ÎN CULISELE *BIG DATA*

O abordare statistică

RAFAEL LAHOZ-BELTRA
Traducere de Anca Dinu

LITERA
București

CUPRINS

Capitolul 3

Capitolul 4

Capitolul 0

Big data și cea de-a doua revoluție a statisticii

„Sărutul provoacă palpitații atât de puternice, încât inima bate în 4 secunde mai mult decât în 3 minute. Statisticile demonstrează că 480 de săruturi ne scurtează viața cu o zi, 2360 de săruturi ne privează de o săptămână și 148 071 de săruturi ne costă, nici mai mult, nici mai puțin de un an din viață."
(Paul Morand, 1888-1976).

Odată cu trecerea timpului, oamenii și-au dat seama de necesitatea de a prezice situații viitoare, de a identifica tendințe, de a recunoaște și a clasifica obiecte, de a compara grupe în cadrul experimentelor și de a rezolva o multitudine de probleme practice de tot felul. Inundațiile provocate de râuri, infracțiunile înregistrate de-a lungul unui an, clasificarea documentelor clinice și a imaginilor medicale, sau compararea eficienței a două medicamente, sunt doar câteva exemple care ilustrează aceste situații. Deși la prima vedere acestea par a fi foarte diferite între ele, toate au o trăsătură comună, și anume că ceea ce se întâmplă se poate exprima cantitativ sau calitativ sub formă de date. Cu alte cuvinte, evenimentele din viața de zi cu zi generează date, iar statistica, disciplină derivată din matematică, este știința chemată să dea seama de toate aceste evenimente de naturi atât de diverse. O caracteristică fundamentală proprie statisticii, căreia i se datorează în bună măsură succesul, a fost stabilirea unor tehnici și a unei metodologii comune pentru rezolvarea unor probleme atât de diferite. Această metodologie constă în organizarea, analizarea și interpretarea datelor generate de o situație concretă din lumea reală, indiferent că este vorba de medicină, biologie, psihologie, economie, politică, sociologie, inginerie, sau de orice alt domeniu.

Un alt factor care explică succesul statisticii este capacitatea de a stabili generalizări pornind de la o mulțime de date, așa-numita „inferență statistică", așa cum se întâmplă, de exemplu, cu sondajele electorale sau cu analizele de piață într-o societate. Astfel, statistica este un instrument extraordinar cu care se pot face prognoze, determina tendințe, clasifica și compara obiecte sau extrage generalizări, ceea ce se dovedește de un real ajutor în procesele decizionale, fie ele de natură politică, clinică, antreprenorială, sau științifică.

Se poate afirma, așadar, că succesul actual al statisticii se datorează unor cauze independente și separate în timp de aproape un secol. Pe de o parte, se datorează revoluției care a avut loc la sfârșitul secolului al XIX-lea, inițiată de două personalități celebre ale statisticii, **Francis Galton** (1822-1911) și **Karl Pearson** (1857-1936), părinții statisticii moderne. Pe de altă parte, succesul statisticii se datorează conceptului mai recent de „big data" (în engleză), termen apărut la sfârșitul secolului XX și numit astfel de către informaticianul de origine americană **John Mashey** (n. 1946).

De la teoria probabilităților la statistică

De-a lungul secolului al XVII-lea, personalități ca Newton au produs o adevărată revoluție a modului de lucru al oamenilor de știință. Din fericire, moștenirea lor nu s-a pierdut, ba chiar a fost continuată de alte minți strălucite. În secolul al XVIII-lea, epoca Iluminismului și a Raționalismului, Revoluția Industrială (1760-1820) și intensificarea relațiilor comerciale cu numeroasele colonii de peste ocean au generat noi nevoi în societate. Tocmai aceste noi necesități au acționat ca stimuli pentru dezvoltarea aparatului matematic, care este astăzi cunoscut sub numele de „teoria probabilităților", unul dintre stâlpii fundamentali pe care s-a fondat și s-a construit statistica. Cu toate că analiza matematică a jocurilor a avut inițial un caracter mai degrabă anecdotic, a constituit un alt factor care a favorizat studiul probabilităților. Una dintre descoperirile fundamentale a aparținut lui

Abraham de Moivre (1667-1754), care a stabilit că distribuția binomială, una dintre cele mai importante legi ale statisticii, se apropie, în anumite condiții experimentale, de ceea ce numim astăzi distribuție normală Gauss, distribuția cea mai importantă din statistică

$$P(x,n,p)=C_x^n p^x q^{n-x} \cong p\left(z=\frac{x-np}{\sqrt{npq}}\right)$$

Deși anumite idei despre probabilitate sunt prezentate în lucrarea *The doctrine of chances* (Teoria riscurilor) a lui De Moivre, publicată în 1718, contribuțiile fundamentale la teoria probabilităților erau deja realizate de ceva timp, de către celebra „familie Bernoulli". **Nicolau Bernoulli** (1695-1726) introdusese așa-numita distribuție Bernoulli, iar **Jacob Bernoulli** (1655-1705) noțiunea de probabilitate. După definiția acestuia, probabilitatea înseamnă gradul de certitudine al unui eveniment și se poate măsura printr-un număr între 0 și 1, aceasta devenind în zilele noastre una dintre axiomele (propoziție al cărei adevăr este acceptat și care, așadar, nu necesită nicio demonstrație) fundamentale ale teoriei probabilităților.

