Curs 5 – Introducere în statistica inferențială

Lect. dr. Adrian Gorbănescu

În primul capitol am făcut distincția între statistica descriptivă și statistica inferențială. În capitolele 3 și 4 am văzut că statistica descriptivă ne oferă informații despre caracteristicile variabilelor măsurate: frecvențe absolute, frecvențe procentuale, media, abaterea standard etc. Deși indicatorii statistici obținuți prin statistica descriptivă ne oferă o imagine foarte amplă a setului de valori, nu putem extinde concluziile obținute și la nivelul populației din care face parte eșantionul analizat. În domeniul Psihologiei, în foarte multe dintre situațiile în care realizăm măsurători suntem interesați să generalizăm rezultatele obținute. De exemplu, un psihoterapeut poate fi interesat să studieze dacă metoda de lucru în tratarea anxietății unei persoane este eficientă și în lucrul cu ceilalți clienți care se confruntă cu problema anxietății. Astfel, pentru a-și îndeplini acest deziderat este nevoit să utilizeze statistica inferențială.

5.1 Scorurile standardizate

Statistics School își pregătește elevii pentru olimpiadă. Printre elevii ce urmează să meargă la olimpiadă se află și S.B, acesta urmând să meargă la proba de matematică. Directorul școlii nu înțelege de ce S.B. nu merge la proba de fizică.

Director: S.B. de ce nu mergi la proba de fizică? Am observat că media la această disciplină este mai mare decât la matematică.

S.B: Așa este domnule director, aveți perfectă dreptate. La matematică am nota 9,66, în timp ce nota la fizică este 10. Însă nu ați luat în calcul și abaterea standard.

Director: ...

Atunci când măsurăm o variabilă obținem un scor observat, care este cunoscut sub denumirea de scor brut. Acest scor nu ne poate spune nimic prin simpla lui valoare numerică. Notele elevilor sunt de la 1 la 10, și directorul școlii a concluzionat că nota 10 este mai mare decât 9,66. Prin urmare consideră că elevul S.B. are o performanță mai bună la fizică. Dar dacă analizând mediile tuturor elevilor observăm că media 9,66 este cea mai mare? Într-o astfel de situație este important să cunoaștem împrăștierea valorilor din distribuție. Chiar dacă ambele note sunt maxime, ele fac parte din distribuții distincte, cu împrăștieri diferite.

Imaginea 5.1 – Forma distribuției pentru notele obținute la Fizică, respectiv Matematică

În imaginea de mai sus putem observa că deși cele două valori se află în extrema dreaptă a distribuției, nota 9,66 face parte dintr-o distribuție cu o împrăștiere mai mare, fiind mai îndepărtat față de medie, comparativ cu nota 10 care face parte dintr-o distribuție omogenă, ceea ce îl face să fie mai apropiat de medie.

Deoarece cele două note sunt obținute la discipline diferite și sunt acordate de profesori diferiți trebuie găsită o soluție care să permită compararea lor. Transformarea fiecărei note în scoruri standard reprezintă soluția prin care se poate spune la ce disciplină este mai performant elevul S.B. Scorul standardizat este un scor ce exprimă cât de mult se îndepărtează un scor brut

față de medie, distanța fiind exprimată în abateri standard (Yin, 1994). Cu alte cuvinte, scorul standard ne spune la câte abateri standard este un scor față de medie. Cele mai utilizate scoruri standard sunt **scorul z** și **scorul T**. *Pentru a afla scorul z al unei valori din distribuție este necesar să cunoaștem media și abaterea standard pentru acea distribuție*. Scorul z se calculează după următoarea formulă:

$$z = \frac{X - m}{s}$$

(formula 5.1)

- **X** – reprezintă orice valoare din distribuție.
- **m** – este media distribuției
- **s** – este abaterea standard a distribuției.

Să ne imaginăm că pentru elevii din *Statistics School* se înregistrează la matematică o medie de 7,15 și o abatere standard de 1,5, în timp ce la fizică media tuturor elevilor este 8 și abaterea standard este 1,4. Mai departe vom transforma notele obținute de elevul S.B. în scoruri standard z.

$$z_{matematică} = \frac{9{,}66 - 7{,}15}{1{,}5} \rightarrow z_{matematică} = \frac{2{,}51}{1{,}5} \rightarrow z_{matematică} = 1{,}67$$

$$z_{fizică} = \frac{10 - 8}{1{,}4} \rightarrow z_{fizică} = \frac{2}{1{,}4} \rightarrow z_{fizică} = 1{,}42$$

În conformitate cu rezultatele obținute mai sus putem observa că nota de la matematică se află la 1,67 abateri standard peste medie, în timp ce nota de la fizică se află la 1,42 abateri standard peste medie. În concluzie, S.B. are o performanță mai ridicată la matematică și a luat decizia corectă atunci când a ales să se înscrie la olimpiadă pentru această disciplină.

Atunci când un scor brut este mai mic decât media, scorul standard obținut va avea o valoarea negativă, ceea ce se traduce prin faptul că acel scor este mai mic decât media. Astfel, înțelegem că semnul „+" în fața unui *scor z* indică faptul că acesta este superior mediei, în timp ce semnul „-" din fața *scorului z* semnalează un scor mai mic decât media distribuției.

Proprietățile scorului z sunt:

- media este întotdeauna egală cu 0.
- Abaterea standard este întotdeauna egală cu 1.

