Introducere

În viaţa de zi cu zi ne întâlnim adesea cu noţiuni din statistică, când se încearcă descrierea numerică a fenomenelor socio-economice („numărul” persoanelor cu dizabilităţi; „rata” şomajului; „procentul” de promovabilitate la un examen, „marja de eroare” a unui sondajelor de opinie; salariul „mediu” al unui angajat şi altele asemenea).

Dar ce este statistica? În mod sigur nu este doar un şir de cifre şi calcule matematice.

Există două răspunsuri care se pot formula:

- statistica este o disciplină de studiu, o „ştiinţă” – pe lângă statistică ca disciplină teoretică (ramură a matematicii) se numără şi multiplele sale aplicaţii empirice: statistică judiciară, statistică economică, statistică medicală, statistică psihologică şi altele;
- statistica este o metodă de cercetare, metodă folosită de multe ştiinţe şi având o contribuţie importantă la dezvoltarea acestora.

În ceea ce priveşte disciplinele ce studiază fenomenele psihosociale, statistica a ajutat pe linia câştigării de către acestea a statutului de „ştiinţe”.

Statistica este o teorie a informaţiei folosită pentru a analiza colecţii mari de date şi pentru a face deducţii valide asupra fenomenelor, pe baza informaţiilor de la nivel de eşantion (Ott, Larson & Mendenhall, 1987).

Ca metodă sau instrument de cercetare, statistica are în vedere în principal centralizarea, gruparea, sintetizarea şi analiza datelor provenind din cercetări empirice, precum şi prezentarea rezultatelor. Ea face ca datele brute să devină informaţii, să comunice ceva – ceva relevant despre fenomenele de masă.

În concepţia lui D. P. Marţian statistica trebuie să studieze fenomenele naturale şi sociale, având drept scop descoperirea legilor care le guvernează. Se evidenţiază apartenenţa sa la curentul de largă circulaţie conform căruia obiectul de cercetare al statisticii trebuie să depăşească graniţele sferei fenomenelor sociale.

Modelul liniar

În statistici, termenul model liniar este folosit în moduri diferite în funcție de context. Cea mai frecventă apariție este legată de modelele de regresie și termenul este adesea luat ca sinonim cu regresia liniară. Cu toate acestea, termenul este folosit și în analiza seriilor temporale. În fiecare caz, denumirea „liniar" este utilizată pentru a identifica o subclasă de modele pentru care reducerea substanțială a complexității teoriei statistice este posibilă.

Modelul liniar specifică o relaţie liniară între o variabliă dependentă şi ***n*** variabile independente, utilizând formula:

$$\mathbf{y = a_0 + a_1 x_1 + a_2 x_2 + Ã‚Â¼+ a_n x_n,}$$

unde *y* este variabila dependentă, $\{x_i\}$ sunt variabilele independente, iar $\{a_i\}$ reprezintă parametrii modelului.

Spre exemplu, pentru un eşantion de 25 de oraşe, următorul model a fost estimat pentru o relaţie între circulaţia unui ziar (variabila dependentă, numită astfel deoarece depinde de

celelalte variabile) şi vânzările ziarului şi densitatea populaţiei oraşului (variabile independente):

$y = 0.381 + 0,067\ x_1 + 0,025\ x_2$

unde: y = numărul de exemplare ale ziarului (x 1000), x_1 = totalul exemplarelor vândute (x 1,000,000) şi x_2 = populaţia pe km^2.

Formula de mai sus se transpune astfel: circulaţia ziarului (explimată în mii de exemplare) este egală cu totalul vânzărilor (exprimat în milioane) înmulţit cu 0,067 plus densitatea populaţiei înmulţită cu 0,025 plus 0,381.

Populaţie

Statistica analizează un număr de entităţi de aceeaşi natură care au un set de caracteristici comune. Mulţimea acestor entităţi (elemente) se numeşte populaţie statistică.

