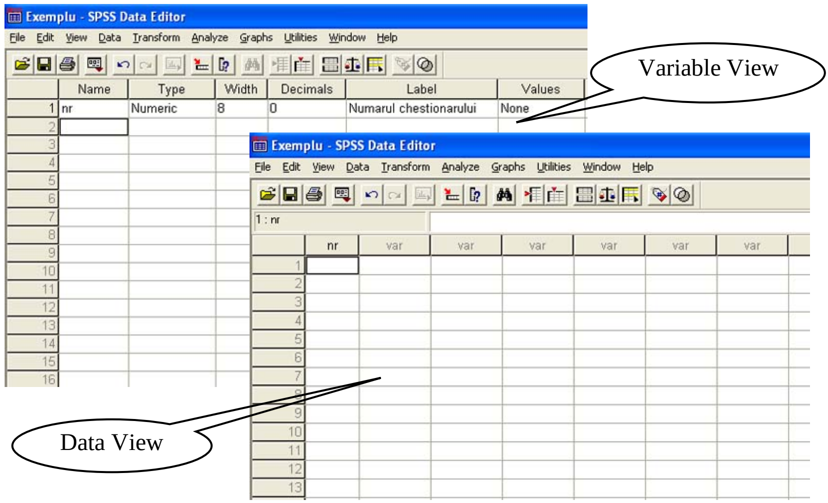


Fig. 7.3. Documentul SPSS după definirea primei variabile

- Se salvează baza de date (**File** → *Save As...);*
- Se continuă definirea celorlalte variabile, din când în când salvându-se baza de date cu **File** → *Save.*

Introducerea datelor se face în foaia *Data View* pe linii, introducând consecutiv numerele de ordine (codurile) ale răspunsurilor marcate din chestionar. Celulele, pentru care în chestionar nu sunt marcate variante de răspuns, rămân necompletate: programul le va considera omise (*missing*).

Introducerea datelor din chestionare poate fi făcută și în fișierele Excel. Pentru aceasta, în prealabil, baza de date elaborată în SPSS se salvează cu *Save As…* în format Excel (*.xls*). Denumirile variabilelor (coloanelor), preluate din SPSS, se vor poziționa în prima linie a foii de calcul, iar datele se vor introduce începând cu linia a doua. Pentru a evita greșelile la introducerea datelor (greșeli ce țin de ieșirea din diapazonul de valori al variabilei), în coloane se introduc restricții asupra numerelor ce urmează a fi introduse prin comanda **Data**→*Data Validation…*

Datele pot fi introduse de câţiva operatori, în final ele putând fi adunate într-o singură bază de date prin comanda **Data**→ *Merge Files*→ *Add Cases...*, dacă ele sunt introduse în fișiere SPSS. Cele introduse în fișiere Excel, pur și simplu, se copie în baza de date SPSS începând cu prima coloană. În final, datele sunt sortate după numărul de ordine al chestionarului (**Data**→*Sort Cases...).*

Verificarea şi corectarea datelor e comod a fi realizată în Excel, după salvarea în SPSS a bazei de date în format Excel (*.xls*) cu comanda **File**→ *Save As...* În continuare:

- în baza de date din Excel se introduce filtrul (**Data** → *Filter*→*AutoFilter*);
- consecutiv, variabilă cu variabilă, se verifică dacă datele introduse nu iese din domeniul de valori al variabilei respective;
- dacă se detectează greşeli, atunci se determină numărul chestionarului pentru care au fost comise (acest număr se găseşte în prima coloană a bazei de date!);

- corectarea greşelilor se face folosind chestionarele originale, în baza de date schimbându-se valorile greşite cu cele marcate în chestionare;
- un următor pas este verificarea variabilelor corespunzătoare întrebărilor de control şi/sau de trecere, filtrând baza de date cu condiţiile respective.

Baza de date, corectată în Excel, se copie prin metoda obişnuită (*selectare date în Excel→ Copy→ Paste*) înapoi în SPSS în prima celulă din *Data View*, astfel devenind pregătită pentru a trece la etapa de prelucrare a datelor.

Observaţia 7.1. Prin comanda **Utilites** → *File Info* programul SPSS afişează informaţia despre variabilele din baza de date (a se vedea exemplul din Figura 7.4).

```
List of variables on the working file
Name                                                        Position
NR         Numarul chestionarului                                  1
           Measurement Level: Scale
           Column Width: 8  Alignment: Right
           Print Format: F8
           Write Format: F8
WINDOWS    Nivelul de cunoastere Windows                           2
           Measurement Level: Scale
           Column Width: 8  Alignment: Right
           Print Format: F8
           Write Format: F8
           Value     Label
              1      practic nu cunosc
              2      slab
              3      suficient
              4      mediu
              5      inalt
...
```

Fig. 7.4. Un fragment de informaţie, afişată de SPSS în urma executării comenzii Utilites→*File Info*

Observăm că o astfel de informaţie permite, de exemplu, a „restabili” chestionarul, în baza căruia a fost elaborată baza de date respectivă. Pentru aceasta este suficient de a copia informaţia afişată într-un editor de text (*Word,* de exemplu) şi de a o redacta, ştergând totul în afară de textul evidenţiat (evidenţierea aparţine autorului, ea cuprinzând etichetele şi valorile variabilelor din baza de date).

Exerciții, întrebări de control

1. Care din următoarele șiruri de caractere nu pot fi folosite în calitate de nume de variabilă în SPSS 11.0: *name_1, name 2, _name3, name.4, name5., name*6, name7?, 8name, name_nine, name.ten* . Explicați, de ce?
2. Este dat un fragment de chestionar:

   …

   Q1. Câte persoane locuiesc în gospodăria Dvs.? ____

   Q2. Unde Vă simţiţi mai bine?
   1. Acasă
   2. În ospeţie

   Q3. Ce obiecte aveţi în gospodărie?
   1. Aparat TV
   2. Aparat de radio
   3. Telefon fix
   4. Fier de călcat
   5. Maşină de spălat
   6. Aragaz
   7. Alte obiecte

   Q4. Indicaţi 3 din cele mai importante scopuri din viața Dvs.?
   1. Să mă căsătoresc
   2. Să cresc copiii
   3. Să lucrez
   4. Să văd lumea
   5. Să învăţ în continuare
   6. Altceva

   Q5. Indicați data nașterii Dvs.: ziua____ luna____ anul____

   …

   Definiți în SPSS variabilele corespunzătoare întrebărilor din chestionar.
3. Pentru baza de date obținută în pct.2 să se elaboreze în Excel forma de introducere a datelor cu validarea lor.

Tema 8

Prelucrarea primară a datelor în *SPSS*. Calcularea frecvenţelor şi a indicatorilor statistici

Prelucrarea primară a datelor are ca scop obținerea unui tablou general al rezultatelor, examinarea suplimentară a variabilelor şi depistarea greşelilor ce n-au fost descoperite prin alte proceduri şi metode. Această prelucrare se face prin determinarea frecvenţelor variabilelor şi prin calcularea indicatorilor statistici ai acestora.

Determinarea frecvenţelor în SPSS se face prin acţionarea comenzii **Analyze**→ *Descriptive Statistics*→ *Frequencies*...Ca rezultat apare caseta de dialog, prezentată în Figura 8.1.

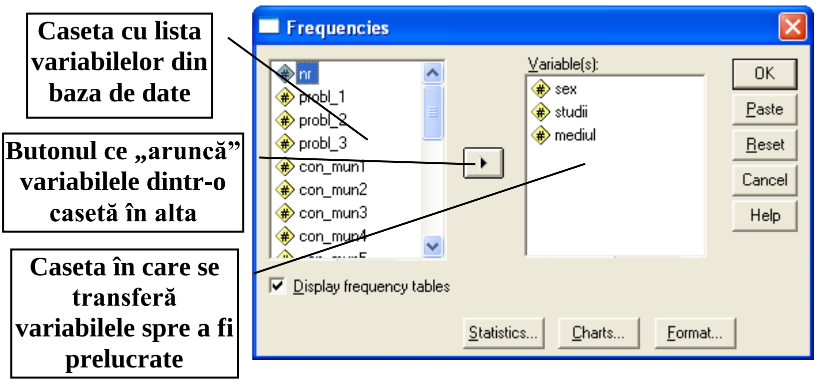
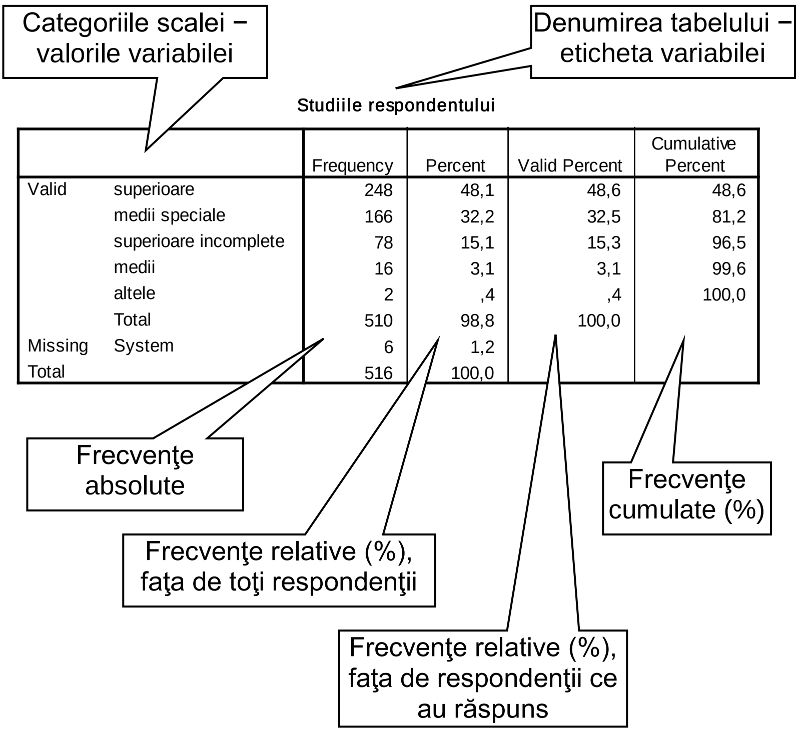


Fig. 8.1. Caseta de dialog *Frequencies*

În continuare, din lista din partea stângă a casetei de dialog se transferă în zona *Variable(s)* din dreapta variabilele, pentru care se determină frecvenţele. Acţionarea butonului **OK** conduce la afişarea rezultatului (a se vedea Figura 8.2), care apare într-un nou tip de document (*.spo*) cu denumirea implicită *Output*, ce poate fi salvat şi păstrat sau din care pot fi copiate, prin metoda obişnuită, rezultatele în alte tipuri de documente (*Word, Excel* etc.).



Studiile respondentului

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	superioare	248	48,1	48,6	48,6
	medii speciale	166	32,2	32,5	81,2
	superioare incomplete	78	15,1	15,3	96,5
	medii	16	3,1	3,1	99,6
	altele	2	,4	,4	100,0
	Total	510	98,8	100,0	
Missing	System	6	1,2		
Total		516	100,0		

Fig. 8.2. Rezultatul determinării frecvenţelor

Observaţia 8.1. Dacă în caseta de dialog ***Frequencies*** se acţionează butonul **Charts...** , atunci suplimentar poate fi construită şi o diagramă a frecvenţelor. Tipul acesteia se indică de către utilizator.

Calcularea indicatorilor statistici pentru variabilele selectate în caseta ***Frequencies*** se va face prin acţionarea butonului **Statistics...** şi bifarea casetelor de validare corespunzătoare indicatorilor solicitaţi spre a fi determinaţi (a se vedea Figura 8.3). În aşa mod pot fi determinate: media (*Mean*), mediana (*Median*), modulul (*Mode*), dispersia sau abaterea standard (*St. deviation*), valorile minimale (*Minimum*) şi maximale (*Maximum*) ale variabilelor analizate.

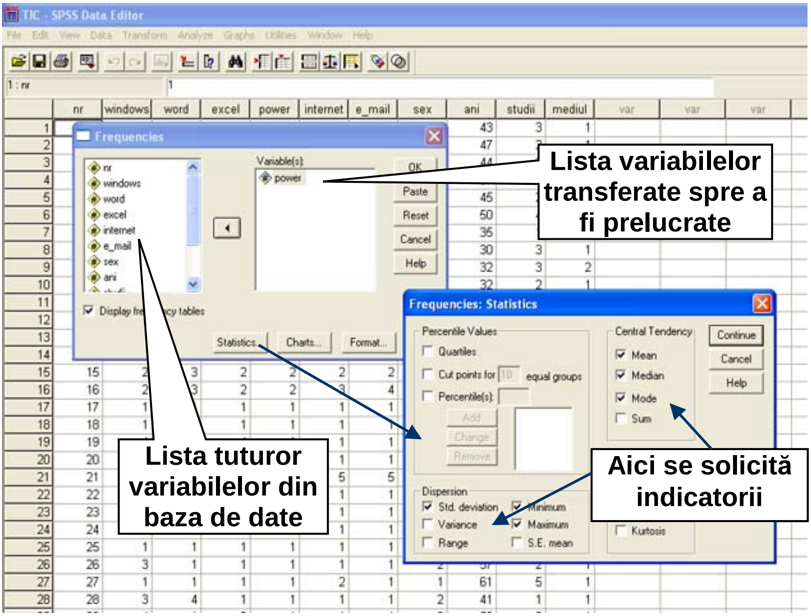
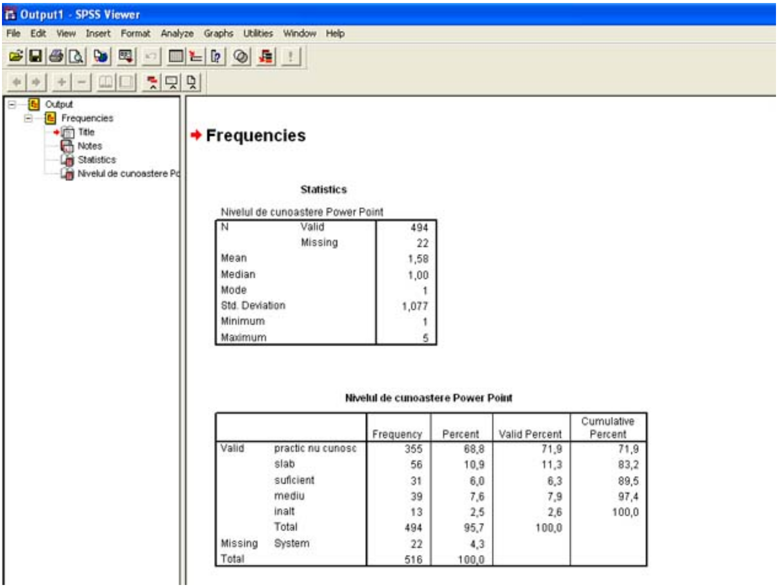


Fig. 8.3. Calcularea indicatorilor statistici în SPSS

Rezultatul va apare sub formă de tabel, intitulat *Statistics*, în acelaşi document de afişare a rezultatelor (a se vedea Figura 8.4).

Observaţia 8.2. Indicatorii statistici pot fi calculaţi şi prin comanda **Analyze**→ *Descriptive Statistics*→ *Descriptives...* Procedura este asemănătoare cu cea descrisă *supra*.

Observaţia 8.3. Prin meniul **Graphs** programul SPSS permite construirea diferitelor diagrame, care pot fi utilizate pentru analiza variabilelor şi a relaţiilor dintre ele.



Statistics

Nivelul de cunoastere Power Point

N	Valid	494
	Missing	22
Mean		1,58
Median		1,00
Mode		1
Std. Deviation		1,077
Minimum		1
Maximum		5

Nivelul de cunoastere Power Point

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	practic nu cunosc	355	68,8	71,9	71,9
	slab	56	10,9	11,3	83,2
	suficient	31	6,0	6,3	89,5
	mediu	39	7,6	7,9	97,4
	inalt	13	2,5	2,6	100,0
	Total	494	95,7	100,0	
Missing	System	22	4,3		
Total		516	100,0		

Fig. 8.4. Rezultatul determinării indicatorilor statistici în SPSS

Observaţia 8.3. Frecvenţele şi indicatorii statistici pot fi calculaţi şi cu ajutorul programului *Excel*. Pentru aceasta este necesar a avea datele în foaia de calcul şi a folosi următoarele funcţii:

=MODE(domeniu) − pentru determinarea modulului;

=MEDIAN(domeniu) − pentru calcularea medianei;

=AVERAGE(domeniu) − pentru calcularea mediei;

=MAX(domeniu) − *MIN(domeniu)* − pentru calcularea amplitudinii;

=STDEV(domeniu) − pentru calcularea abaterii standard;

=FREQUENCY(domeniu, limite) − pentru calcularea frecvenţelor, în care sunt introduse notările:

domeniu – domeniul de celule în care se găsesc datele analizate;

limite − domeniul ce conţine capetele intervalelor, în care se calculează frecvenţele.

Exerciții, întrebări de control

1. În tabelul *infra* sunt prezentate opiniile femeilor și ale bărbaților față de concubinaj (1 – pozitivă, 2 – negativă, 3 – dificil de apreciat).

Opinii față de concubinaj

Femei	2	2	2	2	1	1	2	2	2	1	1	1	3	1
Bărbați	1	1	1	1	2	1	1	3	3	1	1	2	3	2

Introduceți aceste date în programul *SPSS*, determinați frecvențele răspunsurilor și le comparați grafic în *Excel*, sub formă de diagramă cu bare.

2. În Figura 8.5 sunt reprezentate frecvențele răspunsurilor unui grup de respondenți cu privire la aprecierea sănătății lor, calculate în SPSS. Să se interpreteze rezultatele marcate.

Cum apreciati starea sanatatii Dvs. in prezent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excelenta	313	9,2	9,4	9,4
	Buna	1786	52,5	53,4	62,8
	Satisfacatoare	677	19,9	20,3	83,0
	Rea	140	4,1	4,2	87,2
	Nu stiu/nu o pot aprecia	427	12,5	12,8	100,0
	Total	3343	98,2	100,0	
Missing	System	62	1,8		
Total		3405	100,0		

Fig. 8.5. Aprecierea sănătății de către respondenți

3. Construiți în *Excel* o diagramă circulară, care să reprezinte rezultatul din Figura 8.5. În calitate de date pentru diagramă să se folosească frecvențele absolute (*Frequency*), iar cele procentuale să se afișeze în jurul diagramei.
4. Analizați rezultatul aprecierii sănătății numai pentru respondenții care au considerat-o ca fiind de la „rea” până la „excelentă”. Construiți diagrama respectivă în *Excel*, afișând în jurul ei frecvențele procentuale.
5. Interpretați și analizați rezultatele calculării indicatorilor statistici pentru două variabile („numărul de copii dorit de respondent” și „numărul ideal de copii în familie”), prezentați în Figura 8.6.