Exemplu

Un psiholog realizează evaluarea periodică a doi angajați din cadrul companiei Statistics Icecream. Printre probele pe care le aplică se află și câte un test de inteligență. Angajatul **A** obține la testul de inteligență scorul 80, în condițiile în care media pentru acel test este 73, iar abaterea standard 9,5. Angajatul **B** obține la un testul de inteligență scorul 76, în condițiile în care media este 70, iar abaterea standard este 8. Care dintre cei doi angajați a obținut o performanță mai bună?

Având în vedere că celor doi angajați li s-au administrat teste diferite, pentru a realiza o comparație între cele două performanțe trebuie să transformăm scorurile obținute la test în scoruri standard z.

$$z_A = \frac{80-73}{9,5} \rightarrow z_A = \frac{7}{9,5} \rightarrow z_A = 0,73$$

$$z_B = \frac{76-70}{8} \rightarrow z_A = \frac{6}{8} \rightarrow z_A = 0,75$$

Rezultatele obținute ne arată că angajatul **B** a obținut o performanță mai bună, deoarece distanța dintre scorul lui și medie este de 0,75 abateri standard, în timp ce angajatul **A** se îndepărtează față de medie cu 0,73 abateri standard.

Un alt scor standardizat, foarte des utilizat în Psihologie, este **scorul T**. Spre deosebire de *scorul z*, care implică valori negative și zecimale, *scorul T* are valori pozitive și întregi (zecimalele pot fi ignorate), care sunt mult mai ușor de înțeles și interpretat. Media unei distribuții de scoruri T este 50, iar abaterea standard este 10. Pentru a obține un scor T se aplică formula:

$$T = 50 + 10 * \frac{X-m}{s} \rightarrow T = 50 + 10 * z, \text{ unde:}$$

(formula 5.2)

- **X** – este orice scor din distribuție
- **m** – reprezintă media distribuției
- **s** – reprezintă abaterea standard

Întorcându-ne la elevul S.B., putem calcula performanțele la matematică, respectiv fizică și în scoruri T. Astfel:

$$T_{matematică} = 50 + 10 * \frac{9,66-7,15}{1,5} \rightarrow T_{matematică} = 50 + 10 * 1,67 \rightarrow T_{matematică} = 50 + 16,7 \rightarrow T_{matematică} = 66,7$$

$$T_{fizică} = 50 + 10 * \frac{10-8}{1,4} \rightarrow T_{fizică} = 50 + 10 * 1,42 \rightarrow T_{matematică} = 50 + 14,2 \rightarrow T_{matematică} = 64,2$$

Scorul standard T obținut la matematică 66,7 poate fi rotunjit la 67, în timp ce scorul T de la fizică 64,2 poate fi interpretat ca fiind egal cu 64. Scorurile T mai mari decât 50 sunt superioare mediei, iar cele mai mici de 50 sunt inferioare mediei.

Un cercetător este interesat să transforme în scoruri standardizate scorurile obținute pentru variabila Nevrotism. Eșantionul este format din 1781 de participanți, obținând scoruri între 10 și 50 (m =27.04; s = 8.01).

Pentru a transforma o variabilă în scoruri z trebuie să cunoaștem media și abaterea standard a variabilei respective. Așa cum am observat în capitolul anterior, acești indicatori pot fi obținuți cu ajutorul funcțiilor **mean()**, respectiv **sd()**. Să presupunem că suntem interesați să transformăm în scoruri z variabila *Nevrotism*. Folosind funcțiile menționate mai sus putem observa că Nevrotism are media egală cu 27.04, iar abaterea standard egală cu 8.01.

Pentru a obține variabila care conține scorurile z ale nevrotismului vom pargurge pașii de mai jos:

1. Vom vrea variabila care să primească scorurile z ale variabilei *Nevrotism*. Vom denumi această variabilă sugestiv *zNevrotism*.
2. Vom scrie formula de calcul a scorurilor z.

zNevrotism <- (bdcap5$Nevrotism – 27.04)/8.01, unde:

- zNevrotism este variabila care primește scorurile z ale variabilei *Nevrotism*;
- bdcap5 este numele bazei de date;
- Nevrotism indică variabila pe care dorim să o transformăm (ea cuprinde scorurile brute pe care le vom transforma în scoruri z).

Pentru a transforma scorurile variabilei *Nevrotism* obținute de participanți în **scoruri T** vom folosi aceeași procedură, cu mențiunea că numele variabilei care va conține scorurile T se va numi TNevrotism și ne vom folosi de formula de calcul a scorurilor standardizate T. Astfel, linia de cod care va permite calcularea scorurilor standardizate T pentru variabila *Nevrotism* este:

TNevrotism <- 50+10*zNevrotism, unde:

- TNevrotism reprezintă variabila care primește scorurile T ale variabilei *Nevrotism*;
- 50 și 10 reprezintă media, respectiv abaterea standard a scoruilor standardizate T;
- zNevrotism este variabila care conține scorurile standardizate z;
- bdcap5 este numele bazei de date;
- Nevrotism indică variabila pe care dorim să o transformăm.

