International Journal of Materials Science and Applications
2017; 6(6): 297-301
http://www.sciencepublishinggroup.com/j/ijmsa
doi: 10.11648/j.ijmsa.20170606.15
ISSN: 2327-2635 (Print); ISSN: 2327-2643 (Online)

Application of Mathematical Statistics Analysis Algorithm for Chemical Data

Shen Nana*, Lu Xinjian

Nanjing Chem Cyber Technology Company Ltd, Nanjing, China

Email address:

Nana.shen@chemcyber.com (Shen Nana), Xinjian.lu@chemcyber.com (Lu Xinjian)

*Corresponding author

To cite this article:

Shen Nana, Lu Xinjian. Application of Mathematical Statistics Analysis Algorithm for Chemical Data. *International Journal of Materials Science and Applications*. Vol. 6, No. 6, 2017, pp. 297-301. doi: 10.11648/j.ijmsa.20170606.15

Received: August 20, 2017; **Accepted**: August 31, 2017; **Published**: December 6, 2017

Abstract: In this paper, chemical process data is analyzed by variance, the least square algorithm and then comparing the original data and processed data in excel. Through the comparing result, processed data is easier for operators to observe and find out rules and hidden problems in chemical conditions. According to the two algorithms, experienced operators can adjust chemical conditions to be normal. So they are better ways to optimize chemical conditions, as a result, the data analysis algorithm make a contribution to chemical industry.

Keywords: Data Analysis, Excel, Least Squares Method, Chemical Conditions

1. Introduction

The chemical production process is scientific, continuous and stable, and it is also very complex which includes chemical material flowing, chemical material response, super high temperature and high pressure conditions [1]. And the majority of flowing vast are at high risk of flammable and explosive, poisonous and harmful. So it is essential for chemical production to discover the abnormal data to solve the problems in the chemical production process.

It is important to strengthen the construction of chemical enterprise information construction, enterprises should pay more attention to the development data analysis tools, and make full use of data analysis [2]. Now big data era is coming, it will became the new way to improve chemical enterprise benefit. Establishing the management mechanism of "use Numbers to talk, use the data to make decisions, use data to manage and use the data to innovation" will be the development trend of new era. Now relatively large chemical enterprise are inseparable from the chemical data analysis. Using the method of mathematical statistics to the analysis data in the process of chemical production, could find the abnormal data timely.

In this paper, using mathematical statistics method such as least square method and variance method to deal with the chemical data, and then making a comparison in excel [3]. By comparison, found that characteristics of processed data is more representative than the original data, it is easy to observe and handle, as a result, workers can found the problem timely and improve chemical operation process.

In this paper, the experimental data is from heavy oil olefin device in catalytic workshop of one chemical enterprise. Using variance, the least square algorithm to analysis this device data, the number of data in table 1 is 972.

Table 1. *Resources of data.*

No.	Data
1	622.3106689
2	622.6915283
3	623.4661255
4	623.0188599
5	621.3545532
6	619.2012329
7	617.9406738
8	617.9406738
9	617.9406738
10	617.9406738
……	……
972	617.9012451

2. Algorithm of Chemical Data Abnormal Point

2.1. Variance

Variance is proposed by Ronald Fisher in his essay "The Correlation Between Relatives on The Supposition of Mendelian Inheritance". In the statistics description, variance is used to calculate the difference between each variable (observations) and the overall mean.

Variance is a measurement of random variables or a set of discrete degree of measurement data in probability theory. Variance is used to measure the deviation between the expected (average) and random variables. In many practical problems, the variance deviation degree has important significance. Variance is to measure the difference between source data and overall mean. Variance calculation formula:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n} \quad (1)$$

σ^2 is variance, x is variable, μ is average, n is number of data [4].

Using the method of variance, it is easier to find out the data which has the larger changes in chemical production process. For example, variation range of a variable is small, when its value became to be a constant and it does have no longer changes, it is difficult for the operator to find the unusual phenomenon by observing meter. But using variance can clearly see the variance of the original data making a difference, from zero to nonzero.

