DISTRIBUȚIA NORMALĂ ȘI EVALUAREA ACESTEIA

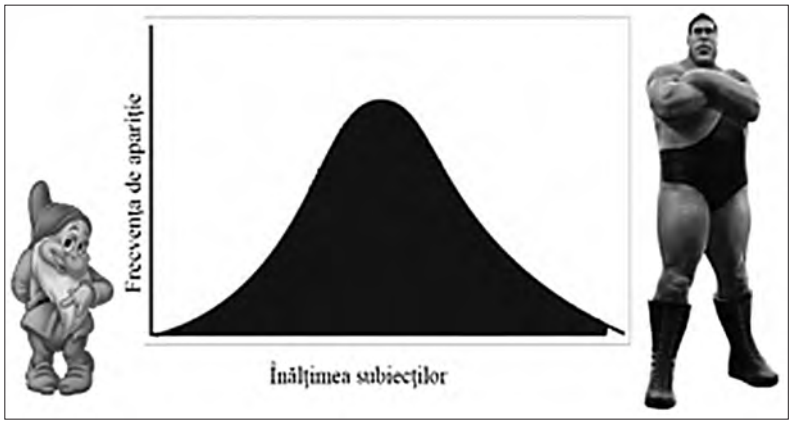
Cristian OPARIUC-DAN

Aproape permanent adusă în discuție, chiar dacă este destul de puțin înțeleasă, distribuția normală este prezentă în toate analizele de date și am putea discuta mult asupra relevanței sale din punct de vedere matematic în anumite ramuri ale disciplinelor socio-umane, însă este cert faptul că reprezintă una dintre cele mai utilizate distribuții de probabilități, uneori fiind și singura studiată mai detaliat. Despre modelul teoretic al distribuției normale am vorbit deja în curs, așadar nu vom reveni asupra unor noțiuni cunoscute, însă *primul lucru pe care va trebui să-l avem în vedere este acela că distribuția normală este o distribuție de probabilități pentru variabile continui, lipsa continuității conducând la utilizarea eronată a acestui model*.

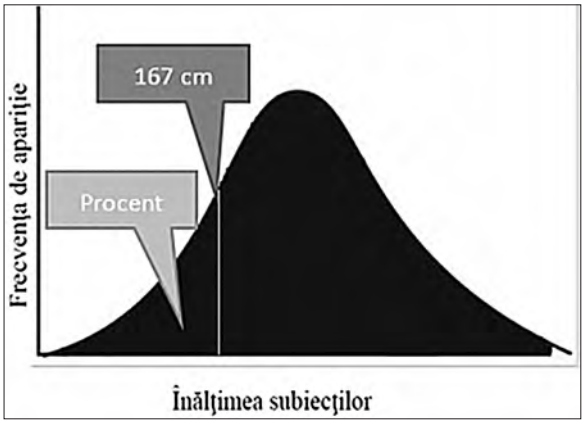
Să reluăm exemplul măsurării înălțimii bărbaților și să considerăm, pur teoretic, că avem acces la toată populația masculină a planetei, ce a împlinit vârsta de 18 ani. În cazul în care am reprezenta grafic, printr-un grafic cu bare, frecvența absolută de apariție a fiecărui scor, barele s-ar îngusta atât de mult între cel mai scund bărbat și cel mai înalt bărbat, încât, la un moment dat, s-ar uni între ele, rezultând o imagine similară cu cea din figura alăturată. Dacă am încerca să trasăm un poligon al frecvențelor, unind vârfurile barelor, ar rezulta o curbă în formă de clopot, simetrică față de tendința centrală, aceasta fiind distribuția normală sau distribuția gaussiană.

Dacă în loc de frecvențe absolute am reprezenta frecvențele relative, exprimate procentual, atunci întreaga zonă neagră ar însemna 100%, totalitatea bărbaților din lume. Deplasându-ne pe axa OX, de la stânga la dreapta, am putea reprezenta, imaginar, frecvențele relative cumulate, adică procentul până la care am întâlni, să spunem, bărbați cu înălțimea de 167 de centimetri.

