

Cuprins

Despre autori	13
Introducere	15
Bookdown	17
Informații despre software	17
1 Concepte de bază în statistică	21
1.1 Ce este statistica?	21
1.2 Începutul demersului statistic	22
1.3 Mărimi relative	27
1.3.1 Mărimi relative de structură	29
1.3.2 Mărimi relative de intensitate	31
1.3.3 Mărimi relative de coordonare	32
1.3.4 Mărimi relative ale dinamicii și ale planului	33
2 Sistematizarea datelor	35
2.1 Metode și procedee de sistematizare	35
2.2 Serii de distribuție	38
2.3 Indicatorii frecvenței	45
2.4 Histograma	50
3 Vizualizarea datelor	55
3.1 Reprezentarea datelor sub formă de tabele	55
3.2 Reprezentarea grafică a datelor	56
4 Indicatorii de nivel. Tendința centrală	65
4.1 Indicatorii tendinței centrale	66
4.1.1 Media	66
4.1.2 Mediana	79
4.1.3 Modul	85
4.2 Quantilele	87

4.2.1	Quartilele	88
4.3	Aplicații - în R	89
5	Indicatorii variației	97
5.1	Indicatorii variației	97
5.1.1	Indicatorii simpli ai variației	97
5.1.2	Indicatorii sintetici ai variației	100
5.2	Indicatorii formei de repartiție	108
5.2.1	Indicatorul de asimetrie - Skewness	109
5.2.2	Indicatorul de boltire - Kurtosis	110
5.3	Indicatorii concentrării	112
5.3.1	Curba lui Lorenz	113
5.3.2	Coeficientul Gini	115
6	Distribuții de probabilitate	117
6.1	Conceptul de probabilitate	117
6.2	Distribuții de probabilitate	118
6.2.1	Distribuții discrete	118
6.2.2	Distribuții continue	131
7	Inferența statistică	147
7.1	Parametri de selecție	148
7.2	Componentele cercetării statistice selective	154
7.2.1	Selecția	154
7.2.2	Extinderea rezultatelor de selecție	157
7.2.3	Mărimea eșantionului	176
8	Indicii statistici	181
8.1	Indicii individuali	182
8.2	Indicii de grup	184
8.2.1	Indicele agregat al variabilei complexe	184
8.2.2	Indicii agregați ai factorilor	186
9	Statistici oficiale	195
9.1	Statistica oficială în România	195
9.2	Statistica oficială în Uniunea Europeană	198
9.3	Surse de date în statistica oficială	198
9.4	Diseminarea informației statistice	201
ANEXA 1	Distribuția Bernoulli	205
ANEXA 2	Distribuția Poisson	209

CUPRINS

7

ANEXA 3 Distribuția Normală Standard

215

ANEXA 4 Distribuția t-Student

217

Lista tabelelor

1.1	Exemple de variabile calitative	24
1.2	Exemple de variabile cantitative discrete	25
1.3	Exemple de variabile cantitative continue	25
1.4	Producția industrială în anul t	29
1.5	Structura producției industriale în anul t	30
1.6	Distribuția salariaților după nivelul salariului brut, în anul t	30
2.1	Exemplul unei serii de distribuție pe variante	39
2.2	Exemplul unei serii de distribuție pe intervale	39
2.3	Corespondența între numărul de observații și numărul de grupe	40
2.4	Distribuția firmelor după numărul de salariați	43
2.5	Indicatorii de frecvență ai distribuției firmelor după numărul de salariați	48
2.6	Frecvențele cumulate ale distribuției firmelor după numărul de salariați	49
2.7	Tabelul frecvențelor distribuției firmelor după cifra de afaceri	50
3.1	Populația ocupată din România, după statutul profesional, pe sexe și medii de rezidență	56
4.1	Distribuția firmelor după cifra de afaceri	69
4.2	Algoritmul de calcul al mediei unei serii de distribuție	73
4.3	Distribuția firmelor după valoarea totală a activelor	77
4.4	Algoritmul de calcul al medianei PIB/locuitor	83
4.5	Distribuția țărilor după numărul utilizatorilor de Internet ce revin la 1000 locuitori	84
4.6	Algoritmul de calcul al medianei numărului de utilizatori Internet	84
4.7	Distribuția populației ocupate, după vârstă	87
4.8	Distribuția persoanelor, după venitul lunar	89
5.1	Abaterile individuale ale cifrei de afaceri	101
5.2	Algoritm de calcul al indicatorilor variației (cazul unei serii simple)	105
5.3	Algoritm de calcul al indicatorilor variației (cazul unei serii simple)	107

