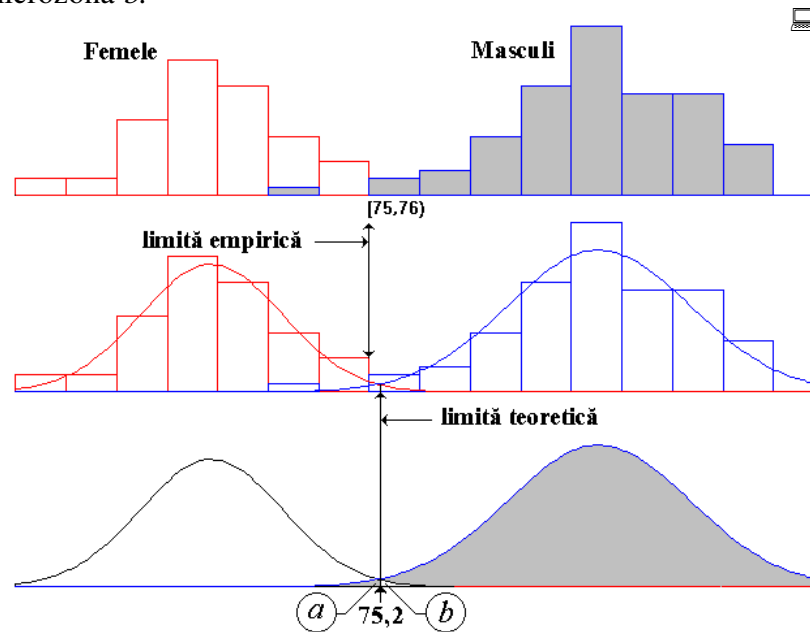


### § 3.7. →<sup>1</sup> Distribuția normală

Acesta este un subiect din planul teoriei probabilităților unidimensionale. Îl tratăm însă aici pentru a pregăti ideea - de statistică inductivă - a avantajului obținut atunci când "întrevedem" în spatele datelor empirice un model teoretic, de exemplu o distribuție normală.

De pildă, dacă în spatele celor două distribuții empirice ale lungimilor craniilor de jderi, din exemplul de la subpunctul 3<sup>o</sup> subparagraful 3.1.2., "vedem" două distribuții normale - ca în desenul exact următor, realizat printr-un program special construit - , obținem situația generală în care există ambele tipuri de erori de diagnostic, precum și o limită de discriminare mai bine fondată (75,2 în loc de 75). Este de presupus că această limită stabilită teoretic va produce, pe alte seturi de date, un procent total de erori - de ambele feluri - mai mic decât limita stabilită empiric. Este de așteptat, totodată, și o repartizare mai echilibrată a celor două feluri de erori: masculi considerați femele - vezi microzona *a* în desenul următor - , respectiv invers - vezi microzona *b*.



Această cale de efort teoretic este un substitut optim al variantei empirice în care, pentru a stabili mai bine o limită de discriminare, am măsura un număr gigantic de cranii de jderi. O asemenea variantă este sau extrem de costisitoare sau, adeseori, imposibilă.

În subparagraful 3.7.5., dedicat aplicațiilor distribuției normale, vom vedea că numai prin intermediul acestei distribuții teoretice putem stabili standardele de normalitate ale parametrilor fiziologici, ale măsurătorilor corporale etc.

Am intitulat paragraful "distribuția normală", dar vom vedea în continuare că, de fapt, este vorba de o familie infinită de distribuții normale.

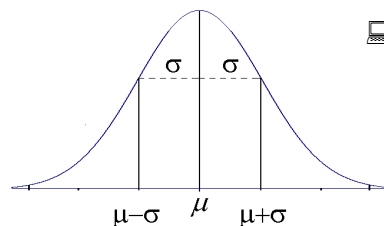
#### 3.7.1. Distribuție normală - descriere

**Sinonime:** clopot al lui Gauss - denumiri utilizate la începutul cursului, **distribuție Gauss, distribuție gaussiană, distribuție Laplace, distribuție Gauss-Laplace.**

<sup>1</sup> Semnul "→" indică un subiect *inserat* între subiectele capitoului dar care aparține altui "plan".

*Descriere:*

- ◆ Este o distribuție continuă, de forma unui clopot (deci unimodală și simetrică), cu două cozi infinite care tind asimptotic la zero.
- ◆ Este caracterizată de două numere  $\mu$  și  $\sigma$ , care sunt doi parametri ai distribuției:  
 $\mu$  este *media aritmetică*, iar  
 $\sigma$  este *abaterea standard*.
- ◆ Are două puncte de inflexiune (schimbare a concavității) situate, evident, simetric față de verticala  $x = \mu$ , la distanța  $\sigma$ . Adică punctele de inflexiune sunt  $\mu - \sigma$  și  $\mu + \sigma$ .
- ✓ Fiind unimodală și simetrică, media coincide cu moda și cu mediana.
- ✓ Întrucât  $\mu$  poate fi orice număr real, iar  $\sigma$  orice număr real strict pozitiv, rezultă că există, de fapt, o infinitate de distribuții normale.