Statistics

		Numarul de copii dorit de respondent	Numarul ideal de copii in familie in opinia respondentului
N	Valid	412	416
	Missing	24	20
Mean		2,437	2,736
Median		2,000	3,000
Mode		2,0	2,0
Std. Deviation		1,0503	1,0964
Minimum		,0	,0
Maximum		7,0	8,0

Fig. 8.6. Numărul dorit și ideal de copii în familie

6. Cum se va schimba numărul mediu dorit de copii din exemplul *supra* (Fig. 8.6), dacă la eșantion s-ar mai adăuga 100 de respondenți, care își doresc să aibă câte 3 copii?

Tema 9

Asocierea variabilelor.
Construirea tabelelor de asociere

De multe ori, este interesant a determina cum se comportă valorile unei variabile faţă de valorile altor variabile. Spre exemplu, cum s-au repartizat răspunsurile la o întrebare din chestionar în funcție de sexul, vârsta, nivelul studiilor etc. celor intervievaţi. În astfel de cazuri, ne vine în ajutor programul SPSS prin comenzile de construire a tabelelor de asociere a variabilelor, rezultatele fiind exprimate atât în frecvenţe absolute, cât şi relative.

Tabelele de asociere se elaborează prin meniul **Analyze**→ *Custom Tables* ▸. Dintre variantele posibile de tabele cele mai simple sunt cele generale (→ *General Tables...*). Pas cu pas, vom demonstra procedura de construire a tabelelor de asociere.

În primul rând, vom conveni asupra următoarelor:

- în tabelele de asociere variabilele le vom diviza în *dependente* şi *independente*;
- vom considera ***dependente*** variabilele ce se analizează (se studiază). De regulă, ele se poziționează în coloanele tabelului de asociere;
- vom considera ***independente*** variabilele faţă de care se analizează (se studiază) cele dependente. Ele sunt acele, care se poziționează în liniile tabelului de asociere.

Lansarea în SPSS a comenzii **Analyze** → *Custom Tables*▸→ *General Tables...* conduce la afişarea pe ecran a casetei de dialog prin care se introduc parametrii viitorului tabel de dependenţe (a se vedea Figura 9.1).

În continuare, în caseta *Rows:* se transferă variabilele independente. În viitorul tabel valorile lor vor apărea în stânga tabelului, în calitate de denumiri ale liniilor. După necesitate, pentru fiecare variabilă independentă, se acţionează butonul **Insert Total**, care va permite calcularea totalurilor pe coloane după fiecare variabilă independentă din tabel.

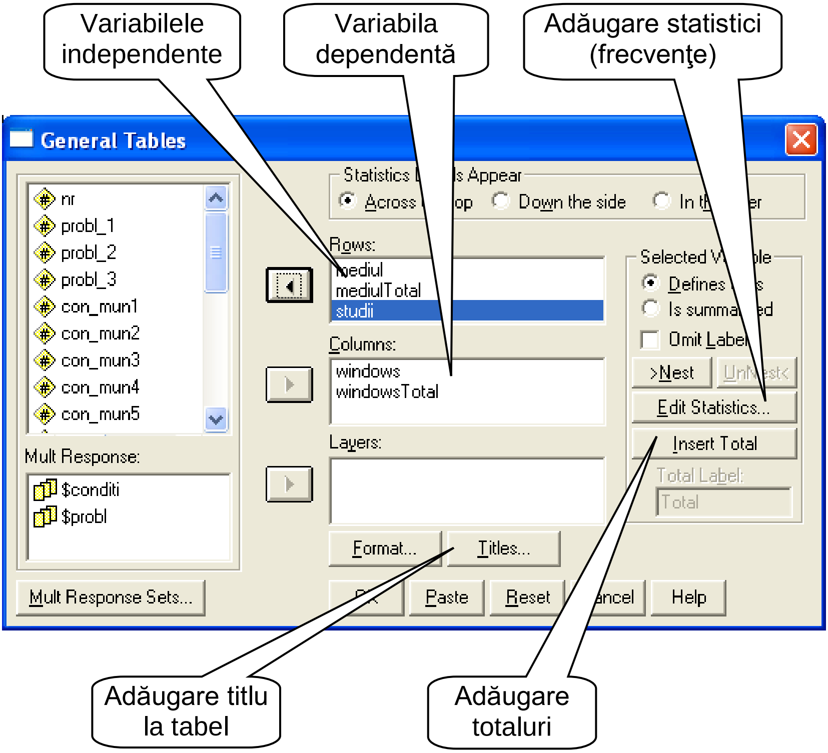


Fig. 9.1. Caseta de dialog de construire a tabelelor generale

Variabila care se studiază (cea dependentă), de regulă − numai una, se transferă în caseta *Columns:* . În viitorul tabel valorile ei vor apărea în „pălăria” tabelului, în calitate de denumiri ale coloanelor. Pentru variabila dependentă, la fel, pentru a calcula totalurile pe linii se poate acţiona butonul **Insert Total**.

Acţionarea butonului **Titles...** permite a culege titlul viitorului tabel. După acţionarea butonului **OK** rezultatul acestor setări − tabelul solicitat − va fi afişat în documentul de afişare a rezultatelor, având forma Tabelului 9.1.

Tabelul 9.1

Cunoasterea *WINDOWS*

		Nivelul de cunoaștere *Windows*					
		practic nu cunosc	slab	suficient	mediu	înalt	Total
Mediul de reședință	sat	202	61	49	63	21	396
	oraș	24	4	18	30	22	98
Total		226	65	67	93	43	494
Studiile respondentului	superioare	87	29	36	57	32	241
	medii speciale	103	20	13	13	6	155
	superioare incomplete	20	13	17	20	5	75
	medii	10	2	1	3		16
	altele	2					2
Total		222	64	67	93	43	489

Observăm că în Tabelul 9.1 sunt afişate numai frecvenţe absolute (numărul de indivizi − persoanele intervievate). Pentru a impune programul SPSS să calculeze şi frecvenţe relative, exprimate în procente, este necesar ca în caseta de dialog *General Tables* (a se vedea Figura 9.1) să se acţioneze butonul **Statistics...** , după care apare caseta de dialog suplimentară *General Tables: Cell Statistics for windows* (a se vedea Figura 9.2). În această casetă de dialog se transferă din stânga în dreapta tipurile de frecvenţe relative, necesar a fi calculate (*Row%* − frecvenţe pe linii, *Col%* − frecvenţe pe coloane, *Count* − frecvenţe absolute etc.). Pentru fiecare din aceste frecvenţe, poate fi setat şi un format de afişare. Acţionarea butonului **Continue** ne întoarce la caseta de dialog *General Tables*, în care se acţionează butonul **OK**.

Un fragment de tabel, ce conţine şi frecvenţe relative, este prezentat în Figura 9.3.

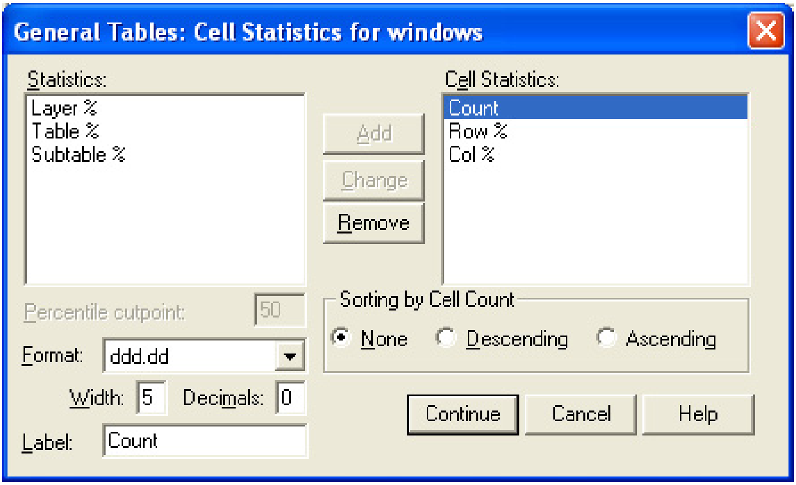


Fig. 9.2. Caseta de dialog prin care se adaugă statisticile

Cunoaşterea Windows

		practic nu cunosc			slab	
		Count	Row %	Col %	Count	Row %
Mediul de resedinta	sat	202	51,0%	89,4%	61	15,4%
	oras	24	24,5%	10,6%	4	4,1%
Total		226	45,7%	100,0%	65	13,2%
Studiile respondentului	superioare	87	36,1%	39,2%	29	12,0%
	medii speciale	103	66,5%	46,4%	20	12,9%
	superioare incomplete	20	26,7%	9,0%	13	17,3%
	medii	10	62,5%	4,5%	2	12,5%
	altele	2	100,0%	,9%		
Total		222	45,4%	100,0%	64	13,1%

Fig. 9.3. Tabel cu statistici (frecvenţe absolute şi relative), elaborat în programul SPSS

Interpretarea rezultatelor obținute sub forma tabelelor de asociere o vom explica în baza tabelului din Figura 9.3.

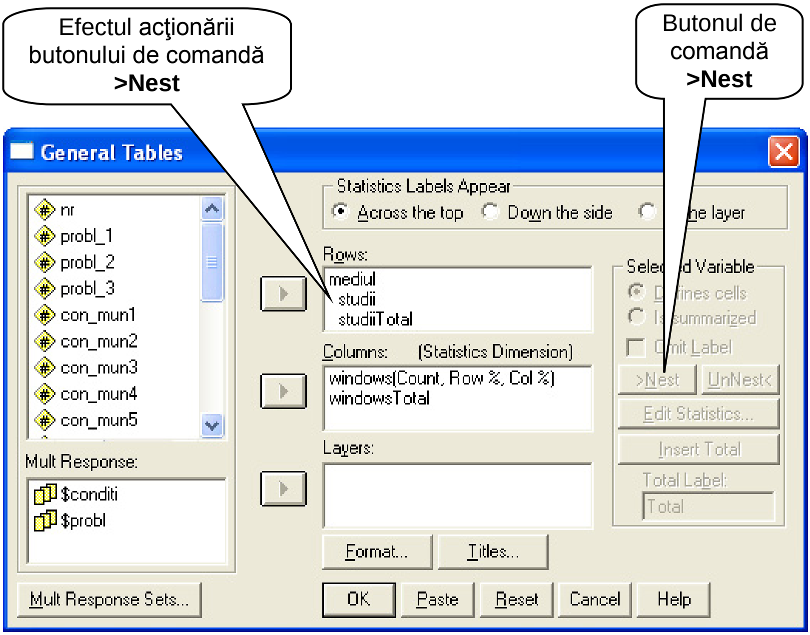
Interpretarea frecvențelor absolute (Count) este evidentă: ele reprezintă numărul de respondenți din categoria din linia respectivă, indicată în stânga tabelului, care au opinia din coloana respectivă, indicată în partea de sus a tabelului.

La interpretarea frecvențelor relative, ne vom baza pe faptul că suma procentelor pe linie (Row%) este egală cu 100%, ceea ce înseamnă că ele exprimă opiniile față de problema cercetată numai ale respondenților din linia respectivă. Tot așa, deoarece suma procentelor pe coloană (Col%) este egală cu 100%, ele reprezintă repartizarea categoriilor de respondenți după opinia indicată în coloana respectivă.

Astfel, de exemplu (a se vedea rezultatele din Figura 9.3):

- 24 de respondenți de la oraș practic nu cunosc programul Windows;
- 29 de respondenți cu studii superioare cunosc slab programul Windows;
- 51,0% din toți respondenții de la sat practic nu cunosc programul Windows;
- 4,1% din toți respondenții de la oraș cunosc slab programul Windows;
- din respondenții care practic nu cunosc programul Windows 89,4% sunt de la oraș și 10,6% sunt de la sat;
- din respondenții ce cunosc slab programul Windows 12,0% au studii superioare, 12,9% – medii speciale etc.

O posibilitate interesantă în procedura de construire a tabelelor de asociere o reprezintă descompunerea valorilor unei variabile independente după valorile alteia, ce permite a diviza și diversifica categoriile de respondenți. Modalitatea şi rezultatul acestei descompuneri sunt demonstrate în Figura 9.4.



				practic nu cunosc			slab	
				Count	Row %	Col %	Count	Row %
Mediul de resedinta	sat	Studiile respondentului	superioare	76	44,7%	34,2%	25	14,7%
			medii speciale	93	65,5%	41,9%	20	14,1%
			superioare incomplete	18	29,0%	8,1%	13	21,0%
			medii	9	60,0%	4,1%	2	13,3%
			altele	2	100,0%	,9%		
		Total		198	50,6%	89,2%	60	15,3%
	oras	Studiile respondentului	superioare	11	15,5%	5,0%	4	5,6%
			medii speciale	10	76,9%	4,5%		
			superioare incomplete	2	15,4%	,9%		
			medii	1	100,0%	,5%		
		Total		24	24,5%	10,8%	4	4,1%

Fig. 9.4. Descompunerea valorilor unei variabile după valorile alteia şi rezultatul acestei descompuneri

Tabelele, elaborate în SPSS, de regulă, nu sunt pregătite pentru a fi folosite în publicaţii, rapoarte, studii etc., având o formă specifică, conţinând termeni în engleză şi informaţie în surplus. Pentru a le utiliza, ele necesită o redactare şi formatare prealabilă, care poate fi făcută eficient în mediul programului Excel. Astfel, apare necesitatea de a transfera unele rezultate din SPSS în Excel.

Transferul rezultatelor (tabelelor) din SPSS în Excel poate fi efectuat prin două metode: prin copiere obişnuită (selectarea tabelului în SPSS → *Copy* → *Paste* în foia de calcul Excel) sau prin exportare.

Vom descrie a doua metodă, care, spre deosebire de prima, păstrează formatul rezultatelor (tabelelor) din SPSS.

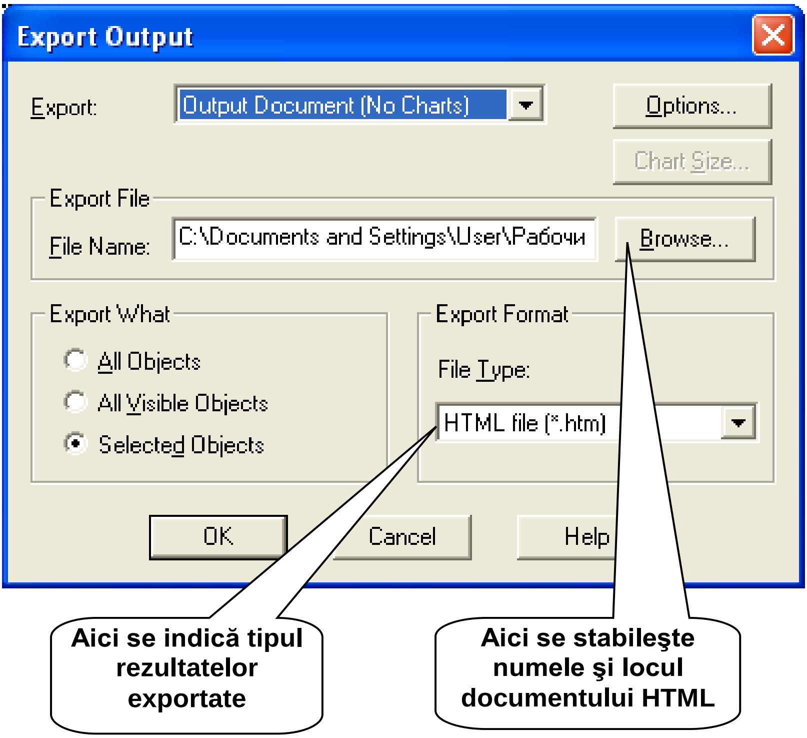


Fig. 9.5. Caseta de export al rezultatelor

Astfel, pentru a exporta rezultatele din SPSS în Excel, se execută următorii paşi:

- în documentul *Output* de afişare a rezultatelor din SPSS se execută un clic drept pe tabelul ce se exportă, urmat de solicitarea comenzii *Export...* din meniul contextual;
- în caseta de dialog ce apare (a se vedea Figura 9.5) se indică locul şi denumirea rezultatului exportat, iar în calitate de tip se indică HTML (în consecinţă, programul generează un document HTML ce se salvează în dosarul indicat de utilizator);
- se deschide documentul HTML, generat de calculator, cu browser-ul *Internet Explorer*;
- din documentul HTML, prin metoda obişnuita (selectare → *Copy* → *Paste*), rezultatul se copie într-o foaie de calcul Excel;
- în Excel rezultatul poate fi prelucrat (redactat, formatat etc.), aducându-l la forma necesară pentru o utilizare ulterioară.

Observaţia 9.1. Rezultatele sub formă de tabele, exportate din SPSS în Excel, pot fi folosite şi pentru construirea diagramelor. Menţionăm aici că Excel-ul, spre deosebire de SPSS, posedă instrumente cu mult mai eficiente de elaborare şi formatare a diagramelor.

Observația 9.2. O altă modalitate de construire a tabelelor de asociere este cea folosită prin comanda: **Analyze** → *Descriptive Statistics* ▸ → *Crosstabs…*

Exerciții, întrebări de control

1. Descrieți structura și conținutul tabelului de asociere a variabilelor. Cum se calculează frecvențele relative pe linii și coloane în astfel de tabele?
2. Prin ce tipuri de diagrame pot fi reprezentate/comparate rezultatele din tabelele de asociere? Argumentați răspunsul prin exemple concrete.
3. Construiți manual trei variante de tabele de asociere pentru datele din următorul tabel: opinia unui grup de respondenți față de concubinaj în

funcție de sex, de mediul de reședință și de combinații ale acestora (bărbați – sat, bărbați – oraș, femei – sat, femei – oraș). Includeți în tabele toate totalurile posibile.

Nr. respondent	**Sex (1 – femeie, 2 – bărbat)**	**Mediul de reședință (1 – sat, 2 – oraș)**	**Opinie față de concubinaj (1 – pozitivă, 2 – neutră, 3 – negativă)**
1	1	1	1
2	2	1	1
3	2	1	2
4	1	1	1
5	1	2	3
6	2	2	3
7	1	1	2
8	1	2	3
9	2	2	1
10	2	1	1
11	1	2	1
12	1	1	3

4. Este dat următorul tabel de asociere a variabilelor:

Utilizaţi programul SPSS la prelucrarea datelor?

