Notiuni introductive

Ce este analiza exploratorie a datelor?

Analiza exploratorie a datelor (*Exploratory Data Analysis* -EDA) este partea Statisticii care se ocupă cu trecerea în revistă, comunicarea şi utilizarea datelor în cazul unui nivel scăzut de informaţie asupra lor.

Reamintim:

Pentru a studia o anumită caracteristică a unei populaţii, nu vom analiza toată populaţia, ci vom considera un anumit eşantion din ea.

Populaţia statistică este mulţimea elementelor, de aceeaşi natură, care au unele însuşiri esenţiale comune, însuşiri caracteristice mulţimii privit ca un tot unitar.
Eşantionul (*sample*) este o submulţime a populaţiei statistice considerate.

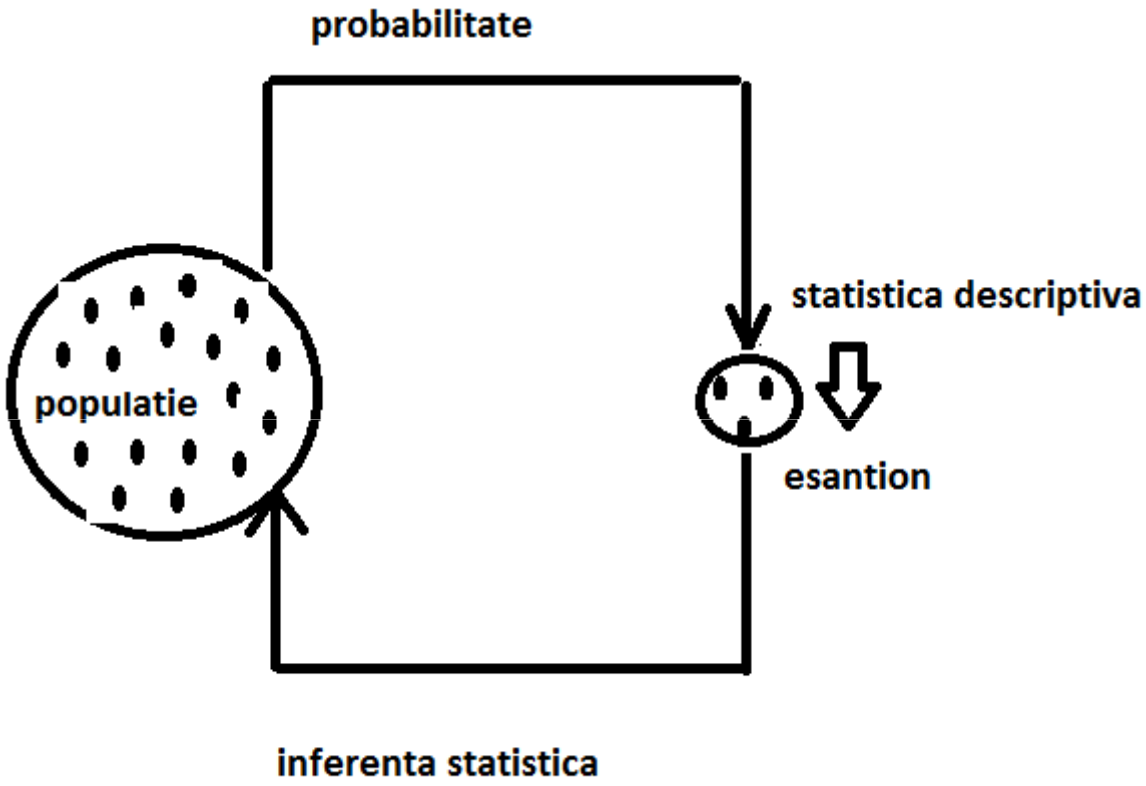
Reamintim:

Inferența statistică este totalitatea metodelor ce permit extragerea de concluzii demne de încredere pornind de la un eșantion statistic.

Prin metodele inferenței statistice deducem caracteristici necunoscute ale unei populații, pornind de la un eșantion obținut din acea populație. Caracteristicile eșantionului reflectă cu o anumită marjă de eroare caracteristicile întregii populații.

De reținut că inferența se aplică întregii populații, ca un tot unitar, și nu fiecărui element al populației respective.

Reamintim:


probabilitate
populatie
statistica descriptiva
esantion
inferenta statistica

Spre deosebire de cazul clasic al testării ipotezelor, utilizat în Statistică pentru a verifica anumite supoziţii apriorice

de exemplu: anumite corelaţii între diferite mărimi/variabile despre care există informaţii că ar fi cumva dependente

în cazul EDA se utilizează diferite tehnici pentru a identifica relaţii sistematice între anumite mărimi/variabile despre care nu există nicio informaţie prealabilă.

John Tukey, 1977

EDA a fost creată şi numită astfel de statisticianul american John Tukey (Tukey J., *Exploratory Data Analysis*, Addison-Wesley, 1977).

Tehnicile computaţionale EDA includ atât metode statistice elementare cât şi altele avansate – *tehnici exploratorii multivariate* – create pentru a identifica anumite pattern-uri ascunse în mulţimi multivariate de date.

Tehnicile EDA sunt utilizate cu scopul de a:

- maximiza cunoașterea intimă a datelor;
- dezvălui structura de bază;
- extrage variabilele importante;
- detecta valorile extreme/excepționale și anomaliile;
- identifica ipotezele fundamentale pentru a fi apoi testate;
- dezvolta modele suficient de simple;
- determina setarea optimă a parametrilor;
- sugera unele ipoteze privind cauzele fenomenelor observate;
- sugera tehnici statistice potrivite datelor disponibile;
- furniza cunoștințe pentru colectarea ulterioară de date pentru cercetare sau experimentare.

-

Odată utilizate tehnicile EDA se verifică rezultatele obţinute.

Explorarea datelor este doar un prim stadiu de analiză a datelor şi rezultatele obţinute vor fi considerate cu titlu experimental atâta timp cât nu sunt validate alternativ.

De exemplu, rezultatul aplicării EDA sugerează un anumit model, atunci acesta trebuie validat aplicându-l la alt set de date, testându-i astfel calitatea predictivă.

In concluzie:

Analiza exploratorie a datelor- EDA poate fi considerată –
în principiu – o filosofie despre modul în care să fie „disecate”,
să fie „privite” şi, în sfârşit, să fie interpretate datele.

Tipuri de date

Obiectele cu care lucrează statistica sunt reprezentate de *date*, adică acele caracteristici numerice sau nenumerice care descriu obiectele/subiecţii unui studiu statistic

Datele obţinute prin măsurare pot fi clasificate în funcţie de tipul de informaţie conţinut:

- ***Datele numerice (cantitative)***
- ***Datele categoriale (calitative)*** sunt acele date care împart obiectele în diferite categorii.

Datele numerice sunt de două tipuri: date ***discrete*** şi date ***continue***.

Datele *discrete* apar atunci când este vorba de observaţii numerice întregi, privitoare la un anumit proces de numărare

Exemple: numărul de copii ai unei familii, pulsul, numărul de consultaţii pe an la care a fost supus un pacient, codul numeric, codul PIN.

Datele numerice *continue* se obţin de obicei în urma unor măsurători. Aceste date sunt de regulă exprimate prin numere reale, spre deosebire de cele discrete care sunt restricţionate la numerele întregi.

Exemple: valoarea contului din bancă sau valoarea acţiunilor tranzacţionate la Bursă; temperatura, presiunea atmosferică, viteza vântului; înălţimea, greutatea, tensiunea arterială, colesterolul unei anumite persoane, etc.

Din date continue se pot obţine date discrete:
Exemplu: evaluarea venitului lunar

venit lunar < 1000 lei,
1000lei < venit lunar < 2000 lei ,
2000 lei < venit lunar

Datele categoriale sunt de două tipuri: date ***nominale*** şi date ***ordinale***

- Datele ***nominale*** sunt acele datele ce reprezinta mai multe categorii.

  Exemple: grupa sanguină (A/B/AB/O), culoarea ochilor, specia de Iris (Iris Setosa, Virginica şi Versicolour).

  Datele nominale pot fi:

  - binare (2 categorii)

Exemple: 0 sau 1; da / nu; adevărat / fals.

  - enumerative (date discrete pentru care nu este definită o ordine).

Exemple: categoriile socio-profesionale sau culoarea ochilor).

Datele nominale pot fi și numerice
Exemplu: codurile poștale.

- Datele ***ordinale*** sunt date enumerative ordonate

Exemple: gradul fumatului (nefumător, fost fumător, fumător ‚amator', fumător ‚înrăit'); ierarhizarea durerii (mică, medie, mare); răspunsurile la o anchetă de opinie (foarte mulţumit, mulţumit, nemulţumit, foarte nemulţumit) etc.

Observaţii

Datele numerice discrete sunt câteodată tratate ca date categoriale
Exemplu: numărul de copii născuţi de o femeie, 0, 1, 2, 3, 4, împart mamele în categoriile corespunzătoare numărului de copii.

Invers, nu este corect să interpretăm datele categoriale ordonate ca date numerice:
Exemplu: la stadiile unor anumite boli, stadiul IV nu este de două ori mai grav decât stadiul II.
Este important ca în această situaţie să se ignore noţiunile de ordine sau de parametri numerici ca, de exemplu, media.

Alte tipuri de date

- ***Rangul*** reprezintă locul pe care il ocupă un subiect într-o ierarhie

Exemple: competiţie sportivă, examinare, preferinţa clienţilor pentru o anumită marcă.

- ***Procentajul*** descrie o anumită proporţie Între două cantităţi.

Exemple: procentajul celor cu studii superioare dintr-o populaţie, procentajul minorilor dintr-o populaţie.

- ***Ratele şi rapoartele*** se referă la frecvenţa observată a unui anumit fenomen sau rapoartele dintre două mărimi, altele decât procentajele.

Exemple: mortalitatea la mia de locuitori, rata de apariţie a unei boli pe arii geografice

- ***Scorul*** este folosit atunci când nu este posibilă o măsurătoare directă, dar trebuie cuantificată o anumită mărime.

Exemple: Scorul Apgar În neonatologie, gravitatea unei boli cuantificată ca uşoară, moderată, severă, foarte severă.

De cele mai multe ori datele sunt stocate sub forma unei matrice.
Prin convenţie valorile variabilelor (atributelor/caracteristicilor) sunt stocate pe coloană în timp ce observaţiile (obiectele) sunt stocate pe linii.

Considerând n vectori de dimensiune p, notaţia x_i^k utilizată de obicei se referă la a i-a variabilă (atribut/caracteristică observată) a vectorului $\mathbf{x}_k$ (observaţia/ obiectul numărul k).
Se pot scrie într-un tabel cele n observaţii şi atributele lor, astfel:

observaţia	atributul 1	..	atributul i	..	atributul p
observaţia 1	x_1^1		x_i^1		x_p^1
.....					
observaţia k	x_1^k		x_i^k		x_p^k
.....					
observaţia n	x_1^n		x_i^n		x_p^n

Reprezentarea datelor sub forma unei matrice **X**, cu n linii şi p coloane:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \dots & x_i^1 & \dots & x_p^1 \\ \dots & \dots & \dots & \dots & \dots \\ x_1^k & \dots & x_i^k & \dots & x_p^k \\ \dots & \dots & \dots & \dots & \dots \\ x_1^n & \dots & x_i^n & \dots & x_p^n \end{pmatrix}$$

1. Exemplu

Bază de date Iris, datorată lui R.A. Fisher, 1936, constă în 50 de exemplare din fiecare specie de iris (*Iris Setosa*, *Virginica* şi *Versicolor*), fiecare exemplar fiind caracterizat de lungimea şi lăţimea sepalelor, respectiv a petalelor.

Prezentăm un tabel cu 7 flori, atributele lor (date *numerice continue*) şi specia de iris /clasa căreia îi aparţine fiecare floare (dată *nominală*). Menţionăm că pentru atribute avem notaţiile: PW lăţimea petalei, PL lungimea petalei, SW lăţimea sepalei, SL lungimea sepalei.

Specie de iris	PW	PL	SW	SL
Setosa	2	14	33	50
Virginica	24	56	31	67
Virginica	23	51	31	69
Setosa	2	10	36	46
Virginica	20	52	30	65
Versicolor	13	45	38	57
Versicolor	16	47	33	63

Prezentăm datele sub formă de matrice, codând datele nominale în felul următor:

0-Setosa, 1-Virginica, 2-versicolor.

$$\begin{pmatrix} 0 & 2 & 14 & 33 & 50 \\ 1 & 24 & 56 & 31 & 67 \\ 1 & 23 & 51 & 31 & 69 \\ 0 & 2 & 10 & 36 & 46 \\ 1 & 20 & 52 & 30 & 65 \\ 2 & 13 & 45 & 38 & 57 \\ 2 & 16 & 47 & 33 & 63 \end{pmatrix}$$

Pentru a accesa baza de date Iris a lui Fisher, in Matlab vom scrie:

```
>>load fisheriris
```

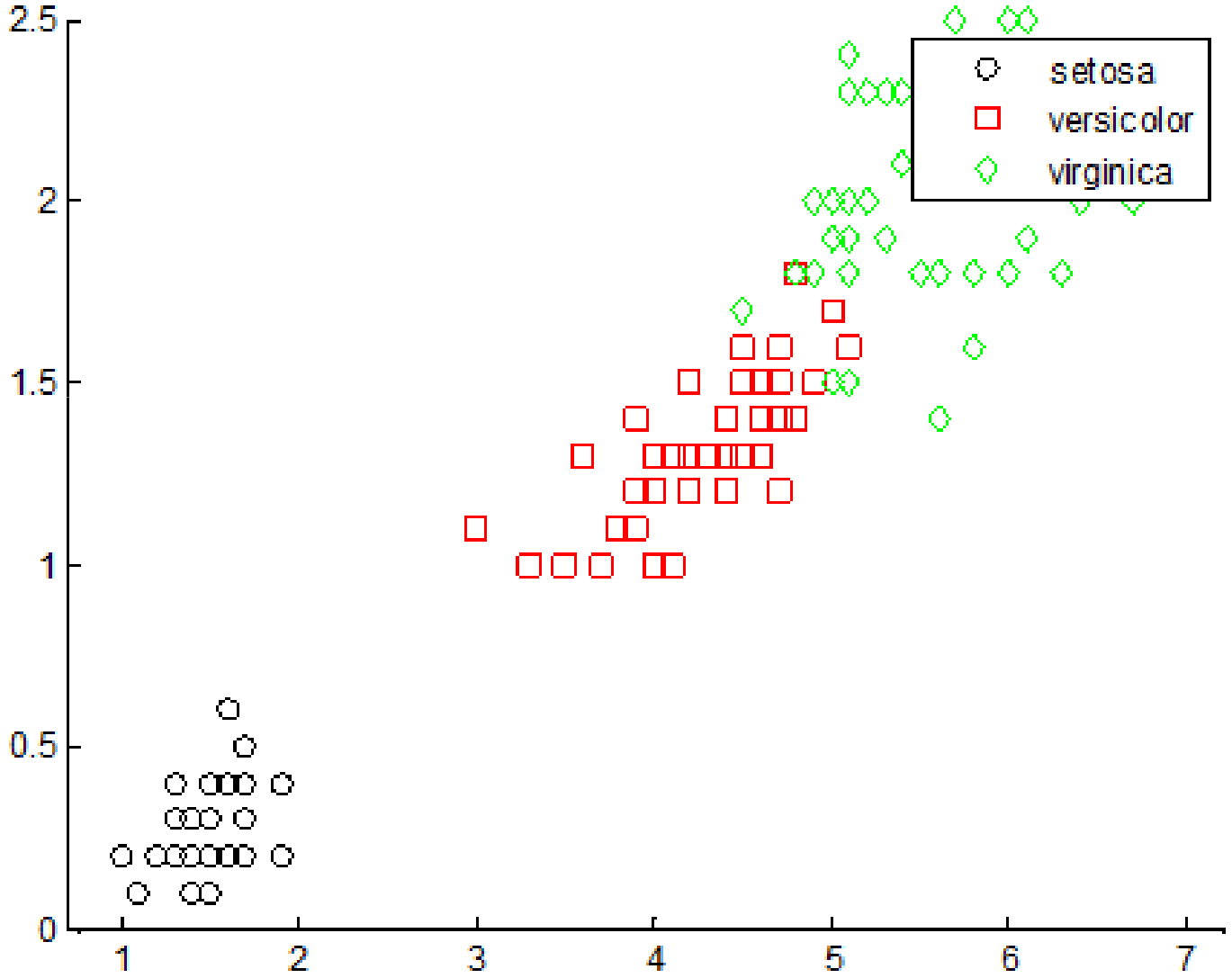
încărcând variabila **meas** (measures) o matrice numerică cu 150 linii şi 4 coloane reprezentând atributele considerate (lungimea sepalei, lăţimea sepalei, lungimea petalei, lăţimea petalei) şi variabila **species**.

Vom prezenta o vizualizare a florilor de Iris, din baza de date a lui Fischer, luând în considerare dimensiunile petalelor.

Notând cu x_1 = lungimea în cm a petalei şi x_2 = lăţimea în cm a petalei, construim vectorul $\mathbf{x} = (x_1, x_2)$, corespunzător unui iris. Exemplarele de flori vor fi reprezentate în spaţiul bidimensional al atributelor.
Să reţinem că lungimea, respectiv lăţimea unei petale diferă şi în cazul aceluiaşi tip de Iris.

Vom specifica atributele folosite meas(:,3), meas(:,4). Impunem ca toate florile, în funcţie de specia căreia îi aparţin, să fie reprezentate prin cerc (setosa), pătrat (versicolor) şi romb (virginica)

```
>> load fisheriris
>> gscatter(meas(:,3), meas(:,4), species,'krg','osd')
```



2. Exemplu

O baza de date medicală conţine datele unor pacienţi, atributele acestora fiind: numărul de consultaţii pe an la care a fost supus pacientul - nr. cons (dată *numerică discretă*), sex-S (dată *nominală*), vârsta - V, indicele masei ponderale - IMC. Glicemia (glyc), colesterolul (Ch), trigligeride(trigl) (date *numerice continue*), gradul fumatului- gr F (dată *ordinală*) şi cele două categorii cărora le aparţin: pacient hipertensiv- Ω_1 sau pacient sănătos - Ω_2 (dată *nominală*).

Prezentăm 3 linii din această bază, corespunzătoare a 3 pacienţi.

nr cons	S	V	IMC	glyc	Ch	trigl.	gr F	clasa
9	F	60	30	92	286	349	fost	Ω_1
2	F	40	23,2	79	180	145	nefumător	Ω_2
5	M	55	25.1	128	230	210	amator	Ω_1

Vom face următoarele codificări:

- pentru feminin/F-0, pentru masculin/M-1;
- gradul fumatului: nefumător - 0, fost fumător -1, fumător amator -2, fumător înrăit - 3;
- pacient hipertensiv Ω_1 -1, pacient sănătos Ω_2 - 0

obţinând astfel următoarea matrice ataşată datelor avute:

$$\begin{pmatrix} 0 & 9 & 60 & 30 & 92 & 286 & 349 & 1 & 1 \\ 0 & 2 & 40 & 23.2 & 79 & 180 & 145 & 0 & 0 \\ 1 & 5 & 55 & 25.1 & 128 & 230 & 210 & 2 & 1 \end{pmatrix}$$

Date cenzurate

Există situaţii în care o anumită dată există, dar din diferite motive nu poate fi bine precizată, situaţii în care considerăm că datele sunt *cenzurate*.

De exemplu în *Analiza supravieţiurii,* tehnică statistică clasică, care studiază dinamica timpului de supravieţiure după o anumită operaţie sau tratament, o parte din subiecţii incluşi în lotul de studiu decedează în perioada de observaţie, dar o altă parte supravieţiuesc acestei perioade sau se retrag benevol.

Alt exemplu când se efectuează anumite măsurători şi aparatul de măsura nu poate înregistra valori inafara scalei sale.

Procesarea datelor în cursul unei analiza statistice impune ca absolut necesară existența așa numitei *variabilități* a datelor.
Prin *variabilitate* se înțelege orice modificare care are loc într-o mulțime de date, indiferent de tipul acestora.

Nu se poate face analiză statistică pe variabile care sunt constante
Cu cât variabilitatea datelor este mai mare cu atât analiza statistică va da rezultate mai consistente.

Definirea datelor in context probabilist

Avem la dispoziţie o anumita mulţime de obiecte/subiecţi (o aşa numită *populaţie statistică*) şi suntem interesaţi de analiza principalelor lor caracteristici/atribute, care reprezintă date statistice.
Din punct de vedere matematic *data* este o aplicaţie definită pe mulţimea ce reprezintă populaţia şi cu valori într-o anumită mulţime, ce depinde de data respectivă

Considerând un câmp de probabilitate (Ω, Σ, P), unde Ω este chiar populaţia considerată, Σ este o σ-algebră de părţi a lui Ω, data X a populaţiei statistice Ω, dacă este numerică, este o variabilă aleatoare pe câmpul de probabilitate (Ω, Σ, P), cunoscută şi sub numele de variabilă statistică.,
În cazul în care X nu ia valori numerice, pe baza unor echivalări numerice ale acestor valori (codificări), putem considera X ca find o variabilă aleatoare.

3. Exemplu

Considerând ca populaţie Ω o clasă de elevi, considerăm o dată X, referitoare la această populaţie, reprezentată prin înălţimea elevilor.
Pentru elevul A ce are Înălţimea de 1,70, avem $X(A) = 1.70$

Statistica descriptiva

Descrierea statistică constă într-o mulţime de diverse metode ce au ca scop rezumarea unui număr mare de observaţii privind datele, punând astfel în evidenţă principalele lor caracteristici.
Există două mari abordări a descrierii statistice a datelor:

- Determinarea unor parametri numerici, interesul fiind focalizat pe proprietăţile lor matematice;
- Diverse reprezentări grafice simple ale datelor, a căror interpretare nu este dificilă, fiind de cele mai multe ori foarte sugestivă. Din punct de vedere strict informaţional interpretare este totuşi limitată.

Pentru a rezuma un mare număr de observaţii privind un set deobiecte/instanţe, punând în evidenţă caracteristicile principale ale atributelor lor, vom utiliza reprezentări numerice care cuprind principalele caracteristici statistice ale datelor respective. (media, mediana, deviaţia standard, modul etc.).

Este vorba de reprezentarea variabilităţii unor date. Această variabilitate poate fi una cu cauze cunoscute, o variabilitate ‚deterministă' care este descrisă statistic pentru a o pune şi mai bine în evidenţă şi a o cuantifica precis, sau poate fi o variabilitate cu cauze doar bănuite sau chiar necunoscute – variabilitatea ‚aleatorie' – şi care, folosind statistica, se speră a fi clarificată cauzal.

Pentru a procesa datele în cursul unei analiza statistice este absolut necesară existența așa numitei variabilități a datelor

Un obiect/o instanță corespunde din punct de vedere probabilist unei variabile aleatoare multidimensionale adică un vector aleator *n*-dimensional, subiect al analizei statistice multivariate.

Fiecare componentă a instanței, reprezentând un anumit atribut, este privită la rândul său ca o variabila aleatoare.

Luăm în considerație în acest caz parametrii numerici (statistici) ce caracterizează o variabilă aleatoare, parametri deosebit de utili în descrierea dinamicii acesteia.

În cele ce urmează, vom considera că fiecare atribut x al unei instanţe reprezintă o valoare pe care o poate lua variabila aleatoare generatoare X, corespunzătoare acelui atribut.

Reamintim că, pentru variabila aleatoare X, funcţia F_X: $\mathbf{R} \to [0, 1]$, definită de:

$$F_X(x) = P\{X < x\} = P\{\omega \in \Omega; X(\omega) \in (-\infty, x)\},$$

se numeşte *funcţia de repartiţie* (*repartiţia de probabilitate*), unde (Ω, Σ, P) este un spaţiu (câmp de probabilitate).

Funcţia de repartiţie a variabilei statistice *X*

În cazul unei analize statistice, unei variabile aleatoare X îi corespunde variabila statistică corespunzătoare, notată în mod firesc tot X. La fel ca în cazul unei variabile aleatoare, şi în cazul unei variabile statistice putem defini noţiunea de funcţie de repartiţie.

Astfel, prin *funcţia de repartiţie* sau *funcţia cumulativă* (*frecvenţa cumulată*) a variabilei statistice X (corespunzătoare variabilei aleatoare X), definită de seria statistică asociată (eşantionul corespondent) $\{x_i\}_{i = 1,\ldots, n}$, înţelegem aplicaţia F: $\mathbf{R} \to [0, 1]$, dată de:

$$F(x) = \frac{f_x}{n}, x \in R,$$

unde f_x reprezintă numărul observaţiilor x_i strict mai mici decât x.

Cuantila

Definim *cuantila* (*quantile* - Kendall, 1940) de ordin α, $0 < \alpha < 1$, a variabilei statistice X, ca fiind numărul q_α cu proprietatea $F(q_\alpha) = \alpha$, adică $f_{q_\alpha} = \alpha \cdot n$, unde f_{q_α} reprezintă numărul observaţiilor strict mai mici decât q_α.

Din punctul nostru de vedere, al explorării datelor, cuantilele implică practic divizarea în $1/\alpha$ submulţimi de mărimi egale a unei mulţimi ordonate de date, ele reprezentând de fapt tocmai frontierele (pragurile) dintre submulţimile consecutive.

În acest context, o *percentilă* (*percentile* -Galton, 1885) reprezintă oricare dintre cele 99 de valori care împart o mulţime ordonată de date în 100 de submulţimi de mărimi egale, consecutive.
De exemplu a 50-a percentilă împarte setul de date în 50% date dedesubtul său şi 50% date deasupra sa.

Astfel:

- *decilele* (*deciles*) reprezintă cele 9 valori care împart o mulţime ordonată de date în 10 submulţimi consecutive de mărimi egale;

- *cuartilele* (*quartiles* - Galton, 1882) reprezintă cele trei valori notate Q_1, Q_2, Q_3, care împart o mulţime ordonată de date în 4 submulţimi consecutive de mărimi egale. Cele mai utilizate cuantile sunt cuartilele Q_1, Q_2, şi Q_3. Astfel, prima cuartilă Q_1 (cuartila inferioară – a 25-a percentilă) are sub ea 25% din date şi deasupra 75% din date; a doua cuartilă Q_2 are sub ea 50% din date şi deasupra tot 50% din date; a treia cuartilă Q_3 (cuartila superioară – a 75-a percentilă) are sub ea 75% din date şi deasupra 25% din date.

Calculul cuantilelor in Matlab- *Statistics Toolbox*

>> Y = quantile(X,p)

unde

Y returnează cuantilele corespunzătoare valorilor lui X,

p este scalarul sau vectorul valorilor probabilităţii şi anume:

dacă dorim să calculăm

- percentilele, luăm p=0.01:0.01:0.99;
- decilele luăm p=0.1:0.1:0.9;
- cuartilele, luăm p=0.25:0.25:0.75.

Dacă X este un vector, Y va avea dimensiunea lui p şi componenta Y(i) va conţine cuantila p(i).