---

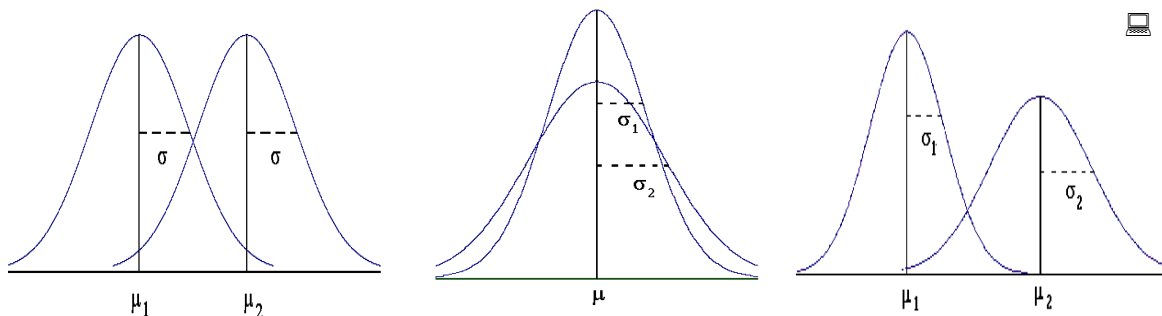
Distribuția normală de medie  $\mu$  și abatere standard  $\sigma$  se notează  $N(\mu, \sigma)$ .

---

Dacă fixăm atât pe  $\mu$ , cât și pe  $\sigma$ , vom obține o unică distribuție normală.

În desenele următoare sunt figurate perechi de distribuții normale care diferă:

- doar prin medii, adică  $\mu_1 \neq \mu_2$ ;
- doar prin abaterile standard, adică  $\sigma_1 \neq \sigma_2$ ;
- prin ambele, adică  $\mu_1 \neq \mu_2$  și  $\sigma_1 \neq \sigma_2$ .



### 3.7.2. Distribuția normală standard și consultarea tabelii corespunzătoare

Dintre toate distribuțiile normale se distinge distribuția cu media  $\mu = 0$  și abaterea standard  $\sigma = 1$  (vezi figura următoare).

---

Distribuția cu media 0 și abaterea standard 1 se numește **distribuția normală standard** și se notează  $N(0, 1)$ .

---

#### 1° Determinarea ariilor la dreapta punctelor și a $\alpha$ -cuantilelor superioare

Acest lucru se poate realiza direct prin consultarea tabelii de  *$\alpha$ -cuantile superioare* din anexa 1. Tabela poate fi utilizată în două moduri:

- a. pentru determinarea proporției de arie  $\alpha$ , sau altfel spus, a ariei relative  $\alpha$  aflate sub distribuția normală standard la dreapta unui punct dat  $z$   
 (sau, exprimându-ne riguros, dându-se  *$\alpha$ -cuantila superioară  $z$*  se determină  $\alpha$ );

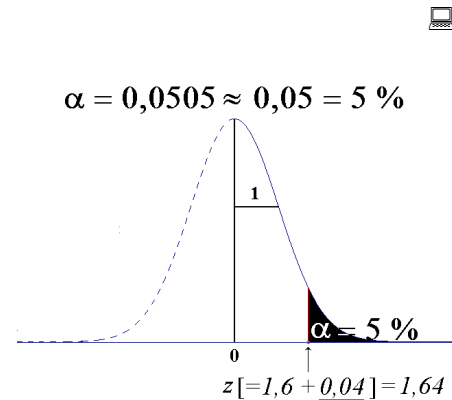
b. invers, pentru determinarea punctului  $z$  care lasă la dreapta sa, sub distribuția normală standard, aria relativă  $\alpha$

(sau , exprimându-ne riguros, dându-se  $\alpha$  se determină  $\alpha$ -cuantila superioară,  $z$ ).

### Exemplul 3.7.2.

a'. Aria relativă  $\alpha$  aflată la dreapta punctului  $z = 1,64$  se obține citind, în tabela a doua din anexa 1, valoarea înscrisă la intersecția liniei  $1,6$  cu coloana  $0,04$  (care sumate dau valoarea  $1,64$ ). Se obține  $\alpha = 0,0505 \approx 0,05 = 5\%$ .

b'. Invers, valoarea  $z$  care lasă la dreapta sa aria relativă  $\alpha = 0,05$  se află căutând în aceeași tabelă, de data aceasta nu pe capetele de linii și coloane, ci în interior, o valoare cât mai apropiată de valoarea  $\alpha$  căutată. În acest caz, aceasta poate fi  $0,0505$  sau  $0,0495$ , ambele fiind la aceeași distanță de  $\alpha = 0,05$ , dar în sensuri diferite. Ne fixăm la una dintre acestea, de exemplu la  $0,0505$ , și citim pe linie valoarea  $1,6$ , iar pe coloana corespunzătoare,  $0,04$ . Valoarea  $z$  va fi suma dintre ultimele două numere:  $z = 1,6 + 0,04 = 1,64$ .



**Notă:** Unii autori stabilesc prin interpolare valoarea  $z = 1,645$ .

- ✓ În anexa 1 sunt marcate prin chenar simplu, dublu, respectiv prin fond întunecat valorile  $\alpha$  utilizate în mod curent pentru  $\alpha$ -cuantile unilaterale superioare, bilaterale superioare, respectiv pentru ambele tipuri.

### +2° Tabela redusă

Datorită simetriei față de zero a distribuției normale standard  $\alpha$ -cuantilele din prima pagină a tabelii din anexa 1 pot fi deduse din cele din pagina a doua sau invers. Adică, în prima situație,  $x_{1-\alpha} = -x_{\alpha}$ , după cum am arătat la punctul 1° din 3.4.4.