	Da			**Nu**			**Total**	
	count	row%	col%	count	row%	col%	count	col%
Sex: fată	**30**				**50%**			
băiat		**70%**				**50%**		
Total:							**160**	

Să se interpreteze rezultatele evidențiate și să se completeze tabelul cu informația lipsă.

5. Este dat următorul tabel de asociere a variabilelor:

Cunoaşterea calculatorului de către studenţi

	Bine		**Suficient**		**Slab**		**Total**	
	Row%	Col%	Row%	Col%	Row%	Col%	Row%	Col%
Fete						25,5%		
Băieţi			15,5%					
Total								

Care din următoarele afirmații sunt corecte:
a) Numai 25,5% dintre fete cunosc calculatorul.
b) 15,5% dintre toţi băieţii cunosc suficient calculatorul.
c) 25,5% dintre toate fetele cunosc slab calculatorul.
d) Din totalul celor ce cunosc suficient calculatorul 15,5% sunt băieţi.
e) Fetele care cunosc slab calculatorul formează 25,5% din toată populaţia cercetată.
f) Băieţii care cunosc suficient calculatorul formează 15,5% din toată populaţia cercetată.
g) Numai 15,5% dintre băieţi cunosc calculatorul.
h) Din totalul celor ce cunosc slab calculatorul 25,5% sunt fete.
i) 41% dintre băieți și fete cunosc între slab și suficient calculatorul.

Tema 10

Prelucrarea întrebărilor cu răspunsuri multiple. Definirea şi utilizarea seturilor de variabile în *SPSS*

După cum s-a menţionat anterior (a se vedea Tema 7), întrebările cu multiple răspunsuri definesc în baza de date atâtea variabile, câte răspunsuri se cer a fi date la ele. Dacă întrebarea presupune un număr determinat de răspunsuri (să zicem − 3), ea defineşte un număr de 3 variabile, care vor fi categoriale, având valori ce coincid cu variantele de răspuns, iar dacă întrebarea presupune orice număr de răspunsuri, numărul de variabile va coincide cu numărul de variante de răspuns, toate fiind dihotomice cu valori, de exemplu, *1 − da, 0 − nu,*. Menţionăm, suplimentar, că şi într-un caz, şi în altul numărul de răspunsuri înregistrate depăşeşte numărul indivizilor chestionaţi. Acest lucru permite a calcula două tipuri de frecvenţe relative: faţă de numărul total de răspunsuri şi faţă de numărul de respondenţi. Exemplul de mai jos (întrebare cu 4 variante de răspuns, dintre care se cer a fi date numai 3, numărul de respondenţi − 5) demonstrează cele spuse (a se vedea Tabelul 10.1):

Tabelul 10.1

Răspunsuri	**Frecvenţe absolute**	**Frecvenţe relative**	
		Faţă de numărul de respondenţi (5)	**Faţă de numărul de răspunsuri (15)**
1 − 1 2 4	1 − de 2 ori	1 − 40%	1 − 13,3%
2 − 2 3 4	2 − de 4 ori	2 − 80%	2 − 26,7%
3 − 2 3 4	3 − de 4 ori	3 − 80%	3 − 26,7%
4 − 1 3 4	4 − de 5 ori	4 − 100%	4 − 33,3%
5 − 2 3 4	**Total − 15 răspunsuri**	**Total − 300%**	**Total − 100%**

În programul SPSS variabilele ce corespund întrebărilor cu răspunsuri multiple pot fi prelucrate pe două căi. În ambele cazuri se definesc aşa-numitele *seturi de variabile*, care în continuare participă la prelucrare (determinare frecvenţe, construire tabele de asociere etc.), asemănător cu variabilele obişnuite.

O primă cale constă în definirea seturilor de variabile prin comanda **Analyze** → *Multiple Response* → *Define Sets...* Lansarea acestei comenzi conduce la afişarea unei casete de dialog, prin care şi se definesc viitoarele seturi de variabile (a se vedea Figura 10.1).

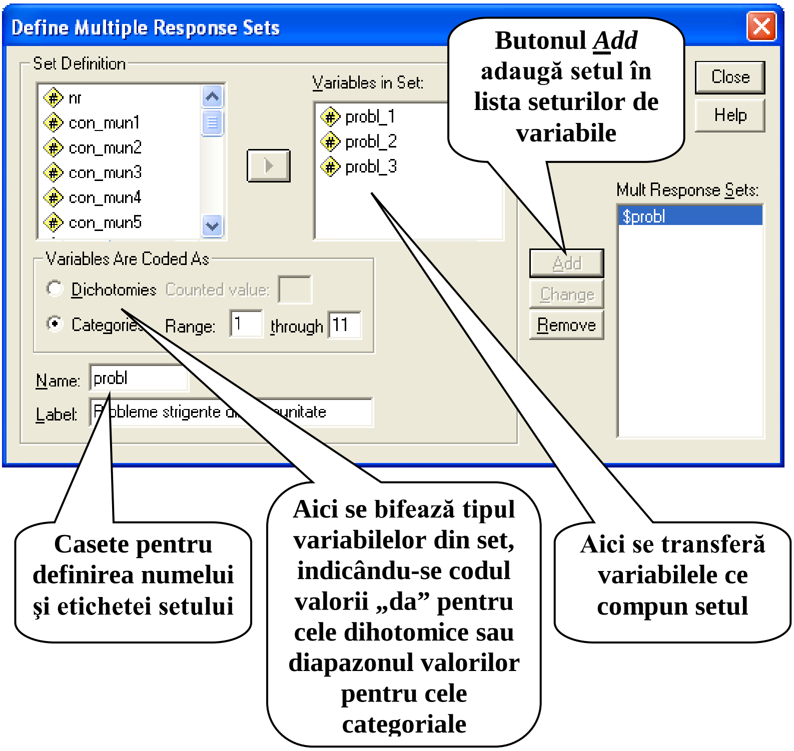


Fig. 10.1. Caseta de definire a seturilor de variabile

Determinarea frecvenţelor absolute şi relative pentru seturile definite de variabile se face prin comanda **Analyze** → *Multiple Responce* → *Frequencies...* Rezultatul afişat de calculator are forma prezentată în Figura 10.2 (este cazul unui set de variabile categoriale). În acest rezultat sunt afişate (de la stânga la dreapta): valorile variabilelor din set, codurile acestor valori, frecvenţele absolute, frecvenţele relative, calculate faţă de numărul total de răspunsuri ale respondenţilor, şi frecvenţele relative, calculate faţă de numărul de respondenţi.

Multiple Response

Group $PROBL Probleme strigente din comunitate

Category label	Code	Count	Pct of Responses	Pct of Cases
abandonul batranilor	1	94	6,6	18,4
abuzul de alcool	2	193	13,5	37,8
violenta in familie	3	101	7,1	19,8
abandonul copiilor	4	73	5,1	14,3
indiferenta APL	5	28	2,0	5,5
migratia populatiei	6	215	15,1	42,1
saracia	7	297	20,8	58,1
sanatatea precara a populatiei	8	109	7,6	21,3
lipsa serviciilor de asistenta sociala	9	76	5,3	14,9
somajul	10	211	14,8	41,3
altele	11	29	2,0	5,7
	Total responses	1426	100,0	279,1

5 missing cases; 511 valid cases

Fig. 10.2. Frecvenţele valorilor seturilor de variabile

Observaţia 10.1. Seturile de variabile definite şi utilizate prin meniul **Analyze** → *Multiple Response* ▸ se păstrează numai pe durata secvenţei de lucru cu baza de date. La închiderea bazei de date ele dispar, nu se salvează.

O altă cale de definire a seturilor de variabile se găsește în componența casetei de dialog de construire a tabelelor de asociere. Acţionarea butonul de comandă **Mult Response Sets...** din colţul stâng-jos al casetei de dialog conduce la afişarea unei alte casete de dialog, asemănătoare cu cea prezentată în Figura 10.1. În continuare definirea seturilor de variabile se face asemănător cu definirea seturilor explicată anterior, numai că ea se finalizează prin acţionarea butonului de comandă $\boxed{\textbf{Save}}$ (acest buton înlocuieşte butonul $\boxed{\textbf{Close}}$ din caseta de dialog din Figura 10.1). Setul de variabile definit astfel se salvează şi se păstrează pentru orice alte secvenţe de lucru cu baza de date.

Observaţia 10.2. Seturile de variabile, definite prin **Analyze**→ *Custom Tables,* pot fi utilizate numai pe loc, la construirea tabelelor de asociere respective. Ele se transferă în casetele *Rows* sau *Columns* asemănător variabilelor obişnuite, adăugându-li-se şi statisticile necesare.

Observaţia 10.3. Frecvenţele, necesar a fi calculate pentru valorile seturilor de variabile (faţă de numărul de respondenţi sau faţă de numărul de răspunsuri), se indică la definirea seturilor prin bifarea butoanelor de opţiune respective, situate în partea de jos a casetei de dialog *Define Multiple Response Sets.*

Observaţia 10.4. Pentru a obţine un rezultat asemănător celui din Figura 10.2 setul de variabile, definit prin **Analyze**→ *Custom Tables,* se transferă în caseta *Rows*, adăugându-i-se statisticile respective (*Col%*).

Exerciții, întrebări de control

1. Formulați câte trei întrebări cu trei variante de răspuns și cu orice număr de răspunsuri la tema de cercetare „Timpul liber al studentului”. Câte variabile generează întrebările formulate? Care vor fi etichetele și valorile acestor variabile?
2. O întrebare din chestionar solicită până la patru răspunsuri. Dacă se calculează frecvențele procentuale ale răspunsurilor față de numărul total de respondenți, pe de o parte, iar pe de alta – față de numărul total de răspunsuri, care va fi suma maximală a lor în ambele cazuri?
3. O întrebare cu 10 variante de răspuns permite orice număr de răspunsuri de la respondenți. Dacă se calculează frecvențele procentuale ale răspunsurilor față de numărul total de respondenți, pe de o parte, iar pe de alta – față de numărul total de răspunsuri, care va fi suma maximală a lor în ambele cazuri?

4. Prin ce se deosebesc seturile de variabile categoriale de cele dihotomice?
5. În SPSS 11.0 seturile de variabile pot fi definite prin câteva comenzi. Când și cum pot fi determinate frecvențele absolute și cele procentuale ale setului de variabile atât față de numărul de respondenți, cât și față de numărul de răspunsuri?
6. La o întrebare cu șase variante de răspuns respondenților li s-a cerut să marcheze cel mult trei răspunsuri. Rezultatele obținute de la 12 respondenți sunt prezentate în tabelul de mai jos, în care sunt indicate numerele variantelor de răspuns (codurile răspunsurilor) marcate de respondenții respectivi:

Respondent	Răspunsul I	Răspunsul II	Răspunsul III
1	2	4	6
2	1	2	
3	2	3	6
4	5		
5	1	2	4
6	1	6	
7	2	6	
8	3	4	6
9	4	5	6
10	2	4	6
11	3	6	
12	4	5	6

Să se calculeze frecvențele absolute, frecvențele procentuale față de numărul respondenților și frecvențele procentuale față de numărul răspunsurilor. Rezultatele să se organizeze sub formă de tabel (asemănător celui din Figura 10.2), la care să se adauge totalurile pe coloane.
7. Ce tipuri de diagrame pot fi utilizate pentru reprezentarea grafică a frecvențelor seturilor de variabile? Argumentați răspunsul prin exemple concrete.

Tema 11

Gestiunea cazurilor în *SPSS*

Amintim că baza de date din SPSS are forma unui tabel ce se păstrează în foaia *Data View*. În acest tabel coloanele corespund caracteristicilor indivizilor şi sunt nu altceva decât variabilele (*Variables*), iar liniile corespund indivizilor, purtând denumirea de cazuri (*Cases*). Şi cu unele, şi cu altele în SPSS pot fi executate un şir de operaţii, care au scopul:

- de a completa baza de date cu cazuri şi variabile suplimentare;
- de a efectua analize mai profunde ale fenomenelor cercetate prin divizarea populaţiei cercetate după una sau câteva caracteristici, prin selectarea şi studierea numai a unei părţi a populaţiei, prin construirea şi introducerea de noi caracteristici etc.;
- de a verifica şi corecta suplimentar datele;
- de a corecta eşantionul în scopul asigurării reprezentativităţii lui etc.

În SPSS majoritatea acestor operaţii se execută cu ajutorul comenzilor din meniurile **Data** şi **Transform.**

În acest compartiment, vom examina un şir de *operaţii cu cazurile* din baza de date.

I. Sortarea cazurilor

Operaţia de sortare a cazurilor poate fi folosită pentru:

- aranjarea cazurilor în ordine crescătoare sau descrescătoare după una sau mai multe variabile;
- verificarea valorilor extreme ale caracteristicilor şi detectarea de valori ieşite din domeniile de valori ale variabilelor;

- aranjarea compactă a cazurilor pentru care nu au fost introduse date pentru unele variabile (nonrăspunsuri) în scopul verificării lor suplimentare.

Lansarea comenzii de sortare a cazurilor (**Data**→ *Sort Cases...)* conduce la afişarea casetei de dialog *Sort Cases*, prin care, în continuare, se fac setările respective (sau necesare) de sortare (a se vedea Figura 11.1).

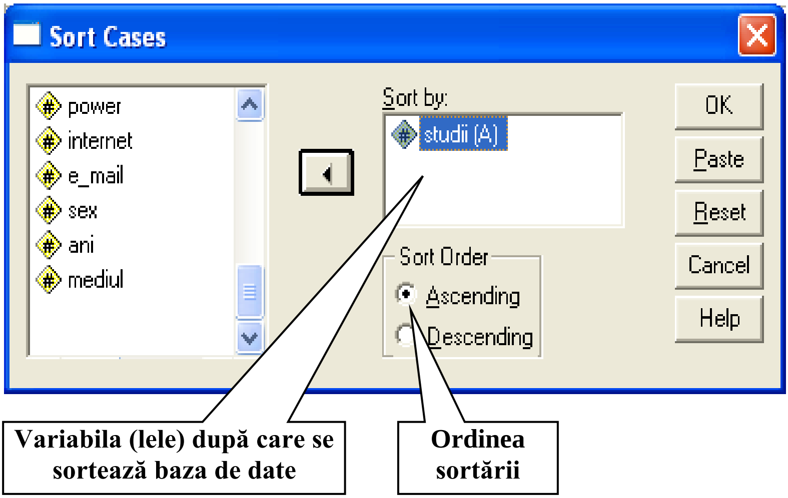


Fig. 11.1. Caseta de dialog *Sort Cases*

Sortarea în creştere a bazei de date rezultante după numărul de ordine al chestionarelor permite a detecta astfel de greşeli, cum ar fi: introducerea multiplă a unuia şi aceluiaşi chestionar (dublarea cazurilor) sau neintroducerea unor chestionare.

II. Adăugarea cazurilor la baza de date

Baza de date din SPSS poate fi completată cu cazuri noi, luate din alte baze de date, identice după structură (acelaşi număr, consecutivitate şi proprietăţi ale variabilelor). Această operaţie se utilizează cel mai frecvent pentru a aduna împreună datele introduse de mai mulţi operatori.

Adăugarea de cazuri noi la baza de date se face prin lansarea comenzii **Data**→ *Merge File*→ *Add Cases...*, care conduce la afişarea casetei de dialog *Add Cases: Read File* (a se vedea Figura 11.2). În continuare, se solicită baza de date cu cazurile necesare a fi adăugate; se finalizează operaţia prin acţionarea butonului de comandă **Open**.

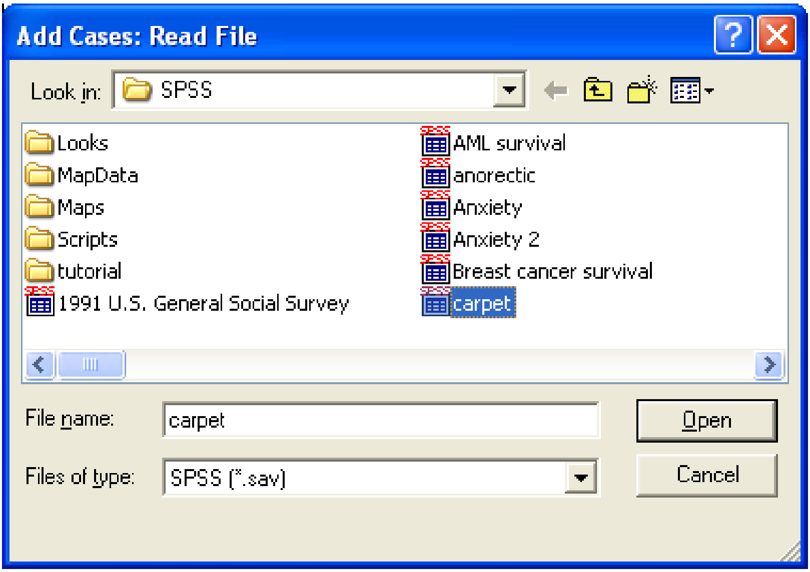


Fig. 11.2. Caseta de dialog pentru adăugarea cazurilor la baza de date

III. Divizarea bazei de date

Divizarea bazei de date are scopul de a descompune baza de date în mai multe subbaze după valorile unei caracteristici (variabile) sau după toate combinaţiile de valori a mai multor caracteristici (variabile). În continuare, toate prelucrările de date (frecvenţe, indicatori, tabele etc.) se efectuează separat pentru fiecare subbază sau subgrup de cazuri. Astfel, apare posibilitatea comparării rezultatelor pentru diferite grupuri de cazuri (indivizi) sau de a obţine rezultatele respective pentru unele grupuri de cazuri (indivizi).

În cazul întrebărilor de control, rezultatele obţinute în urma divizării bazei de date pot ajuta la verificarea sincerităţii răspunsurilor sau la depistarea fraudelor.

Consecutivitatea operaţiilor pentru divizarea bazei de date este următoarea:

1. Se acţionează comanda **Data**→ *Split File...*
2. În caseta de dialog *Split File* (a se vedea Figura 11.3) se bifează butonul de opţiune *Compare groups* pentru a obţine prelucrările ulterioare într-un singur tabel sau *Organize output by groups* pentru a obţine tabele separate pentru fiecare subgrup de cazuri (indivizi).