2.2. Comparison in Excel

EXCEL is one of the suites of office software. It not only has function to create, edit and print form, but also more outstanding, it is in the form data to calculate, sort and classify data, and generate charts [5]. It is convenient to process a large number of daily statistics and various reports, practice has proved that EXCEL playing a big role in data analusis.

2.2.1. VARA Function

VARA function which is inline functions of excel and its principle is variance algorithm. So this paper use VARA function to calculate the data in table 1.

This paragraph introduces the syntax and usage of VARA function in Microsoft Excel. VARA function calculates the variance of the given sample, using method as follows [6]:

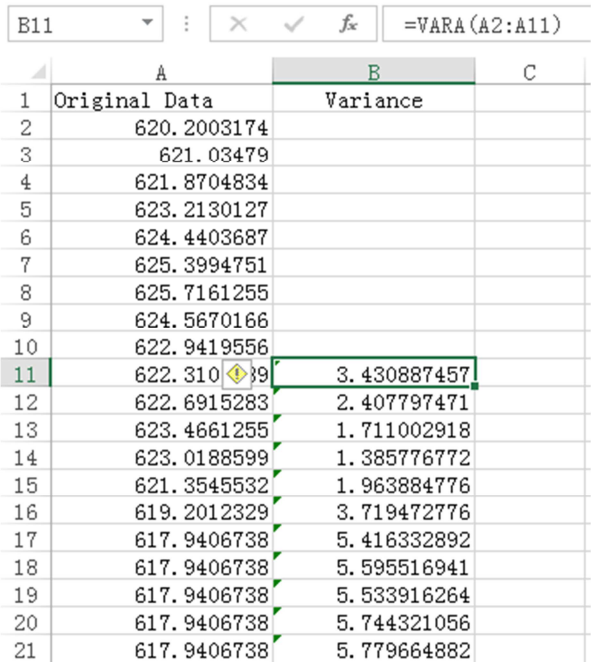
VARA (value1, [value2],...)

VARA function has the following parameters:

1. Value1, value2,... Value1 is required and subsequent values are optional.
2. VARA assumes that its parameters are the overall sample. If the data represents the sample population, must using the function VARPA to calculate the variance.
3. parameters can be as the following form: numerical value; Contains the name, array, or a logical value in a reference, such as TRUE and FALSE.
4. If the parameter that contains TRUE is calculated as 1; if the parameters that contain text or FALSE are evaluated as 0 (zero).
5. if the parameter is an array or reference, only ues the value of parameter. The blank cells and text values in an array or reference can be ignored.
6. if the parameter is an error value or text cannot be converted to numeric, it will result in an error.
7. use the VAR function if you want to make the calculation do not include the logical value in the reference and the text that represents the number.

2.2.2. Result of Comparison

Using the data in table 1 to calculate variance of original data. In this paper, all the calculation and comparison are completed in excel. Using built-in function of excel: VARA function to calculate variance: each 10 original data is a group, calculate the last data's variance and so on. Removing the before nine, there are 963 variances. Detailed data information as is shown in figure 1.



	A	B	C
1	Original Data	Variance	
2	620.2003174		
3	621.03479		
4	621.8704834		
5	623.2130127		
6	624.4403687		
7	625.3994751		
8	625.7161255		
9	624.5670166		
10	622.9419556		
11	622.310 39	3.430887457	
12	622.6915283	2.407797471	
13	623.4661255	1.711002918	
14	623.0188599	1.385776772	
15	621.3545532	1.963884776	
16	619.2012329	3.719472776	
17	617.9406738	5.416332892	
18	617.9406738	5.595516941	
19	617.9406738	5.533916264	
20	617.9406738	5.744321056	
21	617.9406738	5.779664882	

Figure 1. *Process information of Variance calculation.*

After calculation is completed, data curves of the original data and original data variance are generated in excel, which is shown in figure 2, It can be saw that the variance changes are bigger than the original changes, it is convenient for operator to observe abnormal points and summarize the rules for chemical process, it is also better to optimize the operation process.