Nu este suficientă calcularea variabilelor *zNevrotism* și *TNevrotism* pentru ca acestea să fie adăugate în baza de date. Inserarea lor în baza de date necesită utilizarea funcției **data.frame()** precum în linia de mai jos:

bdcap5v2 <- data.frame(bdcap5, zNevrotism)
View(bdcap5v2)

- bdcap5v2 reprezintă numele pe care îl va avea baza de date după ce adăugam variabila zNevrotism;
- data.frame este funcția care creează baza de date;
- bdcap5 se referă la numele bazei de date în care adăugăm noua variabila;
- zNevrotism reprezintă variabila care conține scorurile z;
- View(bdcap5z) afișează noua bază date.

```
zNevrotism <- (bdcap5$Nevrotism - 27.04)/8.01
TNevrotism <- 50+10*zNevrotism
bdcap5v2 <- data.frame(bdcap5, zNevrotism, TNevrotism)
View(bdcap5v2)
```

	Nevrotism	Extraversie	Deschidere	Agreabilitate	Conștiinciozitate	zNevrotism	TNevrotism
1	19	43	43	42	47	-1.003745318	39.9625
2	30	30	20	17	20	0.369538077	53.6953
3	24	37	32	37	34	-0.379525593	46.2047
4	18	40	34	45	43	-1.128589263	38.7141
5	30	31	37	32	30	0.369538077	53.6953
6	27	40	44	37	32	-0.004993758	49.9500
7	29	27	31	36	34	0.244694132	52.4469
8	28	30	30	30	30	0.119850187	51.1985

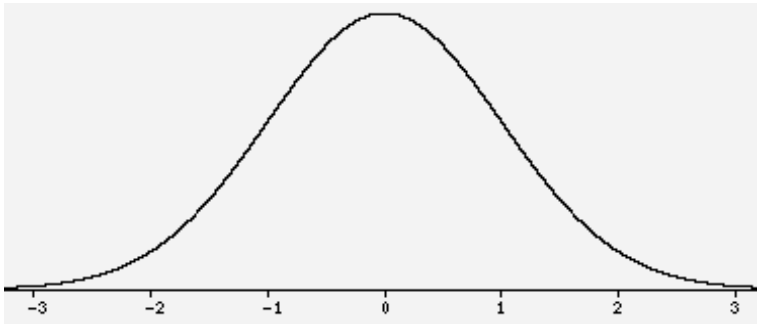
5.2 Distribuția normală (Curba normală)

În capitolele 3 și 4 am prezentat modalitatea prin care poate fi descrisă o distribuție (analiza de frecvențe și indicatorii sintetici ai distribuției). Scorurile cu care am lucrat în exemplele discutate sunt obținute în urma unui proces de măsurare. Distribuțiile cu scoruri observate prin procesele de măsurare sunt cunoscute sub denumirea de **distribuții empirice**. Pe lângă distribuțiile empirice există și **distribuții teoretice** care sunt bazate pe formule matematice.

Una dintre distribuțiile teoretice care își dovedește utilitatea în foarte multe domenii este **distribuția normală**. Această distribuție a fost promovată de Carl Friedrich Gauss cu scopul de a explica erorile aleatorii în observațiile astronomice (Stewart, 1977). La sfârșitul secolului al 19-lea Karl Pearson a dat distribuției normale denumirea de **curbă normală**, deoarece s-a observat că o largă varietate de măsurători au tendința de a se organiza într-o distribuție sub formă de clopot. Unul din promotorii curbei normale a fost Adolphe Quetlet care a arătat că multe măsurători realizate în științele sociale și în biologie sunt distribuite conform curbei normale (Porter, 1986). Caracteristicile curbei normale sunt valabile pentru un număr foarte mare de măsurători, tinzând spre infinit.

Curba normală (curba lui Gauss) are următoarele proprietăți:

- are formă de clopot – cea mai mare parte a valorilor se concentrează în zona medie.
- este perfect simetrică, de fiecare parte a mediei fiind 50% din valorile distribuției.
- poate lua valori oricât de mari sau oricât de mici; marginile curbei nu ating axa Ox.



Imaginea 5.2 – Curba normală

Deși distribuția normală poate avea valori oricât de mari sau oricât de mici, prin convenție, pe curbă sunt prezentate valorile cuprinse între -3 și 3 abateri standard.

Curba normală exprimată în scoruri z se numește curbă standardizată și are aceleași proprietăți precizate mai sus, cu mențiunea că media este 0 și abaterea standard este 1. Curba normală standardizată are următoarele caracteristici:

- între medie (z = 0) și o abatere standard (z = 1) se află aproximativ 34% din scorurile distribuției normale.
- Între -1z și +1z se află 68% din scorurile distribuției.
- Între -2z și +2z se află 95% din scorurile distribuției.
- Între -3z și +3z se află 99% din scorurile distribuției.

Imaginea 5.3 – Curba normală standardizată

Având în vedere aceste caracteristici ale curbei normale standardizate putem răspunde la următoarele întrebări:

- Ce procent de scoruri se află între o anumită valoare din distribuție și medie?
- Ce procent de scoruri se află între două valori ale distribuției?
- Ce procent de scoruri se află între două scoruri standardizate?

Răspunsurile la aceste întrebări se află utilizând un tabel special care cuprinde probabilitățile de sub curba normală **z** (**Anexa 1**). Pentru fiecare scor z (coloana A) în tabel este prezentată probabilitatea de a avea valori cuprinse între medie și scorul respectiv (coloana B), dar și probabilitatea de a avea valori mai mari decât acel scor (coloana C). De exemplu, probabilitatea de a avea scoruri între medie și scorul z = 1,67 este **0,4525**. Cu alte cuvinte, procentul valorilor cuprins între medie și z = 1,67 este **45,25%**. De asemenea, tabelul ne indică faptul că probabilitatea de a avea scoruri mai mari decât z = 1,67 este **0,0475**. În concluzie, **4,75%** din scoruri sunt mai mari decât z = 1,67.

