

Metode de analiză statistică folosite în cercetarea fenomenelor teritoriale

Lector universitar doctor ecolog, doctor în geografie, abilitat în urbanism Alexandru-Ionuț PETRIȘOR

INFORMAȚII DESPRE CURS ȘI ATELIER

Programul de studii	Facultatea de Urbanism, Școala Doctorală de Urbanism
Tipul cursului	Obligatoriu
Nivelul cursului	Doctorat
Număr de credite ECTS	3
Competențe de dezvoltat	1) Înțelegerea principalelor concepte folosite în statistică și analiza datelor 2) Cunoașterea metodelor de lucru folosite în analiza statistică, inclusiv a tehnicilor bazate pe folosirea calculatorului 3) Interpretarea rezultatelor finale ale analizelor din punctul de vedere al relevanței în urbanism 4) Utilizarea corectă a vocabularul tehnic și a termenilor de specialitate
Obiective	Cursul urmărește familiarizarea studenților cu principalele concepte și metode folosite în analiza statistică a datelor din cercetare, cu accent pe aplicarea lor în cercetarea fenomenelor teritoriale și interpretarea rezultatelor în termenii specifici domeniului.
Metoda de predare	Prelegeri, prezentări PowerPoint, demonstrații pe calculator (ArcView, Excel), discuții
Mod de evaluare	Prezența la cel puțin 75% dintre cursuri permite acumularea a 3 credite.
Bibliografie	1) Dragomirescu L. (1998), <i>Biostatistică pentru începători</i> , Editura Constelații, București, 220 pag.; (alternativ) Dragomirescu L., Drane J. W. (2001), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a II-a, Editura Ars Docendi, București, 165 pag.; Dragomirescu L., Drane J. W. (2005), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a II-a., Editura CREDIS, București, 165 pag.; Dragomirescu L., Drane J. W. (2006), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a III-a revăzută și adăugită, Editura CREDIS, București, 206 pag.; Dragomirescu L., Drane J. W. (2007), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a IV-a revăzută și adăugită, Editura CREDIS, București, 186 pag.; Dragomirescu L., Drane J. W. (2008), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a V-a, Editura CREDIS, București, 206 pag.; Dragomirescu L., Drane J. W. (2009), <i>Biostatistică pentru începători</i> , Vol. I. <i>Biostatistică descriptivă</i> , Ediția a VI-a revăzută, Editura CREDIS, București, 207 pag. 2) Petrișor AI (2007), <i>Elemente de statistică urbană cu aplicații rezolvate în Excel</i> , Editura Universitară „Ion Mincu”, București, 51 pag. 3) Petrișor AI (2009), <i>Elemente de regresie aplicată în științele medicale, științele vieții și științele Pământului</i> , Editura Ars Docendi - Universitatea din București, București, 109 pag.

SINTEZA PRELEGERILOR*

Partea I. Introducere în statistică

Statistica este un instrument de cunoaștere științifică, un mod de gândire complex și dificil de asimilat, deoarece cere experiență, fler, o cunoaștere aprofundată a multor subdomenii – cele mai multe aparținând matematicii – precum și o anumită etică profesională, având ca **scop** studiul unei mulțimi de observații efectuate asupra unei mulțimi de obiecte de aceeași natură, denumite unități statistice și care prezintă caracteristici variabile (numite simplu variabile) susceptibile de a fi clasate, ordonate sau măsurate. Mulțimea acestor obiecte se numește serie statistică sau șir statistic.

* Bazată în marea majoritate pe: Dragomirescu L. (1998), *Biostatistică pentru începători*, Editura Constelații, București, 220 pag.

Mulțimile de unități statistice sunt de două tipuri:

- **Populațiile statistice** sunt mulțimi de unități statistice (obiecte, indivizi, fenomene, evenimente, idei, opinii, numere etc.) la care se limitează interesul cercetătorilor. Sunt foarte mari, majoritatea fiind infinite, și de aceea studiul lor exhaustiv nu este economic sau este imposibil.
- **Eșantioanele** sunt submulțimi ale populațiilor statistice considerate pentru a obține informații despre populațiile respective.

Având în vedere tipul de mulțime studiată, se disting:

- **Statistica descriptivă:** (1) sintetizează grafic și numeric informația culeasă exhaustiv dintr-o populație statistică, (2) descrie fără să explice esențialul din date, și (3) se realizează prin sinteze grafice și numerice, în care se renunță la o parte din informație pentru a câștiga în relevanță – paradigma centrală a statisticii, după Paul Benzecri.
- **Statistica inductivă sau inferențială:** (1) cuprinde: teoria eșantionajului, teoria estimației, verificarea ipotezelor statistice și planificarea experimentelor, (2) este o cercetare selectivă, prin eșantioane, și (3) are ca obiectiv obținerea unui maximum de informație asupra populației cu un minimum de efort – îmbunătățirea reprezentativității eșantioanelor.

Conceptul de probabilitate. Probabilitatea este o caracteristică numerică a gradului posibilității de realizare a unui eveniment sau fenomen în condiții bine determinate. Probabilitatea poate fi obiectivă sau subiectivă. **Probabilitatea obiectivă** are două definiții: (1) definiția clasică: raportul dintre numărul de cazuri favorabile producerii unui eveniment și numărul total de cazuri, și (2) definiția empiristă (von Mises): numărul spre care tinde să se stabilizeze frecvența relativă a evenimentului respectiv pe măsură ce evenimentul se repetă de un număr cât mai mare de ori. **Probabilitatea subiectivă** reprezintă o codificare subiectivă de informații efectuată de o persoană interesată de a o evalua – traducerea bunului simț în cifre.

