BIOSTATISTICĂ
NOȚIUNI INTRODUCTIVE

I. Bioinformatica

Bioinformatica este un termen nou inventat și se referă la o ramură nouă a științei care intersectează două domenii tradiționale: biologia și informatica. Pe parcursul timpului, bioinformatica a devenit ea însăși un nou domeniu de cercetare. Prin urmare, bioinformatica se preocupă în primul rând de crearea și aplicarea metodologiilor bazate pe informație pentru analiza datelor biologice și exploatarea ulterioară a informațiilor conținute de acestea. Adoptarea pe scară largă a unei game mari de tehnologii precum microarray-urile, la fel ca și proiectele de mare anvergură precum secvențierea genomului, au avut ca rezultat crearea unor situații în care o cantitate extrem de mare de date a fost generată zilnic, o cantitate prea mare, de fapt, pentru a putea fi examinată manual și exploatată ulterior. Prin urmare, s-a impus dezvoltarea unor instrumente informatice adecvate pentru automatizarea analizei și extracției acestor seturi mari de date. În acest context, bioinformatica a oferit cele mai potrivite soluții.

În plus, sistemele biologice sunt intrinsec zgomotoase. Fundamental, sistemele biologice și procesele care le guvernează sunt neclare (greu de măsurat cu o precizie foarte bună) în natură. Prin urmare, orice date sau observații obținute de acolo vor fi în mod inevitabil la fel de neclare. Având în vedere natura inerent zgomotoasă, tehnicile matematice utilizate pentru a analiza aceste seturi de date biologice trebuie să fie metode capabile de a face față incertitudinii care este invariabil prezentă în date. Din acest punct de vedere metodele statistice sunt soluția naturală pentru astfel de probleme.

II. Biostatistica

Statistica reprezintă o colecție de metode utilizate în proiectarea unui experiment și al analizei datelor cu scopul de a putea formula concluzii cât mai exacte.

II.1. Tipuri de statistică

Statistică descriptivă: în general aceasta caracterizează un set de date elementare prin afișarea informațiilor sub formă grafică sau descrie felul în care acestea sunt distribuite și eventual care este tendința lor.

Statistică deductivă: aceasta încearcă să deducă informații cât mai precise despre un întreg (o populație) folosind date caracteristice unui eșantion.

II.2. Termeni folosiți în statistică

Populație: reprezintă un set unitar de date elementare. Populația reprezintă un termen abstract, înțelesul acestuia variază în funcție de aplicabilitatea lui.

Eșantion: reprezintă o porțiune dintr-o populație selectată pentru analiză.

Parametru: acesta este o caracteristică a întregii populații care, în majoritatea situațiilor, poate fi doar estimat. Parametrii nu trebuie confundați cu variabilele care se referă la măsurători sau atribute reale. Parametri se referă la cantități care definesc un model teoretic, ei pot fi estimați prin metode statistice descriptive sau inferențiale (**μ** media, **σ** deviația standard, **σ^2** varianța ș.a.).

Statistică: aceasta este o caracteristică a unui eșantion, presupusă a fi măsurabilă (**$\bar{x}$** media, **S** deviația standard, **S^2** varianța ș.a.).

Variabilă: informațiile înregistrate despre un eșantion (ex. pacienți) cuprind măsurători cantitative precum vârsta, greutatea, tensiunea arterială și atribute calitative precum grupa sanguină, stadiul bolii etc. Aceste valori variază între subiecții eșantionului supus analizei. În acest context,

tensiunea arterială, greutatea, grupa sanguină ș.a.m.d. reprezintă variabile. Variabilele sunt cantități care variază de la individ la individ.

Datele: reprezintă valorile individuale ale variabilelor în populație.

