# Aspecte privind inferența statistică

**Constantin Anghelache**
*Profesor universitar doctor*
**Georgeta Vintilă**
*Profesor universitar doctor*
**Mădălina Dumbravă**
*Doctorand*
Academia de Studii Economice București

***Abstract.*** *There are two major types of data sources that can be used when a phenomenon or a variable is investigated: the population and the sample. Based upon the sample, various facts concerning the entire population can be deduced; this process is called statistical inference. A major problem of the inference is the variability.*

**Key words:** inference; population; sample; variability.

Ori de câte ori dorim să observăm sau să investigăm un fenomen sau o variabilă, există două tipuri fundamentale de surse de date pe care ar trebui să le utilizăm. În primul rând, ar trebui să avem acces la *populație (colectivitate definită în sens statistic).* Înțelegem prin aceasta să avem acces la toate observațiile posibile, trecute, prezente și viitoare, cu privire la variabila de interes. Eșantionul reprezintă cel de-al doilea tip de surse de date cu care ne-am putea întâlni. Pe baza eșantionului de care dispunem, trebuie să deducem fapte în legătură cu populația din care s-a prelevat eșantionul. Acest proces este cunoscut sub denumirea de *inferență statistică.*

În ansamblul tuturor problemelor legate de inferența statistică, una majoră este cea privind *variabilitatea eșantionării*. Înțelegem prin aceasta că diferitele eșantioane vor conduce la rezultate diferite.

Când prelevarea eșantionului se face de o anumită manieră, variabilitatea de selecție urmează un model sistematic. Aceste eșantioane trebuie să fie *aleatorii.*

Despre un eșantion de mărimea *n* se spune că este *aleatoriu*, *atunci când orice combinație de n unități ale unei populații are șanse egale de a intra în eșantionul care este prelevat.*

- Distribuția mediilor de selecție are media $\mu$, și dispersia $\sigma^2$ . Aceasta înseamnă $E(X) = \mu$ și $Var(X) = \sigma^2$. Dispersia este pur și simplu o măsură a gradului în care câștigurile muncitorilor individuali sunt dispersate sau „împrăștiate" în raport cu media lor, $\mu$ .

$\mu$ și $\sigma^2$ sunt cunoscute sub denumirea de *parametri ai populației*. Aceștia sunt mărimi *fixe* dar, de regulă, *necunoscute.*

Am notat media populației și variația cu aceleași simboluri ca și în cazul utilizării mediei și variației unei distribuții probabilistice. Aceasta se datorează faptului că, în condițiile unei populații atât de mari, putem interpreta frecvența relativă cu care survine un anumit nivel de câștiguri ca fiind o probabilitate. Populația poate fi considerată a fi analoagă cu o distribuție probabilistică pentru variabila *X.*

Să presupunem că din populația respectivă se extrage un eșantion aleatoriu de *n* muncitori. Aceasta se exprimă sub forma:

$$\overline{X} = \frac{\sum X_i}{n},$$

în care $X_i$ reprezintă câștigurile muncitorului *i* din cadrul eșantionului, iar suma acoperă toate valorile *i.*

Un singur eșantion extras din populație poate reprezenta o medie de eșantion $\overline{X_1}$. Dar diferitele eșantioane conduc la rezultate diferite, astfel că un al doilea eșantion ar putea să indice $\overline{X_2}$, un al treilea $\overline{X_3}$, un al patrulea $\overline{X_4}$ etc. Imaginați-vă o situație în care foarte multe, poate mii de eșantioane, toate de aceeași mărime, *n*, au fost extrase din această singură populație. În astfel de condiții ar putea deveni posibil să se construiască o distribuție de frecvențe relative pentru $\overline{X_i}$, media unui eșantion aleatoriu de mărime *n.* De exemplu, o medie de $\overline{X_i}$ poate apărea cu o frecvență relativă de $n_i$. Deoarece s-au extras mai multe eșantioane, astfel de frecvențe relative pot fi interpretate ca probabilități. Putem afirma, de exemplu, că poate să apară de $n_i$ ori, cu probabilitatea $pr_i$. În acest mod, este posibil să se construiască o distribuție de probabilitate pentru $\overline{X_i}$.