Dacă X este o matrice, atunci linia a i-a conţine cuantilele p(i) ale fiecărei coloane a lui X

4. Exemplu: decilele si quartilele corespunzatoare dimensiunilor sepalelor si petalelor florilor de iris

```
>> load fisheriris
>> X=[meas(:,1) meas(:,2) meas(:,3) meas(:,4)];
>> y10=quantile(X,0.1:0.1:0.9)
y10 =
    4.8000    2.5000    1.4000    0.2000
    5.0000    2.7000    1.5000    0.2000
    5.2500    2.8000    1.7000    0.4000
    5.6000    3.0000    3.9000    1.1500
    5.8000    3.0000    4.3500    1.3000
    6.1000    3.1000    4.6500    1.5000
    6.3000    3.2000    5.0000    1.8000
    6.5500    3.4000    5.3500    1.9000
    6.9000    3.6500    5.8000    2.2000
>> y1=quantile(X,0.25:0.25:0.75)
y1 =
    5.1000    2.8000    1.6000    0.3000
    5.8000    3.0000    4.3500    1.3000
    6.4000    3.3000    5.1000    1.8000
```

Valorile tipice corespunzătoare unei analize a datelor

- măsuri tipice ale tendinţei centrale (locaţiei): *modul* (*mode*), *mediana* (*median*), *media* (*mean*, *arithmetic mean*, *average*), *media geometrică* (*geometric mean*) şi *media armonică* (*harmonic mean*);
- măsuri tipice ale împrăştierii (deviaţiei): *dispersia* (*variance*) şi *deviaţia standard/abaterea medie pătratică* (*standard deviation*);
- măsuri tipice ale formei repartiţiei: *asimetria* (*skewness*) şi *excesul* (*kurtosis*).

Media

Media este cel mai comun parametru ce măsoară „tendinţa centrală” a unei serii statistice, reprezentând practic media aritmetică a tuturor observaţiilor:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

In Matlab, pentru calcularea mediei, utilizăm funcţia

```
>>M=mean(X)
```

unde vectorul X secvenţa de date a cărei medie vrem să o calculăm.

Dacă X este o matrice mean(X) consideră coloanele lui X ca fiind vectori, returnând un vector linie ale cărui componente sunt mediile vectorilor coloană

5. Exemplu

Considerăm următoarea secvenţă de date reprezintă valorile înălţimii băieţilor dintr-o clasă de gimnaziu:
{ 1,40; 1,37; 1.57; 1.46; 1.49; 1.46; 1.39; 1.55; 1.29; 1.58; 1.50; 1.33}
Să calculăm înălţimea medie a băieţilor din această clasă:

```
>> X1=[1.40, 1.37, 1.57, 1.46, 1.49, 1.46, 1.39, 1.55, 1.29, 1.58, 1.50, 1.33 ];
>>m1= mean(X1)
m1 =
    1.4492
```

6. Exemplu

Vom da un exemplu de calcul al mediei pentru date discrete, considerând un tabel de distribuire (repartiție) a datelor, în care valorile x_i au frecvențele de apariție n_i:

Un medic de familie are în îngrijire 140 de copii. Prezentam situaţia numărului de vizite ale copiilor la cabinet pe durata unui an:

Nr vizite	0	1	2	3	4	5	6	7
Nr copii	2	38	27	45	18	5	4	1

Să calculăm media numărului de vizite efectuate de un copil, într-un an.

```
>> a=[0 0];
>> b=[1];for i=2:38 b=[1 b];end
>> c=[2];for i=2:27 c=[2 c];end
>> d=[3];for i=2:45 d=[3 d];end
>> e=[4];for i=2:18 e=[4 e];end
>> f=[5];for i=2:5 f=[5 f];end
>> g=[6];for i=2:4 g=[6 g];end
>> h=[7];
>>> X2=[a b c d e f g h]
>> m=mean(X2)
m =
   2.5357
```

Cu toate că media este o caracteristică foarte sugestivă a datelor pe care le reprezintă, existenţa valorilor extreme (*outliers*) îi pot perturba serios capabilitatea de ilustrare a datelor.

Considerăm secvenţă de date din exemplul nr 5, la care adaugăm înălţimea unui alt băiat care suferă de o uşoară forma de nanism:

```
>> X=[1.40, 1.37, 1.57, 1.46, 1.49, 1.46, 1.39, 1.55, 1.29, 1.58, 1.50, 1.33 1.05];
>>m= mean(X)
m =
   1.4185
```

Remarcăm faptul că media m=1.4185 diferă de cea calculată în exemplul nr 5, m1= 1.4492, valori fiind suficient de distincte datorită influenţei unei singure date.

Mediana

Pentru a evita asemenea situaţii, în astfel de cazuri se utilizează mediana în locul mediei. Astfel, *mediana* este definită ca numărul real care împarte în două efective egale seria statistică dată, observaţiile fiind ordonate crescător, cu alte cuvinte mediana este chiar a doua cuartilă Q_2. Formal, mediana este dată de:

$$P\{X \le Q_2\} = P\{X > Q_2\} = 1/2 .$$

In Matlab, pentru calcularea medianei, utilizăm funcţia

```
>>M=median(X)
```

unde vectorul X este secvenţa de date a cărei mediană vrem să o calculăm.

Dacă X este o matrice median (X) consideră coloanele lui X ca fiind vectori, returnând un vector linie ale cărui componente sunt valorile medianelor.

Folosind cuantilele avem:

```
>>M =quantile(X,0.50)
```

Exemple

In exemplul nr 5 avem:

median(X1) = 1.4600 şi mean(X1) = 1.4492

median(X) = 1.4600 şi mean(X) = 1.4185

In exemplul nr 6 avem:

median(X2)=3 şi m=mean(X2)=2.5357

7. Exemplu

O fabrică produce bare de oțel de diametru 20mm. Datorită unor reclamaţii asupra calităţii produselor s-a ales un eşantion măsurându-se diametrul fiecărei bare din acest eşantion; rezultatele sunt prezenrtate în următoarea secvenţă de date:

{ 19.9 19.8 20.1 19.9 19.7 20.1 20 19.6 19.7 20.1 20.4 20 19.9 19.8 20.2 20 19.8 19.6 29.9 20.3}

Vom calcula media, respectiv mediana valorilor diametrului barelor din acest eşantion

```
>> X3=[19.9 19.8 20.1 19.9 19.7 20.1 20 19.6 19.7 20.1 20.4 20 19.9 19.8 20.2
20 19.8 19.6 29.9 20.3]
>> m=mean(X3)
m =
   20.4400
>> M=median(X3)
M =
   19.9500
```

Dacă efectivul seriei statistice este un număr impar $n = 2k + 1$, atunci mediana este a $(k + 1)$-*a* valoare a seriei, iar dacă efectivul este un număr par $n = 2k$, atunci mediana se înlocuieşte cu *intervalul median* dat de valorile a *k*-a şi a $(k + 1)$-*a* (mediana se poate considera astfel ca mijlocul acestui interval –media aritmetică a capetelor sale).

Mediana mai este folositoare atunci când există posibilitatea ca unele valori extreme ale seriei statistice să fie cenzurate. Atunci când există observaţii care nu sunt suficient de exact precizate, nu putem folosi media, înlocuind-o prin mediană dacă **avem valori exacte pentru mai mult de jumătate din observaţii** (cazul măsurătorilor fizico-chimice, când există valori în afara scalei normale a aparatului de măsură).

Ambele măsuri sunt la fel de eficiente şi, cu toate că media este mai frecvent folosită decât mediana, aceasta din urmă poate fi mai valoroasă în anumite circumstanţe.

În exemplul nr 5, al setului de date privind înălţimea unor elevi, să calculăm medianele corespunzătoare celor două cazuri (cu şi fără valoarea extremă). sunt egale, ceea ce dovedeşte că mediana nu este influenţată semnificativ de valori extreme.

Modul (modă)

Modul (*modă*) este o altă măsură a tendinţei centrale care reprezintă cea mai frecventă valoare a secvenţei, des utilizată în cazul datelor categoriale. Este foarte probabil ca modul sa nu fie unic, deoarece pot apărea mai multe valori ale seriei statistice ce au aceeaşi frecvenţă de aparţie. Este vorba de date *plurimodale*

Dacă datele sunt grupate în clase adică valorile aparţin unor intervale, vom numi *clasă modală* orice clasă corespunzând unui maxim al frecvenţei.

In Matlab, pentru calcularea modului, utilizăm funcţia

```
>>M=mode(X)
```

unde vectorul X este secvenţa de date al cărei mod vrem să-l calculăm (vrem să ştim care date sunt cele mai frecvente). Dacă X este o matrice mode (X) consideră coloanele lui X ca fiind vectori, returnând un vector linie ale cărui componente sunt valorile modelor.

Exemple

În exemplul nr 5 avem:

```
>>mode(X1)
ans =
    1.4600
>> mode(X)
ans =
    1.4600
```

În exemplul nr 7 avem:

```
>> mode(X3)
ans =
   19.8000
```

Media geometrica

Media geometrică este dată de formula: $\sqrt[n]{x_1 \cdot x_2 \cdot \ldots x_n}$

Media geometrică este utilizată cu precădere în cazul măsurătorilor cu scală neliniară (e.g. în psihometrie, unde rata intensităţii unui stimul este adesea o funcţie logaritmică în raport cu intensitatea, caz în care se foloseşte media geometrică şi nu media aritmetică).

In Matlab, pentru calcularea mediei geometrice, utilizăm funcţia

```
>>M=geomean(X)
```

unde vectorul X este secvenţa de date a cărei medie geometrică vrem să o calculăm.

Dacă X este matrice, geomean(X) este un vector linie, ale cărui componente sunt mediile geometrice ale coloanelor.

Media armonica

Media armonică, dată de formula: $\frac{n}{\sum_{1}^{n} \frac{1}{x_i}}$

se utilizează câteodată în cazul determinării mediei frecvenţelor.

In Matlab, pentru calcularea mediei armonice, utilizăm funcţia

>>M=harmmean(X)

unde vectorul X este secvenţa de date a cărei medie armonică vrem să o calculăm.

Dacă X este matrice, geomean(X) este un vector linie, ale cărui componente sunt mediile armonice ale coloanelor.

Calculăm pentru secvenţele de date din exemplul nr 5 (X şi X1), respectiv pentru secvenţa din exemplul nr 7 (X3) valorile pentru medie, modă, mediană, medie geometrică şi medie aritmetică.

Principalii parametrii ai tendinţei centrale (cu două zecimale)

Secvenţa de date	medie	mediana	modă	medie geometrică	media armonică
X	1.45	1.46	1.46	1.45	1.44
X1	1.42	1.46	1.46	1.41	1.40
X3	20.44	19.95	19.80	20.35	20.28

Din tabelul de mai sus se observă că în toate cele trei cazuri repartiţia este suficient de „simetrică”, parametrii tendinţei centrale fiind apropiaţi ca valoare.

Dispersia

O altă abordare privind analiza datelor este reprezentată de măsurarea împrăştierii (*dispersiei*) faţă de medie, adică **măsurarea distanţei fiecărei valori a seriei statistice faţă de media eşantionului**.
Plecând de la cazul probabilist clasic al dispersiei, vom defini *dispersia* sau *varianţa* (*variance*) - termen introdus de Fisher, 1918 - corespunzătoare unei serii statistice $\{x_i\}_{i=1,\dots,n}$, cu ajutorul formulei:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \ ,$$

unde m este media cunoscută a variabilei statistice (a populaţiei originare, cu alte cuvinte).

In Matlab, pentru calcularea dispersiei, utilizăm funcţia

```
>>V=var(X)
```

unde vectorul X este secvenţa de date a cărei dispersie vrem să o calculăm.

Dacă X este matrice, var(X) este un vector linie, ale cărui componente sunt dispersiile vectorilor coloană.

De obicei, considerăm că seria statistică cu care lucrăm nu reprezintă toată populaţia ci este doar un eşantion al ei şi astfel media m nu este cunoscută ci putem calcula doar media eşantionului $\bar{x}$.
Vom folosi în locul formulei de mai sus o formulă de aproximaţie (o estimaţie) a dispersiei, înlocuind media m cu media seriei $\bar{x}$ şi împărţind prin $(n-1)$ în loc de n, deci:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Remarcăm aici că, pentru serii statistice de dimensiuni mari, diferenţa dintre valoarea dată de formula de mai sus şi formula clasică:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

este neglijabilă.

Deviaţia standard

De multe ori, este preferabil ca în loc de dispersie să folosim o mărime care este măsurată cu aceeaşi unitate ca şi seria statistică, şi anume *deviaţia standard* (sau *abaterea medie pătratică*), dată de formula:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

In Matlab, pentru calcularea deviaţiei standard, utilizăm funcţia

```
>>s=std(X)
```

unde vectorul X este secvenţa de date a cărei deviaţie standard vrem să o calculăm.
Dacă X este matrice, std(X) este un vector linie, ale cărui componente sunt deviaţiile standard a vectorilor coloană.

Interval de încredere

Deviaţia standard este folosită în statistica descriptivă mai ales pentru definirea unor intervale în care se găsesc marea majoritate a observaţiilor. Astfel, în cazul unor repartiţii rezonabil de simetrice, marea majoritate a observaţiilor ce compun seria statistică (aproximativ 95% din ele) se găsesc în intervalul definit de:

$$\textit{medie} \pm 2 \times \textit{deviaţia standard},$$

numit *interval de încredere* (*confidence interval*).

Cele spuse mai sus nu mai au relevanţă dacă nu avem repartiţii relativ simetrice.

Exemple

Vom ilustra măsurile tipice împrăştierii şi formei repartiţiei în cazul secvenţelor de date din exemplul nr 5 (X şi X1), respectiv pentru secvenţa din exemplul nr 7 (X3) .

```
>> X=[1.40, 1.37, 1.57, 1.46, 1.49, 1.46, 1.39, 1.55, 1.29, 1.58, 1.50, 1.33 ];
>> m=mean(X); v=var(X);s=std(X); [v, s, m-2*s, m+2*s]
ans =
   0.0089    0.0944    1.2603    1.6380
```

Principalii parametrii ai împrăştierii (dispersia, deviaţia standard, intervalul de încredere pentru medie):

Dispersia	Deviaţia standard	Intervalul de încredere
0.0089	0.0944	(1.2603, 1.6380)

```
>> X1=[1.4,1.37, 1.57, 1.46, 1.49, 1.46, 1.39, 1.55, 1.29, 1.58, 1.50, 1.33, 1.05 ];
>> m1=mean(X1); v1=var(X1);s1=std(X1); [v1, s1, m1-2*s1, m1+2*s1]
ans =
   0.0204    0.1429    1.1326    1.7043
```

Principalii parametrii ai împrăştierii (dispersia, deviaţia standard, intervalul de încredere pentru medie):

Dispersia	Deviaţia standard	Intervalul de încredere
0.0204	0.1429	(1.1326 1.7043)

```
>> X3=[19.9 19.8 20.1 19.9 19.7 20.1 20 19.6 19.7 20.1 20.4 20 19.9 19.8 20.2
20 19.8 19.6 29.9 20.3];
>> m3=mean(X3); v3=var(X3);s3=std(X3); [v3, s3, m3-2*s3, m3+2*s3]
ans =
   5.0057    2.2373   15.9653   24.9147
```

Principalii parametrii ai împrăştierii (dispersia, deviaţia standard, intervalul de încredere pentru medie):

Dispersia	Deviaţia standard	Intervalul de încredere
5.0057	2.2373	(15.9653 24.9147)

Remarcă

Înafara parametrilor statistici amintiţi mai sus, se mai utilizează câteodată şi:

- *Domeniul* (*range*) valorilor, reprezentat de diferenţa între maximul şi minimul valorilor datelor;
- *Domeniul inter-cuartile* (*interquartiles range*), definit de diferenţa $Q_3 - Q_1$;
- Media abaterilor absolute de la medie:

$$AAD(x) = \frac{1}{n}\sum_{1}^{n}\left|x_i - \bar{x}\right|$$

- Mediana abaterilor absolute de la medie:

$$MAD(x) = mediana\ \{|x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_n - \bar{x}|\}$$

Reprezentarea grafică a unei mulţimi de date

Reprezentarea grafică elementară a datelor înseamnă conversia datelor în format vizual sau tabular simplu, astfel încât să se poată analiza şi raporta rapid caracteristicile datelor, precum şi relaţiile existente între atribute.

Reprezentările grafice aferente unui set de date (atribute ale unei instanţe în cazul de faţă) depind de natura atributelor: calitative sau cantitative.

Reprezentarea circulară (pie) a datelor calitative

Datele calitative pot fi reprezentate grafic cu ajutorul diferitelor diagrame formate din bastoane verticale sau orizontale, cercuri, etc., bi- sau tri-dimensionale, plecând de la partiţia setului de date pe care atributul o induce.

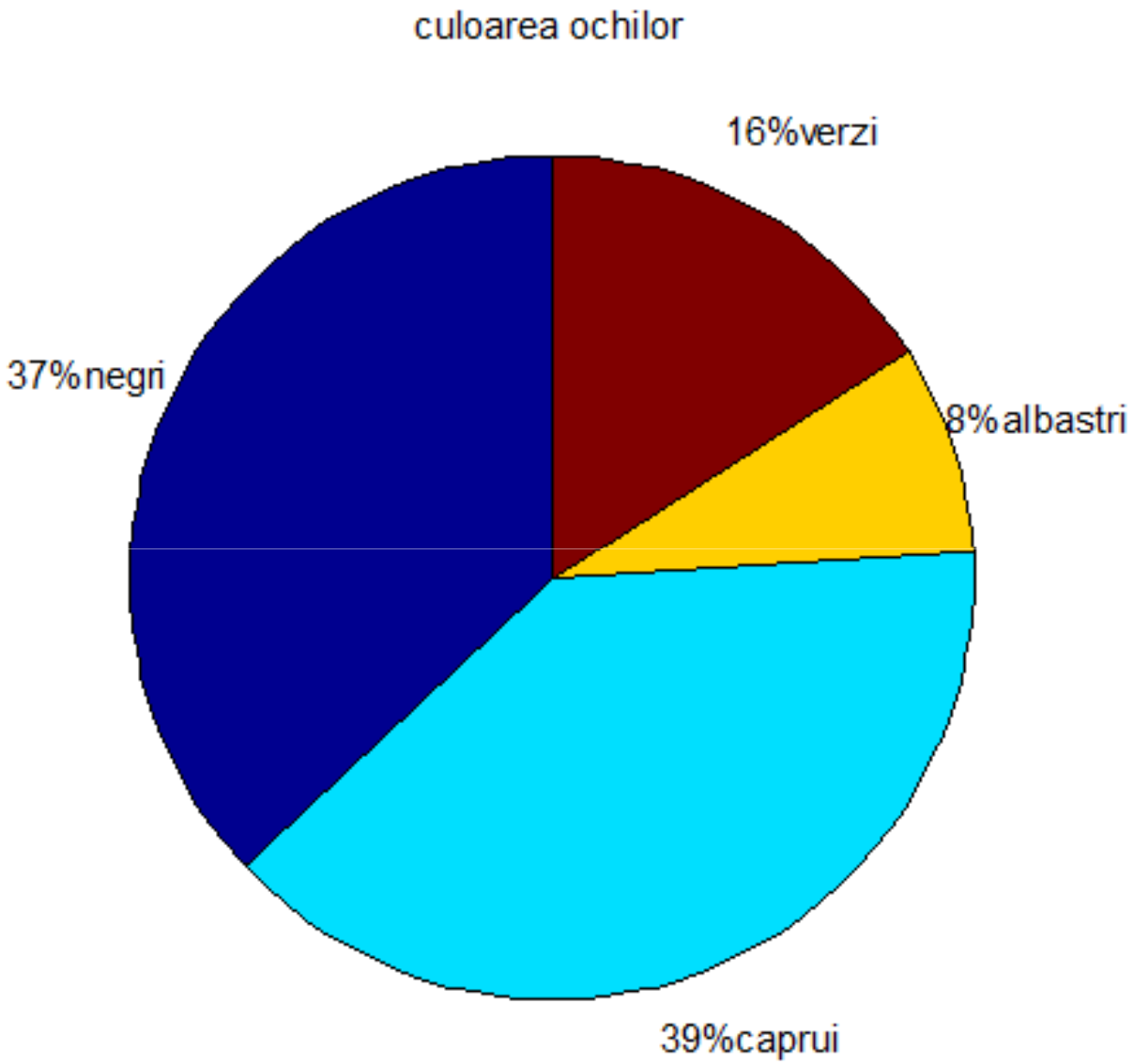
În cazul utilizării *reprezentării circulare* (numită şi *'pie' ~ plăcintă* în engleză), întreaga mulţime de obiecte este reprezentată de cerc, fiecare atribut al obiectelor este reprezentat printr-un sector circular a cărui suprafaţă este proporţională cu numărul obiectelor având atributul respectiv (sau cu procentajul corespunzător).

8. Exemplu

În Matlab pie(X) realizează o reprezentare circulară bidimensională, pe baza datelor din vectorul X, vector ce are drept componente procentajele corespunzatoare unor atribute Fiecare atribut va fi reprezentat de un sector circular.

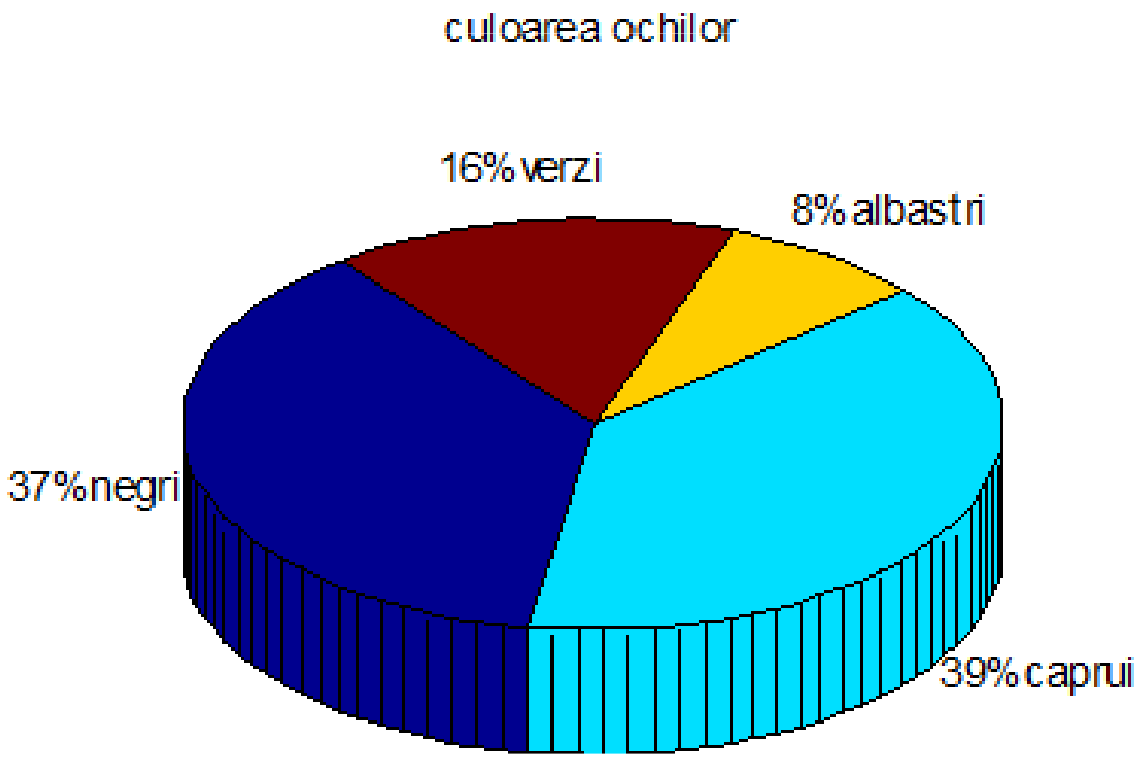
Suntem interesaţi de culoarea ochilor indivizilor - atribut calitativ, ce aparţin unei anumite populaţii. Presupunem că mulţimea culorilor ochilor este *Culoare_ochi* = {negru, albastru, verde, căprui} şi că din studiul efectuat a rezultat că 37% din populaţie are ochii negri, 39% are ochii căprui, 8% are ochii albaştri şi 16% are ochii verzi, aşadar X=[0.37 0.39 0.08 0.16].

```
>> X=[0.37 0.39 0.08 0.16];
>>pie(X,{'37%negri','39%caprui', ' 8%albastri', ' 16%verzi'})
```


culoarea ochilor
16%verzi
8%albastri
39%caprui
37%negri

În Matlab pie3(X) realizează o reprezentare circulară tridimensională, reprezentare ce nu furnizează informație suplimentară fața de cazul bidimensional;

```
>>X=[0.37 0.39 0.08 0.16];
>>pie3(X,{'37%negri','39%caprui', ' 8%albastri', ' 16%verzi'})
```



9. Exemplu

Vom folosi funcţia pie pentru a vizualiza contribuţia ce o aduc trei produse în totalul vânzarilor. Coloanele matricei X conţin vânzările anuale pentru fiecare produs de-a lungul a 5 ani:

```
>>X = [19.3 22.1 51.6;
    34.2 70.3 82.4;
    61.4 82.9 90.8;
    50.5 54.9 59.1;
    29.4 36.3 47.0];
```

Vom calcula vânzările totale pentru fiecare produs în cei 5 ani utilizând funcţia sum.

```
>>x = sum(X);
            194.8000  266.5000  330.9000
```

Folosind argumentul **explode**, putem pune în evidenţă sectorul circular care are cea mai mare contribuţie.
explode este un vector ale cărui componente nenule corespund unui anumit sector circular.
Pentru început vom crea un vector de componente nule:

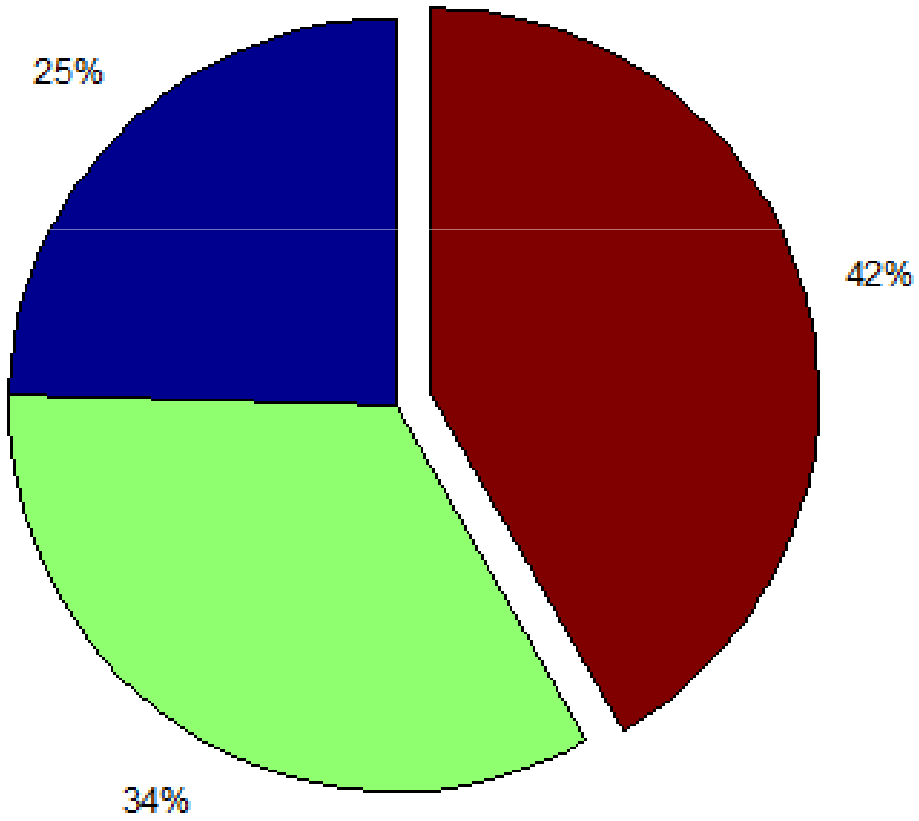
```
>> explode = zeros(size(x))
explode =
    0    0    0
```

Vom găsi produsul care îşi aduce cea mai mare contribuţie şi vom atribui componentei corespunzătoare a argumentului **explode** valoarea 1:

```
>> [c,offset] = max(x)
c =
  330.9000
offset =
    3
>> explode(offset) = 1
explode =
    0    0    1
```

Pentru a crea reprezentarea circulară vom scrie:

```
>>h = pie(x,explode);
```



Reprezentarea prin diagrame cu bastoane (bar) a datelor calitative

Cealaltă modalitate de reprezentarea a unui atribut calitativ – cel al *diagramelor cu bastoane*(*bar graph*) – se raportează la o vizualizare într-un sistem de axe: pe axa absciselor apar atributele iar pe axa ordonatelor apare numărul obiectelor cu atributul respectiv (sau procentajul corespunzător).

