CORINA ILINCA

ANALIZA DE DATE:
EXPLICAȚII, INTERPRETAREA REZULTATELOR ȘI EXERCIȚII REZOLVATE

TRITONIC
Tritonic Books
București | 2024

Corina Ilinca
Analiza de date:
explicații, interpretarea rezultatelor și exerciții rezolvate

TRITONIC
Str. Coacăzelor nr. 5, București
e-mail: editura@tritonic.ro
www.tritonic.ro

Tritonic București apare la poziția 18 în lista cu Edituri de prestigiu recunoscut în domeniul științelor sociale (lista A2) (CNATDCU):
http://www.cnatdcu.ro/wp-content/uploads/2011/11/A2_Panel41.xls

Colecția Comunicare Media este coordonată de lect. univ. dr. Bogdan Hrib.

Descrierea CIP a Bibliotecii Naționale a României
ILINCA, CORINA
Analiza de date : explicații, interpretarea rezultatelor și exerciții rezolvate /
Corina Ilinca. - București : Tritonic Books, 2024
Conține bibliografie
ISBN 978-606-749-769-4

316
36

Coperta: Alexandra Bardan
DTP: Ioan Dorel Radu
Editor: Rebeca Cojocaru
Comanda nr. AS03/2024
Bun de tipar: August 2024
Tipărit în România

CUPRINS

INTRODUCERE

Plecând de la ideea de a prezenta diferite analize de date, urmând pașii de bază în analiză, am creat această carte pentru a exemplifica modalități diferite de a prezenta tendințe existente în date, de a testa ipoteze și de a crea modele mai dezvoltate care să utilizeze diferiți algoritmi. Deși statistica obține în general sentimente de teamă indiferent de domeniul de studiu, mi-ar plăcea ca prin parcurgerea acestei cărți cititorii să prindă curaj și să înțeleagă cu ușurință abordările prezentate. Exemplele prezentate pot fi utile pentru diferite domenii, de la asistență socială, sociologie, psihologie, medicină, la matematică, automatică și informatică.

Fiecare capitol conține la început exemple și explicații introductive în tema propusă, cu referențe către alte materiale ce pot fi consultate pentru dezvoltarea cunoștințelor. Dacă aveți deja cunoștințe dobândite în cadrul unui curs de analiză de date, ar fi util pentru o înțelegere mai dezvoltată a temelor propuse. O strategie de învățare poate fi și prin consultarea răspunsurilor la întrebări și recitirea întrebărilor știind răspunsul corect. Un singur răspuns este corect pentru fiecare întrebare din toată cartea, indiferent câte variante de răspuns sunt propuse. Tabelele au fost realizate folosind programele IBM SPSS Statistics 29, Microsoft 365 Excel 2024, versiunea Desktop și IBM AMOS 29. Pentru a lucra cu programele SPSS și AMOS pe propriul dispozitiv, IBM (2024a) recomandă o listă de vânzători de unde persoanele din mediul academic pot cumpăra o licență cu un preț mai mic (30 zile fiind gratuită utilizarea).

Mulțumesc familiei mele pentru că este alături de mine în parcursul profesional ales, studenților și profesorilor mei alături de care am dezvoltat diverse analize de date de-a lungul timpului și celor care m-au încurajat să continui, în special domnului profesor Doru Buzducea pentru încurajările de a publica această carte. Le adresez mulțumiri tuturor profesorilor mei de la care am

învățat statistică și analiză de date, în special celor pentru care am ținut seminarii de analiză de date în perioada studiilor doctorale: doamnei profesoare Cosima Rughiniș de la care am învățat bazele analizelor univariate, bivariate și corelaționale cu sens teoretic, doamnei profesoare Paula Tufiș de la care am învățat analizele multivariate în SPSS și modele de ecuații structurale în AMOS și domnului profesor Marian Vasile de la care am învățat aspecte de practica cercetării cantitative folosind SPSS.