Tipuri de scale: dacă A și B sunt două unități statistice, iar x o variabilă cu caracteristicile x_A și x_B , (1) scala **nominală** realizează doar o distincție: $x_A = x_B$ sau $x_A \neq x_B$, (2) scala **ordinală** are în plus o ordine: dacă $x_A \neq x_B$, atunci $x_A > x_B$ sau $x_A < x_B$, (3) scala **de intervale egale** are în plus o măsură semnificativă a diferenței dintre două valori: dacă $x_A > x_B$, atunci spunem că A este mai mare decât B cu $x_A - x_B$, (4) scala **de proporții egale** (scala **raport**) are în plus un zero absolut, implicit și o măsură semnificativă a proporției dintre două valori; astfel, putem spune că A este mai mare decât B de x_A/x_B ori.

Clasificarea variabilelor: (1) variabile **calitative** – cele de pe scala nominală, inclusiv cele binare, (2) variabile **tip rang** – cele de pe scala ordinală, (3) variabile **tip măsurătoare/dimensiuni** – cele de pe scalele interval sau raport.

Clasificarea seriilor statistice: (1) în funcție de exhaustivitatea datelor: **populații** și **eșantioane** statistice; (2) în funcție de numărul de variabile luate în considerație simultan: **univariate**, **bivariate** sau **multivariate**.

Partea a II-a. Statistică descriptivă

1. Sintează grafică univariată

Sinteza grafică univariată se face pentru evidențierea intuitivă și aproximativă a aspectelor de **variabilitate** dintr-o serie statistică în două etape: (1) construirea tabelor statistice simple/cu simplă intrare, (2) reprezentări grafice adecvate tipului de variabilă.

Sinteza grafică în tabele statistice se poate face prin grupare fără pierdere de informație – tabelele conțin distribuțiile de frecvențe ale variantelor sau valorilor, sau grupare cu pierdere de informație – tabelele conțin distribuțiile de frecvențe ale claselor sau intervalelor de grupare. O distribuție de frecvențe conține aceeași informație ca și seria din care provine, dar este mai intuitivă. O distribuție de frecvențe grupate conține mai puțină informație decât seria din care provine, dar este mai relevantă.

Sinteza grafică univariată se realizează prin diferite instrumente în funcție de tipul variabilelor. Pentru **variabilele calitative** se recomandă diagramele circulare, pentru cele **tip rang** se recomandă diagramele în batoane sau poligoanele sau curbele de frecvențe, iar pentru variabilele **tip măsurătoare** se recomandă diagramele în batoane, poligoanele sau curbele de frecvențe și, în special, histogramele.

Tipuri de reprezentări grafice

- Se numește **diagramă în batoane** reprezentarea carteziană plană în care pe axa orizontală sunt marcate variantele sau valorile, fiecareia dintre acestea corespunzându-i pe verticală o coloană de înălțime proporțională cu frecvența corespunzătoare.
- Se numește **diagramă circulară** cercul format din sectoare pentru fiecare variantă sau valoare, astfel încât unghiul, respectiv aria fiecărui sector să fie proporțional(ă) cu frecvența respectivă.
- Se numește **histogramă** reprezentarea carteziană plană a unei distribuții grupate (în care se specifică valorile și frecvențele acestora), formată din dreptunghiuri alipite, cu bazele plasate pe intervalele de grupare și cu ariile proporționale cu frecvențele claselor.
- În demografie, volumele celor două sexe (m / f) pe vârste sau grupe de vârstă dintr-o populație umană sunt reprezentate prin două histograme cu bazele – reprezentând vârsta – alipite pe verticală. Datorită faptului că odată cu înaintarea în vârstă, generațiile scad ca volum din cauza mortalității, reprezentarea are aspectul unei piramide numită **piramida vârstelor**. În funcție de forma acesteia, se disting 3 tipuri: (1) baza îngustată indică o populație în declin, fenomen denumit și creștere negativă sau îmbătrânirea populației – volumul total este în scădere; (2) forma ideală a piramidei indică o populație în echilibru staționar – volumul total rămâne constant (3) baza lărgită indică o populație în creștere prin întinerire – volumul total este în creștere.
- Se numește **poligon de frecvențe** linia frântă formată din segmentele care unesc mijloacele laturilor din vârfurile coloanelor consecutive figurate în diagrama prin batoane, fără a mai reprezenta și coloanele, sau mijloacele laturilor superioare ale dreptunghiurilor prin care sunt reprezentate clasele sau intervalele de grupare în histograme, fără a reprezenta și dreptunghiurile.
- Poligoanele de frecvențe tind odată cu creșterea preciziei și a numărului de măsurători către o curbă netedă numită **curbă de frecvențe**. Curbele de frecvențe se „străvăd” în spatele diagramelor prin batoane, histogramelor sau poligoanelor de frecvențe și nu se obțin direct, ci prin creșterea numărului de observații sau a preciziei. Aceste curbe sunt curbe teoretice. Există programe care „netezesc” poligoanele de frecvențe. Se obțin astfel curbe empirice care sunt, de fapt, o formă convențională de reprezentare a poligoanelor de frecvențe, și pot eventual sugera aspectul curbelor teoretice. Curbele obținute prin netezirea poligoanelor de frecvențe sunt mai intuitive, dar curbele teoretice sunt mai relevante.

Curba erorilor de măsurare. Procesul de măsurare este afectat de două tipuri de erori: (1) **erori aleatoare**, care sunt mici, se produc în ambele sensuri și se compensează reciproc, (2) **erori sistematice, grosolane**, care se produc sistematic în același sens și produc o deplasare sistematică a valorii căutate. Curba erorilor de măsurare are o formă de clopot, care ilustrează următoarele aspecte: (1) marea majoritate a măsurătorilor au valori apropiate de centrul distribuției (care este, cel mai probabil, valoarea corectă sau reală), (2) numărul măsurătorilor care se abat de la centru scade odată cu creșterea abaterii de la centru, (3) numărul măsurătorilor cu o anumită abatere pozitivă este relativ egal cu numărul măsurătorilor cu aceeași abatere, dar de semn negativ.

