3.1. DISTRIBUȚII PROBABILISTICE

- Definiția distribuției probabilistice
- Variabilele aleatoare discrete și continue
- Media aritmetică și varianța unei distribuții probabilistice discrete
- Media aritmetică și varianța unei distribuții probabilistice continue
- Distribuții probabilistice discrete: binomială
- Distribuții probabilistice continue: normală
- Distribuția normală standard (redusă)

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O5-1: să identifice caracteristicile unei distribuții de probabilitate

O5-2: să facă distincția între o variabilă aleatoare discretă și una continuă

O5-3: să calculeze media, varianța și deviația standard în cazul unei distribuții discrete

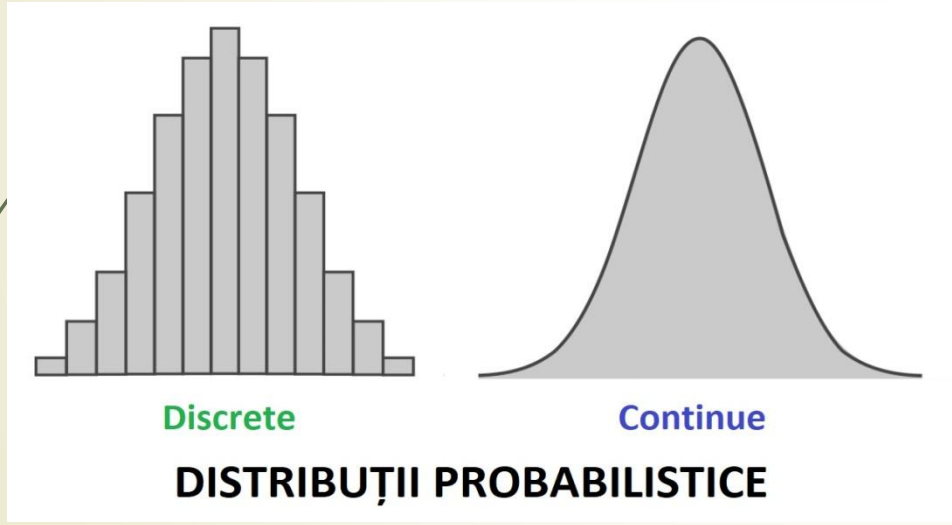
O5-4: să explice caracteristicile unei anumite distribuții discrete și să le aplice în vederea calculării probabilităților

O5-5: să descrie proprietățile distribuțiilor probabilistice continue (normală) și să le utilizeze pentru determinarea probabilităților

O5-6: să descrie proprietățile distribuției normale standard (redusă) și să le utilizeze în calculul probabilităților

INTRODUCERE

În cadrul capitolului anterior au fost introduse conceptele de bază ale teoriei probabilităților alături de metodele principale de determinare a probabilității unui anumit eveniment. În continuare, vom utiliza aceste noțiuni, explorând noi modalități de calcul ale probabilității unui eveniment în condiții mai complexe. Astfel, vom vedea imediat faptul că relația dintre valorile pe care le poate lua o variabilă statistică aleatoare și probabilitățile apariției acestora pot fi sintetizate sub forma unei „DISTRIBUȚII PROBABILISTICE”.



Cunoașterea distribuției probabilistice a unei variabile aleatoare reprezintă pentru un cercetător un instrument important în centralizarea și descrierea unui set de date statistice în vederea formulării unor concluzii pertinente referitoare la o anumită populație (colectivitate generală) pe baza datelor obținute la nivelul unui eșantion reprezentativ extras din aceasta.

O distribuție probabilistică este asemănătoare cu o distribuție a frecvențelor relative, cu precizarea însă că, în loc să descrie evenimente care s-au derulat anterior, este utilizată în vederea obținerii unor estimări a șansei de producere în viitor a unor evenimente.

DISTRIBUȚIA PROBABILISTICĂ

Reprezintă o listă (tabel, grafic) care cuprinde întreaga plajă de rezultate posibile ale unui experiment împreună cu toate nivelurile probabilităților asociate fiecărui rezultat în parte.

Proprietățile unei distribuții probabilistice

1) **Probabilitatea realizării fiecărui eveniment este cuprinsă în intervalul [0,1];**
2) **Evenimentele sunt incompatibile (se exclud reciproc)**
3) **Evenimentele formează un sistem complet de evenimente, astfel încât suma probabilităților lor de realizare este 1**

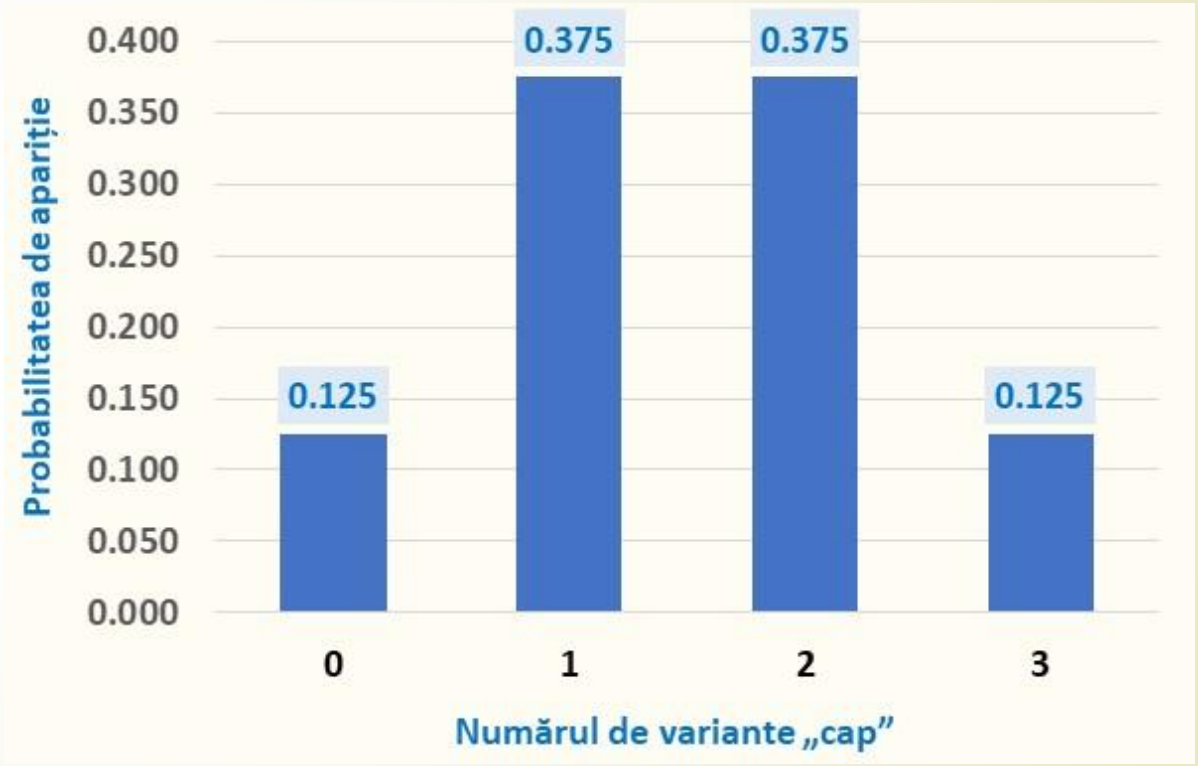
Exemplu: suntem interesați în realizarea unui experiment care presupune aflarea numărului de variante „cap” care pot apare în situația aruncării de 3 ori a unei monede. Aplicăm regula multiplicării și avem 2 x 2 x 2 = 8 rezultate posibile în derularea experimentului, fiecare cu un număr diferit de apariții. Centralizând, obținem doar 4 rezultate diferite (evenimente): să avem 0, 1 ,2 sau 3 variante „cap” în urma aruncării monedei.

Rezultate posibile	Aruncarea monedei			Numărul de var. „cap”
	Prima	A doua	A treia	
1	P	P	P	0
2	P	P	C	1
3	P	C	P	1
4	P	C	C	2
5	C	P	P	1
6	C	P	C	2
7	C	C	P	2
8	C	C	C	3

În continuare realizăm sub formă tabelară distribuția probabilistică pentru cele 4 evenimente (0,1,2,3 variante „cap”) rezultate în urma aruncării de trei ori a unei monede. Deoarece unul dintre aceste rezultate trebuie să aibă loc, suma probabilităților tuturor evenimentelor posibile este egală cu 1.

Numărul de variante „cap” x	Probabilitatea de apariție P(x)
0	1 / 8 = 0,125
1	3 / 8 = 0,375
2	3 / 8 = 0,375
3	1 / 8 = 0,125
Total	8 / 8 = 1,000

Informațiile privind cele 4 evenimente și probabilitățile aferente acestora pot fi înfățișate și sub formă grafică:



VARIABILE ALEATOARE

În orice experiment care implică șansa, rezultatele apar în mod întâmplător. De exemplu, aruncarea unui zar este un eveniment, în care oricare dintre cele șase fețe poate să apară. O parte dintre experimente pot avea rezultatele evidențiate numeric fiind exprimate sub forma unor variabile cantitative (greutatea, înălțimea, glicemia, numărul de copii, salariile lunare) iar altă parte pot genera rezultate sub forma unor tipuri calitative (categorii) diferite fiind exprimate sub forma unor variabile calitative (grupa de sânge, genul, culoarea ochilor, nivelurile de calificare, mediul social).

O variabilă aleatoare reprezintă o variabilă măsurată sau observată ca rezultat al unui experiment, care în mod cu totul întâmplător poate lua diferite valori.

Există două tipuri de variabile aleatoare: discrete și continue. Variabilele discrete pot lua doar un număr limitat de valori izolate (numărul de copii dintr-o familie, notele studenților la examenul de statistică). Variabilele continue însă pot lua absolut orice valoare în cadrul unui anumit interval (temperatura corpului uman, înălțimea, greutatea).

1. VARIABILE ALEATOARE DISCRETE

Considerăm cazul unei variabile aleatoare X. În urma realizării unui experiment se obțin valorile $x_1, x_2, x_3, \dots x_n$ cu probabilitățile de apariție $P(X = x_1) = p_1, P(X = x_2) = p_2, P(X = x_3) = p_3, \dots, P(X = x_n) = p_n$.

Repartiția unei variabile aleatoare discrete reprezintă enumerarea tuturor valorilor (variantelor) posibile împreună cu probabilitățile de obținere corespunzătoare (tabel sau sub grafic).