În Matlab pentru a vizualiza bidimensional diagramele cu bastoane folosim funcţia

```
>>bar(Y)
```

unde Y este vectorul ale cărui componente reprezintă numărul obiectelor cu atributul respectiv sau procentajul corespunzător
Funcţia

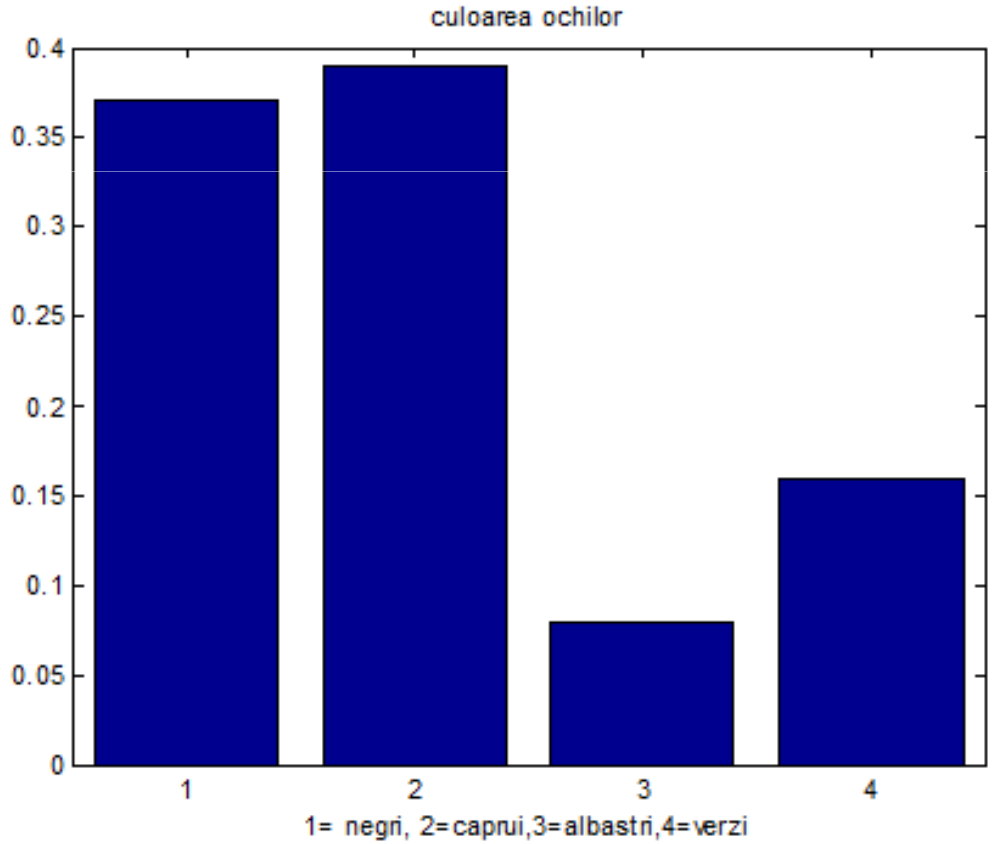
```
>>bar3(Y)
```

desenează tridimensional fiecare element ca un paralelipiped, cu elementele distribuite pe axa Oy

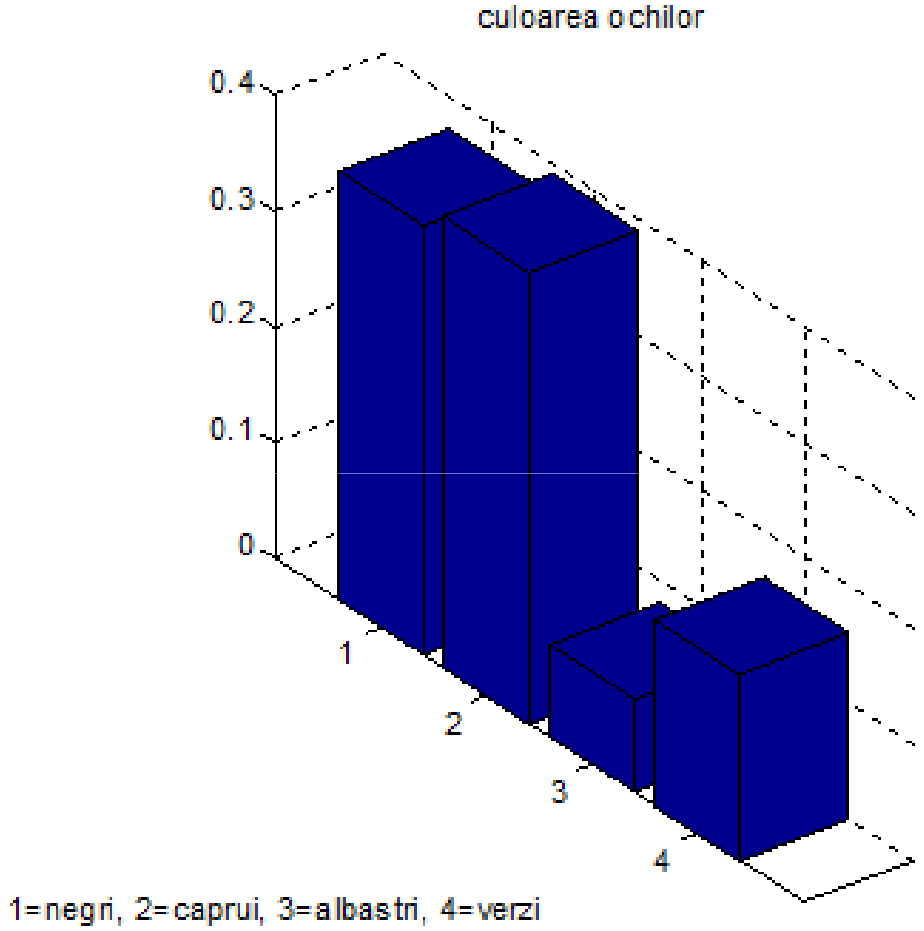
Exemplu

Reluăm exemplul nr 8, lucrând cu reprezentarea prin bastoane a frecvenţei de apariţie a diferitelor tipuri de culori ale ochilor în populaţia respectivă.

```
>> Y=[0.37 0.39 0.08 0.16];bar(Y)
```



```
>> Y=[0.37 0.39 0.08 0.16];
>> bar3(Y)
```



Remarcă:

În acest context se mai foloseşte pentru aceste diagrame şi termenul de histogramă, dar, reprezentarea prin histograme se referă la reprezentarea grafică a frecvenţelor tabulate – cu referire directă la datele (atributele) cantitative.
Matematic vorbind, *histograma* reprezintă aplicaţia ce produce numărul de observaţii (valori ale datelor) ce aparţin unor anumite intervale (echivalent, frecvenţa observaţiilor aparţinând intervalelor de valori).
Vizual, prin histogramă (bidimensională) se înţelege o reprezentare grafică a acestei aplicaţii, adică a repartiţiei numărului de valori (frecvenţelor) unui anumit atribut numeric în care fiecare baston (coloană) reprezintă un anumit interval de valori ale atributului iar înălţimea sa este proporţională cu numărul (frecvenţa) valorilor din intervalul respectiv. Termenul a fost introdus de Pearson -1895.

Reprezentarea grafică în cazul atributelor numerice, cantitative. Histograme

În cazul unor date numerice, întâlnim cele două moduri de reprezentare grafică prin *histograme*, corespunzătoare felului datelor: discrete sau continue.

În cazul datelor *discrete*, reprezentarea grafică este asemănătoare cazului datelor calitative, cu toate că există diferența conceptuală subliniată anterior. Dacă considerăm diagramele cu bastoane, lungimea acestora are o semnificație numerică precisă. Concret pe axa absciselor sunt reprezentate valorile variabilei (datei) discrete, în timp ce pe axa ordonatelor este reprezentată frecvența relativă a apariției fiecărei valori.

Problema se complică atunci când este vorba de date numerice *continue*. Aici, pentru trasarea unei histograme, este necesară împărţirea (gruparea) datelor numerice în anumite clase (de regulă intervale), cărora să le corespundă pe cealaltă axă frecvenţa relativă de apariţie (sau numărul de observaţii), corespunzătoare fiecărei clase. În general, pe axa absciselor sunt reprezentate clasele /intervale de valori, iar pe cea a ordonatelor este reprezentat numărul corespunzător de observaţii, înălţimea dreptunghiului fiind egală cu numărul de elemente din clasa respectivă

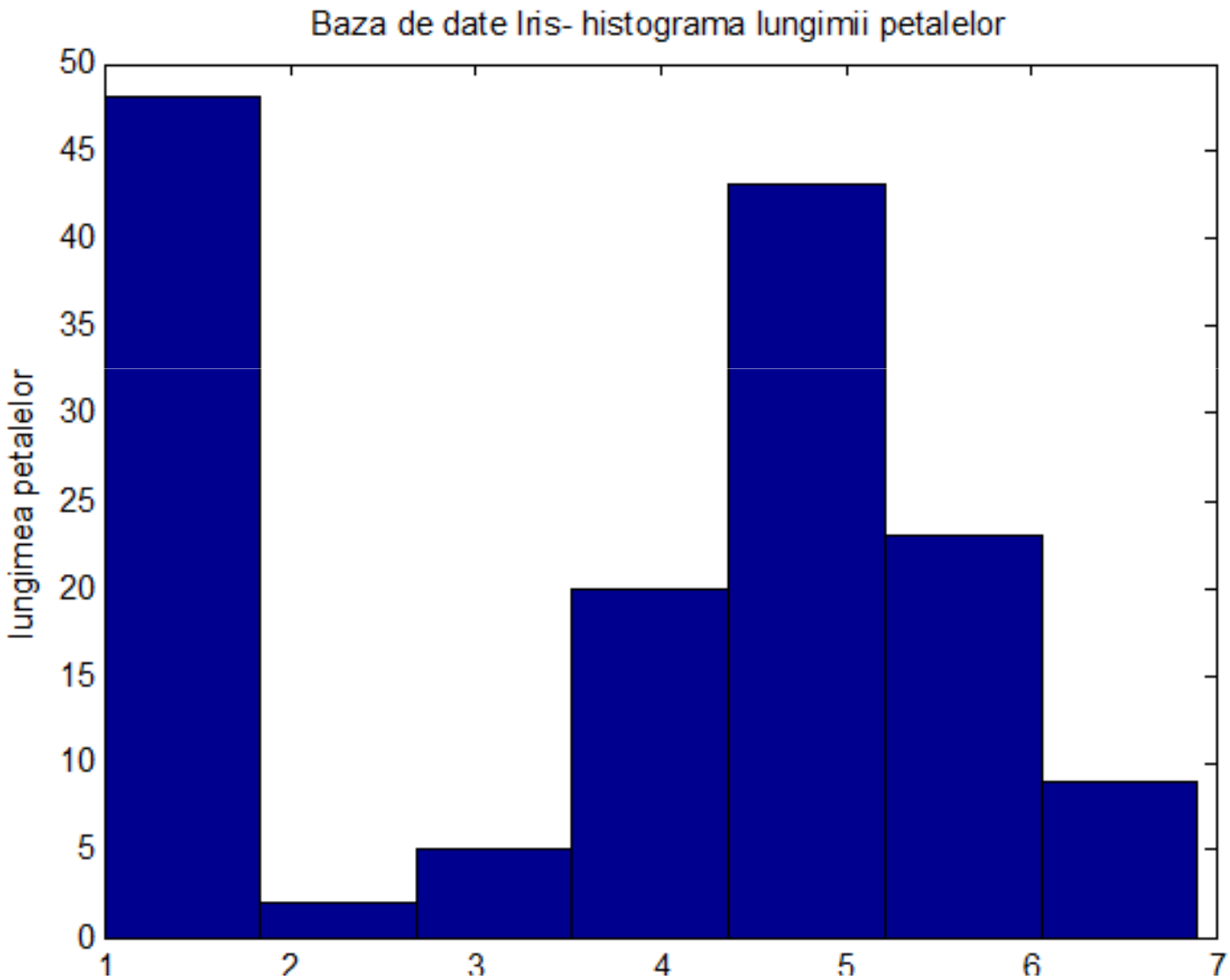
În Matlab putem utiliza funcţiile:

hist(x) care desenează o histogramă cu 10 intervale pentru elementele vectorului **x;**
hist(x,n) care desenează o histogramă cu n intervale pentru elementele vectorului **x;**

10. Exemplu

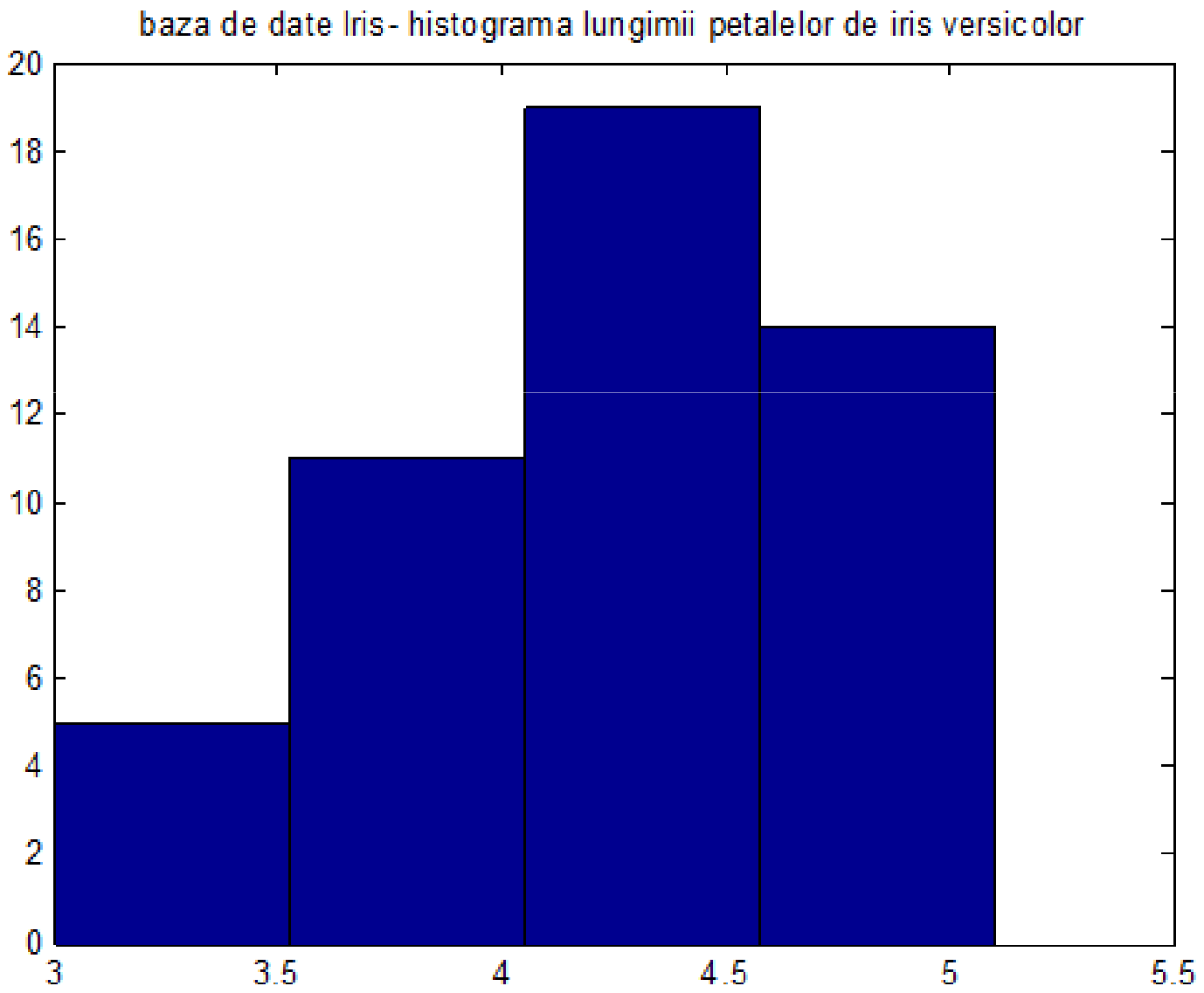
Vom desena histograma pentru vectorul ce reprezintă lungimea petalelor florilor de iris din baza de irişi a lui Fisher. Vom calcula valoarile extreme ale acestor lungimi şi în funcţie de aceasta vom decide numărul de intervale.

```
>> load fisheriris
>> x=meas(:,3);
>> min(x)
ans =
    1
>> max(x)
ans =
   6.9000
>> hist(x,7)
```


Baza de date Iris- histograma lungimii petalelor
lungimea petalelor
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7

Vom desena histograma lungimii petalelor de iris versicolor:

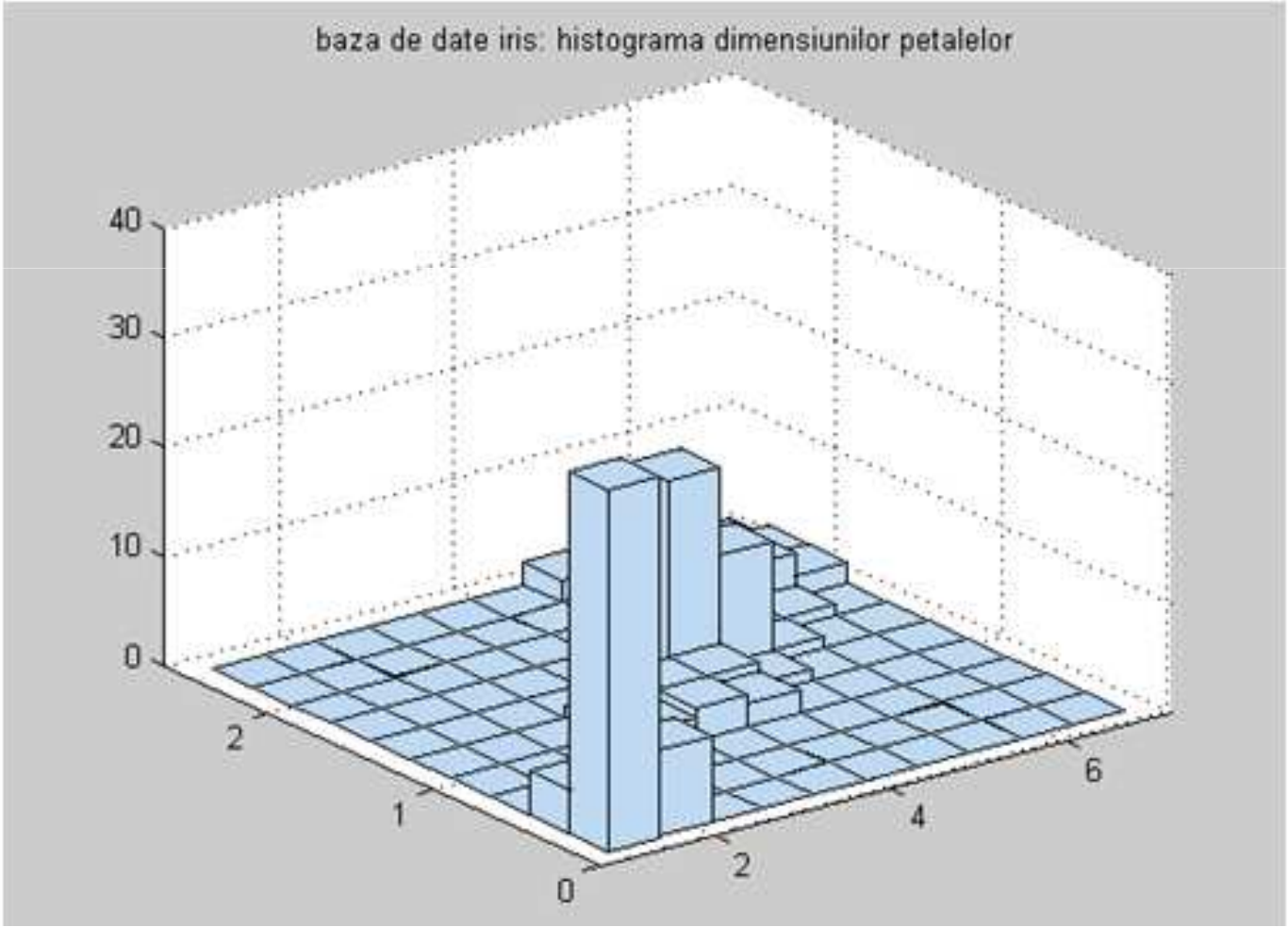
```
>> load fisheriris
>> versicolor_indices = strcmp('versicolor',species);
>> versicolor = meas(versicolor_indices,:);
>> versicolorT = versicolor(1:49,:);
>> x=versicolorT(:,3);
>> min(x)
ans =
     3
>> max(x)
ans =
   5.1000
>> hist(x,4)
```


baza de date Iris - histograma lungimii petalelor de iris versicolor
20
18
16
14
12
10
8
6
4
2
0
3
3.5
4
4.5
5
5.5

hist3(X) depozitează elementele unei matrici de tip $m \times 2$ rețea de tip 10×10, formată din pătrate egale şi desenează o histogramă.
Fiecare coloană a matricei X corespunde unei dimensiuni a reţelei
hist3(X, [n1,n2]) desenează histograma având în spaţiu xOy o rețea de tip $n_1 \times n_2$.

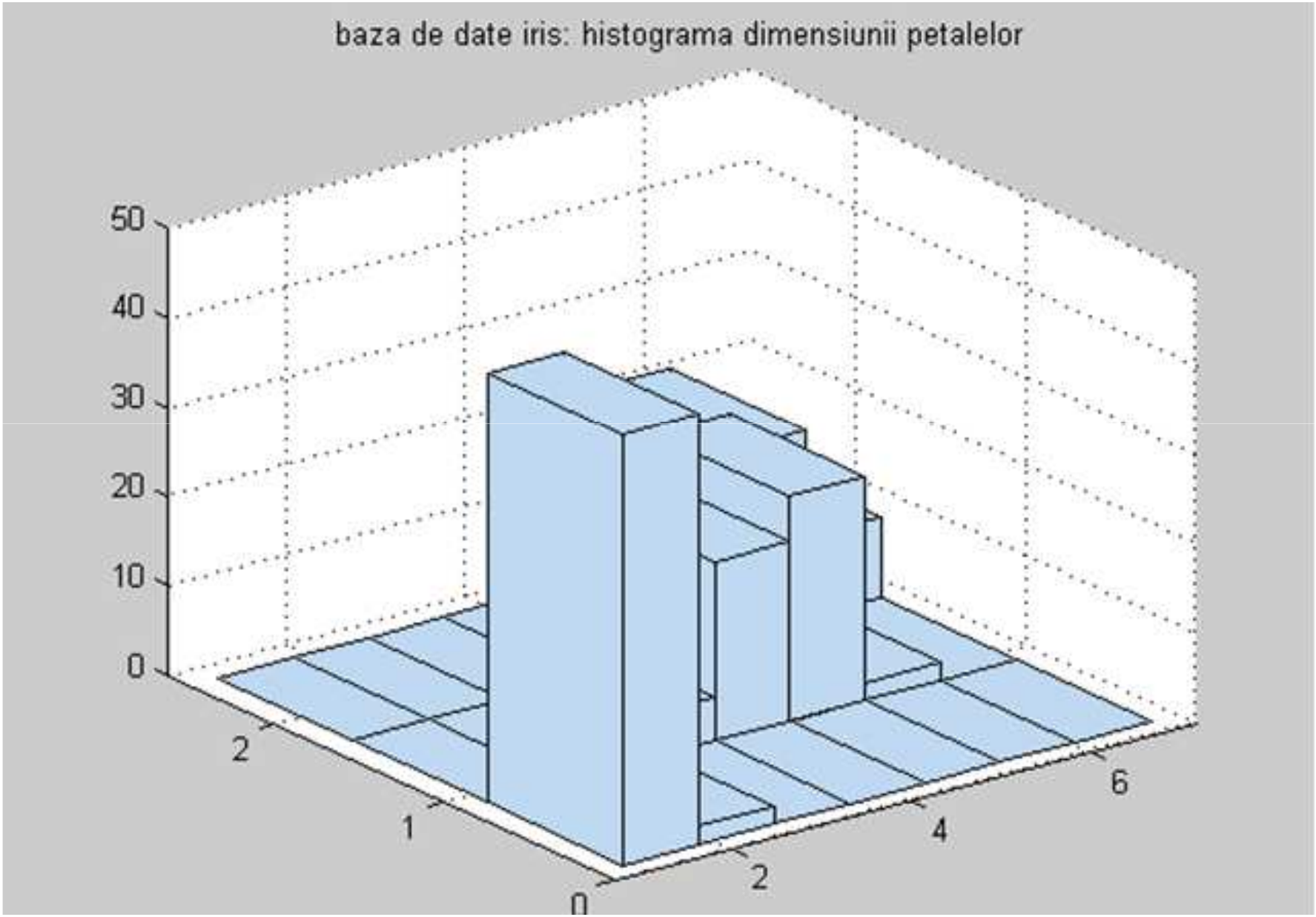
Desenăm histogramele tridimensionale pentru dimensiunea petalelor în cazul bazei de date Iris:

```
>> load fisheriris
>> x=meas(:,3:4);
>> hist3(x)
```



```
>> load fisheriris
>> x=meas(:,3:4);
>> min(meas(:,3))
ans =
     1
>> max(meas(:,3))
ans =
   6.9000
>> min(meas(:,4))
ans =

   0.1000
>> max(meas(:,4))
ans =
   2.5000
>> hist3(x,[7 3])
```


baza de date iris: histograma dimensiunii petalelor
50
40
30
20
10
0
2
1
0
2
4
6

Examinarea repartiţiilor variabilelor

Asimetria(skewness)

În cazul în care repartiţia datelor nu este simetrică este necesară analiza abaterii de la simetrie.
Definim *asimetria* (*skewness* – Pearson, 1895) ca fiind măsura deviaţiei repartiţiei date de la simetrie. Formula de calcul a asimetriei este:

$$Asimetria = \frac{n \cdot \sum_{1}^{n} (x_i - \bar{x})^3}{(n-1)(n-2) \cdot \sigma^3}.$$

Dacă asimetria este diferită de zero, atunci repartiţia este asimetrică.