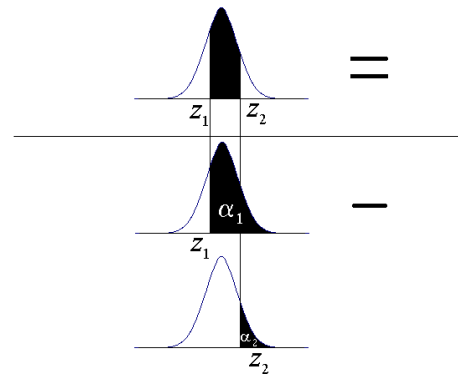
De exemplu,  $1,64$  fiind  $x_{0,0505}$ , punctul  $-1,64$  va fi cuantila superioară  $x_{1-0,0505}$  care lasă la dreapta sa aria  $1 - 0,0505$ , adică  $0,9495$ ; ceea ce se verifică și în tabelă. În concluzie, este suficientă tabela redusă la pagina a doua.

În cazul utilizării tabelii reduse trebuie adăugate următoarele reguli.

- a''. Dacă scorul  $z$  este negativ, consultăm tabela redusă (pagina a doua din anexa 1) pentru valoarea absolută a lui  $z$  și calculăm aria prin complementul față de 1 al ariei citite.
- b''. Dacă aria relativă,  $\alpha$ , este mai mare decât  $0,5$ , calculăm diferența  $1 - \alpha$  și consultăm tabela redusă pentru această nouă arie, după care punem semnul minus în fața cuantilei citite în tabelă.

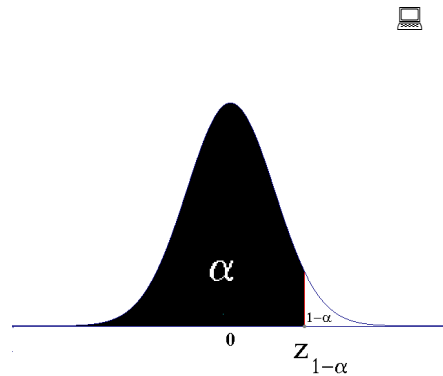
### 3° Determinarea ariilor cuprinse între două puncte

Figura alăturată indică modul în care se poate calcula aria dintre două puncte, sub distribuția normală standard.



### 4° Determinarea ariilor la stânga punctelor și a $\alpha$ -cuantilelor inferioare

- Determinarea ariei relative  $\alpha$ , aflate sub distribuția normală, la stânga unui punct  $z$ , se face ținând cont de faptul că  $z$  va lăsa la dreapta sa aria  $1 - \alpha$ .
- Invers, punctul  $z$  care lasă la stânga sa, sub distribuția normală standard, aria relativă  $\alpha$  este punctul care lasă la dreapta sa aria relativă  $1 - \alpha$ , adică  $z_{1-\alpha}$ . Exprimându-ne riguros, oricare ar fi  $\alpha$  subunitar,  $\alpha$ -cuantila inferioară este  $(1 - \alpha)$ -cuantila superioară.



#### Exemplul 3.7.2'.

- Aria relativă  $\alpha$ , aflată la stânga punctului  $z = 1,64$ , va fi 0,9495.
- Invers, valoarea  $z$  care lasă la stânga sa aria relativă  $\alpha = 0,9495$  va fi 1,64.

Reținem, deci, că aria relativă aflată la dreapta unui punct, sub distribuția normală standard este tabelată, iar aria la stânga este complementul față de 1 al ariei tabelate.

### 3.7.3. Standardizare

Ariile relative  $\alpha$ , respectiv  $\alpha$ -cuantilele (*inferioare* ori *superioare*) pe o distribuție normală oarecare se află cu ajutorul aceleiași tabele, deoarece orice distribuție normală de medie  $\mu$  și abatere standard  $\sigma$  poate fi transformată în distribuția normală standard,  $N(0,1)$ .

Într-adevăr,

- dacă vom executa translația  $x' = x - \mu$ , numită **centrare**, noua distribuție normală va avea media 0 - conform observației 1 de la 3.6.2. - și
- dacă vom aplica și comprimarea sau dilatarea  $x'' = x' / \sigma$ , numită **reducere**, vom obține și abaterea standard 1 - conform observației 3 de la 3.6.2.
- Deci aplicarea simultană a ambelor transformări

$$x'' = (x - \mu) / \sigma,$$

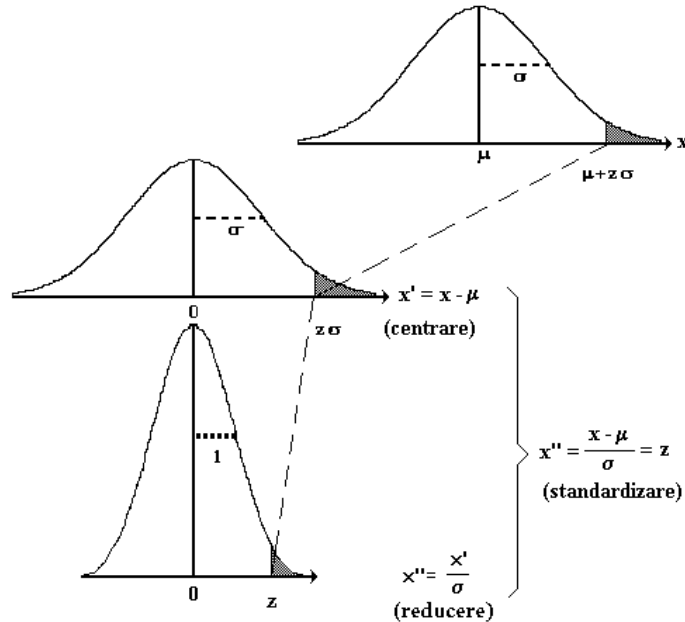
produce *distribuția normală standard*.  $x''$  se notează adesea  $z$ .

Transformarea  $z = (x - \mu) / \sigma$  se numește **standardizare** și rezultatul se numește **scor z**.