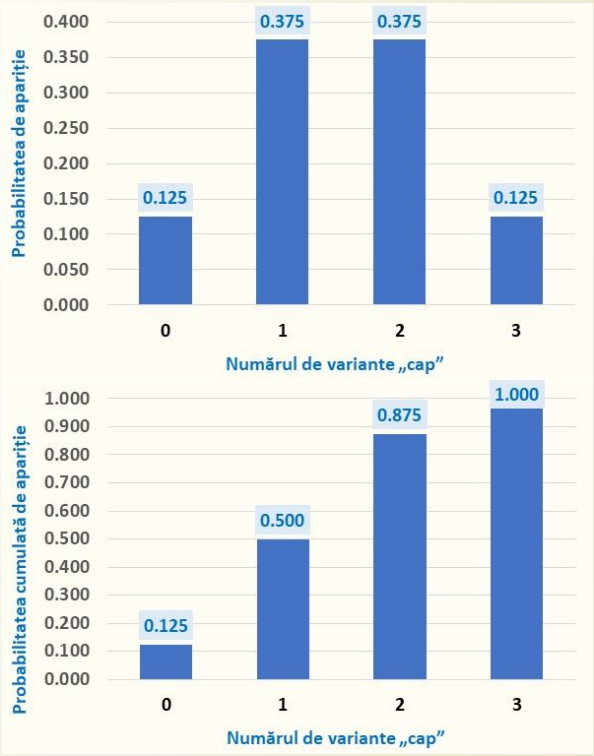
Legea de repartiție (FUNCȚIA DE PROBABILITATE) a unei variabile aleatoare X este:

$$f(x_i) = P(X = x_i) = P(x_i) = p_i$$

În unele situații este necesară determinarea probabilității ca valoarea variabilei aleatoare să fie mai mică decât un anumit nivel prag x_i

Vom determina astfel, FUNCȚIA DE REPARTIȚIE:

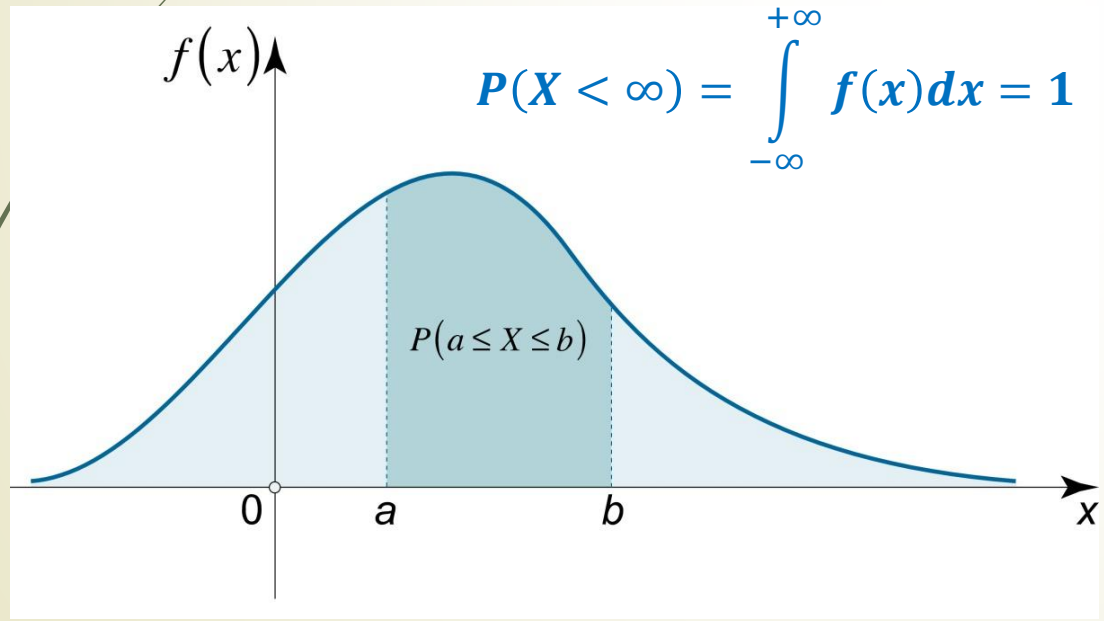
$$F(x_i) = P(X \leq x_i) = \sum_{i=1}^{k} P(X = x_i) = \sum_{i=1}^{k} p_i$$



2. VARIABILE ALEATOARE CONTINUE

Variabilele continue pot lua un număr „infinit” de valori în cadrul unui anumit interval. Din acest considerent, evidențierea sub formă tabelară a tuturor valorilor pe care le poate avea o variabilă aleatoare continuă nu este posibilă. Astfel, se impune utilizarea probabilității evenimentului $X < a$, operațiile fiind efectuate asupra unor intervale de valori și nu asupra unor valori fixe.

Funcția de probabilitate, specifică variabilelor discrete, este înlocuită în acest caz cu FUNCȚIA DENSITATE DE PROBABILITATE. Suprafața totală cuprinsă între curbă și axa orizontală este egală cu 1, reprezentând probabilitatea ca x să se afle în intervalul (-∞, +∞).



FUNCȚIA DE REPARTIȚIE se determină după formula:

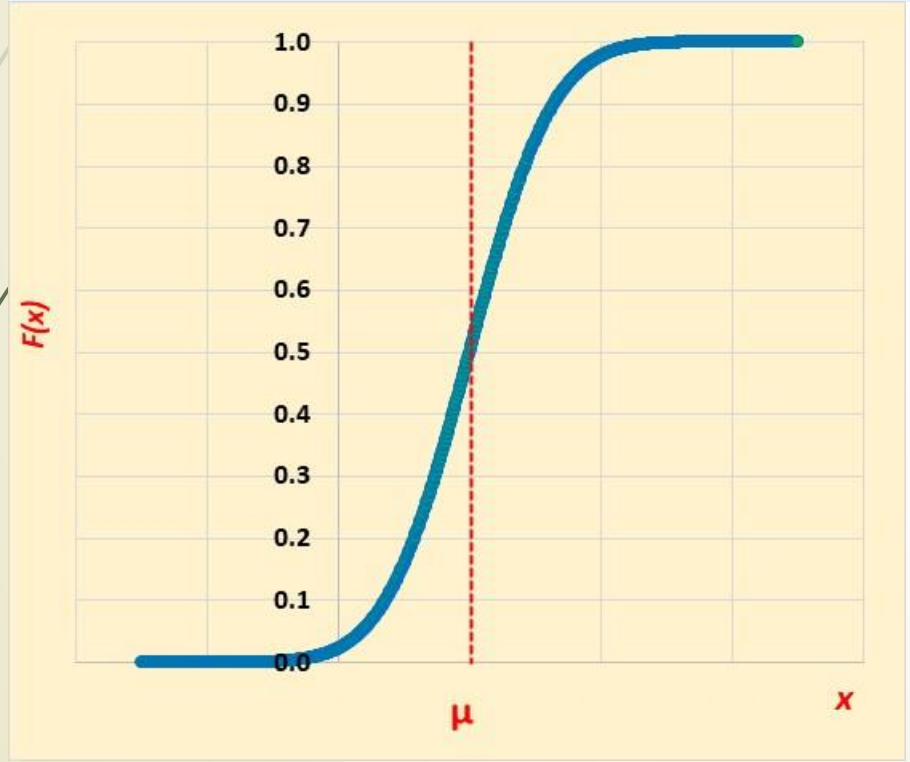
$$F(x) = P(X < a) = \int_{-\infty}^{a} f(x)dx$$

O curbă a densității de probabilitate este curba care are același aspect cu curba către care tinde poligonul frecvențelor relative, atunci când numărul de valori dintr-o serie tinde la infinit, iar lungimea fiecărei clase tinde la 0.

Probabilitatea ca variabila aleatoare să ia valori în intervalul (a, b) este egală cu suprafața cuprinsă între cele 2 valori. Astfel:

$$P(a,b) = P(a < x < b) = P\big((x < b) \cap (x > a)\big) = P(x < b) - P(x < a)$$

$$P(a,b) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$



Avem două drepte asimptote $F(x) = 0$ și $F(x) = 1$

Observăm că sunt respectate relațiile:

$$F(-\infty) = 0$$

$$F(+\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

Dacă $x_1 \leq x_2$, atunci $F(x_1) \leq F(x_2)$,
Prin urmare funcția de repartiție este crescătoare.

MEDIA ARITMETICĂ ȘI VARIANȚA UNEI DISTRIBUȚII PROBABILISTICE DISCRETE

1. MEDIA ARITMETICĂ A UNEI DISTRIBUȚII DISCRETE

Media este o valoare tipică, reprezentativă a tendinței centrale în cazul unei distribuții probabilistice, uneori făcându-se referire la ea ca fiind „valoarea așteptată”.

Dacă avem o distribuție statistică discretă, în care fiecare nivel individual al variabilei x_i apare cu frecvența relativă f_i, atunci media aritmetică se calculează după formula clasică:

$$\mu_{x_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{1} = \sum_{i=1}^{n} x_i \cdot f_i$$

Conform legii numerelor mari, dacă un experiment se repetă, în aceleași condiții, de un număr suficient de mare de ori, atunci *frecvența relativă a unui eveniment* f_i prezintă o anumită stabilitate, variind în jurul probabilității de realizare a acestuia p_i

Rezultă imediat faptul că, MEDIA ARITMETICĂ A UNEI DISTRIBUȚII PROBABILISTICE DISCRETE se poate determina ca o medie aritmetică ponderată, în care toate valorile posibile ale unei variabile aleatoare x_i sunt multiplicate cu probabilitățile corespunzătoare de apariție ale lor $P(x_i)$:

$$\mu = \sum_{i=1}^{n} [x_i \cdot P(x_i)] = \sum_{i=1}^{n} x_i \cdot p_i$$

2. VARIANȚA UNEI DISTRIBUȚII DISCRETE

Media este o valoare reprezentativă spre care tind toate valorile individuale ale unei distribuții probabilistice, cunoscută din acest motiv sub numele de „speranța matematică" a unei variabile. Cu toate acestea, media nu poate cuantifica nivelul împrăștierii (variației) dintr-o distribuție statistică, acesta fiind evidențiat de varianță.

Formula de calcul a VARIANȚEI ÎN CAZUL UNEI DISTRIBUȚII PROBABILISTICE DISCRETE:

$$\sigma^2 = \sum_{i=1}^{n} \left[(x_i - \mu)^2 \cdot P(x_i) \right] = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot p_i$$

Ex: Un agent de vânzări înregistrează de obicei cel mai mare număr de automobile vândute în ziua de vineri. El și-a creat următoarea distribuție probabilistică a numărului de mașini pe care se așteaptă să le vândă într-o zi de vineri:

Nr. mașini vândute x_i	Prob. de vânzare $P(x_i)$	$x_i \cdot P(x_i)$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 \cdot P(x_i)$
0	0.1	0.0	0 – 2.1	4.41	0.441
1	0.2	0.2	1 – 2.1	1.21	0.242
2	0.3	0.6	2 – 2.1	0.01	0.003
3	0.3	0.9	3 – 2.1	0.81	0.243
4	0.1	0.4	4 – 2.1	3.61	0.361
Total	1.0	2.1	-	-	1.290

$\mu = 2,1 \quad \sigma^2 = 1,29 \quad \sigma = 1,136$

MEDIA ARITMETICĂ ȘI VARIANȚA UNEI DISTRIBUȚII PROBABILISTICE CONTINUE

1. MEDIA ARITMETICĂ A UNEI DISTRIBUȚII CONTINUE

Este un indicator reprezentativ și mai poartă numele de speranță matematică, valoare așteptată sau medie teoretică. În cazul variabilelor continue, media aritmetică se calculează după formula:

$$\mu = M[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

2. VARIANȚA UNEI DISTRIBUȚII CONTINUE

Dispersia sau varianța, în calitate de indicator sintetic al împrăștierii variantelor unei variabile continue în jurul mediei aritmetice se determină astfel:

$$\sigma^2 = D[X] = M\left[(x-\mu)^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)dx$$

DISTRIBUȚII PROBABILISTICE DISCRETE

1. DISTRIBUȚIA BINOMIALĂ

Distribuția binomială (Bernoulli) reprezintă una dintre cele mai des întâlnite distribuții probabilistice discrete (reprezintă modelul extragerii bilelor din urnă cu revenire – compoziția urnei rămâne nemodificată pe tot parcursul experimentului; astfel fiecare unitate aleasă și inclusă în eșantion este returnată în populația de referință, înaintea selectării următoarei unități).