Acest procent nu reprezintă altceva decât aria de sub curba distribuției normale corespunzătoare punctului în care înălțimea ajunge la valoarea 167. Dacă acest procent ar fi, să spunem, 35,4%, am susține că 35,4% dintre bărbați au cel mult înălțimea de 167 de centimetri, în termeni de probabilități afirmând că, în populație,



Distribuția frecvențelor absolute ale unei variabile continui



Reprezentarea zonelor de sub distribuția normală

avem o probabilitate de 0,354 să întâlnim o persoană cu înălțimea de cel puțin 167 de centimetri. Într-adevăr, deoarece toată aria de sub curba de distribuție normală reprezintă întregul, 100%, atunci există o corespondență între frecvența relativă și probabilitate și este perfect logic să discutăm despre probabilitatea cu care putem întâlni, într-o populație, un anumit interval de scoruri, continuitatea inducând ideea unei *densități de probabilitate*. Desigur, dus la extrem, probabilitatea cu care am întâlni în populație o persoană cu înălțimea până la 300 de centimetri se apropie de 1 (sau de 100%), deoarece orice om are mai puțin de 3 metri înălțime. Totuși, fiind o variabilă continuă, oricând există șanse infime (dar există) să se nască un gigant. Pe baza acestei logici, *distribuția normală (sau altele derivate din distribuția normală) este o distribuție intens utilizată în inferența statistică*.

Evaluarea practică a distribuției normale

Orice tip de analiză de date începe cu *inventarul statistic de bază* (indicatorii statistici de start) și continuă cu *analiza formei distribuției*, decizia asupra acesteia încheind capitolul *analizelor univariate*.

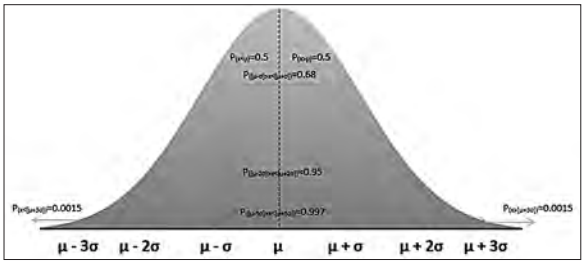
Să considerăm, de exemplu, coeficientul de inteligență, ce se presupune distribuit normal la nivelul populației, media sa fiind 100 și abaterea standard 15:

IQ~N(100, 15)

Probabilitatea cu care am întâlni în populație o persoană cu un coeficient de inteligență situat între o abatere standard la stânga mediei (100 – 15 = 85) și o abatere standard la dreapta mediei (100 + 15 = 115) este de 0,68, adică sunt 68% șanse ca, în mod întâmplător, mergând pe stradă, coeficientul de inteligență al unei persoane să fie cuprins între 85 și 115.

Ne-ar surprinde acest lucru? Evident că nu. Spunem că o persoană cu un IQ de la 85 la 100 are o inteligență medie, la fel cum putem spune și despre persoana cu un coeficient de inteligență cuprins între 100 și 115, doar că în primul caz inteligența este mediu-inferioară, iar în cel de-al doilea caz, mediu-superioară. Din acest motiv, *zona distribuției cuprinsă între o abatere standard la stânga mediei și o abatere standard la dreapta mediei poartă numele de* **zona valorilor medii** sau **zona probabilității comune, obișnuite, normale de apariție a evenimentului**. Pe măsură ce ne îndepărtăm de această zonă, lucrurile se modifică destul de serios.

Probabilitatea cu care am întâlni în populație o persoană cu un coeficient de



Reprezentarea distribuţiei standard normale

inteligenţă situat între două abateri standard la stânga mediei (100 – 2x15 = 70) şi două abateri standard la dreapta mediei (100 + 2x15 = 130) este de 0,954, adică sunt 95,4% şanse ca, în mod întâmplător, mergând pe stradă, coeficientul de inteligenţă al unei persoane să fie cuprins între 70 şi 130.