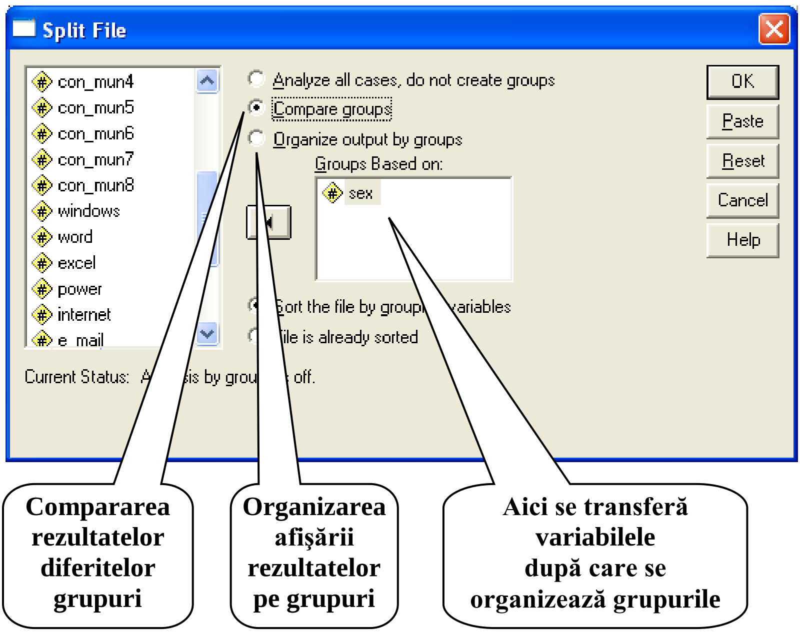


Fig. 11.3. Caseta de dialog *Split File*

3. În zona *Groups Based on:* se transferă variabila (variabilele) după valorile cărora se doreşte divizarea bazei de date şi se acţionează butonul $\boxed{\textbf{OK}}$.

Observaţia 11.1. Anularea divizării bazei de date se face prin lansarea comenzii **Data**→ *Split File*...şi bifarea butonului de opţiune *Analyze all cases, do not organize groups.*

IV. Selectarea cazurilor

Selectarea cazurilor reprezintă o procedură de selectare din baza de date a unei subbaze (subpopulaţii) în conformitate cu condiţiile formulate de utilizator. În continuare toate operaţiile de prelucrare se efectuează numai cu datele din subbaza selectată. Această procedură se foloseşte, de exemplu, pentru a determina diferiţi indicatori, frecvenţe etc. pentru subpopulaţia selectată.

Selectarea (deselectarea) cazurilor se realizează prin comanda **Data** → *Select Cases*... Caseta de dialog ce se afişează în urma lansării acestei comenzi e demonstrată în Figura 11.4.

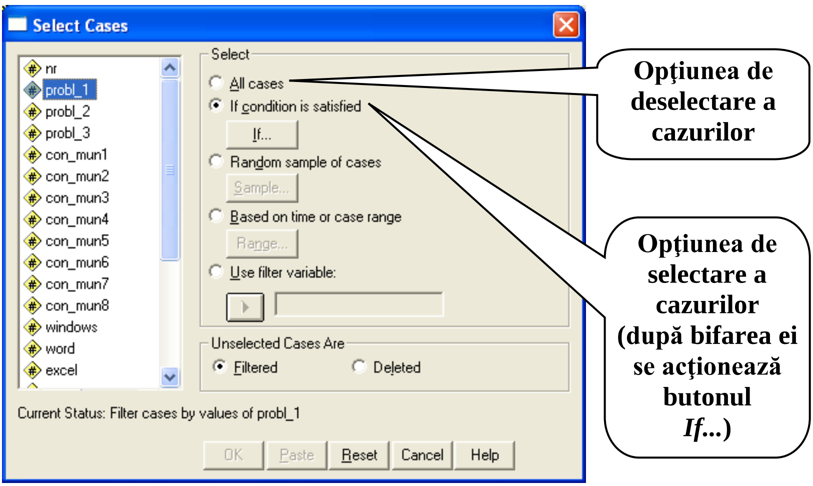


Fig. 11.4. Caseta de dialog *Select Cases*

La acţionarea butonului de comandă **If...** apare o altă casetă de dialog (a se vedea Figura 11.5), în care se introduce condiţia de selectare a cazurilor – o expresie logică. Acţionarea, în continuare, a butoanelor **Continue** şi **OK** exclude din baza de date acele cazuri, care nu satisfac condiţia de selectare (numerele liniilor în baza de date apar tăiate, însă datele nu dispar!), iar prelucrările ulterioare se fac cu cazurile rămase.

Fig. 11.5. Caseta de dialog *Select Cases: If*

Observaţia 11.2. Restabilirea bazei de date (deselectarea cazurilor) se face prin comanda **Data** → *Select Cases...*, urmată de bifarea butonului de opţiune **All cases**.

În continuare, vom defini şi analiza noţiunii de expresie logică, necesară în SPSS într-un şir de situaţii de gestiune a bazei de date (selectarea cazurilor, construirea variabilelor etc.).

Def. 11.1. Se numeşte ***expresie logică*** expresia formată din *condiţii*, legate între ele prin operatorii logici & („şi"), | („sau"), ~ („nu").

Def. 11.2. ***Condiţia*** reprezintă o construcţie de forma:

A <*semn de comparare*> **B**,

unde **A** şi **B** reprezintă expresii aritmetice ce conţin constante, variabile şi funcţii, iar semnul de comparare poate fi: = („egal”), ~= („aproximativ”), < („mai mic”, <= („mai mic sau egal”), > („mai mare”), >= („mai mare sau egal).

Menţionăm că în expresiile logice (ca şi în cele aritmetice) pot fi utilizate paranteze simple.

Vom aduce un exemplu de compunere a expresiilor logice.

Exemplul 8.1. Fie 3 variabile ce caracterizează o populaţie:

- **sex**={1 – femeie, 2 – bărbat} – sexul individului (variabilă nominală);
- **ani** − vârsta individului în ani întregi cu valori, de exemplu, de la 18 până la 80 de ani (variabilă numerică);
- **comp**={1 – deloc, 2 – slab, 3 – mediu, 4 – bine, 5 – excelent} – nivelul de cunoaştere de către individ a calculatorului (variabilă ordinală).

Atunci expresiile logice de mai jos definesc următoarele subpopulaţii:

sex=1 – indivizi de sex feminin (femei);

sex=1 & ani >=57 – femei de vârstă pensionară;

sex=2 & ani >=62 − bărbaţi de vârstă pensionară;

ani < 30 – tineret (indivizi cu vârsta sub 30 de ani);

(sex=1 & ani >=57) | (sex=2 & ani >=62) − indivizi de vârstă pensionară;

ani < 30 & (comp=2 | comp=1) – tineri ce cunosc slab sau deloc calculatorul;

(sex=1 & ani >=57 | sex=2 & ani >=62) & comp>3 – indivizi de vârstă pensionară ce cunosc bine şi excelent calculatorul.

V. Ponderarea bazei de date

Reamintim că la construirea eșantionului prin stratificare se poate întâmpla ca el să nu respecte structura populației din care a fost extras și, respectiv, să nu fie reprezentativ (a se vedea Tema 6). Salvarea situației sau „repararea” eșantionului neproporțional și transformarea lui în unul proporțional se găsește în determinarea coeficienților de ponderare și ponderarea cu ajutorul lor a bazei de date. Vom demonstra acest lucru în cazul unei baze de date din SPSS.

Presupunem că se realizează o cercetare într-o populație stratificată după caracteristica *sex* (femei, bărbați) și se cunoaște repartizarea straturilor în populație: 52% – femei, 48% – bărbați. După culegerea și introducerea datelor s-au calculat frecvențele variabilei *sex* din eșantion (a se vedea Figura 11.6). Este clar că eșantionul nu respectă condiția de reprezentativitate după caracteristica *sex*: în eșantion avem 58,2% – femei și 41,8% – bărbați, valori ce diferă de cele din populație. Astfel ajungem la situația în care este necesar a calcula coeficienții de ponderare și a pondera baza de date.

Sexul respondentului

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	femeie	1981	58.2	58.2	58.2
	barbat	1424	41.8	41.8	100.0
	Total	3405	100.0	100.0	

Fig. 11.6. Frecvenţele variabilei *sex* până la ponderare

Vom calcula coeficienții de ponderare după formulele (6.5) cu ajutorul comenzii **Transform** → *Compute…* (a se vedea Figura 11.7). Valorile acestor coeficienți vor forma o variabilă numerică nouă, suplimentară, în baza de date, având numele *pond_sex*, atribuit în procesul construirii variabilei. (Modalitatea utilizării comenzii **Transform** → *Compute…* va fi examinată pe larg în Tema 12).

În continuare, cu ajutorul comenzii **Data** → *Weight Cases…*, vom pondera baza de date, trecând variabila *pond_sex* în câmpul Weight Cases by și acționând butonul OK (a se vedea Figura 11.8).

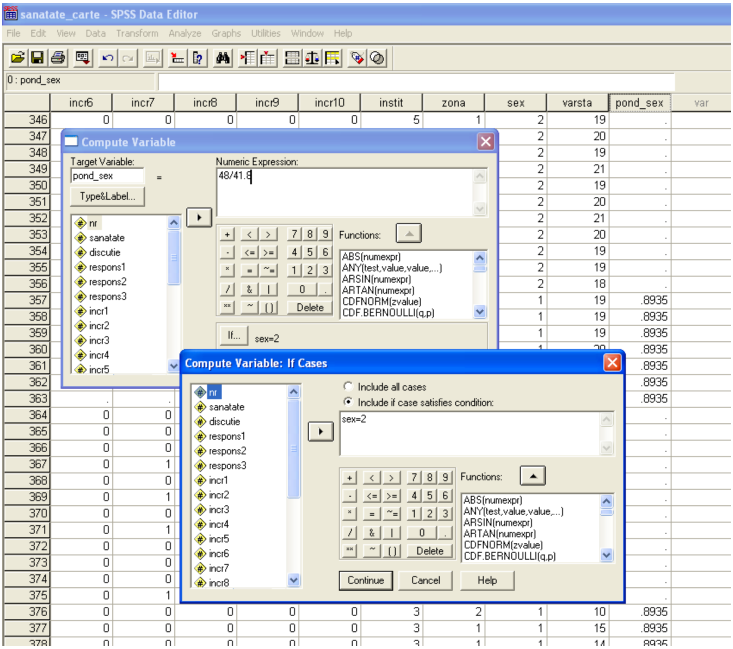


Fig. 11.7. Calcularea ponderilor în SPSS

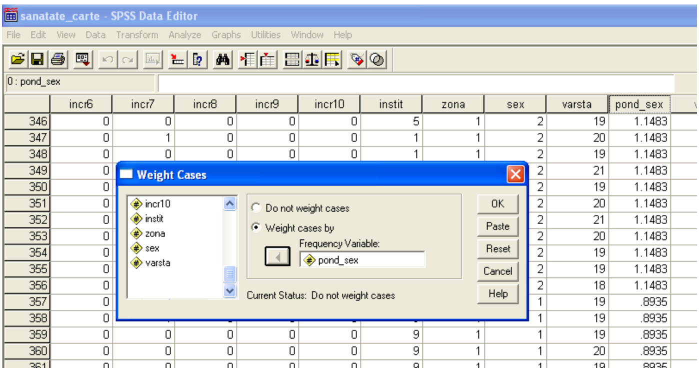


Fig. 11.8. Ponderarea bazei de date prin Data→ ***Weight Cases…***

În baza de date ponderată frecvențele procentuale ale variabilei *sex* din eșantion coincid cu cele din populație, ce demonstrează faptul că eșantionul este reprezentativ după caracteristica *sex* (a se vedea Figura 11.9).

Sexul respondentului

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	femeie	1770	52.0	52.0	52.0
	barbat	1635	48.0	48.0	100.0
	Total	3405	100.0	100.0	

Fig. 11.9. Frecvența variabilei *sex* după ponderare

Aprecierea sănătății (până la ponderarea datelor)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excelenta	313	9.2	9.4	9.4
	Buna	1786	52.5	53.4	62.8
	Satisfacatoare	677	19.9	20.3	83.0
	Rea	140	4.1	4.2	87.2
	Nu stiu/nu o pot aprecia	427	12.5	12.8	100.0
	Total	3343	98.2	100.0	
Missing	System	62	1.8		
Total		3405	100.0		

Aprecierea sănătății (după ponderarea datelor)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excelenta	327	9.6	9.8	9.8
	Buna	1804	53.0	54.0	63.7
	Satisfacatoare	663	19.5	19.8	83.6
	Rea	137	4.0	4.1	87.7
	Nu stiu/nu o pot aprecia	412	12.1	12.3	100.0
	Total	3343	98.2	100.0	
Missing	System	62	1.8		
Total		3405	100.0		

Fig. 11.10. Efectul ponderării bazei de date

În Figura 11.10 se compară frecvențele unei variabile (aprecierea propriei sănătăți de către respondenți) până și după ponderarea bazei de date. Se observă o modificare a tuturor tipurilor de frecvențe, cauzată de „redistribuirea” indivizilor din eșantion după caracteristica *sex* ca rezultat al ponderării datelor (numărul femeilor s-a micșorat de la 1.981 la 1.770, iar al bărbaților – s-a mărit de la 1.424 la 1.635).

În cazul tabelelor de asociere, ponderarea datelor va avea efect asupra frecvențelor calculate pe coloane (Col%) și nu va afecta frecvențele calculate pe linii (Row%).

Exerciții, întrebări de control

1. Cum ar putea fi folosită comanda de sortare a cazurilor pentru verificarea datelor?
2. Datele culese într-un sondaj au fost introduse de câțiva operatori, o parte din ei folosind programul SPSS, alta – programul Excel. Propuneți metoda de adunare într-o singură bază de date SPSS a tuturor datelor introduse. Cum s-ar putea verifica, dacă au fost introduse datele din toate chestionarele și dacă unele n-au fost introduse de câteva ori?
3. Poate o bază de date să fie divizată în 36 de subbaze de date? În caz de răspuns afirmativ, propuneți câteva variante cu variabile concrete.
4. O populaţie este descrisă prin următoarele caracteristici:
   sex={1 – femeie, 2 – bărbat}; *varsta* (în ani întregi);
   sta_civ={1 – căsătorit, 2 – necăsătorit};
   auto={1 – posedă automobil, 0 – nu posedă automobil}.
   Să se compună expresia logică care să definească subpopulaţia: *femei necăsătorite cu vârsta până la 30 de ani inclusiv ce nu posedă automobil împreună cu bărbaţi căsătoriţi cu vârsta de la 35 la 40 de ani inclusiv ce posedă automobil.*
5. Următoarea expresie folosește variabilele din pct.4:
   (sex=1 | auto=1) & varsta > 20 & varsta < 25 | varsta =60 & sta_civ = 2
   Descrieți populația selectată prin ea.
6. Să se pondereze baza de date din pct.3 (Tema 9) după mediul de reședință, știind că în toată populația cercetată numărul indivizilor de la oraș este egal cu cel al indivizilor de la sat. Să se compare grafic opiniile față de concubinaj, determinate fără/și cu ponderarea datelor.

Tema 12

Gestiunea variabilelor în *SPSS*

Dacă operaţiile cu cazurile din baza de date conduc la schimbarea ordinii liniilor, adăugarea unor noi linii, sau excluderea din examinare a unora din ele, atunci operaţiile cu variabilele modifică coloanele bazei de date (atât numărul, ordinea, cât şi conţinutul lor). Vom analiza câteva din aceste operaţii.

I. Construirea variabilelor noi prin calculare

Construirea variabilelor noi se utilizează pentru determinarea unor caracteristici ale indivizilor ce n-au fost culese direct din populaţie (li se mai spune − *caracteristici derivate* sau *auxiliare*), şi care, la rândul lor, conduc la diversificarea rezultatelor obţinute prin prelucrarea acestor caracteristici.

De exemplu, într-o cercetare au fost înregistrate notele la examene dintr-o sesiune a unei grupe de studenţi. O caracteristică suplimentară a studenţilor ar fi nota medie de la sesiune, care nu se culege, dar poate fi calculată uşor având pentru fiecare student notele la toate examenele din sesiunea respectivă (pentru fiecare student, nota medie reprezintă media aritmetică a notelor obţinute la toate examenele). În continuare, aceste medii de la sesiune pot fi utilizate pentru a analiza însuşita studenţilor în funcție de alte caracteristici, a compara însuşita grupei cu însuşita altor grupe etc.

Procedura de construire a variabilelor noi constă din două etape:

1) *calcularea valorilor* noii variabile prin una din opţiunile meniului **Transform** (valori calculate apar într-o coloană suplimentară în foaia *Data View*);
2) *definirea* noii variabile prin introducerea proprietăţilor ei în foaia *Variable View*.

Una dintre modalităţile de calculare a variabilelor noi o oferă comanda **Transform** → *Compute...*

Consecutivitatea paşilor în acest caz este următoarea:

1. Se lansează comanda **Transform** → *Compute...* Drept rezultat, calculatorul afişează caseta de dialog *Compute Variable* (a se vedea Figura 12.1).

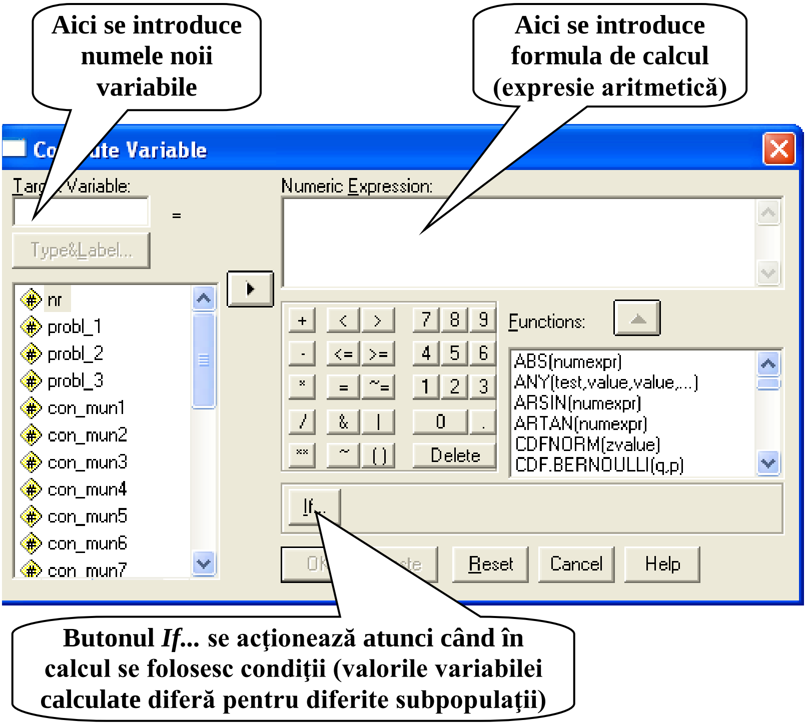


Fig. 12.1. Caseta de dialog *Compute Variable*

2. În caseta de text *Target Variable:* se introduce numele variabilei noi (atenţie, el nu trebuie să coincidă cu numele altor variabile din baza de date!).
3. În caseta de text *Numeric Expresiuon:* se scrie expresia aritmetică după care se calculează valorile variabilei. (Observăm că în caseta de dialog sunt prezente toate

accesoriile pentru culegerea unei expresii aritmetice: cifre, semne ale operaţiilor aritmetice, funcţii.)