Distribuția probabilistică pentru $\overline{X_i}$ este cunoscută sub denumirea de *distribuție a mediei de selecție* pentru un eșantion aleatoriu de mărime *n.* Distribuțiile de selecție de acest fel au o importanță crucială în inferența statistică.

Desigur, în practică, distribuțiile de selecție sunt rareori construite de o manieră empirică. În mod normal, se extrage un singur eșantion.

▪ Dacă un parametru al populației este necunoscut există două modalități prin care acesta poate fi *estimat.* În primul rând, putem estima respectivul parametru printr-o singură valoare (*estimare punctuală)* sau, în al doilea rând, putem specifica un interval în cadrul căruia suntem siguri că se găsește parametrul real.

*Estimările punctuale* sunt exprimate printr-o singură valoare. De exemplu, am putea estima media câștigurilor lunare ale salariaților din comerțul cu amănuntul la o valoare $\overline{X}$. De fapt, modalitatea evidentă de a estima o medie necunoscută a populației, μ, constă în a cunoaște media eșantionului $\overline{X}$. Există un avantaj din utilizarea *estimatorului* $\overline{X}$. Știm că E($\overline{X}$) = μ. Aceasta înseamnă că, dacă am fi extras „mai multe" eșantioane din populație, am fi putut obține o distribuție de selecție și „în medie", am fi obținut a valoare egală cu valoarea reală, dar necunoscută a lui μ. Deși în practică extragem numai un eșantion, este important să apreciem că nu există nicio eroare sistematică sau *interferență* în procedura de estimare.

Deoarece E($\overline{X}$) = μ, se spune că $\overline{X}$ este un *estimator punctual nedeplasat* pentru μ.

De asemenea, vor exista situații când dorim să estimăm o dispersie a populației, $\acute{o}^2$. Estimarea punctuală evidentă pentru $\acute{o}^2$ este aceeași cu dispersia dată de formula:

$$v^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n}$$

De exemplu, având datele cu privire la câștigurile lunare ale unui eșantion aleatoriu format din *n* unități, folosim pur și simplu expresia aferentă dispersiei unui set de *n* numere. Problema privind $n^2$ constă în aceea că, similar lui $\overline{X}$, valorile pentru eșantioane diferite for fi și ele diferite și se poate demonstra că:

$$E\left(v^2\right) = \frac{n-1}{n}\delta^2 \neq \delta^2$$

În acest caz, dacă s-ar fi extras „mai multe" eșantioane atunci $v^2$ ne-ar fi dat o valoare mai degrabă mai mică decât valoarea reală a lui $\sigma^2$. De aceea, în acest caz, există o tendință sistematică spre eroare, iar despre $v^2$ se spune că este *un estimator punctual deplasat* pentru $\sigma^2$.

Pentru a depăși această problemă a interferenței, $\sigma^2$ este, în mod normal, estimată prin relația:

$$s^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1}$$

Aceasta datorită faptului că $s^2 = [n/(n-1)]v^2$, astfel că:

$$E\left(s^2\right) = E\left(\frac{n}{n-1}v^2\right) = \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \sigma^2$$

Astfel, $s^2$ devine o estimare punctuală nedeplasată pentru $\sigma^2$.

▪ Uneori se va pune problema că un singur număr, sau estimare punctuală, pentru un parametru nu este suficient. Este posibil să vrem să specificăm într-un fel și nivelul de încredere care se regăsește în estimarea noastră. Una dintre căile pentru a realiza acest lucru constă în a încerca să găsim un „interval" de valori în cadrul căruia suntem „convinși în proporție de 95%" că se regăsește respectivul parametru. Abordăm această problemă considerând media populației, μ, în felul următor:

Să presupunem că dorim să găsim un interval de valori cuprinse între $\overline{X} + E$ și $\overline{X} - E$, astfel încât, înainte de a extrage eșantionul, există o probabilitate de 0,95 ca intervalul stabilit în cele din urmă să includă parametrul necunoscut μ.

Deoarece $\overline{X}$, respectiv eșantionul mediei, este un estimator nedeplasat al parametrului μ, a-l plasa în centrul intervalului pe care îl căutăm capătă sens. *E* este pur și simplu o expresie, sau formulă, pe care trebuie să o găsim.