Exemple de populaţie statistică:

– un grup de persoane dintr-un spaţiu geografic, demografic sau social (studenţii unei universităţi) care au o anumita vârstă (mai tineri de 24 de ani) şi sunt de un anumit gen (masculin);

– o mulţime de obiecte (totalitatea vaccinurilor antigripale produse într-o perioadă de timp la un institut);

– o mulţime de măsurători (tensiunea arterială măsurată la o persoană sau grup de persoane de o anumită vărstă într-un interval de timp).

Numărul de entităţi (elemente) ale unei populaţii reprezintă *volumul*, *efectivul*, *dimensiunea* sau *talia* populaţiei. Entităţile unei populaţii statistice se numesc unităţi statistice sau indivizi. Stabilirea indivizilor unei populaţii se bazează pe două tipuri de criterii:

1. criterii de includere ce stabilesc condiţiile ce trebuiesc îndeplinite pentru ca un individ să aparţină unei populaţii;
2. criterii de excludere ce stabilesc condiţiile în care un individ nu aparţine unei populaţii.

În practică de multe ori cercetătorul utilizează o populaţie de subiecţi disponibili în locul populaţiei iniţiale (populaţie ţintă) numită şi populaţie de selecţie. Din populaţia de selecţie, dacă este reprezentativă pentru populaţia ţintă, se extrag ***eşantioane*** ce vor fi folosite în studiile statistice.

În statistică, în cadrul studiilor, nu se utilizează întreaga populaţie din unul mai multe din inconveniente:

– volumul populaţie poate fi uneori foarte mare;

– timpul efectiv de studiu creşte proporţional cu numărul elementelor studiate;

– costurile şi resursele alocate cresc proporţional cu numărul entităţilor studiate;

– de multe ori se distrug entităţile studiate;

– există situaţii în care nu se pot culege informaţii despre toţi indivizii populaţiei;

– precizia rezultatele este invers proporţională cu numărul entităţilor studiate.

Eşantion

Plecându-se de la considerentele de mai sus, apare necesitatea de a cuantifica (reduce) populaţia statistică la o submulţime de elemente cu un efectiv redus, submulţime ce numeşte ***eşantion*** sau ***selecţie***.

Eşantioanele sunt imagini reduse cât mai fidele ale întregii populaţii statistice care permit studierea corectă a unui set de caracteristici. În acest caz, spunem despre un eşantion că este reprezentativ pentru populaţie, în caz contrar este nereprezentativ.

Un eşantion reprezentativ pentru o populaţie statistică îndeplineşte două condiţii principale:

1. o condiţie *cantitativă* în sensul că talia sa trebuie să fie suficient de mare pentru studii statistice;
2. o condiţie *calitativă* în sensul că extragerea indivizilor din populaţie trebuie făcută aleator. Eşantioane în care pentru fiecare individ al populaţiei se cunoaşte probabilitatea (şansa) de a fi inclus în eşantion se numesc *eşantioane probabiliste*. În situaţia în care nu se cunoaşte probabilitatea de selectare a indivizilor avem de a face cu *eşantioane nonprobabilistice*.

O serie statistică este pur și simplu o listă de valori din același set , în care ordinea termenilor nu este semnificativă (spre deosebire de o serie de timp).

O astfel de listă este obținută în general de la o populație (în sensul statistic al termenului, adică elemente ale unei categorii pe care se dorește să o studieze (oameni, alte ființe vii, organizații, produse economice, etc).

Pentru fiecare individ selectat, se măsoară mai multe variabile sau caractere. Fiecare

variabilă poate fi cantitativă (numerică și reprezentând o cantitate cum ar fi cantitatea, dimensiunea, costul) sau calitativă (culoare, sex, orientare politică etc.).

O serie statistică poate fi obținută, de asemenea, utilizând simulare computerizată sau fenomene fizice. Studiul unei serii statistice depinde de numărul de variabile identificate și de natura acestora.

Pentru o serie statistică cu o singură variabilă cantitativă, definim indicatori clasici care sunt media, mediana, modul, quartilele, decilele și alte cuantile, precum și indicatorii de dispersie, cum ar fi intervalul intercuartil, varianța, abaterea standard.

Un indicator de dispersie măsoară variabilitatea valorilor unei serii statistice. Este întotdeauna pozitiv și cu atât mai mare cu cât valorile seriei sunt răspândite. Cele mai frecvente sunt varianța , abaterea standard și intervalul intercuartil .