Iată cum, absolut întâmplător, mergând pe stradă, sunt 68% şanse să dăm peste o persoană cu inteligenţă medie şi 95,4% şanse să găsim persoane cu inteligenţă medie, persoane supradotate şi persoane întârziate mintal. Evident, între întârziatul mintal şi supradotat găsim aproape toată populaţia, dar care ar fi şansele ca, întâmplător, să găsim o persoană supradotată? Tot ceea ce avem de făcut este să ne referim doar la zona situată între o abatere standard şi două abateri standard la dreapta mediei:

$$95{,}4\% - 68\% = {}^{27{,}4\%}/_{2} = 13{,}7\%$$

Această zonă, cuprinsă între un IQ de 115 (**μ + σ**) şi un IQ de 130 (**μ + 2σ**) reprezintă **zona valorilor peste medie, zona scorurilor superioare** sau **zona probabilităţii de apariţie a unui eveniment rar**, existând, iată, 13,7% şanse ca, din întâmplare, să dăm peste o persoană supradotată.

Aceleaşi şanse le avem să dăm şi peste o persoană întârziată, având un IQ cuprins între 70 (**μ - 2σ**) şi 85 (**μ - σ**), regiunea fiind cunoscută drept **zona valorilor sub medie, zona scorurilor inferioare** sau tot **zona probabilităţii de apariţie a unui eveniment rar**, de data aceasta în rândul scorurilor mici.

Iată cum într-o distribuţie normală există 27,4% şanse să apară un eveniment rar, împărţite în mod egal, 13,7% şanse ca evenimentul rar să apară la nivelul valorilor mici, iar 13,7% şanse ca evenimentul rar să apară la nivelul valorilor mari. Cum atât inteligenţa la limită, cât şi şi cea superioară reprezintă excepţii, aceste zone poartă şi numele de **zona scorurilor accentuate**.

- Probabilitatea cu care am întâlni în populaţie o persoană cu un coeficient de inteligenţă situat între trei abateri standard la stânga mediei (100 – 3x15 = 55) şi trei abateri standard la dreapta mediei (100 + 3x15 = 145) este de 0,997, adică sunt 99,7% şanse ca, în mod întâmplător, mergând pe stradă, coeficientul de inteligenţă al unei persoane să fie cuprins între 55 şi 145.

Integrând şi această informaţie, spunem că, mergând pe stradă, avem 68% şanse să întâlnim persoane cu o inteligenţă medie, 95,4% şanse să întâlnim persoane cu inteligenţă medie, un intelect de limită sau cu o inteligenţă superioară şi 99,7% şanse să întâlnim persoane cu inteligenţă medie, cu intelect de limită, inteligenţă superioară, dar şi imbecili sau genii.

Ne preocupă, desigur, să aflăm şansele cu care putem găsi, în populaţie, un imbecil şi vom găsi 2,15%:

$$99{,}7\% - 95{,}4\% = {}^{4{,}3\%}/_{2} = 2{,}15\%$$

Cel puţin teoretic, 2,15% din populaţie prezintă o inteligenţă situată între două abateri standard şi trei abateri standard la stânga mediei, adică între un IQ de 55 (**μ - 3σ**) şi unul de 70 (**μ - 2σ**), aceasta fiind **zona valorilor extreme mici** sau **zona probabilităţii de apariţie a unui eveniment extrem de rar**.

Aceleași șanse le avem să găsim și un geniu, adică o persoană cu un IQ cuprins între 130 (**μ + 2σ**) și 145 (**μ + 3σ**), formând **zona valorilor extreme mari** sau **zona probabilității de apariție a unui eveniment extrem de rar**, însă în regiunea valorilor mari.

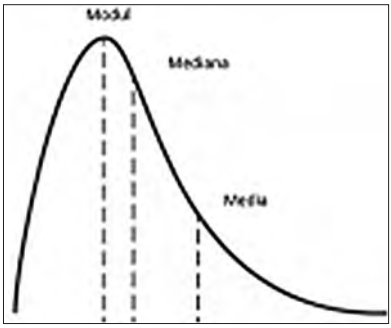
Deoarece distribuția normală este asimptotică la minus și plus infinit, înseamnă că, teoretic, numărul abaterilor standard este, și el, infinit, însă din considerente practice ne rezumăm la 3 abateri standard, deoarece toate evenimentele care depășesc trei abateri standard la stânga sau la dreapta au o probabilitate insignifiantă de apariție.