MODELUL PROBABILISTIC BINOMIAL presupune îndeplinirea următoarelor condiții:

- **Fiecare experiment are doar două rezultate posibile, complementare, denumite generic succes A și insucces $\overline{A}$ (de ex. rezultatul unui proces de vânzare prin telefon: un client poate cumpăra sau nu respectivul produs, dar cele 2 rezultate nu pot coexista);**
- **Probabilitățile evenimentului A și al opusului său $\overline{A}$ rămân constante pe parcursul derulării experimentelor (de ex. în cadrul unui test experiment cu 20 de întrebări, care au sugerate 4 răspunsuri posibile, din care doar unul este corect, avem 20 de probe, iar probabilitatea de a selecta în mod aleator răspunsul corect pentru fiecare întrebare este mereu aceeași: 0,25);**
- **Fiecare probă din cadrul experimentului este independentă de toate celelalte, rezultatul ei neafectând în niciun fel restul probelor (răspunsul corect sau greșit la o întrebare nu influențează în niciun fel răspunsurile la toate celelalte întrebări din test).**

Notăm probabilitatea unui succes $P(A) = p$ și probabilitatea unui insucces $P(\overline{A}) = q$

Astfel, probabilitatea ca efectuând de *n* ori experiența, să se obțină de *k* ori un succes (evenimentul A) și implicit de *n* - *k* ori un insucces (evenimentul $\overline{A}$) este:

$$P(X = x_k) = C_n^k \cdot p^k \cdot q^{n-k}$$

Evenimentele A și $\overline{A}$ sunt independente și conform regulii speciale de înmulțire a probabilităților vom avea:

$$P = P(A \cap A \cap A \cap \cdots \cap A \cap \overline{A} \cap \overline{A} \cap \overline{A} \cap \cdots \cap \overline{A}) =$$
$$= P(A) \cdot P(A) \cdot P(A) \cdot \cdots \cdot P(A) \cdot P(\overline{A}) \cdot P(\overline{A}) \cdot \cdots \cdot P(\overline{A}) =$$
$$= p \cdot p \cdot p \cdot \cdots \cdot p \cdot q \cdot q \cdot q \cdot \cdots \cdot q = p^k \cdot q^{n-k}$$

Deoarece trebuie să ținem cont de toate grupurile de k elemente de tip *A* (succese) pot fi alese din totalul celor n, indiferent de ordinea (poziția) acestora, avem un număr de C_n^k variante posibile.

Media distribuției binomiale este $\mu = n \cdot p$

Varianța (dispersia) distribuției binomiale: $\sigma^2 = n \cdot p \cdot q = n \cdot p \cdot (1 - p)$

Distribuția binomială este biparametrică, adică este caracterizată de doi parametri *n* și *p*

DISTRIBUȚII PROBABILISTICE CONTINUE

DISTRIBUȚIA NORMALĂ

Distribuția normală, cunoscută și sub denumirea de distribuție Gausiană, este probabil cea mai importantă distribuție probabilistică utilizată pentru a descrie o variabilă aleatoare continuă, având multiple aplicații practice (greutatea, înălțimea, scorurile IQ etc.). Pentru a utiliza distribuția probabilistică normală, variabila aleatoare analizată trebuie să fie continuă.

Formula de calcul în cazul distribuției normale a fost publicată pentru prima dată de matematicianul francez Abraham de Moivre în anul 1733. În domeniul teoriei probabilităților și statisticii cea mai semnificativă contribuție a lui de Moivre a constat în aproximarea distribuției binomiale prin distribuția normală în cazul unui număr mare de încercări, cu alte cuvinte forma distribuției binomiale discrete converge către curba continuă a distribuției normale, atunci când numărul de probe ale experimentului *n* tinde spre infinit.

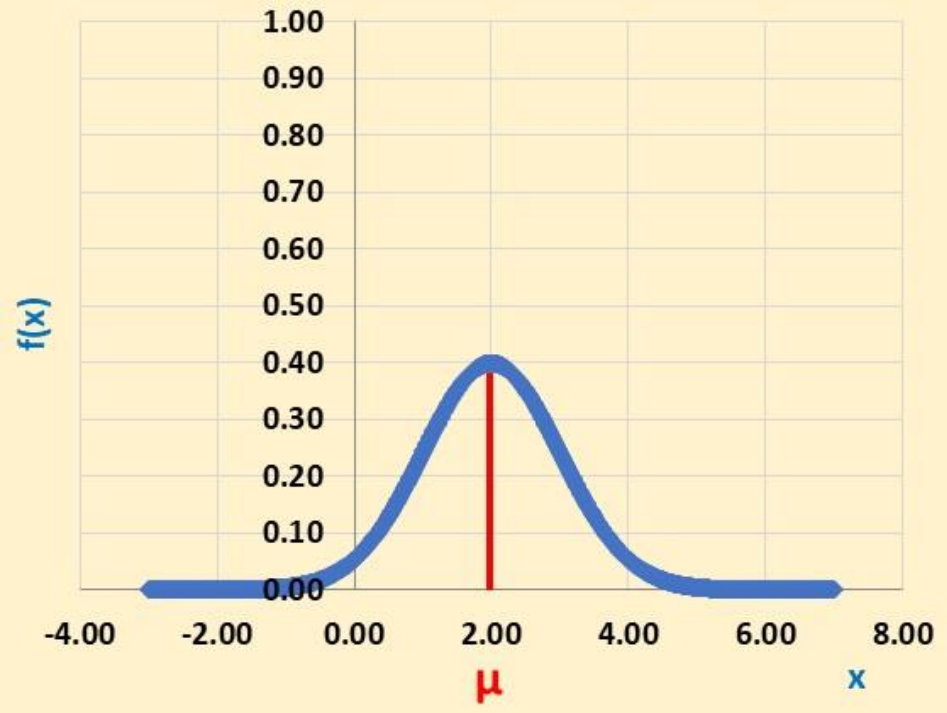
Matematicianul și fizicianul german Johann Carl Friedrich Gauss a introdus distribuția normală, în forma în care o utilizăm și în prezent, prezentând în mod explicit modalitatea în care probabilitatea poate fi reprezentată grafic printr-o curbă în formă de clopot. Această curbă atinge punctul de maxim în jurul valorii medii (așteptate) și scade treptat, pe măsură ce ne îndepărtăm de nivelul mediu, în direcția celor două limite (-∞ și +∞).

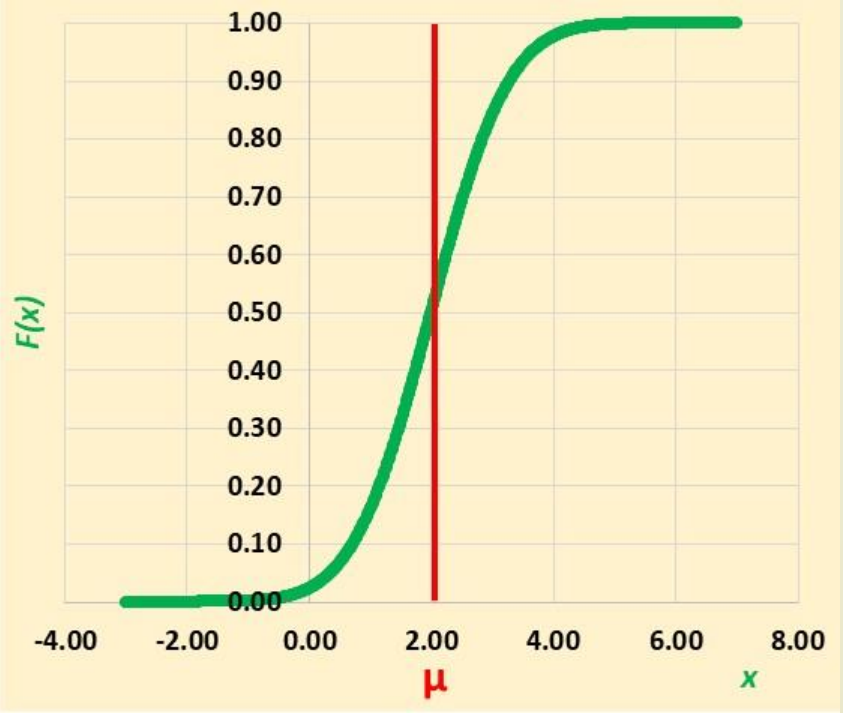
Funcția Gauss-Laplace este caracterizată de doi parametri: media (μ), și dispersia (σ^2): $N(\mu, \sigma^2)$

Expresia densității de probabilitate este dată de formula: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}}$$

unde X este o variabilă continuă: $-\infty < x < +\infty$

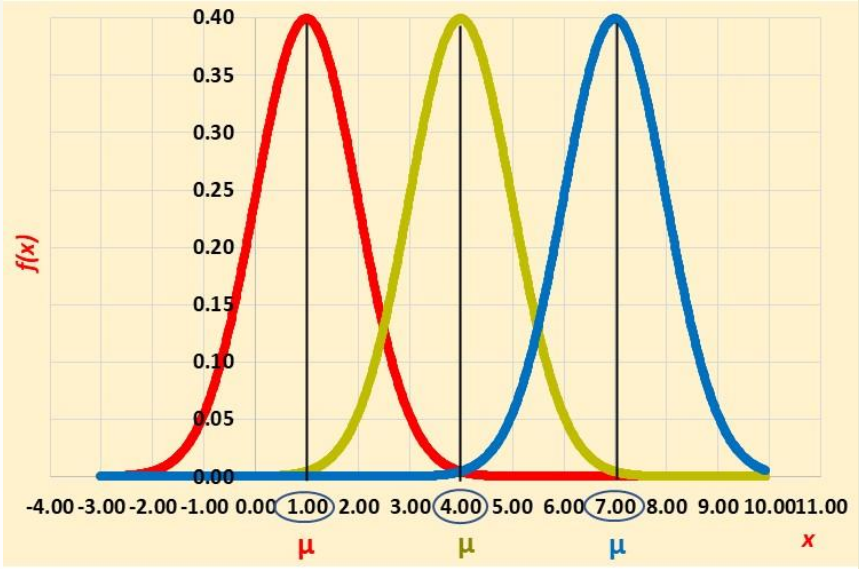
Funcția de repartiție se calculează astfel: $$F(x) = \int_{-\infty}^{x} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}} dx$$