Acești indicatori completează informațiile furnizate de indicatorii de poziție sau de tendință centrală , măsurați prin medie sau mediană.

Varianţa

Varianța este o măsură a dispersiei valorilor într-un eșantion sau distribuție de probabilitate. Exprimă media pătratelor abaterilor de la medie, egală și cu diferența dintre media pătratelor valorilor variabilei și pătratul mediei, conform teoremei König-Huygens.

Astfel, cu cât deviația față de medie este mai mare, cu atât este mai preponderentă în

calculul total (a se vedea funcția pătrată) a variației, ceea ce ar da, prin urmare, o idee bună despre dispersia valorilor.

Varianța este întotdeauna pozitivă și dispare numai dacă valorile sunt egale. Rădăcina sa pătrată definește *abaterea standard* σ, de unde notația.

Varianța este pătratică și invariantă.

Formula varianţei este:

$$\sigma^2 = V = \mathbb{V}(X) = \mathbf{Var(X)}$$

Când seria statistică ia valorile x 1 , x 2 , ..., x n cu frecvențele f 1 , f 2 , ..., f n , varianța sa este:

Varianța este un indicator al dispersiei valorilor , adică este întotdeauna pozitivă, dispare doar pentru o serie statistică în care toți termenii au aceeași valoare, este cu atât mai mare cu cât valorile sunt răspândite, și invariant prin adăugarea unei constante.

Calculul său poate părea mai complicat decât cel al altor indicatori de dispersie, cum ar fi intervalul interquartilei sau deviația absolută medie, dar spre deosebire de acesta din urmă, este cumulativ: dacă colectăm *k* serii statistice într-una singură, varianța globală poate fi calculată din numărul *ni* , varianța *Vi* și media fiecărei serii inițiale prin formula:

Cu alte cuvinte, varianța generală este suma variației mediilor și a mediei variațelor, chiar dacă această a doua componentă este adesea neglijată.

Abaterea standard (standard deviation)

Abaterea standard (deviaţia standard – din engleză "standard deviation") este o măsură a dispersiei valorilor într-un eșantion statistic sau distribuție de probabilitate . Este definit ca rădăcina pătrată a variației sau, în mod echivalent, ca rădăcina medie pătrată a abaterilor de la medie . Este scris în general cu litera greacă σ („sigma"). Este omogenă cu variabila măsurată.

Abaterile standard sunt întâlnite în toate domeniile în care sunt aplicate probabilitățile și statisticile, în special în domeniul anchetelor, fizicii, biologiei sau finanțelor. În general, acestea fac posibilă sintetizarea rezultatelor numerice ale unui experiment repetat. Atât în probabilități, cât și în statistici, este folosit pentru a exprima alte concepte importante, cum ar fi coeficientul de corelație, coeficientul de variație sau distribuția optimă a lui Neyman.

Când abaterea standard a unei populații este necunoscută, valoarea acesteia este aproximată folosind estimatori.

Abaterea medie pătratică

Abaterea medie pătratică a unei variabile aleatoare este rădăcina pătratică din dispersia variabilei aleatoare.

Grade de libertate

Numărul gradelor de libertate, în statistică, reprezintă diferența dintre numărul de experiențe adoptat la proiectarea experimentelor și numărul de coeficienți (sau constante) care au fost calculați ca rezultat al acestor experimente, în mod independent unii de alții. Alţi autori (C. Moineagu și alţii) consideră că numărul gradelor de libertate reprezintă numărul variabilelor independente dintr-un experiment a căror variație nu suferă nici o restricție.

Numărul de grade de libertate se referă la numărul de valori care pot varia fără nici o restricție. Conceptul este folosit în contextul analizelor de regresie.

Există şi o altă definiţie, astfel: numărul gradelor de libertate este egal cu numărul de observații experimentale minus numărul de parametri care trebuie estimaţi.

Estimările de parametri statistici se pot baza pe diferite cantități de informații sau date.