În repartiţiile asimetrice media şi mediana sunt diferite.

11. Exemplu: Calculul asimetriei in Matlab

Pentru vectori skewness(x) este asimetria elementelor lui x.
Pentru matrice skewness(X) este un vector linie ale cărui elemente sunt asimetriile fiecărei coloane.

O parte din elevii unei clase au dat bacalaureatul la literatura română, matematică şi informatică. Notele sunt prezentate sub forma unei matrice, X, prima coloană reprezentând notele de la română, a doua pe cele de la matematică iar a treia reprezintă notele de la informatică. Să studiem simetria acestor rezultate, calculând valoarea skewness(X):

```
>> X=[7 7 6; 9 5 7; 8 4 5; 9 6 8;10 8 9; 6 9 8;9 4 5; 8 5 4; 9 8 7;10 10 8];
>> skewness(X)
ans =
   -0.6873    0.1862   -0.2951
```

Dacă repartiţia este asimetrică, atunci ea va avea o “coadă”fie în stânga, fie în dreapta.
Dacă această “coadă” este la dreapta, spunem că repartiţia are o asimetrie pozitivă, iar în celălalt caz, al “cozii” la stânga, asimetria este negativă.

12. Exemplu

Notele unei subgrupe de studenţi la un examen mai dificil sunt
{ 2 2 3 3 3 4 5 5 9 10}

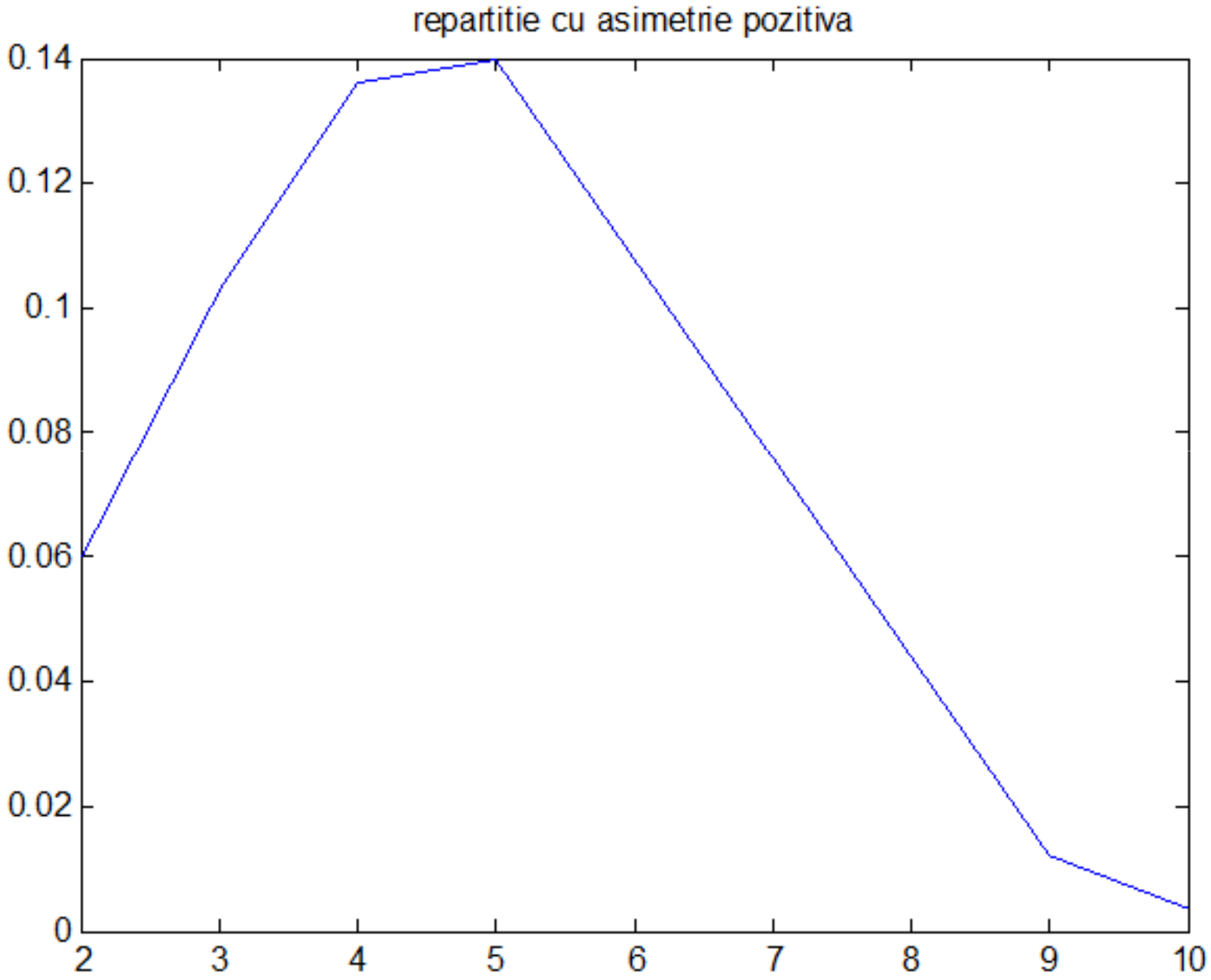
```
>> x=[2, 2, 3, 3, 3, 4, 5, 5, 9, 10];
>> m=mean(x); M=median(x); [m M]
ans =
   4.6000    3.5000
```

Repartiţia este asimetrică.

```
>> skewness (x)
ans =
   1.0446
```

Repartiţia are o asimetrie pozitivă

```
>> s=std(x);
>> f=1/(s*sqrt(2*pi))*exp(-((x-m)./s).^2);plot(x,f)
```


repartitie cu asimetrie pozitiva
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2
3
4
5
6
7
8
9
10

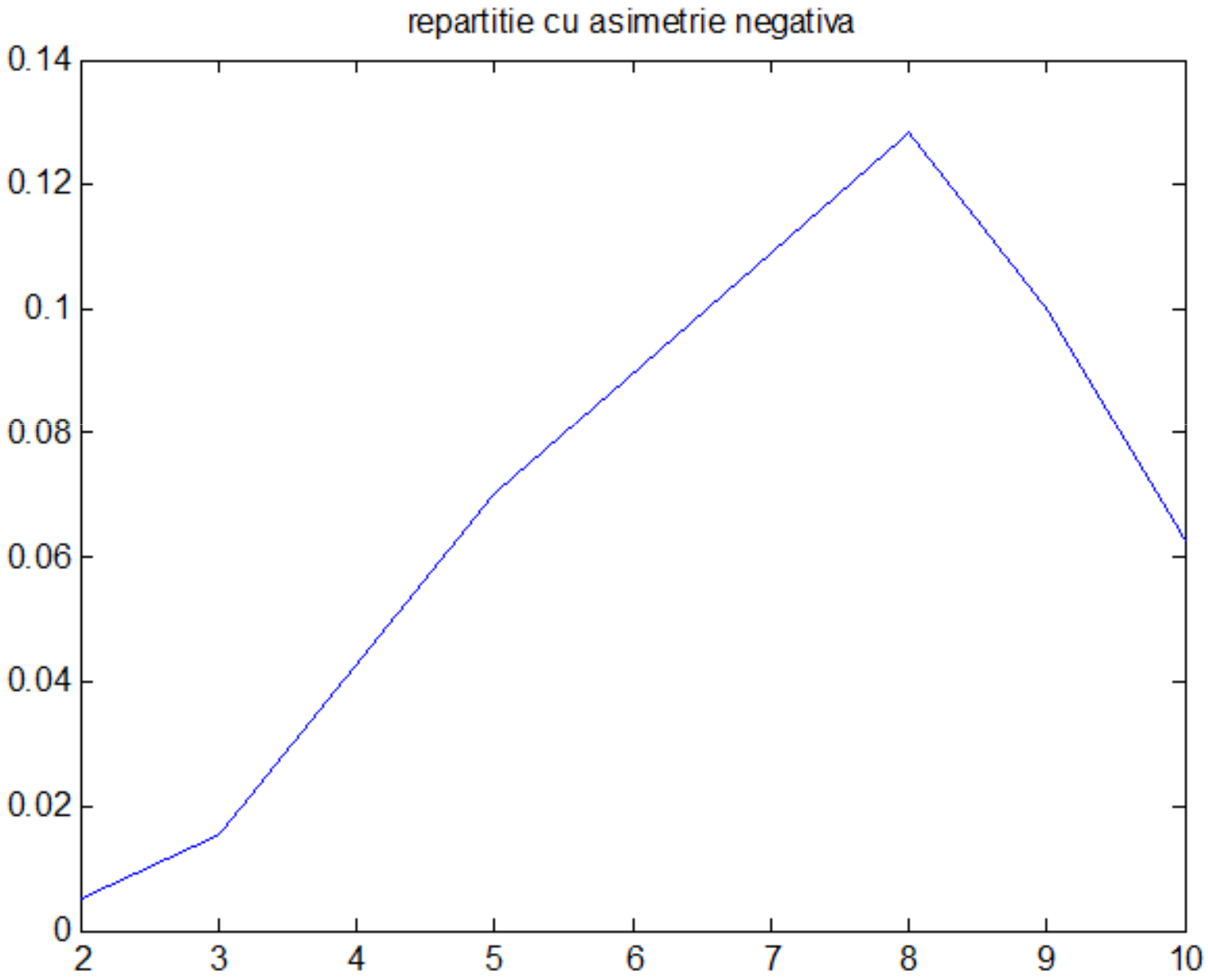
13. Exemplu

Notele aceleiaşi subgrupe de studenţi la un examen mai uşor sunt:
{2 3 5 8 8 9 9 10 10 10 }

```
>>y=[2 3 5 8 8 9  9 10 10 10 ];
>> m1=mean(y);M1=median(y);  [m1 M1]
ans =
   7.4000    8.5000

>> skewness(y)
ans =
  -0.8559

>> s1=std(y);
>> g=1/(s1*sqrt(2*pi))*exp(-((y-m1)./s1).^2);plot(y,g)
```


repartitie cu asimetrie negativa
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2
3
4
5
6
7
8
9
10

Repartiţia normală (gaussiană) este perfect simetrică, reprezentând modelul generic de simetrie. Reamintim că funcţia de repartiţie, cunoscută în acest caz sub numele de funcţia gaussiană este definită de:

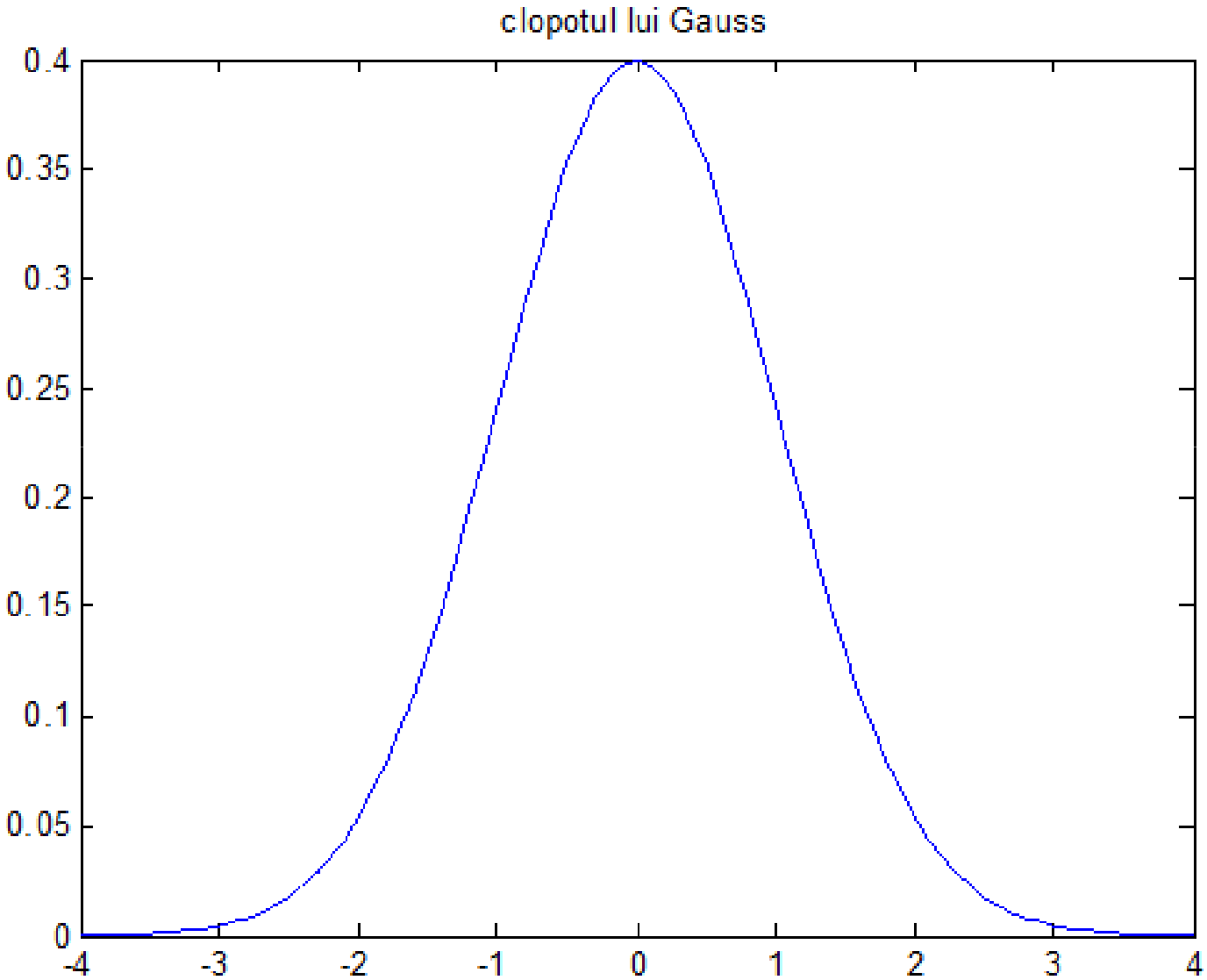
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{\sigma^2}},$$

unde μ este media , σ^2 este dispersia şi σ deviaţia standard.

Repartiţia cu $\mu = 0$ şi $\sigma = 1$ este repartiţia normală standard

Figura următoare reprezintă „clopotul lui Gauss”, adică cea mai „simetrică” repartiţie.

```
>>x=-4:.1:4; plot(x,1/sqrt(2*pi)*exp(-x.^2/2))
```


clopotul lui Gauss
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4

Excesul (kurtosis)

Cealaltă măsură tipică formei repartiţiei, *excesul* (*kurtosis* Pearson, 1905), măsoară „ascuţimea”unei repartiţii, în sensul cât este aceasta de dispusă de a avea valori extreme (outliers).

Dacă excesul este diferit de zero, atunci repartiţia este ori mai „plată” ori mai „ascuţită” decât repartiţia normală, care este chiar zero.
Excesul se calculează utilizând formula următoare:

$$Exces = \frac{n \cdot (n+1) \cdot \sum_{1}^{n} (x_i - \bar{x})^4 - 3 \cdot (n-1) \left[\sum_{1}^{n} (x_i - \bar{x})^2 \right]^2}{(n-1) \cdot (n-2) \cdot (n-3) \cdot \sigma^4}$$

Această formulă este folosită pentru calculele în Excel.

În Matlab se utilizează formula

$$K = \frac{\sum_{1}^{n} (x_i - \bar{x})^4}{\sigma^4}$$, unde $\bar{x}$ este media iar σ este deviatia standard.

Avem egalitatea K=*Exces* + 3

Funcţia kurtosis(x) returnează în cazul vectorului x, excesul K al lui x,

iar în cazul matricei X, un vector linie, ale cărui componente sunt valorile Kurtosis-ului pentru fiecare coloană în parte.

Exemple

Reluăm exemplul nr 11 (cu notele de la bacalaureat) calculând asimetria şi excesul:

```
X=[7 7 6; 9 5 7; 8 4 5; 9 6 8;10 8 9; 6 9 8;9 4 5; 8 5 4; 9 8 7;10 10 8];
>> skewness(X)
ans =
        -0.6873    0.1862   -0.2951
>> kurtosis(X)
ans =
         2.5981    1.7103    1.8363
```

Reluăm exemplele nr 12 (cu notele la un examen dificil) şi nr 13 (cu notele la un examen uşor) calculănd asimetria şi excesul.

```
>> x=[2, 2, 3, 3, 3, 4, 5, 5, 9, 10];
>> skewness (x)
ans =
   1.0446
>> kurtosis(x)
ans =
   1.7996

>>y=[2 3 5 8 8 9  9 10 10 10 ];
>> skewness(y)
ans =
  -0.8559
>> kurtosis(y)
ans =
   2.1793
```

Datele date întâlnite frecvent în realitate sunt caracterizate de repartiţii non-normale (negaussiene), care sunt mai mult sau mai puţin îndepărtate de repartiţia normală.

Vom prezenta, patru asemenea repartiţii, cele mai des întâlnite.

Repartiţia exponenţială

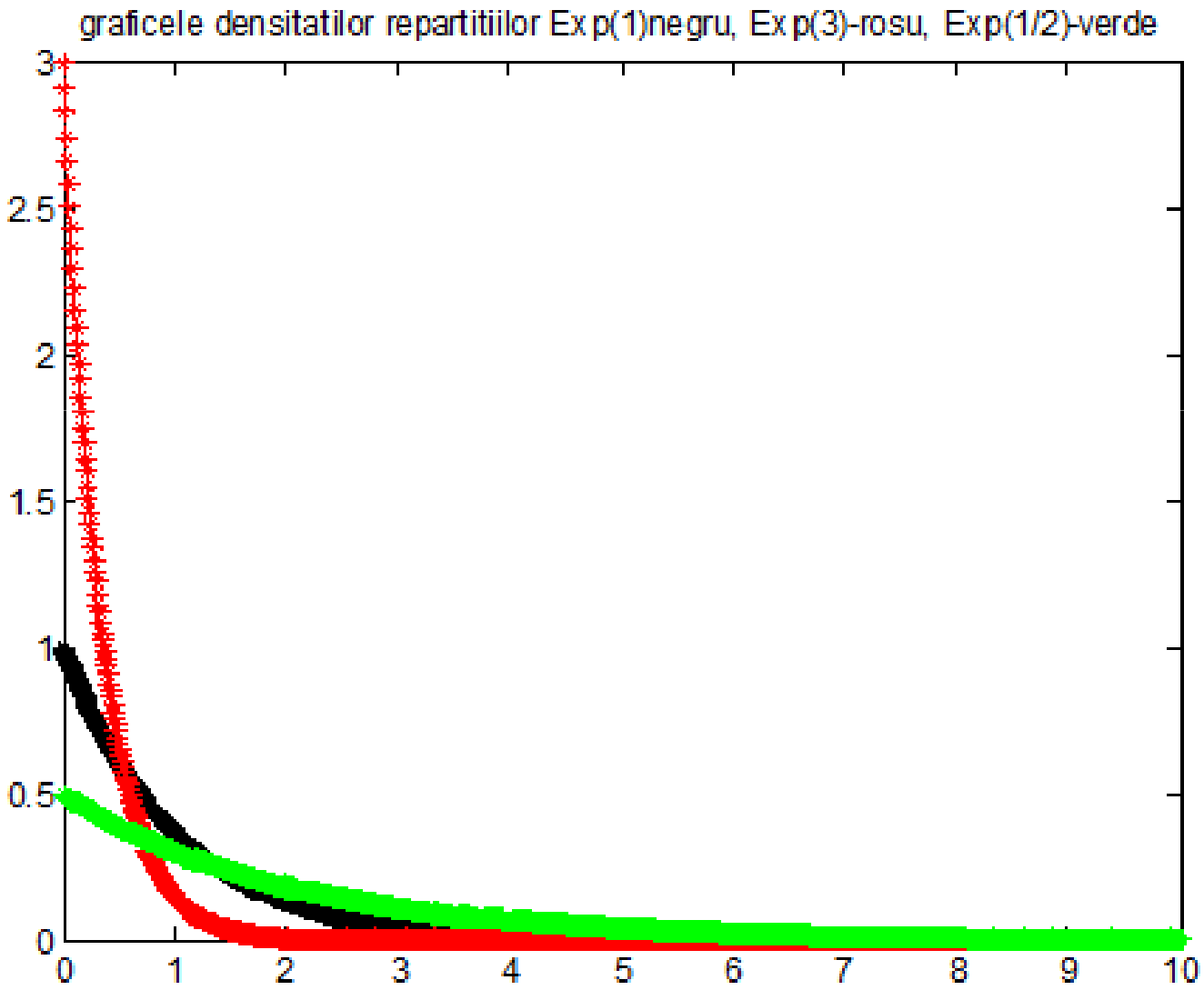
Repartiţia *exponenţială* $X \sim Exp(\lambda)$, descrie timpii dintre evenimentele unui proces Poisson, (proces în care evenimentele apar continuu şi independent cu o rată medie constantă).

Repartiţia *exponenţială* de parametru $\lambda > 0$ este caracterizată de densitatea $f_X(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$.

Media corespunzătoare este $E(X) = \frac{1}{\lambda}$.

Dispersia corespunzătoare este $D^2(X) = \frac{1}{\lambda^2}$

```
>> x=0:.01:10;plot(x,exp(-x),'k*',x,3.*exp(-3.*x),'r*',x,1./2.*exp(-x./2),'g*')
```

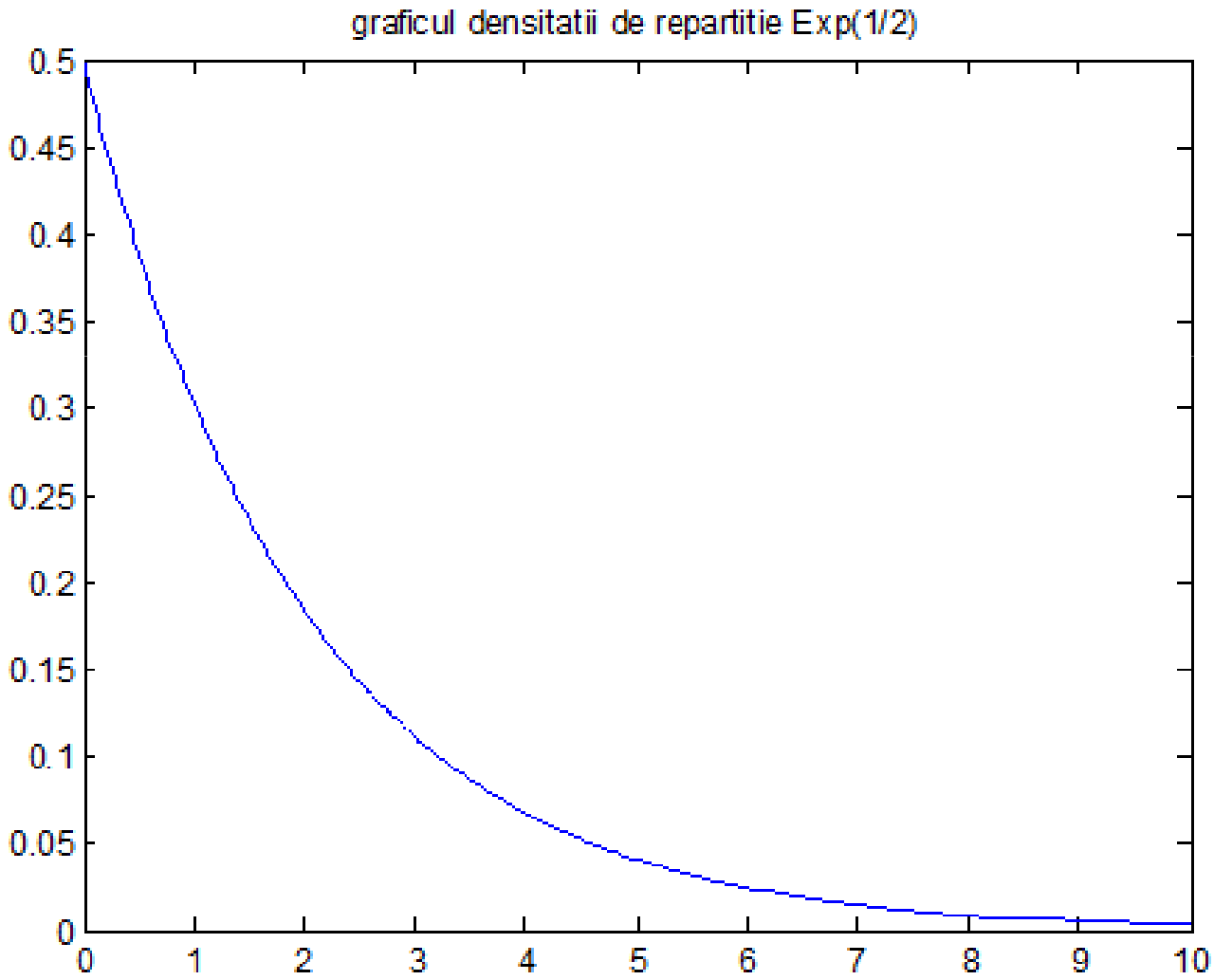


Graficul densitaţii de repartiţie (*probability density function* – pdf) în cazul repartiţiei exponeţiale poate fi desenat în Matlab folosind sintaxa:

```
Y = exppdf(X,mu)
```

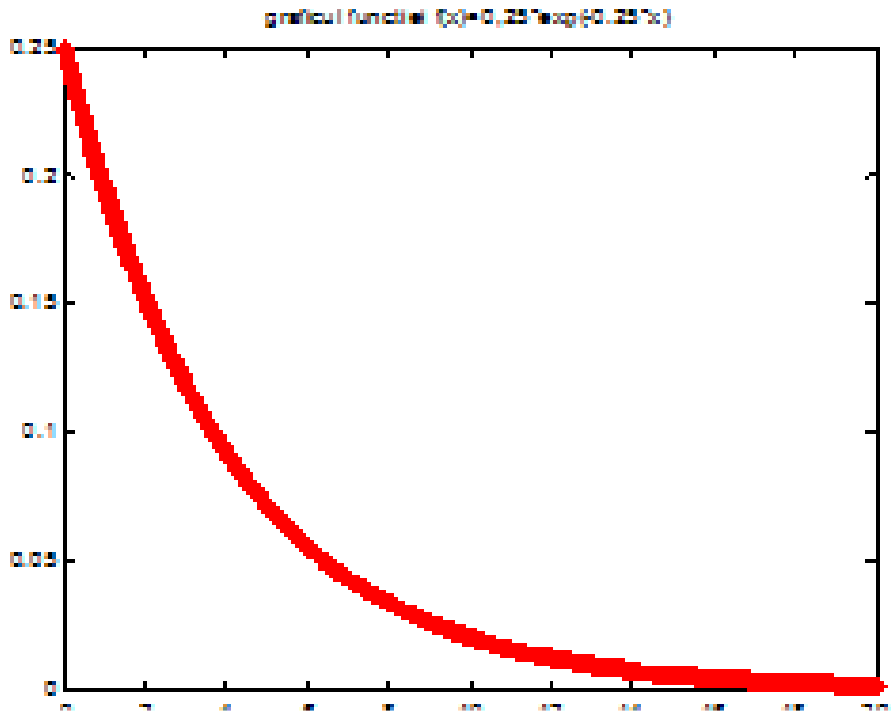
care returnează valorile în X ale densităţii de repartiţie exponenţială de parametru pozitiv mu, unde mu reprezintă media lui X

```
>> X = (0:0.02:10);
>> Y = exppdf(X,2);
>> plot(X,Y)
```