- Probabilitatea cu care am întâlni în populație o persoană cu un coeficient de inteligență situat după trei abateri standard la stânga mediei (< 55) și după trei abateri standard la dreapta mediei (> 145) este mai mică de 0,003, adică sunt 0,3% șanse ca, în mod întâmplător, mergând pe stradă, coeficientul de inteligență al unei persoane să fie mai mic de 55 și mai mare de 145.

Care ar fi șansele să întâlnim, pe stradă, o persoană precum Albert Einstein sau Stephen Hawking? Ei bine, doar 0,15%, la fel ca și șansele să întâlnim un idiot, regiunea fiind cunoscută drept **regiunea scorurilor aberante** sau **zona probabilității de apariție a evenimentelor aberante**.

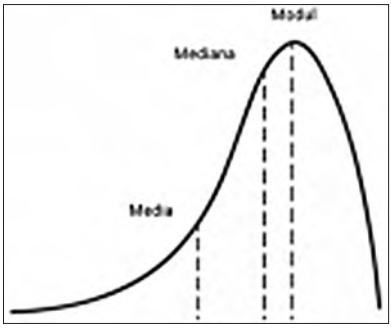
Indicatori statistici ai distribuției normale

Așadar, am decis că pentru a vorbi despre o distribuție normală, *suprafața din dreapta mediei trebuie să fie egală cu suprafața din stânga mediei*. Să ne imaginăm însă că într-o cercetare privind nivelul venitului, eșantionul îl include și pe Bill Gates. Ce se va întâmpla cu distribuția? Ei bine, va exista o persoană cu venituri mult mai mari în comparație cu

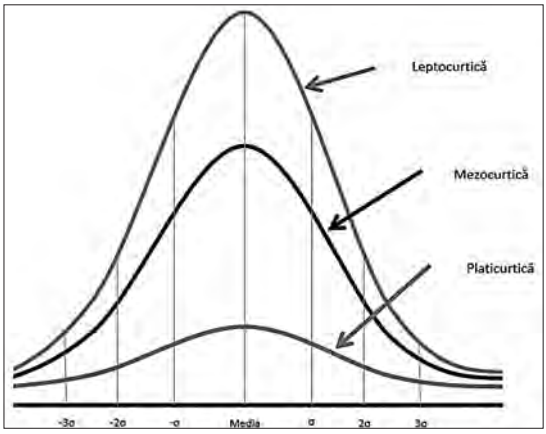


Distribuție asimetrică pozitiv

celelalte persoane, spunând că, în medie, toți sunt milionari, lucru, evident, fals. Venitul lui Bill Gates va „trage" media în sus, rezultând o așa-numită **distribuție asimetrică pozitiv** sau **distribuție skewness pozitiv**, distribuție în care avem de a face cu scoruri extreme mari și pentru care media nu este un bun indica-tor al tendinței centrale. Privind cu atenție această distribuție, constatăm că într-o distribuție asimetrică pozitiv, media are valoarea cea mai mare, lucru absolut firesc deoarece a fost inclus venitul lui Bill Gates. Acest indicator este urmat de mediană și, la final, de valoarea modală. Cu cât scorul extrem este mai mare în comparație cu celelalte scoruri, cu atât distanța dintre medie și mediană este mai mare, crescând și asimetria. Iată și motivul pentru care media nu mai este un bun indicator statistic al tendinței centrale, preferându-se mediana.



Distribuție asimetrică negativ



Distribuție leptocurtică (sus), mezocurtică (mijloc) și platicurtică (jos)

Dacă însă cercetarea include nivelul veniturilor celor mai bogați oameni de pe planetă, iar eșantionul va cuprinde și un om de afaceri din România care are o fabrică de pâine, probabil că veniturile acestuia sunt consistente, nimic de zis, însă sunt foarte departe de cele al lui Jeff Bezos, Bill Gates, Warren Buffett, Bernard Arnault și alții din acest grup. Ne aflăm în situația opusă, în care nivelul venitului omului de afaceri român se comportă ca un scor extrem mic, vorbind despre o **distribuție asimetrică negativ** sau **distribuție skewness negativ**. De această dată media, „trasă" în jos de omul de afaceri din România va avea cea mai mică valoare, fiind, la fel, un indicator nerelevant pentru tendința centrală, urmată, desigur, de mediană și mod.