4. Dacă variabila nouă primeşte valori diferite pentru diferite subpopulaţii, atunci se acţionează butonul **If...**, se bifează butonul de opţiune *Include if case satisfies condition:*, după care se introduce expresia logică de selectare a subpopulaţiei (a se vedea Figura 12.2).
5. Se acţionează consecutiv butoanele **Continue** şi **OK** (sau numai **OK**, dacă procedura se limitează la pasul 3).

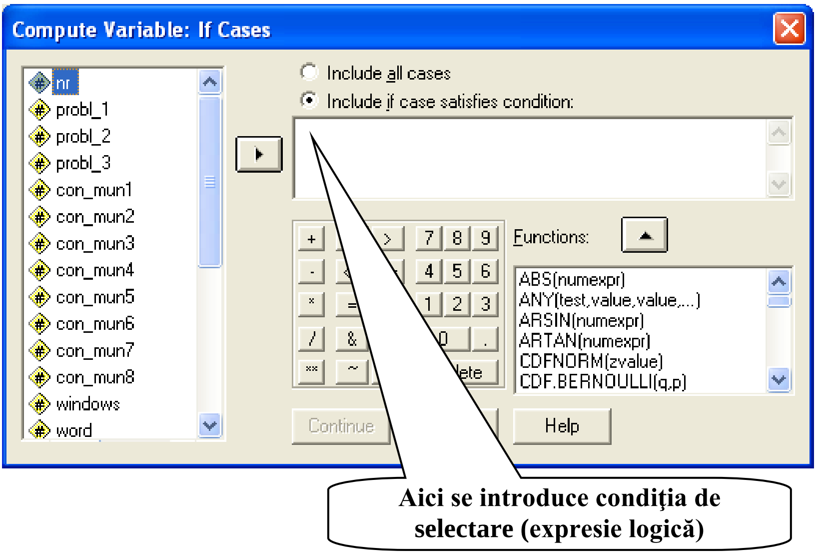


Fig. 12.2. Caseta de dialog *Compute Variable: If Cases*

Observaţia 12.2. Dacă variabila calculată primeşte valori diferite pentru diferite subpopulaţii, paşii 1, 3-5 de calculare a valorilor ei se repetă pentru fiecare subpopulaţie.

Observaţia 12.3. Amintim că după calcularea variabilei prin această metodă se trece la foaia *Variable View* pentru a-i atribui o etichetă, valori (dacă ea nu este numerică), alte proprietăţi.

II. Construirea variabilelor noi prin recodificare

Această operaţie, de regulă, se utilizează, atunci când e necesar a construi o nouă variabilă folosind valorile altei variabile existente în baza de date (o vom numi *variabilă sursă*). Spre exemplu, prin această metodă poate fi transformată vârsta indivizilor exprimată în ani întregi (variabilă numerică) în vârstă pe grupe de vârstă (variabilă ordinală).

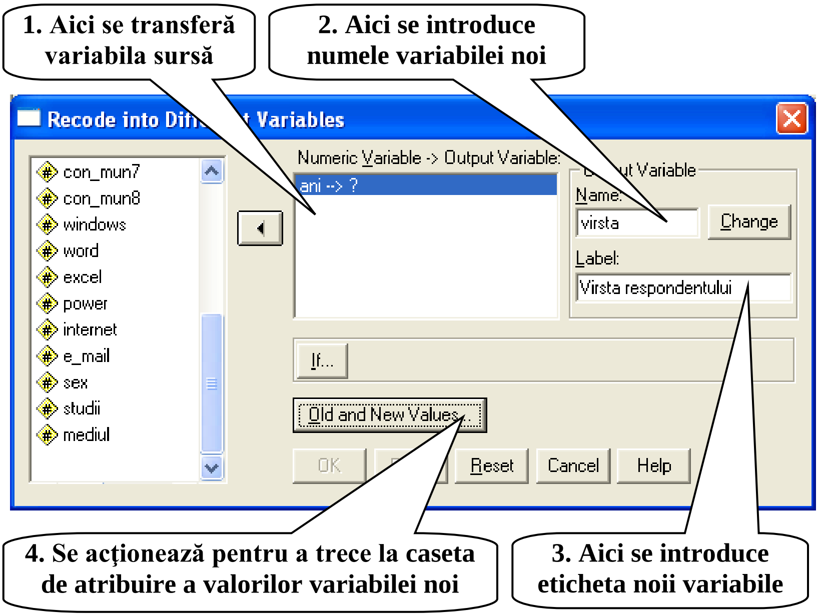


Fig. 12.3. Caseta de dialog *Recode into Different Variables*

Consecutivitatea paşilor de calculare a valorilor noii variabile prin această metodă este următoarea:

1. Se lansează comanda **Transform**→ *Recode*→ *Into Different Variable...* Ca rezultat calculatorul afişează caseta de dialog *Recode into Different Variables* (a se vedea Figura 12.3).

2. În caseta *Numeric Variable* se transferă variabila sursă.
3. În casetele *Name:* şi *Label:* se introduc numele şi eticheta variabilei noi.
4. Se acţionează butonul **Old and New Values...** La ecran apare o nouă casetă de dialog, numită *Old and New Values* (a se vedea Figura 12.4)

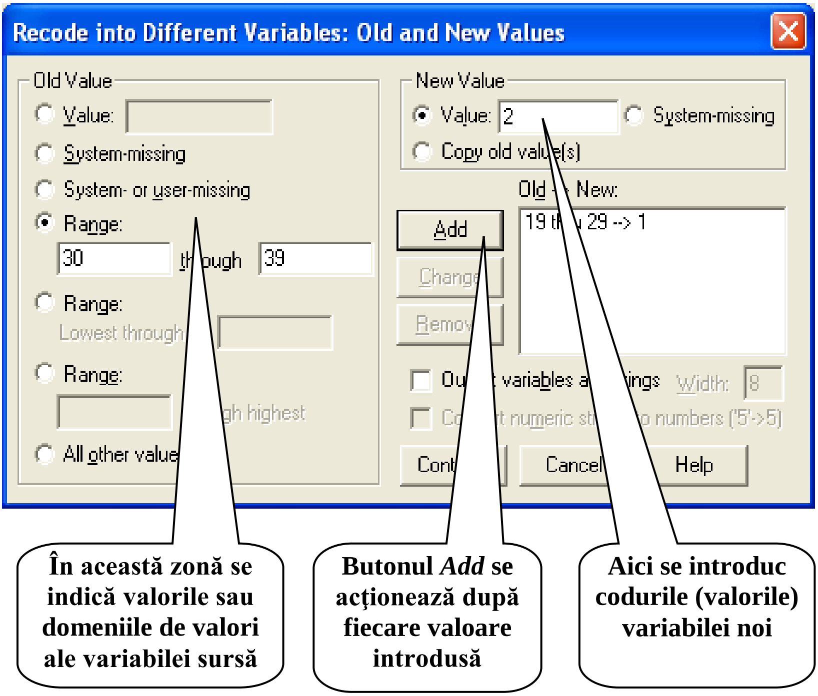


Fig. 12.4. Atribuirea de valori variabilei noi

5. În continuare, în jumătatea din stânga a casetei de dialog *Old and New Values*, se introduc valorile sau domeniile de valori ale variabilei sursă, iar în jumătatea din dreapta – se înscriu codurile valorilor variabilei noi. Butonul de comandă **Add** se acţionează după fiecare atribuire de valori variabilei noi.

6. Lucrul se finalizează cu acţionarea consecutivă a butoanelor $\boxed{\textbf{Continue}}$ → $\boxed{\textbf{Change}}$ → $\boxed{\textbf{OK}}$.

Observaţia 12.4. Spre deosebire de metoda precedentă de calculare a variabilelor noi (**Transform** → *Compute...)*, prin care o variabilă nouă poate fi construită din una sau câteva variabile sursă, prin comanda **Transform**→ *Recode* variabila nouă se construieşte dintr-o singură variabilă sursă.

Observaţia 12.5. Recodificarea valorilor în aceeaşi variabilă (comanda **Transform**→ *Recode* → *Into Same Variable...*) conduce la modificarea ireversibilă a variabilei sursă. Se recomandă de a fi utilizată numai în cazul când variabila sursă nu va mai fi folosită în varianta iniţială.

Observaţia 12.6. În SPSS există şi alte posibilităţi de introducere (construire) a variabilelor noi. Printre acestea menţionăm:

- **Transform**→ *Categorize Variables...* − divizează valorile variabilei sursă pe intervale de valori, codificându-le automat cu 1, 2, 3,... pentru variabila nouă. Numărul de categorii se indică de către utilizator, iar în fiecare categorie calculatorul plasează aproximativ acelaşi număr de cazuri din baza de date.
- **Transform**→*Automatic Recode...* − formează din variabila sursă una nouă, recodificând cu 1, 2, 3,... valorile variabilei sursă aranjate în creştere sau descreştere.
- **Transform**→ *Count...* − formează o variabilă nouă, ale cărei valori reprezintă numărul valorilor de acelaşi fel, întâlnite într-un caz (la un individ) pentru variabilele sursă indicate de către utilizator.

III. Adăugarea la baza de date a variabilelor din alte baze de date

Această operaţie permite a completa baza de date cu variabile din alte baze de date din calculator. Ea poate fi folosită, de exemplu, în cazul când baza de date se elaborează de câteva persoane pentru a aduna împreună variabilele definite de acestea.

Paşii, care se cer a fi întreprinşi pentru adăugarea de variabile la baza de date curentă, sunt următorii:

1. Se lansează comanda **Data** → *Merge File* → *Add Variables...* Ca rezultat, programul afişează o casetă de dialog de tipul celei din Figura 11.2, prin care se solicită baza de date „donatoare”.

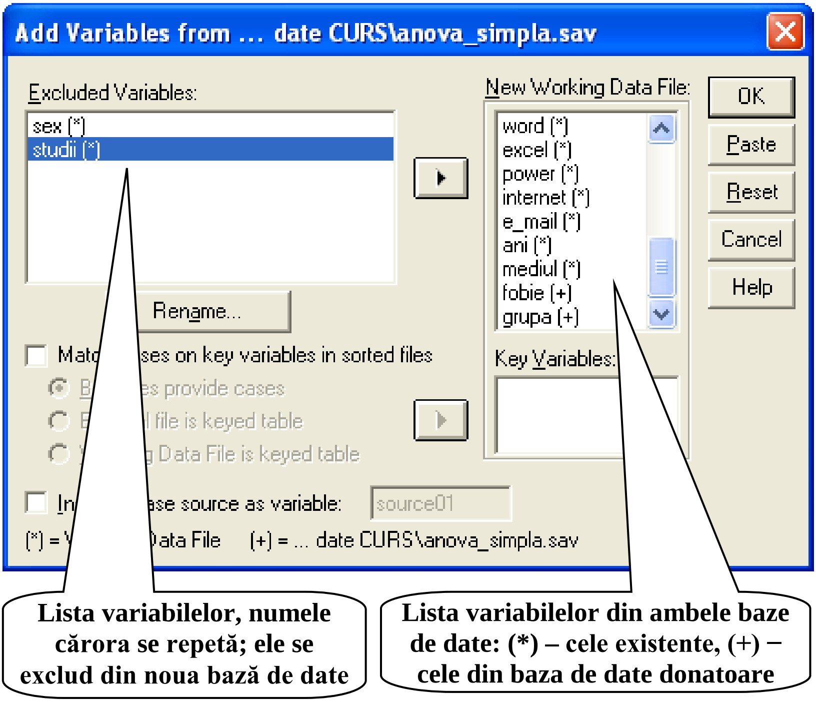


Fig. 12.5. Adăugarea de variabile la baza de date

2. Se acţionează butonul $\boxed{\textbf{Open}}$, care deschide o altă casetă de dialog: *Add Variables from...* (a se vedea Figura 9.5). Variabilele din ambele baze de date apar în lista *New Working Data File:*, iar dintre cele ce se dublează − câte un exemplar în lista *Excluded Variables:* .

3. Se transferă din lista *New Working Data File:* în lista *Excluded Variables:* variabilele, de care nu este nevoie sau care nu se vor transfera, în prima listă rămânând cele ce se vor adăuga la baza de date. (Observăm că în lista *New Working Data File:* variabilele existente în baza de date curentă sunt marcate cu semnul (*), iar cele din baza de date „donatoare” − cu semnul (+)).
4. Se acţionează butonul **OK**. Drept rezultat, la baza de date curentă vor fi adăugate din baza de date „donatoare” variabilele solicitate.

Exerciții, întrebări de control

1. În Tabelul 12.1 este adusă informația parțială despre notele studenților unei facultăți la diferite discipline (două testări – *test1* și *test2*, media notelor din auditoriu – *aud*, nota pentru lucrul individual – *indiv* și anii de studii – *an*).

Tabelul 12.1

nr	an	test1	test2	aud	indiv	media	reusita	rest
1	1	4	9	2	6			
2	2	2	6	8	4			
3	2	6	4	6	10			
4	1	4	8	4	2			
5	3	1	8	8	4			
6	1	8	7	9	4			
7	2	6	10	2	7			
8	1	6	8	4	8			
9	3	6	3	6	4			
10	2	8	8	3	4			
11	1	9	10	10	6			
12	1	10	4	10	8			
13	1	10	6	6	8			
…	…	...	...	....	...			

a) Să se construiască variabila *media*, calculată după formula:

*media = (test1 + test2)/2 * 0,3 + aud * 0,3 + indiv * 0,4*

numai pentru studenții ce au cele patru note mai mari sau egale cu 5. Rezultatul să se rotunjească până la întreg. Pentru studenții care au cel puțin o notă mai mică decât 5 – câmpul *media* rămâne necompletat (gol).

b) Câmpului *reusita* i se vor atribui următoarele valori:

1 – eminent, dacă *media* = 9 sau *media* = 10;

2 – reușită medie, dacă *media* = 7 sau *media* = 8;

3 – reușită joasă, dacă *media* = 5 sau *media* = 6;

4 – restanțier, dacă câmpul *media* este gol.

c) În câmpul *rest* se va calcula și se va include numărul de restanțe (numărul de note ale studentului mai mici ca 5). Pentru studenții nerestanțieri în câmpul *rest* se va pune 0.

2. Cum pot fi mutate variabilele dintr-un loc în altul într-o bază de date SPSS?

3. Cum pot fi create copii ale variabilelor în una și aceeași bază de date SPSS?

4. Aduceți exemple de situații, în care la baza de date deschisă se adaugă variabile dintr-o altă bază de date.

Tema 13

Corelația și regresia datelor

În viață există fenomene, situații etc., care depind unele de altele. La fel, în cazul variabilelor se întâmplă ca și ele să depindă unele de altele, modificarea valorilor unora să conducă la modificarea valorilor altora sau, cum se mai spune, variabilele să coreleze.

Def. 13.1. ***Corelația*** poate fi definită ca 1) legătură reciprocă între lucruri sau fenomene; 2) relație în care unul dintre termeni nu poate exista fără celălalt sau 3) dependență reciprocă între două procese sau fenomene*.

Def. 13.2. Prin ***corelație statistică*** se înțelege intensitatea și direcția legăturii statistice dintre două sau mai multe variabile.

Legătura dintre variabile de diferite tipuri este observată, în particular, în cazul tabelelor de asociere a variabilelor, analizate în Tema 9. Însă această legătură nu poate fi apreciată ca intensitate și direcție.

Statistica dispune de mai multe metode de studiere a dependențelor dintre două sau mai multe variabile. Printre acestea sunt și cele cuprinse în compartimentul *Corelația și analiza de regresie*. În cadrul acestuia se studiază dependența dintre o variabilă rezultativă (Y), numită și dependentă, și una sau mai multe variabile independente (X). Cu toate că pot corela între ele variabile de diferite tipuri, chiar și cele nominale, cele mai expresive exemple de corelație le găsim în cazul variabilelor numerice.

Fie, de exemplu, două variabile numerice X=$\{x_1, x_2,\ldots, x_n\}$ și Y=$\{y_1, y_2, \ldots y_n\}$. Perechile de valori (x_i, y_i) în axele de coordonate din plan reprezintă niște puncte, care, dacă sunt multe, formează un „nor” (a se vedea Figura 13.1). Dacă se întâmplă că acest nor are o formă alungită și poate fi înconjurat cu o elipsă, atunci se spune că

* A se vedea, de exemplu, *https://dexonline.ro/definitie/corela%C8%9Bie*

variabilele corelează liniar, iar legitatea după care ele corelează este dată de dreapta ce aproximativ coincide cu axa mare a elipsei. Mai mult decât atât, în funcție de înclinarea dreptei se poate concluziona cum corelează variabilele respective. De exemplu, în situația prezentată în Figura 13.1 corelația este negativă: creșterea valorilor lui X duce la descreșterea valorilor lui Y. În cazul orientării dreptei de la stânga-jos spre dreapta-sus corelația este pozitivă: creșterea valorilor lui X duce la creșterea valorilor lui Y.

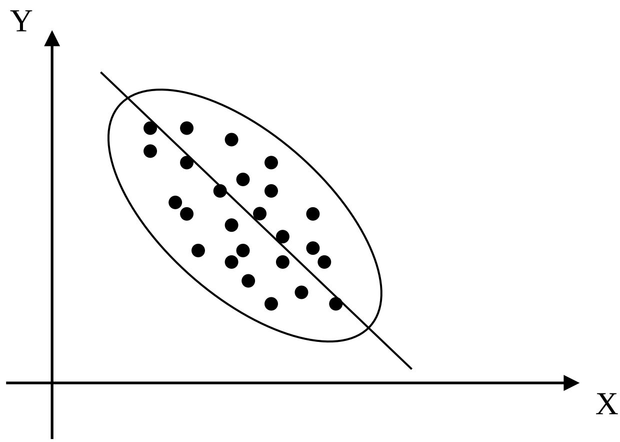


Fig. 13.1. Perechile de valori (x_i, y_i) ca puncte în plan

Menționăm că variabilele X și Y practic nu corelează, dacă norul de puncte are o formă circulară sau dacă axa mare a elipsei este orizontală sau verticală.

