Laborator 4
Analiza și prelucrarea datelor cu erori aleatorii sau cu evolutie aleatorie

1. Notiuni teoretice

Erorile aleatorii ***(întâmplatoare/accidentale)*** se caracterizeaza prin faptul ca, la repetarea masurarilor, în conditii identice, asupra aceluiasi masurand, apar diferente atât ca sens cât si ca valoare, nefiind posibil sa se stabileasca reguli de predeterminare pentru ele; caracterizarea lor poate fi facuta numai în sens probabilistic.

Pentru caracterizarea erorilor aleatorii, continute în datele experimentale, se **aplica metoda selectiei (esantionului)**, în sensul ca se considera un sir de masurari individuale V_1, V_2, ..., V_n, în care n se numeste **volumul selectiei (o selectie de volum n).**

Se parcurg următoarele etape:

A) Deoarece, în forma obtinuta prin experiment, datele constituie o multime dezordonata de valori, **se ordoneaza datele V_i în sens crescator, formându-se *seria variationala*;** rezulta ca V_1 apare de n_1 ori, V_2 de n_2 ori,..., V_k de n_k ori, cu:

$$V_1 < V_2 ... V_k \quad \text{si} \quad \sum_{i=1}^{k} n_i = n$$

B) Se reprezinta rezultatele masurarilor individuale printr-o ***histograma*** sau printr-un ***poligon de frecvente***.

Histograma este o reprezentare in plan, pe ordonata considerandu-se *frecventa absoluta de aparitie* n_i, sau *frecventa relativa de aparitie* $f_i=n_i/n$, iar pe abscisa se împarte domeniul valorilor $[V_{min}, V_{max}]$ în intervale elementare de lungime Δ, denumite *intervale de grupare sau de clasa*, Δ fiind dat de **formula lui Sturges**:

$$\Delta = \frac{V_{max} - V_{min}}{1 + 3{,}22 \cdot \lg n}$$

Ca punct de plecare în constructia histogramei se considera - în general - valoarea $\mathbf{V}_{min}$, astfel ca datele se grupeaza în intervalele $[V_{min} + i\Delta; V_{min} + (i+1)\Delta)$ deschise la dreapta.

C) R*epartitia empirica este prelucrata statistic* în vederea obtinerii unor **valori tipice de selecție** ca: **media** $\boldsymbol{m_v}$, modul (dominanta) *Mo*, mediana *Me*, dispersia de selectie μ, dispersia reala σ, **estimația** $\hat{\sigma}$, eroarea medie absoluta θ etc, care reprezinta indicatori sintetici esentiali pentru evaluarea erorilor aleatorii.

Valoarea medie $\boldsymbol{m_v}$ este foarte importanta întrucât ea se considera, în mod obisnuit, ca marime de referinta (valoare conventionala). Se mai numeste și valoarea cea mai probabilă.

$$m_V = \frac{1}{n} \sum_{i=1}^{n} V_i$$

Estimația erorii standard (deviația standard) reprezinta valoarea medie a patratelor erorilor aparente:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (V_i - m_V)^2$$

Dispersia de selectie μ^2 reprezinta media aritmetica a patratelor erorilor aparente:

$$\mu^2 = \frac{1}{n} \sum_{i=1}^{n} \Delta V_i^2 = \frac{1}{n} \sum_{i=1}^{n} (V_i - m_V)^2$$

Eroarea medie absoluta de selectie θ reprezinta media aritmetica a erorilor reale în modul, dar in practică se foloseste $\hat{\theta}$ adica: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|V_i - m_V|$

Mai multe valori tipice de selecție sunt definite în cursul nr. 3.

2. Realizarea unei histograme in MATLAB folosind generatorul de numere aleatoare

In cazul in care nu avetti la dispozitie date experimentale pentru realizarea unei histograme în MATLAB, puteti folosi functia ***random*** pentru a genera numere aleatoare cu o distributie de probabilitate data. Acest mod de lucru este util în special pentru a realiza simulari. In acest caz vom simula un set de date pentru a exersa realizarea histogramelor.

Sintaxa pentru apelarea ***functiei random*** este:

- random(NUME,A) returneaza un vector de numere aleatoare, alese din distributia de probabilitate cu un parametru, specificata de NUME, cu valoarea parametrului A.
- random(NUME,A,B) sau random(NUME,A,B,C) returneaza un vector de numere aleatoare, alese dintr-o distributie de probabilitate cu valorile parametrilor A, B (si C).

Exista multe tipuri de distributii caree pot alese la utilizarea acestei functii. Introduceti ***help random*** în linia de comanda pentru a lansa o lista completa cu numele si parametrii de intrare.

Distribuţie	Parametrul de intrare A	Parametrul de intrare B
'bino' sau ' Binomial'	n: numărul de încercări	p: probabilitatea de succes pentru fiecare încercare
'exp' sau 'Exponential'	μ: media	-
'norm' sau 'Normal'	μ: media	σ: deviaţia standard

Odata creat setul de date cu ajutorul generatorului de numere aleatoare, puteti reprezenta grafic datele, folosind functia histograma.