14. Exemplu

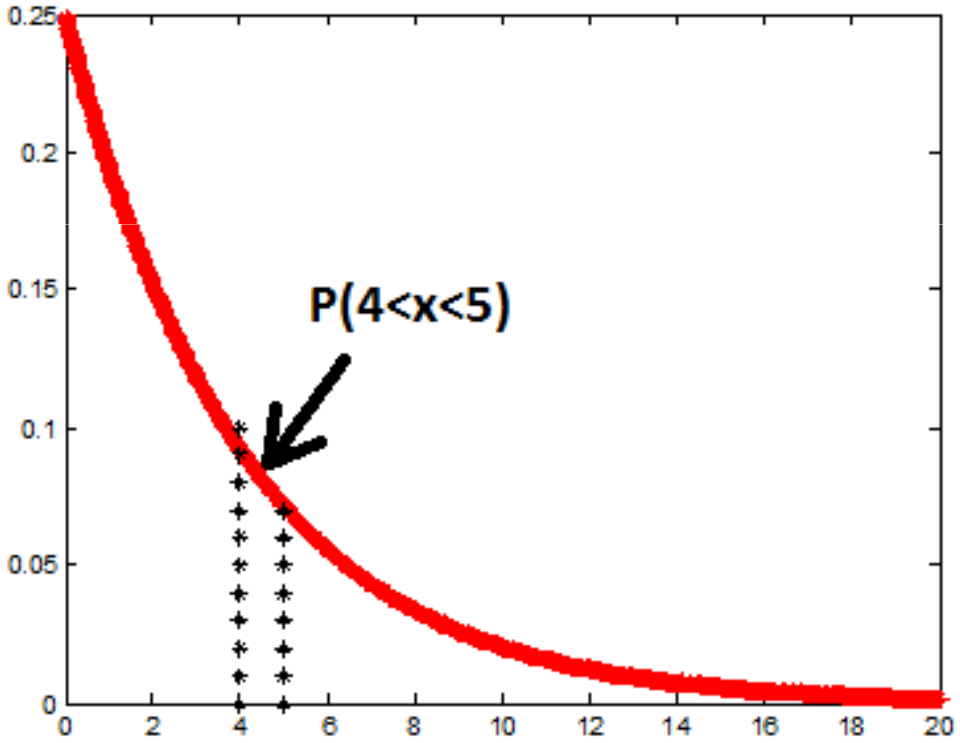
X reprezintă timpul, în minute, pe care un funcţionar de la ghişeul de informaţii al unei bănci îl alocă fiecărui client. Se ştie că aceşti timpi au o repartiţie exponenţială, iar timpul mediu alocat unui client este de 4 minute. Astfel $\lambda = \frac{1}{\mu}$, unde $\mu = 4$. Funcţia de repartiţie este $f(x) = \frac{1}{4} e^{-\frac{1}{4}x}$ pentru $x \geq 0$



Să calculăm probabilitatea ca funcţionarul să aloce unui client Între 4 şi 5 minute, adică $P(4 < x < 5)$

Ştim că $P(x < a) = 1 - e^{-\lambda a}$ şi astfel

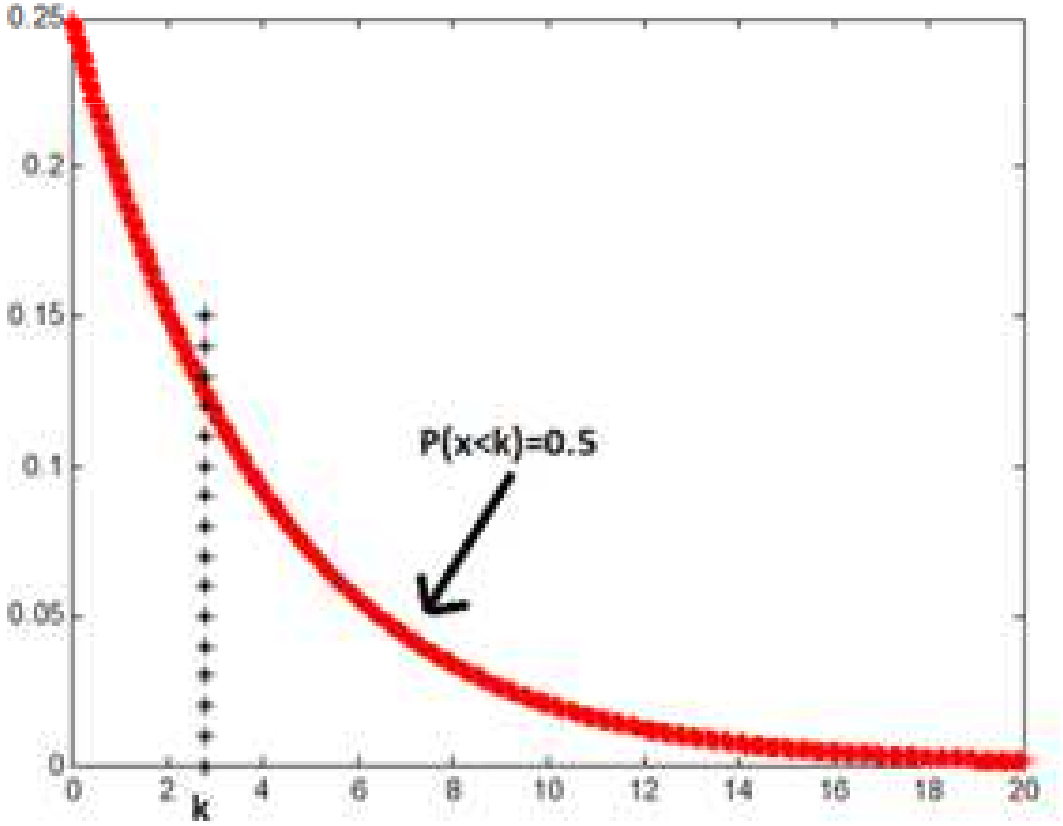
$P(x < 4) = 1 - e^{-0.25 \cdot 4} = 0.6321$ şi $P(x < 5) = 1 - e^{-0.25 \cdot 5} = 0.7135$



Ordonăm crescător timpii alocaţi fiecărui client. Care este media de timp alocată fiecărui client din prima jumătate? Această înseamnă să găsim a 50-a percentilă, rezolvând ecuaţia $P(x < k) = 1 - e^{-0.25 \cdot k} = 0.5$

```
>> syms k
>>solve(0.5-exp(-0.25*k))
ans =
   4*log(2)
```

Avem: k=2.7726



Care este mai mare media sau mediană?

Media= 4 (enunţul problemei);
Mediana =2.7726 (50-a percentilă).

Generarea de numere aleatoare exponenţial repartizate - Matlab

R = exprnd(mu) generează aleator numere exponenţial repartizate; mu poate fi un vector sau o matrice. Dimensiunea lui R este egală cu dimensiunea lui mu

```
>> R = exprnd(1:8)
R =
   1.2840   6.1508   6.9951   0.7770   1.8205   6.8912   0.3574  26.9469
>> R = exprnd(1:8)
R =
   0.8238   1.9270   0.8016   0.9166   8.3866   4.2830   5.6586   3.4918

>> A=[1 3; 2 4; 7 5];
>> R = exprnd(A)
R =
   1.8476   2.1685
   0.0597   0.8912
   0.3064   9.7636
>> R = exprnd(A)
R =
   0.8633   0.1241
   0.1761   1.6880
```

Repartiţia gamma

Repartiţia *gamma* $X \sim \text{GAM}(\lambda, k)$, de parametri $\lambda > 0$ şi $k > 0$, are densitatea dată de:

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)},\ x \geq 0 \text{ şi } f_X(x) = 0,\ x < 0,$$

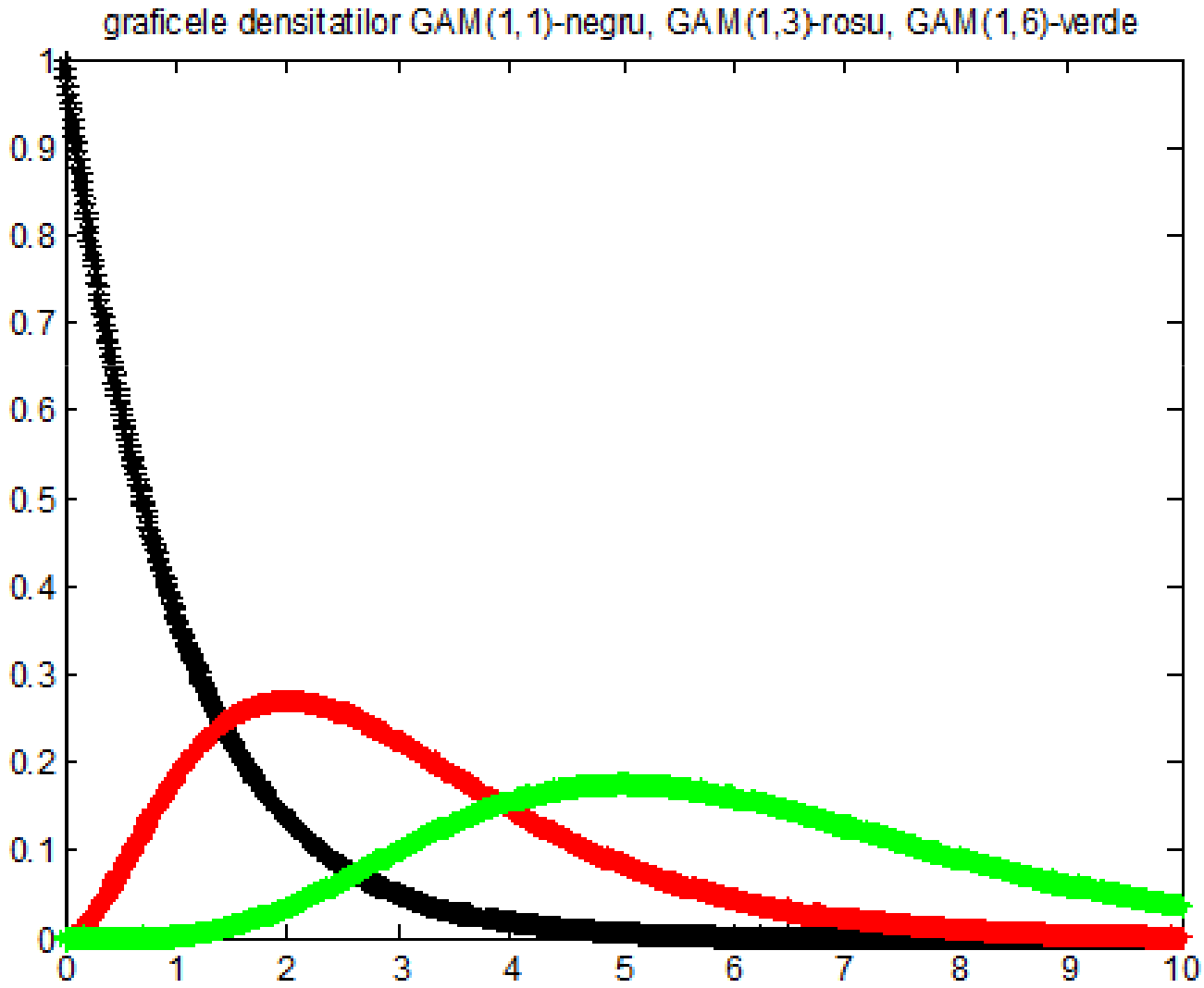
unde $\Gamma(k) = \int_0^\infty e^{-x} x^{k-1} dx$ este funcţia *gamma*.

Media corespunzătoare este $E(X) = \frac{k}{\lambda}$.

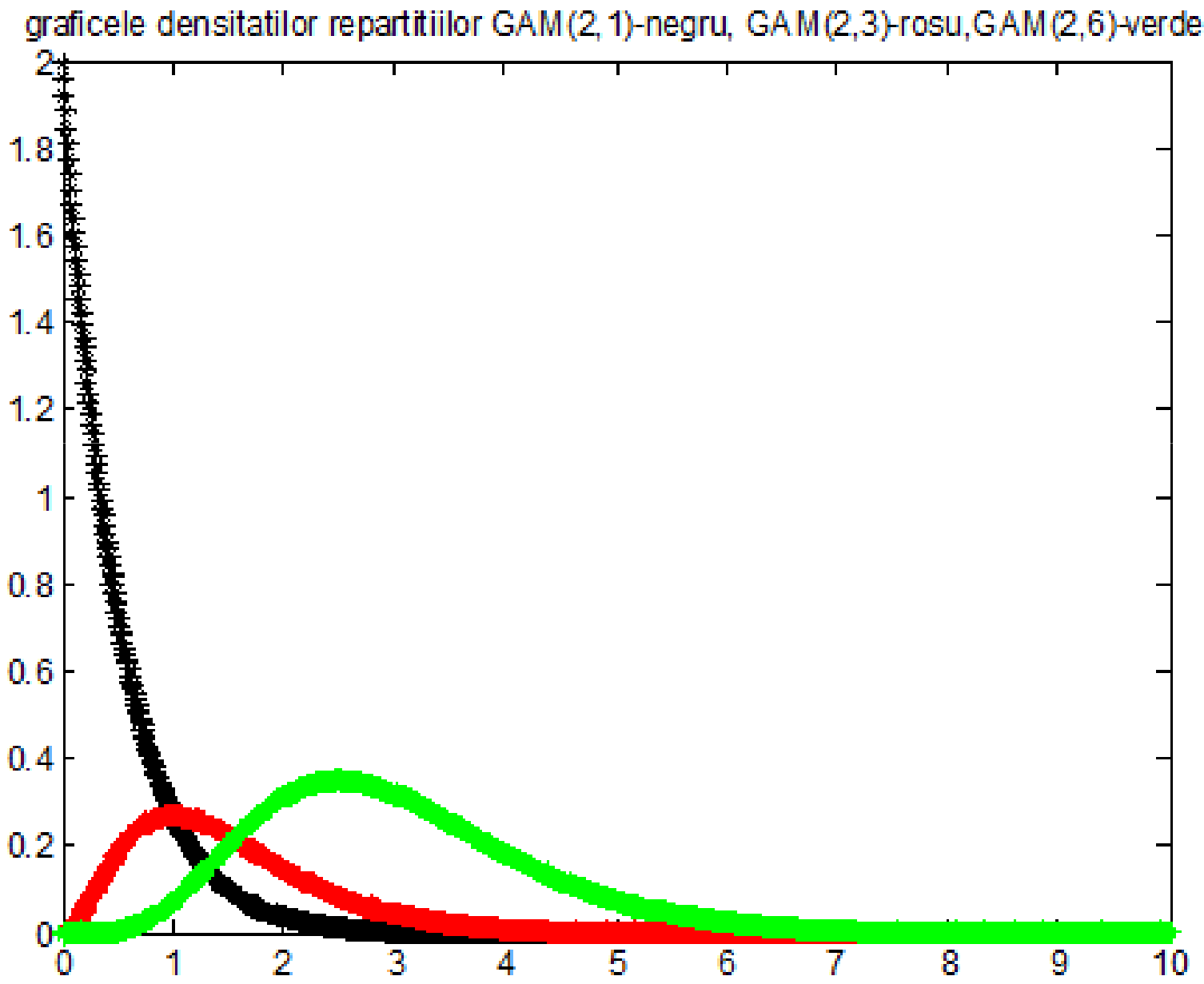
Dispersia corespunzătoare este $D^2(X) = \frac{k}{\lambda^2}$

$\text{GAM}(\lambda, 1) = \lambda \cdot e^{-\lambda x} = Exp(\lambda)$

```
>> syms x
>> k=1; gam(1)=int(exp(-x),0,inf); for k=2:6 gam(k)=int(x^(k-1)*exp(-x),0, inf);end
>> x=0:.01:10;plot(x,exp(-x)/gam(1),'k*',x,(x.^2).*exp(-x)/gam(3),'r*',
x,(x.^5).*exp(-x)/gam(6),'g*')
```



```
>> x=0:.01:10;plot(x,2*exp(-2*x)/gam(1),'k*',x,2^2*(x.^2).*exp(-2*x)/gam(3),'r*',
x,2^6*(x.^5).*exp(-2*x)/gam(6),'g*')
```



Graficul densitaţii de repartiţie (*probability density function* – pdf) în cazul repartiţiei gamma poate fi desenat folosind sintaxa

```
Y = gampdf(X,k,lamda)
```

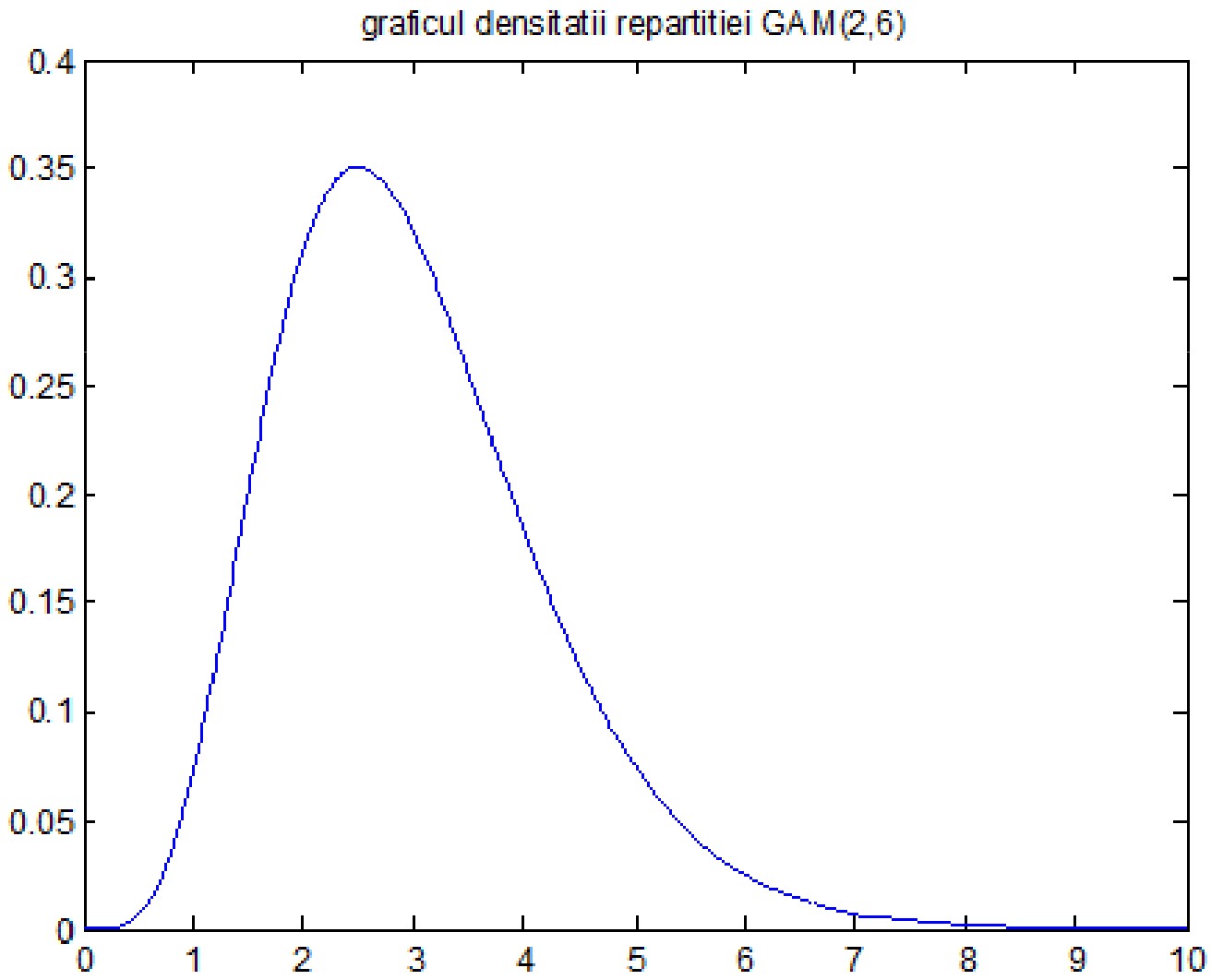
care returnează valorile în X ale densităţii repartiţiei GAM(λ , k).
Este obligatoriu de reţinut că în Matlab funcţia de repartiţie gamma are formula:

$$f_X(x) = \frac{x^{k-1} e^{-\frac{x}{\lambda}}}{\lambda^k \Gamma(k)}$$

care este o altă exprimare pentru formula anterioarăprin înlocuirea lui λ cu $\frac{1}{\lambda}$.

Să desenăm densitatea repartiţiei GAM(2,6):

```
>> X=0:.01:10;Y=gampdf(X,6,1/2);plot(X,Y)
```



Dacă $X_1, X_2, \ldots, X_k$ sunt variabile aleatoare exponențiale, independente și identic repartizate de parametru λ, atunci variabila sumă

$$X_1 + X_2 + \ldots + X_k$$

este o variabilă gamma de parametrii λ și k.

Generarea de numere aleatoare gamma repartizate - Matlab

R = gamrnd(A,B,m,n) generează aleator numere cu gamma repartizate de parametri A şi B , scalarii m şi n reprezentând numărul de linii, respectiv coloane a lui R.

```
>> R = gamrnd(1,1,1,7)
R =
   2.8643    1.0182    2.0584    0.6936    2.0296    0.7393    0.2257
>> R = gamrnd(1,1,1,7)
R =
   5.0164    1.4830    1.6873    0.0521    1.1500    0.5191    0.9342

>> R = gamrnd(2,5,1,9)
R =
  23.3573   13.6873    6.8571    2.8459    9.0288   38.6431    9.6020    1.0473
1.2877
>> R = gamrnd(2,5,1,9)
R =
   3.8150    5.2894    5.0279   14.0750    7.1420    4.0038    3.1517    9.1600
2.7154
```

Repartiţia Weibull

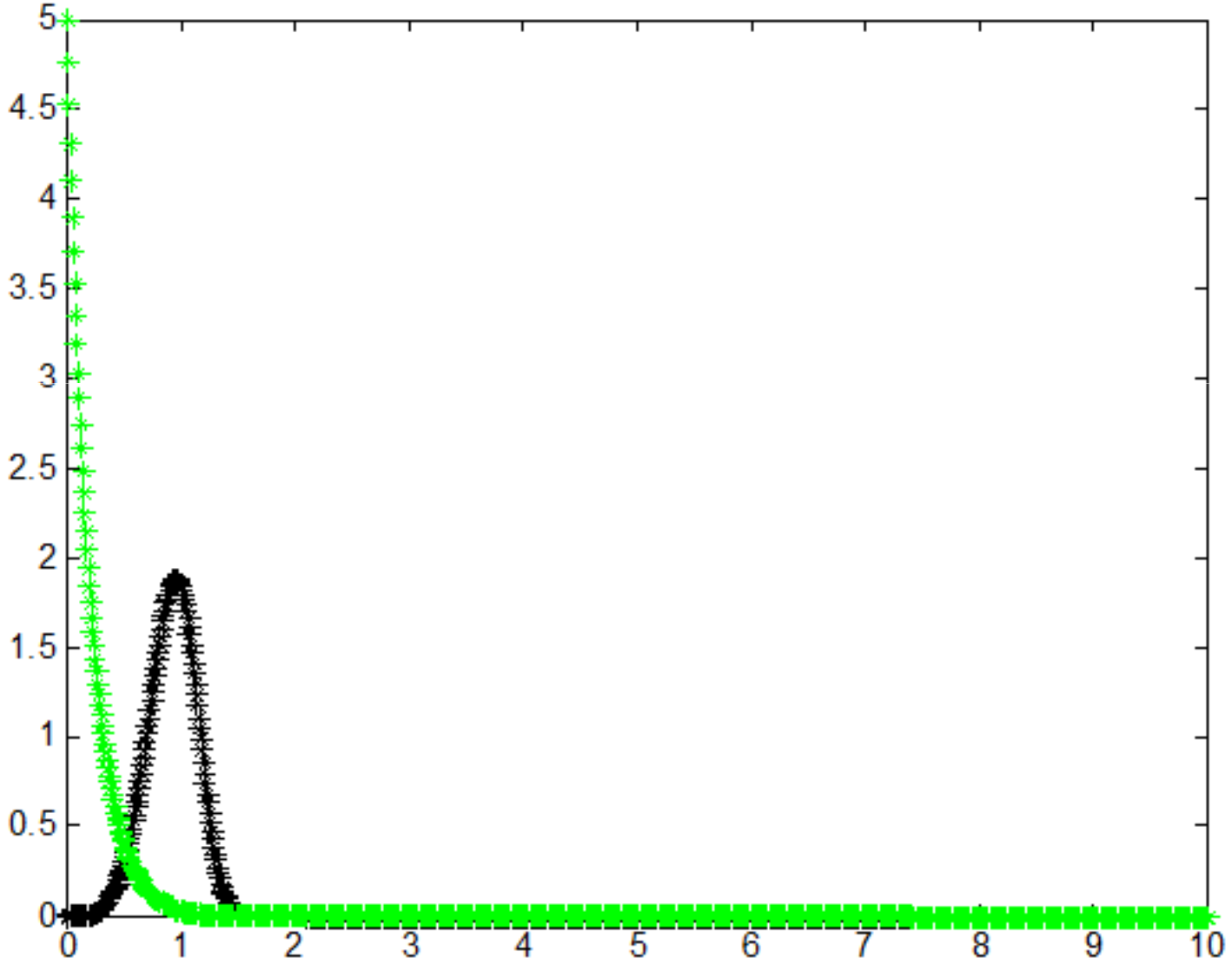
Repartiţia *Weibull*, $X \sim \text{WEI}(\alpha, \beta)$, de parametri $\alpha > 0$ şi $\beta > 0$ are densitatea dată de:

$$f_X(x) = \alpha\beta(\alpha x)^{\beta-1} e^{-(\alpha x)^{\beta}}, x \geq 0 \text{ şi } f_X(x) = 0, x < 0,$$

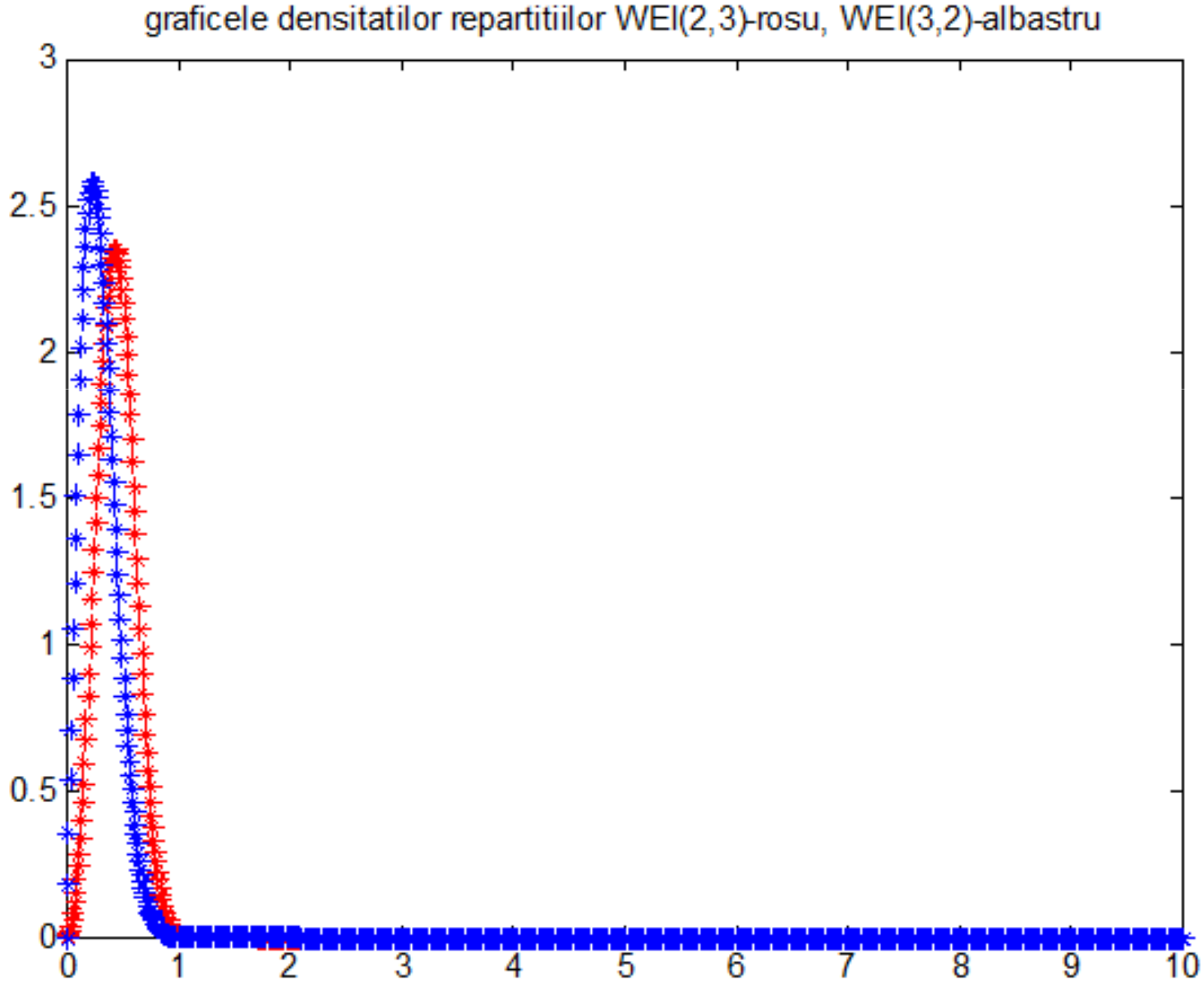
Este un instrument în modelarea rezistenţei materialor şi în modelarea timpilor de supravieţuire.