Karl Pearson a propus pentru măsurarea intensității și direcției legăturii statistice liniare dintre două variabile numerice coeficientul care-i poartă numele, calculat după formula:

$$k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n\sigma_x\sigma_y},$$

unde:

k – coeficientul de corelație liniară Pearson;

X={$x_1, x_2,\ldots, x_n$}, Y={$y_1, y_2, \ldots y_n$} – variabile;

$\overline{x}$, $\overline{y}$ – mediile valorilor variabilelor X și Y;

σ_x, σ_y – abaterile standard ale valorilor variabilelor X și Y.

Coeficientul de corelație liniară Pearson poate înregistra valori cuprinse intre -1 si +1. Valorile negative ale coeficientului indică o dependență invers proporțională dintre variabile, iar cele pozitive – o dependență direct proporțională. Intensitatea corelației depinde de valoarea absolută a coeficientului de corelație. O clasificare a intensității corelației este următoarea:

$0 < |k| \leq 0{,}2$ – corelație foarte slabă;

$0{,}2 < |k| \leq 0{,}5$ – corelație slabă;

$0{,}5 < |k| \leq 0{,}7$ – corelație moderată;

$0{,}7 < |k| \leq 0{,}9$ – corelație puternică;

$0{,}9 < |k| \leq 0{,}1$ – corelație foarte puternică.

Corelațiile pot fi clasificate în funcție de următoarele criterii:

1) După numărul variabilelor care intervin într-un sistem de interdependență statistică, se disting:

• *corelații simple*, când sistemul considerat cuprinde o variabilă independentă (cauza) și o variabilă dependentă (efect);

• *corelații multiple* – o variabilă dependentă și două sau mai multe variabile independente.

2) După sensul sau direcția corelației, pot exista:

• *corelații directe,* când modificarea într-un anumit sens a valorilor variabilei cauză determină modificarea în același sens a valorilor variabilei efect;

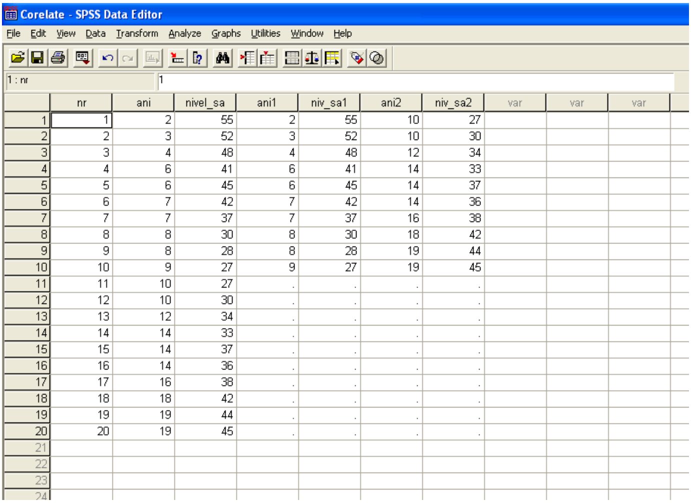
• *corelații indirecte (inverse),* când modificarea într-un anumit sens a valorilor variabilei cauză determină modificarea în sens invers a valorilor variabilei efect (situația din Figura 13.1).

3) După forma analitică, legăturile de interdependență pot fi:

• *corelații liniare*, când perechile de valori (x_i, y_i) – puncte în plan, pot fi aproximate cu o dreaptă $y = ax+b$ (a și b – constante);

• *corelații neliniare,* când perechile de valori (x_i, y_i) – puncte în plan, pot fi aproximate cu orice altă linie curbă $y = f(x)$.

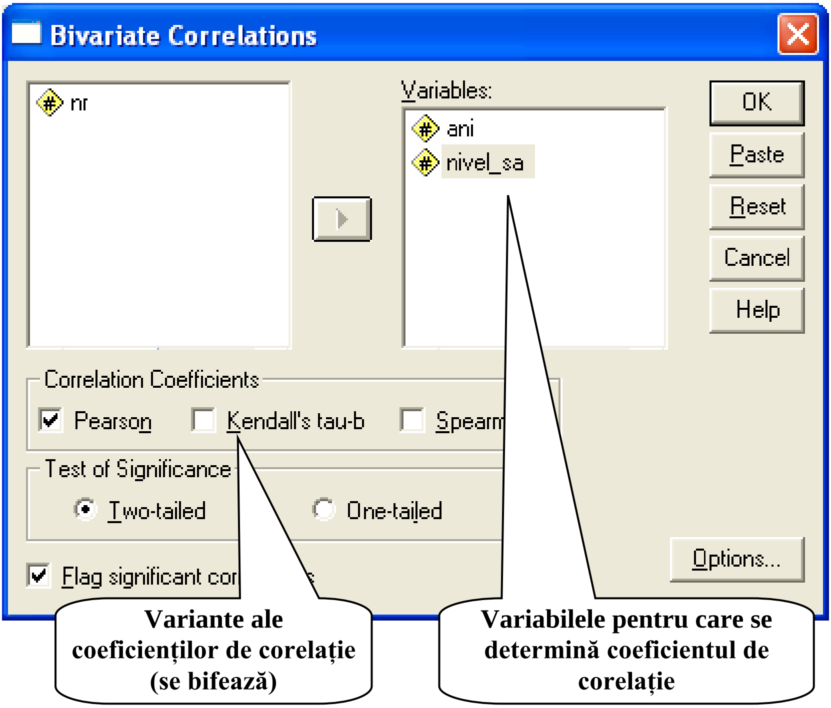
Observația 13.1. În cazul variabilelor ordinale corelația, liniară dintre ele este descrisă de coeficienții de corelație Spearman sau Kendall, având proprietăți asemănătoare coeficientului de corelație Pearson. Corelația variabilelor ordinale mai poartă denumirea de *corelație a rangurilor* (sau codurilor valorilor).



	nr	ani	nivel_sa	ani1	niv_sa1	ani2	niv_sa2	var	var	var
1	1	2	55	2	55	10	27			
2	2	3	52	3	52	10	30			
3	3	4	48	4	48	12	34			
4	4	6	41	6	41	14	33			
5	5	6	45	6	45	14	37			
6	6	7	42	7	42	14	36			
7	7	7	37	7	37	16	38			
8	8	8	30	8	30	18	42			
9	9	8	28	8	28	19	44			
10	10	9	27	9	27	19	45			
11	11	10	27	.	.	.	.			
12	12	10	30	.	.	.	.			
13	13	12	34	.	.	.	.			
14	14	14	33	.	.	.	.			
15	15	14	37	.	.	.	.			
16	16	14	36	.	.	.	.			
17	17	16	38	.	.	.	.			
18	18	18	42	.	.	.	.			
19	19	19	44	.	.	.	.			
20	20	19	45	.	.	.	.			
21										
22										
23										
24										

Fig. 13.2. Datele exemplului pentru analiza corelațională

Vom demonstra în continuare, printr-un exemplu concret, analiza corelațională cu ajutorul programului SPSS. Vom analiza cum depinde satisfacția față de viața de familie (variabila *nivel_sa*) în funcție de durată căsătoriei (variabila *ani*), în presupunerea că ele corelează liniar. Datele au fost culese de la 20 de familii și sunt reprezentate în Figura 13.2.



Correlations

		Durata casatoriei	Nivelul de satisfactie
Durata casatoriei	Pearson Correlation	1	-.234
	Sig. (2-tailed)	.	.321
	N	20	20
Nivelul de satisfactie	Pearson Correlation	-.234	1
	Sig. (2-tailed)	.321	.
	N	20	20

Fig. 13.3. Calcularea coeficienților de corelație liniară

În SPSS coeficienții de corelație liniară se calculează prin comanda **Analyze** → *Correlate* ▸ → *Bivariate*… După lansarea acestei comenzi, introducerea în câmpurile respective a variabilelor ce se analizează, alegerea coeficientului de corelație ce va fi calculat (Pearson, în cazul nostru) și tastarea butonului de comandă OK obținem rezultatul din Figura 13.3. Coeficientul de corelație liniară Pearson calculat (-0,234) demonstrează o corelație slabă indirectă (invers proporțională) între variabilele studiate. Așa oare să fie?

Vom încerca totuși să vedem cum sunt aranjate în plan punctele (x_i, y_i), corespunzătoare variabilelor *ani* și *nivel_sa*. Pentru aceasta construim, tot în SPSS, diagramă de împrăștiere (*scatter plot*) prin comanda **Graphs** → *Scatter..* → *Simple*, plasând valorile duratei căsătoriei pe axa *X*, iar cele ale nivelului de satisfacție – pe axa *Y* (a se vedea Figura 13.4).

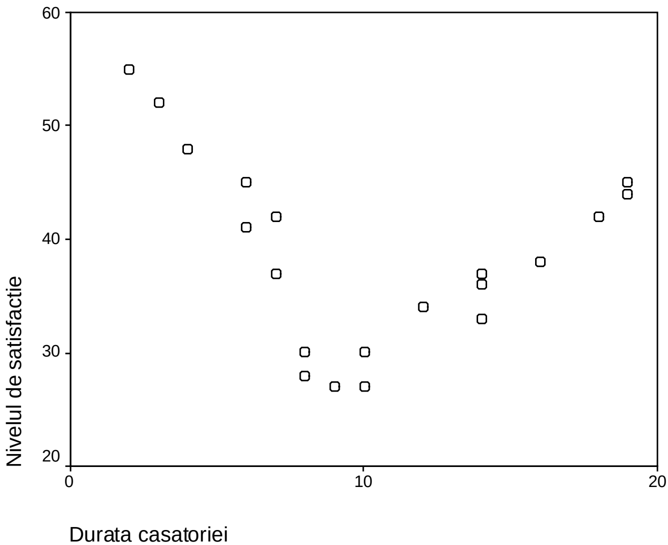


Fig. 13.4. „Norul” de puncte format de perechile de valori ale variabilelor *nivel_sa* și *ani*

Rezultatul obținut demonstrează totuși dependențe, aproape liniare, între variabilele cercetate pe două porțiuni de timp: până la 10 ani de căsătorie și peste 10 ani de căsătorie. Divizând variabilele inițiale în patru variabile, corespunzătoare acestor porțiuni de timp (a se vedea Figura 13.2), și calculând coeficienții de corelație liniară Pearson pentru perechile noi de variabile, obținem că pe porțiunile respective de timp variabilele corelează foarte puternic, coeficienții de corelație fiind apropiați în valoare absolută de 1 (a se vedea Figura 13.5).

Correlations

		Durata casatoriei (pana la 10 ani)	Nivelul de satisfactie
Durata casatoriei (pana la 10 ani)	Pearson Correlation	1	-.952**
	Sig. (2-tailed)	.	.000
	N	10	10
Nivelul de satisfactie	Pearson Correlation	-.952**	1
	Sig. (2-tailed)	.000	.
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

		Durata casatoriei (peste 10 ani)	Nivelul de satisfactie
Durata casatoriei (peste 10 ani)	Pearson Correlation	1	.969**
	Sig. (2-tailed)	.	.000
	N	10	10
Nivelul de satisfactie	Pearson Correlation	.969**	1
	Sig. (2-tailed)	.000	.
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 13.5. Coeficienții de corelație Pearson pentru durata căsătoriei până la 10 ani (-0,952) și peste 10 ani (0,969)

Drept rezultat, se poate concluziona că până la 10 ani de viață de familie nivelul de satisfacție scade liniar (k = *-0,952*), iar după 10 ani – crește, la fel liniar (k = *0,969*). Poate din această cauză familiile divorțează mai frecvent la începutul vieții de familie?

Pentru a vedea care este legitatea de dependență a variabilelor, vom apela la analiza de regresie, care ne arată cum (în ce formă sau după ce formulă) o variabilă este dependentă de o altă variabilă (sau de alte variabile).

Def. 13.3. Activitatea desfășurată pentru obținerea unui model statistic al corelației se numește ***analiză de regresie.*** Scopul principal al acestei activități este de a identifica relația matematică dintre o variabilă dependentă și una sau mai multe variabile independente.

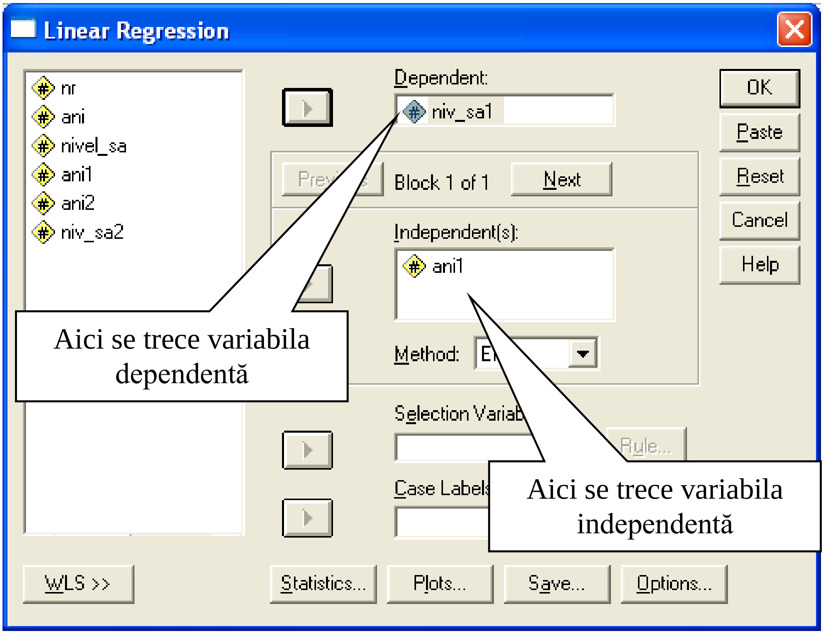
Regresia statistică este folosită pentru modelarea legăturilor statistice dintre variabile. Modelele construite prin regresie pot fi folosite apoi la realizarea de predicții (prognoze) statistice.

Prin regresia statistică se modelează legăturile statistice dintre una sau mai multe variabile *endogene* (denumite și variabile *prezise, explicate* sau *dependente*), notate de obicei cu ***Y*,** și una sau mai multe variabile *exogene* (denumite și variabile *predictoare*, *explicative* sau *independente*), notate de obicei cu *X*.

Pentru regresia statistică, sunt disponibile modele *liniare*, construite pe baza unor funcții matematice liniare, si modele *neliniare*, construite pe baza unor funcții matematice neliniare.

Modelele construite cu o singură variabilă explicată sunt modele de regresie *univariată*, iar modelele construite cu mai multe variabile explicate sunt modele de regresie *multivariată*. Modelele de regresie univariată pot fi, la rândul lor, modele de regresie *simplă*, construite pentru o singură variabilă explicativă, și modele de regresie *multiplă*, care implică mai multe variabile explicative în relație cu variabila explicată considerată.

Pe același exemplu (a se vedea Figura 13.2) vom demonstra analiza de regresie cu ajutorul programului SPSS. Consecutiv, vom construi dreptele de regresie liniară pentru cele două porțiuni de durată a căsătoriei, folosind comanda **Analyze** → *Regression* ▸ → *Linear…* (a se vedea Figura 13.6). Rezultatul executării acestei comenzi sunt cei doi coeficienți ai dreptei de regresie, evidențiați în Figura 13.6: $b = 65{,}000$ și $a + -4{,}083$. Rezultatul pentru toată perioadă vieții de familie examinată în formă grafică este prezentat în Figura 13.7.



Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	65.000	2.955		21.999	.000
	Durata casatoriei (pana la 10 ani)	-4.083	.463	-.952	-8.827	.000

a. Dependent Variable: Nivelul de satisfactie

Fig. 13.6. Regresia liniară pentru variabilele *niv_sa* și *ani1* (durata căsătoriei până la 10 ani) și rezultatul – coeficienții dreptei de regresie *y = -4,083x+65,000*

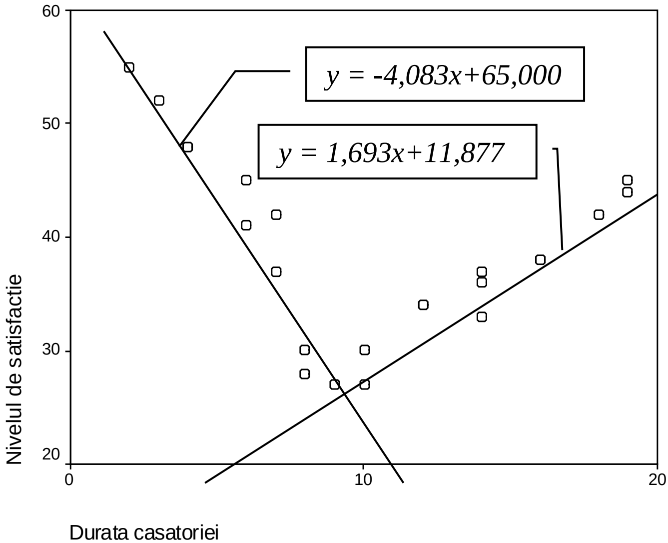


Fig. 13.7. Aproximarea „norului” cu drepte (regresie liniară)

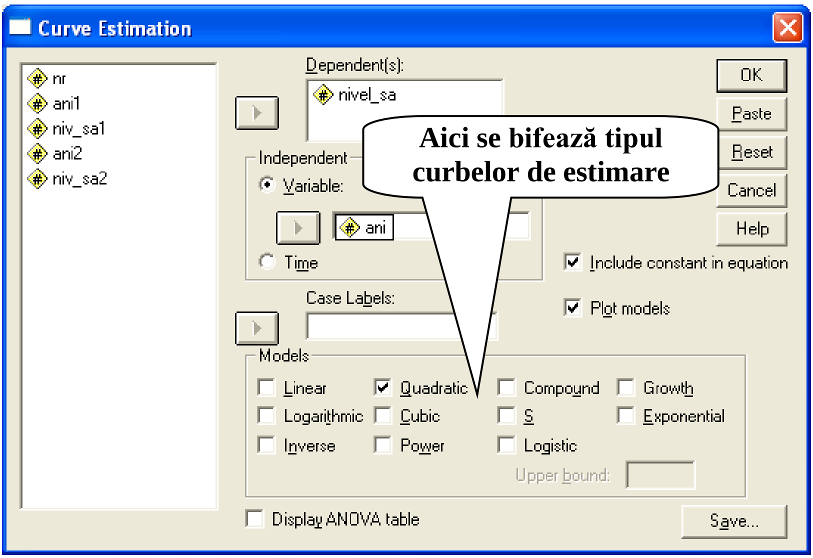


Fig. 13.8. Estimarea cu diferite curbe a relațiilor dintre variabile

Desigur, dependența dintre durata căsătoriei și nivelul de satisfacție poate fi analizată ca o dependență neliniară. Cu alte cuvinte, norul de valori (x_i, y_i) poate fi aproximat cu o singură linie curbă, care ar reprezenta această dependență. Pentru a face acest lucru în SPSS, lansăm comanda **Analyze** → *Regression* ▸ → *Curve Estimation…* (a se vedea Figura 13.8). În caseta de dialog respectivă se bifează, de regulă, mai multe variante ale curbelor de estimare, ca apoi să se selecteze cea mai potrivită. Rezultatul executării comenzii va conține coeficienții curbei de estimare respective.