Limbajul repartițiilor

(1) De ce grupăm?

Pentru a câștiga în relevanță.

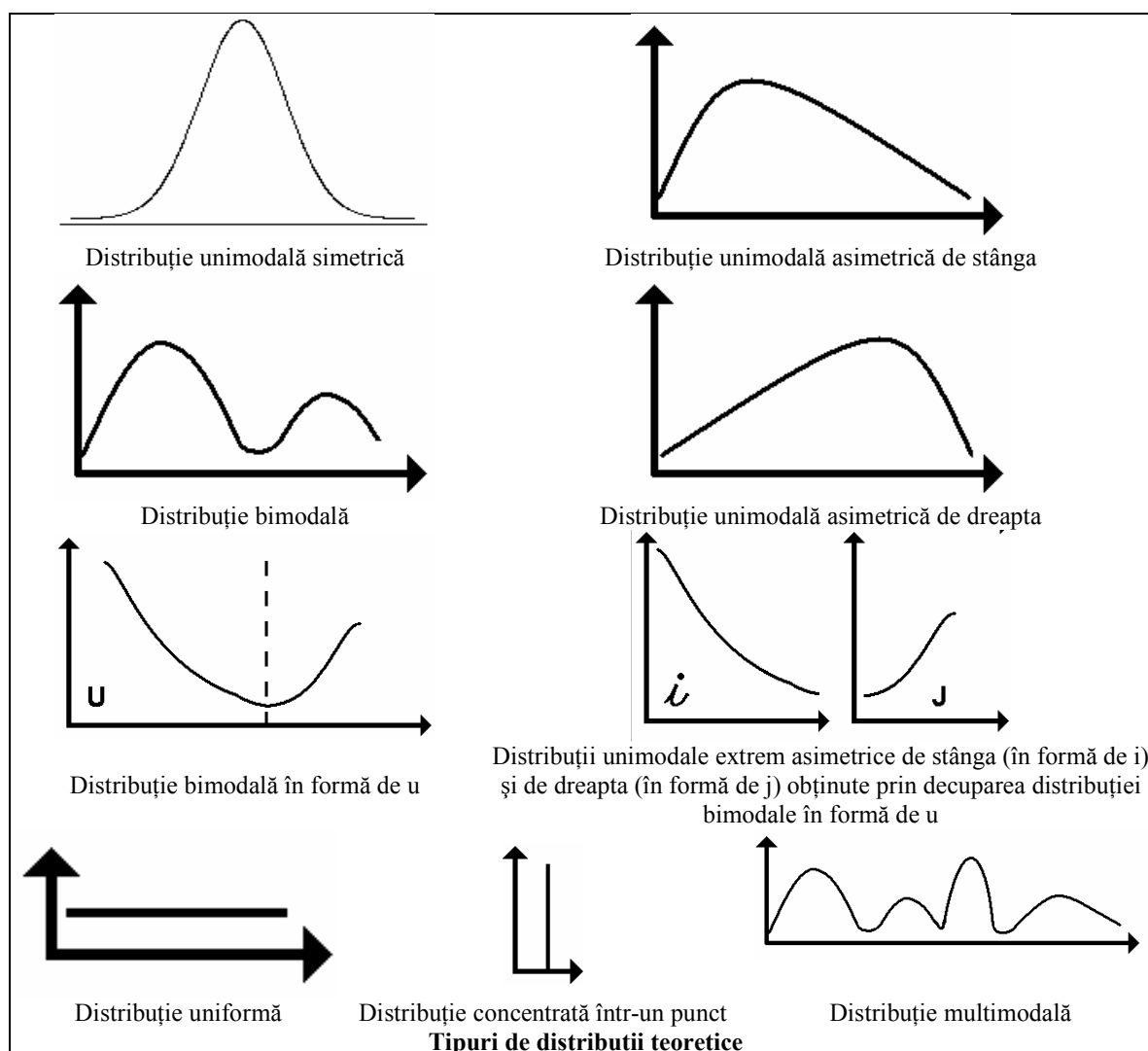
(2) Pentru ce grupăm?

Pentru a sesiza o anumită formă a distribuției (a se vedea mai jos), pe baza căreia se vor diferenția analizele statistice.

O **modă** este un **punct de maxim local**. O distribuție **unimodală** are o singură modă, o distribuție **bimodală** prezintă două mode, iar distribuțiile **multimodale**, două sau mai multe mode. În cazul distribuțiilor **uniforme**, fiecare valoare este modă.

Interpretare: distribuțiile **concentrate într-un punct** exprimă omogenitate absolută; distribuțiile **simetrice** exprimă cel mai bine o tendență centrală; distribuțiile **bi – sau multimodale** exprimă eterogenitate ca amestec de distribuții omogene; distribuțiile **uniforme** exprimă eterogenitate absolută.

Consecințe: distribuțiile **multimodale** vor fi „decupate” în distribuții **unimodale**, care vor fi analizate separat, iar distribuțiile **asimetrice** vor fi transformate în distribuții **simetrice** cu ajutorul unor instrumente matematice.



2. Sintează numerică univariată

Sinteza numerică univariată se face pentru evidențierea obiectivă și exactă a aspectelor esențiale ale variabilității unei serii statistice, percepută ca împrăștiere în jurul unei tendințe centrale.

2.1. Indicatori de tendință centrală

- **Media aritmetică:** $M = \frac{\sum_{j=1}^N X_j}{N} = \sum_{i=1}^p f_i \times X_i$
- **Mediana** se notează cu Me și este punctul care împarte aria de sub curba de frecvențe în două arii egale sau o valoare care împarte seria statistică ordonată în 2 subserii de volume egale, măsurate ca număr de unități statistice sau fracțiuni ale acestora.