Prezentăm graficele densităţilor de repartiţie pentru diferite valori ale parametrilor α şi β

```
>>x=0:.01:10;plot(x,(5*x.^4).*exp(-x.^5),'k*',x,5*exp(-5*x),'g*')
```



```
>>x=0:.01:10;plot(x,6*(2*x).^2.*exp((-2*x).^3),'r*',x,6*(3*x).*exp(-(3*x).^2),'b*')
```

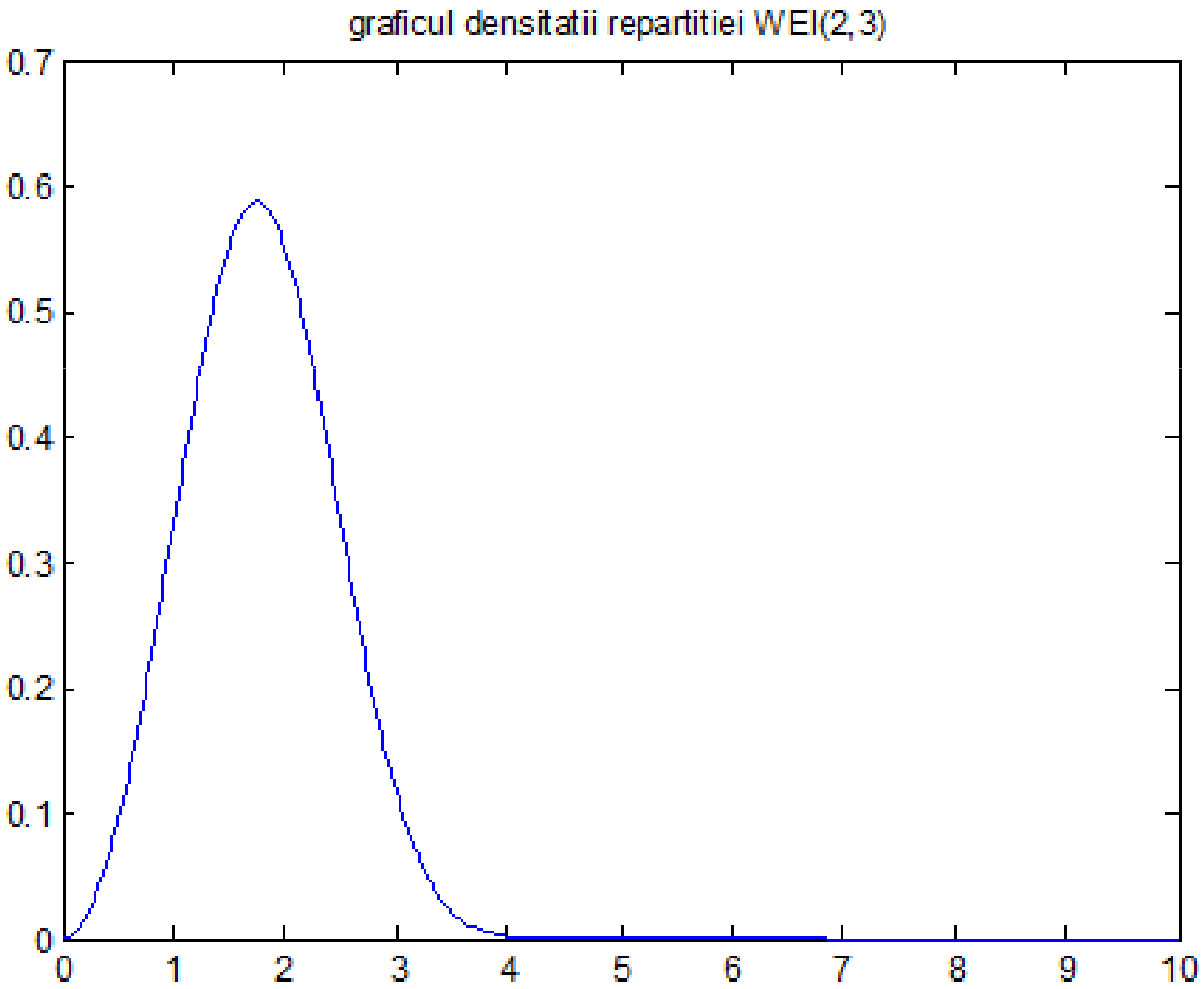


Graficul densitaţii de repartiţie (*probability density function* – pdf) în cazul repartiţiei Weibull poate fi desenat folosind sintaxa

```
Y  = wblpdf(X,A,B)
```

care returnează valorile în X ale densităţii de repartiţie Weibull de parametrii A şi B.

```
>> X = (0:0.02:10);
>> Y  = wblpdf(X,2,3);
>> plot(X,Y)
```



Generarea de numere aleatoare Weibull repartizate - Matlab

R = wblrnd(A,B,m,n) generează aleator numere Weibull distribuite de parametri A şi B , scalarii m şi n reprezentând numărul de linii, respectiv coloane a lui R.

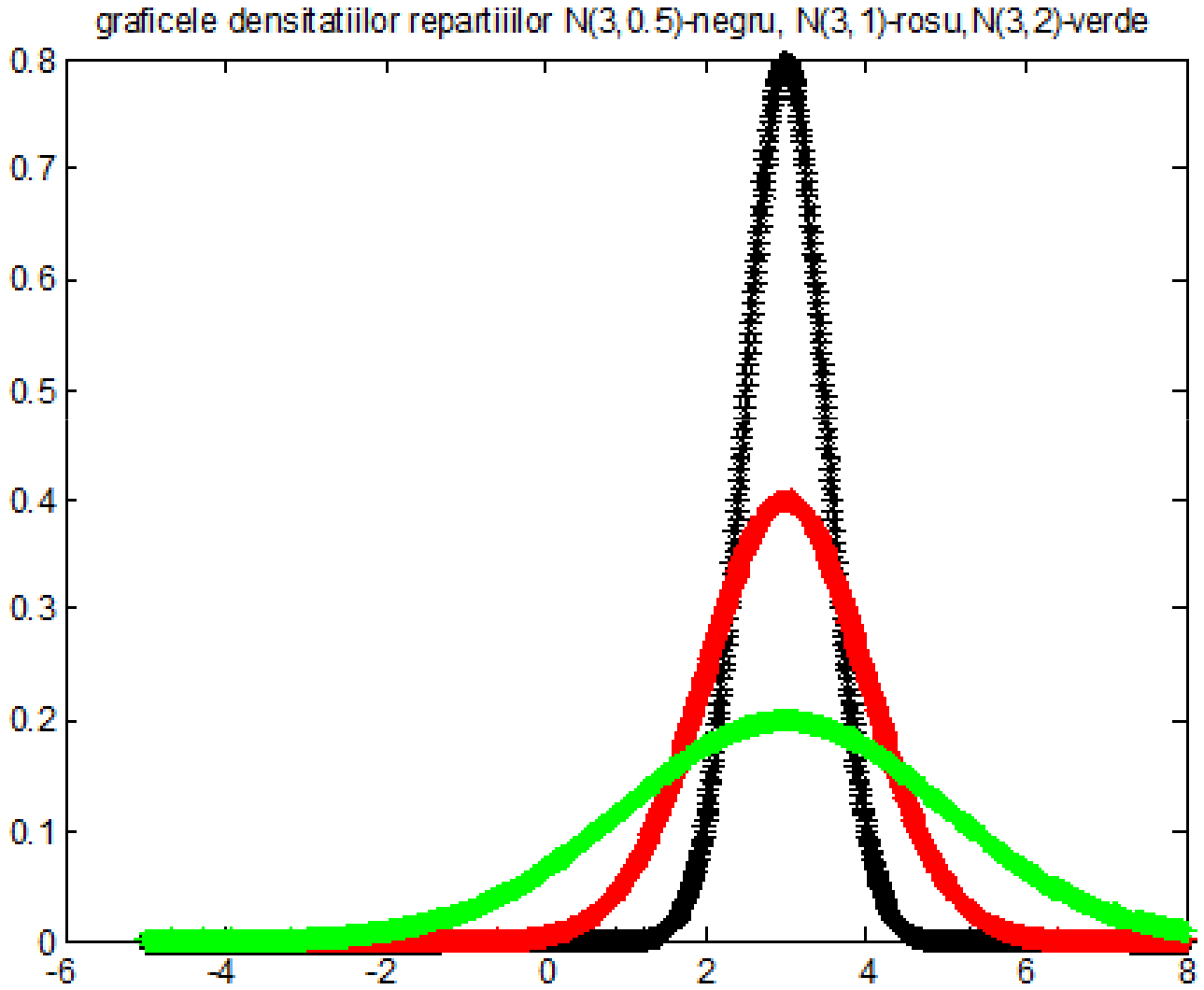
```
>> R = wblrnd(1,5,1,7)
R =
   0.7283   0.6296   1.1559   0.6186   0.8555   1.1841   1.0503
>> R = wblrnd(1,5,1,7)
R =
   0.9039   0.5340   0.5136   1.1306   0.4954   0.5349   0.9371

>> R = wblrnd(2,5,1,8)
R =
   1.4812   2.2864   1.9421   1.2301   1.4944   1.0576   1.6830   2.5444
>> R = wblrnd(2,5,1,8)
R =
   1.3924   1.1692   1.6546   1.5476   1.5687   1.9737   1.6833   2.2407
```

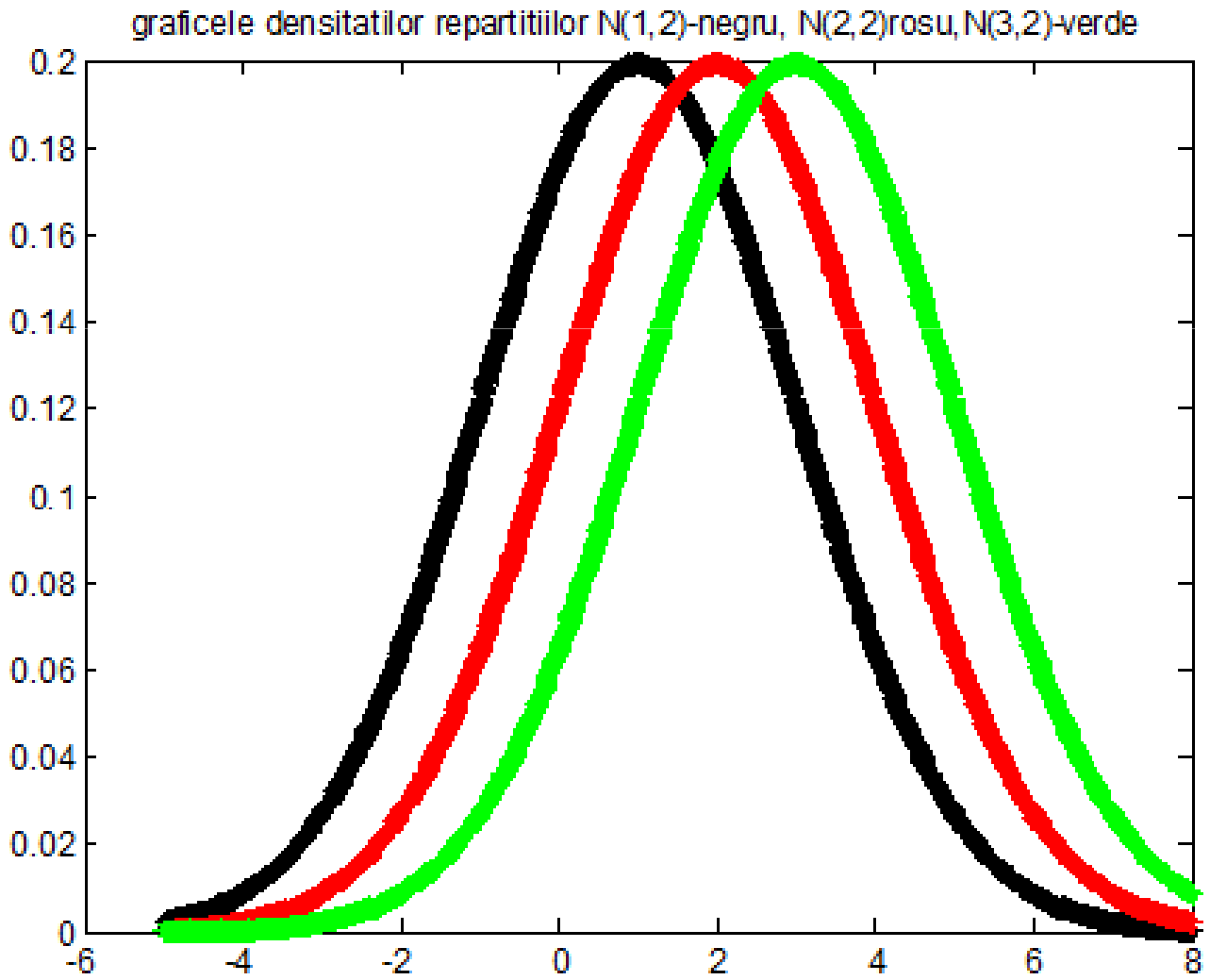
Repartiţia normală

Repartiţia *normală* (*legea lui Gauss*) $X \sim N(\mu, \sigma^2)$ de parametri μ şi σ^2 are densitatea de repartiţie $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{\sigma^2}}$, unde μ este media, σ^2 este dispersia şi σ deviaţia standard.

```
>>x=-5:.01:8;plot(x,1/(0.5*sqrt(2*pi))*exp(-(x-3).^2/(2*0.5^2)),'k*',
x,1/sqrt(2*pi)*exp(-(x-3).^2/2),'r*',x,1/(2*sqrt(2*pi))*exp(-(x-3).^2/(2*2^2)),'g*')
```



```
>>x=-5:.01:8;plot(x,1/(2*sqrt(2*pi))*exp(-(x-1).^2/(2*2^2)),'k*',
x,1/(2*sqrt(2*pi))*exp(-(x-2).^2/(2*2^2)),'r*',
x,1/(2*sqrt(2*pi))*exp(-(x-3).^2/(2*2^2)),'g*')
```

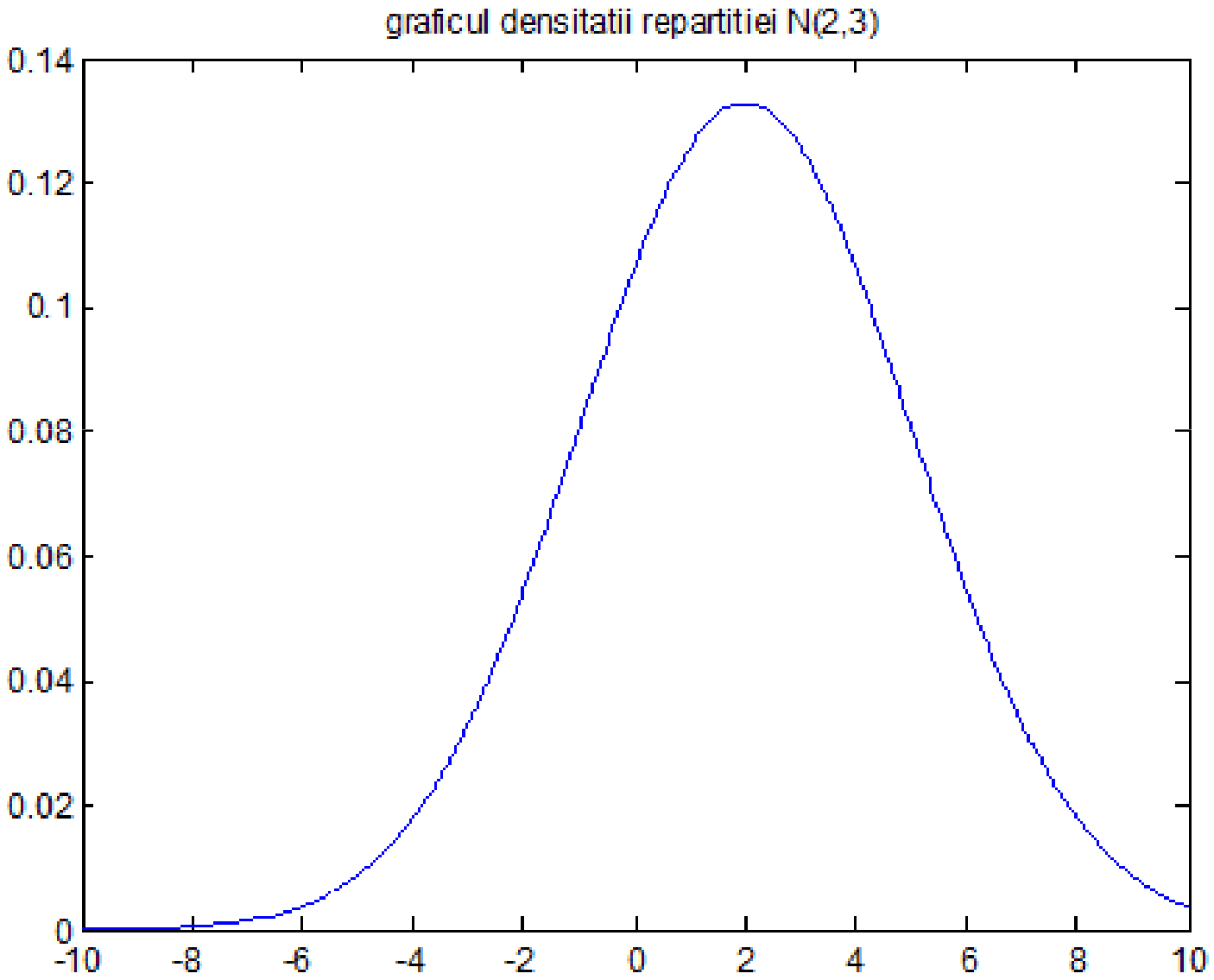


Pentru a desena în Matlab graficul densitaţii repartiţiei normale se foloseşte sintaxa

Y = normpdf(X,mu,sigma)

care calculează valorile în X a densităţii de repartiţie (pdf) pentru distribuţia normală de medie mu (μ) şi deviaţie standard sigma (σ) .

```
>> X=-10:.01:10;Y=normpdf(X,2,3);plot(X,Y)
```



Generarea de numere aleatoare normal repartizate - Matlab

R = normrnd(mu,sigma,m,n) generează aleator numere normal distribuite de parametri mu (media) şi sigma(dispersia), scalarii m şi n reprezentând numărul de linii, respectiv coloane a lui R.

```
>> R = normrnd(2,3,1,8)
R =
   4.6652   -1.4412   -1.2066   -0.4285   -6.8329    6.3151    2.9756   -0.2648
>> R = normrnd(2,3,1,8)
R =
   6.1109   -3.1345    1.6933    1.2757    2.9576    2.9386   -0.5946    1.9098

>> R = normrnd(3,5,1,9)
R =
   2.1756    6.1385    8.4663    8.5464   -1.3183    3.3868   -3.0706   -2.5675
2.9658

>> R = normrnd(3,5,1,9)
R =
  10.6632   -0.8483    4.8569    1.8721    8.5868   -2.4453    3.1628    5.7626
8.5031
```

Graficul probabilităţii normale- Matlab

Graficul probabilităţii normale este o tehnică care testează vizual dacă un set de date are sau nu o repartiţie normală.

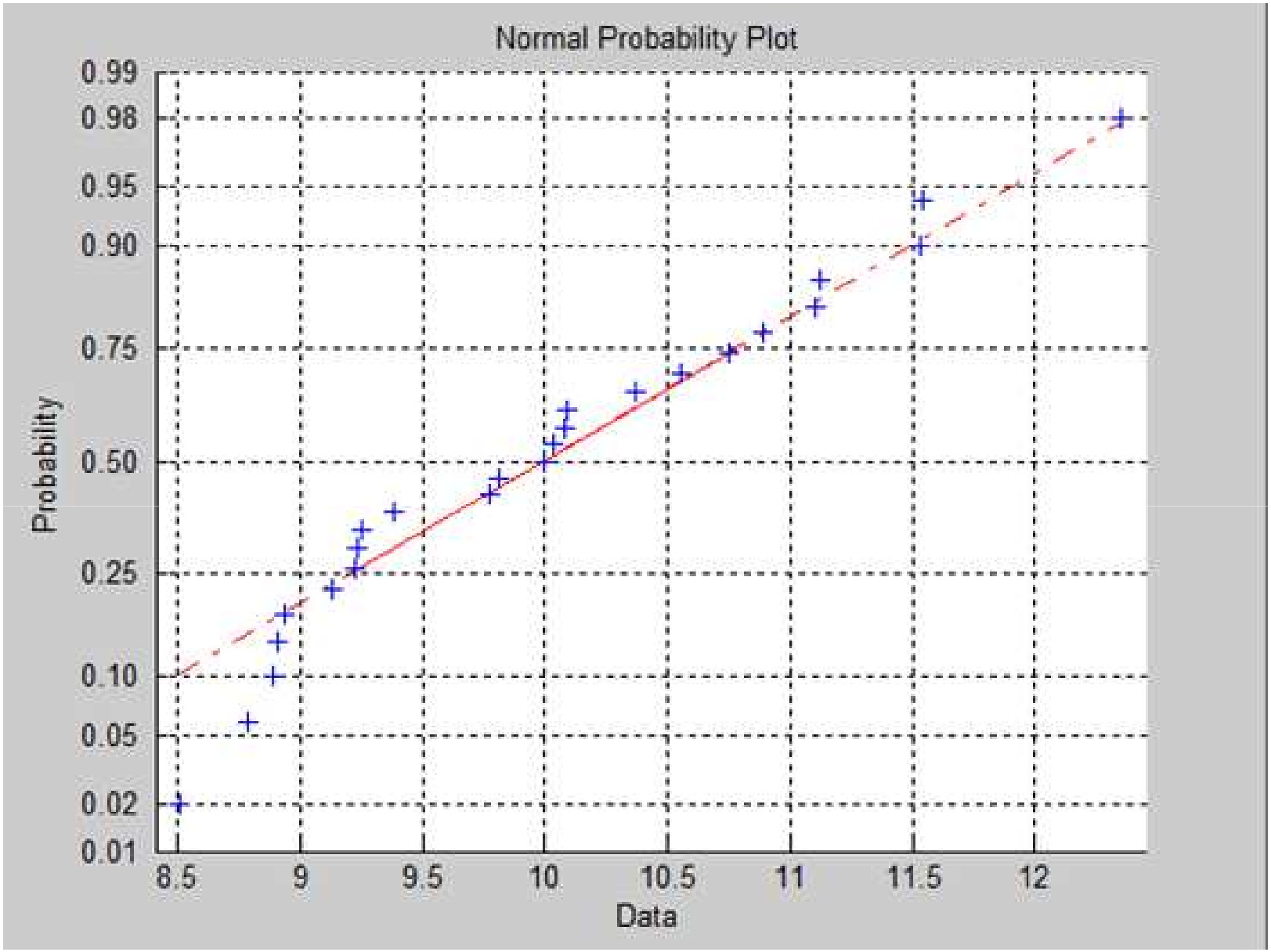
h = normplot(X) afişeaza un grafic al datelor din X.
Dacă X este o matrice, normplot afişează căte un grafic pentru fiecare coloană a lui X.
Graficul prezintă datele eşantionului notate cu '+'. Peste acest grafic este suprapusă o linie ce uneşte prima şi a treia cuartilă a lui X, linie ce este extrapolată la tot eşantionul.