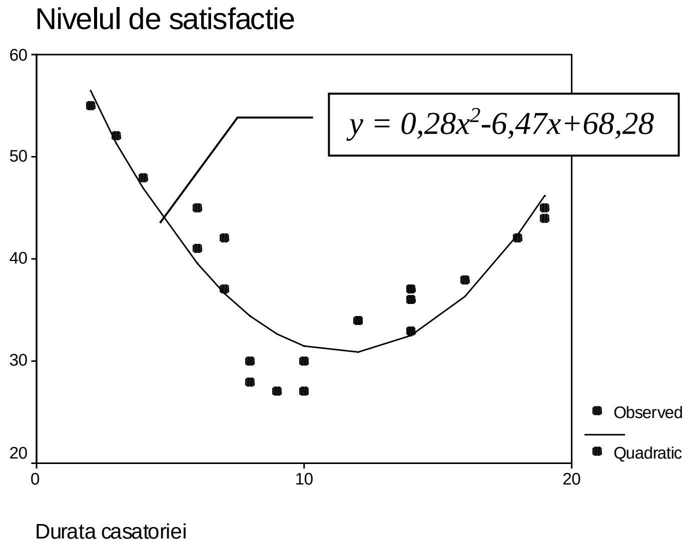


Fig. 13.9. Variantă de regresie neliniară (aproximarea „norului” de puncte cu o parabolă)

În particular, pentru cazul examinat *supra*, perechile de valori (x_i, y_i) au fost aproximate cu o curbă de gradul 2 (parabolă), formula și forma căreia sunt prezentate în Figura 13.9. Cu ajutorul acestei formule, se poate prezice, de exemplu, aproximativ, care va fi nivelul de satisfacție a soților după 21-25 de ani de căsătorie, dar nu mai mult…

Exerciții, întrebări de control

1. în Tabelul 13.1 sunt prezentate ratele inflației din anii ’90 în trei țări: Moldova, Rusia și Slovenia. Să se determine coeficienții de corelație liniară Pearson pentru ratele inflației din fiecare pereche de țări și să se tragă concluziile respective. Dacă între oarecare două țări există o corelație liniară puternică, să se construiască dreapta de regresie respectivă.

Tabelul 13.1

Ani	Moldova	Rusia	Slovenia
1990	4,2	5,3	549,7
1991	98	92,7	117,7
1992	1276,4	1526	207,3
1993	788,5	875	32,9
1994	329,7	311,4	21
1995	30,2	197,7	13,5
1996	23,5	47,8	9,9
1997	11,8	14,7	8,4
1998	7,7	27,6	8
1999	39,3	86,1	6,1
2000	32	20,7	8,6

2. În Tabelul 13.2 sunt aduși coeficienții nivelului intelectului fumătorilor (*IQ*) în funcție de numărul mediu de țigări fumate pe zi (*Nr*). Să se studieze, dacă aceste două variabile corelează între ele și să se determine coeficientul Pearson de corelație liniară pentru confirmare sau infirmare.

Tabelul 13.2

Nr	7	49	41	38	37	19	35	40	1	10	18	21	25	7	38
IQ	10	6	15	5	12	4	19	11	3	3	22	17	12	9	13

3. În Tabelul 13.3 este prezentat numărul de vizite ale pacienților la medicul de familie (*Nr*) în diferite luni ale anului (*Luna*). Să se realizeze analiza corelațională a acestor variabile și, dacă ele corelează, să se găsească ecuația liniei de regresie:

Tabelul 13.3

Luna	*Ian*	*Feb*	*Mar*	*Apr*	*Mai*	*Iun*	*Iul*	*Aug*	*Sep*	*Oct*	*Noe*	*Dec*
Nr	200	170	100	80	80	60	50	50	70	60	90	140

4. În Tabelul 13.4 sunt dați indicii de percepție a corupției (IPC) și cei ai globalizării (IG) pentru câteva ţări ale lumii pentru anul 2011. Determinați în ce măsură corelează aceştia şi formulați concluzia respectivă.

Tabelul 13.4

	Suedia	Canada	Germania	SUA	Polonia	Georgia	Moldova	Rusia	Uzbekistan	Afganistan
IPC	9.4	8.7	8	7.1	5.5	4.1	2.9	2.4	1.6	1.5
IG	89.26	85.8	85.1	79.83	79.66	60.71	62.22	65.91	41.07	30.57

5. Dependența dintre vârsta femeii și numărul mediu de copii născuți până la această vârstă (evoluția ratei totale de fertilitate – RTF) pentru generația anului 1960 din Moldova este prezentată în Tabelul 13.5.

Tabelul 13.5

Vârsta	20	25	30	35	40	45	50
RTF	0,25	1,13	1,76	2,06	2,16	2,21	2,24

Să se determine dacă aceste două variabile corelează, cum corelează, iar dacă corelează, să se construiască linia de regresie.

Tema 14

Principiile analizei factoriale și analizei cluster

Analiza factorială este un instrument statistic, folosit pe larg în psihologie, sociologie, marketing, medicină etc. pentru determinarea unor caracteristici latente ale obiectelor sau caracteristici ce nu pot fi măsurate direct. Ideile principale ale analizei factoriale au fost formulate de psihologul şi antropologul englez F.Galton (1822-1911). Printre savanţii care au contribuit la dezvoltarea şi aplicarea în practică a analizei factoriale se număra: Ch.Spearman, R.Cattell, K.Pearson, H.Hotelling, H.Eysenck.

Analiza factorială permite cercetătorului să rezolve două probleme importante: să descrie obiectul studiat multilateral şi, în acelaşi timp, compact. Cu ajutorul analizei factoriale, pot fi determinaţi factorii variabili latenţi, responsabili de existenţa unor relaţii statistice de corelare între variabilele observabile. Astfel, pot fi evidenţiate două scopuri ale analizei factoriale: determinarea relaţiilor reciproce dintre variabile (clasificarea variabilelor) şi micşorarea numărului de variabile necesare pentru descrierea obiectelor.

Def. 14.1. ***Analiza factorială*** (engl. *Factor analysis*) − procedură prin care un număr mare de variabile observabile (direct măsurabile), ce caracterizează obiectele dintr-o mulţime, se reduce la un număr mai mic de variabile independente, diferite de cele observabile, numite *factori*.

Astfel, un factor adună în sine variabilele ce corelează între ele puternic, pe când variabilele din factori diferiţi corelează între ele slab.

Scopul analizei factoriale este deci de a determina acei factori complecşi, care, pe cât se poate mai deplin, să explice relaţiile dintre variabilele observabile. Un exemplu simplu ar fi factorul „inteligenţa”, care direct nu poate fi măsurat, însă conţine în sine aşa componente (variabile observabile) ca „nivelul de înţelegere a materialului”,

„nivelul de însuşire a materialului”, „calitatea vocabularului”, toate putând fi măsurate cu o scală de la 1 (foarte mic) la 5 (foarte mare).

Analiza factorială poate fi *exploratorie* şi *confirmatorie*. Analiza factorială exploratorie permite a determina factorii latenţi, fără a cunoaşte numărul şi ponderea lor, iar cea confirmatorie este destinată verificării ipotezelor cu privire la numărul şi ponderea factorilor. Menţionăm că în practică este aplicată mai frecvent analiza factorială exploratorie, pentru aceasta utilizându-se un şir de programe pe calculator. În programul *SPSS* analiza factorială poate fi realizată prin comanda **Analyze**→ *Data Reduction* ▸ → *Factor...*

Condiţiile realizării unei analize factoriale sunt următoarele:

- variabilele observabile trebuie să fie numerice (în unele cazuri pot fi şi dihotomice sau chiar ordinale);

- numărul de observaţii (cazuri, obiecte studiate) trebuie să fie cel puţin de două ori mai mare decât numărul variabilelor observabile;

- variabilele observabile trebuie să fie omogene, măsurate cu aceleaşi scale (de exemplu, scale Likert);

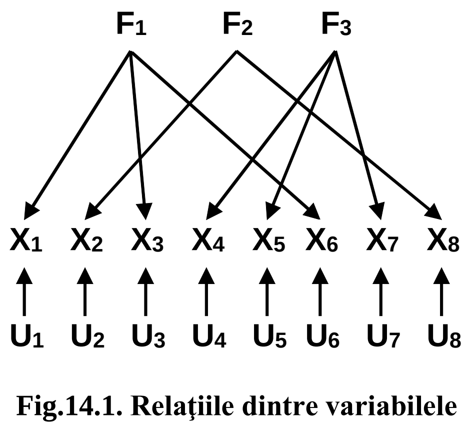


Fig.14.1. Relaţiile dintre variabilele observabile, factorii comuni şi cei unici în analiza factorială

- distribuţia normală a variabilelor observabile reprezintă un avantaj, soluţiile obţinute în aşa caz sunt mai bune;

- între variabilele observabile există corelaţii (dacă între acestea nu există nicio legătură teoretică, variabilele latente nu vor avea niciun sens).

În limbaj matematic analiza factorială poate fi descrisă în felul următor (a se vedea Figura 14.1): fie $X_1, X_2,...,X_n$ − un set de variabile observabile cunoscute, între care nu există relaţii directe, şi se doreşte a determina alt set de variabile, numite *factori comuni*, $F_1,F_2,...,F_m$ (m<n), în aşa fel ca $X_i = p_{i1}F_1+p_{i2}F_2+...+P_{im}F_m+U_i$ (i=1,2,...,n), unde p_{ij} − ponderile factorilor comuni, iar U_i − nişte variabile neobservabile (*factori unici*), care nu corelează între ele şi nici cu factorii comuni F_i.

Def.14.2. ***Analiza cluster*** (engl. *Cluster analysis*) − metodă de descompunere a unei mulţimi de obiecte (indivizi, evenimente) în submulţimi, numite *clustere*, în aşa fel încât fiecare cluster să conţină obiecte similare, pe când obiectele din diferite clustere să difere esenţial.

Analiza cluster ţine de prelucrarea statistică a datelor, este aplicată pe larg într-un şir de domenii: în sociologie − divizarea respondenţilor în grupe omogene; în medicină − clasificarea pacienţilor, preparatelor, simptomelor; în marketing − segmentarea concurenţilor, consumatorilor; în management − descompunerea personalului în diferite grupe după nivelul motivaţiei; în filologie − gruparea limbilor, dialectelor şi altele.

Există circa 100 de algoritmi de clusterizare a mulţimilor, însă cel mai des sunt folosiţi doi dintre ei: *analiza cluster ierarhică* (Hierarchical Cluster Analysis) şi *analiza cluster prin metoda k-mediilor* (K-Means Cluster Analysis). În ambele cazuri, se defineşte o metodă de măsurare a „distanţei” dintre elementele mulţimii, dintre un cluster şi un element al mulţimii sau dintre două clustere pentru a le putea grupa pe cele „apropiate”. Prin următoarele exemple, vom demonstra ambele metode de clusterizare.

În Figura 14.2 este demonstrată ***analiza cluster ierarhică***, care este realizată prin patru paşi. La primul pas cele 8 elemente ale mulţimii se grupează câte două cele mai „apropiate” între ele în sensul distanţei definite. Se obţin patru clustere: a_1, a_2, a_3, a_4. La pasul următor clusterele a_1 şi a_2 se grupează, formând clusterul b_1, întreaga mulţime divizându-se după aceasta în trei clustere: b_1, a_3 şi a_4. La pasul al treilea se mai grupează clusterele b_1 şi a_3, formându-se clusterul c_1, iar mulţimea devine divizată în două clustere: c_1 şi a_4. E clar, că la pasul următor se unesc între ele clusterele c_1 şi a_4, obţinându-se un singur cluster ce coincide cu toată mulţimea.

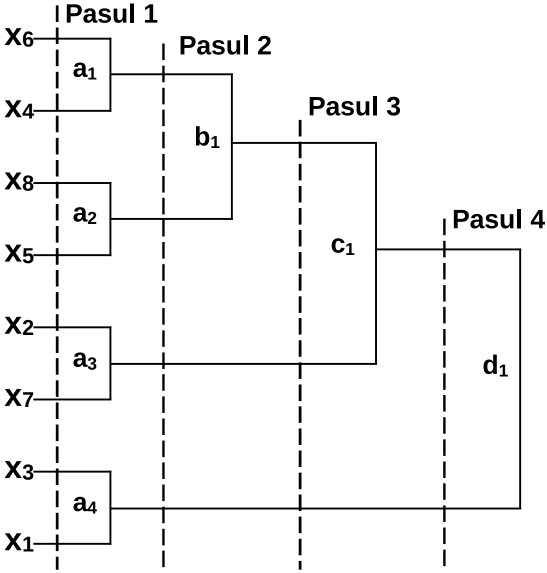


Fig.14.2. Metoda ierarhică de clasificare

Numărul de clustere pe care le alegem ca rezultat al analizei efectuate depinde de problema care se rezolvă: dacă alegem multe clustere (de exemplu, cele obţinute după Pasul 1), atunci se pierde credibilitatea, valorile caracteristicilor acestor clustere sunt multe şi dispersate. Dacă alegem puţine clustere (de exemplu, cele obţinute după Pasul 4), atunci fiecare din acestea vor grupa elemente cu caracteristici destul de neomogene, ceea ce diminuează însăşi ideea clasificării. Astfel, se recomandă ca în urma analizei cluster ierarhice să se utilizeze varianta de mijloc: se aleg nici prea multe şi nici prea puţine clustere.

Cu aceasta, procedura de clasificare prin ierarhizare se termină. Drept rezultat al ei se cunoaşte cum a fost divizată mulţimea, câte elemente şi care din ele le conţine fiecare cluster, dar cel mai important − se cunoaşte numărul clusterelor în care a fost divizată mulţimea.

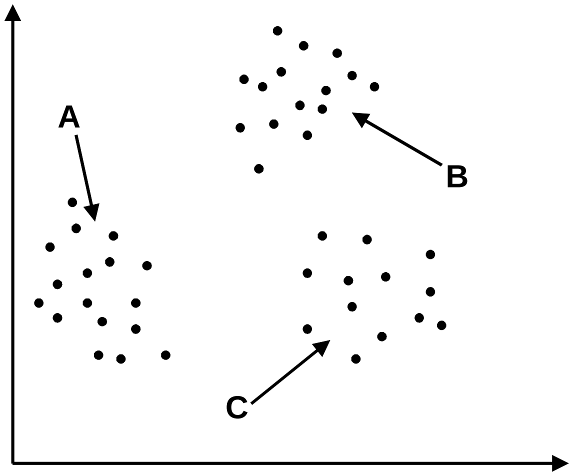


Fig.14.3. Exemplu de divizare a unei mulţimi în trei clustere

Analiza cluster prin metoda k-mediilor este schematic reprezentată în Figura 14.3. Chiar din start metoda presupune divizarea mulţimii în K clustere (K este numărul de clustere, specificat de utilizator). Această procedură începe prin folosirea iniţială a oricăror K elemente ale mulţimii în calitate de estimări temporare ale K centre ale viitoarelor clustere. În continuare, pe rând, fiecare element următor al mulţimii se atribuie unui cluster cu cel mai apropiat centru (în sensul distanţelor definite) ca imediat să se determine noul centru al clusterului. Apoi este folosit un proces iterativ pentru a găsi centrele finale ale clusterelor. La fiecare iteraţie elementele sunt grupate în grupul cu cel mai apropiat centru şi centrele clusterelor sunt recalculate. Acest proces continuă până ce nu mai au loc schimbări în centrele grupurilor sau până când este atins numărul maxim de iteraţii.

Desigur, în prezent analiza cluster nu se realizează manual, pentru aceasta existând programe speciale (cum ar fi cele de analiză statistică a datelor). În particular, programul SPSS permite acest lucru prin opţiunile **Analyze** → *Classify* ▸ → *K-Means Cluster...* şi **Analyze** → *Classify* ▸ → *Hierarchical Cluster...*

Menționăm că analiza cluster o urmează, de regulă, pe cea factorială, conducând la clasificarea populației după variabilele latente, determinate în cadrul analizei factoriale.

Exerciții, întrebări de control

1. Explicați sensul analizei factoriale.
2. Ce proprietăți trebuie să posede variabilele implicate în analiza factorială?
3. Ce se obține în urma analizei factoriale?
4. Descrieți sensul analizei cluster.
5. Ce particularități comune au și prin ce se deosebesc analiza factorială și analiza cluster?
6. De ce, de regulă, analiza cluster o urmează pe cea factorială?

Tema 15.

Reprezentarea rezultatelor

Rezultatele cercetărilor sociologice cantitative pot fi reprezentate atât sub formă de tabele, cât și sub formă de diagrame. Dacă tabelele conțin o cantitate mare de informație, care nu poate fi percepută în întregime și înțeleasă la prima vedere, atunci diagramele, fiind mai sărace în informație (ele reprezintă numai unele laturi ale multitudinii de rezultate), sunt mai ușor percepute și, de regulă, reflectează partea principală, cea mai importantă, a rezultatului.

Desigur, cercetătorului nu-i este interzis să folosească în rapoarte, studii, prezentări ale rezultatelor atât tabele, cât și diagrame. Dacă la folosirea tabelelor în rapoarte, în alte publicații, pot să apară probleme la aranjarea lor pe pagini (ele pot să nu încapă nici în lățime, nici în înălțime pe pagină, iar redimensionarea, micșorarea lor – să reducă mărimea fonturilor, făcând dificil a vedea datele din celule etc.), atunci la utilizarea diagramelor problemele pot fi de altă natură, printre care și suspiciunea: diagramele construite reprezintă corect ceea ce vrea să demonstreze cercetătorul? În continuare, vom încerca să explicăm această situație, demonstrând prin exemple concrete când și ce tipuri de diagrame e mai bine să utilizăm. Nu vom explica metodele de formatare a diagramelor, deoarece ele țin de programul, în care se construiesc (de exemplu, de Excel).