II.3. Nivele de măsurare

Metoda experimentală folosită depinde în mod fizic de măsurarea unor parametri. Conceptul de măsurare a fost dezvoltat în conjuncție cu conceptul de unitate de măsură (numărare). Statistica categorisește măsurătorile în concordanță cu nivelurile. Fiecare nivel se caracterizează printr-un mod în care măsurătoarea poate fi tratată matematic. Din acest punct de vedere există:

- **Date de tip nominal**: acestea nu sunt ordonate. Sunt reprezentate de denumiri sau etichete reprezentative pentru diferite categorii (ex. culorile).
- **Date de tip ordinal**: acestea sunt ordonate, dar intervalul dintre măsurători nu este esențial și nu are nici o importanță, altfel spus, nu au o unitate de măsură constantă.
- **Date de tip interval**: acestea sunt ordonate, intervalul este semnificativ și au o unitate de măsură constantă. Specific acestui tip de date este că nu au un punct precis de origine. Ex. Scalele de temperatură Celsius și Fahrenheit, valoarea zero este aleasă arbitrar. (0°C = 32°F; 0°F = -17,8°C; x°F = 32+(9/5)y°C).
- **Date de tip proporțional sau de raport**: acestea integrează cel mai înalt nivel de măsurare. Sunt ordonate. Raportul dintre măsurători ca și intervalele sunt semnificative deoarece există o origine absolută (de regulă zero). Ex. Scala de măsură a gradelor Kelvin (0°K = -273°C), lungimea, masa, viteza ș.a.m.d.

II.4. Caracterizarea variabilelor și a datelor

II.4.1. Tipuri de date

Date calitative: non-numerice (culoare, strălucire, tipuri de materiale).

Date cantitative: numerice (orice valoare numerică).

La rândul lor, datele de tip cantitativ (numeric) pot fi:
- ***discrete***: acestea au un număr finit de valori posibile. Formează o mulțime de valori discontinuă. Întotdeauna, când datele reprezintă numărători (înregistrări) atunci ele sunt discrete (ex. numărul de păsări, numărul de larve ș.a.m.d.).
- ***continue***: cu posibilități infinite și fără discontinuități (ex. mulțimea numerelor reale, volumul, lungimea, timpul, masa ș.a.m.d.).

Structura și natura datelor vor afecta în mod direct alegerea metodei de analiză. Termenul de structură se referă la faptul că datele pot fi perechi de măsurători (măsurători combinate).

II.4.2. Tipuri de variabile

Variabile calitative: non-numerice (culoare, strălucire, tipuri de materiale).

Variabile cantitative: numerice (orice valoare numerică).

La rândul lor, variabilele de tip cantitativ (numeric) pot fi:
- ***discrete***: pot lua doar anumite valori distincte, valorile intermediare fiind inexistente (ex. numărul de animale, numărul de larve ș.a.m.d.).
- ***continue***: pot lua orice valoare între anumite limite (ex. volumul, lungimea, timpul, masa viteza, cantitatea de substanță ș.a.m.d.).

Atât variabilele discrete cât și cele continue se apreciază pe o scală de tip interval sau de tip proporțional pentru că:

- au o unitate de măsură constantă și definită,

- diferența dintre categorii și observații este cuantificabilă (ex. diferență dintre 1 și 2 este egală cu diferență dintre 7 și 8 sau dintre oricare alte valori aparținând scalei de măsură.

II.5. Eșantionarea statistică

Înainte ca datele să fie analizate statistic, este foarte important de verificat dacă acestea au fost prelevate corect. Analiza statistică este ultima etapă din cadrul desfășurării unui experiment mai mult sau mai puțin complex. De aceea trebuie ținut cont de câteva aspecte:

- siguranța că dimensiunea eșantionului este suficient de mare. Obținerea unui eșantion foarte mare nu este însă o garanție de evitare a variabilității interne a parametrului studiat.
- siguranța unor date prelevate sau măsurate corect.
- siguranța că eșantionul stabilit este reprezentativ pentru populația luată în studiu.
- siguranța că toți indivizii din populație au o șansă egală de a participa la formarea eșantionului.

II.5.1. Metode de eșantionare

Eșantionarea este metoda fundamentală folosită în scopul de a deduce informații despre o întreagă populație, fără să fie necesară măsurarea fiecărui individ din populație. Utilizarea unor tehnici de eșantionare adecvată va avea o influență esențială în acuratețea rezultatelor.