Media de viață. Mortalitatea populației în funcție de vârstă pe o curbă de frecvențe are o **mediană** care reprezintă vârsta până la care au murit 50% din indivizii populației respective, și care se numește **media de vârstă a populației**. Calculul acesteia reprezintă o situație de excepție în care se calculează un indicator de tendință centrală pentru o distribuție multimodală.

- **Moda (Mo)** este un punct de maxim local în cazul unei curbe de frecvențe sau valoarea cu frecvența maximă locală în distribuția de frecvențe.

Pentru distribuțiile unimodale simetrice, cei trei indicatori coincid; la distribuțiile unimodale ușor asimetrice, mediana este plasată între modă și medie, distanța față de modă fiind dublul celei față de medie. Sensul asimetriei poate fi detectat aproximativ prin echivalența: asimetrie de stânga – $Mo < Me (M)$, asimetrie de dreapta – $Mo > Me (M)$.

2.2. Indicatori de împrăștiere

Indicatorii de împrăștiere se bazează pe **indicatorii de tendință centrală** – dispersia, abaterea standard și coeficientul (procentual) de variație, sau pe **indicatorii de tendință extremă** – amplitudinea.

- Dispersia: $S^2 = \frac{\sum_{j=1}^N (X_j - M)^2}{N} = \sum_{i=1}^p f_i \times (X_i - M)^2$
- Abaterea standard se notează cu S și este rădăcina pătrată a dispersiei: $S = \sqrt{S^2}$
- Coeficientul (procentual) de variabilitate, notat cu CV, reprezintă procentul reprezentat de abaterea standard S din media M: $CV = 100 \times \frac{S}{M}$.

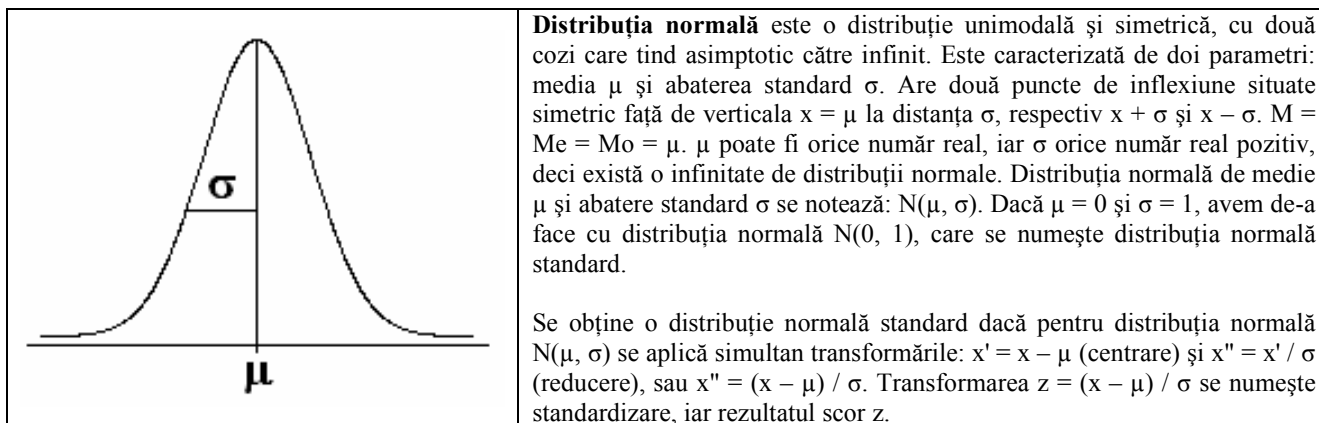
Reguli empirice referitoare la coeficientul (procentual) de variație: $CV\% < 10\%$ – populație omogenă, $CV\% > 30\%$ – populație eterogenă, $10\% < CV\% < 20\%$ – populație relativ omogenă sau chiar omogenă, în funcție de variabilă, $20\% < CV\% < 30\%$ – populație relativ eterogenă.

- Amplitudinea reprezintă diferența dintre valoarea maximă și valoarea minimă dintr-o serie.

2.3. Eliminarea valorilor aberante

Sunt sensibile la valorile aberante media, dispersia, abaterea standard și coeficientul (procentual) de variație. Prin urmare, aceste valori trebuie eliminate dacă dorim să folosim acești indicatori (și dorim, datorită proprietăților matematice, de exemplu aditivitatea). Pentru a le elimina, este necesară identificarea lor. Pentru serii de volum mare ($N > 30$) se utilizează regula „**3 sigma**” de eliminare a valorilor aberante. Regula se bazează pe inegalitatea lui Cebâșev: „orice distribuție se întinde între

media sa plus / minus 6 abateri standard”, și pe faptul că aria cuprinsă între media unei distribuții normale plus / minus 3 abateri standard reprezintă circa 99,7% din aria totală. Valorile situate în afara acestor limite sunt improbabile sau aberante, și neglijabile. Aplicarea regulii presupune existența unei distribuții normale a datelor / fenomenului. Dacă acest lucru nu este cunoscut, se aplică regula „6 sigma”.



Măsurarea gradului de concordanță cu o distribuție normală: (1) prin sinteză grafică – reprezentăm datele sub forma unei histogramme, suprapunem peste histogramă distribuția normală cu aceeași medie și cu aceeași abatere standard cu distribuția datelor și apreciem pe baza graficului

concordanța distribuțiilor; (2) prin sinteză numerică: $\chi^2 = \sum_{j=1}^p \frac{(o_j - t_j)^2}{t_j}$, unde o_j sunt frecvențele

observate, corespunzătoare distribuției analizate, iar t_j sunt frecvențele teoretice, corespunzătoare distribuției normale. Pentru a calcula frecvențele teoretice, avem nevoie de tabele ale distribuției normale standard, după care, prin anumite calcule, vom determina valorile teoretice pentru o distribuție normală cu media și abaterea standard egale cu cele ale distribuției analizate.