Această curbă este foarte importantă în domeniul Psihologiei, nu numai pentru analiza statistică, ci și pentru alte tipuri de activități, cum ar fi cele de evaluare.

Exemple

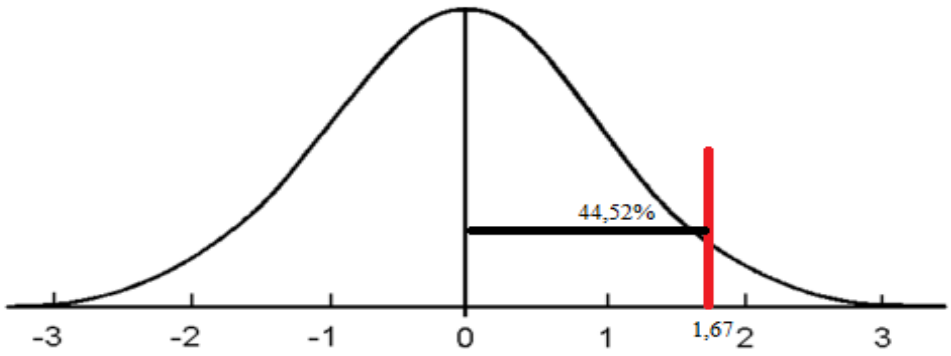
În exemplul anterior am putut observa că profesorul de statistică a dat parțial studenților săi înregistrându-se o medie de 6,59 și abaterea standard 2,12.

1. Ce procent de studenți se află între medie (6,59) și nota 10?

Pentru a răspunde la această întrebare trebuie să transformăm nota 10 în scor z, folosind formula de calcul:

$$z = \frac{X - m}{s} \rightarrow z = \frac{10 - 6{,}59}{2{,}12} \rightarrow z = \frac{3{,}41}{2{,}12} \rightarrow z = 1{,}60$$

În tabelul din **Anexa 1** vom citi probabilitatea cuprinsă între medie și scorul z = 1,60. Astfel, vom obține probabilitatea 0,4452. Această valoare o transformăm în procent și vom concluziona că procentul de studenți cu note între medie și 10 este de 44,52%.

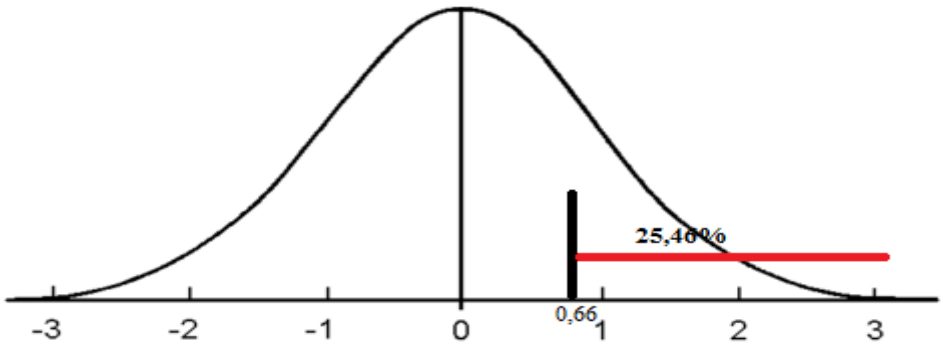


2. Care este procentul studenților care au note peste 8?

Vom calcula scorul z pentru nota 8.

$$z = \frac{X - m}{s} \rightarrow z = \frac{8 - 6{,}59}{2{,}12} \rightarrow z = \frac{1{,}41}{2{,}12} \rightarrow z = 0{,}66$$

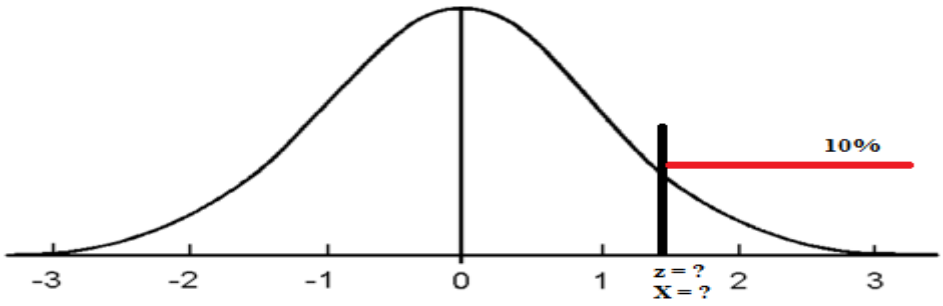
În tabelul din Anexa 1, pentru z = 0,66, vom citi valoarea de pe coloana care ne indică probabilitatea de a avea valori mai mari decât scorul z și anume **0,2546**. În concluzie, **25,46%** dintre studenți au note mai mari sau egale cu 8.