Această carte conține date diverse, din diferite surse, fie agregate deja de la Institutul Național de Statistică, fie date la nivel de individ procesate de mine (World Values Survey 2018 Haerpfer et al., 2022; Health and Retirement Study, 2024; HRS oferit de RAND, 2024; RAND HRS Longitudinal File 2020 (V2), 2024; Special Eurobarometer 398 (EB 79.2, European Commission, 2014; ESS Round 6: European Social Survey Round 6 Data (2012); date de recensământ perioada 1977–2011, România, descărcate de la IPUMS (Ruggles et al., 2024) pentru a le utiliza în exemplificarea unor diferite analize statistici și a putea face exerciții de înțelegere și interpretare a rezultatelor.

Începem cu tipuri de variabile, statistici descriptive, pregătirea datelor, tabele de contingență, regresii liniare multiple, corelații, analize cauzale simple, analize cauzale cu variabile mediatoare, analize factoriale exploratorii, modele cu variabile latente, analize bayesiene (algoritmul Markov Chain Monte Carlo) și încheiem cu structura unei lucrări de cercetare.

Pentru cei care doresc să înțeleagă cum funcționează algoritmii de bază de învățare automată (machine learning) prin exemple simple și explicate, ar putea să vadă cum funcționează învățarea automată nesupravegheată (unsupervised machine learning, a se vedea și Babu, 2020) prin analiza factorială exploratorie sau învățarea automată supravegheată (supervised machine learning, a se vedea și Li et al., 2021) prin algoritmii de regresie liniară multiplă, de regresie binară logistică sau de modele de ecuații structurale.

Corina Ilinca

CAPITOLUL A.

NIVELUL DE MĂSURARE AL VARIABILELOR

Ce fel de întrebări și variabile avem ne interesează pentru a putea decide ce fel de analize sunt necesare pentru a putea vedea ce există în date. Dacă avem o listă de caracteristici imposibil de ordonat, atunci nu putem realiza o numărare care să ne permită să vedem valori medii, de exemplu. Dacă am avea o listă de țări sau o listă de obiecte, atunci nu avem o ordine. Dacă avem o întrebare cu răspunsuri de la o cantitate mică la una mare, de exemplu de la un nivel scăzut de satisfacție la un nivel ridicat, putem spune că avem o ordine și astfel și media ar putea avea sens, însă fiind un număr mic de răspunsuri varianta de raportare care ne ajută mai mult în a înțelege cum sunt datele este cea în care prezentăm procentul de răspunsuri pentru fiecare răspuns și numărul total de cazuri. O abordare similară o avem și pentru variabilele de tip listă care nu pot fi ordonate. Dacă avem direct numere, cu valoarea 0 inclusă chiar, atunci putem număra exact de la un element la altul, iar media are sens și cu posibilitatea ca 0 să însemne că avem o absență a caracteristicii, de exemplu numărul de copii ai unei persoane sau venitul în ultima lună în lei. Pentru a identifica nivelul de măsurare al variabilelor, avem nevoie să știm ce fel de răspunsuri avem și să ne gândim dacă avem o ordine între categorii. Dacă nu avem o ordine, atunci variabila este una nominală. Dacă avem o ordine, dar nu avem numere cu semnificație cantitativă, atunci avem o variabilă ordinală. Dacă avem o variabilă cu o semnificație cantitativă cum este temperatura în grade Celsius, dar 0 nu înseamnă absența caracteristicii, atunci avem o variabilă de tip interval. Dacă avem o variabilă numerică unde 0 înseamnă

absența caracteristicii cum este numărul de limitări în activitățile de zi cu zi, atunci avem o variabilă de tip raport. Mai multe detalii și exemple puteți găsi în capitolul 7 al cărții scrise de doamna profesoară Cosima Rughiniș (2007). Mai există și o variabilă de tip dummy care are două variante de răspuns 1 pentru prezența unei caracteristici și 0 pentru absența caracteristicii, de exemplu, important este să avem doar două coduri și acestea să fie 0 și 1, iar media pentru această variabilă este proporția asociată codului 1 (Garavaglia & Sharma, 1998).

Sursa pentru variabile introduse în acest capitol este World Values Survey 2018 (Haerpfer et al., 2022).