Partea a III-a. Statistică inferențială

Studiile efectuate asupra populațiilor de către statistica descriptivă produc rezultate certe, iar cele efectuate asupra eșantioanelor de către statistica inductivă/inferențială produc rezultate incerte. Exprimarea științifică a gradului de incertitudine este făcută în statistica inductivă prin inferență statistică. **Inferența statistică** reprezintă extrapolarea judecăților obținute pe eșantioane (prelevate prin anumite procedee statistico-matematice) la populațiile statistice din care au fost extrase.

Statistica inferențială este o cercetare selectivă, prin eșantioane, care are ca obiectiv obținerea unui maximum de informație asupra populației cu un minimum de efort – îmbunătățirea reprezentativității eșantioanelor, și cuprinde (1) teoria eșantionajului (sondajului, selecției), (2) teoria estimăției, (3) verificarea ipotezelor statistice, (4) planificarea experimentelor.

Mic dicționar de statistică inferențială

- **A estima** – a calcula aproximativ sau a ne forma o idee aproximativă despre ceva
- **Estimare** – procesul de aproximare a unei valori, de formare a unei idei aproximative despre ceva
- **Estimație** – rezultatul estimării, valoare numerică sau interval de încredere (confidență), care are asociat(ă) un anumit grad (nivel) de încredere

- **Estimată** – valoare sau parametru necunoscut, supus procesului de estimare
- **Estimator** – persoană sau instrument care efectuează o estimare: formulă sau procedură, variabilă aleatoare care apare ca rezultat al aplicării formulei la datele unui eșantion aleator
- **Parametru** – indicator al unei **populații statistice**
- **Statistică** – indicator al unui **eșantion**

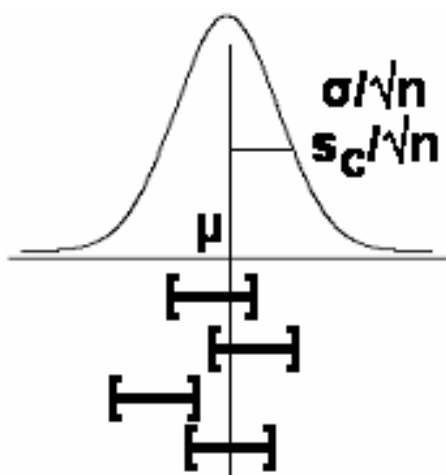
1. Teoria eșantionajului

(a) Modalități de extragere a eșantioanelor

- **Fără revenire** – odată inclus într-un eșantion, un subiect nu mai poate fi inclus în altul
- **Cu revenire** – odată inclus într-un eșantion, un subiect poate fi inclus în alte eșantioane

(b) Tipuri de eșantioane

- **Eșantion aleator simplu** – extragerea se realizează cu ajutorul numerelor aleatoare. Acestea apar din procese cu adevărat aleatoare, cum ar fi descompunerea izotopilor radioactivi. Pentru extragerea eșantioanelor se consultă tabele de numere aleatoare. Numerele pseudoaleatoare sunt generate prin programe ale calculatoarelor. Numerele generate de o persoană „într-o doară” sau „la întâmplare” nu sunt o soluție, deoarece realizarea aleatorismului este unul din cele mai grele lucruri pentru om.
- **Eșantion sistematic** – se extrage elementul k dintr-un anumit cadru – de exemplu, fiecare a Y -a persoană de pe pagina Z , $Z + 10$, $Z + 20$ etc. din cartea de telefon, Y și Z fiind determinate aleator
- **Eșantion stratificat** – eșantion aleatoriu simplu extras din fiecare strat al unei populații
- **Eșantion de tip cluster** – asemănător celui stratificat, doar că în loc de elemente se extrag clustere (grupuri) care apar în mod natural



2. Teoria estimăției

2.1. Proprietățile unui estimator

Un estimator se numește **just (nedeplasat)** dacă media sa coincide cu parametrul estimat. În caz contrar se numește *injust* sau *deplasat*. În practică, o estimăție „corectă” este asigurată de folosirea unui estimator just și de asigurarea aleatorismului extragerii eșantionului.