So variance algorithm can be used in chemical process. There are a lot of small changes data in this algorithm, through variance algorithm, operator can get the abnormal situation obviously and deal with processing problems in time.

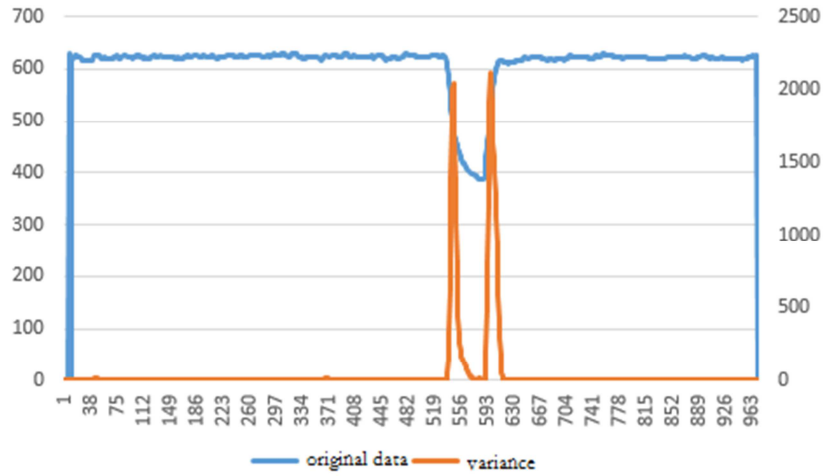


Figure 2. *Comparison between variance and the original data.*

3. Algorithm of Chemical Data Rate

The changes of date rate is earlier happened than the change of the data itself [7], so it is a better way through the rate of change to predict the data itself changes, and the way can provide operation prediction basis for chemical process. The way could be expressed in excel, using least squares function which is the built-in function of excel to calculate.

3.1. Last Square

Last Square Method is put forward by Mali. Legendre in 1806. It through the minimum sum of squares of error to find the best matching function of data. It is easier to obtained unknown data by least square method. meanwhile the sum of the squares of the error between actual data and calculated data is smallest, the least square method can also be used for curve fitting.

Supposing there is a series of data points, $(x_i, y_i)(\mathrm{i} = 1, \dots, \mathrm{m})$, $\mathrm{h}(x_i)$ is estimator of the fitting function $\mathrm{h(x)}$, residual error is $r_i = h(x_i) - y_i$, which would evaluate the fitting degree of fitting function and solving function. And then introducing three kinds of norm:

(1) ∞- norm: a maximum of absolute value of residual error $\max_{1 \le i \le m} |r_i|$,

(2) 1- norm: a absolute value of residuals $\sum_{i=1}^{m} |r_i|$,

(3) 2- norm: a sum of squared residuals $\sum_{i=1}^{m} r_i^2$,

The first two norms are easy to obtain, but under the condition of large amount of data, they need a lot of calculation and are not operable. Therefore 2 – norm is general used. Fitting degree is the similarity between fitting function h (x) and solving function y, the smaller 2 –norm is, the higher similarity is.

Thus, the definition of least square method is:

For a given data $(x_i, y_i)(\mathrm{i} = 1, \dots, \mathrm{m})$, in the assumption space h, to solve the $\mathrm{h(x)} \in \mathrm{h}$, making 2 – norm of residuals $r_i = h(x_i) - y_i$ is minimum, namely:

$$\mathrm{h(x)}\ s.t.\ min \sum_{i=1}^{m} (h(x_i - y_i))^2 \qquad (2)$$

In the aspect of geometric, formula (2) is to search $\mathrm{y = h(x)}$ whose distance from given points $(x_i, y_i)(\mathrm{i} = 1, \dots, \mathrm{m})$ is minimum. $\mathrm{h(x)}$ is called the fitting function or the least-square solution, the method of solving fitting function h(x) is the least squares.