5.4	Distribuția populației ocupate, pe grupe de vârstă, în anul 2013 . . .	113
5.5	Algoritmul de calcul pentru construirea curbei de concentrare a populației ocupate, pe grupe de vârstă	114
5.6	Algoritmul de calcul al coeficientului Gini	116
6.1	Distribuția Bernoulli de probabilitate pentru exemplul celor nouă clienți	123
7.1	Distribuția pe sexe și grupe de vârstă a salariaților unei firme . . .	159
7.2	Extinderea datelor din eșantion prin coeficienți de extindere	160
8.1	Vânzările de produse alimentare ale magazinului XX în t_0 și t_1 . . .	186
8.2	Vânzările de produse alimentare și indicii individuali ai valorii, prețurilor și cantităților	186
8.3	Algoritm de calcul al indicilor agregați	190
9.1	Distribuția Bernoulli	206
9.2	Distribuția Bernoulli - continuare	207
9.3	Probabilitățile de distribuție	209
9.4	Probabilitățile cumulate de distribuție	210
9.5	Probabilitățile cumulate de distribuție - continuare	211
9.6	Probabilitățile cumulate de distribuție - continuare	212
9.7	Probabilitățile cumulate de distribuție - continuare	213
9.8	Probabilitățile cumulate de distribuție - continuare	214
9.9	Tabelul distribuției normale standard	215
9.10	Tabelul distribuției t	217

Lista figurilor

2.1	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (frecvențe absolute)	50
2.2	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (frecvențe relative)	51
2.3	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (frecvențe cumulate crescător)	51
2.4	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (frecvențe cumulate descrescător)	51
2.5	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (poligonul frecvențelor)	52
2.6	Reprezentarea grafică a distribuției firmelor după cifra de afaceri (curba frecvențelor cumulate)	52
3.1	Productivitatea muncii în Uniunea Europeană	57
3.2	Speranța de viață la naștere, pe sexe, în perioada 2000-2016	58
3.3	Ponderea populației de 0-14 ani și a populației de 65 ani și peste în totalul populației rezidente, la 1 ianuarie, în perioada 2008-2017	58
3.4	Numărul de șomeri în orașul M (la sfârșitul trimestrului), în perioada 2009-2012 (persoane)	59
3.5	Numărul maxim de ore de studiu în instituțiile de învățământ, pe niveluri de educație	60
3.6	Distribuțiile valorilor observate și estimate ale unei variabile aleatorii	60
3.7	Structura pe grupe de vârstă a populației ocupate	60
3.8	Structura populației plecată la studii în străinătate pentru 12 luni și peste, pe niveluri de educație	61
3.9	Structura populației active, plecată în străinătate pentru 6-12 luni, după participarea la activitatea economică	62
3.10	Evoluția indicelui parității de gen pentru persoanele ocupate plecate la lucru în statele UE-27 pentru 6-12 luni	62
3.11	Numărul căsătoriilor din România, după grupa de vârstă a soților	63
3.12	Densitatea demografică, pe județe	63