Numărul de informații independente care intră în estimarea unui parametru se numesc gradele de libertate. În general, gradele de libertate ale unei estimări a unui parametru sunt egale cu numărul de independenți scoruri care intră în estimare minus numărul de parametri folosiți ca pași intermediari în estimarea parametrului însuși (de cele mai multe ori varianța eșantionului are N – 1 grad de libertate, deoarece este calculat din N scoruri aleatorii minus singurul 1 parametru estimat ca pas intermediar, care este media eșantionului).

Termenul este cel mai des folosit în contextul modelelor liniare (regresie liniara, analiza variatiei), unde anumiți vectori aleatori sunt constrânși să se afle în subspatii liniare, iar numărul de grade de libertate este dimensiunea subspatiului.

Distorsiunea

Distorsiunea este reprezentarea incorectă (intenţionată sau nu) a unui set de date (informaţii statistice), prin diverse procedee (cel mai cunoscut fiind aşa zisul, din engleză “cherry picking” = “alegerea cireşelor”, care presupune selectarea de informaţii statistice nereprezentative dintr-un set de date, astfel încât din setul de date selectat se trage o concluzie eronată asupra setului de date în întregimea sa.

Distribuţia de eşantionare a mediilor aritmetice

Eșantionarea statistică este utilizată destul de des în statistici. În acest proces, ne propunem să determinăm ceva despre o populație. Deoarece populațiile sunt de obicei de dimensiuni mari, formăm un eșantion statistic prin selectarea unui subset al populației care are o dimensiune prestabilită. Studiind eșantionul putem folosi statistici inferențiale pentru a determina ceva despre populație.

Un eșantion statistic de mărime *n* implică un singur grup de *n* indivizi sau subiecți care au fost aleși aleatoriu din populație. Strâns legată de conceptul de eșantion statistic este o distribuție de eșantionare.

O distribuție de eșantionare apare atunci când formăm mai multe eșantioane simple aleatorii de aceeași dimensiune dintr-o populație dată. Aceste eșantioane sunt considerate independente una de cealaltă. Deci, dacă un individ se află într-un eșantion, atunci are aceeași probabilitate de a fi în eșantionul următor, care este luat.

Calculăm o anumită statistică pentru fiecare eșantion. Aceasta ar putea fi o medie a eșantionului, o varianță a eșantionului sau o proporție de eșantion. Deoarece o statistică depinde de eșantionul pe care îl avem, fiecare eșantion va produce de obicei o valoare diferită pentru statistica de interes. Gama valorilor care au fost produse este ceea ce ne oferă distribuția noastră de eșantionare.

Pentru a exemplifica, vom lua în considerare distribuția eșantionării pentru medie. Media unei populații este un parametru care este de obicei necunoscut. Dacă selectăm un eșantion de mărimea 100, atunci media acestui eșantion este ușor calculată prin adăugarea tuturor valorilor împreună și apoi împărțirea la numărul total de puncte de date, în acest caz, 100. Un eșantion de dimensiunea 100 ne poate da o medie 50. Un alt astfel de eșantion poate avea o medie de 49. Un alt eșantion 51 și un alt eșantion ar putea avea o medie de 50,5.

Distribuția acestor mijloace de eșantionare ne oferă o distribuție de eșantionare. Am dori să luăm în considerare mai mult de doar patru eșantioane de mijloace, așa cum am făcut mai sus. Cu câteva alte mijloace de eșantionare, am avea o idee bună despre forma distribuției de eșantionare.

Să presupunem că începem cu o populație cu o medie de μ și abaterea standard de σ. Abaterea standard ne oferă o măsurare a distribuției. Vom compara acest lucru cu o distribuție de eșantionare obținută prin formarea de probe simple aleatorii de mărimea *n*. Distribuția de eșantionare a mediei va avea în continuare o medie de μ, dar abaterea standard este diferită. Abaterea standard pentru o distribuție de eșantionare devine $\sigma / \sqrt{n}$.

Eroarea standard a mediei

Eroarea standard este o estimare a variaţiei valorii statisticii unui test de la un eşantion la altul. Este o măsură a incertitudinii statisticii testului. Pentru eroarea standard se poate folosi abrevierea SE (din engleză – „standard error”).