În *Tema 8* ne-am întâlnit cu reprezentări grafice ale unor valori numerice, mai exact – ale frecvențelor valorilor variabilelor (sau ale distribuțiilor de frecvențe). În continuare, vom dezvolta mai pe larg această posibilitate, folosind un exemplu simplu: un șir de date numerice (în particular, ele pot fi și niște frecvențe) cu numele „**Șir**” și etichetele valorilor **a**, **b** și **c** (a se vedea Tabelul 15.1).

Tabelul 15.1

	Șir
a	32
b	43
c	21

Reprezentarea grafică a unui astfel de șir numeric poate avea mai multe tipuri (a se vedea Figura 15.1):

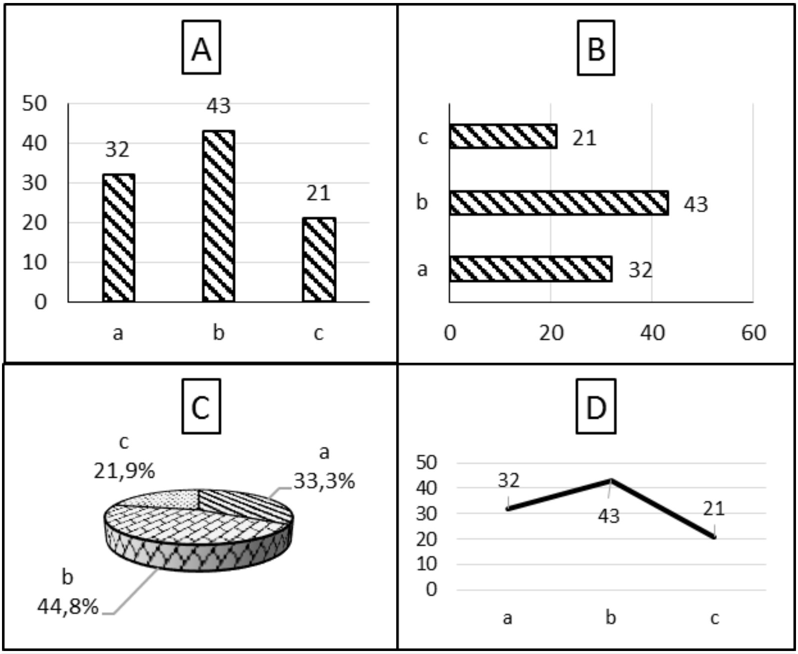


Fig.15.1. Reprezentări grafice ale unui singur șir numeric

Tipul **A** – diagrama cu bare verticale, se folosește cel mai frecvent și e potrivit pentru a compara, pur și simplu, valorile numerice din șir. Tipul **B** – diagrama cu bare orizontale, este asemănător tipului **A**, însă este comod a fi folosit, atunci când etichetele valorilor **a**, **b** și **c** sunt voluminoase (lungi). Și într-un caz, și în altul se recomandă a plasa deasupra, lângă sau chiar pe bare valorile numerice, astfel obținându-se o imagine completă a șirului reprezentat.

Tipul **C** – diagrama circulară sau „plăcinta” (în engl. – pie), se folosește pentru a evidenția valorile șirului numeric ca părți procentuale ale sumei lor sau ale unui tot întreg (100%). Calculatorul, la dorința utilizatorului, singur determină aceste procente și, împreună cu etichetele valorilor șirului, le plasează în jurul „plăcintei”, astfel dispărând necesitatea de a folosi legenda pentru a explica căror valori ale șirului aparțin diferite sectoare ale „plăcintei”.

Tipul **D** – linia frântă, este folosită pentru a reprezenta evoluția în timp a valorilor șirului (în astfel de situații, axa orizontală *X* a diagramei este axa timpului). Tot cu un astfel de tip de diagramă ar putea fi reprezentate frecvențele variabilelor ordinale, scala de valori a cărora în așa caz se poziționează pe axa *X* în creștere, de la stânga la dreapta.

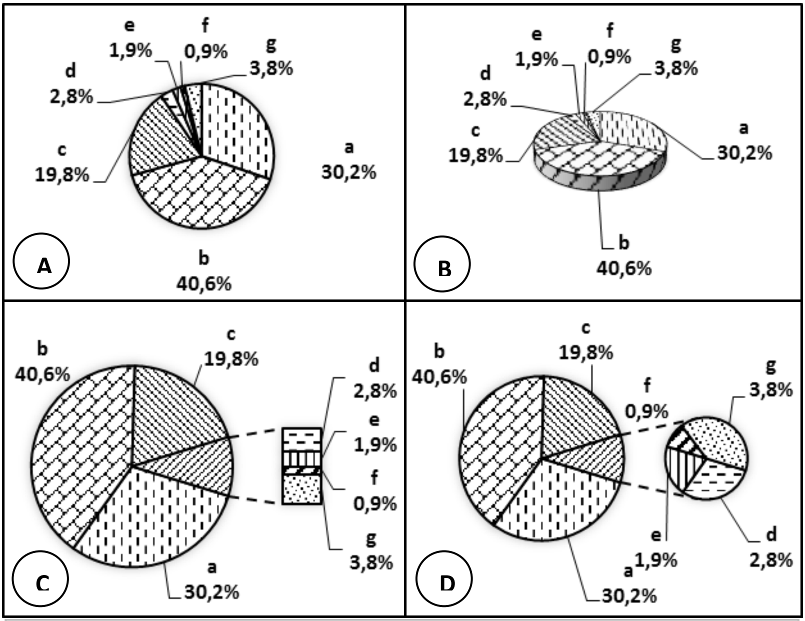


Fig.15.2. Reprezentări grafice ale unui singur șir numeric prin diferite diagrame circulare (pie)

Dacă șirul numeric conține valori care se deosebesc între ele la nivel de ordine, atunci diagramele arată neestetic, cu elemente grafice puțin lizibile (a se vedea Figura 15.2, valorile **d**, **e**, **f** și **g** din diagramele **A** și **B**). Acestea însele pot fi totuși evidențiate prin diagrame perechi „plăcintă” – bară sau „plăcintă” – „plăcintă” (a se vedea Figura 15.2, diagramele **C** și **D**).

Pentru a compara 2 sau mai multe șiruri de valori numerice diagramele circulare nu pot fi folosite. În așa caz cel mai frecvent se folosesc diagramele cu bare și cu linii frânte. Vom examina exemplul a 2 șiruri numerice **Șir1** și **Șir2**, ce se conțin în Tabelul 15.2.

Tabelul 15.2

	Șir1	**Șir2**
a	32	41
b	43	23
c	21	33

Există 3 variante dintre cele mai populare de comparare a șirurilor numerice cu ajutorul diagramelor cu bare (verticale sau orizontale), tipul alegându-se, după cum s-a menționat anterior, ținându-se cont de lungimea etichetelor (a se vedea Figura 15.3, diagramele **A**, **B** și **C**).

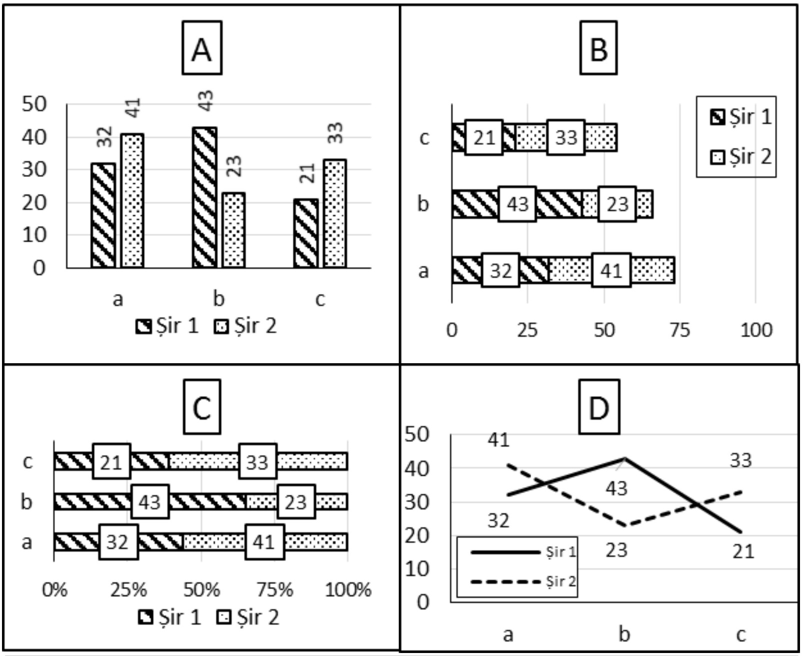


Fig.15.3. Compararea grafică a două șiruri numerice

În prima variantă (**A**), cea mai utilizată, șirurile numerice se compară prin bare alăturate, fiecare grup de bare corespunzând valorilor de același fel al șirurilor, iar fiecare bară din grup – valorii concrete din unul dintre șiruri. Legenda, prezentă obligatoriu în diagramă, indică apartenența valorilor la șiruri.

Varianta a doua (**B**), în afară de compararea valorilor șirurilor, reprezintă și suma acestora, egală cu lungimea totală a barelor suprapuse (bare verticale) sau alăturate (bare orizontale).

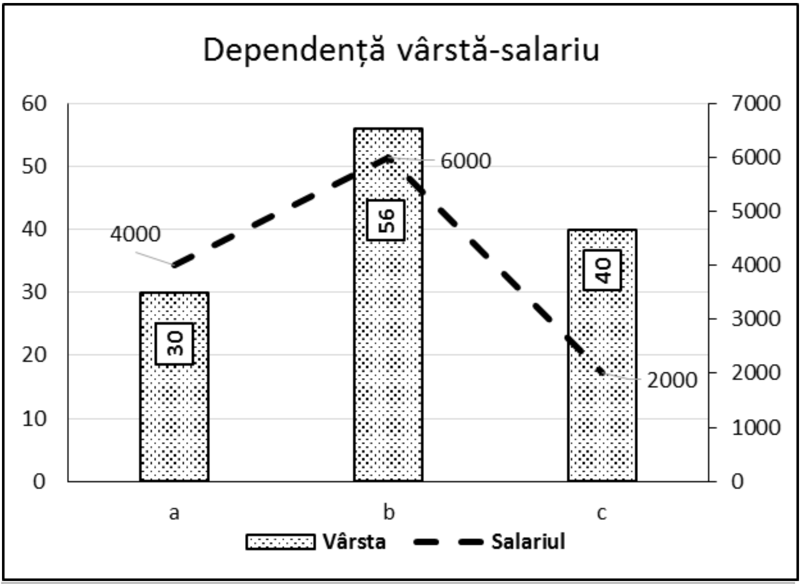


Fig.15.4. Compararea grafică a șirurilor numerice de diferită natură

Varianta a treia (**C**) compară nu numai valorile perechi, dar le transformă și în procente față de suma lor, lungimea sumară a barelor suprapuse (alăturate) fiind de 100%.

În sfârșit, pentru a compara șiruri ce evoluează în timp sau valori ale frecvențelor variabilelor ordinale pot fi folosite diagramele cu linii frânte (**D**), prin care se compară nu numai valorile șirurilor, dar și evoluția lor.

În practică se întâlnesc situații când este necesar a compara în timp două șiruri numerice de diferită natură sau variabile având scale de măsură diferite (de exemplu, Indicele Dezvoltării Umane și Indicele de Percepție a Corupției, salariul angajatului și coeficientul lui de inteligență IQ, rata inflației și costul coșului minim de consum etc.). În așa caz, ne vin în ajutor diagramele combinate, un exemplu din ele (bare – linie frântă) e demonstrat în Figura 15.4, prin care se compară salariul cu vârsta angajaților. Deseori astfel de reprezentări ale șirurilor ajută a observa dacă ele corelează.

În final, aducem câteva recomandări practice:

a) Elementele diagramelor pentru publicațiile alb-negru se fac monocolore, „culorile" suprafețelor reprezentându-se prin diferite ornamente (uzoare, patterne), iar ale liniilor – prin diferite stiluri ale lor.

b) Elementele textuale ale diagramelor, construite pentru prezentări Power Point, trebuie să aibă mărimi ce le fac vizibile clar pe ecranul de proiecție.

c) Variantele 3D ale diagramelor pot fi utilizate, chiar și în publicații, numai în variantă color.

Exerciții, întrebări de control

1. Este dat tabelul de asociere, care reprezintă autoaprecierea sănătății de către diferite categorii de respondenți (a se vedea Tabelul 15.3).

Tabelul 15.3

		Autoapreciere sănătate			
		Bună	**Satisfăcătoare**	Rea	Total
Sex	Feminin	1.097	449	400	1.946
	Masculin	1.002	228	167	1.397
	Total	2.099	677	567	3.343
Mediu de reședință	Sat	1.055	369	291	1.715
	Oraș	524	168	165	857
	mun. Chișinău	520	140	111	771
	Total	2.099	677	567	3.343

a) Să se reprezinte structura respondenților după sex și după mediul de reședință prin diagrame circulare.

b) Să se reprezinte autoaprecierea sănătății de către respondenți printr-o diagramă cu linie frântă

c) Să se compare autoaprecierea sănătății de către femei și bărbați, de către respondenții de la sat, de la oraș și din mun. Chișinău prin diagrame cu bare de diferite subtipuri.

2. În Tabelul 15.4 sunt reprezentați Indicele de Percepție a Corupției (IPC) și Indicele Globalizării (IG) pentru câteva țări ale lumii (anul 2011), iar în Tabelul 15.4 – evoluția valorilor globale ale acestora în anii 2000-2009.

Tabelul 15.4

	IPC	IG
Suedia	9,4	89,26
Canada	8,7	85,80
Germania	8,0	85,10
SUA	7,1	79,83
Polonia	5,5	79,66
Georgia	4,1	60,71
Romania	3,6	71,25
Moldova	2,9	62,22
Rusia	2,4	65,91
Ukraina	2,3	65,71
Uzbekistan	1,6	41,07
Afganistan	1,5	30,57

Tabelul 15.5

	IG	IPC
2000	64,5	4,98
2001	65,1	4,97
2002	65,2	4,96
2003	65,9	4,94
2004	66,9	4,96
2005	67,2	5,03
2006	67,9	5,06
2007	69,1	5,06
2008	68,9	5,05
2009	68,6	5,01

a) Să se reprezinte grafic IPC și IG pentru țările menționate.

b) Să se compare IPC și IG printr-o diagramă combinată (bare – linie frântă).

c) Determinați în ce măsură corelează IPC cu IG pentru țările menționate, calculați coeficientul de corelație Pearson pentru perechea de variabile IPC și IG și faceți concluzia respectivă.

d) Să se reprezinte evoluția valorilor globale ale IPC și IG în anii 2000-2009 printr-o diagramă combinată și să se determine în ce măsură există o corelație între acești indici.

3. *Infra* sunt aduse trei variante ale unei întrebări din chestionare cu același sens:

I. Cât de importante sunt următoarele lucruri pentru Dvs.?

(marcați câte un răspuns pe fiecare linie)

	Foarte important	**Puţin important**	**Deloc important**
1. Familia	1	2	3
2. Lucrul	1	2	3
3. Studiile	1	2	3
4. Prietenii	1	2	3
5. Copiii	1	2	3
6. Timpul liber	1	2	3

II. Care dintre următoarele lucruri este *cel mai important* pentru Dvs.?

1. Familia
2. Lucrul
3. Studiile
4. Prietenii
5. Copiii
6. Timpul liber

III. Care trei dintre următoarele lucruri *sunt cele mai importante* pentru Dvs.?

1. Familia
2. Lucrul
3. Studiile
4. Prietenii
5. Copiii
6. Timpul liber

Suplimentar, se cunoaște sexul respondenților (femeie, bărbat). Propuneți și argumentați variante de diagrame pentru următoarele rezultate:

a) Nivelul de importanță al lucrurilor pentru respondenți.

b) Cel mai important lucru pentru respondenți.

c) Cele mai importante lucruri pentru respondenți.

d) Nivelul comparativ de importanță a lucrurilor pentru femei și bărbați.

e) Cel mai important lucru pentru femei în comparație cu cel pentru bărbați.

f) Cele mai importante lucruri pentru femei în comparație cu cele pentru bărbați.

Bibliografie recomandată

1. CLOCOTICI, V. et al. *Statistică aplicată în psihologie.* Iaşi: Polirom, 2001.
2. CULIC, I. *Metode avansate în cercetarea socială.* Iaşi: Polirom, 2004.
3. HOWITT, D. et al. *Introducere în SPSS pentru psihologie.* Iaşi: Polirom, 2006.
4. LABĂR, A.V. *SPSS pentru ştiinţele educaţiei.* Iaşi: Polirom, 2008.
5. LUNGU, O. *Ghid introductiv pentru SPSS 10.0.* Iaşi: Polirom, 2001.
6. RATEAU, P. *Metodele şi statisticile experimentale.* Iaşi: Polirom, 2004.
7. ROTARIU, T. et al. *Ancheta sociologică şi sondajul de opinie*. Iaşi: Polirom, 1997.
8. ROTARIU, T. et al. *Metode statistice aplicate în ştiinţele sociale.* Iaşi: Polirom, 1999.
9. БЮЮЛЬ, А., ЦЕФЕЛЬ, П. *SPSS: исскуство обработки информации.* Москва, СПб, Киев, 2002 (*www.crras.usm.md*)
10. НАСЛЕДОВ, А. *SPSS: компьютерный анализ данных в психологии и социальных науках.* Москва, 2007.
11. ПАЦИОРКОВСКИЙ, В.В., ПАЦИОРКОВСКАЯ, В.В. *SPSS для социологов. Учебное пособие.* Москва: ИСЭПН РАН, 2005. (*http://csl.isc.irk.ru/BD/Books/spss%20для%20 социологов.pdf*)
12. ФАРАХУТДИНОВ, Ш.Ф., БУШУЕВ, А. С. *Обработка и анализ данных социологических исследований в пакете SPSS 17.0. Курс лекций*. Тюмень: ТюмГНГУ, 2011. (*http://frima.org/soc/INFT/SPSS.pdf*)

Oleg BULGARU

APLICAȚII STATISTICE
în cercetarea sociologică

Suport de curs

Redactare – ***Antonina Dembițchi***
Machetare computerizată – ***Oleg Bulgaru***

Bun de tipar 20.06.2018
Formatul $60x84^{1}/_{16}$.
Coli de tipar 9,4. Coli editoriale 4,6.
Comanda 14. Tirajul 50 ex.

Centrul Editorial-Poligrafic al USM
str. Al. Mateevici, 60, Chişinău, MD-2009.