International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 3, July 2014, PP 15-20
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

An Intelligent System for Eliminating the Suspicious Experiment Data

Jiang Xingfang[1], Chen Dongdong[2], Fu Haiming[2], Huang Shuang[2], Liu Hongfei[3]

[1]Jiangsu Key Laboratory for Solar Cell materials and Technology, Changzhou University, Changzhou 213164, China
[2]School of Mathematics and Physics, Changzhou University, Changzhou 213164, China
[3]School of Huaide, Changzhou University, Changzhou 213016, China

Abstract: *There is a distinct suspicious experiment datum (outlier) in a group data and they could be eliminated by the method of three times standard deviation. For the problem the method of two times standard deviation has been proposed and the intelligent system for eliminating the outlier had been developed. The intelligent system has four stages. The first one is that the general calculation software has been developed for calculating the mean and the uncertainity in the ten data. The second one is that the calculation software is for filling data randomly by user. The thrid one is suit for filling data randomly by user and there are outliers in the experiment data. The fourth one is that the intelligent system is suit for filling randomly by user, there are outliers, and the minimum digit of decimal places being different in experimental data. The statistics method of " add 1 for not empty" for filling randomly by user has been adopted. The confidence possibility of 95.5% that it is the judgment method of two times standard deviation range near the mean has been proposed. The method of "reading characters" for calculating the minimum digit of decimal places has been applied. The intelligent system is good for all the data processing in Physics experiment and it could be spreaded to other subjects for data processing.*

Keywords: *Intelligent system; Outlier; Add 1 for not empty; Method of reading characters; Multimedia ToolBook*

1. INTRODUCTION

In the process of experimental measurement, the outlier is inevitably. There are many factors such as the environmental conditions change suddenly, the equipment failure suddenly, and the unexpected reasons to read wrong number, write wrong number, mistype a fault number. The outliers must be eliminated or it will seriously affect the result of the experiment. Therefore it is important in the treatment of experimental data and the work of eliminating outlier need to be careful. The methods for eliminating outlier are Paüta criterion [1-4], Grubbs criterion [5-6], Dixon criterion [7-9], Cauvenet criterion and so on [10]. For an array of 0.896, 0.891, 0.893, 0.893, 0.894, 0.892, 0.895, 0.899, 0.897, 0.891, the mean that is the best estimate [1] is 0.8941 and the uncertainity is three times standard deviation of the mean. The experiment result is 0.894±0.003 according to the rules that the uncertainty is a digit and the digit is alignment with the last digit of the signification figure, the uncertainty is only carried on and the signification figure is carried on when it is lagre than 0.5. When the fifth datum is outlier of 0.794, the mean of the data is 0.8841 and three times standard deviation is 0.090414. The difference between the outlier and the mean is |0.794-0.8841|=0.901 that is less than three times standard deviation. When the fifth datum is outlier of 1.894, the difference between the outlier and the mean is less than three times standard deviation as well as. The outlier does not eliminated by Paüta criterion like 0.794.

2. The Effective Eliminating Method for the Outlier

There are many measurement values for the measured physical quantities under the certain condition and the data change randomly with unpredictable way. The experimental data obey the normal distribution. According to mathematical statistics theory for the measured value *x*, the density function is

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},\quad \sigma=\lim_{n\to\infty}\sqrt{\frac{\sum_{i=1}^{n}(x_i-a)^2}{n}} \tag{1}$$

Here σ is standard deviation and the true value is a. The possibility that the measured value x_i is in the range of one time standard deviation near the mean is 68.3%, the possibility that the measured value x_i is in the range of two times standard deviation near the mean is 95.5% and the possibility that the measured value x_i is in the range of three times standard deviation near the mean is 99.7% . The function of the normal distribution is shown in Fig. 1.