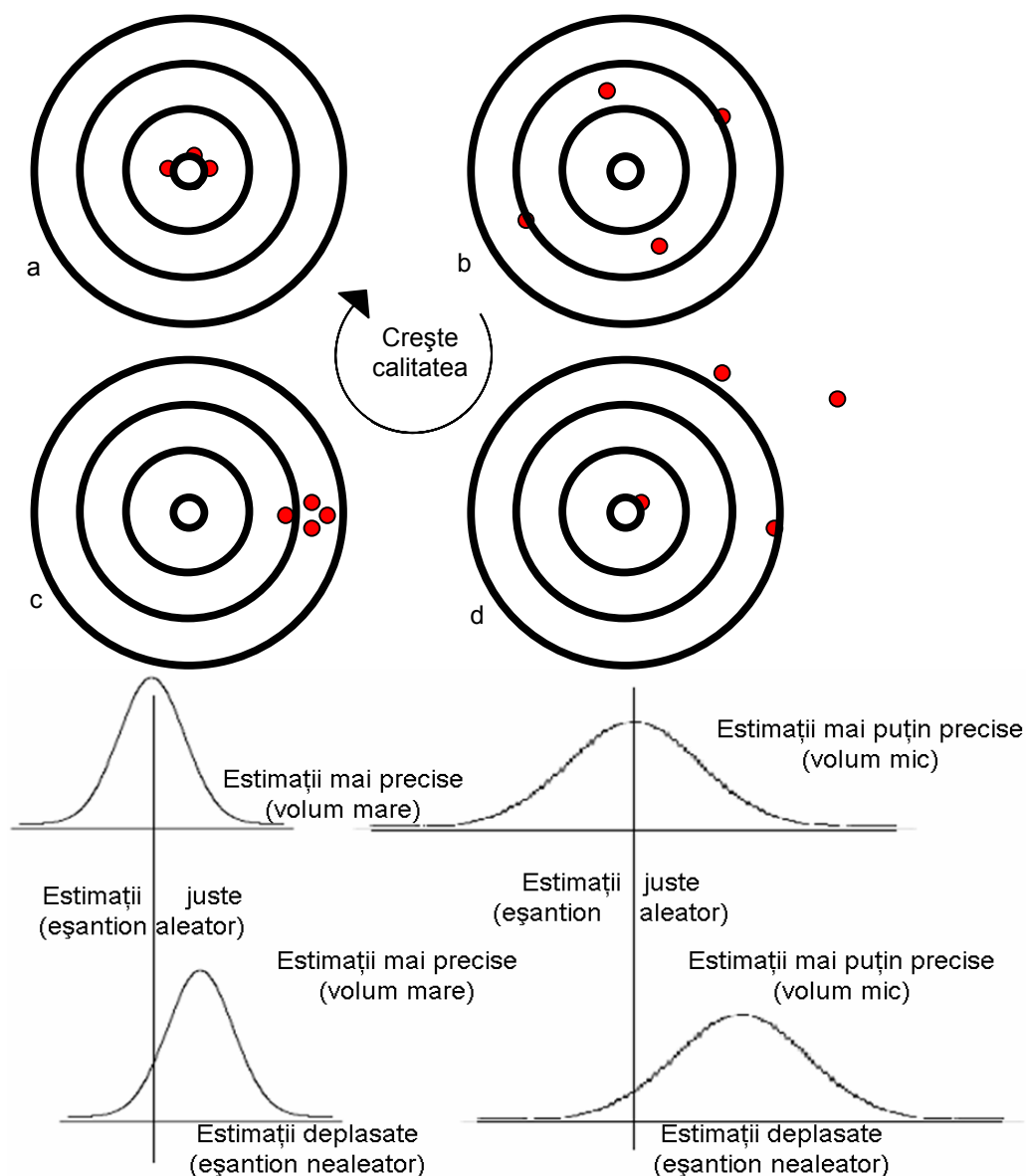
Un estimator este **mai efecace decât alt estimator** dacă are o eroare standard mai mică. Cel mai efecace estimator al aceluiași parametru studiat se numește **estimator eficient**.

Asigurarea justeței și preciziei estimării

Eșantion aleator	și estimator just	estimăție „corectă”
În plus: eșantion de volum mare	și estimator mai efecace	estimăție mai precisă
În particular: eșantion de volum fixat	și estimator eficient	cea mai precisă estimăție

Analogie între proprietățile unui estimator și cele ale unui țințaș

Un bun țințaș face o tragere de probă pentru a vedea dacă pușca are „bătaie”. În acest caz, corectează punctul de ochire pentru a o compensa. Doi trăgători care lovesc fără bătaie diferă prin precizie (cazurile c și d).



În mod analog, în primul rând un estimator trebuie să fie în primul rând just, apoi cât mai eficace. Dacă oricine identifică situația a ca fiind cea ideală, cea mai defavorabilă situație nu este d, cum s-ar putea crede la o primă vedere, ci c. Dintre cele două obiective, cel de a lovi ținta este cel mai important și se poate observa că în situația d există o șansă de a lovi ținta, care se „pierde” în situația c.

3. Testarea ipotezelor statistice

Tentativa de a explica una sau mai multe observații științifice se numește **ipoteză științifică**. Pentru a fi susținute, ipotezele științifice au nevoie de seturi de date (observații, experimente) și de o susținere statistică. **Ipotezele statistice** sunt afirmații privind una sau mai multe populații statistice pentru verificarea ipotezelor științifice. O ipoteză statistică este formată dintr-o **ipoteză nulă** (formată de regulă sub forma „nu există diferențe”) și o **ipoteză alternativă** care o contrazice și care corespunde ipotezei științifice. În urma aplicării unui **test statistic**, se respinge ipoteza nulă dacă au fost detectate diferențe semnificative sau nu, în caz contrar. **Diferențele semnificative** sunt atât de mari (comparate cu un **nivel de semnificație** notat cu α), încât ele nu pot fi atribuite fluctuațiilor datorate extragerii aleatoare a eșantioanelor, ci provin dintr-un motiv semnificativ, mai exact cel precizat de ipoteza științifică.

Etapele aplicării unui test statistic sunt:

1. Formularea clară a problemei pentru care se dorește o decizie
2. Identificarea variabilei (tip, scală), eșantionului (mod de extragere, volum), informațiilor despre distribuția variabilei în populație (gaussianitate, transformabilitate)

3. Stabilirea distribuției de eșantionaj a statisticii testului.
4. Formularea ipotezelor de studiu, determinarea tipului de test (unilateral etc.).
5. Stabilirea nivelului de semnificație pentru respingerea ipotezei nule (regula de decizie)
6. Efectuarea calculelor, obținerea deciziei prin calcul manual (al statisticii) sau automat (al probabilității de respingere).