- **Eșantionarea aleatorie** (randomizată). În acest caz membrii unei populații sunt aleși în așa fel încât fiecare ***să aibă o șansă egală de a fi măsurat***. Ex. Un caz aparent randomizat este selecția unor persoane după cartea de telefon. Dar, în această situație, există persoane care au mai multe telefoane sau care nu au deloc, deci, nu este perfect aleatorie pentru că nu satisface condiția fundamentală de a da o șansă egală tuturor.
- **Eșantionarea sistematică**. În acest caz este ales întotdeauna al n-lea membru din populația studiată.
- **Eșantionarea prin stratificare**. Populația este divizată în două sau mai multe straturi, fiecare astfel de sub-populație obținută este la rândul ei eșantionată aleatoriu. Ex. eșantionarea realizată pe criteriul de vârstă.
- **Eșantionarea prin grupare**. O populație este divizată în două sau mai multe grupuri, de regulă printr-un procedeu aleatoriu (randomizat), ulterior aceste grupuri sunt complet analizate (fiecare individ în parte).
- **Eșantionare prin convenienț**ă. Fiecare element este ales după propria lui dorință, acesta decide participarea sau neparticiparea la eșantion. Acest mecanism este utilizat în special în cadrul populațiilor umane.

III. Elemente de statistică descriptivă

Descrierea probelor presupune surprinderea a două trăsături majore ale acestora:

- **Tendința centrală:** reprezintă o valoare (o condiție) care tipizează datele (ex. majoritatea florilor unei specii sunt de culoare roșie, tendința este culoarea roșie).
- **Variabilitatea:** reprezintă gradul de împrăștiere a valorilor individuale în jurul tendinței centrale (ex. diametrul florilor este cuprins între 1 și 3 cm. 90% dintre flori au diametrul doar cu 0,2 cm diferit de cel al mediei).

III.1. Măsurarea tendinței centrale

În funcție de tipul variabilei (discretă sau continuă) și de scala de măsură (nominale și ordinale) există trei funcții care pot fi utilizate: modul, mediana și media aritmetică.

- **Modul**[1] reprezintă valoarea cea mai frecventă dintr-un set de date și reprezintă singura măsură a tendinței centrale aplicabilă și pentru variabile nominale (ex. din totalul de 20 de flori, 15 sunt roșii și 5 sunt roz, în acest caz cea mai frecventă culoare este roșu, deci modul este roșu). În aplicația Excel[2] funcția statistică corespunzătoare este =MODE.SNGL(v_1, v_2, v_3, ... , v_n).
- **Mediana** reprezintă valoarea din mijlocul[3] unui set de date **ordonate crescător**. Se utilizează pentru variabile ordinale (discrete și continue). Funcția statistică corespunzătoare în Excel este =MEDIAN(v_1, v_2, v_3, ..., v_n).
- **Media aritmetică** (μ, $\overline{x}$) reprezintă suma tuturor valorilor împărțită la numărul acestora. Media aritmetică se aplică atât variabilelor discrete cât și celor continue. Ca notații convenționale, μ se referă întotdeauna la media întregii populații[4] și reprezintă un parametru, iar $\overline{x}$ se referă întotdeauna la media unui eșantion dintr-o populație și este o statistică. În Excel funcția corespunzătoare este =AVERAGE(v_1, v_2, v_3, ... , v_n).

III.2. Măsurarea variabilității

Probabilitatea ca toți indivizii unei populații să fie identici din punct de vedere al unui caracter este extrem de mică (aproape imposibilă). Însă, majoritatea valorilor individuale ale unei populații se concentrează în jurul valorii tendinței centrale (media și mediana) și din ce în ce mai puține valori se găsesc la distanțe mai mari față de tendința centrală, distanța cea mai mare corespunzând valorilor extreme (minima și maxima). Cuantificarea variabilității se poate face cu ajutorul următoarelor funcții:

- **Amplitudinea** reprezintă diferența dintre cea mai mare și cea mai mică valoare dintr-un set de date. Întrucât în Excel nu există o funcție pentru calculul direct al amplitudinii se recurge la identificarea maximei și minimei cu funcțiile =MAX (v_1, v_2, ... , v_n) respectiv =MIN(v_1, v_2, ..., v_n) și se efectuează diferența dintre rezultatele returnate. Deoarece calculul amplitudinii se bazează numai pe două valori extreme, precizia acesteia este scăzută, ea nu poate furniza nici un fel de detaliu despre modul în care celelalte valori din setul de date contribuie la variabilitate.
- **Varianța sau dispersia**, (σ^2, s^2) în teoria probabilităților și în statistică, reprezintă o măsură a gradului de împrăștiere a unor valori (numere) în cadrul unui set de date. O varianță egală cu zero indică faptul că toate valorile sunt identice. Varianța este întotdeauna pozitivă. O varianță cu valoare foarte mică indică faptul că valorile sunt foarte apropiate de medie și de asemenea sunt valori foarte apropiate și între ele. O varianță foarte mare indică faptul că valorile sunt foarte împrăștiate, deci foarte îndepărtate atât unele față de altele cât și față de medie. O măsură echivalentă este rădăcina pătrată din varianță numită deviație standard, aceasta având aceeași dimensiune ca și datele, este comparabilă cu

[1] Se pronunță módul cu accentul pe litera o, a nu se confunda cu modúl, funcția matematică ce returnează valoarea absolută a unui număr. În Excel acesta este =ABS.

[2] Pentru exemplificare va fi folosită aplicația *Microsoft Excel* datorită faptului că este cel mai răspândit și mai accesibil program pentru analiză de date. Exemplele sunt compatibile numai cu versiunile 2010, 2013 și 2016 ale aplicației. Funcțiile matematice și statistice discutate au o implementare diferită față de versiunea 2007, versiune ce a reprezentat un punct de cotitură în evoluția pachetului MS Office.

[3] În cazul în care un set de date este compus dintr-un număr impar de valori, atunci există o singură valoare ce se situează la mijlocul acestora atunci când este sunt ordonate crescător sau descrescător. În cazul în care setul de date este compus dintr-un număr par de valori există două valori ce se situează la mijlocul acestora, în acest caz mediana este media aritmetică a celor două valori identificate.

[4] În marea majoritate a cazurilor experimentale, efectuate în laborator, se consideră întreaga populație. Aceste situații apar mai ales în cazul experimentelor care au loc pe organisme (animale sau plante) crescute în condiții de laborator și implicit sunt ținta unui tratament, rezultatele fiind evaluate statistic pe baza existenței unui/unor lot/loturi martor (control) și unui/unor lot/loturi probă. Aceste organisme nu sunt ținta variabilității naturale datorită izolării, ci doar a variabilității interne. Dacă se înregistrează diferențe semnificative, acesta sunt, aproape evident, datorate tratamentului.

deviațiile față de medie. Prin analogie cu notațiile folosite în cazul mediei aritmetice, în cazul variației σ^2 reprezintă varianța întregii populații și este un parametru, iar s^2 reprezintă varianța unui eșantion și este o statistică. Pentru cele două situații, aplicația Excel are două funcții distincte, =VAR.P(v_1, v_2, v_3, ... , v_n) respectiv =VAR.S(v_1, v_2, v_3, ... , v_n).

