

Laborator 5

Teste statistice pentru date afectate de erori grosiere

1. Notiuni teoretice

Erorile grosiere sunt acelea ale caror valori conduc la denaturari ale rezultatelor, acestea fiind apreciabil diferite de cele obtinute prin masurari similare. Determinarile afectate de erori grosiere nu participa la procesul de prelucrare, de aceea se procedeaza la depistarea si inlaturarea acestora.

Exista două categorii de **teste pentru depistarea erorilor grosiere**:

✓ teste pentru selectii de volum ridicat:

▲ **Chauvenet**

▲ **3 σ**

✓ teste pentru selectii de volum scazut:

▲ **Grubbs-Smirnov**.

Testul 3 σ consta în verificarea încadrării datelor experimentale într-un interval de încredere $[m_v - 3\sigma; m_v + 3\sigma]$ în care sunt admise erorile aleatorii (s-a presupus ca erorile sistematice au fost eliminate). Testul presupune urmatoarele etape:

P1. ordonarea datelor V_1, V_2, \dots, V_n în sens crescator, având ca extreme V_{\min} si V_{\max} ;

P2. calculul valorii medii m_v si a deviației erorii standard $\hat{\sigma}$;

P3. calculul intervalului de valori în care trebuie sa se încadreze datele:
 $[m_v - 3\sigma; m_v + 3\sigma]$

P4. toate datele care sunt în afara intervalului dat de relatia de la pasul 3 se elimina fiind considerate ca afectate de erori grosiere;

P5. se reia testul cu noile date ramase de la pasul 2, pâna când nici un rezultat experimental nu se găsește în afara intervalului dat de relatia de la pasul 3.

Testul Grubbs - Smirnov consta în verificarea încadrării valorilor extreme ale datelor experimentale si presupune urmatoarele etape:

P1. ordonarea datelor V_1, V_2, \dots, V_n în sens crescator (seria variatională), având ca extreme V_{\min} si V_{\max} ;

P2. calculul valorii medii m_v si a deviației erorii standard $\hat{\sigma}$;

P3. daca valoarea suspectata de a fi afectata de erori grosiere este V_{\max} (V_{\min}) atunci se calculeaza:

$$v = \frac{V_{\max} - m_v}{\hat{\sigma}}; \quad \text{sau} \quad v = \frac{m_v - V_{\min}}{\hat{\sigma}}$$

P4. se compara v cu valoarea $v_{n,\alpha}$ din tabelul Grubbs-Smirnov, în care n este volumul selectiei de date, iar α reprezinta **riscul ales** (la masurarile uzuale riscul $\alpha = 0,05$ sau $\alpha = 0,01$ pentru *masurarile de foarte mare precizie* $\alpha = 0,005$ sau $\alpha = 0,001$)

daca $v \geq v_{n,\alpha}$ atunci V_{\max} (V_{\min}) este afectata de o eroare grosiera si se elimina din sirul de determinari, în caz contrar se mentine în sirul de rezultate;

P5. se reia testul cu noile date ramase de la pasul 2, pâna când $v < v_{n-i,\alpha}$.

2. Aplicarea testului 3 σ pentru un set de date

Exemplul 1. Se consideră un set de date care constituie notele obținute de studenti la un examen. Aceste date se gasesc in fisierul "Note_Stud.xls". Realizați **histograma** aferenta acestui set de date, utilizând instrucțiunea "histplot" și salvati-o în figura "Hist_note_init.scg". **Verificați existența erorilor grosiere** cu ajutorul **testului 3 σ** . Salvați datele afectate de erori grosiere într-un vector y . Realizați noua histograma după eliminarea erorilor grosiere.

Indicații:

```
%citire date
[date,text] = xlsread('Note_stud.xls')
%extragere date de temperatura
x = date';
%calcul medie aritmetica
med_aritm = mean(x)
%estimatie=deviatia standard
estimatie = std(x)
dimensiune = length(x)
%creare histograma
figure(1)
hist(x)
title('Histograma cu datele initiale')
%testul 3sigma
y = [];
xx = [];
for i = 1:dimensiune
    if x(i) > med_aritm + 3 * estimatie
        y = [y x(i)];
    elseif x(i) < med_aritm - 3 * estimatie
        y = [y x(i)];
    else xx = [xx x(i)];
    end
end
%calcul medie aritmetica noua
med_aritm_nou = mean(xx)
%estimatie=deviatia standard noua
estimatie_nou = std(xx)
dimensiunex = length(xx)
%creare histograma
figure(2)
hist(xx)
title('Histograma dupa testul 3*sigma')
```

3. Exerciții propuse

3.1. Se consideră setul de date "Date_temp.xls". Să se determine dacă datele sunt afectate de erori grosiere. Dacă astfel de date eronate există, aceste valori vor fi extrase într-un vector **z1**. Să se construiască o histogramă pentru datele initiale ("Histograma_31_initial" și să se ridice noua histogramă pentru datele corecte rămase cu numele "Histograma_31_date_corecte". Care este **valoarea medie a datelor corectate**? Dar **deviația standard** pentru acestea?

3.2. Prin măsurarea repetată a tensiunii unei surse s-au obținut valorile: 304,5; 305,2; 303,3; 304,9; 304,8; 305,0; 304,6; 305,1; 304,7; 304,9 [mV]. Să se verifice dacă valorile obținute prin măsurări sunt sau nu afectate de erori grosiere, cu ajutorul testului Grubbs Smirnov. Din tabel (Anexa 1), adoptând un risc $\alpha = 0,05$, inițial rezultă $v_{n,\alpha} = v_{10,0.05} = 2,176$ cu care se vor compara valorile extreme. Valorile afectate de erori grosiere vor fi extrase în vectorul **z2**, iar noile valori de selecție după excluderea datelor eronate se vor calcula: valoarea medie **mv2** și deviația **sig2**.

Anexa 1. Valorile parametrului $v_{n,\alpha}$ pentru testul Grubbs – Smirnov**Valorile parametrului $v_{n,\alpha}$ pentru aplicarea
testului Grubbs - Smirnov**

n	α			
	0,005	0,01	0,05	0,1
3	1,155	1,155	1,153	1,148
4	1,496	1,492	1,463	1,425
5	1,764	1,749	1,672	1,602
6	1,973	1,944	1,822	1,729
7	2,139	2,097	1,938	1,828
8	2,274	2,221	2,032	1,909
9	2,387	2,323	2,110	1,977
10	2,482	2,410	2,176	2,036
11	2,561	2,485	2,234	2,088
12	2,636	2,550	2,285	2,134
13	2,699	2,607	2,331	2,175
14	2,755	2,659	2,371	2,213
15	2,806	2,705	2,409	2,247
16	2,852	2,747	2,443	2,279
17	2,894	2,785	2,475	2,309
18	2,932	2,821	2,504	2,335
19	2,968	2,854	2,532	2,361
20	3,001	2,884	2,557	2,385
30	3,236	3,103	2,745	2,563
40	3,381	3,240	2,866	2,682
50	3,483	3,336	2,956	2,768
100	3,754	3,600	3,207	3,011