Exerciții

A1. Întrebarea „Vă rugam să ne spuneți cât de importante sunt următoarele lucruri în viața dvs.: Timpul liber", având variantele de răspuns 1 pentru „Foarte important", 2 pentru „Destul de important", 3 pentru „Puțin important" și 4 pentru „Deloc important", are un nivel de măsurare:

a. Nominal
b. Ordinal
c. Interval
d. Raport

A2. Întrebarea „Pe lista următoare sunt trecute diferite grupuri de persoane. Ați putea, vă rugăm, să alegeți pe aceia pe care nu i-ați dori ca vecini? (notează un răspuns pentru fiecare grup): Cupluri necăsătorite care trăiesc împreună", având variantele de răspuns 1 pentru „Menționat" și 2 pentru „Nemenționat", are nivelul de măsurare:

a. Dihotomic
b. Ordinal
c. Interval
d. Raport

A3. Întrebarea „În ce măsură sunteți de acord cu următoarele afirmații? Unul dintre principalele mele scopuri în viață este să îmi fac părinții mândri de mine", cu variantele de răspuns 1 pentru „În foarte mare măsură", 2 pentru „În mare măsură", 3 pentru „În mică măsură" și 4 pentru „În foarte mică măsură", are nivelul de măsurare:

a. Dihotomic
b. Ordinal
c. Interval
d. Raport

A4. Întrebarea „Vă voi citi o listă cu unele schimbări care ar putea să apară în modul nostru de viață în viitorul apropiat. Dacă acestea s-ar produce, ați spune că ar fi bine, rău sau v-ar fi indiferent? Importanța muncii în viața noastră să scadă" cu variantele de răspuns 1 pentru „Bine", 2 pentru „Indiferent" și 3 pentru „Rău", are nivelul de măsurare:

a. Nominal
b. Ordinal
c. Interval
d. Raport

A5. Întrebarea „Luând în considerare toate aspectele vieții dvs., ați spune că sunteți... (citește toate variantele și notează un singur răspuns)", având variantele de răspuns 1 pentru „Foarte fericit", 2 pentru „Destul de fericit", 3 pentru „Nu prea fericit" și 4 pentru „Deloc fericit", are nivelul de măsurare:

a. Nominal
b. Ordinal
c. Interval
d. Raport

A6. Întrebarea „Ați putea să îmi spuneți anul nașterii?" are un nivel de măsurare:

a. Nominal
b. Ordinal

c. Interval
d. Raport

A7. Întrebarea „Sexul respondentului" cu variantele de răspuns 1 pentru „Femeie" și 0 pentru „Bărbat" are un nivel de măsurare:
a. Dummy
b. Ordinal
c. Interval
d. Raport

A8. Întrebarea „Sunteți cetățean/cetățeană român/ă?" având variantele de răspuns 1 pentru „Da, sunt cetățean/ă roman/ă" și 0 pentru „Nu, nu sunt cetățean/ă român/ă" are nivelul de măsurare:
a. Dummy
b. Ordinal
c. Interval
d. Raport

A9. Întrebarea „Câți copii aveți?" are nivelul de măsurare:
a. Nominal
b. Ordinal
c. Interval
d. Raport

A10. Întrebarea „Ce limbă vorbiți în mod obișnuit acasă?", cu variantele de răspuns 1 pentru „română", 2 pentru „maghiară", 3 pentru „romanes", 4 pentru germană, are nivelul de măsurare:
a. Nominal
b. Ordinal
c. Interval
d. Raport

A11. Întrebarea „Cât de des vă uitați la TV sau citiți ziare sau site-uri sau ascultați radio din alte țări?", având variantele de răspuns 1

pentru „Niciodată", 2 pentru „Mai rar", 3 pentru „Lunar", 4 pentru „Săptămânal" și 5 pentru „Zilnic", are nivelul de măsurare:
a. Nominal
b. Ordinal
c. Interval
d. Raport

A12. Întrebarea „Aceasta înseamnă că aveți vârsta de ____ ani (notați vârsta din două cifre)" are nivelul de măsurare:
a. Nominal
b. Ordinal
c. Interval/ raport