- **Deviația standard sau abaterea standard** (σ, s) reprezintă media abaterilor valorilor individuale față de media valorilor respective, cu alte cuvinte, este o măsură folosită în scopul cuantificării variației sau dispersiei unor valori din cadrului unui set de date. O deviație standard aproape de zero indică un set de date omogen, valorile fiind foarte apropiate de medie și foarte apropiate între ele. O valoare mare a deviației standard indică un eșantion eterogen, cu valori îndepărtate de medie și mai ales unele față de altele. Deviația standard a unui set de date corespunzător unei variabile aleatorii, populații statistice sau a unei distribuții probabilistice este de fapt rădăcina pătrată a variației acestuia. Din punct de vedere algebric este mai simplă, dar în practică este mai puțin robustă decât media deviației absolute[1]. O proprietate utilă a deviației standard, spre deosebire de varianță, este expresia ei în aceeași unitate de măsură ca și datele. Pe lângă exprimarea variabilității în cadrul unei populații, deviația standard este adesea folosită ca măsură a gradului de încredere (confidență) în emiterea unor concluzii statistice. Prin analogie cu notațiile folosite în cazul mediei aritmetice sau a variației, în cazul deviației standard σ reprezintă deviația standard a întregii populații și este un parametru, iar S reprezintă deviația standard a unui eșantion și este o statistică. Pentru cele două situații, aplicația Excel are două funcții distincte, =STDEV.P(v_1, v_2, v_3, ... , v_n) respectiv =STDEV.S(v_1, v_2, v_3, ... , v_n).
- **Eroarea standard a mediei** reprezintă deviația standard a mediei eșantionului folosit în scopul estimării mediei unei întregi populații. De asemenea, poate fi văzută și ca deviația standard a erorii mediei eșantionului în raport cu media reală a populației atât timp cât media eșantionului este un estimator imparțial. Calculul erorii standard a mediei se poate face pe baza următoarei formule: $ES_{\overline{\mu}} = \frac{\sigma}{\sqrt{n}}$ sau $ES_{\overline{x}} = \frac{s}{\sqrt{n}}$ unde, similar cu situațiile descrise anterior, există două abordări, eroarea standard a întregii populații sau a eșantionului. În primul caz *σ* reprezintă deviația standard a întregii populații și *n* reprezintă numărul de indivizi ce alcătuiesc populația, iar în cel de-al doilea caz, *s* este deviația standard a eșantionului și *n* reprezintă numărul de valori care-l compun. Această formulă derivă din ceea ce se știe despre varianța sumei unor variabile aleatorii independente[2]. Trebuie însă, notat că eroarea și deviația standard a unui eșantion mic tinde sistematic să subestimeze eroarea și deviația standard a populației. Eroarea standard a mediei este un estimator sistematic al erorii standard a populației. În cazul *n* = 2 subestimarea ajunge la aproximativ 25%, dar pentru un *n* = 6 subestimarea este de doar 5%. Gurland și Tripathi (1971) au oferit o soluție sub forma unei corecții (o ecuație) pentru acest efect, iar Sokal și Rohlf (1981) au oferit o ecuație pentru factorul de corecție al eșantioanelor mici (n < 20). În practică, scăderea incertitudinii în estimarea mediei cu un factor de doi necesită măsurarea a de patru ori mai multe observații în cadrul stabilirii eșantionului. Sau, scăderea erorii standard cu un factor de zece necesită măsurarea a o sută de ori mai multe observații în stabilirea eșantionului.

  În multe cazuri practice, adevărata valoare a deviației standard (*σ*) este necunoscută. Ca