Exemplul 1

Realizati o rutina MATLAB pentru a genera o selectie aleatoare de 1000 de puncte de date dintr-o distributie Gaussiana cu $m_v = 1$ si $\hat{\sigma} = 0.5$.

```
media=1;
dispersia=0.5;
N=1000;
date=random('Norm',media,dispersia,1,N);
figure(1);
hist(date)
title('Histograma cu media=1, dispersia=0.5') ;
```

Completati rutina Matlab cu instrucțiunile aferente pentru a obține alte 2 grafice în care se modifică, pe rând, valoarile pentru medie si apoi pentru estimație. Care sunt efectele modificării m_v ? Dar la mărirea valorii $\hat{\sigma}$?

3. Determinarea histogramei pentru un set de date

Exemplul 2

Se consideră un set de date care constituie notele obținute de studenti la un examen. Aceste date se gasesc in fisierul "Student/Note Stud.xls".

Se dorește găsirea valorilor tipice de selecție media m_v și estimația $\hat{\sigma}$. În vederea determinării acestor valori, datele din fișier trebuiesc importate ca și vector.

Folosiți Help-ul Matlab-ului pentru a intelege sintaxa instrucțiunilor ”xlsread”, ”mean”, ”std” pentru **determinarea m_v, $\hat{\sigma}$.** Realizați **histograma** aferenta acestui set de date, utilizând instrucțiunea ”hist”. Aceasta trebuie să rezulte ca în **figura 1**.

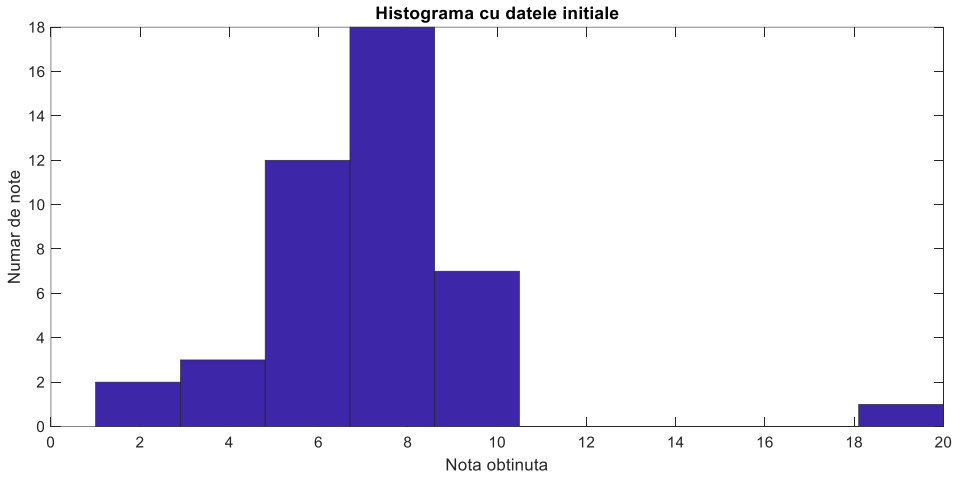

Fig.1. Histograma Exemplul 2

4. Exercitii propuse

4.1. La măsurarea dozelor anuale de radiații din incinta unei facilități de cercetare în domeniul nuclear sunt utilizate 10 aparate de măsură (dozimetre) situate în locații diferite. Doza maximă acceptată de radioactivitate pentru o persoană obișnuită, aprobată de Comisia Națională pentru Controlul Activităților Nucleare și de Organizația Mondială a Sănătății este de 1mSv (miliSievert). Datele înregistrate de cele 10 dozimetre se găsesc în tabelul următor:

Dozimetru	Nivel de radioactivitate [mSv]
1	$0.121*10^{-3}$
2	$0.217*10^{-3}$
3	$0.325*10^{-3}$
4	$0.517*10^{-3}$
5	$0.476*10^{-3}$
6	$0.192*10^{-3}$
7	$0.088*10^{-3}$
8	$0.235*10^{-3}$
9	$0.217*10^{-3}$
10	$0.112*10^{-3}$

Să se determine următoarele valori tipice de selectie pentru măsurătorile obținute de cele 10 dozimetre:

- media aritmetica *mv1*;
- mediana de selecție *me*;
- dispersia de selectie μ^2 *niu*;
- estimația $\hat{\sigma}$ *sig1*
- eroarea medie absoluta de selectie *teta*.

4.2. Numărul de mesaje trimise pe oră printr-o rețea de calculatoare este exprimat prin frecventa de apariție f_i și are următoarea distribuție:

x=numar de mesaje	10	11	12	13	14	15
$f_i(x)=n_i/n$	0.08	0.15	0.3	0.2	0.2	0.07

Calculați media aritmetică *mv2* și estimația standard *sig2* pentru numărul de mesaje trimise orar.

4.3. Se consideră setul de date ”Date_temp.xls”. Să se determine:
- media m_v *mv3* și estimația $\hat{\sigma}$ *sig3*
- histograma