Eroarea standard este calculată folosind devierea standard a distribuţiei de eşantionare pentru statistica testului. Distribuţia de eşantionare este distribuţia tuturor eşantioanelor posibile.

Dacă se efectuează un sondaj şi se aleg la întâmplare 1.000 de persoane ca participanţi, acest grup este un eşantion. Se pot alege un număr diferit de eşantioane. Apoi se poate calcula media pentru fiecare eşantion. Distribuţia acestor medii de eşantion este distribuţia de eşantionare. Prin calcularea devierii standard a acestei distribuţii, se obţine *eroarea standard a mediei*. Când eroarea standard este scrisă fără aplicarea unui calificativ, se presupune că este eroarea standard a mediei.

Relaţia dintre abaterea standard şi eroarea standard este astfel: pentru o anumită mărime a unui eşantion, eroarea standard este egală cu abaterea standard împărţită la rădăcina pătrată a mărimii eşantionului. Eroarea stndard este de asemenea invers proporţională cu mărimea eşantionului. Cu cât este mai mare eşantionul, cu atât este mai mică eroarea standard (deoarece statistica se va apropia de valoarea reală).

Teorema limitei centrale

Teorema limitei centrale explică relația dintre o populație și distribuția mijloacelor de eșantionare găsite prin prelevarea tuturor eșantioanelor posibile de o anumită dimensiune din populația inițială, găsirea mediei fiecărui eșantion și aranjarea lor într-o distribuție.

Distribuirea eșantionării mediilor este un concept ușor. Să presupunem că avem o populație de *x*-uri. Se ia un eșantion de *n* din acele *x*-uri și se găseşte media eșantionului respectiv, obținând un *x*. Apoi se ia un alt eșantion de aceeași dimensiune, *n*, și se găseşte *x*-ul său. Se face acest lucru din nou și din nou până când se vor fi ales toate eșantioanele posibile de dimensiunea *n*. Se generează astfel o nouă populație, o populație de *x*-uri. Se aranjează această populație într-o distribuție și se obţine distribuția eșantionării mediilor. Se pot găsi distribuția eșantionării mediilor, sau variante, sau alte statistici ale eșantionului, colectând toate eșantioanele posibile de o anumită dimensiune, *n*, găsind mediile, varianța sau alte statistici despre fiecare eșantion și aranjându-le într-o distribuție.

Teorema limitei centrale se referă la distribuția prin eșantionare a mediilor. Conectează distribuția eșantionării *x*-urilor cu distribuția originală a *x*-urilor. Teorema ne spune două lucruri:

1. Media mediilor eșantionului este egală cu media populației inițiale, $\mu_x = \mu$. Aceasta este ceea ce face din *x* un estimator imparțial al lui μ.
2. Distribuția *x*-urilor va avea formă de clopot, indiferent de forma distribuției originale a *x*-urilor.

Aceasta înseamnă că doar o mică parte din eșantioane au medii care sunt departe de media populației. Pentru ca un eșantion să aibă o medie care să fie departe de μ_x, aproape toți membrii acestuia trebuie să se afle din coada dreaptă a distribuției lui *x* sau aproape toți trebuie să fie din coada stângă. Există mult mai multe eșantioane cu majoritatea membrilor lor din mijlocul distribuției sau cu unii membri din coada dreaptă și unii din coada stângă, iar toate aceste eșantioane vor avea un *x* apropiat de μ_x.

Intervalul de confidenţă (intervalul de încredere)

Un interval de încredere, în statistică, se referă la probabilitatea ca un parametru al populației să se încadreze într-un set de valori pentru o anumită proporție de ori.

Un interval de încredere afișează probabilitatea ca un parametru să se încadreze între o pereche de valori în jurul mediei. Intervalele de încredere măsoară gradul de incertitudine sau certitudine într-o metodă de eșantionare. Acestea sunt cel mai adesea construite folosind niveluri de încredere de 95% sau 99%, uneori putându-se folosi şi 90%.

