Identifying Data Affected by Aberrant Errors. Applied Program

Mihai-Radu COSTESCU
University of Craiova

Statistical survey has become a very powerful tool for understanding reality and interpreting it and prediction. Nevertheless, even with the accepted margin for errors, a survey's results may be inconcludent. This is mostly due to sample data quality. In this article, we refer to two tests to identify and eliminate aberrant errors, and at the end we present a program for applying these tests.
***Keywords**: Chauvenet, Young, program.*

The quality of research data can be influenced by two types of errors: systematical errors, with unilateral action, and random errors, with action both ways, due to a majority of factors whose individual influence is negligible.
Systematical errors as well as values affected by absurd errors must be discovered and eliminated, because the unfavorably influence the result of the investigation.
In case of discovering values affected by absurd errors, meaning data homogenization, the „standard" elimination possibilities of these values are numerous. We present the *Chauvenet test*, which, as opposed to other tests, (*Grubbs – Smirnov*, *Irwin*) does not assume certain parameters of the population which the sample comes from.
Discovering and eliminating systematical errors practically proves to be more difficult due to all the factors that condition themselves and that is why eliminating these errors has a very complex and varied character. We expose further on the *Young test*, which does not offer the possibility of eliminating systematical errors, but only appreciating the influence of systematical causes upon research data.

The Chauvenet Test
It is given a line of experimental values $x_1, x_2, ..., x_n$, it is considered that value x_i is being affected by aberrant errors if this condition is being verified (Chauvenet criterion):

$$\left|x_i - \bar{x}\right| > z \cdot \sigma$$

where $\bar{x}$ and σ represent the arithmetical average, respectively the standard deviation of the line of experimental values and the magnitude z is chosen from *table 1* according to the number n of values in the line.
From obvious reasons, it is enough that verifying the above relation to be made only for extreme values (minimal and maximal) within the sample.
The value of standard deviation of the line of experimental values is determined in this case with the expression:

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

Value z in table *1* can be determined with the help of the relation:

$$z = \frac{0{,}435 - 0{,}862 \cdot a}{1 - 3{,}604 \cdot a + 3{,}213 \cdot a^2}\ ,$$

where $a = \frac{2 \cdot n - 1}{4 \cdot n}$.

Table1 Values z for the Chauvenet Test

n	*z*	*n*	*z*	*n*	*z*
5	1,64	14	2,10	27-29	2,37
6	1,73	15	2,12	30-33	2,41
7	1,80	16	2,14	34-38	2,46
8	1,87	17	2,17	39-45	2,51
9	1,91	18	2,20	46-55	2,58

10	1,96	19	2,23	56-71	2,65
11	2,00	20-21	2,26	72-100	2,75
12	2,04	22-23	2,29	101-166	2,88
13	2,07	24-26	2,33	167-500	3,09

If, after applying the test, one of the tested values is affected by aberrant errors, that value is eliminated from the sample, we re-calculate the values of the average and standard deviation for the remained values and we start again with verifying the initial condition, the algorithm is applied until that condition is no longer verified for any of the two external values of the sample.

The Young Test

It is given the line of experimental values $x_1, x_2, ..., x_n$, we calculate the magnitude

$$\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(x_{i+1} - x_i\right)^2$$

and the magnitude $M = \frac{\delta^2}{\sigma^2}$.

Magnitude M is determined in this way with the values *CIV* (critical inferior value) and *CSV* (critical superior value), chosen from *table 2*, and it is considered that the line of experimental values has a random character, with α probability, if the following condition is fulfilled:

$$CIV < M < CSV$$

Table 2 Values CIV and CSV for the Young Test

n	*CIV*		*CSV*	
	$\alpha = 0{,}95$	$\alpha = 0{,}99$	$\alpha = 0{,}95$	$\alpha = 0{,}99$
4	0,78	0,53	3,22	3,47
5	0,82	0,54	3,18	3,46
6	0,89	0,56	3,11	3,44
7	0,94	0,61	3,06	3,39
8	0,98	0,66	3,02	3,34
9	1,02	0,71	2,98	3,29
10	1,06	0,75	2,94	3,25
11	1,10	0,79	2,90	3,21
12	1,13	0,83	2,87	3,17
15	1,21	0,92	2,79	3,08
20	1,30	1,04	2,70	2,96
25	1,37	1,13	2,63	2,87

It can be observed that the test can only be applied for samples containing at the most 25 experimental values. Parameter α from *table 2* has the meaning of a trustworthy coefficient, and can be chosen informatively, according to the sample's amount, in *table 3*. If the sample's amount is between two values in *table 3*, it is indicated to chose value α corresponding to a smaller sample's amount.

Table3

n	5	6	7	8	9	10	12	14
α	0,960	0,970	0,976	0,980	0,983	0,985	0,988	0,990
n	16	18	20	25	30	50	100	150
α	0,991	0,992	0,993	0,994	0,995	0,996	0,997	0,9973

Choosing the trustworthy coefficient in *table 3* can replaced with its determination with the help of the relation:

$$\alpha = \frac{1{,}5057 + 0{,}9968 \cdot n^{1{,}7404}}{2{,}1803 + n^{1{,}7404}}.$$

If the chosen or calculated value of the trustworthy coefficient is between the values available in *table 2*, it is indicated to choose the inferior available value.