Principalele tipuri de teste statistice sunt:

1. Clasificarea ca teste binare: teste **bilaterale** (orice diferențe) și **unilaterale** (stânga sau dreapta).
2. Clasificarea după modul de tratare a variabilelor cantitative: ca parametri – **teste parametrice** și ca ranguri sau variabile calitative – **teste neparametrice**.
3. Clasificarea după tipul ipotezelor verificate
 - **Teste de conformitate** – compară un eșantion cu o populație pe baza unui indicator particular
 - **Teste de concordanță** – compară o populație cu un eșantion pe baza unui maxim de informație posibilă (de exemplu, testul χ^2 , prezentat anterior – acesta se extinde alegând ca referință orice distribuție teoretică sau empirică).
 - **Teste de egalitate (omogenitate, comparație)** – compară două sau mai multe populații prin tot atâtea eșantioane pe baza unui indicator
 - **Teste de independență versus asociere** – studiază asocierea caracterelor, relații cauzale, compară două metode etc.

În testarea ipotezelor statistice se produc două tipuri de erori, cărora le corespund două tipuri de riscuri:

	H₀ acceptată	H₀ respinsă
H₀ adevărată	Decizie corectă (probabilitate: $1 - \alpha$)	Eroare de speța I (risc α , riscul bețivului, riscul furnizorului)
H₀ falsă	Eroare de speța a doua (risc β , riscul chiorului, riscul beneficiarului)	Decizie corectă (probabilitate: $\pi = 1 - \beta$)

α este riscul de speța I, riscul de a respinge ipoteza nulă atunci când aceasta este adevărată, deci de a detecta diferențe care nu există în realitate, ci se datorează fluctuațiilor de extragere a eșantioanelor. Cu cât diferențele sunt mai mari, cu atât probabilitatea ca acestea să se datoreze fluctuațiilor de eșantionaj și să nu aibă o bază reală scade. De aceea în practică se folosesc pragurile din tabelul următor.

În ecologie, frecvent se consideră semnificativ $\alpha = 10/100$, chiar $20/100$ (fiind mai important ca valoarea β să fie redusă – se preferă, de exemplu, „alarmă falsă” a depășirii concentrației unor poluanți, exprimată de riscul α , nedetectării depășirii reale a acestora – riscul β).

Semnificația statistică a diferențelor se traduce fie în valori ale testului statistic mai îndepărtate de cele din condițiile ipotezei de nul (corespunzătoare unui nivel prestabilit al riscului α), fie, având în vedere utilizarea programelor specializate de statistică pentru a testa ipotezele statistice, în „**valoarea p**”, reprezentând probabilitatea de a obține (datorită fluctuațiilor de eșantionaj) o valoare a testului statistic egală cu cea obținută pentru datele analizate sau chiar și mai îndepărtată de aceasta în condițiile în care ipoteza nulă este adevărată – măsură a incertitudinii datorate unui studiu pe baza eșantioanelor. Interpretarea valorilor **p** corespunde scalei Michelin (tabelul de mai jos).

Valoarea p	Denumire
$0,01 < p \leq 0,05$	diferență semnificativă (*)
$0,001 < p \leq 0,01$	diferență înalt semnificativă (**)
$p \leq 0,001$	diferență foarte înalt semnificativă (***)
$p \geq 0,05$	diferență nesemnificativă
$0,1 > p \geq 0,05$	zona de incertitudine/diferență cu semnificație marginală: date suplimentare (creșterea volumului eșantionului) pot crește semnificația peste pragul de 0,05

β este riscul de speța a II-a, riscul de a accepta ipoteza nulă atunci când aceasta este falsă, deci de nu sesiza diferențe care există în realitate. Relația dintre α și β este complexă, dar, simplificat, poate fi exprimată astfel: dacă α crește, β scade.

Puterea unui test statistic este $\pi = 1 - \beta$. Deși redundant exprimat, reprezintă puterea unui test statistic de a detecta diferențe atunci când acestea există în realitate. Ea crește cu α și cu volumul eșantionului, acest aspect fiind relevant atât pentru teoria eșantionajului, cât și pentru planificarea experimentelor. Anumite programe calculează, în condiții bine determinate, puterea testului fiind stabilite α și volumul eșantionului, sau, cunoscând α , volumul minim al eșantionului necesar pentru a obține o anumită putere a testului statistic.

Statistică confirmatorie și exploratorie. În statistica inductivă se disting două ramuri. **Statistica confirmatorie** cuprinde toate mecanismele în care, plecând de la o ipoteză, se realizează designul experimental, se extrag eșantioane, se prelucrează datele și în urma prelucrărilor statistice se resping sau nu ipotezele inițiale, se fac estimări și predicții, în condiții bine determinate impuse distribuției datelor. **Statistica exploratorie** (numită și **analiză exploratorie a datelor**) pornește de la datele experimentale pentru a identifica prin simple calcule aritmetice și reprezentări grafice anumite structuri ale acestora, a oferi descrieri parțiale, a vedea ce se află „în spate”, testarea unor ipoteze privind aceste elemente fiind posibilă într-un număr redus de situații și sub rezerva unor limite foarte mari în extrapolarea rezultatelor. Un astfel de instrument permite identificarea unor modele pornind de la eșantion, permițând specialistului ca, în cazul în care observă o repetabilitate a acestor structuri, să testeze validitatea modelelor prin alte metode. Statistica exploratorie a apărut deoarece se pierde anumite informații dacă se lucrează în condiții controlate de presupuneri despre care se știe că nu pot fi verificate și se bazează pe principiul că este importantă înțelegerea a ceea ce se poate face înainte de a învăța metoda de a măsura cât de bine pare a fi făcut ceva. Orice metodă care simplifică descrierea îi crește utilitatea, și orice metodă care arată ceva în plus face descrierea mai efektivă. Obiectivul statisticii exploratorii este, astfel, de a explica și nu de a confirma.