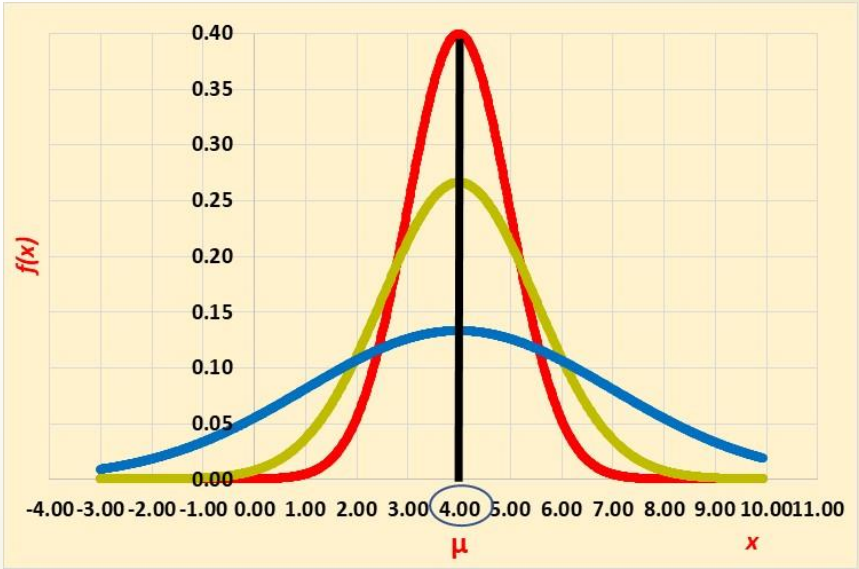




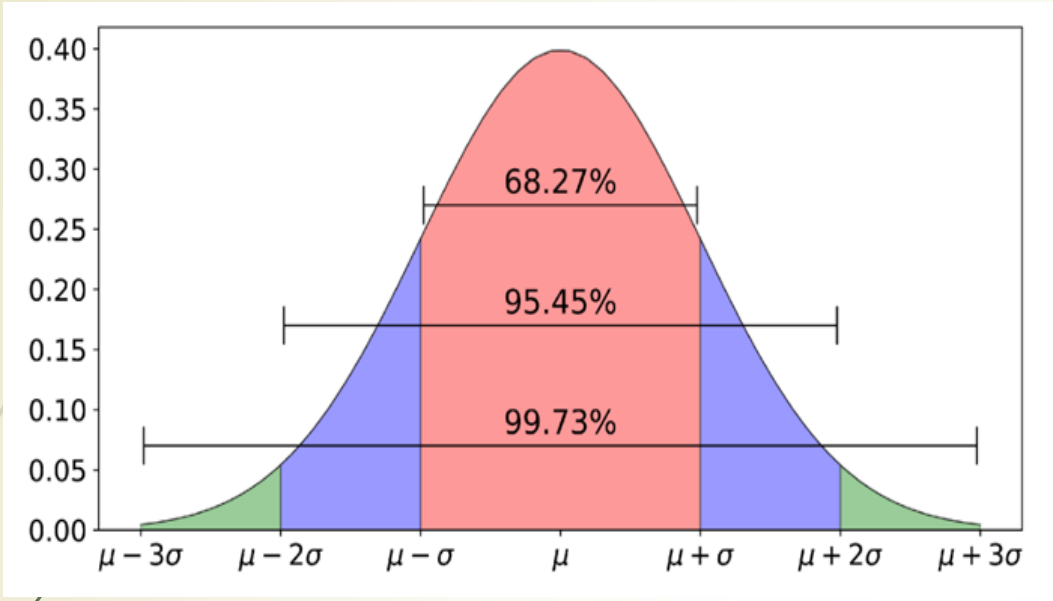
PROPRIETĂȚILE REPARTIȚIEI NORMALE

- admite un singur punct de maxim ($x = \mu$), fiind unimodală
- este simetrică in raport cu dreapta x = μ
- densitatea de probabilitate are formă de clopot → „clopotul lui Gauss”
- media aritmetică, mediana și modul coincid
- suprafața totală cuprinsă între curba si axa absciselor este 1, reprezentând probabilitatea evenimentului sigur
- punctele de pe abscisă, de valoare μ-σ, μ+σ, sunt puncte de inflexiune
- in punctele $x \to \infty$ si $x \to -\infty$, funcția $f(x)$ tinde la zero
- coeficienții de asimetrie și boltire pentru o repartiție normală au valoarea zero
- modificarea mediei conduce la o translare a curbei de-a lungul axei absciselor
- curba densității de probabilitate este cu atât mai boltită cu cât valoarea dispersiei este mai mică

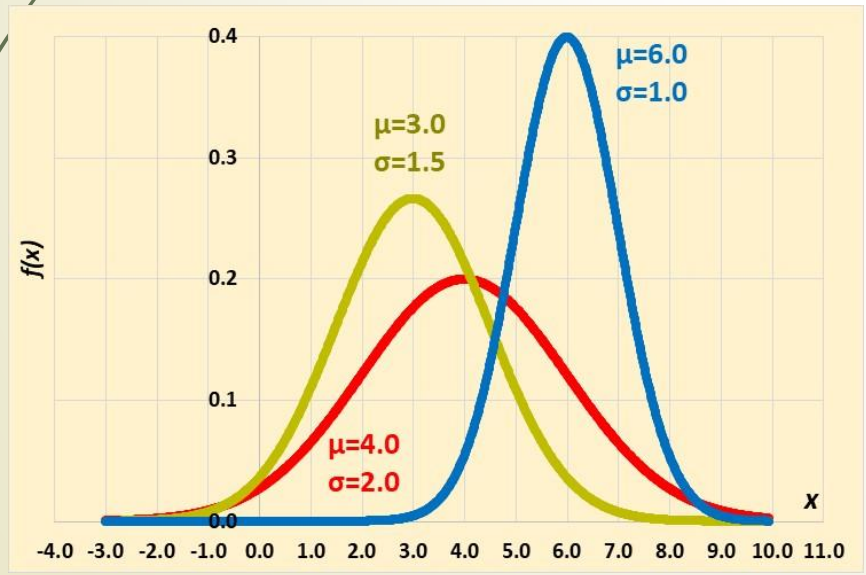




REGULA EMPIRICĂ



Descrie modul în care valorile individuale ale datelor studiate ar fi distribuite sub curba de distribuție, dacă acestea ar fi normal distribuite (are la baza media μ și abaterea standard a datelor σ). De exemplu, 68,27% reprezintă suprafața cuprinsă între dreptele *x1=μ-σ*, *x2=μ+σ*, curba densității de probabilitate și axa orizontală a absciselor -> există o probabilitate de 68,27% ca o valoare observată într-o distribuție normală să aparțină acestui interval.



Datorită acestor caracteristici, putem afirma faptul că media μ reprezintă un parametru de localizare și deviația standard σ un parametru al formei distribuției normale.

Având în vedere variația parametrilor *μ* și *σ* putem vorbi de un număr practic nelimitat de repartiții normale, fiecare având media și deviația sa standard particulare.

DISTRIBUȚIA NORMALĂ STANDARD (REDUSĂ)

Caracteristicile prezentate anterior arată faptul că, distribuția normală este „o familie de distribuții”, în care fiecare membru se deosebește de altul în baza valorilor μ și σ. Cel mai important membru al acestei familii este distribuția normală standard $N(0,1)$, care are o medie egală cu zero și o deviație standard egală cu 1.

Orice distribuție probabilistică normală poate fi convertită într-o distribuție normală standard, scăzând media acesteia din fiecare valoare observată și împărțind această diferență la deviația standard. Rezultatele obținute prin această schimbare de variabilă poartă denumirea de valori z sau scoruri z.

Astfel, presupunând că avem o variabilă aleatoare distribuită normal X, valorile z_i asociate fiecărui nivel x_i al lui X ne arată distanța acestuia față de media aritmetică μ a variabilei exprimată în unități ale deviației standard σ:

$$z = \frac{x_i - \mu}{\sigma}$$

Densitatea de probabilitate:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-z^2}{2}} \quad \text{unde: } -\infty < z < +\infty$$

Funcția de repartiție Laplace:

$$F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{\frac{-z^2}{2}} dz$$

Valorile funcției $F(z)$ sunt tabelate, fiind disponibile în majoritatea manualelor de statistică. Ele se pot determina ușor și în programul Excel prin utilizarea funcției =NORM.S.DIST(z,TRUE).

Localizăm, de exemplu, valoarea z=1,50 în tabelul alăturat și citim valoarea corespunzătoare a probabilității cumulate de 0,9332. Valoarea de 93,32% reprezintă probabilitatea ca o valoare z extrasă aleator din populația Z să aibă o valoare cuprinsă între -∞ și 1,50. Aceasta poate fi interpretată și ca proporția (frecvența relativă de apariție) a valorilor lui z cuprinse între -∞ și 1,50.

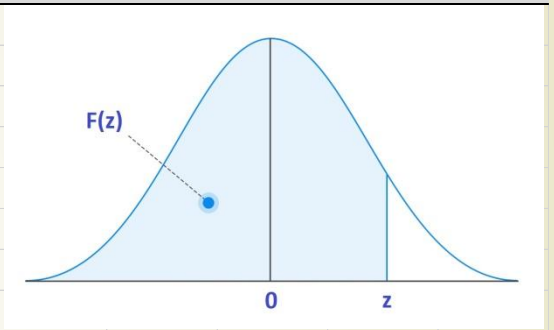
Probabilitatea ca o anumită variabilă de interes X să ia valori într-un interval (x_1, x_2) se determină, utilizând schimbarea de variabilă z:

$$P(x_1 < x < x_2) = P\left(\frac{x_1 - \mu}{\sigma} < z < \frac{x_2 - \mu}{\sigma}\right) =$$

$$= P\left(-\infty < z < \frac{x_2 - \mu}{\sigma}\right) - P\left(-\infty < z < \frac{x_1 - \mu}{\sigma}\right)$$

DISTRIBUȚIA NORMALĂ STANDARD

z	Probabilități cumulate						
	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.00	0.5000	0.5040					
0.10	0.5398	0.5438					
0.20	0.5793	0.5832					
0.30	0.6179	0.6217					
0.40	0.6554	0.6591					
0.50	0.6915	0.6950					
0.60	0.7257	0.7291					
0.70	0.7580	0.7611					
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846



3.2. METODE DE EȘANTIONARE ȘI TEOREMA LIMITĂ CENTRALĂ

- **Metode de eșantionare:** eșantionarea simplă aleatoare, eșantionarea sistematică, eșantionarea stratificată aleatoare, eșantionarea tip cluster (ciorchine) și eșantionarea de convenienţă
- **Distribuțiile de sondaj și eroarea de sondaj**
- **Distribuția de sondaj a mediilor eșantioanelor**
- **Teorema limită centrală**

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O6-1: să explice de ce este utilă eșantionarea unei populații și să descrie cele 4 metode de eșantionare

O6-2: să definească distribuțiile de sondaj și eroarea de sondaj

O6-3: să explice distribuția de sondaj a mediilor eșantioanelor

O6-4: să explice distribuția de sondaj a proporțiilor eșantioanelor

O6-5: să definească teorema limită centrală și să o aplice în calculul probabilităților

O6-6: să înțeleagă conceptele de bază ale eșantionării cu și fără revenire

AVANTAJELE EȘANTIONĂRII

În cadrul studierii caracteristicilor unei populații, pentru a afla valoarea anumitor indicatori statistici (media, proporția, varianța, etc.) există o serie de considerente de natură practică datorită cărora preferăm să selectăm, observăm și măsurăm anumite porțiuni (eșantioane) ale colectivității generale, care să prezinte „în miniatură" trăsăturile esențiale ale acesteia.