Choosing values *CIV* and *CSV* in *table 2* can be replaced with their determination with the help of the relation:

$$CIV = \begin{cases} 0{,}491 + 0{,}081 \cdot n - 0{,}003 \cdot n^2 & pentru \quad \alpha = 0{,}95 \\ \dfrac{192{,}883 + 1{,}269 \cdot n^{2{,}336}}{411{,}427 + n^{2{,}336}} & pentru \quad \alpha = 0{,}99 \end{cases}$$

$$CSV = \begin{cases} 3{,}317 - 1{,}057 \cdot e^{-8{,}919 \cdot n^{-0{,}941}} & pentru \quad \alpha = 0{,}95 \\ 3{,}484 - 0{,}882 \cdot e^{-33{,}574 \cdot n^{-1{,}399}} & pentru \quad \alpha = 0{,}99 \end{cases}$$

The program, made in Borland Pascal language, is presented hereby:

```
program eliminare_valori_chauvenet_young;
{Author: Mihai Radu Costescu
     The program checks:
- the presence of aberrant errors on Chauvenet test basis
- the presence of aberrant errors on Young test basis
 Parameters:
             n = run of values dimension
             x = values' vector
             z = confidence level (z=0.95 or z=0.99)}
type vector=array[1..500] of real;
var n,i,flag:integer;
    x:vector;
    a,z,vci,vcs,alfa,media,sigma,delta,xmax,xmin:real;
function med(n:integer;x:vector):real;
    var sum:real;
    begin
         sum:=0;
         for i:=1 to n do
             sum:=sum+x[i];
         med:=sum/n;
end;
function sig(n:integer;media:real;x:vector):real;
    var sp:real;
    begin
         sp:=0;
         for i:=1 to n do
             sp:=sp+sqr(x[i]-media);
         sig:=sqrt(sp/(n-1));
end;
function deltap(n:integer;x:vector):real;
    var delt:real;
    begin
         delt:=0;
         for i:=1 to n-1 do
             delt:=delt+sqr(x[i+1]-x[i]);
         deltap:=delt/(n-1);
end;
procedure max_min(n:integer;x:vector;var xmax,xmin:real);
          begin
               xmax:=x[1];
               xmin:=x[1];
               for i:=2 to n do
                   begin
                        if x[i]<xmin then xmin:=x[i]
                        else if x[i]>xmax then xmax:=x[i];
                   end;
          end;
```

```
begin {program principal}
     repeat
           writeln('Introduce the run of values dimension from 5 to 500');
           write('n='); readln(n);
     until (n>=5) and (n<=500);
     repeat
           writeln(' introduce confidence level alfa=0.95 or alfa=0.99');
           write('alfa='); readln(alfa);
     until (alfa=0.95) or (alfa=0.99);
     for i:=1 to n do
         begin
              write('x(',i,') = ');
              readln(x[i]);
         end;
     flag:=0;
     a:=(2*n-1)/(4*n);
     z:=(0.435-0.862*a)/(1-3.604*a+3.213*a*a);
     if alfa=0.95 then
                           begin
                                  vci:=0.491+0.081*n-0.003*n*n;
                                  vcs:=3.317-1.057*exp(-8.919*exp(-0.941*ln(n)))
                           end
                   else
                           begin

       vci:=(192.883+1.269*exp(2.336*ln(n)))/(411.427+exp(2.336*ln(n)));
                                  vcs:=3.484-0.882*exp(-33.574*exp(-1.399*ln(n)))
                           end;
     media:=med(n,x);
     sigma:=sig(n,media,x);
     delta:=deltap(n,x);
     max_min(n,x,xmax,xmin);
     if abs(xmin-media)>z*sigma
        then begin
                  flag:=1;
                  writeln('Eliminated value:',xmin:10:3);
             end;
     if abs(xmax-media)>z*sigma
        then begin
                  flag:=1;
                  writeln('Eliminated value:',xmin:10:3)
               end;
     if flag=0 then writeln(' There were no aberrant values ');
     if (delta/sqr(sigma)>vci) and (delta/sqr(sigma)<vcs)
                           then writeln(' There were no systematic causes ')
                           else writeln(' There were systematic causes ');
end.
```

Bibliography

1. M.R.Costescu, N.Vasilescu, C.Ionaşcu, *Statistics and elements of research theory (revised and improved edition)* – Universitaria Publishing House, Craiova, 2001

2. N.Vasilescu, M.R.Costescu, C.Ionaşcu, G.Babucea, V.Tomiţă, D.Stuparu, *Statistics* – Universitaria Publishing House, Craiova, 2003

3. M.R.Costescu, A.Ionescu, *Informational processing of measurement data* – Universitaria Publishing House, Craiova, 2004

4. M.R.Costescu, *Statistical methods applied in social sciences* – Publishing House „Libertatea”, Panciova – Serbia