În consecință, exprimându-ne *riguros*:

$\mu + z \cdot \sigma$  pentru distribuția normală de medie  $\mu$  și abatere standard  $\sigma$  fixate și  $z$  pentru distribuția normală standard

sunt  $\alpha$ -cuantile cu același  $\alpha$  (fie inferioare, fie superioare).



### 3.7.4. Determinarea a diverse arii sub o distribuție normală oarecare

1° În cazul  $\alpha$ -cuantilelor superioare, respectiv al **ariilor  $\alpha$  la dreapta**:

- pentru determinarea ariei relative  $\alpha$  aflate, la dreapta unui punct dat,  $x$ , sub o distribuție normală oarecare de medie  $\mu$  și abatere standard  $\sigma$ , calculăm scorul  $z = (x - \mu) / \sigma$  și citim valoarea lui  $\alpha$  corespunzătoare lui  $z$  - vezi a' din exemplul 3.7.2. - în tabela pentru distribuția normală standard din anexa 1;
- pentru determinarea punctului  $x$  care lasă la dreapta sa sub o distribuție normală  $N(\mu, \sigma)$  proporția de arie  $\alpha$  aflăm cu ajutorul aceleiași table a distribuției normale standard, valoarea lui  $z$  corespunzătoare lui  $\alpha$  - vezi b' din exemplul 3.7.2. - după care  $x$  se obține din formula anterioară:  $x = \mu + z \cdot \sigma$ .

#### Exemplele 3.7.4.

Ariile **la dreapta unor puncte**, după standardizare, se citesc direct în tabela cu  $\alpha$ -cuantile superioare ale distribuției normale standard din anexa 1.

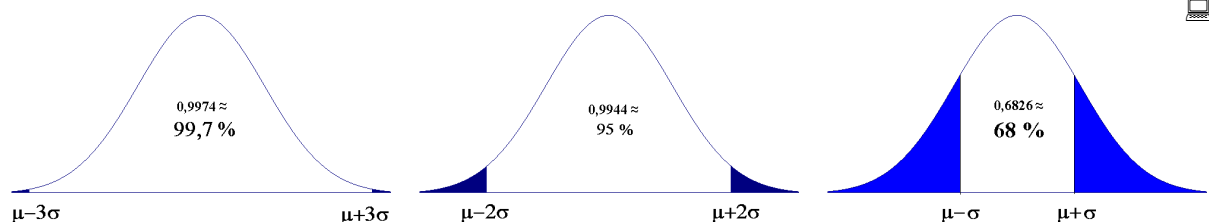
Distribuția normală de medie $\mu$ și abatere standard $\sigma$ $N(\mu, \sigma)$	Distribuția normală standard, $N(0, 1)$	Din tabela distribuției normale standard rezultă:
Aria la dreapta lui $\mu - 3\sigma =$	= (conform standardizării) = aria la dreapta lui $-3 =$	= 0,9987
Aria la dreapta lui $\mu + 3\sigma =$	= (conform standardizării) = aria la dreapta lui $3 =$	= 0,0013

## 2° Determinarea ariilor aflate între două puncte

### Exemplul 3.7.4'.

Distribuția normală de medie $\mu$ și abatere standard $\sigma$ , $N(\mu, \sigma)$	Distribuția normală standard, $N(0, 1)$	Din tabelul anterior rezultă:
Aria cuprinsă între $\mu - 3\sigma$ și $\mu + 3\sigma$ =	= aria cuprinsă între -3 și 3 = (aria la dreapta lui -3) - (aria la dreapta lui 3) =	= 0,9987 - 0,0013 = <b>0,9974</b>

În mod analog se calculează și ariile **intervalelor sigmatice** ( $\mu - 2\sigma$ ,  $\mu + 2\sigma$ ) și ( $\mu - \sigma$ ,  $\mu + \sigma$ ). Rezultatele sunt figurate în continuare.



## 3° Determinarea ariilor aflate la stânga punctelor

### Exemplele 3.7.4''.

Distribuția normală de medie $\mu$ și abatere standard $\sigma$ , $N(\mu, \sigma)$	Distribuția normală Standard, $N(0, 1)$	Din tabela de la punctul 1° rezultă:
Aria la stânga lui $\mu - 3\sigma$ (= 1 - aria la dreapta lui $\mu - 3\sigma$ ) =	1 - aria la dreapta lui -3	= 1 - 0,9987 = 0,0013
Aria la stânga lui $\mu + 3\sigma$ (= 1 - aria la dreapta lui $\mu + 3\sigma$ ) =	1 - aria la dreapta lui 3	= 1 - 0,0013 = 0,9987

## 4° $\alpha$ -cuantile remarcabile

Următoarele  $\alpha$ -cuantile sunt utilizate în mod frecvent în statistica inductivă tratată în capitolele 5 și 6. O prezentare sintetică a acestora poate fi făcută în următorul tabel preluat din [6], în care *procentajul* înseamnă  $\alpha$ , iar *deviația normală* înseamnă  $\alpha$ -cuantila unilaterală superioară:

Procentaj:	10%	5%	2,5%	1%	0,5%	0,1%	0,05%
Deviația normală:	1,28	1,64	1,96	2,33	2,58	3,09	3,29

## 3.7.5. Aplicații în biologie

### 1° Scări sigmatice în antropologie

În antropologie se folosesc frecvent mai multe tipuri de **scări sigmatice** cu trei, respectiv cinci *clase* sau *trepte*. Recomandăm utilizarea acestora și nu a celor bazate pe centile (vezi 3.4.2.) atunci când distribuția dimensiunii respective este unimodală simetrică, asimilabilă, deci, cu o distribuție normală. Cea mai utilizată scară este următoarea.