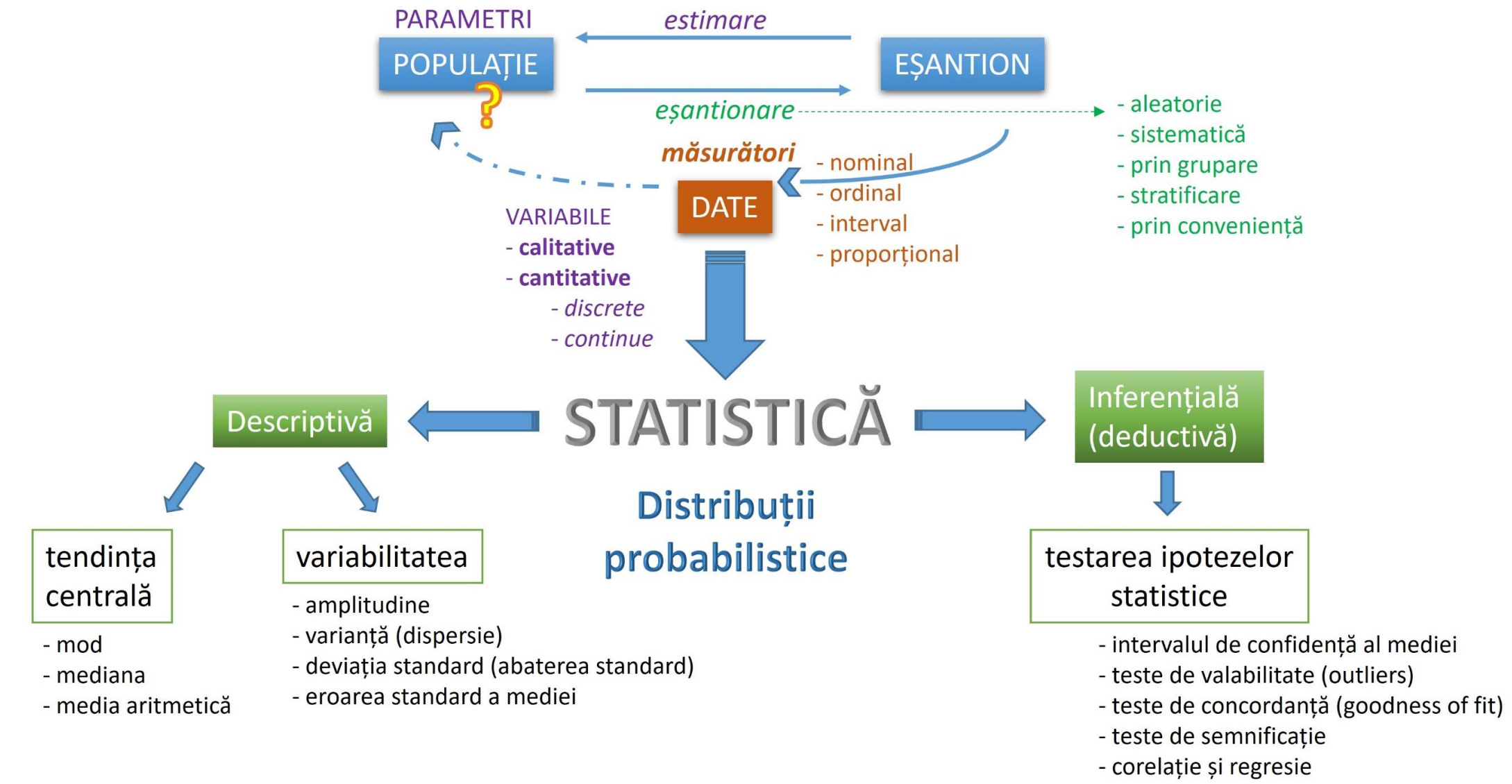
[1] Media deviației absolute a unui set de date reprezintă media deviațiilor absolute față de punctul central. Este un sumar statistic a dispersiei statistice sau a variabilității. În această descriere generală, punctul central poate fi media, mediana, modul sau rezultatul provenit dintr-o altă măsurătoare ce exprimă tendința centrală.

[2] Dacă $X_1 + X_2, \ldots, X_n$ sunt *n* observații independente dintr-o populație cu o medie μ și o deviație standard σ, atunci varianța totalului (întregului) $T = (X_1 + X_2 + \ldots + X_n)$ este $n\sigma^2$. În acest caz varianța lui T/n trebuie să fie $(1/n^2)n\sigma^2$ deci este egală cu σ^2/n, iar deviația standard a lui T/n trebuie să fie $\sigma/\sqrt{n}$. În cazul descris, este evident că T/n reprezintă media eșantionului $\overline{x}$.