4.1	Mediana volumului vânzărilor	81
4.2	Mediana cifrei de afaceri	82
4.3	88
5.1	Reprezentarea grafică a formelor de distribuție	111
5.2	Reprezentarea grafică a curbei de concentrare a populației ocupate, pe grupe de vârstă	114
6.1	Reprezentarea grafică a distribuției binomiale	122
6.2	Reprezentarea grafică a funcției densitate de probabilitate în cazul distribuției binomiale	122
6.3	Reprezentarea grafică a distribuției binomiale (de tip Bernoulli) pentru exemplul celor nouă clienți	124
6.4	Reprezentarea grafică a distribuției de probabilitate Poisson	130
6.5	Reprezentarea grafică a funcției densitate de probabilitate în cazul distribuției Poisson	130
6.6	Reprezentarea grafică a comparației între distribuția binomială și distribuția Poisson	131
6.7	Reprezentarea grafică funcției densitate de probabilitate a unei variabile aleatorii care urmează o distribuție uniformă	133
6.8	Reprezentarea grafică a funcției densitate de probabilitate pentru o variabilă aleatorie uniform distribuită în intervalul $[0, 1]$	136
6.9	O reprezentare utilă a distribuției normale standard ($\mu = 0, \sigma^2 = 1$)	142
6.10	Reprezentarea grafică a distribuției unei variabile aleatorii într-o rețea de probabilitate (normal probability plot)	144
6.11	Distribuția exponențială - reprezentarea grafică a probabilităților cumulate	145

Despre autori

Nicoleta Caragea este conferențiar universitar la Facultatea de Management Financiar, Universitatea Ecologică din București și expert în cadrul Institutului Național de Statistică.

Titlul de doctor în Economie l-a obținut sub egida Academiei Române, Institutul de Economie Națională. Activitatea sa didactică se concentrează, în principal, în domeniul statisticii, prin cursuri și seminarii la programele de licență și masterat (Statistică, Statistică economică, Statistică socială, Analiză economico-financiară).

În activitatea de cercetare a participat ca expert național în diverse proiecte de cercetare științifică, workshop-uri și conferințe organizate de instituții internaționale de prestigiu cu activitate statistică (EUROSTAT, OCDE, OMS, Banca Mondială, IMF, UNICEF-UIS).

Inițiatoare a grupului de cercetare R-omanian team, a organizat diferite workshopuri și conferințe internaționale, având ca scop principal introducerea mediului de analiză statistică (R) în statistica oficială din România (<http://www.r-project.ro>).

Rezultatele activității sale de cercetare au fost publicate în numeroase reviste prestigioase din țară și din străinătate (Revista Română de Economie, Revista Romană de Statistică, Economic Computation and Economic Cybernetics Studies and Research), precum și în baze de date internaționale recunoscute (Clarivate Analytics, RePEC, Scopus, DOAJ, Index Copernicus, Elsevier etc.).



Scopus Author ID: 56299742400
ResearcherID: C-9002-2018
<https://orcid.org/0000-0003-3199-4186>

Ciprian Alexandru este conferențiar universitar la Facultatea de Management Financiar, Universitatea Ecologică din București și expert în cadrul Institutului Național de Statistică.

Titlul de doctor în Economie l-a obținut sub egida Academiei Române, Institutul de Economie Națională. A participat la un program de studii postdoctorale în care a implementat utilizarea software-ului R ca instrument de analiză a evoluției indicilor bursieri.

Activitatea sa didactică se concentrează, în principal, în domeniul burselor de valori, prin cursuri și seminarii la programele de licență și masterat (Piețe de capital, Managementul Portofoliului, Piețe internaționale de capital).

A participat la diverse proiecte de cercetare, workshop-uri, conferințe naționale și internaționale. Activitatea de cercetare a fost pusă în valoare prin publicarea studiilor în reviste din țară și din Europa, precum și în baze de date internaționale recunoscute (RePEC, DOAJ, EBSCO).

În prezent, în cadrul Institutului Național de Statistică, participă ca expert în proiecte BigData și utilizează software-ul de analiză statistică R pentru Data cleaning, Data Matching, Web Scraping, analize de date și vizualizare, Data mining, Data integration, data processing, data validation, dar și utilizarea datelor din sursele administrative pentru realizarea de statistici oficiale.