#### Scară cu cinci trepte:

Interval de variație a dimensiunii $x$ :	Diagnostic (clasă) dimensiune:
$\mu - 3\sigma \leq x < \mu - 2\sigma$	foarte mică
$\mu - 2\sigma \leq x < \mu - \sigma$	mică
$\mu - \sigma \leq x < \mu + \sigma$	medie

$$\begin{array}{l} \mu + \sigma \leq x < \mu + 2\sigma \quad \left| \begin{array}{l} \text{mare} \\ \text{foarte mare} \end{array} \right. \\ \mu + 2\sigma \leq x < \mu + 3\sigma \end{array}$$

Valorile  $x$  mai mici decât  $\mu - 3 \cdot \sigma$  sau mai mari decât  $\mu + 3 \cdot \sigma$  sunt excluse din serie conform regulii prezentate în continuare, la punctul 2<sup>o</sup>.

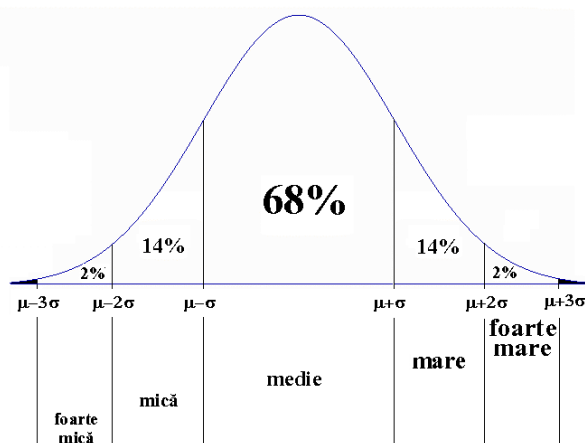
### Problema 3.7.5.

Ținând cont că marea majoritate a dimensiunilor corporale sau cefalo-faciale umane sau transformate ale acestora se repartizează gaussian pentru loturi omogene de volume mari, să se calculeze care sunt procentele de așteptat pentru fiecare clasă din scara sigmatică (aplicată variabilelor sau transformatelor acestora).

*Răspuns:*

Următoarea figură conține, grafic și numeric, răspunsul. Procentele au fost rotunjite la valori întregi, astfel ca suma lor să fie 100%. Se consideră că valorile sub  $\mu - 3 \cdot \sigma$ , respectiv, peste  $\mu + 3 \cdot \sigma$  sunt neglijabile (0,0013 fiecare).

Cititorul va face un bun exercițiu calculând aceste procente după exemplele cuprinse în 3.7.4.



O ilustrare a avantajelor utilizării unui model teoretic "întrevăzut în spatele" unei distribuții empirice este dată de următoarea problemă. Unul dintre avantajele va fi acela că vom putea preciza ce înseamnă o persoană - de sex feminin - scundă, înaltă, foarte scundă, foarte înaltă etc.

### Problema 3.7.5'.

Măsurându-se înălțimea (statura) fiecărei fete dintr-o populație statistică omogenă formată din  $N = 10\,000$  unități statistice s-a observat, după cum era de așteptat, o bună concordanță între histograma distribuției empirice și o curbă normală, ambele având media  $\mu = 165$  cm și abaterea standard  $\sigma = 5$  cm. Uitănd distribuția empirică, să se determine:

- proporția și numărul de fete cu înălțimea  $\geq 175$  cm;
- proporția și numărul de fete cu înălțimea cuprinsă între 165 cm și 175 cm;
- proporția și numărul de fete cu înălțimea  $< 175$  cm;
- proporția și numărul de fete cu înălțimea de 162 cm;
- intervalele sigmatice pentru scara de înălțimi prezentată mai sus;
- diagnosticul de înălțime pe scara sigmatică anterioară pentru o fată de 173 cm.

*Rezolvare:*

Ariile de sub o distribuție normală oarecare se determină după transformarea datelor prin standardizare (centrare și reducere):

a) Urmăm punctul 1<sup>o</sup> din 3.7.4.:

Notăm cu  $x_a = 175$  cm. Calculăm  $z_a = (x_a - \mu) / \sigma = (175 \text{ cm} - 165 \text{ cm}) / 5 \text{ cm} = 2$ .

Deci	$p(x \geq 175)$	=	$p(z \geq 2)$	=	0,0228 (din anexa 1).
<i>Citește:</i>	proporția de fete cu înălțimea $x \geq 175$		proporția fetelor cu scorul $z$ în distribuția normală standard $\geq 2$		

Numărul va fi  $N_a = N \cdot p(x \geq 175 \text{ cm}) = 10000 \cdot 0,0228 = 228$ .

b) Urmăm punctul 2<sup>o</sup> din 3.7.4.:

Notăm cu  $x_b = 165$  cm. Calculăm și  $z_b = (x_b - \mu) / \sigma = (165 \text{ cm} - 165 \text{ cm}) / 5 \text{ cm} = 0 \text{ cm} / 5 \text{ cm} = 0$ .

Astfel,  $p(165 \text{ cm} \leq x < 175 \text{ cm}) = p(x \geq 165 \text{ cm}) - p(x \geq 175 \text{ cm}) =$   
 $= p(z \geq 0) - p(z \geq 2) = 0,5 - 0,0228 = 0,4772$ .

Deci  $N_b = N \cdot p(165 \text{ cm} \leq x < 175 \text{ cm}) = 10000 \cdot 0,4772 = 4772$ .