Dacă eșantionul este mare, atunci știm din teoria referitoare la limita centrală $\overline{X}$ că este N(μ, $\sigma^2/n$). Prin urmare, putem spune că:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$ urmează o distribuție N(0,1)

Utilizarea tabelelor distribuției normale indică faptul că:

$$Pr(-1{,}96 < Z < 1{,}96) = 0{,}95$$

Utilizând $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ pentru substituirea lui *Z* în formula de mai sus ajungem la expresia:

$$\Pr\left(-1{,}96 \angle \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \angle 1{,}96\right) = 0{,}95$$

Ultima relație se mai poate scrie și sub forma:

$$\Pr\left(\overline{X}+1{,}96\frac{\sigma}{\sqrt{n}} > \mu > \overline{X}-1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

Constatăm că am găsit exact ceea ce căutam: un interval care să garanteze cu probabilitatea de 0,95 că va conține valoarea necunoscută μ. Expresia *E* pe care am căutat-o este de fapt egală cu $1{,}96\ \sigma/\sqrt{n}$.

Intervalul pe care l-am obținut este denumit, în mod normal, *intervalul de încredere de 95%* pentru μ.

Singura problemă în legătură cu acest interval constă în aceea că $E = 1{,}96\ \sigma/\sqrt{n}$ depinde de valoarea σ care, ca și μ, este o necunoscută. În practică, atunci când se calculează un interval de încredere, σ trebuie să fie înlocuit prin *s*, abaterea standard a eșantionului. Prin urmare, intervalul mare de încredere, 95%, al eșantionului se poate rescrie sub forma $\overline{X} \pm E$, sau:

$$\overline{X} \pm 1{,}96\frac{s}{\sqrt{n}}$$

Este posibil să dorim să fim „mai mult de 95% confidenți" asupra faptului că intervalul nostru va conține valoarea μ. Pentru a modifica nivelul de încredere, folosim valoarea corespunzătoare din tabelul distribuției normale standardizate. Pentru a garanta un interval de încredere 99%, înlocuim valoarea de 1,96 cu 2,58 și obținem:

$$\overline{X} \pm 2{,}58\frac{s}{\sqrt{n}}$$

Odată stabilite expresiile de bază, intervalele de încredere sunt ușor de calculat. De exemplu, dacă, în cazul nostru, un eșantion de mărimea *n = 80* ar trebui să conducă la o medie , cu *s = 94,* atunci, la un interval de încredere 95% rezultă:

$$574 \pm 1{,}96\frac{94}{\sqrt{80}} = 574 \pm 20{,}6$$

Cheia acestei probleme rezidă în a ne aminti că diferitele eșantioane vor conduce la diferite medii $\overline{X}$ și la diferite abateri standard, *s.* Prin urmare, diferitele eșantioane vor prezenta *diferite intervale de încredere.* Dacă s-ar extrage mai multe eșantioane, 95% dintre aceste intervale ar conține necunoscuta μ, dar 5% nu ar conține-o. *Intervalele diferă de la eșantion la eșantion, dar* μ *este fix.*

După cum vom vedea, adesea calculăm intervale de încredere pentru parametri ai populației alții decât media μ. Totuși, procesul este întotdeauna similar cu cel prezentat mai sus.

Abaterea standard a distribuției de selecție a unui estimator este cunoscută sub denumirea de *eroare standard a estimării.*

De exemplu, eroarea standard a estimării pentru $\overline{X}$ este $s_{\overline{X}} = s/\sqrt{n}$, respectiv abaterea standard a distribuției sale de selecție. La o estimare punctuală nedeplasată dată și, cu condiția ca distribuția sa de selecție să fie simetrică, intervalele de încredere sunt de forma:

*Estimare punctuală(valoare critică)* ± *(eroarea standard a estimării)*

„Valoarea critică" este luată din tabelele de valori de distribuție, cum ar fi tabelul distribuției normale standardizate.

Vor exista situații când, în loc de a dori să estimăm un parametru al populație, am putea fi interesați să stabilim dacă acest parametru ia sau nu o anumită valoare.

În primul rând, formulăm așa-numita *ipoteză nulă*, conform căreia media câștigurilor populației nu a crescut în anul precedent.