Sunt tabele care prezintă doar probabilitatea de a avea valori între medie și scorul z. Pentru o astfel de situație, citim probabilitatea scorurilor dintre medie și scorul z = 0,66 (0,2454). Știm că probabilitatea de a avea valori mai mari decât medie este 0,50. Astfel, probabilitatea de a avea valori mai mari decât z = 0,66 este:

$$0{,}50 - 0{,}2454 = 0{,}2546.$$

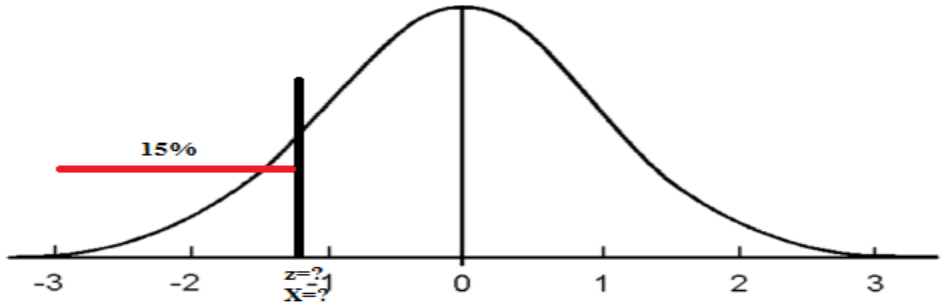
3. Ce notă trebuie să obțină un student pentru a fi printre cei mai buni 10%?



Pentru această situație știm că între medie și nota studentului trebuie să se afle 40% dintre valori. Astfel, el trebuie să obțină o notă, care transformată în scor z, să ne indice o arie a probabilității între medie și scor de 0,40. Vom căuta în Anexa 1 primul scor z care indică probabilitatea de a avea valori între medie și z de 0,40. Scorul cu această probabilitate este **z = 1,29**. Folosind formula de calcul a scorului z vom afla nota (X) pe care trebuie să o obțină studentul.

$$z = \frac{X - m}{s} \rightarrow X - m = z * s \rightarrow X = z * s + m \rightarrow X = 1{,}29 * 2{,}12 + 6{,}59 \rightarrow X = 9{,}32$$

4. Ce scor trebuie să obțină un student pentru a fi printre cei mai slabi 15%?



Din datele problemei știm faptul că aria de sub scorul z al notei pe care ar trebui să o obțină studentul este de 15%, exprimată în probabilități 0,15. În Anexa 1, pe coloana ce indică aria de sub un scor vom căuta probabilitatea 0,15. Scorul z corespunzător acestei probabilități este 1,03. Deoarece performanța studentului este sub medie, scorul standardizat al studentului este **z = -1,03**. Folosind formula de calcul a scorului z vom afla nota (X) pe care trebuie să o obțină studentul.

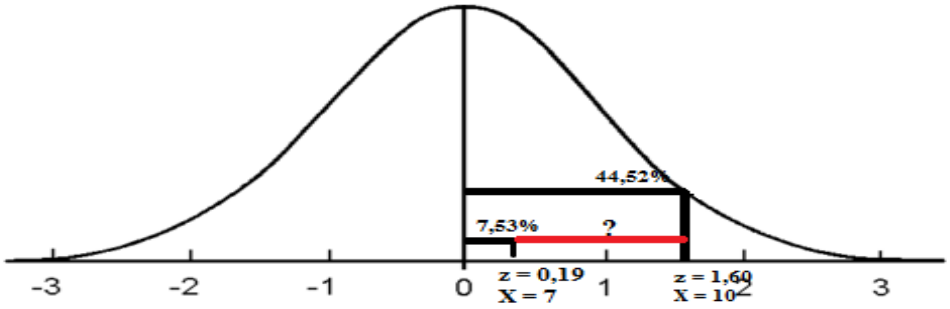
$$z = \frac{X - m}{s} \rightarrow X - m = z * s \rightarrow X = z * s + m \rightarrow X = -1{,}03 * 2{,}12 + 6{,}59 \rightarrow X = 4{,}40$$

5. Ce procent de studenți au obținute note între 7 și 10?

Acest exercițiu ne cere să aflăm ce procent de studenți au obținut note între 7 și 10. Prin procedurile de calcul existente nu putem răspunde direct la această întrebare. În schimb putem afla distanța procentuală de la medie până la nota 7, respectiv 10 și apoi să calculăm procentul dintre cele două note.

$$z_7 = \frac{X - m}{s} \rightarrow z_7 = \frac{7 - 6{,}59}{2{,}12} \rightarrow z_7 = \frac{0{,}41}{2{,}12} \rightarrow z_7 = 0{,}19$$
$$z_{10} = \frac{X - m}{s} \rightarrow z_{10} = \frac{10 - 6{,}59}{2{,}12} \rightarrow z_{10} = \frac{3{,}41}{2{,}12} \rightarrow z_{10} = 1{,}60$$

Conform tabelului, probabilitatea de a avea un scor între medie și z = 0,19 este 0,0753. Cu alte cuvinte, procentul studenților care au note cuprinse între medie și 7 este 7,53%. Din exemplul 1 știm că procentul studenților cu note cuprinse între medie și nota 10 este 44,52%.