Assuming expression of $\mathrm{h(x)}$ is:

$$\mathrm{h(x, w)} = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n \qquad (3)$$

$\mathrm{h(x, w)}$ is a polynomial, w is its parameters. Least squares method is to find such a set $\mathrm{w} = (w_0, w_1, \dots, w_n)$ to make $\sum_{i=1}^{m} (h(x_i) - y_i)^2$ is minimum. Using the calculus principle to calculate the value of $\mathrm{w} = (w_0, w_1, \dots, w_n)$.

3.2. Comparison in Excel

3.2.1. Linest Function

LINEST function which is inline functions of excel and its principle is variance algorithm. So this paper use LINEST function to calculate the data in table 1.

This paragraph introduces the syntax and usage of LINEST function in Microsoft Excel.

The LINEST function [8] computes the statistical value of a line by using the least square method to calculate the line with the best fit of the existing data, then returns an array that describes the line.

The formula of the line is:

$$\mathrm{y = mx + b} \qquad (4)$$

or

$$\mathrm{y} = m_1 \times 1 + m_2 \times 2 + \cdots . b \qquad (5)$$

If x values in multiple regions, the dependent variable y is a function of independent variable x. m is corresponding coefficient to each x, and b is constant. Otherwise, y, x and m can be vectors. The LINEST function returns an array of $\{\mathrm{b}m_n, m_{n-1}, \dots m_1, b\}$. The LINEST function can also return the additional regression statistics.

LINEST (known_y 's, [known_x' s], [const], [stats])

The LINEST function has the following parameters:

1. Known_y 's is required
2. If known_y's corresponds to a cell region in a single column, each column of known_x's is treated as an independent variable.
3. If known_y's corresponds to a cell region in a single row, each row of known_x's is treated as an independent variable.
4. Known_x 's is optional
5. The cell area corresponding to known_x's can contain one or more sets of variables. If only one variable is used, then known_x 's and known_y's have the same dimension, and they can be any shape. If multiple variables are used, known_y's must be a vector (that is, a row or column).
6. If known_x's is omitted, assume that the array is {1, 2, 3,... whose size is the same as known_y's.
7. const is optional. A logical value that specifies whether the constant b is set to 0.
8. If const is TRUE or omitted, b will be computed normally.
9. If const is FALSE, b will be set to 0, and the m value will be adjusted to make y = mx.
10. stats is optional. A logical value that specifies whether additional regression statistics are returned.
11. If the stats is TRUE, the LINEST function returns the attached regression statistics, and the returned array is $\{m_n, m_{n-1}, \ldots m_1, b$; Sen sen – 1,…, se1, seb; r2, sey; F, df; ssreg ssresid$\}$.
12. If stats is FALSE or omitted, the function LINEST returns only the coefficients m and constant b.

3.2.2. Result of Comparison

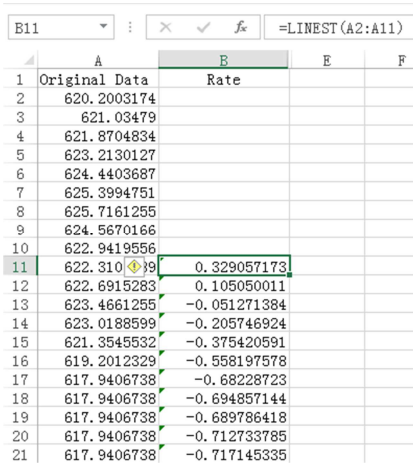


Figure 3. *Process information of rate calculation.*

Using excel built-in function: LINEST function to find out original data rate. And the calculation process is similar with the variance process, each 10 data is a group, calculating the last data's variance and so on. Detailed data information as is shown in figure 3. Data curves of the original data and original data rate are generated in excel, which is shown in figure 4. From the picture. It can be saw that rate changes are happened in the 556th point and original data changes are happened in the 630th point. Namely, rate changes are earlier happened than original data changes. Through rate changes, operator can predict the changes of original data and according to the changes of original data, chemical conditions can be adjusted and optimized.