A13. Întrebarea „Locuiți împreună cu părinții sau cu socrii dvs. ? (Un singur răspuns)", având variantele de răspuns 1 pentru „Nu", 2 pentru „Da, împreună cu părinții", 3 pentru „Da, împreună cu socrii" și 4 pentru „Da, împreună cu părinții și cu socrii" are nivelul de măsurare:
a. Nominal
b. Ordinal
c. Interval
d. Raport

A14. Întrebarea „Din motive de securitate, ați făcut vreunul din următoarele lucruri? Am preferat să nu ies noaptea din casă", având variantele de răspuns 1 pentru „Da" și 0 pentru „Nu", are nivelul de măsurare:
a. Dummy
b. Ordinal
c. Interval
d. Raport

A15. Întrebarea „În general vorbind, ați spune că se poate avea încredere în cei mai mulți oameni sau că e mai bine să fii atent în relațiile cu oamenii? (Un singur răspuns)", cu variantele de răs-

puns 1 pentru „Se poate avea încredere în cei mai mulți oameni" și 0 pentru „E mai bine să fii atent în relațiile cu oamenii", are nivelul de măsurare:

a. Dummy
b. Ordinal
c. Interval
d. Raport

A16. Este incorect să calculăm media pentru o variabilă măsurată la nivel:

a. Nominal
b. Interval
c. Raport

A17. Care este nivelul de măsurare al unei variabile care măsoară gradul de informare cu privire la obligațiile unei cetățean având variantele de răspuns 1 pentru neinformat(ă), 2 pentru nici neinformat(ă), nici informat(ă), 3 pentru informat(ă)?

a. Raport
b. Interval
c. Ordinal
d. Nominal

A18. Media variabilei de tip dummy (coduri 1 și 0) este egală cu valoarea procentului de cazuri care au codul 1 înmulțită cu 100. De exemplu, dacă media variabilei genul respondentului este 0.51, atunci avem 51% femei (codul 1). Afirmația este:

a. adevărată
b. falsă

Răspunsuri corecte Capitol A

A1. b
A2. a
A3. b
A4. b
A5. b
A6. c
A7. a
A8. a
A9. d
A10. a
A11. b
A12. c
A13. a
A14. a
A15. a
A16. a
A17. c
A18. a

CAPITOLUL B.

STATISTICI DESCRIPTIVE

Pentru a observa tendințele de răspuns la o întrebare, avem nevoie să prezentăm măsura în care anumite răspunsuri au fost selectate. Dacă avem o întrebare cu un număr mic de variante de răspuns, să zicem chiar două: 1. Da, 0. Nu, atunci prezentăm procente pentru fiecare răspuns și numărul total de cazuri. Dacă avem trei, patru, cinci variante de răspuns, poate și încă puțin mai multe, atunci am putea proceda similar, să prezentăm procentele și numărul de cazuri. Dacă însă avem întrebări cu răspunsuri numerice variate, de exemplu dacă întrebăm vârsta sau valoarea veniturilor, atunci vom aveam multe răspunsuri diferite și vom prezenta următoarele statistici: media (mean sau average), abaterea standard/ A.S. (standard deviation, S.D. în limba engleză – cât de depărtate sunt valorile față de medie), minimum (min), maximum (max) și numărul de cazuri (N). Putem adăuga mediana dacă avem valori extreme ce dorim să le păstrăm, astfel observând diferențe între medie și mediană.

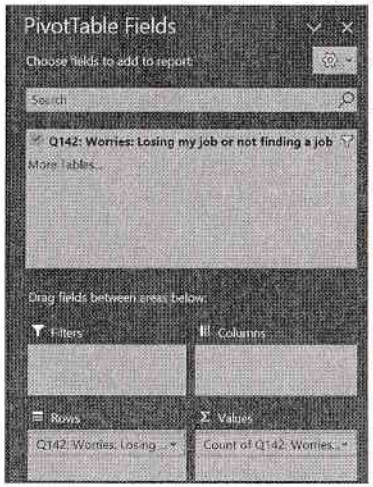
Ne interesează să știm ce măsoară întrebarea noastră, care este conceptul analizat. Astfel, pe lângă întrebarea adresată, trebuie să ne uităm și la variantele de răspuns, în ce ordine sunt și care sunt codurile atașate. De ce este important? Deoarece putem avea o întrebare care se referă la mulțumirea față de serviciile primite, dar care are variantele de răspuns ordonate de la 1 foarte mulțumit/ă la 4 foarte nemulțumit/ă, astfel că avem un cod mare negativ, ce ne arată că noi măsurăm prin întrebare și variante de răspuns nemulțumirea față de servicii având codul mare 4 referitor la nemulțumire. În ce măsură oamenii sunt nemulțumiți cu