- **Contactarea și măsurarea caracteristicii statistice din programul cercetării la nivelul întregii populații ar putea implica o perioadă imensă de timp (de ex., dacă un potențial candidat la președinția țării ar dori să determine șansele de a fi ales, acesta poate apela la o companie specializată în domeniul sondajelor de opinie, care va intervieva un eșantion reprezentativ de persoane, operațiune care ar dura câteva zile; în situația în care ar dori ca aceeași firmă de specialitate să investigheze toată populația țării cu drept de vot, ar fi nevoie de câțiva ani pentru a finaliza cercetarea)**

- **Costul studierii tuturor unităților statistice care compun populația generală ar putea fi imposibil de achitat (de ex., pentru a testa un produs nou, utilizarea unui eșantion reprezentativ de 1000 de persoane la care să fie trimise mostrele și colectate apoi rezultatele reprezintă un obiectiv tangibil, comparativ cu testarea produsului pe întreaga populație a țării)**
- **Imposibilitatea fizică de a verifica toate unitățile statistice dintr-o anumită populație (anumite populații pot fi considerate ca fiind „infinite”; este imposibilă de ex. verificarea întregii cantități de apă dintr-un anumit lac pentru a determina nivelul bacteriilor, soluția fiind în schimb utilizarea mai multor eșantioane din diferite locații ale acestuia)**
- **Natura distructivă a unor teste (de ex. dacă într-o anumită casă de vinuri se dorește o verificare a calității produselor, iar persoanele care efectuează testarea ar utiliza în analiză toată populația, nu ar mai rămâne niciun produs disponibil în vederea comercializării)**
- **Rezultatele obținute din analiza unui eșantion reprezentativ sunt adecvate (chiar dacă de ex. resursele financiare, de timp și umane ar fi disponibile nelimitat, în vederea determinării indicelui lunar al prețurilor de consum pentru produsele alimentare, analiza tuturor magazinelor de specialitate din România nu ar afecta semnificativ valoarea indicelui, deoarece prețurile laptelui, pâinii și altor bunuri de consum alimentare nu înregistrează variații însemnate între marile lanțuri de magazine)**

METODE DE EȘANTIONARE

Așa cum am văzut anterior, rolul statisticii descriptive este centralizarea și prezentarea într-o formă convenabilă a ceea ce deja cunoaștem, iar cel al statisticii inferențiale este „să învățăm elemente noi, necunoscute încă, pornind de la ceea ce cunoaștem”. Dacă teoria probabilităților reprezintă fundația care stă la baza întregi teorii statistice, teoria sondajelor poate fi privită ca fiind elementul structural esențial necesar pentru definitivarea întregii forme a construcției cunoașterii statistice. Astfel, teoria sondajelor deține un rol important în precizarea ipotezelor de lucru pe care se bazează analiza statistică inferențială ulterioară.

1. EȘANTIONAREA ALEATORIE SIMPLĂ

Sondajul întâmplător simplu se caracterizează prin următoarele trăsături specifice:

- fiecare unitate a populației are asigurată aceeași probabilitate de a face parte din eșantion;
- extragerea fiecărei unități statistice din cadrul populației este independentă de oricare alta;
- volumul populației generale rămâne același pe parcursul tuturor extragerilor de constituire a eșantionului.

Principiile de constituire a eșantionului în cazul unui sondajului aleatoriu simplu se bazează pe schema probabilistică clasică a urnei lui Bernoulli cu bila revenită sau cea a bilei nerevenite (schema hipergeometrică). În funcție de schema probabilistică ce se utilizează pentru constituirea eșantionului, sondajul aleator simplu poate fi repetat sau nerepetat.

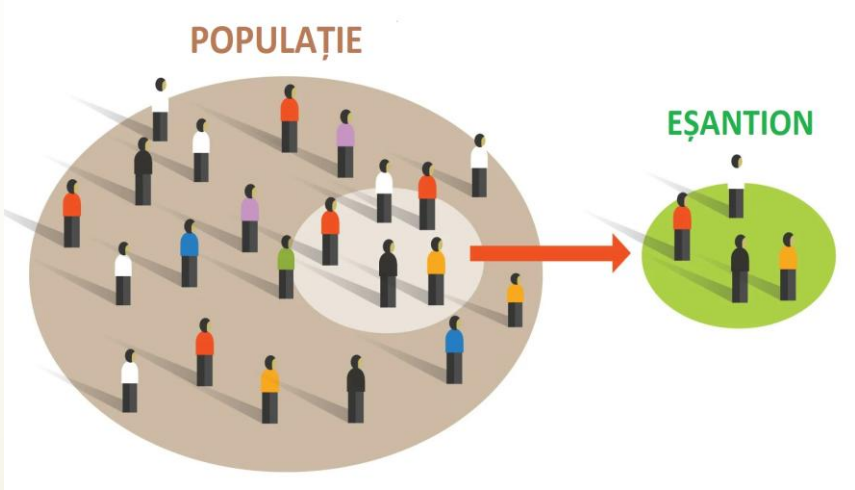
În cazul **sondajului aleator simplu repetat**, fiecare element component al populației (obiect, persoană, etc.) de volum *N* are aceeași șansă de a fi selectat și introdus în eșantionul de volum *n*, respectiv $p = 1 / N$.

Sondajul aleator simplu nerepetat are loc atunci când unitatea statistică extrasă nu mai revine în cadrul populației *N* din care a fost extrasă. Aceasta înseamnă că, pe măsura constituirii eșantionului, volumul populației se diminuează cu unitățile selectate. Astfel, după prima extragere, populația va fi compusă din *N-1* elemente, după a doua extragere, din *N-2* și așa mai departe, încât după extracția *n* vor rămâne în cadrul colectivității generale *N - n* unități. Probabilitatea de selecție crește ușor pe măsura constituirii eșantionului, astfel, pentru prima unitate este 1 / N , pentru a doua 1 / (N – 1) , iar pentru extracția *n* va fi $1 / (N – n + 1)$.

În practică, **o distribuție hipergeometrică poate fi de obicei aproximată printr-o distribuție binomială**, întrucât în situația în care volumul eșantionului nu depășește 5% din cel al populației, există o diferență redusă între eșantionarea cu și fără revenire.

De ex., dintr-o populație de 2000 de studenți din anul I ai unei universități, dorim să extragem un eșantion reprezentativ de 50 de persoane. Putem nota numele fiecărui student, din cei 2000 care compun populația de referință, pe câte o bucată de hârtie și introduce aceste hârtii într-o cutie, unde, după ce le amestecăm bine, putem extrage cei 50 de studenți care vor constitui eșantionul de lucru, fără să returnăm fiecare bucată de hârtie în cutie. Observăm faptul că probabilitatea selectării fiecărui element nou în eșantion crește foarte puțin, datorită selecției cu revenire, dar volumul eșantionului fiind sub 5% din cel al populației (50/2000 = 2,5%) diferențele nu influențează semnificativ rezultatele. Ca variantă alternativă, pentru selectarea eșantionului pot fi utilizate tabelele cu numere aleatoare, după realizarea în prealabil a unei liste cu cei 2000 de studenți. Se alege la întâmplare o căsuță din tabel pentru primul student extras și ulterior se selectează și restul de studenți din componența eșantionului, alegând în ordinea în care apar în tabel toate numerele cuprinse între 1 și 2000.

Astfel eșantionul va fi reprezentativ pentru populația de referință, putând fi astfel utilizat în analiza statistică ulterioară. În practică, este relativ ușor să extragem un eșantion dintr-o populație de referință, prin utilizarea programului Excel (Modulul „Data Analysis” → „Sampling”: la Input Range se introduce lista cu numerele de ordine ale elementelor populației generale, apoi se precizează metoda de eșantionare „Random” și volumul dorit al eșantionului „Number of Samples”).



2. EȘANTIONAREA ALEATORIE SISTEMATICĂ

Poate reprezenta o alternativă la eșantionarea aleatorie simplă în situația când dispunem de o listă completă a tuturor elementelor care compun o populație. În realizarea efectivă a selecției, se alege un punct de start (oricare element al listei) și prin adunarea unei constante la acesta se obține următorul element din eșantion, și așa mai departe, procedeul repetându-se până la completarea întregului volum al lotului de studiu. Observăm o anumită periodicitate în selectarea elementelor componente ale eșantionului, care în final, în anumite situații, poate conduce la erori de reprezentativitate a lotului extras. În cazul când lista care conține toate unitățile populației pe care o utilizăm prezintă și ea o ciclicitate asemănătoare, eșantioanele generate pot avea erori de deplasare (un ESTIMATOR ESTE NEDEPLASAT, dacă valoarea așteptată a distribuției sale de sondaj este egală cu valoarea reală, dar necunoscută a parametrului populației pe care acesta îl estimează → de exemplu, media mediilor tuturor eșantioanelor de un anumit volum n este egală cu media reală a populației din care acestea au fost extrase, reprezentând un estimator nedeplasat al acesteia; în schimb, media varianțelor acestor eșantioane este întotdeauna mai mică decât varianța populației generale, constituind un estimator deplasat al acesteia, în special în cazul eșantioanelor de volum redus). Eșantionarea sistematică poate fi o soluție adecvată în situația când eșantionarea aleatorie simplă nu poate fi aplicată. De exemplu, un supermarket dorește să studieze cât timp petrec clienții în magazinele sale. Deoarece nu avem o listă exhaustivă cu toți clienții, atribuirea de numere aleatoare acestora este imposibilă. O soluție ar putea fi, în acest caz, procedeul de eșantionare aleatorie sistematică, elementele care vor intra în componența eșantionului fiind alese la distanțe egale după o origine stabilită (punctul de start).

3. EȘANTIONAREA ALEATOARE STRATIFICATĂ

În situația când o anumită populație poate fi împărțită în mai multe grupuri sau straturi în funcție de o variabilă statistică de interes ne aflăm în situația eșantionării stratificate, în care eșantionul de lucru se obține prin selectarea în mod aleator a unui subeșantion din fiecare strat. Păstrarea proporției straturilor din populație în cadrul eșantionului conduce la o îmbunătățire a preciziei comparativ cu eșantionarea simplă aleatorie. Eșantionarea stratificată poate fi în anumite situații mai eficientă comparativ cea simplă aleatorie, în special când anumite subpopulații sunt rare. De exemplu, în studiul schizofreniei este indicat să se împartă populația de referință în 2 părți: schizofrenici și neschizofrenici, iar apoi să se selecteze aleator un număr egal de persoane din fiecare grup. Dacă am fi selectat aleator persoanele din populația generală, eșantionul ar fi conținut extrem de puține persoane cu schizofrenie, încât studiul ar fi fost lipsit de relevanță.