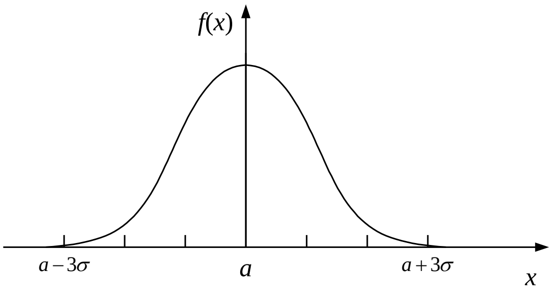


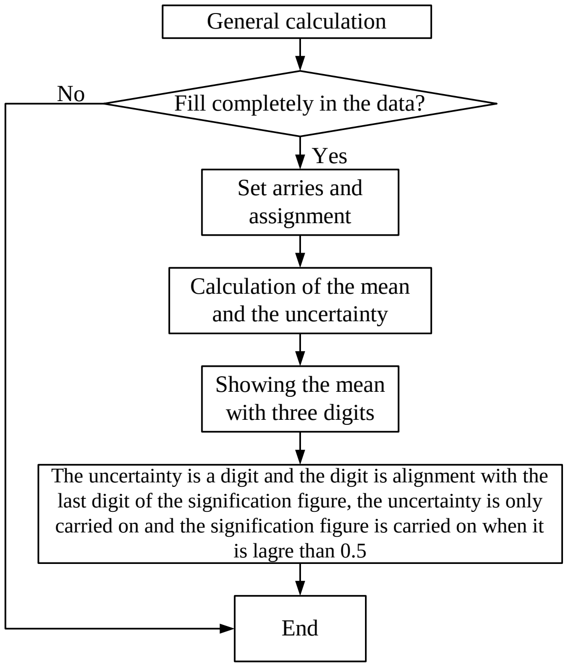
Fig1. *The function of the normal distribution*

Because of the Paüta criterion that is the method of three times standard deviation is too conservative. The outlier like 0.794 could not be eliminated. According to the Reference [2-9] about Grubbs criterion, Dixon criterion, and Cauvenet criterion, the general confidence probability is took 95%, the method of two times the standard deviation near mean is very effective for eliminating the outlier of 0.794 and the confidence probability is 95.5% [9].

3. Making of Intelligent System

3.1. General Calculation

In the first stage the software is for calculation the mean and the uncertainty such as 0.896, 0.891, 0.893, 0.893, 0.894, 0.892, 0.995, 0.999, 0.997, 0.991. The data are from the diameter of the holes on solar cell that is drilled by laser with metal wrap through (MWT) technology. The advantages are precise and fast to calculate the mean and the uncertainty for ten data and each datum has three digits. The result is strictly according to the rules that the uncertainty is a digit and the digit is alignment with the last digit of the signification figure, the uncertainty is only carried on and the signification figure is carried on when it is lagre than 0.5. The Flow-chart for the button of "Calcultion" is shown in figure 2. The disadvantages are three points. The first point is that it could not calculate when the number of the data is less than ten. The second point is that the outliers do not eliminated so the uncertainty is too big. The thrid point is that the calcultion system is not suit for the digits being less than three.


General calculation
Fill completely in the data?
No
Yes
Set arries and assignment
Calculation of the mean and the uncertainty
Showing the mean with three digits
The uncertainty is a digit and the digit is alignment with the last digit of the signification figure, the uncertainty is only carried on and the signification figure is carried on when it is lagre than 0.5
End

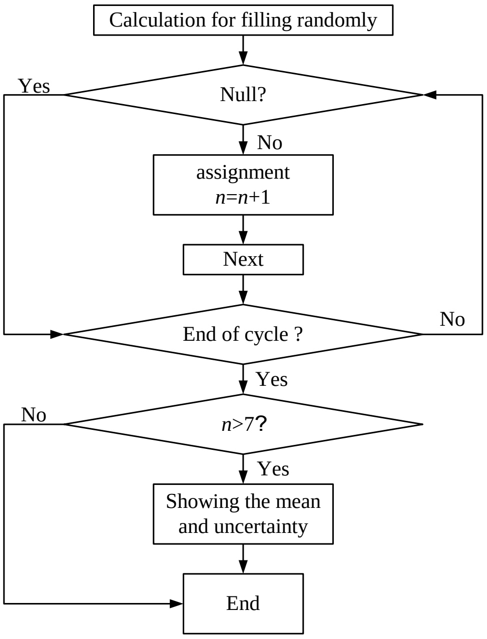

Calculation for filling randomly
Null?
Yes
No
assignment
n=n+1
Next
End of cycle ?
No
Yes
n>7?
No
Yes
Showing the mean and uncertainty
End