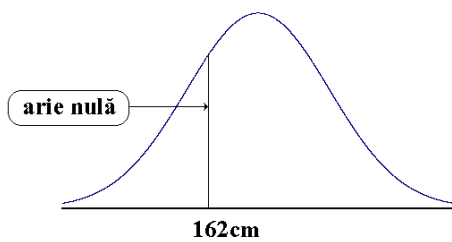
c) Urmăm punctul 3<sup>o</sup> din 3.7.4.:

$p(x < 175 \text{ cm}) = p(z < 2) = 1 - p(z \geq 2) = 1 - 0,0228 = 0,9772$ .

Deci  $N_c = N \cdot p(x > 175 \text{ cm}) = 10000 \cdot 0,9772 = 9772$ .

d) Ce înseamnă a avea înălțimea egală cu o anumită valoare întregă de cm, de exemplu 162 ?

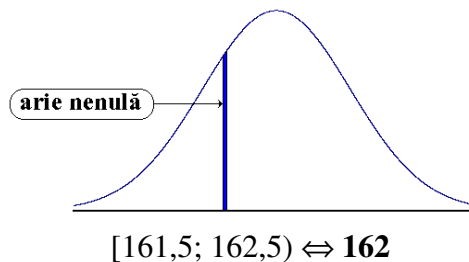
- a avea exact 162 cm ?



În acest caz nu putem folosi doar informațiile de mai sus, căci:

$$p(x = 162 \text{ cm}) = p(162 \text{ cm} \leq x < 162 \text{ cm}) = p(x \geq 162 \text{ cm}) - p(x \geq 162 \text{ cm}) = 0.$$

- a avea aproximativ 162 cm (cu aproximație de  $\pm 0,5$  cm),



adică  $p(x = 162 \text{ cm}) = p(161,5 \leq x < 162,5)$ .

A doua situație este cazul real. Aria fiind nenulă cazul poate fi rezolvat numai cu informațiile din problemă și metoda este cea de la punctul b. Astfel:

$$z_{d1} = (x_{d1} - \mu) / \sigma = (162,5 \text{ cm} - 165 \text{ cm}) / 5 \text{ cm} = (-2,5 \text{ cm}) / 5 \text{ cm} = -0,5$$

$$z_{d2} = (x_{d2} - \mu) / \sigma = (161,5 \text{ cm} - 165 \text{ cm}) / 5 \text{ cm} = (-3,5 \text{ cm}) / 5 \text{ cm} = -0,7$$

$$p(x = 162) = p(161,5 \text{ cm} \leq x < 162,5 \text{ cm}) = p(x \geq 161,5 \text{ cm}) - p(x \geq 162,5 \text{ cm})$$

$$= p(z \geq -0,7) - p(z \geq -0,5) =$$

(conform tabeli) = 0,758 - 0,6915 = 0,0665.

Deci  $N_d = N \cdot p(x = 162 \text{ cm}) = 10000 \cdot 0,0665 = 665$ .

e) Calculăm valorile limitelor intervalelor sigmatice:

$\mu - 3 \cdot \sigma = 165 - 3 \cdot 5 = 165 - 15 = 150$	$\mu + \sigma = 165 + 5 = 170$ $\mu + 2 \cdot \sigma = 165 + 2 \cdot 5 = 165 + 10 = 175$ $\mu + 3 \cdot \sigma = 165 + 3 \cdot 5 = 165 + 15 = 180$
$\mu - 2 \cdot \sigma = 165 - 2 \cdot 5 = 165 - 10 = 155$	
$\mu - \sigma = 165 - 5 = 160$	

Intervalele și denumirile vor fi:

$150 \leq x < 155$	foarte mică de statură
$155 \leq x < 160$	mică de statură
$160 \leq x < 170$	medie de statură

$170 \leq x < 175$	mare de statură
$175 \leq x < 180$	foarte mare de statură

f) Deoarece  $170 \leq 173 < 175$  fata este mare ca statură.

### Morala problemei:

Dacă dispunem de următoarele 4 informații:

- 1) faptul că distribuția empirică concordă cu o distribuție normală, ambele având:
- 2) media  $\mu$  și
- 3) abaterea standard  $\sigma$ ,
- 4) precum și volumul  $N$ , al distribuției empirice,

putem cunoaște întreaga distribuție  $(x_j, N_j)$ .

Dintr-o informație calitativă (punctul 1) și trei numerice (punctele 2 - 4) putem, deci, deduce un număr oricât de mare de perechi de numere  $(x_j, N_j)$ .

### INVERS:

Pornind de la distribuția empirică  $(x_j, N_j)$  care concordă cu o distribuție normală, putem spune că am sintetizat un număr oricât de mare de numere în trei valori și o formă (grafică). Problema anterioară ilustrează deci avantajul sintezei (grafice și numerice) a datelor, efectuate de statistică.

### 2° Regula "trei sigma" de eliminare a valorilor aberante

Pentru serii de volume mari<sup>2</sup> ( $N > 30$ ), în antropologie (și nu numai) se utilizează regula "trei sigma" pentru eliminarea valorilor aberante. Ea are sens oriunde putem considera că dacă am mări foarte mult numărul de unități ale seriei ne vom apropia de o distribuție normală. Pragul de "trei sigma" a fost stabilit bazându-ne pe proprietatea distribuției normale semnalată în exemplul 3.7.3': "aria cuprinsă între medie minus trei sigma și medie plus trei sigma este cca. 99,7% din întreaga arie". Altfel spus ceea ce depășește într-un sens sau altul aceste limite este neglijabil. De aceea, valorile din afara acestor limite se pot considera extrem de improbabile, deci aberante. Aceste limite sunt valabile pentru distribuția normală, deși pentru orice distribuție se pot stabili analog două limite peste care putem considera că distribuția are "cozi" neglijabile. Conform unei teoreme denumită **inegalitatea lui Cebâșev** și formulată aproximativ, "orice distribuție (teoretică sau empirică) se întinde practic între media sa plus / minus șase abateri standard (ale sale)". În consecință, eliminarea valorilor aberante în condițiile în care nu cunoaștem forma distribuției subiacente fenomenului, ar trebui să se execute conform regulii "șase sigma". Aplicând regula "trei sigma" înseamnă, în consecință, că știm sau presupunem că în spatele datelor se ascunde o distribuție normală.