ResearcherID: V-2168-2017

<https://orcid.org/0000-0001-8215-6671>



Introducere

“... acum nu mai e nimic nou de descoperit;
tot ce rămâne e doar măsurătoarea din ce în ce mai precisă”

— Lord Kelvin (1894)

Cartea tipărită merită răsfodită. Trăim în vremea în care internetul facilitează comunicarea globală, informația fiind disponibilă oricând și oricum. Toată lumea, de la oameni de știință și până la copii de vârstă școlară primesc și oferă informații și propagă idei pe calea internetului. Tirajele publicațiilor, cărților și manualelor tipărite sunt în scădere în întreaga lume, în timp ce postările online captează atenția omenirii.

Obiectivul principal al cărții pe care o propun este de a fi un ghid cuprinzător, în termeni de concepte și tehnici, reprezentativ și, mai ales, practic, în ceea ce privește utilizarea instrumentelor software de analiză statistică, R fiind principalul software utilizat pentru aplicațiile propuse. Ca abordare generală, cartea prezintă principalele concepte utilizate în statistică, cu exemple și explicații descriptive. Exemplele din viața economică - cele mai multe dintre ele bazate pe date statistice reale - problemele rezolvate, dar și cele propuse, acoperă o arie cuprinzătoare de tematici, cititorul având șansa de a fi introdus în sfera aplicativă a conceptelor teoretice parcurse.

Cartea este destinată tuturor celor care doresc să înțeleagă, prin mijloace științifice, fenomenele economice și sociale, sub aspectul măsurării cantitative și din perspectiva determinării cauzale. Deși se adresează, în principal, studenților care se pregătesc să devină specialiști în științele economice, lucrarea este utilă și celor care își propun să cunoască un domeniu atât de frumos și de captivant. Tocmai nevoia de informații, din ce în ce mai complexe, dar și posibilitățile de calcul avansat cu ajutorul soft-urilor tot mai performante, au condus la crearea unui bazin imens de date care pot fi cu ușurință exploatare pe baza analizei statistice. Poate că acesta este și motivul pentru care statistica rămâne o disciplină percepută ca fiind adesea prea matematizată, destinată specialiștilor. Pentru mulți cititori, mai ales dintre cei care nu au o formare bazată pe un aparat matematic, studiul fenomenelor economice prin metode statistice și matematice, presupune un efort

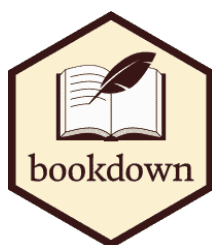
deosebit. Din acest motiv, am încercat să tratez aspectele teoretice, dar și problemele cu aplicație practică din sfera economică, într-o manieră simplă, accesibilă. Așadar, lucrarea are menirea de a facilita înțelegerea conceptelor fundamentale cu care operează statistica, utilizarea adecvată a metodelor de analiză statistică, precum și interpretarea corectă a rezultatelor, în vederea cunoașterii modului de manifestare a fenomenelor.

Nicoleta Caragea
Septembrie, 2018

Bookdown

Această carte a fost editată cu ajutorul pachetului R **bookdown** (Xie, 2015). Cartea are la bază manualul *Statistică - concepte și metode de analiză a datelor* (Caragea, 2015).

Pachetul R **bookdown** este integrat R Markdown (<http://rmarkdown.rstudio.com>). Documentele elaborate pe baza acestui tip de instrumentar de editare sunt pe deplin reproductibile și dau posibilitatea creării unor formate de ieșire diverse (PDF/HTML/Word/...). Informații suplimentare referitoare la utilizarea pachetului **bookdown** se pot găsi la adresa: <https://bookdown.org>.