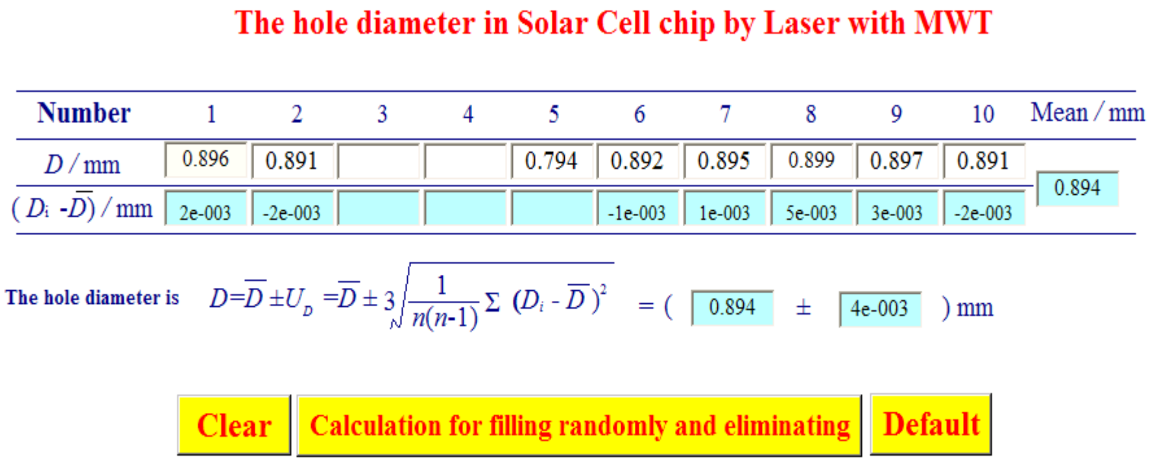
Fig2. *The flow-chart of "General calculation"* **Fig3.** *The flow-chart of "Calculation for filling randomly"*

3.2. Calculation for filling randomly

Because the user may not have in order to fill in the data, the general calcultion system could be improved for allowing the user to fill in randomly as long as the number of data not less than 7. The second stage is that the technology of "Add 1 for not empty" was used for statistics the number of the data. If the number that input by the user is less than 7, the system appears the prompt of "Fill the experimental data, please!". The flow-chart of the button of "Calculation for filling randomly" is shown in figure 3.

3.3. Calculation for filling randomly and eliminating suspicious data

Although it uses "Add 1 for not empty" in section 3.2 the case that the uncertainty is too big because there are outliers. The third stage that the software system is suit for eliminating the outliers with the method of two times standard deviation.


The hole diameter in Solar Cell chip by Laser with MWT

Number	1	2	3	4	5	6	7	8	9	10	Mean / mm
D / mm	0.896	0.891			0.794	0.892	0.895	0.899	0.897	0.891	0.894
$(D_i - \bar{D})$ / mm	2e-003	-2e-003				-1e-003	1e-003	5e-003	3e-003	-2e-003	

The hole diameter is $D=\bar{D}\pm U_D=\bar{D}\pm 3\sqrt{\frac{1}{n(n-1)}\Sigma\,(D_i-\bar{D})^2}$ = (0.894 ± 4e-003) mm

Fig4. *The interface of "Calculation for filling randomly and eliminating suspicious data"*

The effective number of the data is large than 7 or the prompt of "Fill the experimental data, please!" would be occured. The calculated result is shown in Fig. 4 and the flow-chart of "Calculation for filling randomly and eliminating suspicious data" is shown in Fig. 5.