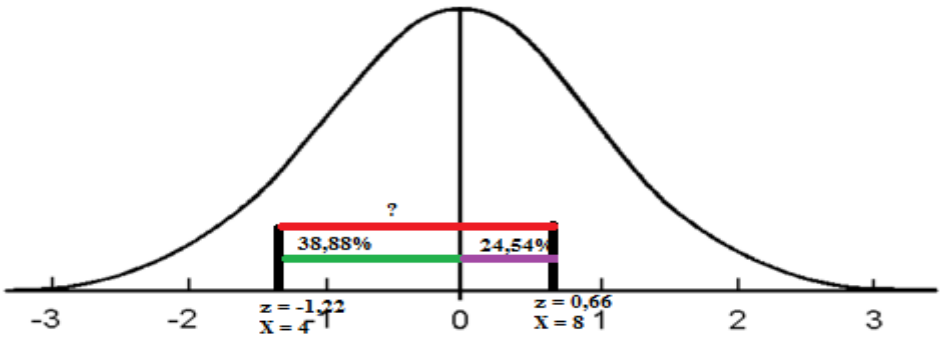
Pentru a afla procentul studenților cu note între 7 și 10 vom calcula diferența dintre 44,52% și 7,53%. Concluzionăm că 36,99% dintre studenți au note 7 și 10.

6. Ce procent de studenți au obținut note între 4 și 8?

Precum în exercițiul anterior, vom calcula scorurile z pentru cele două note. Scorul z pentru nota 8 a fost calculat și în exercițiul 2. Astfel, z_8 = 0,66 și procentul de studenți cuprinși între medie și nota 8 este de 24,54%

$$z_4 = \frac{X - m}{s} \rightarrow z_4 = \frac{4 - 6{,}59}{2{,}12} \rightarrow z_4 = \frac{-2{,}59}{2{,}12} \rightarrow z_4 = -1{,}22$$

Probabilitatea de a avea scoruri între medie și z_4 = -1,22 este 0,3888. Astfel, între medie și nota 4 avem 38,88% din studenți.



Pentru a afla procentul studenților cu note între 4 și 10 vom calcula suma dintre 38,88% și 24,54%. Concluzionăm că 63,42% dintre studenți au note 4 și 8.

Din exercițiile rezolvate putem observa că distanța dintre media distribuției și un anumit scor este exprimată în termeni de probabilitate. Pentru a transforma probabilitatea în procent am apelat la înmulțirea cu 100. Prin urmare, pentru a transforma un procent în probabilitate vom face împărțirea valorii procentuale la 100. Atunci când aruncăm o monedă probabilitatea

de apariție a fiecărei fațete este de 0,50. Altfel spus, fiecare fațetă a monedei are o șansă de apariție de 50%. În exercițiul 6 am calculat că 63,42% dintre studenți au note cuprinse între 4 și 8. Acest rezultat poate fi exprimat și prin faptul că există o probabilitate de 0,63 de a avea studenți cu note între 4 și 8.

5.3 Eșantionarea aleatorie

Eșantionarea aleatorie este o metodă de a obține un eșantion care are o probabilitate foarte mare de a fi reprezentativ pentru populație. Faptul că un eșantion este aleatoriu înseamnă că în selecția participanților nu a intervenit hazardul sau întâmplarea (Wilkinson, 1999). În cele mai multe situații de cercetare prin **eșantion aleatoriu** ne referim la un eșantion de volum (mărimea eșantionului) **N** care are o probabilitate de a fi selectat egală cu a oricărui alt eșantion de aceeași mărime din populația de referință. Pentru a obține un eșantion aleatoriu trebuie parcurse următoarele etape (Spatz, 1992):

- Definiți (stabiliți) populația.
- Identificați numărul de persoane din populație.
- Selectați participanții astfel încât fiecare să aibă o probabilitate egală de a fi selectat (1/N).

Mai departe vom demonstra aceste etape utilizând exemplul cu cei 271 de studenți prezenți la parțial la statistică, de unde vom selecta un eșantion de 50 de studenți.

O primă metodă de a alege aleatoriu eșantionul dorit este aceea de a scrie bilete cu numele tuturor studenților care au dat parțialul și de a le adăuga într-o urnă. După amestecarea celor 271 de bilete vom extrage 50, acestea reprezentând studenții selectați pentru a face parte din eșantion. Această metodă se dovedește a fi utilă în situațiile în care populația este de dimensiuni mici. Atunci când populația are dimensiuni foarte mari metoda aceasta este incomodă sau chiar imposibil de aplicat.

O altă metodă de a obține eșantioane aleatorii constă în utilizarea tabelului cu numere aleatorii (**Anexa 2**). Pentru a utiliza acest tabel, fiecărui membru al populației trebuie să îi oferim un cod de identificare. Vom alege din tabel, la întâmplare, o linie și o coloană. De exemplu, alegem linia 35 și coloana 20-24. La intersecția acestora se află numărul 10061. Deoarece dorim să selectăm un eșantion de volum N = 50 din cei 271 de studenți, vom alege doar primele trei unități ale valorilor indicate de tabel. Astfel, primul student selectat pentru a face parte din eșantion este cel cu numărul 100. Mai departe, fie vom coborî, fie vom urca pe coloana 20-24. Dacă alegem să coborâm pe această coloană, următoarea valoare citită va fi la intersecția liniei 36 cu coloana 20-24. Este vorba de valoarea 10683. Al doilea student selectat este cel cu numărul 106. Folosind același algoritm următoarea valoare tabelară este 43233, deci vom selecta studentul cu numărul 432. Dacă valoarea tabelară ne indică un număr care nu face parte din populație, cum este și cazul nostru, vom merge la următoarea valoare. Dacă am terminat valorile de pe o coloană și încă nu am terminat de selectat eșantionul, vom relua procedura prin selecția unei alte linii și coloane. Apoi vom coborî sau vom urca pe acea coloană. Dacă tabelul ne va indica un caz care se repetă, vom ignora acel caz a doua oară.