Statistică parametrică și neparametrică. În statistică se deosebesc metode **parametrice** și **neparametrice**. Metodele parametrice sunt mai puternice, dar mai puțin robuste; cele neparametrice sunt mai puțin puternice, dar mai robuste, în consecință rezultatele obținute cu ajutorul primelor sunt chestionabile dacă nu se respectă condițiile de aplicare. Cel mai adesea, condițiile vizează distribuția normală a datelor. În contextul acestor condiții se vorbește despre robustețea unui test sau estimator (capacitatea de a opera corect în condițiile în care nu sunt îndeplinite condițiile de aplicare) pentru a descrie capacitatea sa de a opera corect atunci când nu sunt îndeplinite condițiile necesare folosirii metodelor parametrice. Testele de normalitate se folosesc pentru a analiza distribuția și a alege metoda potrivită. Indiferent de domeniul de studiu, foarte puțini cercetători folosesc metode neparametrice, chiar dacă distribuția datelor se abate de la normalitate.

Corelație. Cauzalitatea determină clasificarea relațiilor în **false (aparente)** (acestea există în date, dar nu au susținere în plan fenomenologic și pot fi, la rândul lor, **artefactuale (fictive)** sau **indirecte**) și **cauzale (etiologice)**; cele artefactuale se datorează fluctuațiilor de eșantionaj, iar cele indirecte, confundării (cele două variabile sunt corelate cu o terță variabilă, neluată în calcul, și din această cauză instrumentarul statistic poate identifica o corelație inexistentă în plan fenomenologic). Relațiile cauzale pot fi necesare și suficiente, necesare dar insuficiente, suficiente dar nenesecare, insuficiente și nenesecare. Cauzalitatea este analizată pe baza următoarelor elemente: succesiunea temporală, tăria asocierii, specificitatea, diverse analogii, existența relației doză-răspuns, replicarea rezultatelor, plauzibilitatea biologică, considerarea explicațiilor alternative, evoluția la momentul încetării expunerii și după aceasta, și corelarea cu alte studii.

Există corelație în plan fenomenologic (din motive de specialitate)	→	Există corelație în date (în condiții corecte de design experimental)
Există corelație în date	→	Poate exista corelație în plan fenomenologic (nu este obligatoriu – a se vedea corelațiile false)
Nu există corelație în date	→	Nu există corelație în plan fenomenologic

Analiza factorială și analiza componentelor principale. Analiza factorială exploratorie urmărește să identifice natura unor construcții care influențează un set de variabile de răspuns, iar analiza factorială confirmatorie testează dacă un anumit set de construcții influențează într-o manieră prestabilită răspunsurile. În toate cazurile modelul este unul asemănător cu al regresiei, cu deosebirea că în analiza factorială variabila sau variabilele dependente nu sunt măsurate și, în unele cazuri, nici măcar cunoscute. Din acest punct de vedere, principiul analizei factoriale se aseamănă cu cel al metodei analizei componentelor principale, diferența constând în faptul că ultima metodă are ca obiectiv reducerea datelor, în sensul obținerii unui set de variabile care să explice cât mai mult din variabilitatea observată. Există două diferențe între cele două metode: (1) dacă analiza factorială exploratorie presupune că variabilele măsurate se datorează influenței unor factori, analiza componentelor principale presupune că aceste componente se bazează pe variabilele măsurate și (2) analiza factorială exploratorie presupune că variația poate fi descompusă în cea explicată de factorii comuni și cea explicată de factorii unici, în timp ce componentele principale sunt combinații liniare ale variabilelor măsurate și conțin variația comună și pe cea unică.

3. Planificarea experimentelor

Planificarea experimentelor ia în considerație teoria eșantionajului, estimării și testării ipotezelor statistice pentru a asigura respectarea condițiilor statistico-matematice impuse de fiecare metodă de analiză a datelor, respectiv pentru asigurarea corectitudinii statistico-matematice a desfășurării experimentelor științifice. Acest proces presupune următoarele etape: (1) formularea ipotezelor statistice, (2) determinarea condițiilor experimentale (variabile independente), a celor externe (zgomot) care trebuie controlate, (3) specificarea volumului eșantionului și definirea populației, (4) specificarea procedurii de alocare a subiecților condițiilor experimentale, (5) determinarea măsurătorilor (variabile dependente) și a metodelor de analiză statistică. Una dintre problemele esențiale în planificarea experimentelor este stabilirea volumului eșantionului. În cazul estimărilor, acesta depinde de eroarea-standard dorită, iar în cazul testării ipotezelor statistice, volumul eșantionului este direct proporțională cu π , puterea testelor statistice. Intervin, însă, și alte condiții – tipul de eșantion, dar, mai ales, constrângerile financiare.

Din toate aceste considerente rezultă următorul algoritm de determinare a volumului eșantionului: (1) în cazul estimării – dacă se cunoaște variabilitatea caracterului în populație (σ), atunci se stabilește o anumită eroare standard ($\sigma/\sqrt{n}$) și se determină volumul eșantionului, n , în funcție de aceasta; dacă variabilitatea caracterului nu este cunoscută, se extrage un pre-eșantion pe baza căruia se estimează variabilitatea cu ajutorul dispersiei corectate de eșantionaj (s_c^2) și, utilizând această valoare, se procedează ca și în cazul anterior. (2) în cazul testării ipotezelor statistice, volumul nu poate fi determinat decât în câteva situații pe baza alegerii unui anumit nivel de încredere și cunoscând sau estimând variabilitatea caracterului în populație (populații): testele z și t , compararea proporțiilor și ANOVA. Influența constrângerilor financiare dictează următorul algoritm: mai întâi se determină volumul necesar pentru o anumită eroare standard sau putere, apoi se estimează volumul posibil în funcție de cost și se determină eroarea standard, respectiv puterea care pot fi obținute pe baza lui.

Notă. Acest document reprezintă o sinteză a cursurilor realizată pentru studenții și cadrele didactice ale Universității de Arhitectură și Urbanism „Ion Mincu”, București. Pentru utilizarea sa în oricare alte scopuri, solicitați permisiunea autorului.