Termenul „skewness" nu reprezintă cine știe ce concept sofisticat, însemnând, în limba engleză, „oblic", „înclinat", „asimetric". În limbaj statistic, prin **indicator al simetriei (skewness)** înțelegem *diferența dintre medie și mediană, raportată la abaterea standard* (McNemar, 1969), iar *o distribuție perfect simetrică are valoarea acestui indicator egală cu zero (skewness = 0*). *Valorile mai mari de zero (skewness > 0) reprezintă distribuții asimetrice pozitiv, în timp ce valorile mai mici de zero (skewness < 0) sunt caracteristice distribuțiilor asimetrice negativ*.

Din păcate, în practică nu întâlnim distribuții perfect simetrice, așadar întotdeauna valoarea acestui indicator va fi diferită de zero. Ca **reguli de bună practică**, *coeficienții de simetrie cuprinși între -0,5 și +0,5 indică o distribuție simetrică*. Dacă *valorile acestora se încadrează între -1,0 și -0,5 sau între +0,5 și +1,0, atunci vorbim despre o asimetrie moderată*, iar dacă *sunt mai mici de -1,0 și mai mari de +1,0, atunci asimetria este severă*.

Indicatorul de simetrie este un indicator estimat, prin urmare va fi însoțit și de eroarea de estimare, iar dacă dorim să fim mai riguroși în interpretarea sa, nu ne vom rezuma doar la regulile de bună practică, ci vom furniza măsuri mai relevante ale simetriei, pe care le vom discuta imediat.

Așadar, pornind de la indicatorii tendinței centrale, am descoperit relația dintre ei și am ajuns să o evaluăm sub aspectul simetriei, găsind **distribuții simetrice** sau **asimetrice** (pozitiv sau negativ), **distribuția normală fiind o distribuție simetrică**.

Au rămas însă indicatorii dispersiei, mai exact abaterea standard, iar pe baza acesteia putem evalua cât de „ascuțită" sau cât de „turtită" este distribuția (variabilita-

tea probabilității de apariției a scorurilor în jurul tendinței centrale), intrând în domeniul **indicatorilor boltirii**.

Să presupunem că într-o clasă, mediile elevilor la matematică sunt cuprinse între 5,7 și 6,3, fiind distribuite simetric. Se observă imediat că doar 0,7 puncte diferențiază competențele la matematică ale tuturor elevilor, nivelul acestora fiind extrem de omogen, iar abaterea standard, foarte mică.

Deja știm că dincolo de trei abateri standard la stânga sau la dreapta ne aflăm în zona evenimentelor extrem de rare, dar ce înseamnă, mai exact acest „dincolo de trei abateri standard la stânga sau la dreapta" atâta vreme cât cel mai mic scor este 5,7, iar cel mai mare scor este 6,3? Nimic altceva decât că o persoană cu media 5,5 la matematică este „total idioată" la această disciplină, iar una cu media 6,5 este „absolut genială". O astfel de distribuție se numește în termenii analizei de date **distribuție ascuțită** sau **distribuție leptocurtică** (vezi curba albastră de sus), iar probabilitatea de a se obține scoruri extreme este foarte mare. În definitiv, o medie de 7 va reprezenta un caz extrem foarte mare (situat la peste trei abateri standard la dreapta), iar o medie de 5 va reprezenta un caz extrem foarte mic (situat la peste trei abateri standard la stânga). Făcând o paralelă cu inteligența, este ca și cum media 5 ar însemna un idiot clinic, iar media 7 un geniu, fapt, desigur, mult exagerat. Iată și principala problemă a distribuției leptocurtice – *există o probabilitate foarte mare ca valori comune ale distribuției teoretice să pară drept scoruri extreme, acest lucru afectând puternic inferența statistică*. Pe baza acestei analize am putea generaliza că cel cu media 7 la matematică este un geniu, concluzie falsă, deoarece într-o altă clasă, media 7 la matematică ar putea corespunde, după cum este și normal, unui nivel mediu de cunoștințe.