4. EȘANTIONAREA DE TIP CLUSTER (CIORCHINE)

Este utilizată deseori în scopul reducerii costurilor privind eșantionarea unei populații răspândite pe o arie geografică extinsă. În acest scop, populația studiată este împărțită în zone (grupuri sau clustere) utilizând de obicei limitele naturale ale regiunilor respective. Ulterior se selectează în mod aleator un anumit număr de clustere din care va fi extrasă de asemenea în mod aleator fiecare unitate care va intra în componența eșantionului.

5. EȘANTIONAREA CONVENABILĂ

Unitățile componente ale viitorului eșantion sunt alese într-un mod convenabil de către cercetător și nu selectate în mod aleator din populația de interes. Reprezintă o metodă mai accesibilă și ușor de aplicat, care nu garantează însă reprezentativitatea eșantionului obținut în acest fel. Deoarece permite obținerea mai rapidă a rezultatelor are rolul de a ajuta la conturarea unei prime imagini asupra fenomenului studiat. Estimările pot prezenta erori mari de deplasare, motiv pentru care rezultatele vor fi folosite doar ca studii pilot care ajută la schițarea unor noi direcții de cercetare.

DISTRIBUȚIILE DE SONDAJ

Conform LEGII NUMERELOR MARI (Jacob Bernoulli, 1713) dacă un experiment se repetă, în aceleași condiții, de un număr suficient de mare de ori, atunci frecvența relativă a unui eveniment prezintă o anumită stabilitate, variind în jurul probabilității de realizare a acestuia, bazată pe definiția clasică. Aplicând această lege mediilor eșantioanelor (media putând fi privită ca frecvența de apariție a unui eveniment) se poate afirma faptul că pe măsură ce volumul eșantionului devine mai mare ($n \to \infty$), media acestuia tinde să se apropie de valoarea adevărată a mediei populației din care acesta provine ($\bar{x} \to \mu$).

Din punct de vedere tehnic, Legea numerelor mari poate fi aplicată oricărui indicator statistic CALCULAT la nivelul unui eșantion SUB FORMA UNEI MEDII DE VALORI INDEPENDENTE, cum ar fi de exemplu varianța eșantionului (media pătratelor abaterilor fiecărui nivel individual de la nivelul mediu al eșantionului – momentul de ordinul 2), asimetria (momentul de ordinul 3), boltirea (momentul de ordinul 4) sau coeficientul de corelație (media produsului abaterilor standard ale celor două variabile). Valorile minimă și maximă ale unui eșantion însă nu pot fi cuantificate sub forma unor mărimi medii și din acest motiv nu intră sub incidența legii numerelor mari.

În situația concretă a unui studiu trebuie însă să cunoaștem anumite elemente privitoare la comportamentul mediei eșantionului de lucru când aceasta este calculată dintr-un set de date de multe ori modest.

DISTRIBUȚIA DE SONDAJ reprezintă mulțimea tuturor valorilor pe care le poate lua un indicator statistic calculat pentru toate eșantioanele de volum *n*, extrase în mod aleator din cadrul aceleiași populații de volum *N*.

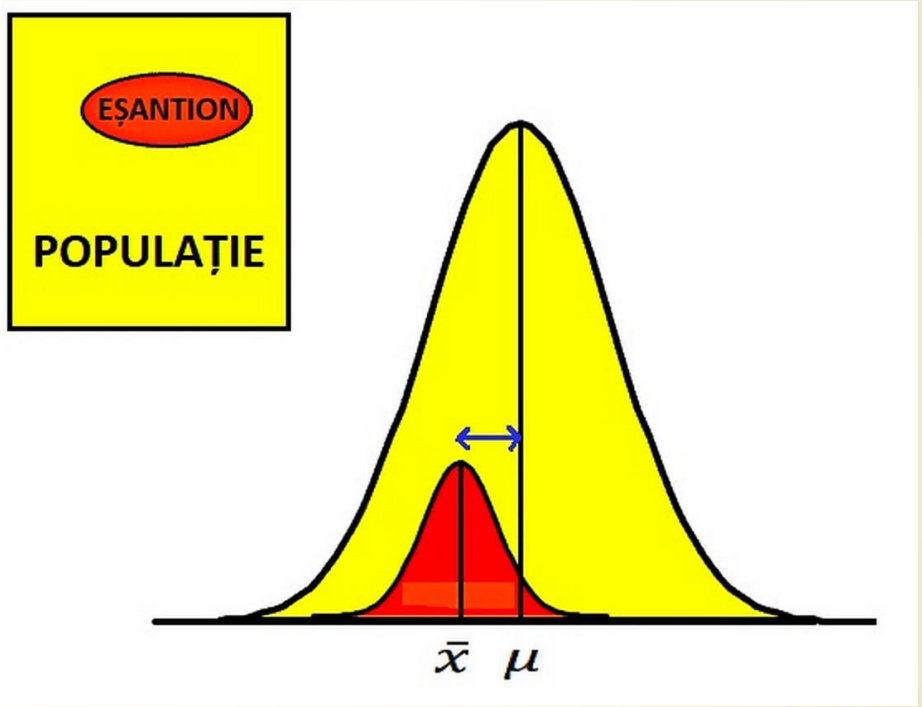
Construcția unei distribuții de sondaj se realizează extrem de dificil dacă populația de referință are un volum considerabil și reprezintă o sarcină imposibilă dacă populația este infinită. În aceste situații, distribuțiile de sondaj vor putea fi aproximate prin extragerea unui număr mare de eșantioane având același volum n.

În cazul distribuțiilor de sondaj, suntem de obicei interesați să cunoaștem media, varianța și forma funcțională a acestora. Printre cele mai des întâlnite distribuții de sondaj, în practică întâlnim: distribuția mediilor de sondaj, distribuția proporțiilor de sondaj, distribuția diferențelor dintre 2 medii de sondaj și distribuția diferențelor dintre 2 proporții de sondaj.

	PARAMETRU	STATISTICĂ
MEDIA	μ	$\bar{x}$ $\hat{\mu}$
DEVIAȚIA STANDARD	σ	S $\hat{\sigma}$
VARIANȚA	σ^2	s^2 σ^2
PROPORȚIA	π	p $\hat{\pi}$

EROAREA DE SONDAJ

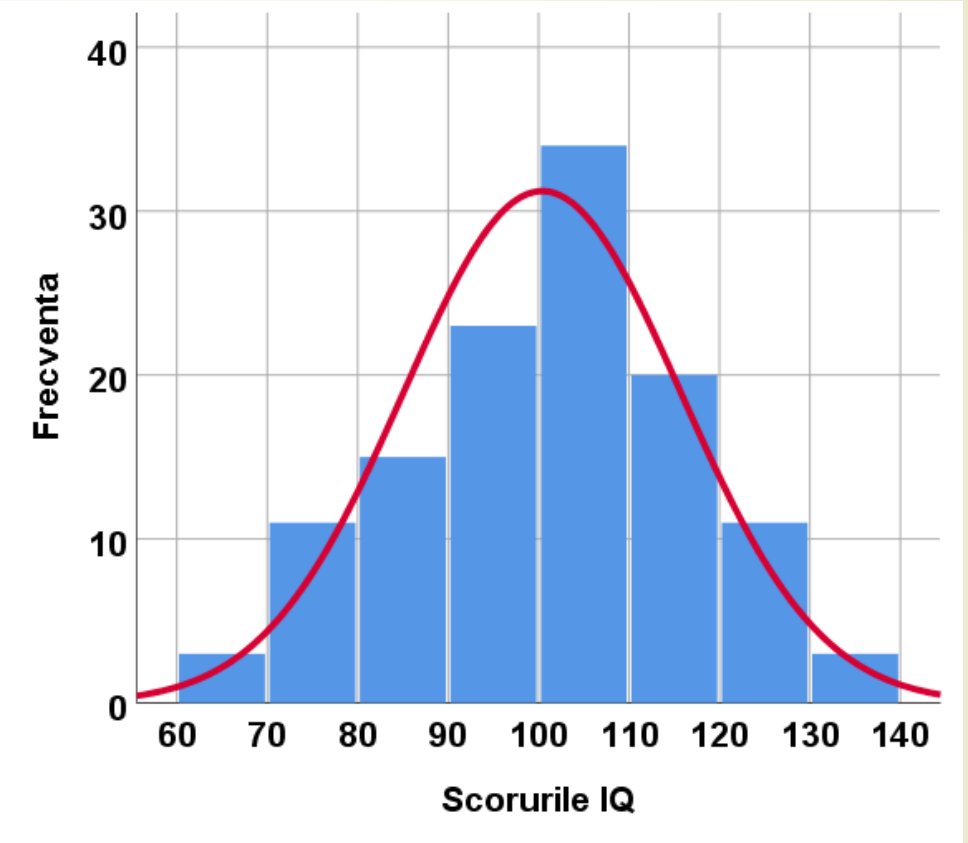
Eșantioanele sunt utilizate pentru a estima caracteristicile populației supuse studiului, de exemplu media unui eșantion este folosită în vederea estimării mediei populației din care acesta provine. Având în vedere faptul că eșantionul reprezintă doar o mică parte a populației generale, este foarte puțin probabil ca media calculată la nivel de eșantion să fie egală cu cea a întregii populații. În mod similar, deviația standard calculată pentru datele eșantion nu va fi perfect egală cu deviația standard a întregii populații. Din acest considerent, ne așteptăm să existe o diferență între statistica calculată la nivelul eșantionului și parametrul corespunzător care reprezintă întreaga colectivitate sau populație de referință. Această diferență poartă numele de EROARE DE SONDAJ.



$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \neq \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

DISTRIBUȚIA DE SONDAJ A MEDIILOR EȘANTIOANELOR

Presupunem că avem o „populație fictivă” de N = 120 de persoane, pentru care dorim să analizăm nivelul de inteligență, cuantificat prin scorurile IQ. Variabila statistică IQ urmează o distribuție aproximativ normală având o medie de 100,39 și o deviație standard de 15,27. Nivelurile variabilei studiate sunt cuprinse între o valoare minimă de 67 și o valoare maximă de 132. Histograma variabilei IQ din populația generală este:



Dorim să estimăm această valoare medie a scorurilor IQ din populația de referință pornind de la media unui eșantion de volum n = 5 extras aleator din aceasta și, în același timp, să putem evalua cât de precisă este această estimare.

Datele ordonate crescător privind scorurile IQ ale celor 120 de persoane care compun populația de referință sunt:

Distribuția de sondaj a mediilor eșantioanelor reprezintă o distribuție probabilistică a tuturor mediilor eșantioanelor de volum n extrase aleator dintr-o populație de volum N.