Informații despre software

Software-ul R a devenit în prezent unul dintre cele mai utilizate instrumente de analiză statistică, fiind utilizat în statisticile oficiale, în mediile universitare și de cercetare academică, dar și în mediul de afaceri. Acest manual este destinat tuturor celor care doresc să învețe statistica, fiind un material introductiv de studiu, care prezintă un spectru larg de exemple, prezentări grafice și analiză a datelor, dezvoltate cu ajutorul R.

Aplicațiile din această carte utilizează R, ceea ce înseamnă că pentru reproducerea acestora va fi nevoie de instalarea R pe calculatorul pe care lucrați.

R este un sistem pentru analize statistice și reprezentare grafică creat de către Ross Ihaka și Robert Gentleman, profesori de statistică la Universitatea Auckland din Noua Zeelandă¹.

R este considerat un dialect al limbajului S creat de AT&T Bell Laboratories. S este disponibil sub forma software-ului S-PLUS, comercializat de compania Insightful. Există diferențe importante între cele două limbaje, R și S: acestea sunt documentate de către Ihaka & Gentleman (1996) sau se regăsesc în R-FAQ².

Astfel, numele limbajului R provine de la inițiala prenumelui creatorilor, dar este

¹Ihaka R. & Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314.

²R-FAQ

totodată și un omagiu adus limbajului S.

În primul rând, R este open-source, fiind distribuit în mod gratuit sub licență *GNU - General Public Licence*³; dezvoltarea și distribuirea sunt în grija câtorva profesori și statisticieni, afiliați companiilor și universităților, cunoscuți sub denumirea generică de *R Development Core Team*.

Conform filosofiei *GNU*⁴, software-ul open-source este caracterizat de libertatea acordată utilizatorilor săi de a-l utiliza, copia, distribui, studia, modifica și îmbunătăți. Mai exact, este vorba de patru forme de libertate acordate utilizatorilor (Dușa et al., 2015):

- Libertatea de a utiliza programul, în orice scop (libertatea 0);
- Libertatea de a studia modul de funcționare a programului, și de a-l adapta nevoilor proprii (libertatea 1). Accesul la codul-sursă este o condiție pentru aceasta;
- Libertatea de a redistribui copii, în scopul ajutorării aproapelui tău (libertatea 2);
- Libertatea de a îmbunătăți programul, și de a pune îmbunătățirile la dispoziția publicului, în folosul întregii societăți (libertatea 3). Accesul la codul-sursă este o condiție pentru aceasta.

Faptul că este gratuit atrage automat avantajul competitiv în fața altor software-uri de analiză statistică, precum Stata, SAS și SPSS. Astfel, costurile alocate licenței de software dispar. R este denumit de către Norman Nie, unul dintre fondatorii SPSS și CEO al Revolution Analytics, “cel mai puternic și flexibil limbaj de programare statistică din lume” (în engleză *“the most powerful and flexible statistical programming language in the world”*).⁵ Dovadă a succesului pe care R îl are în știința datelor, s-au dezvoltat medii de integrare a acestuia în SAS și chiar SPSS. Este vorba despre modulul SAS/IML⁶, care integrează limbajul R în SAS, și despre *translate2R*, un serviciu de traducere a codului SPSS direct în R dezvoltat de compania *eoda*⁷. R are susținerea comunității științifice, dar și a multor companii internaționale. Dintre acestea, menționăm: Google, Facebook, Mozilla, Twitter, The New York Times, The Economist, NewScientist, Lloyd’s, Bing, Johnson&Johnson, Pfizer, Shell, Bank of America, Ford.⁸ R este susținut și de mediul academic. Marile universități din lume sprijină R, la fel cum sprijină și alte inițiative sau software-uri open-source, precum sistemul de operare Linux sau sistemul de preparare a documentelor L^AT_EX.

³GNU

⁴GNU Philosophy

⁵Smith, D., 2010, "R is Hot", Revolution Analytics

⁶SAS/IML Module

⁷translate2R - eoda

⁸Revolution Analytics, "Companies Using R"