Exercițiu

Folosind tabelul de la anexa 2 selectați un eșantion aleatoriu de N = 15 dintr-o populație de 330 de studenți.

5.4 Distribuția de eșantionare

Dintr-o anumită populație pot fi extrase o „infinitate" de eșantioane de volum N. Pentru eșantioanele obținute putem calcula anumiți indicatori statistici: medie, abatere standard, dispersie etc. Să ne imaginăm că pentru fiecare eșantion posibil extras din populație calculăm media. Aceste medii formează, la rândul lor, o distribuție care se numește **distribuție mediei de eșantionare (distribuție de eșantionare)**. Distribuția de eșantionare va avea, la rândul ei, o medie care poartă numele de **medie de eșantionare** și se calculează după următoarea formulă:

$$\mu = \frac{m_1 + m_2 + m_3 + \cdots + m_k}{k}$$

(formula 5.3)

În formula prezentată mai sus:

- **μ** reprezintă media populației.
- **m** reprezintă media fiecărui eșantion extras.
- **k** reprezintă numărul de eșantioane extrase din populație.

Să presupunem că extragem 15 eșantioane de volum N = 50 din populația de studenți care au participat la parțial. Știm că media populației (μ) este 6,59, iar abaterea standard (σ) este 2,12. Pentru fiecare eșantion vom calcula media rezultatelor de la parțial. Tabelul de mai jos prezintă mediile celor 15 eșantioane.

$m_1 = 6{,}67$	$m_2 = 6{,}49$	$m_3 = 6{,}59$	$m_4 = 6{,}67$	$m_5 = 6{,}47$
$m_6 = 6{,}62$	$m_7 = 6{,}34$	$m_8 = 6{,}90$	$m_9 = 6{,}87$	$m_{10} = 6{,}83$
$m_{11} = 6{,}53$	$m_{12} = 6{,}43$	$m_{13} = 7{,}03$	$m_{14} = 6{,}76$	$m_{15} = 5{,}93$

Pe baza mediilor celor 15 eșantioane extrase din populația de studenți prezenți la test putem realiza o distribuție. Media acestei distribuții (media mediilor) este egală cu **6,60**, valoarea care se apropie foarte mult de media populației (**μ = 6,59**). Pe măsură ce numărul eșantioanelor extrase va crește, tinzând spre „infinit", media mediilor se va apropia de media populației. După cum se poate observa, media fiecărui eșantion oscilează în jurul mediei populației.

În cadrul distribuției de eșantionare, ca în orice distribuție, putem de vorbi de împrăștierea valorilor. Cu cât eșantioanele extrase din populație sunt mai mari, cu atât media lor se apropie de media populației, ceea ce face ca abaterea standard a distribuției de eșantionare să fie mai mică. Valorile se apropie de medie, deci nu există împrăștierea valorilor. Abaterea standard a distribuției de eșantionare ne indică cât de mult se abat mediile față de media populației, motiv pentru care mai poartă numele de **eroare standard a mediei** și se calculează astfel:

$$s_m = \frac{\sigma}{\sqrt{N}}$$

(formula 5.4)

- $\mathbf{s_m}$ este eroarea standard a mediei.
- **σ** este abaterea standard la nivelul populației.
- **N** reprezintă volumul eșantionului.

În cazul exemplului nostru, eroarea standard a mediei este:

$$s_m = \frac{\sigma}{\sqrt{N}} \rightarrow s_m = \frac{2{,}12}{\sqrt{271}} \rightarrow s_m = \frac{2{,}12}{16{,}46} \rightarrow s_m = 0{,}32$$

De cele mai multe ori abaterea standard la nivelul populației nu este cunoscută, motiv pentru care eroarea standard a mediei va fi estimată cu ajutorul abaterii standard a eșantionului. Având în vedere că eroarea standard a mediei ne spune cât de mult se îndepărtează media unui eșantion de media populației din care a fost extras, vom înțelege că cu cât eroarea standard a mediei este mai mică cu atât media eșantionului se apropie mai mult de media populației. De asemenea, pe baza formulei de calcul, vom înțelege că pe măsură ce volumul eșantionului crește are loc scăderea erorii standard a mediei.

5.5 Teorema limitei centrale

În cercetările pe care le desfășurăm nu reușim niciodată să investigăm toate eșantioanele posibile dintr-o populație. Din acest motiv selectăm un eșantion pe care îl supunem analizelor și, în final, concluzionăm care sunt caracteristicile populației din care a fost extras. În spatele generalizării unor rezultate obținute la nivelul unui eșantion asupra populației se află foarte multă matematică și acesta nu este obiectivul principal al lucrării de față. Totuși, există o „autoritate" în domeniu care susține acest tip de inferențe – **teorema limitei centrale**. Această teoremă importantă susține că: *pentru orice populație de scoruri, indiferent de formă, distribuția de eșantionare a mediei se va supune legilor curbei normale cu cât volumul eșantioanelor este mai mare. În plus, cu cât numărul eșantioanelor este mai mare, cu atât media distribuției de eșantionare se apropie de media populației.*

Afirmațiile teoremei limitei centrale sunt valide doar atunci când eșantioanele au același volum și sunt selectate în mod aleatoriu.