Pe parcursul secolului al XVIII-lea, apare necesitatea de estimare a erorilor de măsurare și se dezvoltă așa-numita „teorie a erorilor".

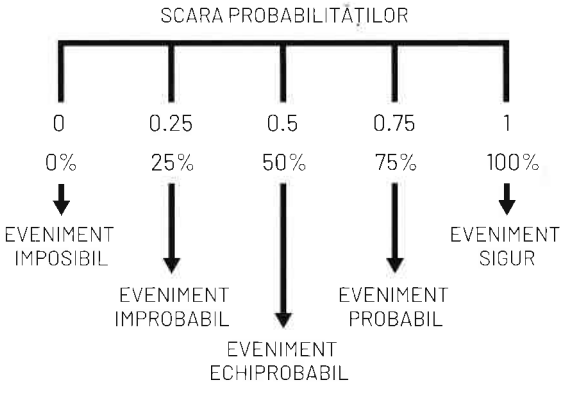

Figura 1.
Scara Probabilităților.

De pildă, dacă se măsoară masa unui pahar cu apă, erorile apărute din cauza lipsei de precizie a instrumentului de măsurare sau a celui care măsoară duc la apariția unei diferențe, ε_i, între masa obținută prin măsurare x_i și cea reală x, adică $\varepsilon_i = |x_i - x|$. Pentru rezolvarea acestui tip de problemă se caută o funcție matematică prin care să se exprime frecvența cu care apare o anumită eroare. Astfel, se va cunoaște probabilitatea de a obține o anumită eroare de măsurare. Una dintre primele funcții propuse îi aparține lui **Johann Heinrich Lamber** (1728-1777), un erudit din Europa secolului optsprezece. În 1765, Lamber propune o distribuție semicirculară a erorilor:

$$f(x) = \frac{1}{2}\sqrt{(1-x^2)}$$

similară cu cea propusă, câțiva ani mai târziu, de italianul naturalizat francez **Joseph-Louis Lagrange** (1736-1813). În locul funcției propuse de Lamber, Lagrange introduce în 1776 o funcție parabolică, cea mai potrivită pentru exprimarea distribuției erorilor:

$$f(x) = \frac{3}{4}(1-x^2).$$

Figura 2.
Pierre-Simon Laplace.

Dar pasul cel mai important către maturizarea statisticii a fost făcut cu puțin timp înaintea tentativei lui Lagrange, de data aceasta aparținându-i lui **Pierre-Simon Laplace** (1749-1827). Francezul este considerat continuatorul operei lui Newton, care murise la începutul secolului, în 1727. Autor al lucrării *Teoria analitică a probabilităților* (1812), Laplace nu numai că a introdus o definiție care permite calcularea probabilității unui eveniment, dar a și deschis drumul care a dus

la apariția distribuției normale. Datorită lui, cunoaștem astăzi cum să calculăm probabilitatea *a priori* a evenimentelor echiprobabile, adică a evenimentelor cu probabilitatea egală. De exemplu, într-un experiment E, care constă în aruncarea unui zar, rezultatele, sau *cazurile posibile*, sunt $\Omega = \{1, 2, 3, 4, 5, 6\}$ și toate au aceeași probabilitate. Probabilitatea de a obține un număr par, adică aparținând mulțimii {2, 4, 6}, care sunt *cazurile favorabile*, se obține astfel:

$$p(\text{par}) = \frac{\text{cazuri favorabile}}{\text{cazuri posibile}} = \frac{3}{6} = \frac{1}{2}$$

În 1778, Laplace face o descoperire fundamentală, demonstrând că frecvența cu care se obține o eroare de măsurare dată este direct proporțională cu pătratul valorii ei, propunând următoarea funcție:

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

în care b este un parametru de ponderare ($b > 0$). În orice caz, un geniu al matematicii, **Carl Friedrich Gauss** (1777-1855), a fost acela care a fost *creditat* pentru (re)descoperirea acestei funcții, care îi poartă numele, la care ajunseseră și alți matematicieni. Funcția $f(x)$ reprezintă celebra curbă sub formă de clopot, care exprimă distribuția erorilor:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

Dacă secolul al XVIII-lea a fost cel în care s-au făcut primii pași și în care s-au pus bazele statisticii, în cursul secolului al XIX-lea putem spune că această disciplină ajunge la maturitate. Legendre, în 1805, și Gauss, în 1809, publică, independent unul de celălalt, *metoda celor mai mici pătrate*, una dintre tehnicile fundamentale ale