67	76	85	90	95	99	102	104	107	111	116	122
68	78	85	91	95	99	102	104	108	112	117	123
69	78	86	91	95	100	102	104	109	112	117	124
71	79	86	92	96	100	103	104	109	113	117	124
72	80	87	93	96	100	103	105	109	113	118	125
73	81	88	93	96	100	103	105	109	114	119	127
73	82	88	93	97	100	103	106	110	114	120	129
75	83	89	93	98	101	103	106	110	115	121	130
76	84	89	94	98	101	103	106	111	115	121	131
76	84	90	95	98	101	103	107	111	115	121	132

Revenind la exemplul prezentat anterior, din „populația” de volum *N = 120* persoane analizate prin prisma scorurilor IQ putem extrage în mod aleator un număr de $C_{120}^{5} = \frac{120!}{5!(120-5)!} = 190.578.024$ eșantioane posibile de volum *n = 5* persoane.

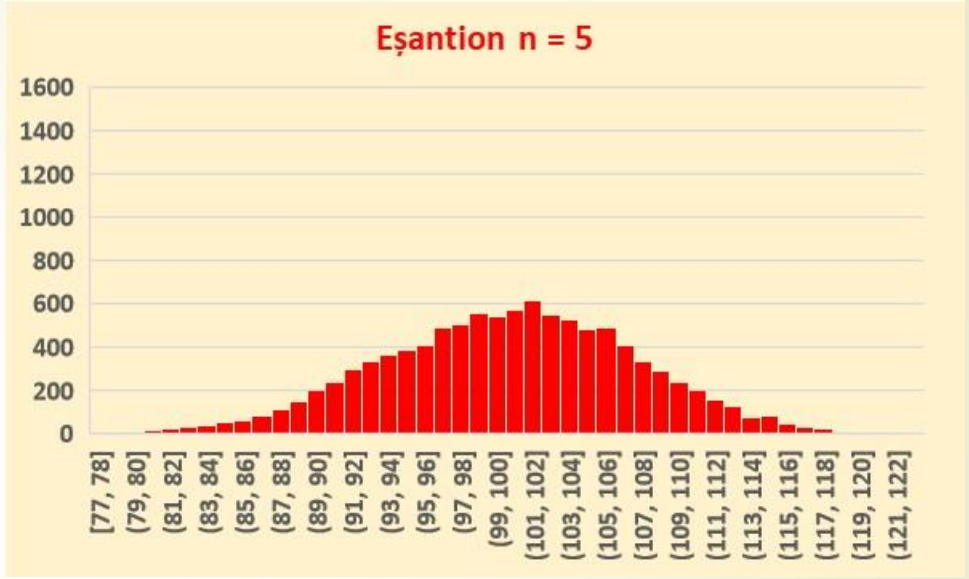
Reprezentând grafic distribuția celor 190.578.024 medii ale tuturor eșantioanelor de volum n = 5 extrase aleator din populația generală de 120 persoane vom obține distribuția de sondaj a acestora.

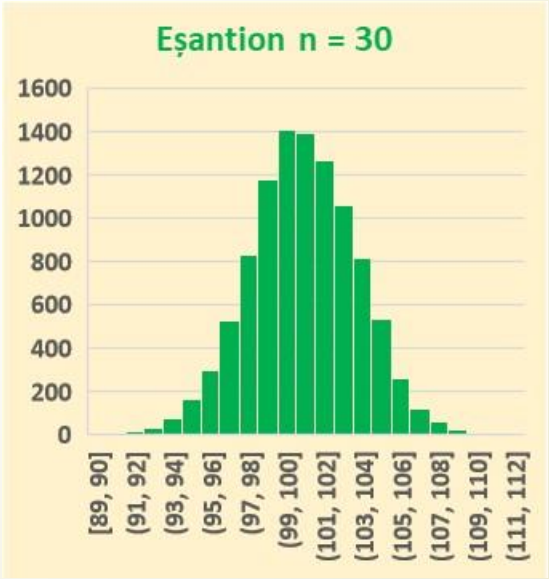
Deoarece construirea în acest mod a distribuției de sondaj a mediilor eșantioanelor ar presupune un efort enorm, fiind aproape imposibil de realizat în practică, ne vom rezuma la utilizarea unui număr mai redus, de doar 10000 de eșantioane, având fiecare un volum *n = 5* persoane, generate în mod aleator prin utilizarea programului Excel, de exemplu.

Eșantioane	Scorurile IQ individuale					Media IQ
	Pers. 1	Pers. 2	Pers. 3	Pers. 4	Pers. 5	
Eșantion 1	93	119	100	87	118	103.4
Eșantion 2	89	111	102	103	98	100.6
Eșantion 3	117	117	119	115	95	112.6
Eșantion 4	104	103	102	90	87	97.2
Eșantion 5	110	112	106	100	108	107.2
Eșantion 6	103	85	93	108	92	96.2
Eșantion 7	117	84	121	101	83	101.2
Eșantion 8	129	102	121	109	115	115.2
Eșantion 9	92	125	110	121	100	109.6
Eșantion 10	95	110	117	91	93	101.2
…	…	…	…	…	…	…
Eșantion 9998	68	103	84	104	105	92.8
Eșantion 9999	103	83	87	117	129	103.8
Eșantion 10000	112	83	89	96	91	94.2

Observăm faptul că distribuția mediilor celor 10000 de eșantioane de volum *n = 5* este aproximativ normală având o medie de 100,44, valoare care este aproape identică cu media populației generale din care acestea provin (100,39). Mediile fiecărui eșantion din cele 10000 extrase sunt cuprinse între o valoare minimă de 77,2 și o valoare maximă de 123,0. Observăm faptul că experimentul nostru cu eșantioanele de volum *n = 5* nu este foarte precis, mediile eșantioanelor având o variabilitate destul de mare.

În situația în care utilizăm un eșantion mai consistent, de volum *n = 30*, media mediilor IQ ale celor 10000 de eșantioane va fi de 100,37 din nou extrem de apropiată de media populației generale (100,39) din care loturile au fost extrase.



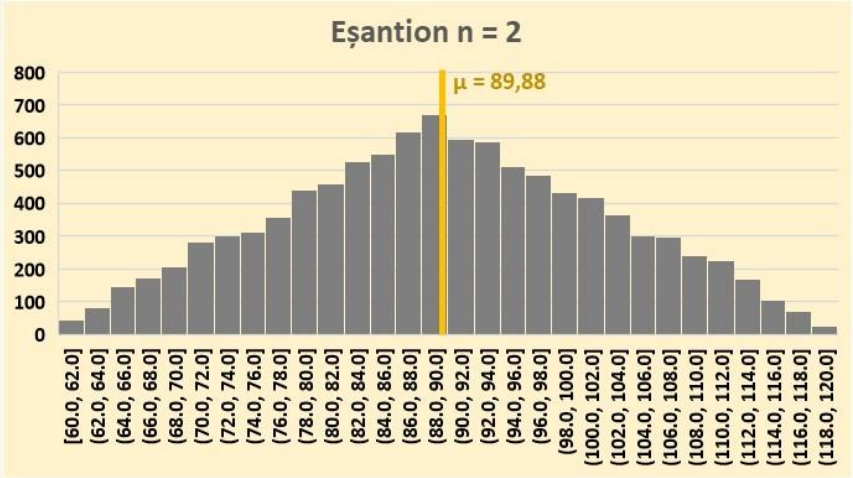


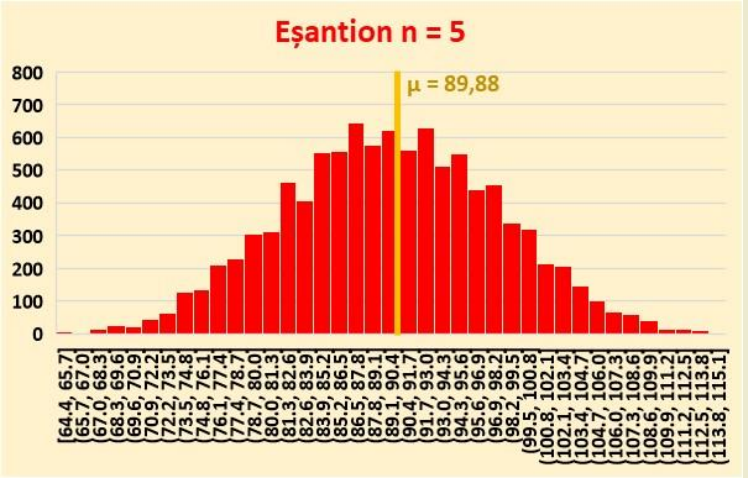
În același timp, remarcăm faptul că distribuția mediilor eșantioanelor este și în acest caz aproximativ normală, prezentând un nivel mult mai redus de împrăștiere a valorilor medii ale eșantioanelor comparativ cu cazul anterior și implicit o tendință de concentrare mai mare în jurul valorii reale a mediei populației de referință (de la 88,9 la 111,3).

Considerăm un al doilea exemplu, cel al unei populații compuse din N = 2000 de unități statistice, care urmează în acest caz o distribuție uniformă, fiind repartizată liniar constant în intervalul [60, 120], având o medie de 89,88 și o deviație standard de 17,92 (poate fi utilizată în acest sens funcția Excel: =RANBETWEEN(60,120)).

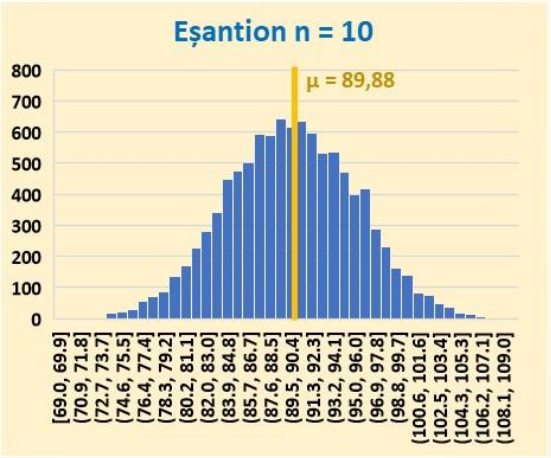
Histograma variabilei studiate din populația generală este:

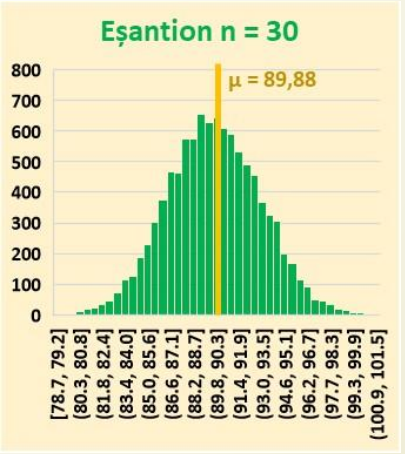
Vom extrage în mod aleator 10000 de eșantioane de volum *n = 2*, 10000 de eșantioane de volum *n = 5*, 10000 de eșantioane de volum *n = 10*, 10000 de eșantioane de volum *n = 30* și 10000 de eșantioane de volum *n = 50* din populația generală, compusă din 2000 de unități statistice. Distribuțiile de sondaj ale mediilor celor 10000 de eșantioane, în funcție de volumul acestora, se prezintă astfel:

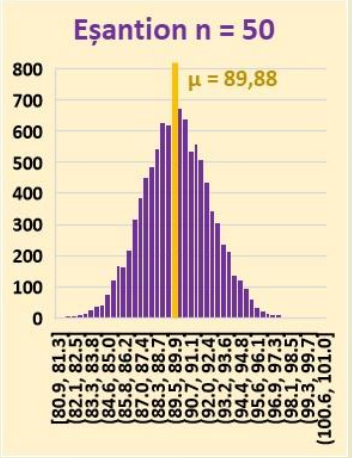




Și în al doilea exemplu observăm că distribuțiile de sondaj ale mediilor eșantioanelor sunt aproximativ normale în toate cele 5 cazuri, chiar dacă populația generală din care sunt extrase nu urmează o distribuție normală.