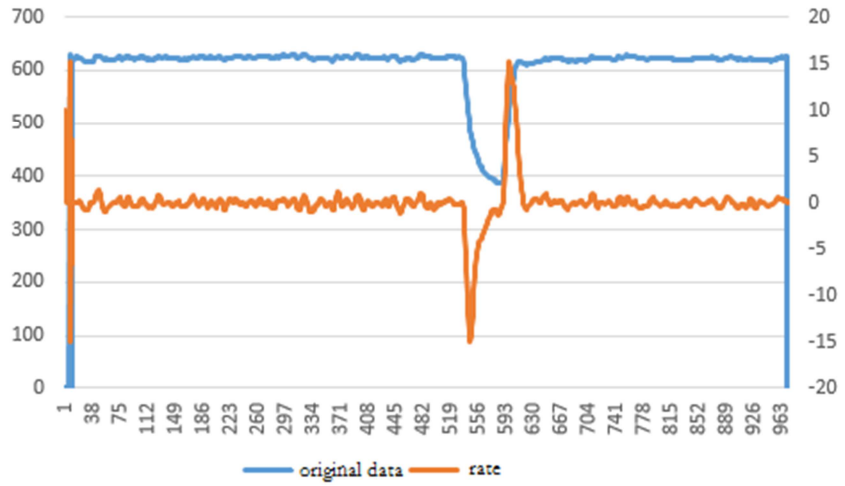


Figure 4. *Comparison between rate and the original data.*

3.3. Algorithm of Rate Range

Assuming that a certain variable's normal scope is 0-100 in a stable chemical process, its value of upper limit warning is 80, its value of lower limit warning is 20, when its real time value is not between 20 and 80, operators will be reminded that the chemical process is in risk. when its value occurring from 25 to 70, because all the value are between the value of its lower and upper limit warning [9], it will not generate alarm, but according to working experience, big changes show that the chemical process would be in risk, this abnormal phenomenon is easy to be ignored. To capture this abnormal keenly, it is necessary to find out the rate range, if some point's rate scope beyond its normal range, and then generating alarm. According to working experience, rate range calculation formula is [10]:

$$y_1 = \mu - n * \sigma \tag{6}$$

$$y_2 = \mu + n * \sigma \quad (7)$$

y_1 is rate lower limit of rate and y_2 is upper limit of rate, u is the average of rate, σ is the standard deviation of rate, n is coefficient, the smaller n is, the better rate of scope is, but it is not too small. Through trial the value of n is 3.

(1) Average of rate: u

Using excel built-in function: AVERAGE function to calculate the average of rate, first insert AVERAGE function in excel and select the column B:Rate data and get the result of u which is in cell C11 as is shown in figure 5.

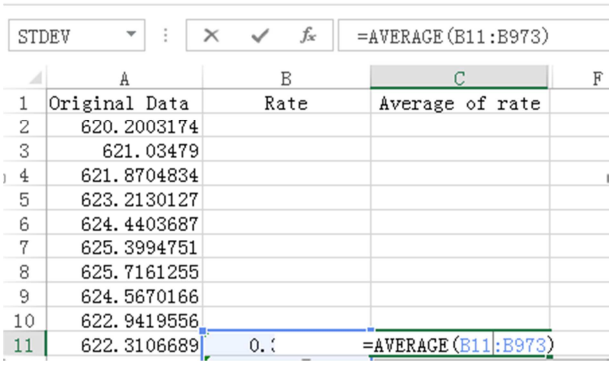


Figure 5. *Process information of calculation of average of rate.*

(2) Standard deviation of rate: σ

Using excel built-in function: STDEV function to calculate standard deviation of rate, the specific operation is similar to B, insert STDEV function in excel and select the column B:Rate data and get the result of σ which is in cell D11 as being shown in figure 6.