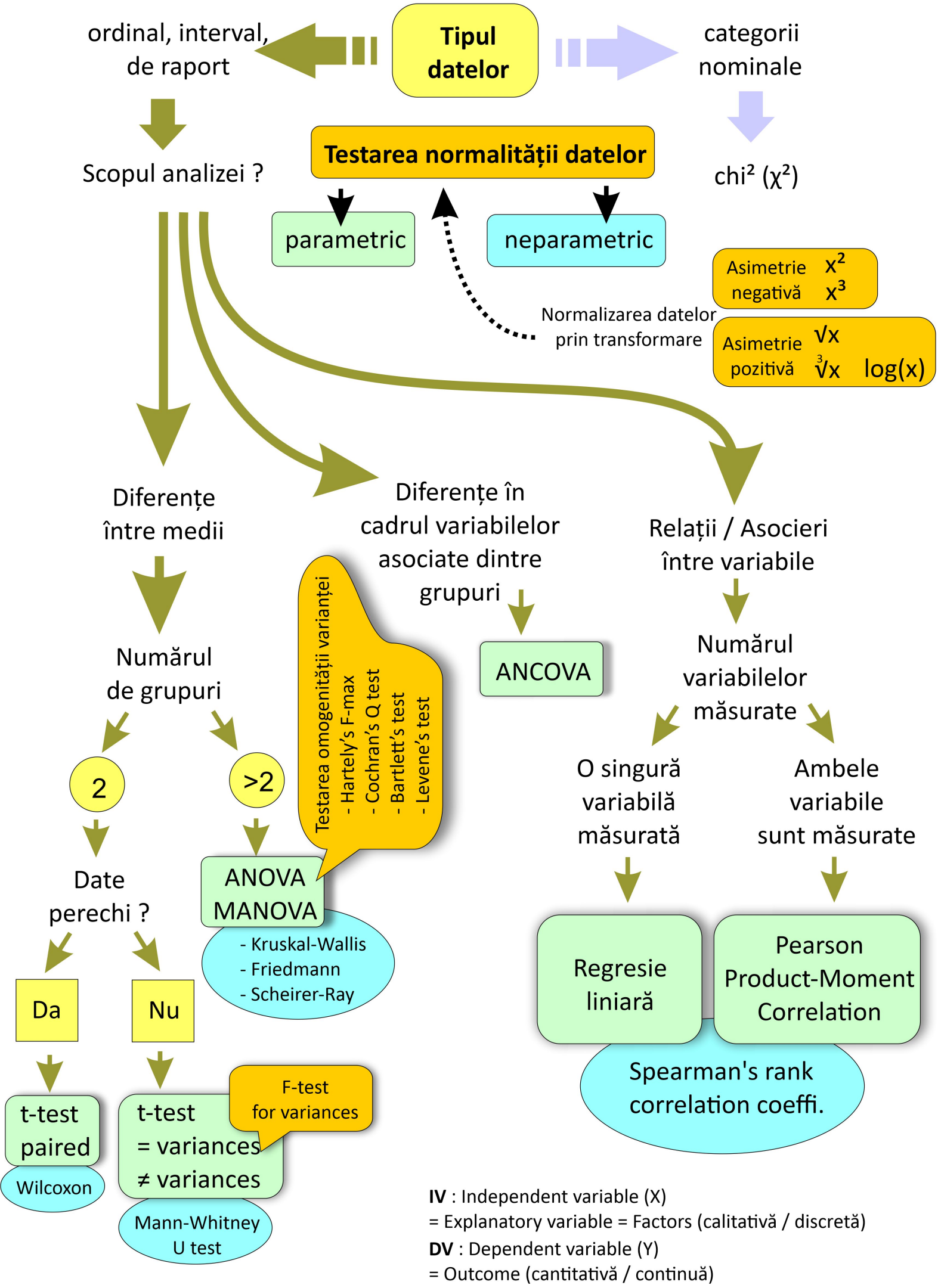
rezultat, este necesară utilizarea unei distribuții care să ia în considerare împrăștierea posibilelor deviații standard. Atunci când distribuția subiacentă este cunoscută că fiind distribuția *Gauss*, deși deviația standard (σ) este necunoscută, distribuția estimată rezultată urmează întotdeauna o distribuție *t Student*. În acest caz, eroarea standard este deviația standard a distribuției *t Student*. Distribuțiile *t* sunt oarecum diferite față de cea *Gauss* și variază în funcție de dimensiunea eșantionului. Pentru a estima eroarea standard a unei distribuții *t* este suficientă utilizarea deviației standard *s* corespunzătoare eșantionului, în detrimentul lui σ, și se poate folosi această valoare pentru calcularea intervalelor de confidență. Probabilitatea distribuțiilor *Student* poate fi o bună aproximare a distribuției *Gauss* atunci când dimensiunea eșantionului depășește 100 de valori.
În literatura științifică și tehnică, datele experimentale sunt adesea reduse la utilizarea mediei în combinație cu deviația standard sau la utilizarea mediei în combinație cu eroarea standard. Această situație conduce adesea la confuzii în raport cu interschimbabilitatea lor. Totuși, media și deviația standard reprezintă o statistică descriptivă, în timp ce eroarea standard e mediei descrie limite în cazul unui proces de eșantionare aleatorie. În ciuda micilor diferențe în ecuația deviației standard și a erorii standard, aceste mici diferențe schimbă înțelesul a ceea ce se raportează de la o descriere a variației măsurătorilor la o afirmație cu caracter probabilistic despre cum numărul eșantioanelor vor furniza o mai bună limită în estimările mediei populației în lumina teoriei limitei centrale. Cu alte cuvinte, eroarea standard a unui eșantion reprezintă o estimare a probabilității de apropiere sau îndepărtare a mediei acestuia față de media populației, în timp ce deviația standard a unui eșantion reprezintă gradul de diferențiere a valorilor individuale din cadrul eșantionului față de media acestuia. Dacă deviația standard a populației este finită, pe măsură ce crește dimensiunea eșantionului (crește numărul de valori din componența sa) atunci eroarea standard a eșantionului va tinde spre zero pentru că estimarea mediei populației se îmbunătățește, iar deviația standard a eșantionului va tinde spre valoarea deviației standard a populației.
În Excel nu există o funcție dedicată erorii standard e mediei, de aceea, ea poate fi calculată, fie prin utilizarea funcțiilor implementate în program ce compun formula descrisă mai sus, fie se poate accesa modulul de analiză de date (*Data Analysis* din cadrul secțiunii *Data*) și se apelează componenta *Descriptive Statistics* care va furniza o scurtă statistică descriptivă.