### Exemplul 3.7.5.

În subparagraful 3.1.1. punctul 2° am prezentat seria S4 de măsurători ale lungimii palmei unui voluntar. Media respectiv abaterea standard pentru acest șir sunt  $M \cong 188,94$ , respectiv  $S \cong 2,91$ . În consecință  $M - 3 \cdot S \cong 180,2$ . Volumul seriei fiind mare ( $N = 36 > 30$ ) și presupunând o distribuție normală în spatele procesului de măsurare, putem aplica *regula 3 sigma*. Astfel rezultă că valorile 179 și 180 sunt valori aberante fiind mai mici decât limita inferioară 180,2. Trecerea la seria S4' prin eliminarea executată de noi, doar pe baze intuitive, în subparagraful menționat a fost, deci, perfect justificată.

---

<sup>2</sup> Problema detectării și eliminării valorilor aberante este o direcție specială a statisticii matematice (inductive). Pentru volume mici ( $N \leq 30$ ) sunt construite teste speciale în funcție de distribuția teoretică presupusă a fi "ascunsă în spatele" datelor de observație sau din experiment.

### 3° Stabilirea limitelor de normalitate biomedicală sau regula "doi sigma"

După cum am menționat și la începutul acestui paragraf (3.7.), modalitatea cea mai bine justificată de a stabili limite principiale este să apelăm la distribuția teoretică care se profilează sau măcar se presupune în spatele datelor empirice. Diverse studii executate pe volume mari de date au arătat că alături de multe măsurători corporale sau cefalo-faciale și mulți parametri biochimici umani sau, mai general, animalii se distribuie gaussian. Desigur, există și distribuții asimetrice ale unora dintre aceștia dar, așa cum am mai menționat, ele pot fi gaussianizate prin transformări adecvate. În concluzie, limitele de normalitate se stabilesc pe baza unor distribuții normale corespunzătoare (de unde vine probabil și denumirea de limite de normalitate). Mai precis, se parcurg următoarele etape:

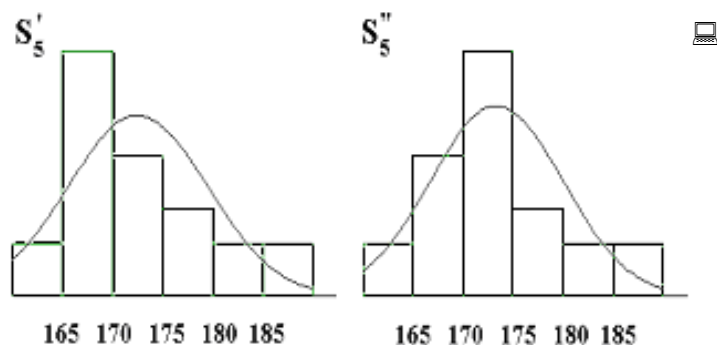
- A. Se colectează un volum cât mai mare de date de la indivizi considerați normali, dintr-o anumită populație biologică.
- B. Se calculează media  $M$  și abaterea standard  $S$  a seriei respective.
- C. Se verifică apoi concordanța distribuției caracterului sau a unei transformate a acestuia cu distribuția normală de aceeași medie și abatere standard.
- D. În final, se calculează drept limite de normalitate ale variabilei sau ale transformatei care o simetrizează, valorile  $M - 2 \cdot S$ , respectiv  $M + 2 \cdot S$ .

Rațiunea de alegere a două intervale sigmatice de o parte și de alta a mediei este dată exclusiv de faptul că între aceste limite trebuie să se afle (sub o distribuție normală) cca. 95% (mai exact 95,45% - vezi exemplul 3.7.3'.) din indivizii populației respective. Implicit considerăm astfel că cca. 5% din indivizii normali pot fi considerați totuși anormali, bolnavi etc. după caz. Este, evident, un risc pe care suntem obligați să ni-l asumăm, întrucât altfel nu putem tranșa acest tip de problemă.

#### 3.7.6. Măsurarea gradului de concordanță cu o distribuție normală

Am invocat de mai multe ori până acum ideea concordanței între o distribuție empirică și una teoretică, de exemplu cea normală (vezi de pildă punctul C din subparagraful anterior). Acest lucru se poate realiza în statistică pe cele două căi cunoscute: cea grafică, intuitivă, și cea numerică, exactă. Vom prezenta în continuare ambele căi aplicându-le pe seria grupată  $S_5'$  (de la 3.1.1. punctul 2°) și o nouă serie notată  $S_5''$  și obținută din  $S_5'$  prin permutarea frecvențelor din cea de-a doua, respectiv cea de-a treia clasă.

Pentru aprecierea grafică a concordanțelor desenăm, deasupra fiecărei histograme care reprezintă o distribuție empirică, distribuția normală care are aceeași medie ( $M=172,5$  respectiv  $173,33$ ) și abatere standard ( $S=6,97$ , respectiv  $6,61$ ) cu distribuția empirică corespunzătoare  $S_5'$ , respectiv  $S_5''$ .