Intervalele de încredere măsoară gradul de incertitudine sau certitudine într-o metodă de eșantionare. Ele pot lua orice număr de limite de probabilitate, cele mai frecvente fiind un nivel de încredere de 95% sau 99%. Intervalele de încredere sunt realizate folosind metode statistice, cum ar fi testul t.

Un *test t* este un tip de statistică inferențială utilizat pentru a determina dacă există o diferență semnificativă între mijloacele a două grupuri, care pot fi legate de anumite caracteristici. Se folosește mai ales atunci când seturile de date, (spre exemplu setul de date înregistrat ca rezultat al aruncării unei monede de 100 de ori), ar urma o distribuție normală și ar putea avea variante necunoscute. Un *test t* este utilizat ca instrument de testare a ipotezelor, care permite testarea unei ipoteze aplicabile unei populații.

Un test t analizează statistica *t*, valorile *distribuției t* și gradele de libertate pentru a determina semnificația statistică. Pentru a efectua un test cu trei sau mai multe mijloace, trebuie să utilizați o analiză a variației.

Statisticienii folosesc intervale de încredere pentru a măsura incertitudinea într-o variabilă eșantion. De exemplu, un cercetător selectează eşantioane diferite, în mod aleatoriu, din aceeași populație și calculează un interval de încredere pentru fiecare probă pentru a vedea cum poate reprezenta valoarea reală a variabilei populației. Seturile de date rezultate sunt

toate diferite; unele intervale includ parametrul adevărat al populației, iar altele nu.

Un interval de încredere este un interval de valori, delimitate deasupra și sub media statisticii, care probabil ar conține un parametru necunoscut al populației. Nivelul de încredere se referă la procentul de probabilitate sau certitudine că intervalul de încredere ar conține parametrul adevărat al populației atunci când trageți de multe ori un eșantion aleatoriu. Sau, în limba populară, „suntem siguri 99% (*nivel de încredere)* că majoritatea acestor eșantioane *(intervale de încredere)* conțin parametrul populației adevărat".

Să presupunem că un grup de cercetători studiază înălțimile jucătorilor de fotbal dintr-un liceu. Cercetătorii iau un eșantion aleatoriu din populație și stabilesc o înălțime medie de 74 inci.

Media de 184 de cm. este o estimare punctuală a populației. O estimare punctuală în sine are o utilitate limitată, deoarece nu relevă incertitudinea asociată cu estimarea; nu aveți o bună înțelegere a cât de departe ar putea fi acest eșantion mediu de 184 de cm de media populației. Ceea ce lipsește este gradul de incertitudine din acest eșantion unic.

Intervalele de încredere oferă mai multe informații decât estimările punctuale. Prin stabilirea unui interval de încredere de 95% folosind media și *deviația standard* a eșantionului și presupunând o *distribuție normală* așa cum este reprezentată de curba clopotului, cercetătorii ajung la o limită superioară și inferioară care conține media reală 95% din timp.

Să presupunem că intervalul este între 180 de cm și 188 de cm. Dacă cercetătorii iau 100 de eșantioane aleatorii de la populația de jucători de fotbal din liceu în ansamblu, media ar trebui să scadă între 180 și 188 de centimetri în 95 dintre aceste eșantioane.

Dacă cercetătorii doresc încredere și mai mare, ei pot extinde intervalul până la încredere de 99%. Dacă faceți acest lucru, invariabil se creează o gamă mai largă, deoarece face loc pentru un număr mai mare de mijloace de eșantionare. Dacă stabilesc intervalul de încredere de 99% ca fiind între 178 și 190 de cm, se pot aștepta ca 99 din 100 de probe evaluate să conțină o valoare medie între aceste numere.

Un nivel de încredere de 90%, pe de altă parte, implică faptul că ne-am aștepta ca 90% din

estimările intervalului să includă parametrul populației.

Estimarea corectă a unui parametru statistic se face cu ajutorul intervalului de încredere. Intervalul de încredere depinde de volumul eşantionului şi de eroarea standard. Cu cât eroarea standard este mai mare cu atât intervalul de încredere este mai larg. Cu cât volumul eşantionului este mai mic cu atât intervalul de încredere este mai larg.

Material realizat de Isabela Anca Tucanu