Să trecem în cealaltă extremă și să ne imaginăm că elevii din clasa respectivă obțin medii la matematică între 1,0 și 10,0, acoperind întreaga amplitudine teoretică, iar abaterea standard este mare. Nivel performanțelor la această disciplină va fi foarte eterogen, iar această împrăștiere mare în jurul tendinței centrale face ca probabilitatea de a se obține media 5 să fie redusă, iar pe măsură ce ne apropiem de extreme, probabilitatea să scadă dramatic. Deja problemele acestei distribuții, denumită **distribuție plată** sau **distribuție platicurtică** (vezi curba roșie de jos) sunt evidente – *probabilitatea de apariție a unor scoruri extreme este aproape inexistentă*. Vor fi destui care au media 1, cea mai mică medie posibilă, astfel încât nu vom putea identifica „idioții" la această disciplină și, de asemenea, suficienți cu media 10, pentru a nu putea vorbi despre genii. Cu o astfel de distribuție, din nou, nu putem efectua inferențe statistice folosind logica testării semnificației statistice a ipotezei nule. Această logică implică utilizarea unui prag al semnificației statistice, arhicunoscutul *„p", care nu reprezintă altceva decât probabilitatea cu care s-ar putea obține un scor extrem în condițiile unei ipoteze nule adevărată*. Dacă nu avem scoruri extreme, atunci, în mod evident, orice raționament statistic am efectua pe baza acestei probabilități va fi, din start, eronat.

O distribuție normală este o **distribuție mediu boltită** sau **mezocurtică** (vezi curba neagră de la mijloc), acest lucru însemnând că trebuie să existe o probabilitate redusă (0,003) de a se obține un scor extrem foarte mare și un scor extrem foarte mic, nici prea mare (ca în cazul distribuției leptocurtice) și nici prea mică (precum am întâlnit la distribuțiile platicurtice), indicatorul statistic al boltirii fiind numit și **indicator kurtosis**.

Termenul „kurtosis" provine din grecescul κυρτός și înseamnă „curbat", „arcu-

it", indicându-se astfel curbura distribuției și fiind o măsură a probabilității de apariție a scorurilor extreme la ambele capete ale distribuții, în relație cu variabilitatea. Similar simetriei, boltirea are, în cazul unei distribuții perfect normale, valoarea 3, din acest caz distribuția normală numindu-se și „distribuția celor 3 grade de boltire". În cazul unei *valori mai mari de 3, distribuția devine leptocurtică, ascuțită*, iar dacă *valoarea este mai mică de 3, distribuția devine platicurtică, turtită*. Desigur, și acestea sunt **reguli de bună practică**, valoarea indicatorului fiind estimată și rezultând o eroare de estimare pe care o vom utiliza pentru obținerea unor măsuri mai riguroase.

Pentru a se armoniza interpretarea, cele mai multe aplicații de analiză de date scad valoarea 3 din indicatorul kurtosis, interpretându-l ca pe cel al simetriei, aceasta fiind o practică universală, din acest motiv vom găsi foarte rar cercetători care să se raporteze la nivelul 3 al boltirii.

Evaluarea statistică a normalității distribuției

După cum s-a constatat, distribuția normală este una dintre cele mai importante distribuții statistice pentru variabile continui, foarte multe fenomene fizice, sociale, economice sau psihologice putând fi modelate pe baza ei, de aici și rolul său central în inferența statistică, precum și accentul pus pe corecta sa evaluare. Ea **este descrisă în mod complet de *medie* și de *abatere standard* și evaluată pe baza *simetriei* și *boltirii*, încheind prima etapă a oricărei cercetări științifice, cea a analizelor univariate descriptive**.

Reamintim faptul că folosim metode statistice în analiza datelor nu pentru a descrie un lot de cercetare sau un eșantion, ci pentru a face inferențe la nivelul populației, iar *foarte multe tehnici inferențiale de analiză a datelor pleacă de la premisa că, la nivelul populației, fenomenul studiat se distribuie normal*. Această premisă, denumită și **asumpția normalității**, are două consecințe majore asupra cercetării.