5.6 Intervalul de încredere al mediei populației

Intervalul de încredere (lb. engleză **confidence interval**), prezentat în literatura de specialitate ca **CI**, este utilizat cu scopul de a estima un parametru al populației pe baza unui indicator al eșantionului. De exemplu, putem fi interesați să estimăm care este salariul mediu pentru locuitorii din București folosind un eșantion de 1000 de angajați. Deoarece rezultatele eșantionului pot varia (vezi distribuția de eșantionare) este necesar să stabilim un interval în care poate varia parametrul pe care dorim să îl estimăm. Variabilitatea pe care ne-o asumăm se numește marjă de eroare. Media obținută la nivel de eșantion plus sau minus marja de eroare ne vor indica intervalul de valori între care se va situa media populației - cu alte cuvinte ne prezintă intervalul de încredere.

Marja de eroare nu reprezintă probabilitatea de a comite erori în estimarea parametrului. Ea ne indică variabilitatea generată de șansă în cadrul eșantionului. Deoarece eșantionul nu cuprinde toată populația, ne așteptăm ca media acestuia să cuprindă o doză de hazard. În același timp, media eșantionului s-ar putea modifica dacă alegem un alt eșantion. Aceste rezultate sunt exacte numai într-un anumit interval, calculat ținând cont de marja de eroare. Această marjă de eroare se numește eroare standard (lb. engleză, **standard error**).

Formula de calcul pentru intervalul de încredere al mediei populației este:

$$CI = m \pm z * s_m$$

(formula 5.5)

- m – reprezintă media eșantionului.
- z – este valoarea critică tabelară.

- s_m – indică eroarea standard a mediei.

Eroarea standard a mediei se calculează astfel:

$$s_m = \frac{\sigma}{\sqrt{N}}$$

(formula 5.6)

- σ – este abaterea standard a populației.
- N – indică volumul eșantionului.

Atunci când facem o estimare folosind intervalul de încredere ne dorim ca aceasta să fie cât mai precisă. Cu cât nivelul de eroare este mai mic, cu atât intervalul de încredere este mai restrâns, iar estimarea este mai exactă. Cum ne dăm seama dacă un interval de încredere este suficient de precis? Acesta este un aspect la care trebuie să ne gândim înainte de a începe colectarea datelor. Iată trei factori care pot avea un efect asupra intervalului de încredere (Rumsey, 2010):

- Nivelul de încredere.
- Mărimea eșantionului.
- Abaterea standard la nivelul populației. Atunci când nu avem acces la abaterea standard a populației o putem estima folosind abaterea standard a eșantionului.

Trebuie să nu pierdem din vedere faptul că valoarea indicatorului statistic (rezultatul obținut pe eșantion) nu are legătură cu mărimea intervalului de încredere. Acesta reprezintă valoarea din mijlocul intervalului de încredere, nu mărimea lui.

Nivelul de încredere al intervalului corespunde procentului care ne arată de câte ori putem să obținem o estimare corectă dacă selectăm eșantioane aleatorii. De obicei, se selectează un nivel de încredere de 95% sau 99%. Se pot utiliza și alte niveluri de încredere (de exemplu, 99,9%), dar nu mai mici de 95%. Nivelul de încredere determină valoarea critică tabelară (z_{critic}), adică de câte ori vom scădea și vom aduna eroarea standard la media eșantionului. Astfel, înțelegem că în funcție de nivelul de încredere asumat valoarea lui z se modifică.

Tabel 5.1 – Valorile z pentru nivelul de încredere selectat

Nivel de încredere (%)	**α**	**Valoarea lui z**
95	0,05	1,96
99	0,01	2,58
99,9	0,001	3,27

Nivelul de încredere poate fi scris sub forma (1-α), unde α reprezintă procentul în care intervalul de încredere nu este corect. Astfel, dacă alegem un nivel de încredere de 95%, α este 0,05. Valoarea lui α ne indică și probabilitatea de comite o eroare de tip I (vezi capitolul 6.5).

Exemplu de calcul

Psihologul unui penitenciar este interesat să studieze stabilitatea emoțională a persoanelor private de libertate și dorește să estimeze, la un nivel de încredere de 95%, care este media stabilității emoționale pentru întreaga populație de deținuți din penitenciarul în care lucrează. Astfel, selectează un eșantion aleatoriu (N = 182), aplică chestionarul de evaluare a stabilității emoționale și obține m = 65,10 și s = 12,69.

Deoarece nivelul de încredere este 95%, valoarea lui z este 1,96. Întrucât nu avem acces la abaterea standard a populației (σ), ne folosim de abaterea standard a eșantionului (s).

$$CI = m \pm z * s_m$$

$$s_m = \frac{\sigma}{\sqrt{N}} \rightarrow s_m = \frac{12{,}69}{\sqrt{182}} \rightarrow s_m = 0{,}94$$

$$\lim\inf CI95\% = m - z * s_m \rightarrow \lim\inf CI95\% = 65{,}10 - 1{,}96 * 0{,}94$$
$$\rightarrow \lim\inf CI95\% = 63{,}26$$
$$\lim\sup CI95\% = m + z * s_m \rightarrow \lim\inf CI95\% = 65,10 + 1{,}96 * 0{,}94$$
$$\rightarrow \lim\inf CI95\% = 66{,}94$$

Cu o precizie de 95% media stabilității emoționale a populației de deținuți din penitenciarul în care lucrează psihologul este cuprinsă între 63,26 și 66,94.