Aceasta presupune că μ este în continuare egal cu $\overline{X}$. O ipoteză nulă este notată, de regulă, prin $H_0$. Astfel, avem:

*Ipoteza nulă* $H_0$ *:* μ = *540 (nicio modificare)*

A se reține că valoarea μ este media populației în anul inițial considerat.

În faza următoare formulăm *ipoteza alternativă,* notată prin $H_A$, care acoperă toate alternativele rezonabile la cea nulă $H_0$. Întrucât anii considerați au fost inflaționiști, vom face abstracție, pentru moment, de posibilitatea ca respectivele câștiguri să fi scăzut și adoptăm ca alternativă ipoteza că acestea au crescut:

*Ipoteza alternativă* $H_A$*:* μ > *540 (creștere a câștigurilor)*

Problema devine astfel una de a alege între $H_0$ și $H_A$, respectiv între ipoteza nulă și cea alternativă. Trebuie să facem acest lucru pe baza informațiilor date de un eșantion de mărime *n* = 100.

Odată eșantionul extras, vom cunoaște valoarea mediei eșantionului, $\overline{X}$. Este evident că *a respinge* ipoteza nulă $H_0$, conform căreia câștigurile nu au crescut, capătă sens dacă se dovedește că $\overline{X}$ are o valoare „mai mare" decât valoarea din anul precedent. O întrebare importantă în context se referă la cât de mare trebuie să fie $\overline{X}$ înainte de a respinge $H_0$ și de *a accepta* alternativa $H_A$, conform căreia câștigurile populației au crescut.

Un instrument de care dispunem pentru a soluționa această problemă este Teorema Limită Centrală. Întrucât eșantionul nostru este unul relativ mare, știm că distribuția mediilor de selecție pentru $\overline{X}$ urmează o distribuție normală, $N(\mu, \sigma^2/n)$.

Cantitatea $(\overline{X} - 540/(\sigma/\sqrt{n})$ este cunoscută sub denumirea de *test statistic (TS)*. Punctul crucial referitor la acest test statistic este dat de faptul că are *o distribuție N(0,1) numai atunci când ipoteza nulă $H_0$ este adevărată.*

Dacă $H_0$ nu este adevărată, ci *falsă,* atunci relația nu se va verifica, deoarece μ nu va lua altă valoare decât 540 RON.

După cum se poate observa, distribuția normală standardizată sau N(0,1) este centrată în jurul valorii zero. Dacă ipoteza nulă este adevărată, în condiții $H_0$ există o mare probabilitate ca TS să ia o valoare în jurul valorii zero. Dacă ar lua o valoare diferită de zero, atunci vom fi înclinați să ne îndoim de faptul că $H_0$ este adevărată. Dacă $H_0$ este falsă, nu există niciun motiv pentru care TS nu ar trebui să ia o valoare depărtată de zero. Prin urmare, testul statistic oferă un mijloc de „testare" a măsurii în care $H_0$ este adevărată.

## Bibliografie


Anghelache, C-tin (2004). *Statistică teoretică și economică – teorie și aplicații,* Editura Economică, București

Anghelache, C-tin (2004). *Sistemul European al Conturilor – note de curs*, Editura ARTIFEX, București

Anghelache, C-tin, Capanu, I. (2003). *Indicatori macroeconomici – calcul și analiză economică*, Editura Economică, București

Andrei, T. (2003). *Statistică și econometrie*, Editura Economică, București

Biji, M., Biji, M.E., Lilea, E., Anghelache, C-tin (2002). *Tratat de statistică*, Editura Economică, București

Capanu, I., Anghelache, C-tin (2003). *Indicatori economici pentru managementul micro și macroeconomic – calcul, prezentare, analiză*, Editura Economică, București

Capanu, I., Wagner, P., Mitruț, C-tin (2004). *Sistemul Conturilor Naționale și Agregate macroeconomice*, Editura ALL, București

Dobrescu, E. (1996). *Macromodels of the Romanian Transition Economy,* Editura Expert, București

Gilbert, M., Kravis, I. (1954). *An International Comparison of National Product and Purchasing Power of Currencies*, OEEC, Paris

Isaic-Maniu, Al., Mitruț, C., Voineagu, V. (1995). *Macroeconomie și analiză macroeconomică,* Editura „Constantin Brâncoveanu", Rm. Vâlcea