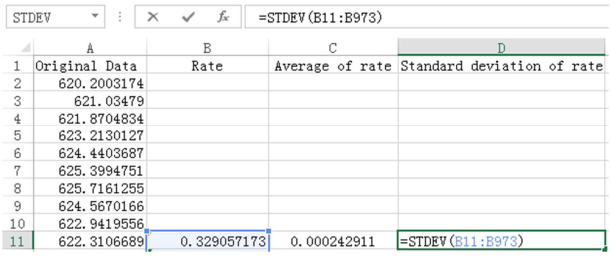


Figure 6. *Process information of calculation of standard deviation of rate.*

And then according to the formula (4), formula (5) and the parameters σ and u, the rate of change can be calculated. Through the rate range, it is easier to measure rate change is normal. And operator can according to rate change to predict original date changes and then to adjust chemical conditions. As a result, the rate range play an important role in algorithm of chemical data rate.

4. Conclusion

For chemical process data, data analysis algorithm of variance method and least squares method are proposed in this paper. Through those analysis algorithm, it is convenient for operator to find out abnormal conditions, which are hidden in chemical process and would not be found in traditional method.

Through variance algorithm, abnormal and small changes in chemical conditions can be found timely and then if those changes be found, operator would take action to prevent changes from spreading.

Algorithm of chemical data rate play an important role in chemical conditions. It has prediction function, it can predict whether the original data will change by the change of rate. For experienced operator, it's a better way to know how to optimized chemical conditions through original data changes

Those analysis algorithm have been used in UT-CLOUD which is a cloud analysis platform of Nan Jing Chem Cyber Technology Company Ltd. UT-CLOUD applies those analysis algorithms to make on-line analysis for chemical data, and it provide customers with accurate information of real-time working condition. So those data analysis algorithm do made a contribution to optimize the chemical operation.

References

[1] Couper, James R., W. Roy Penney, and James R. Fair. *Chemical Process Equipment-Selection and Design (Revised 2nd Edition)*. Gulf Professional Publishing, 2009.

[2] Macfarlane, Robert, et al. *The NJOY Nuclear Data Processing System, Version 2016*. No. LA-UR-17-20093. Los Alamos National Laboratory (LANL), 2017.

[3] Weaver, Kathleen F., et al. "Basics in Excel." An Introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences, First (2018), pp. 523-537.

[4] Parsons, Luke A., et al. "Temperature and precipitation variance in CMIP5 simulations and paleoclimate records of the last millennium." *Journal of Climate* 2017 (2017).

[5] Pickles, Anthony J. "To Excel at bridewealth, or ceremonies of Office." *Anthropology Today* 33.1 (2017), pp. 19-22.

[6] Jacobs, Perke, and Wolfgang Viechtbauer. "Estimation of the biserial correlation and its sampling variance for use in meta‐analysis." *Research synthesis methods* 8.2 (2017), pp. 161-180.

[7] Duník, Jindřich, Ondřej Straka, and Miroslav Šimandl. "On autocovariance least-squares method for noise covariance matrices estimation." *IEEE Transactions on Automatic Control* 62.2 (2017), pp. 967-972.

[8] Laboure, Vincent M., Ryan G. McClarren, and Yaqi Wang. "Globally Conservative, Hybrid Self-Adjoint Angular Flux and Least-Squares Method Compatible with Voids." *Nuclear Science and Engineering* 185.2 (2017), pp. 294-306.

[9] Benelli, Giovanni. "Commentary: data analysis in bionano science—issues to watch for." *Journal of Cluster Science* (2017), pp. 1-4.

[10] Tyanova, Stefka, et al. "The Perseus computational platform for comprehensive analysis of (prote) omics data." *Nature methods* 13.9 (2016), pp. 731-740.

[11] Ammann, M., et al. "IUPAC Task Group on Atmospheric Chemical Kinetic Data Evaluation." (2016).

[12] Berger, Elisabeth, et al. "Field data reveal low critical chemical concentrations for river benthic invertebrates." *Science of the Total Environment* 544 (2016): 864-873.