- În primul rând, *dacă fenomenul nu se distribuie normal la nivelul populației, toate concluziile pe care le tragem sunt eronate din start*. Din păcate, un astfel de lucru nu poate fi verificat, deoarece nu avem acces la populație. Noi doar presupunem că inteligența se distribuie normal, la fel anxietatea, depresia, nivelul venitului, interesul pentru reclame publicitare și așa mai departe, dar dacă nu este așa? Sunt mari semne de întrebare că anumite funcții psihice ar urma o astfel de distribuție, asta ca să nu mai discutăm despre accentuările psihologice, cum ar fi anxietatea. La fel, am putea greși în presupunerea că interesul pentru publicitate se distribuie normal, cu atât mai mult nivelul venitului, însă această greșeală ne-o asumăm, altminteri cercetarea științifică devine imposibilă.
- În al doilea rând, *dacă acceptăm faptul că fenomenul se distribuie normal la nivelul populației, atunci, pentru a fi modelat, el trebuie să se distribuie normal și la nivelul lotului de cercetare sau al eșantionului*, altfel inferența statistică nu mai este validă, acesta fiind motivul evaluării, încă de la început, a **îndeplinirii asumpției normalității univariate**. Dacă această asumpție nu este îndeplinită, atunci indiferent de tehnicile inferențiale utilizate, concluziile nu pot fi decât eronate.

Evaluarea îndeplinirii asumpției normalității univariate sau, pe scurt, evaluare normalității distribuției se poate realiza în **patru feluri:***(a) pe baza reperelor de bună practică, (b) pe baza graficelor dia-*

gnostice, (c) pe baza indicatorilor simetriei și boltirii și *(d) pe baza testelor statistice de comparare a dis*tribuției empirice cu distribuția normală teoretică.

Dacă doriți să aflați, în mod concret, modul în care puteți efectua astfel de analize în IBM SPSS Statistics și în R, vă invit să vă **faceți un cont pe acest site** și să urmați cursurile **„Introducere în analiza datelor"** apoi **„Probabilități și distribuții statistice"** pentru a dobândi competențe reale și solide.

https://new-skills.eu/ro/blogul-nostru/distributia-normala-si-evaluarea-acesteia
Accesat, 28 iunie, 2019

REFLECȚII ASUPRA BIBLIOFILIEI

Umberto ECO

Să le vorbești despre bibliofilie bibliofililor este cu totul altceva decât să le vorbești despre același subiect unor persoane obișnuite, ca să spunem așa. Adevărata dilemă a unui colecționar de cărți este aceasta: dacă ar colecționa tablouri din Renaștere sau porțelan chinezesc, le-ar putea ține în sufragerie și toți oaspeții ar rămâne extaziați în fața lor; în schimb, bibliofilul nu știe niciodată cui să arate comorile sale. Cei care nu sunt bibliofili aruncă o privire distrată și nu pricep cum de o cărțulie din secolul al XVII-lea, in 12°, cu filele îngălbenite, poate fi mândria cuiva care a reușit să-și procure ultimul exemplar de pe piață; iar ceilalți bibliofili nutresc adesea sentimente de invidie (ar vrea și ei să aibă acea carte și se înfurie) sau de dispreț (cred că au în biblioteca proprie exemplare mult mai rare sau colecționează cărți cu tematică diferită de a voastră - adică un colecționar de cărți despre arhitectură, datând din Renaștere, poate rămâne indiferent în fața celei mai prețioase culegeri de pamflete rozicruciene din secolul al XVII-lea).

Cauza principală a dezinteresului oamenilor obișnuiți este că bibliofilia este considerată o pasiune costisitoare, care poate fi practicată numai de persoane foarte bogate. Desigur, este adevărat că există cărți vechi care costă sute de milioane și că ultimul exemplar în circulație al *Divinei Comedii* în prima ediție tipărită a fost adjudecat la licitație pentru suma de un miliard și jumătate de lire italiene, dar dragostea pentru carte se poate manifesta și prin colecții de ediții princeps moderne, care adesea sunt achiziționate la prețuri foarte accesibile de la tarabe. Tot căutând pe tarabe, un student de-al meu colecționa numai ghiduri turistice din orice epocă sau țară; la fel, un tânăr cu o situație economică modestă poate găsi mici ediții din secolul al XVI-lea care încă mai