Scopul acestei reprezentări este de a vizualiza dacă datele din X sunt normal repartizate. Dacă da, graficul va fi liniar. Alt tip de repartiţie face ca datele să fie pe o anumită curbă.
Dacă avem date cenzurate se utilizeaza probplot

15. Exemplu

Să generăm un eşantion de 25 de numere aleatoare, normal repartizate N(25,1) şi să desenăm graficul probabilităţii normale, corespunzător acestui eşantion

```
>>x = normrnd(10,1,1,25);
>>normplot(x)
```


Normal Probability Plot
Probability
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
8.5
9
9.5
10
10.5
11
11.5
12
Data

16. Exemplu

în Matlab bază de date gas.mat conţine două eşantioane cu preţurile galonului de benzină într-un număr de staţii, aleator alese din statul Massachusetts, în 1993. Primul eşantion price1, conţine 20 de observaţii aleatoare din aceaşi zi de ianuarie, staţiile fiind pe tot cuprinsul statului, în timp ce al doilea eşantion price2, conţine 20 de observaţii aleatoare din aceaşi zi, o lună mai târziu, staţiile fiind pe tot cuprinsul statului.

```
>> normplot(x)
>> load gas
>> prices = [price1 price2]
>> normplot(prices)
```

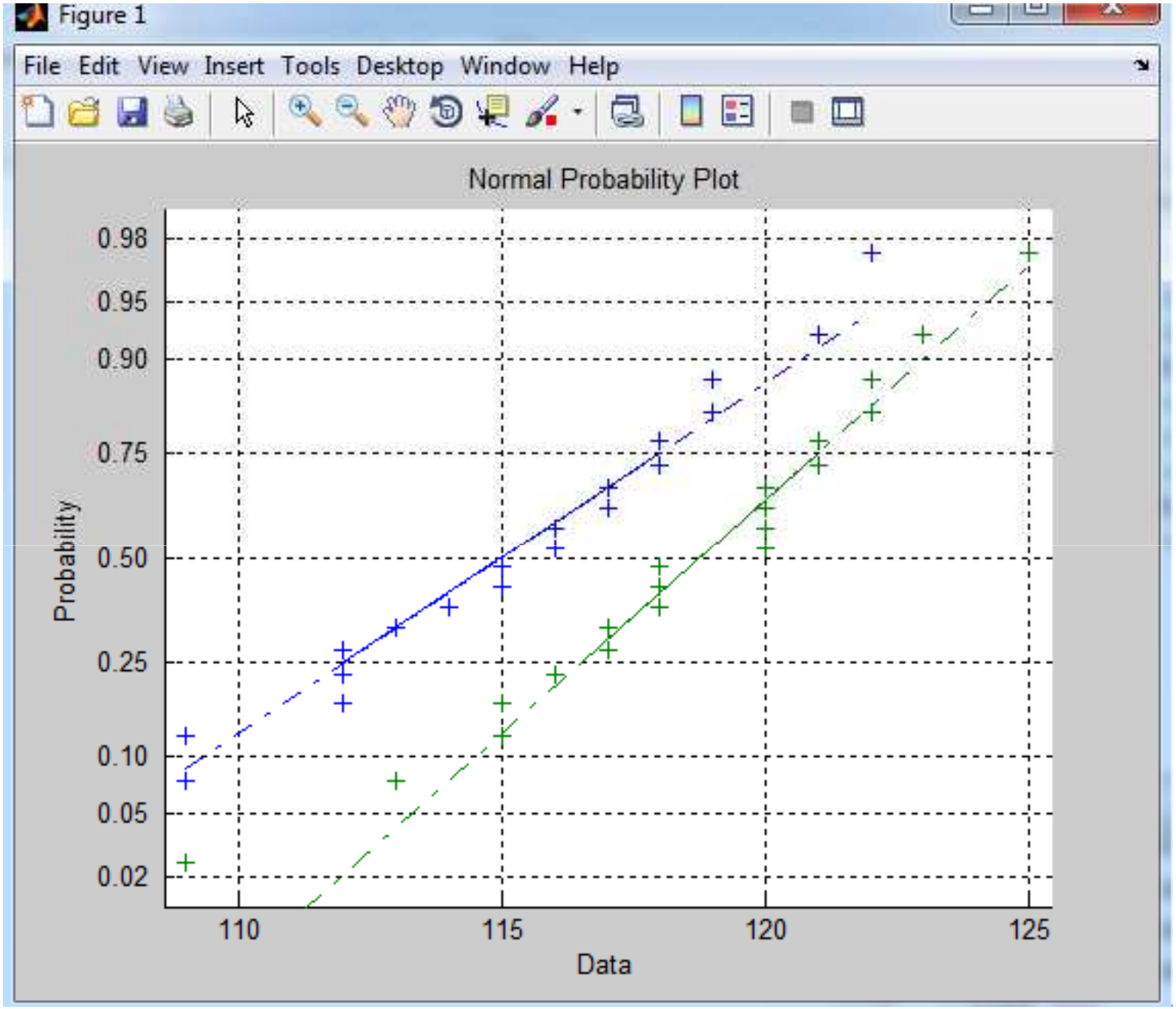


Figure 1
File Edit View Insert Tools Desktop Window Help
Normal Probability Plot
Probability
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
110
115
120
125
Data

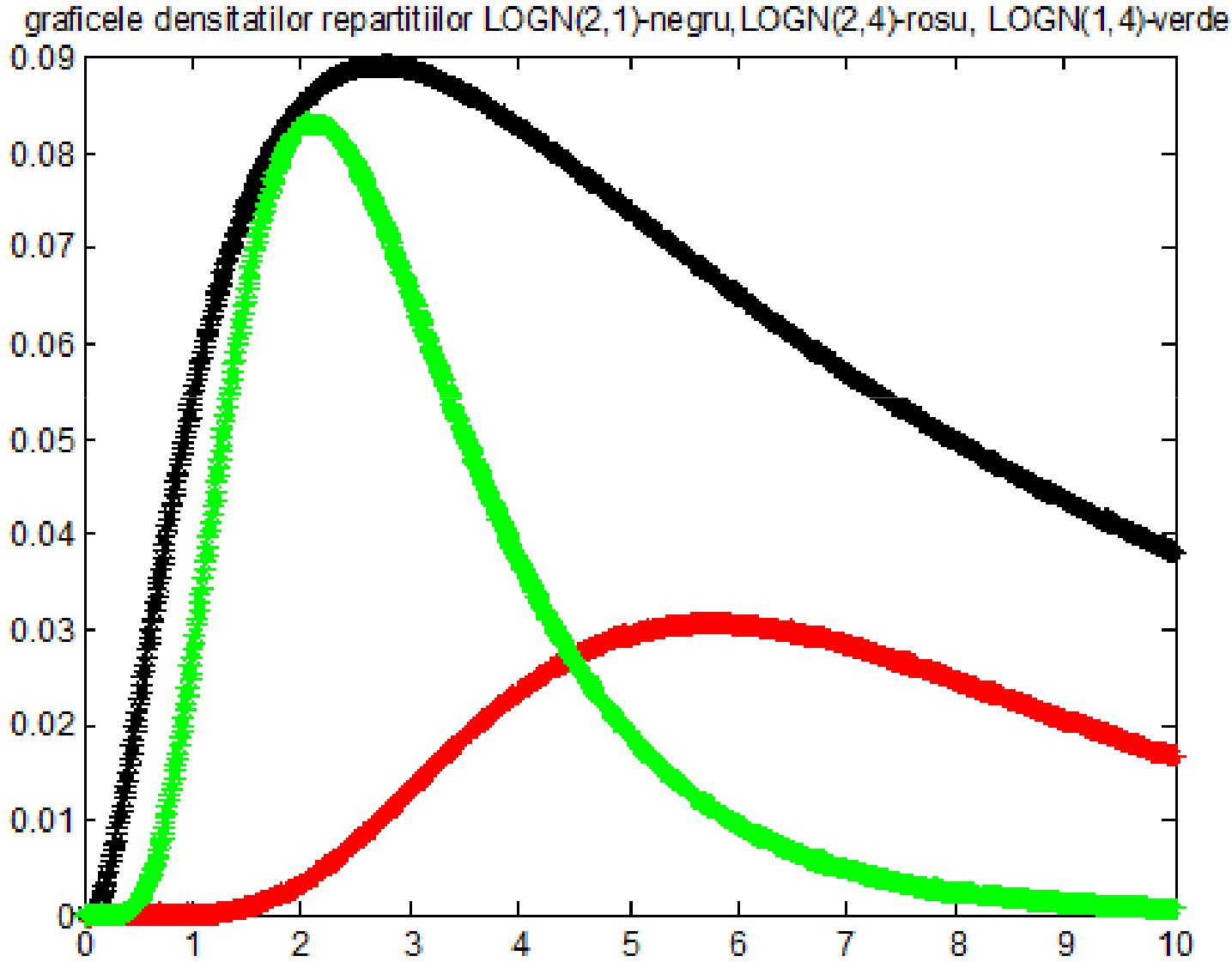
Repartiţia log-normală

Repartiţia *log-normală* $X \sim LOGN(\mu, \sigma^2)$: variabila $X \geq 0$ este *log-normal repartizată* dacă variabila $ln(X)$ să fie normal repartizată.

Densitatea de repartiţie este:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \; x > 0 \, ,$$

```
>> x=0:.01:10; plot(x,(1./(sqrt(2*pi)*x)).*exp(-(log(x)-2).^2/2),'k*',
x,(1./(sqrt(2*pi)*2*x)).*exp(-(log(x)-2).^2/2*2^2),'r*',
x,(1./(sqrt(2*pi)*2*x)).*exp(-(log(x)-1).^2/2*2^2),'g*')
```

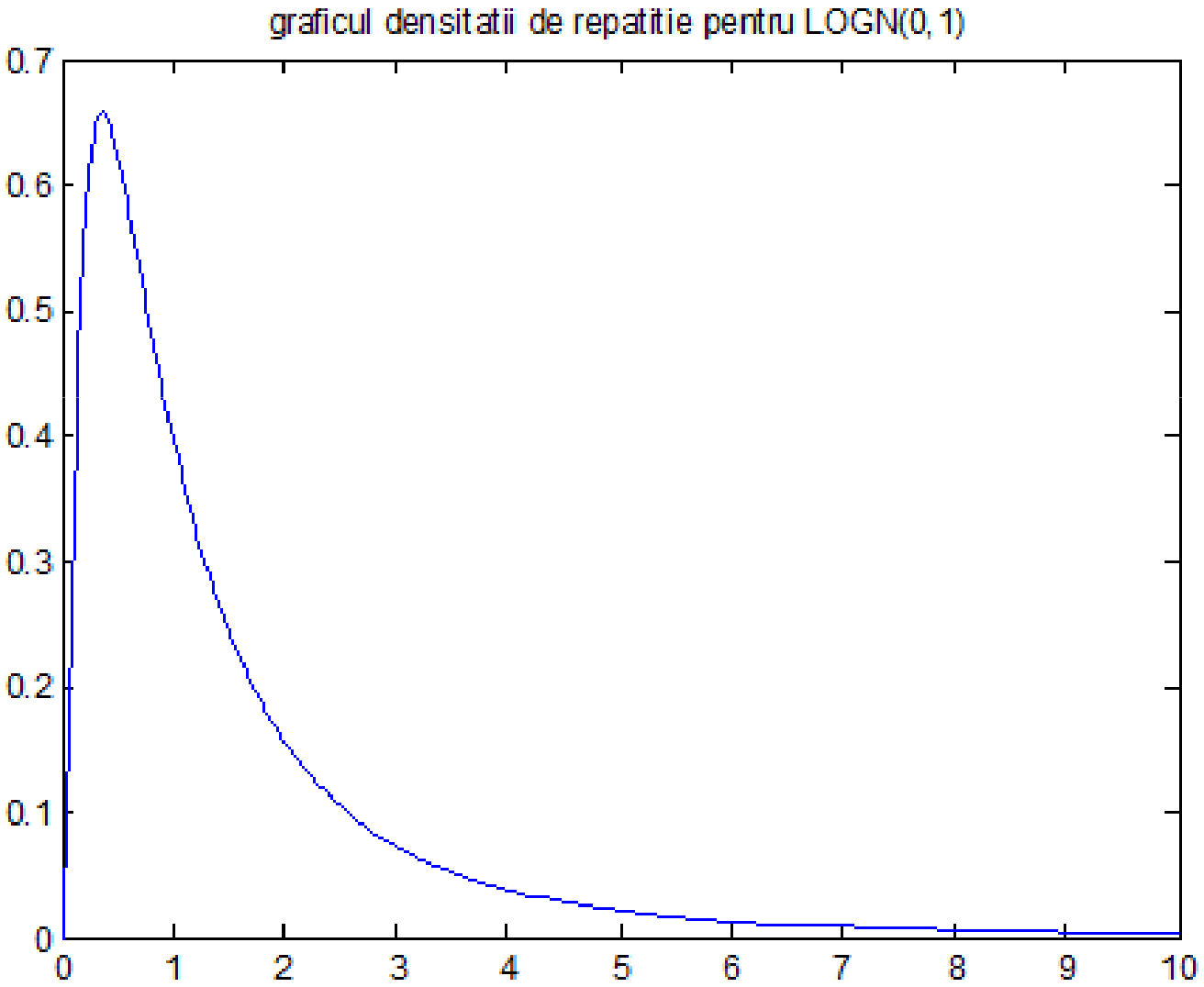


Graficul densitaţii de repartiţie (*probability density function* – pdf) în cazul repartiţiei lognormale poate fi desenat folosind sintaxa

```
Y = lognpdf(X,mu,sigma)
```

care returnează valorile în X ale densităţii de repartiţie lognormale de parametri mu şi sigma. Reamintim că mu şi sigma reprezintă media şi deviaţia standard a repartiţiei normale asociate X.

```
>> x = (0:0.02:10);
>> y = lognpdf(x,0,1);
>> plot(x,y)
```



Prezentăm formulele de calcul ale mediei şi dispersiei variabilei lognormale X:

$$m = e^{\mu + \frac{\sigma^2}{2}};$$

$$D = e^{(2\mu+\sigma^2)\cdot(e^{\sigma^2}-1)}$$

O repartiţie lognormală de medie m şi dispersie are ca parametrii:

$$\mu = \ln \frac{m^2}{\sqrt{D+m^2}}$$

$$\sigma = \sqrt{\ln\left(\frac{v}{m^2}+1\right)}$$

17. Exemplu

Pentru a calcula media şi dispersia unei variabile lognormale în Matlab folosim funcţia lognstat.

Să generăm 15 numere aleatoare lognormal repartizate având media 1 şi dispersia 2:

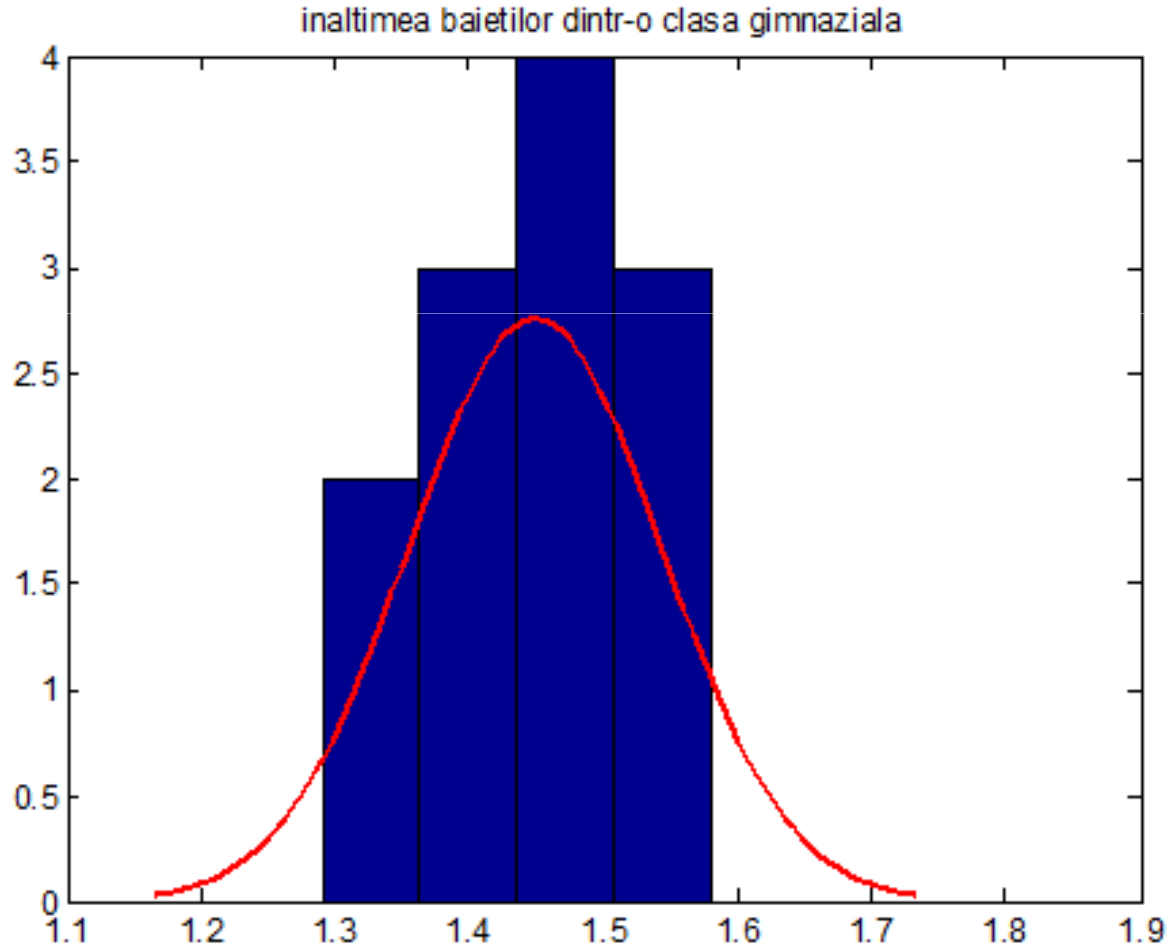
```
>> m = 1;v = 2;mu = log((m^2)/sqrt(v+m^2))
mu =
   -0.5493
>> sigma = sqrt(log(v/(m^2)+1))
sigma =
    1.0481

>> X = lognrnd(mu,sigma,1,15)
X =
  Columns 1 through 9
    0.5187    0.4483    0.8068    0.8014    0.2332    0.5594    0.4857    1.1147
1.8159
  Columns 10 through 15
    1.8467    0.2335    0.6261    0.1617    0.1797    0.5732
>> MX = mean(X);VX = var(X);[MX VX]
ans =
    0.6937    0.2809
```

histfit

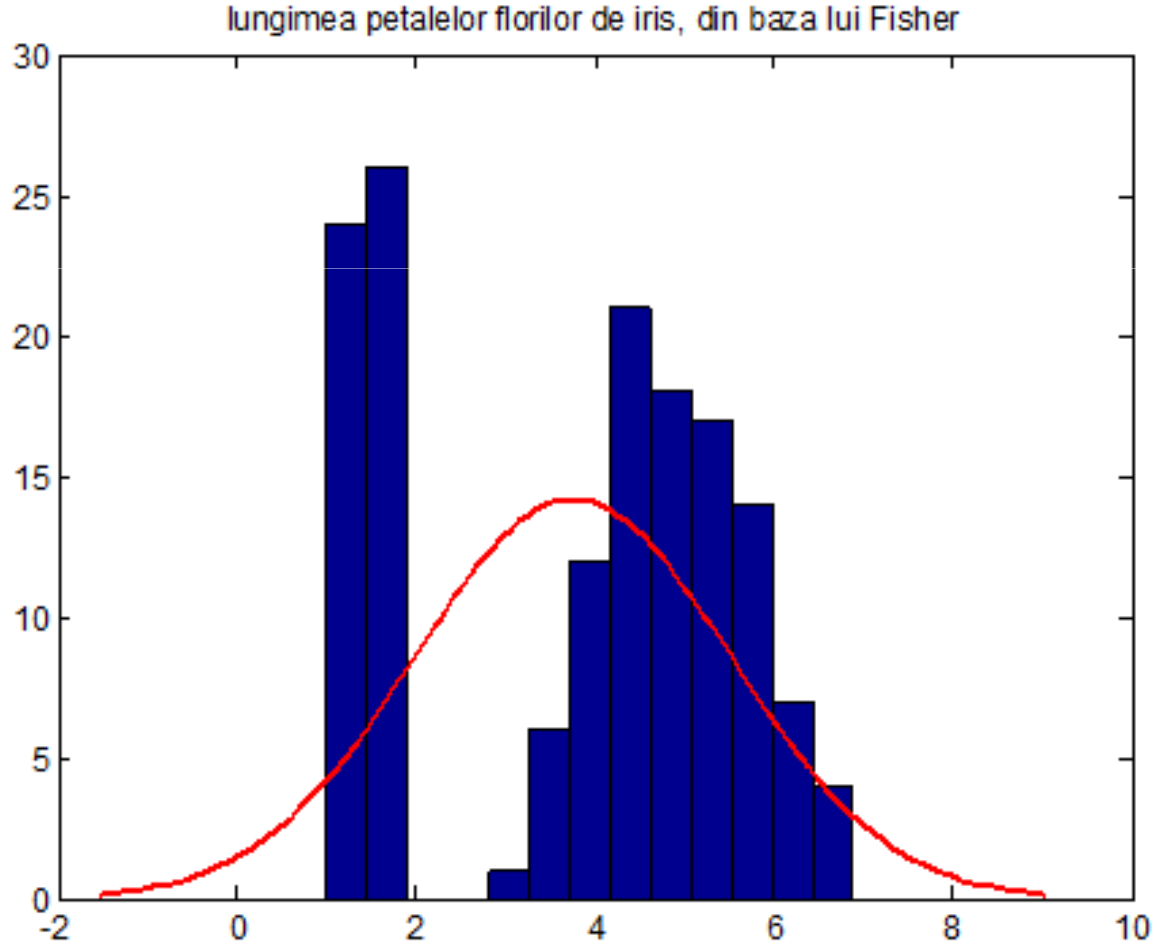
histfit(X) crează o histogramă a valorilor vectorului X, utilizând un număr de intervale egal cu [$\sqrt{\dim X}$] +1, căreia îi suprapune o densitate de repartiţie normală. Reluăm exemplul nr 5:

>> X1=[1.40, 1.37, 1.57, 1.46, 1.49, 1.46, 1.39, 1.55, 1.29, 1.58, 1.50, 1.33]; histfit(X1)



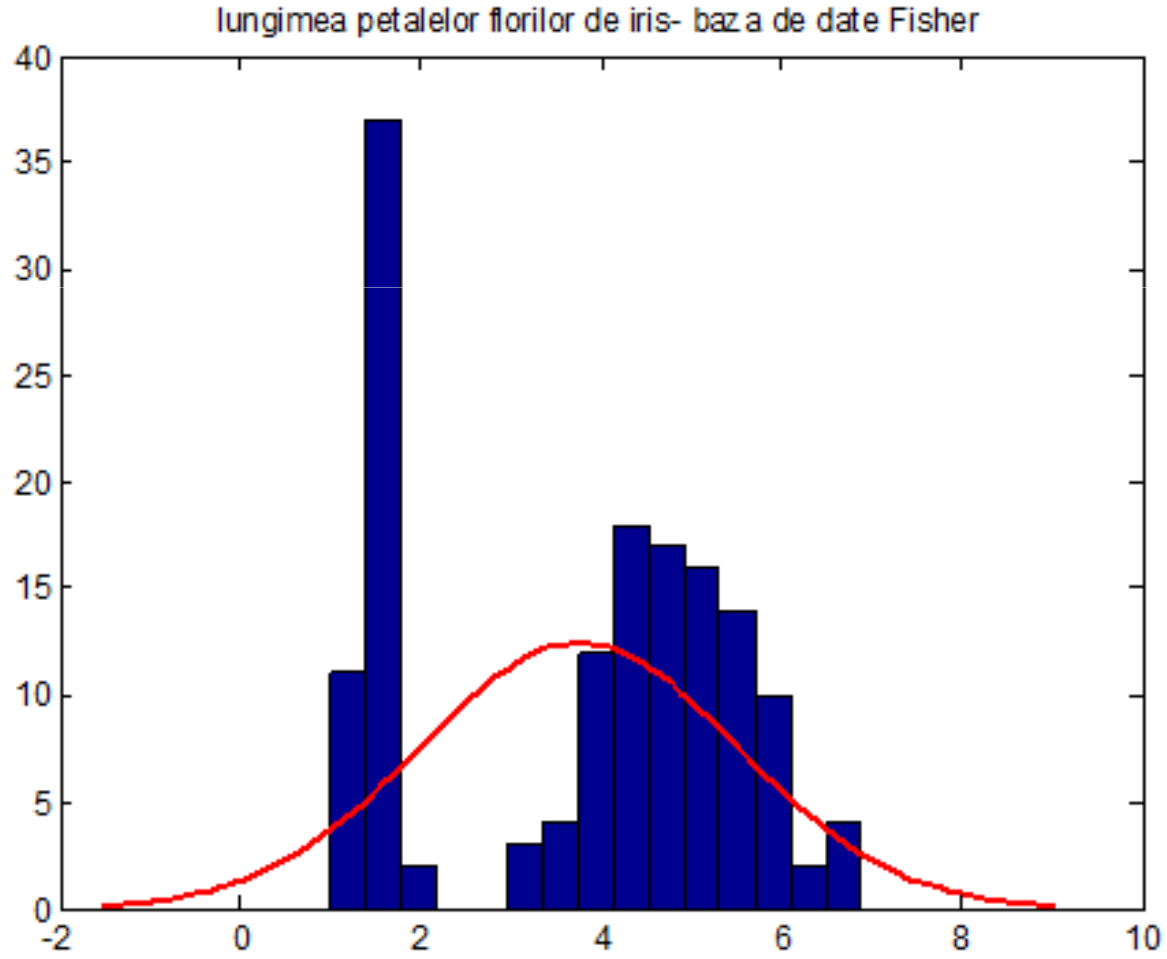
Exemplificăm cu vectorul ce reprezintă lungimea petalelor florilor de iris (Fisher)

```
>> load fisheriris
>> x=meas(:,3); histfit(x)
```



histfit(X,n) crează o histogramă a valorilor vectorului X, utilizând *n* intervale, căreia îi suprapune o densitate de repartiţie normală. Exemplificăm cu vectorul ce reprezintă lungimea petalelor florilor de iris (Fisher)

```
>> load fisheriris
>> x=meas(:,3); histfit(x,15)
```

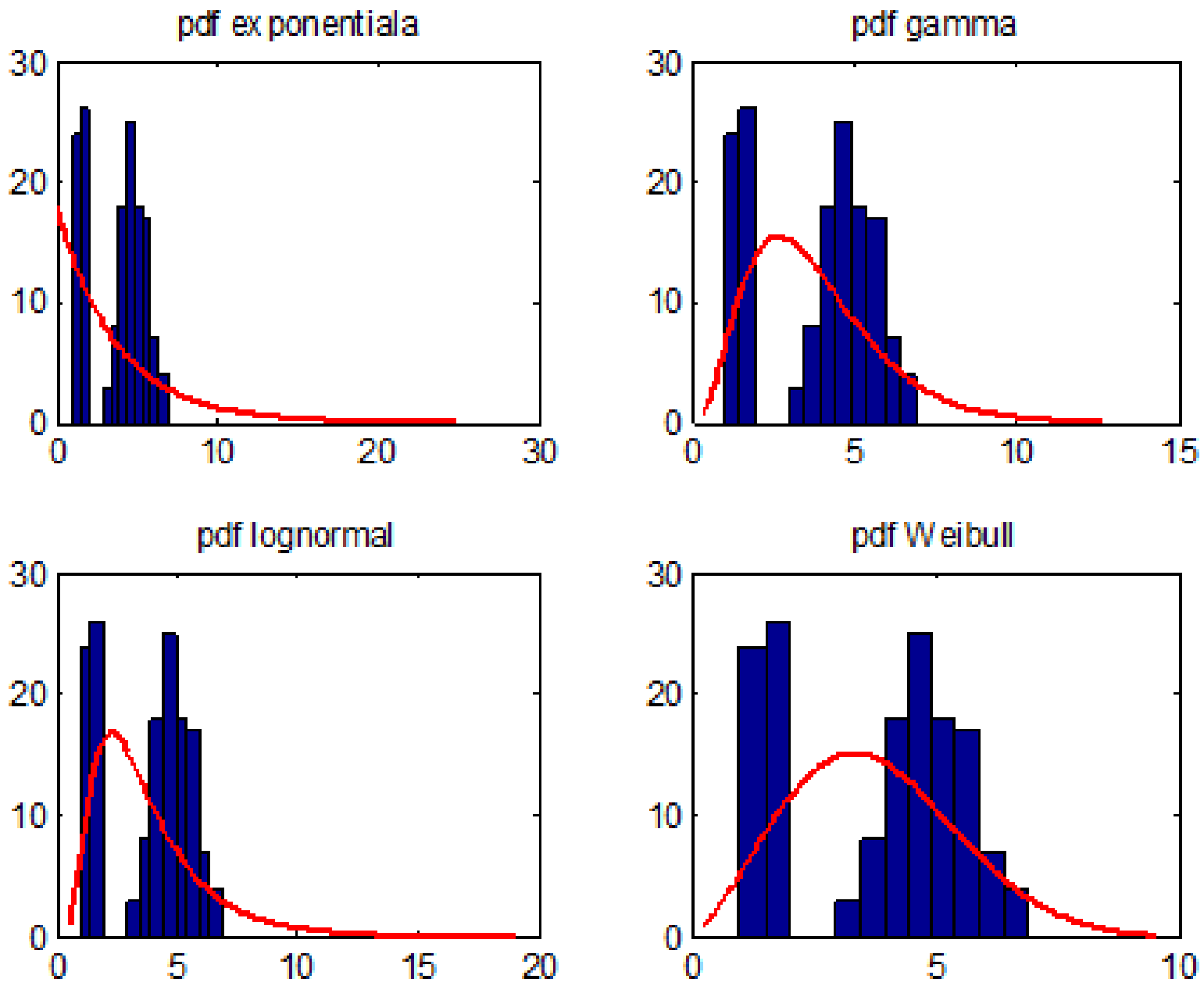


histfit(X,n,'repartitie') crează o histogramă a valorilor vectorului X, utilizând *n* intervale, căreia îi suprapune o densitate de repartiţie impusă:

- 'exponential'
- 'gamma'
- 'lognormal'
- 'normal'
- 'weibull' sau'wbl'

Prentăm histograma lungimii petalelor irişilor din baza de date Fisher, suprapunând câte o densitate de repartiţie:

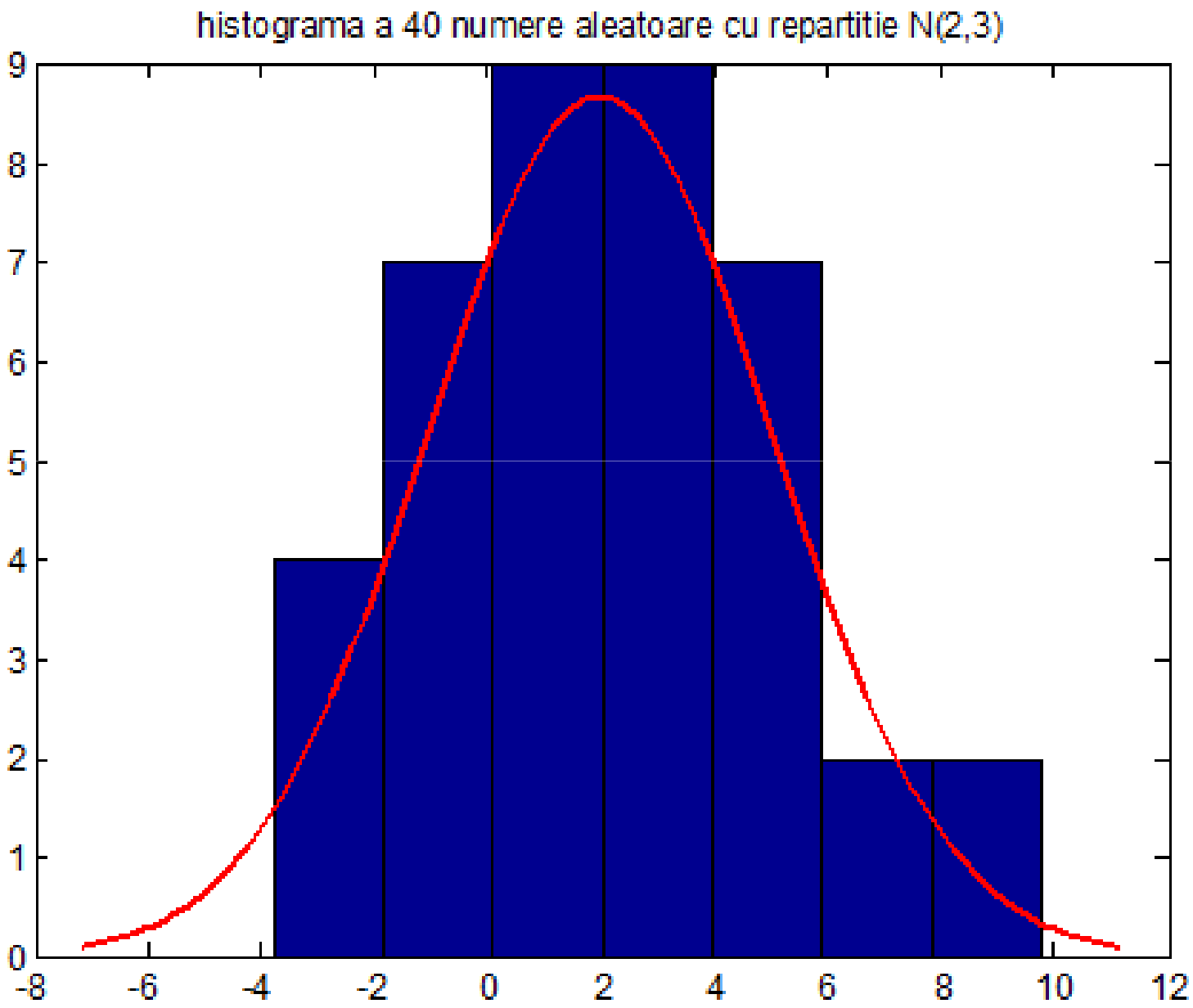
```
>> load fisheriris
>> x=meas(:,3);
>> subplot(221);histfit(x,12,'exponential')
>> subplot(222);histfit(x,12,'gamma')
>> subplot(223);histfit(x,12,'lognormal')
>> subplot(224);histfit(x,12,'wbl')
```


pdf exponentiala
pdf gamma
pdf lognormal
pdf Weibull
0
10
20
30
0
5
10
15
20

18. Exemplu

Generăm aleator 40 de numere, cu repartiţie normală N(2,3), desenăm histograma lor şi suprapunem graficul densităţii de repartiţie normale:

```
>> R = normrnd(2,3,1,40);
>> histfit(R)
```


histograma a 40 numere aleatoare cu repartitie N(2,3)
9
8
7
6
5
4
3
2
1
0
-8
-6
-4
-2
0
2
4
6
8
10
12

Teste statistice

Plecând de la realitatea înconjurătoare, formulăm diferite ipoteze pe baza informaţiilor primite, ipoteze trebuiesc verificate (testate). După testare, care este un proces complex, ipoteza respectivă devine ipoteză *adevărată*.

Din punct de vedere statistic ipotezele fac referire la fenomene şi relaţii ce privesc populaţiile statistice. Este vorba de ipoteze privind repartiţii de probabilitate, valoarea unor parametri statistici sau legăturile dintre aceştia, plecând de la observaţiile făcute pe un eşantion de populaţie.

Există două tipuri de teste statistice:

- testele parametrice se referă la ipoteze statistice ce privesc parametrii statistici (medie, dispersie);
- testele non-parametrice se folosesc pentru studiul datelor care nu sunt guvernate de repartiția normală, aplicate mai ales la date care sunt guvernate de legi de probabilitate asimetrice.

Formularea şi testarea ipotezelor este o parte esenţială a inferenţei statistice dinspre eşantion spre populaţia generală.

În scopul testării se formulează o ipoteză de studiu pe care trebuie să o demonstrăm. Poate fi vorba despre un raţionament care este considerat corect sau care poate fi folosit ca bază de comparaţie.

De exemplu o ipoteză poate fi:

un tratament nou este mai bun decât tratamentul folosit curent pentru aceeaşi afecţiune.

O testare statistică se bazează pe confruntarea a două ipoteze:

- ipoteza nulă H_0 care afirmă că nu există nicio diferenţă semnificativă între datele comparate, diferenţa obţinută datorându-se doar întâmplării.
  - Când comparăm media de sondaj cu media cunoscută a populaţiei, H_0 afirmă că nu există diferenţă semnificativă între cele două valori, deci eşantionul este reprezentativ pentru populaţia originară.
  - In cazul a două eşantioane, în contextul comparaţiei unor anumiţi parametri, H_0 afirmă că cele două eşantioane aparţin aceleiaşi populaţii.

- ipoteza alternativă H_1 care afirmă că efectul comparaţiei este nenul, existând diferente semnificative, nedatorate hazardului privind datele considerate.

De exemplu, în studiul unui nou medicament se spune:

H_0 = *noul medicament nu este mai eficient, în medie, decât medicamentul curent*.

Trebuie acordată o mare importanță ipotezei nule pentru că este în strânsă relație cu ipoteza alternativă, care trebuie confirmată.

Neacceptarea ipotezei nule, confirmă aplicarea ipotezei alternative H_1 .care este în fapt scopul studiului:

H_1 : *noul medicament este mai eficient, în medie, decât medicamentul curent*.

Concluzia finală, după aplicarea unui test de semnificaţie se exprimă întotdeauna în termenii ipotezei nule:

Respingem H_0, caz ce sugerează că ipoteza alternativă poate fi adevărată

Nu respingem H_0, caz ce sugerează că nu există destule argumente pentru respingerea ipotezei nule.

Ipoteza alternativă este o afirmaţie pe care aplicarea testului de semnificaţie statistică doreşte să o confirme, prin neinfirmare.
Formularea acestei ipoteze poate lua trei forme:

$H_1 \neq H_0$, *de exemplu: efectul noului medicament este diferit de cel curent*

$H_1 > H_0$, *de exemplu: efectul noului medicament este mai bun decât cel curent*

$H_1 < H_0$, *de exemplu: efectul noului medicament este mai prost decât cel curent*

Intrebări:

- Care este riscul de a respinge ipoteza nulă, ea fiind totuşi adevărată?
- Care este riscul de a o accepta , ipoteza nulă dovedindu-se în final falsă?

Pot fi făcute două tipuri de erori, noţiunile statistice următoare precizând acest fapt:

- Eroarea (riscul) de primă speţă, notată α , este probabilitatea ca să respingem ipoteza nulă, atunci când ea este adevărată; valoarea lui α este considerată uzual 5% şi este aleasă înainte de a incepe analiza. Decizia este greşită, fiind un rezultat *fals pozitiv*.
- Eroarea (riscul) de speţa a doua, notată β , se refera la probabilitatea de a accepta ipoteza nulă, atunci când ea este falsă; valoarea lui β depinde de diferenţa obţinută în analiza cât şi de volumul eşantionului. Decizia este greşită, fiind un rezultat *fals negativ*.

Puterea testului, notată cu π şi egală cu 1- β reprezintă probabilitatea de a respinge ipoteza nulă, atunci când este falsă. Un interval de încredere mare indică o putere scăzută.

Procedura de testare constă în maximizarea puterii testului, ceea ce înseamnă minimizarea erorii de speţa a doua, atunci când eroarea de prima speţă este limitată a priori arbitrar.

- se stabileşte că ipoteza nulă este adevărată, adică nu există nicio diferenţă semnificativă în comparaţia făcută;
- interpretarea rezultatului obţinut după procesarea datelor: Care este probabilitatea de a fi obţinut rezultatul respectiv, dacă ipoteza nulă ar fi adevărată? Această probabilitate este cunoscută sub numele de *nivelul de semnificaţie p*.

Presupunem că s-au colectat date din două eşantioane şi că mediile acestora sunt diferite. A observa că cle două medii sunt diferite nu este suficient pentru a concluziona că populaţiile au medii diferite.
E posibil ca populaţiile să aibă aceeaşi medie, iar diferenţa observată să fie o coincidenţă în urma eşantionării aleatorii.
Nu există nici un mod de a fi siguri că diferenţa observată reflectă o diferenţă reală sau este o consecinţă a eşantionării aleatorii.

Valoarea *p* răspunde la această întrebare: dacă populaţiile au avut într-adevăr aceeaşi medie, care este probabilitatea de a observa o asemenea diferenţă (sau una mai mare) între mediile eşantioanelor într-un experiment de dimensiunea acestuia?
Dacă valoarea *p* este mică, concluzia este că diferenţa are şanse mici să fie provocată de eşantionarea aleatorie şi se poate concluziona că populaţiile au medii diferite.

Pentru detectarea semnificaţiei statistice trebuie urmate etapele de inferenţă:

1. Stabilirea unei valori prag, înainte de a efectua experimentul; în mod tradiţional valoarea - prag (numită α) este stabilită la 0,05 sau 0,01.
2. Definirea ipotezei nule.
3. Se aplică testul statistic aferent pentru a calcula valoarea *p*.
4. Se compară valoarea *p* cu valoarea – prag.
   - Dacă valoarea p este mai mică decât pragul stabilit, se declară că *se respinge ipoteza nulă* şi că diferenţa *e semnificativă din punct de vedere statistic*.
   - Dacă valoarea *p* este mai mare decât pragul, se declară că *nu se respinge ipoteza nulă* şi că diferenţa *nu e semnificativă din punct de vedere statistic*. Nu se poate concluzia că ipoteza nulă este adevărată. Se poate spune doar că nu avem probe suficiente pentru a respinge ipoteza nulă.

Testul de semnificaţie calculează valoare *p*, probabilitatea de a obţine o valoare corectă a parametrului la nivel de populaţie, prin extrapolare (inferenţă) de la valoarea estimatorului (calculat pe baza eşantionului).

Testarea ipotezelor implică posibile erori:

Decizia statistică	H_0 real adevărată	H_0 real falsă
H_0 respinsă	Eroare tip I (α)	Corect
H_0 acceptată	Corect	Eroare tip II (β)

Eroarea de tip I este numită nivel de semnificaţie, valoarea ei fiind stabilită de cercetător.
Populaţiile sunt identice, deci nu există de fapt nici o diferenţă. Din întâmplare s-au obţinut valori mai mari în cazul unui grup şi mai mici în cazul celuilalt.

Când apare un rezultat semnificativ din punct de vedere statistic – în cazul în care populaţiile sunt identice ($H_0 = H_1$) înseamnă că a intervenit o eroare de tip I.

Dacă se defineşte $p < 0.05$ ca fiind semnificaţia statistică, va apare eroare de tip I în 5% din experimentele în care nu există nici o diferenţă.

Nivelul de semnificaţie p este o cuantificare a influenţei hazardului asupra diferenţei obţinute prin comparaţie, eşantionul fiind presupus reprezentativ pentru populaţia originară.

$p = 0.05$ nivel statistic semnificativ la limită

$p = 0.01$ nivel statistic semnificativ

$p = 0.005$ nivel înalt semnificativ

Reamintim că:

Regiunea critică reprezintă mulţimea valorilor x ale variabilei statistice corespunzătoare X pentru care se respinge ipoteza nulă.

Regiunea de încredere a unui parametru este mulţimea valorilor θ_0 pentru care se acceptă ipoteza nulă.

Problema practică:

Cum putem calcula probabilitatea P de a obţine efectul respectiv, plecând de la datele eşantionului, presupunând ipoteza nulă adevărată?

Se calculează testul statistic, care este o funcţie de variabilele de sondaj, plecând de la o anumită cantitate calculată pe baza sondajului şi comparată cu o valoarea ipotetică, propusă, presupunând că ipoteza nulă este adevărată.

$$test\ statistic = \frac{valoare\ de sondaj - valoare\ ipotetica}{\frac{deviatia\ st.}{\sqrt{n}}}$$

unde prin valoare de sondaj înţelegem estimatorul parametrului căutat.

Algoritm de lucru

1. Se formulează setul de ipoteze H_0, H_1.
2. Se calculează dintr-un eşantion o statistică (test statistic)
3. Dacă în lumea reală are loc un eveniment, se calculează în ipoteza H_0, probabilitatea p de apariţie a valorii calculate.
4. Dacă p este mică apare o contradicţie, pentru rezolvarea căreia se va respinge H_0 în favoarea lui H_1 (p este mică deoarece în calculul acesteia s-a acceptat ipoteza H_0).
5. Dacă p este mare nu se respinge H_0, neexistând nici un motiv pentru a lua decizia contrară.

Rămâne o singură întrebare: începând de unde o probabilitate este Considerată drept "mică"? Pentru a nu introduce subiectivismul în această decizie, se fixează, anterior deciziei în test, un prag sub care o probabilitate este considerată "mică".

Această valoare se numește prag de semnificație și se notează uzual cu α.

Regula de decizie în test poate fi formulată atunci:

• dacă P≤ α, atunci se respinge ipoteza nulă, H_0, în favoarea ipotezei alternative, H_1;

• dacă P > α . atunci nu se respinge ipoteza nulă H_0.

Reamintim:Repartiţia t-Student - W.S Gosset alias *student* 1908

Repartiţia Student (t -*distribution*) este o familie de curbe ce depind de un singur parametru n, care reprezintă gradele de libertate. Dacă $n \to \infty$, atunci repartiţia t se apropie de repartiţia normală standard

Dacă x este un eşantion de dimensiune n dintr-o repartiţie normală de medie μ, atunci testul statistic $t = \dfrac{\overline{X} - \mu}{\dfrac{\overline{\sigma}}{\sqrt{n}}}$ are o repartiţie Student cu n-1 grade de libertate.

Funcţia de repartiţie Student este:

$$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

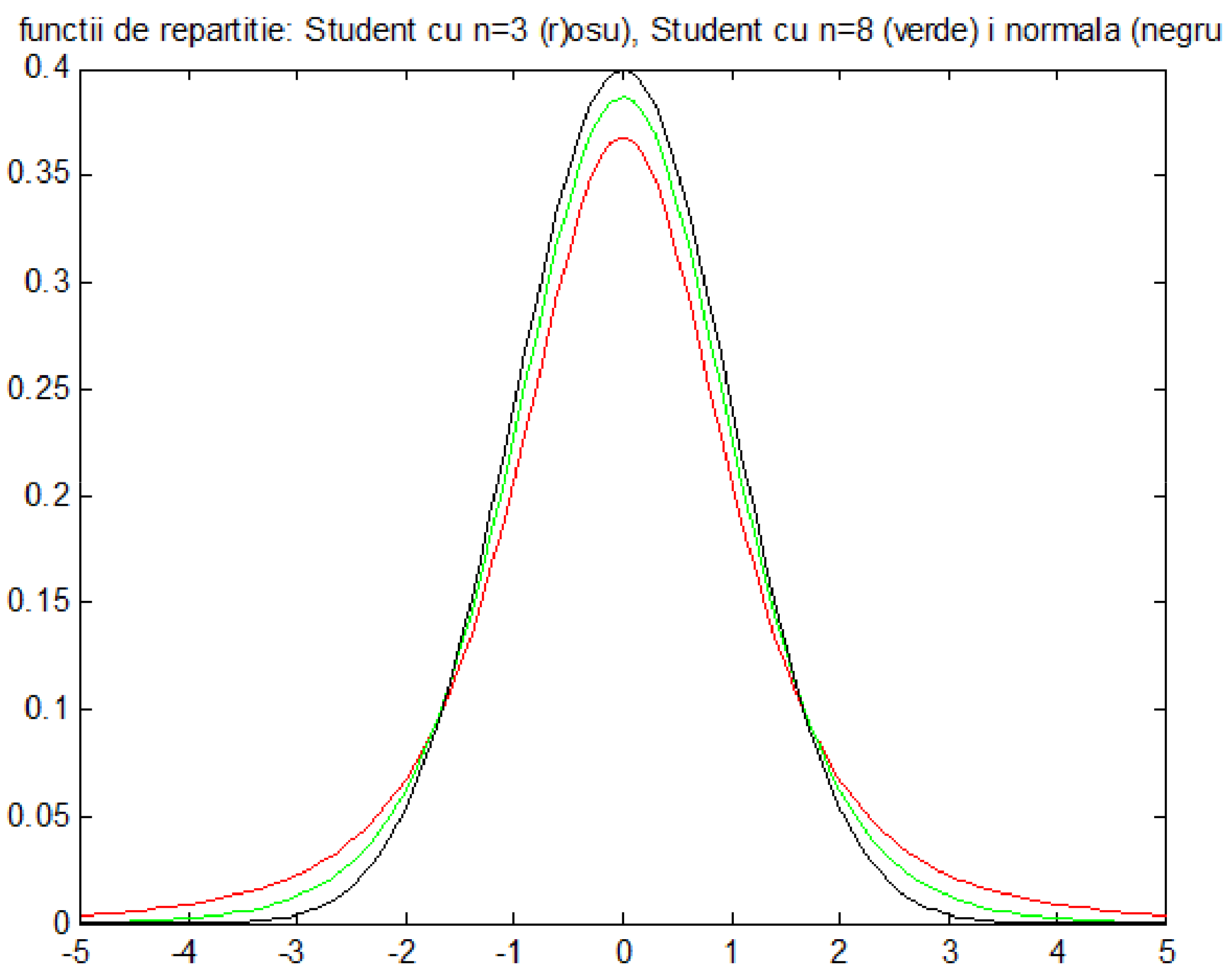
unde $\Gamma(n)=\int_0^{\infty}t^{n-1}\cdot e^{-t}dt$, (funcţia gamma)

19.Exemplu

Vom desena în acelaşi cadran functia de repartiţie Student, pentru n = 3, respectiv pentru n = 8 şi funcţia de repartiţie normală.

y = tpdf(x,n) calculează valoarea funcţiei de repartiţie Student, în fiecare componentă a vectorului x, folosind numărul de grade de libertate date de n.

```
>> x = -5:0.1:5;
>> y1 = tpdf(x,3);
>> y2 = tpdf(x,8);
>> z = normpdf(x,0,1);
>>plot(x,y1,'r',x,y2,'g',x,z,'k')
```


functii de repartitie: Student cu n=3 (r)osu), Student cu n=8 (verde) i normala (negru
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

20. Exemplu

In cazul testării dacă media populaţiei originare are o anumită valoare ipotetică, propusă într-un studiu statistic, testul statistic este:

$$t = \frac{\overline{X} - \mu}{\frac{\overline{\sigma}}{\sqrt{n}}}, \text{ unde:}$$

- $\overline{X}$ este media de sondaj,
- μ este valoarea ipotetică a mediei populaţie;
- $\overline{\sigma}$ este deviaţia stansard.

Concret, se alege se alege ca medie ipotetică propusă o anumită valoare μ_0, presupunându-se adevărată ipoteza nulă $H_0 : \mu = \mu_0$, ceea ce implică faptul că testul statistic urmează o lege Student cu un anumit număr de grade de libertate.

Plecând de la valoarea concretă a lui t, se obţine valoarea corespunzătoare a lui P, din tabelul repartiţiei Student.

Concret, dacă $\bar{x} = 6753.6, \quad \mu_0 = 7725, \quad \bar{\sigma} = 1142.1, \quad n = 11,$

obţinem $t = 2.821.$

Prezentăm o parte dintr-un tabel corespunzător unei repartiţii Student cu grade de libertate între 1 şi 14.

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	**2.764**	**3.169**	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

În cazul nostru având 10 grade de libertate, valoarea lui *t* corespunde la un nivel de semnificaţie (two tailed) între 98% şi 99%.

Dintr-un tabel complet am calculat P = 0.018 , deci diferenta obţinută este semnificativă statistic.

Alegerea numărului de grade de libertate se bazează pe faptul că ipoteza H_0 este adevărată.

Odată testul statistic calculat, alegerea lui P se poate face în două moduri: *one-tailed* sau *two-tailed*, în funcţie de coada sau cozile repartiţiilor de probabilitate ce apar pentru valorile extreme.

Testul chi pătrat de independenţă - test neparametric

Reamintim că dacă $X_1,\ldots,X_n$ sunt variabile gaussiene standard, atunci variabila $X^2 = X_1^2 + \ldots + X_n^2$ urmează legea χ^2 cu n grade de libertate. Densitatea sa de repartiţie este

$$f_n(x) = \frac{2^{1-\frac{n}{2}} \cdot x^{n-1} \cdot e^{-\frac{x^2}{2}}}{\Gamma\left(\frac{n}{2}\right)}$$

Tabelul de contingenţă reprezintă o clasificare a datelor în funcţie de 2 criterii în cadrul carora datele sunt în continuare divizate in 2 sau mai multe categorii discrete şi mutual exclusive.

De exemplu:
Considerăm o populaţie statistică formata din copii. împărţiţi pe sexe: fetiţe si băieţi şi vom studia care este preferinţa acestora faţa de trei tipuri de jucăii A, B şi C, pe baza analizei unui eşantion de copii.
Tabelul de contingenţă corespunzător este:

		Tipuri de jucării		
		A	B	C
Sex	Fetiţe	50	20	15
	Băieţi	0	30	54

Analiza acestui tabel se bazează pe testarea ipotezelor.

Ipoteza nulă se referă la presupunerea că nu există nicio relaţie semnificativă la nivelul populaţiei de copii între cele două clasificări : sexul copiilor şi tipurile de jucării, adică variabilele sunt considerate independente.

In acest caz vom compara frecvenţele observate cu cele aşteptate, dacă ipoteza nulă ar fi adevarată.

Notăm cu O_{ij} frecvenţele observate şi cu E_{ij} frecvenţele aşteptate dacă ipoteza nulă ar fi adevarată, unde *i* indică numărul liniei, iar *j* indică numărul coloanei din tablou.

Sub auspiciile ipotezei nule statistica $\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

urmează o repartiţie χ^2 cu $(p-1)\cdot(q-1)$ grade de libertate, unde *p* este numărul de linii şi *p* este numărul de coloane din tabelul de contingenţă.

Prezentăm o parte din tabelul ataşat testului χ^2 de independenţă, pâna la 16 grade de libertate.

n	Probability less than the critical value 0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252

21. Exemplu

Tabelul de contingenţă următor ilustrează legătura între statutul marital şi consumul de cafea:

		Consum cofeină (mg/zi)				total
		0	1-150	151-300	>300	
Status marital	căsătorit	652	1537	598	242	3029 77.5%
	divorţat sau văduv	58	46	38	21	163 4.2%
	necăsătorit	218	327	106	67	718 18.3%
total		928	1910	742	330	3910

Din cei ce nu consumă cafea

- 928*77.5/100 ar trebui să fie căsătoriţi,
- 928*4.2/100 ar trebui să fie divorţaţi sau văduvi
- 928*18.3/100 ar trebui să fie necăsătoriţi

s.a.m.d

Obţinem astfel tabelul cu frecvenţele aşteptate :

		Consum cofeină (mg/zi)			
		0	1-150	151-300	>300
Status marital	căsătorit	719	1480	575	256
	divorţat sau văduv	39	80	31	14
	necăsătorit	170	350	136	60
total		928	1910	742	330

Aplicând formula de mai sus obținem: $\chi^2 = 51.61$, ceea ce corespunde pentru repartiția χ^2 cu $3 \cdot 2 = 6$ grade de libertate, unui nivel de semnificatie P=0.00, ceea ce înseamnă că există o asociere înalt semnificativăîntre statutul marital şi consumul de cafea (variabile dependente).