- **Intervalul de confidență sau intervalul de încredere** reprezintă un tip de interval estimativ al unui parametru populațional. Intervalul de confidență este un interval observat (este calculat pe baza observațiilor/măsurătorilor) care diferă de la eșantion la eșantion și care include în mod frecvent valoarea unui parametru de interes neobservabil, în cazul în care experimentul este repetat. Această noțiune a fost introdusă în statistică de către Jerzy Neyman în 1937. Pe baza nivelului de încredere (confidență) sau al coeficientului de încredere (confidență) se poate determina cât de frecvent apare parametrul în intervalul observat. Mai exact, înțelesul termenului *nivel de încredere* este acela că dacă intervalele de confidență sunt realizate de-a lungul a mai multor procese de analiză de date separate a unor experimente repetate (și posibil diferite), proporția unor astfel de intervale care conțin valoarea reală a parametrului se va potrivi cu nivelul de încredere dat. Întrucât limitele de confidență bilaterale (situate de o parte și de alta) formează un interval de confidență, omologii lor unilaterali sunt denumiți limite de confidență inferioară sau superioară.
  Intervalele de încredere constau dintr-o serie de valori care acționează în calitate de estimări bune ale parametrului necunoscut al populației. Totuși, în cazuri rare, se poate întâmpla ca niciuna dintre aceste valori să nu includă valoarea parametrului. Gradul de

confidență este stabilit de către cercetător și în nici un caz nu depinde de datele măsurate. În practică, intervalul de confidență cel mai des folosit este de 95%.


PARAMETRI
estimare
POPULAȚIE
EȘANTION
?
eșantionare
- aleatorie
- sistematică
- prin grupare
- stratificare
- prin convenienţă
măsurători
- nominal
- ordinal
- interval
- proporțional
DATE
VARIABILE
- calitative
- cantitative
- discrete
- continue
STATISTICĂ
Descriptivă
Inferențială (deductivă)
Distribuții probabilistice
tendința centrală
- mod
- mediana
- media aritmetică
variabilitatea
- amplitudine
- varianță (dispersie)
- deviația standard (abaterea standard)
- eroarea standard a mediei
testarea ipotezelor statistice
- intervalul de confidență al mediei
- teste de valabilitate (outliers)
- teste de concordanță (goodness of fit)
- teste de semnificație
- corelație și regresie

Alegerea unui test statistic



IV : Independent variable (X)
= Explanatory variable = Factors (calitativă / discretă)

DV : Dependent variable (Y)
= Outcome (cantitativă / continuă)