Media mediilor eșantioanelor este pentru fiecare situație în parte aproximativ egală cu cea a populației. În schimb, media variațelor eșantioanelor pentru fiecare caz în parte este în mod constant mai mică decât varianța populației inițiale. De asemenea, observăm și în acest exemplu un nivel mult mai redus de împrăștiere a valorilor medii ale eșantioanelor de la media mediilor acestora pe măsura creșterii volumului eșantionului de lucru.

Exemplele anterioare ilustrează o serie de relații importante care se conturează între distribuția populației generale și distribuția de sondaj a mediilor eșantioanelor de diferite volume extrase din aceasta:

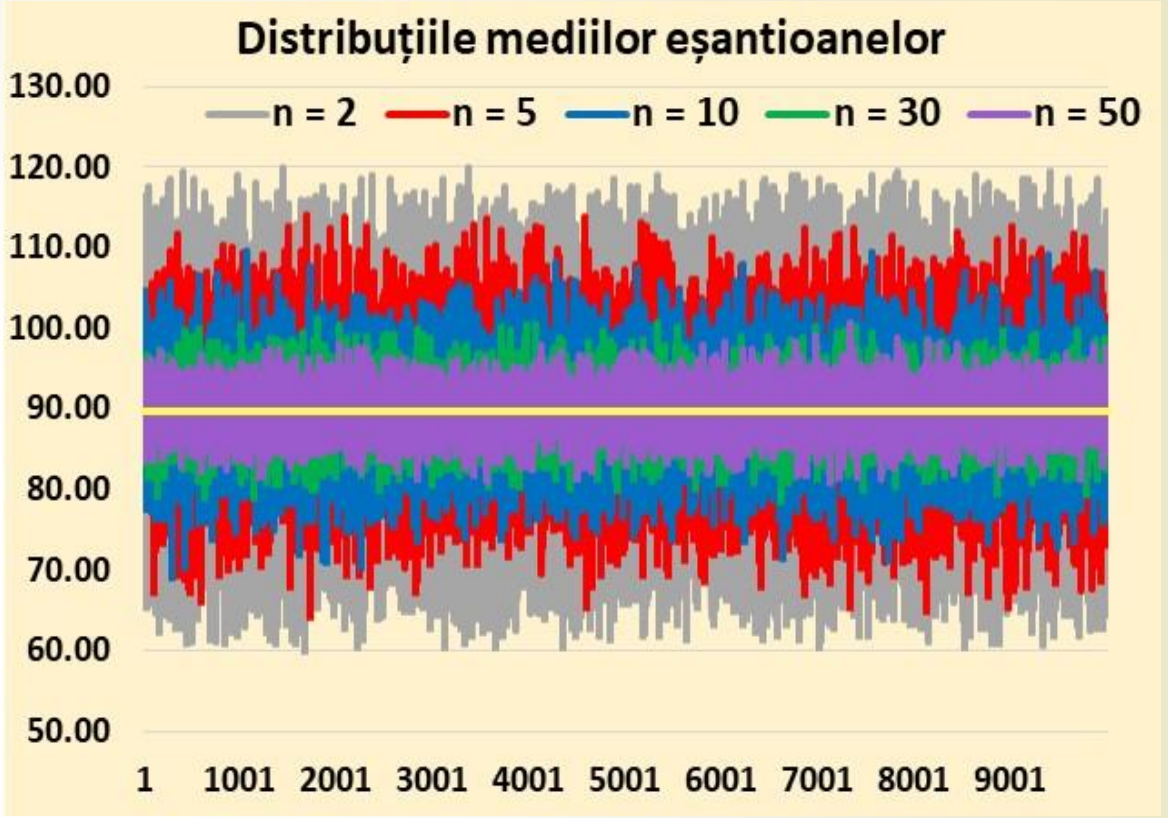
Volum eșantion	Media mediilor eșantioanelor	Media variațelor eșantioanelor	Varianța distribuției de sondaj a mediilor eșantioanelor
n = 2	89.88	163.83	158,16
n = 5	89.84	255.78	65,14
n = 10	89.94	286.65	32,42
n = 30	89.91	309.79	10,67
n = 50	89.91	314.82	6,37
Populație	89,88	320.99	-

a) **Media mediilor eșantioanelor este aproximativ egală cu media populației generale (în cazul în care am putea utiliza în calcul mediile tuturor eșantioanelor de același volum extrase aleator din cadrul populației de referință cele 2 valori ar fi perfect egale). Acest aspect atestă faptul că media mediilor eșantioanelor este un estimator nedeplasat al parametrului pe care îl estimează – media populației generale;**

b) **Media variantelor eșantioanelor în fiecare caz este mai mică comparativ cu varianța populației generale, fapt care ne arată că acesta este un estimator deplasat al parametrului pe care îl estimează – varianța populației generale;**

c) **Varianța distribuției de sondaj a mediilor eșantioanelor este mai redusă decât cea existentă în populația generală, fiind din ce în ce mai scăzută pe măsura creșterii volumului eșantionului de lucru;**

d) **Distribuția de sondaj a mediilor eșantioanelor poate fi aproximată printr-o distribuție probabilistică normală în situația când volumul eșantioanelor este mare ($n \geq 30$). Pentru $n \leq 30$ distribuția de sondaj a mediilor poate fi considerată normală doar daca populația generală este normal distribuită.**



Observăm tendința de diminuare a împrăștierii (dispersiei) mediilor celor 10000 de eșantioane de volume 2, 5, 10,30 și 50 unități în jurul nivelului mediu al lor (aproximativ 89,88) pe măsura utilizării unor eșantioane cu volum superior.

TEOREMA LIMITĂ CENTRALĂ

1. DISTRIBUȚIA DE SONDAJ A MEDIILOR

Atunci când extragem eșantioane de un anumit volum *n* din cadrul oricărei populații de volum *N*, indiferent de tipul de distribuție al acesteia, distribuția de sondaj a tuturor MEDIILOR eșantioanelor tinde către repartiția normală (Gauss-Laplace), aproximarea fiind din ce în ce mai bună pe măsura creșterii volumului acestora ($n \geq 30$).

- **Dacă populația generală de volum N are o medie μ și o deviație standard σ, atunci distribuția de sondaj a mediilor tuturor eșantioanelor de volum n extrase aleator din aceasta are de asemenea o medie μ și o eroare standard *SEM* (deviație standard a distribuției de sondaj a mediilor eșantioanelor $\sigma_{\bar{x}}$) egală cu:**

$$SEM = \frac{\sigma}{\sqrt{n}}$$

- **Întotdeauna avem nivel mai redus al dispersiei în distribuția de sondaj a mediilor eșantioanelor comparativ cu cel înregistrat în populația de referință. Observăm faptul că pe măsura creșterii volumului eșantionului de studiu n, eroarea standard a mediei se reduce.**

Eroarea standard a fost determinată anterior având la bază presupunerea faptului că eșantionarea s-a efectuat cu repetare sau eșantionul a fost extras dintr-o populație „infinită”. În general, procesul de eșantionare se efectuează fără revenire și în cele mai multe situații din cadrul unei populații de volum finit.

Astfel, atunci când eșantionul este extras fără revenire dintr-o populație finită de volum *N*, distribuția de sondaj a mediilor tuturor eșantioanelor de volum *n* extrase aleator din aceasta va avea media μ și eroarea standard:

$$SEM = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Elementul care apare în acest caz suplimentar în formulă, respectiv $\sqrt{\frac{N-n}{N-1}}$, poartă denumirea de factor de corecție pentru o populație finită și poate fi ignorat atunci când volumul eșantionului este foarte mic în raport cu cel al populației din care provine (de obicei când $\frac{n}{N} \leq 0,05$).

APLICAȚII PRACTICE ALE DISTRIBUȚIEI DE SONDAJ A MEDIILOR EȘANTIOANELOR

Având la dispoziție informații despre media μ, abaterea standard a unei populații σ și dimensiunea unui eșantion extras în mod aleator din aceasta n, putem obține distribuția mediilor de sondaj și în același timp putem calcula și probabilitatea ca media aritmetică a eșantionului să fie situată într-un anumit interval. Așa cum am văzut anterior, DISTRIBUȚIA DE SONDAJ A MEDIILOR EȘANTIOANELOR TINDE CĂTRE REPARTIȚIA NORMALĂ în cazul în care:

- **Eșantioanele sunt extrase din populații normal distribuite, situație în care volumul eșantionului nu joacă un rol important;**
- **Atunci când forma distribuției populației generale nu este cunoscută, dimensiunea eșantionului joacă un rol important; în practică distribuția de sondaj a mediilor se va apropia de o distribuție normală, în situația extragerii unor eșantioane de cel puțin 30 de observații.**

Din capitolul precedent cunoaștem faptul că orice distribuție probabilistică normală poate fi convertită într-o distribuție normală standard prin scăderea mediei din fiecare nivel individual al variabilei analizate și raportarea acestei diferențe la deviația standard a seriei:

$$z_i = \frac{x_i - \mu}{\sigma}$$

În practică însă suntem interesați însă de distribuția mediilor eșantioanelor $\bar{x}$ în locul distribuției variantelor individuale ale variabilei x. În al doilea rând, în cadrul acestei distribuții a mediilor eșantioanelor vom utiliza în locul deviației standard a populației generale σ, deviația standard a distribuției de sondaj a mediilor tuturor eșantioanelor extrase aleator din aceasta sau eroarea standard a mediei:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Singura valoare care rămâne în aceeași formă în cadrul formulei noi este MEDIA MEDIILOR TUTUROR EȘANTIOANELOR de volum *n* ($\mu_{\bar{x}}$) extrase în mod aleator din populația de volum *N*, care este un ESTIMATOR NEDEPLASAT AL MEDIEI POPULAȚIEI GENERALE (μ), având aceeași valoare cu aceasta ($\mu_{\bar{x}} = \mu$).

Astfel, atunci când dorim să aflăm șansa ca media populației din care a fost extras eșantionul să fie cuprinsă într-un anumit interval, utilizăm următoarea formulă pentru a afla valoarea z aferentă, iar apoi citim din tabelele distribuției z valoarea probabilității corespunzătoare acesteia; putem folosi în acest scop și funcția Excel NORM.S.DIST(z,cumulative=TRUE):

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$