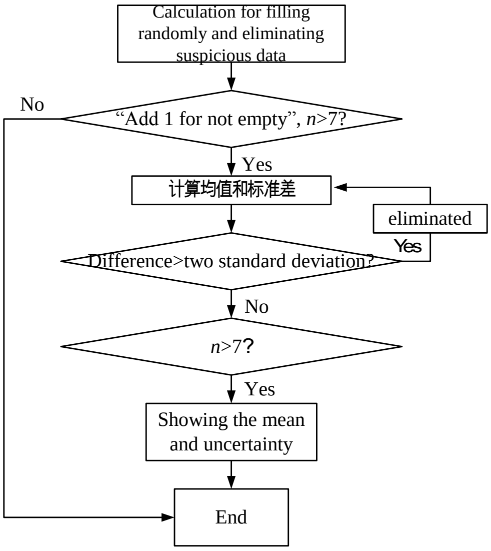

Calculation for filling randomly and eliminating suspicious data
No
"Add 1 for not empty", n>7?
Yes
计算均值和标准差
eliminated
Yes
Difference>two standard deviation?
No
n>7?
Yes
Showing the mean and uncertainty
End

Fig5. *The flow-chart for "Calculation for filling randomly and eliminating suspicious data"*

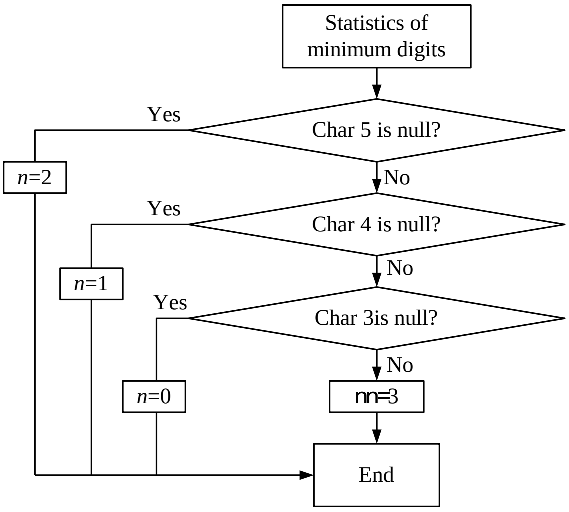


Fig 6. *The flow-chart for statistics of minimum digits*

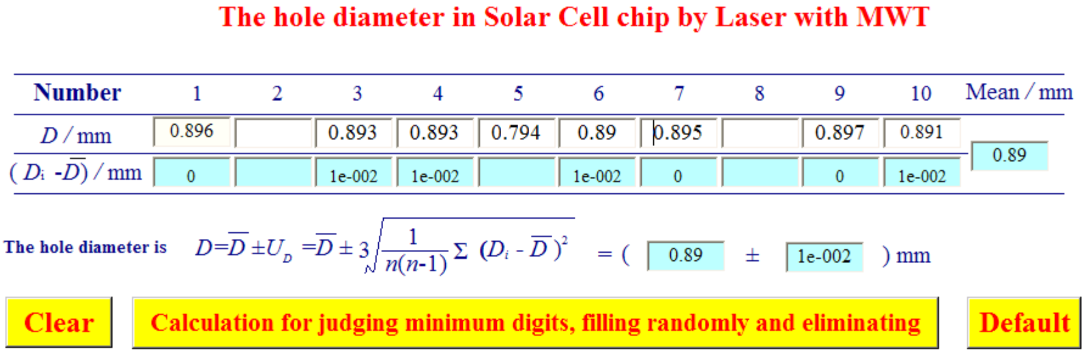
3.4. Calculation for judging the minimum digits, filling randomly and eliminating suspicious data

Sometimes the user fills the data are not three digits. The software made in section 3.3 could not show correctly the mean and the uncertainty. The intelligent system must finish the statistics of the minimum digits. The key is "reading characters" for calculating the minimum digit of decimal places. The program is as follow.

```
step i from 1 to 10
    if char 3 of text of field ("field"&(110+i)) = null
        n=0
    else
        if char 4 of text of field ("field"&(110+i)) = null
            n=1
        else
            if char 5 of text of field ("field"&(110+i)) = null
                n=2
```

```
          else
             n=3
          end
       end
    end
 end
```