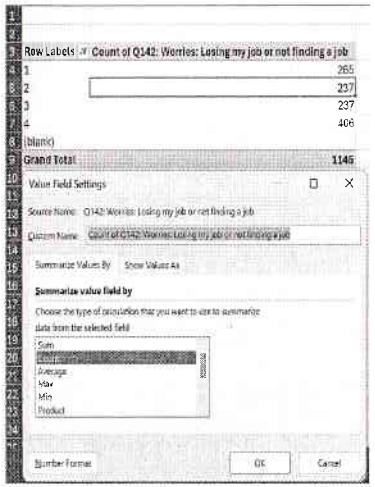
serviciile, de la nemulțumire mică, la nemulțumire mare – această idee o includem în analiză în acest mod de măsurare. Dar dacă noi vrem să analizăm mulțumirea față de servicii, atunci trebuie să inversăm scala, să avem codul mic 1 pentru mulțumire scăzută, iar codul mare 4 pentru mulțumire ridicată.

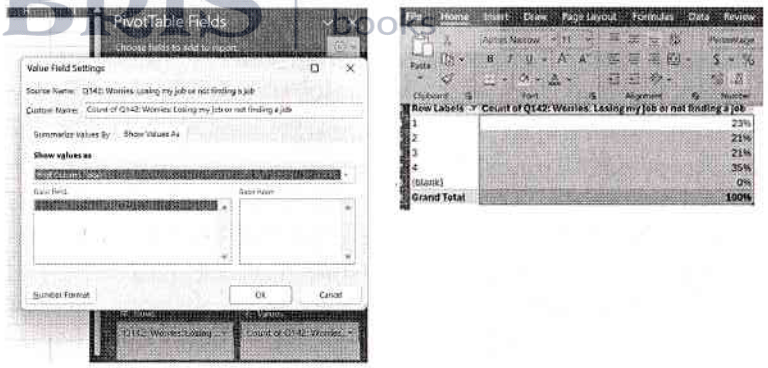
Tabelul de frecvență rezultat din SPSS ne arată numărul efectiv de cazuri pentru fiecare variantă de răspuns („Frequency), procentul („Percent") pentru variantă de răspuns inclusiv pentru valorile lipsă definite în prealabil în program („Missing"), procentele valide care se raportează de fapt în lucrări/ articole științifice („Valid Percent") ce nu iau în considerare datele pentru valorile lipsă/ missing data și procentele cumulate/ adunate până la fiecare cod valid inclus („Cumulative Percent"), pentru interpretarea lor trebuie să ținem astfel cont că sunt adunate toate variantele incluse până atunci și semnificația lor.

Pentru a realiza tabele de frecvență în Excel, putem folosi de la meniul Insert, opțiunea Pivot tabel. Selectăm în prealabil datele, fie tot tabelul cu date, fie doar coloana/ coloanele de interes, apoi mergem la meniul respectiv, selectăm opțiunea Pivot Tabel. Pentru că am selectat datele deja nu trebuie să îi mai spunem la apariția ferestrei după selecția opțiunii unde sunt datele, iar a doua opțiune în fereastră se referă la rezultat și unde dorim să îl avem: într-o filă nouă sau lângă date. O filă/ sheet nouă este mai utilă în general, fiind și opțiunea de bază. Astfel, după ce selectăm datele, apoi opțiunea Pivot tabel, putem cere să mergem mai departe apăsând butonul OK din fereastra apărută. Avem apoi în foaia nouă o zonă în dreapta unde putem vedea lista de variabile/ întrebări. Pe „Rows"/ linii, putem trage o variabilă din listă, apoi tragem din nou aceeași variabilă la „Values" (valori). La „Values" apăsăm pe săgeata din dreapta denumirii variabilei, găsim „Value Field Settings", iar la „Summarize Values By" să avem „Count" sau număr de cazuri (nu „Sum" sau sumă), iar dacă dorim procente, adăugăm la „Show Values As" în loc de „No calculation", „% of Column Total"). Din tabel, unde apare Row sau Column Labels, putem selecta codurile ce dorim să apară în tabel. Putem scoate