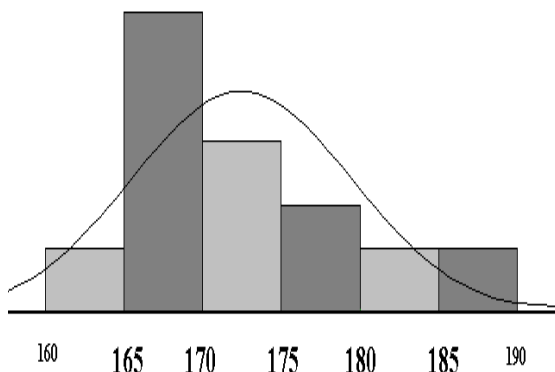


În figură se observă că noua distribuție concordă mai bine cu distribuția normală corespunzătoare. Să vedem în continuare cum putem măsura gradul de concordanță pentru a exprima numeric ceea ce am observat grafic.

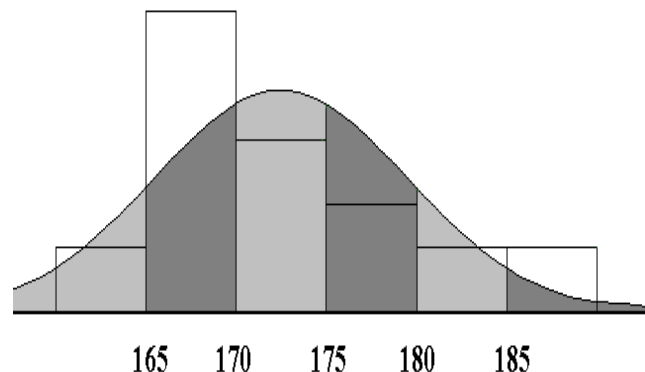
Pentru aceasta, mai întâi vom calcula frecvențele teoretice corespunzătoare distribuției normale pentru clasele unei distribuții, de exemplu S5'. Calculele se derulează întocmai ca în problema 3.7.5'. Prima clasă va fi intervalul  $(-\infty, 165)$ , iar ultima va fi  $[185, +\infty)$ . Se obține distribuția de frecvențe teoretice în ipoteza de normalitate (notate  $t_j$ ), formată din coloanele 1 și 3, respectiv 4 și 6, din tabela următoare. În coloanele 2 și 5 s-au reînscris frecvențele distribuției empirice, adică frecvențele observate renotate cu  $o_j (= N_j)$ .

Intervalul de clasă	$o_j$	$t_j$	intervalul de clasă	$o_j$	$t_j$
$(-\infty, 165)$	3	5,04	[175, 180)	5	7,89
[165, 170)	14	7,89	[180, 185)	3	3,72
[170, 175)	8	10,12	[185, $+\infty)$	3	1,32
$N = 36$ Total = 35,98					

- Frecvențele observate  $o_j$  sunt ariile dreptunghiurilor umplute cu nuanțe de gri:



- Frecvențele teoretice  $t_j$  sunt ariile de sub curba normală umplute cu nuanțe de gri:



Pentru măsurarea depărtării distribuției observate de cea teoretică se construiește un indicator asemănător dispersiei, notat  $\chi^2$  și citit “hi pătrat”:

$$\chi^2 = \sum_{j=1}^p \frac{(o_j - t_j)^2}{t_j} = \sum_{j=1}^p \frac{o_j^2}{t_j} - N$$

Prima formulă este cea teoretică (de definiție) iar cea de-a doua este formula de calcul rapid și exact. În formula teoretică se observă că acest  $\chi^2$  cumulează abaterile frecvențelor observate de la cele teoretice corespunzătoare, ridicate însă la pătrat pentru a nu se compensa cele de semn contrare (întocmai ca la dispersie) dar, în plus, fiecare pătrat de abateri este divizat cu frecvența teoretică corespunzătoare pentru ca ordinul de mărime al fiecărei frecvențe să nu influențeze suma.

Calculul lui  $\chi^2$  se poate organiza în următoarea tabelă aplicată exemplului nostru.

$o_j$	$o_j^2$	$t_j$	$o_j^2 / t_j$
3	9	5,04	1,79
14	196	7,89	24,84
8	64	10,12	6,32
5	25	7,89	3,17
3	9	3,72	2,42
3	9	1,32	6,82
Suma = 45,36			

În final obținem  $\chi^2 = \sum_{j=1}^p \frac{o_j^2}{t_j} - N =$   
 $45,36 - 36 = \mathbf{9,36}$

Cititorul este invitat să aplice aceeași rețetă de calcul seriei S5". Se va obține valoarea **5,35** care este mai mică decât cea corespunzătoare seriei S5' ceea ce confirmă numeric ceea ce am observat pe grafice, în mod intuitiv.

Cum apreciem dacă o valoare  $\chi^2$  este suficient de mică pentru ca să considerăm că avem o bună concordanță este, însă, o problemă de statistică inductivă care va fi tratată în volumul II, dedicat biostatisticii inductive.

- ✓  $\chi^2$  este un indicator de concordanță general, adică se poate aplica pentru măsurarea concordanței cu orice altă distribuție teoretică, nu numai cu distribuția normală. De exemplu, dacă se dorește măsurarea concordanței cu distribuția uniformă, se va lua pe post de distribuție teoretică distribuția uniformă a celor 36 de unități statistice în cele 6 clase. Deci toate frecvențele teoretice vor avea valoarea 6.