The flow-chart for the button of “Calculation for judging the minimum digits, filling randomly and eliminating suspicious data” is shown in Fig. 6. The calculation result is shown in Fig. 7. In Fig. 7 the second datum and the eighth datum are not filled, the fifth datum is an outlier, and the number of digit in sixth datum is 2 while in other’ number are 3. The result is 2 digits for the mean.



The hole diameter in Solar Cell chip by Laser with MWT

Number	1	2	3	4	5	6	7	8	9	10	Mean / mm
D / mm	0.896		0.893	0.893	0.794	0.89	0.895		0.897	0.891	0.89
$(D_i - \overline{D})$ / mm	0		1e-002	1e-002		1e-002	0		0	1e-002	

The hole diameter is $D=\overline{D}\pm U_D=\overline{D}\pm 3\sqrt{\frac{1}{n(n-1)}\Sigma\,(D_i-\overline{D})^2}$ = (0.89 ± 1e-002) mm

Clear
Calculation for judging minimum digits, filling randomly and eliminating
Default

Fig7. *The interface of “Calculation for judging the minimum digits, filling randomly and eliminating suspicious data”*

4. Conclusion

To develop an intelligent system is system engineering. For eliminating the outlier like 0.794, the developed intelligent system has stages. The first stage is for general calculation according to the uncertainty is a digit and the digit is alignment with the last digit of the signification figure. The second stage is that the developed system is suit for filling randomly. The third stage is suit for filling randomly and eliminating suspicious data. The intelligent system is suit for judging the minimum digits, filling randomly and eliminating suspicious data. The technologies of “Add 1 for not empty” , the method of two times standard deviation that the confidence probability is 95.5%, and the “ reading characters” for calculating the minimum digit of decimal places have been applied. For the case of more complex, the more technologies would be proposed.

Acknowledgment

This work is supported by Open Issues of Jiangsu Key Laboratory for Solar Cell materials and Technology (No. 201207) and State Key laboratory of Satellite Ocean Environment Dynamics (No. SOED1201).

References

[1] Jiang Xingfang, Xie Jiansheng, Tang Li. Physics Experiment (second vision). Beijing: Science Press. 2013

[2] Zhang Min, Yuan Hui. The Paüta criterion and rejecting the abnormal value [J]. Journal of Zhengzhou University of technology. 1997, 18(1): 84-88

[3] Wim Coucke, Bernard China, Isabelle Delattre, et al. Comparison of different approaches to evaluate External Quality Assessment Data Original Research Article [J]. Clinica Chimica Acta. 2012, 413(5-6): 582-586

[4] Chen Rui, Zhou Shumin. The application of improved Grubbs' Criterion for Inspecting the Count of Radon concentration [J]. Nuclear Electronics & Detection Technology. 2009, 29(1): 113-115

[5] Chen Jianhua. The suspicious data processing with mathematical statistics method in detection. China Applicance Technology. 2008, 1: 61-63

[6] Ram B. Jain. A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data [J]. Clinical Biochemistry. 2010, 43: 1030-1033

[7] Li Ke, Ye Yingzhi. Amelioration of Outlier Detection in Linear Regression and Its Application [J]. Chinese Journal of Scientific Instrument. 2004, 25(4): 723-725

[8] Origin Used in Comparison the methods of eliminating the excrescent data [J]. Experiment Science and Technology. 2012, 10(1): 74-76+118

[9] Ye Chuan, Wu Chuanhui, Zhang Jiayi. Comparison about how to get rid of abnormal data in metrology & measurement. Metrology & Measurement Technique. 2007, 34(7): 26-28

[10] Wang Wenzhou. t Test-the superlative test to discard abnormal values with σ unknown [J]. Journal of Sichuan University of Science and Technology. 2000, 19(3): 84-86