zecimalele de la meniul Home, butonul  . Pentru realizarea statisticilor dedicate variabilelor numerice în Excel, putem folosi funcțiile =average(B2:B10) pentru medie, =stdev(B2:B10) pentru abatere standard, =min(B2:B10) pentru minimum, =max(B2:B10) pentru maximum, =count(B2:B10) pentru număr de cazuri N, =percentile(B2:B10,0.05) pentru percentila 5% sau =percentile(B2:B10,0.9) pentru percentila 90%, considerând că datele sunt în căsuțele de la B2 la B10 și fiecare formulă este trecută în căsuța ei în foaia cu date (pentru formula percentilei sau altele, dacă folosim punct, virgulă sau punct și virgulă, trebuie să verificați că exemplul oferit de program este similar sau punctul se inversează cu virgula pentru marcarea zecimalelor sau virgula cu punct și virgulă pentru marcarea diferențierii elementelor în funcție).







Analizele realizate în acest capitol și prezentate în formă tabelară utilizează datele din World Values Survey (WVS), valul 7, România, anul 2018 (Haerpfer et al., 2022). Procentele pot fi prezentate și în grafice sau alte forme vizuale, însă acestea tot pe bază de procente sunt create, iar în articolele științifice formatul tabelar este prevalent. Pentru crearea de grafice, trebuie să aveți în vedere și modalitatea de diseminare a materialelor, dacă aceasta este color sau nu, mărimea fontului, verificând astfel dacă rezultatul este ușor de citit și corect realizat.

Exerciții

Q3 Important în viață: Timpul liber

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Foarte important	561	44.6	44.7	44.7
	2 Destul de important	507	40.3	40.4	85.1
	3 Puțin important	167	13.3	13.3	98.4
	4 Deloc important	20	1.6	1.6	100.0
	Total	1255	99.8	100.0	
Missing	-1 Nu știu	2	.2		
Total		1257	100.0		

Sursa datelor: WVS, 2018 (Haerpfer et al., 2022)

B1. Din tabelul rezultat din programul SPSS, fără alte editări, putem spune:

a. Cei mai mulți respondenți consideră că timpul liber nu este important.
b. Numărul total de respondenți este 100.
c. 44.7% dintre respondenți consideră că timpul liber este foarte important.

B2. Variabila Q3 este preluată din baza de date așa cum au fost datele colectate. Care dintre următoarele variante de răspuns este corectă?

a. 98.4% dintre respondenți consideră că timpul liber este puțin important.
b. Variabila măsoară lipsa importanței timpului liber în varianta prezentată în tabel

Q47 Cum ați descrie starea dvs. de sănătate în prezent?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Foarte bună	235	18.7	18.8	18.8
	2 Bună	630	50.1	50.3	69.1
	3 Nu prea bună	305	24.3	24.4	93.5
	4 Proastă	44	3.5	3.5	96.9
	5 Foarte proastă	38	3.1	3.1	100.0
	Total	1253	99.6	100.0	
Missing	-2 Fără răspuns	2	.2		
	-1 Nu știu	2	.2		
	Total	4	.4		
Total		1257	100.0		

Sursa datelor: WVS, 2018 (Haerpfer et al., 2022)

B3. Din tabelul oferit de SPSS, putem spune:

a. Jumătate dintre respondenți consideră că starea lor de sănătate în prezent este bună.
b. 7 din 10 respondenți consideră că au o stare de sănătate precară.
c. 38 de respondenți consideră că au un nivel de sănătate excelent.
d. Numărul total de respondenți cu răspunsuri valide este 1257.