QUALITY TESTING OF CHROMATOGRAPHIC DATA WITH THE AID OF A STATISTICAL CRITERION*

GEORGE LEDIN, JR., WALDEMAR R. GUSTAVSON AND ARTHUR FURST
Institute of Chemical Biology, University of San Francisco, San Francisco, Calif. (U.S.A.)
(Received October 22nd, 1965)

INTRODUCTION

An exhaustive mathematical treatment of the chromatographic process, even if of high theoretical interest, is usually of no practical value for the practicing chemist. The interpretative mathematics of biology and chemistry may provide a reasonably close explanation of the phenomena involved, but are generally inaccessible tools for the biologist or chemist who need to analyze the experimental data on hand and obtain an answer of validity. Many good theories[1] have been written and chromatography has been presented as a convolution process, or a Poisson process[2], but again almost no literature is available on the practical aspects of evaluating the data upon completion of the experiment[3].

The criterion here derived is the result of the authors' work on the statistics of the chromatography of vitamin B_6 in which it was desirable to establish a simple test for reproducibility. This criterion is a fast test for dispersion (% error) that provides a narrow confidence band and in many cases will prove to be easier to use, quicker, and better than the very well known and often misused *t*-Student's and Chi-Square tests.

THEORY

Although the R_F of a specific compound at a fixed pH should be a constant value, in actual practice these figures vary from experiment to experiment. This variation is bounded:

$$0 \leqq x_k \leqq 1 \tag{1}$$

where x_k is the k-th R_F value.

The expression x_k, of course, can be applied to all R_F values, which naturally will have bounds within the closed interval [0,1].

For statistical purposes it is convenient to normalize the x_k by defining a new variable, the variability ratio:

$$Y_k = \frac{x_k}{X} \tag{2}$$

* Supported by Contract No. AF 33 (615)-2332. Further reproduction is authorized to satisfy the needs of the United States Government.

where X is the mean R_F value. One of the common forms of the coefficient of dispersion is:

$$d = \frac{s}{X} \tag{3}$$

$$d = \sqrt{\frac{\sum_{k=1}^{N} (Y_k - 1)^2}{N - 1}} \tag{4}$$

where s is the standard deviation of the sample, and N is the number of readings.

Optimum reproducibility (replicability) conditions call for a maximum coefficient of dispersion, the magnitude of which will depend upon the difficulty of the separations. The researcher who needs to check the reproducibility of the experimental data may obviously use formulas (3) or (4); but, generally speaking, this is a tedious and time-consuming process. A much quicker way of checking if the coefficient of dispersion is within the established interval is the "Y-test".

The Y-test

In order to have $d \leqslant d_0$, it is necessary to have*

$$Y_{\max} = \frac{x_{\max}}{X} \leq 1 + d_0 \quad \text{and} \quad Y_{\min} = \frac{x_{\min}}{X} \geq 1 - d_0$$

Proof. Formula (4) may be rewritten in the following approximate form:

$$d = \sqrt{\frac{N}{N-1}}\,(Y_{\max} - 1) \tag{5a}$$

and, solving for $Y_{\max}$:

$$Y_{\max} = 1 + \sqrt{1 - \frac{1}{N}} \cdot d \tag{5b}$$

where $Y_{\max}$ is the maximum variability ratio (or, what is the same, the ratio of the highest allowable reading to the mean) permitted for a given dispersion d. Using the binomial expansion, (5b) may be rewritten as follows:

$$Y_{\max} = 1 + \left(1 - \frac{1}{2N} - \frac{1}{8N^2} - \cdots\right)d \tag{6}$$

Depending on how large N is, one of the two following approximate expressions can be used:

$$Y_{\max} = 1 + \left(1 - \frac{1}{2N}\right)d \tag{7a}$$

$$Y_{\max} = 1 + d \tag{7b}$$

---

* Necessary, but not sufficient. The speed of this test is obtained at the expense of some accuracy.

The same line of reasoning may be applied for Y_{min}. In this case, formulas (7a) and (7b) would be replaced by:

$$Y_{min} = 1 - \left(1 - \frac{1}{2N}\right)d \qquad (7a')$$

$$Y_{min} = 1 - d \qquad (7b')$$

Naturally, formulas (7) can be used to determine the Y-test for any given dispersion d. Particularly, if $d = d_0$, then, using (7b) and (7b'), $Y_{max} = 1 + d_0$, and $Y_{min} = 1 - d_0$.

EXAMPLES

To illustrate the theoretical results obtained above, two examples of vitamin B_6-amine-5PO_4 (synthetic compound and from mouse brain) will be analyzed. This technique of analysis is of course also applicable to pyridoxols and pyridoxals, and, in general, to any R_F data.

TABLE I

ANALYSIS OF THE R_F VALUES OF THE SYNTHETIC COMPOUND VITAMIN B_6-AMINE-5PO_4

Solvent at pH 6.5.

k	x_k	$\|X - x_k\|$	$\|X - x_k\|^2 \cdot 10^{-6}$
1	0.09	0.03	900
2	0.11	0.01	100
3	0.11	0.01	100
4	0.11	0.01	100
5	0.11	0.01	100
6	0.11	0.01	100
7	0.11	0.01	100
8	0.11	0.01	100
9	0.12	0.00	0
10	0.12	0.00	0
11	0.12	0.00	0
12	0.12	0.00	0
13	0.12	0.00	0
14	0.12	0.00	0
15	0.13	0.01	100
16	0.13	0.01	100
17	0.13	0.01	100
18	0.13	0.01	100
19	0.13	0.01	100
20	0.15	0.03	900

Here $N = 20$, $X = 0.12$. Let us set $d = 0.10 = 10\%$ (expecting 90% of our data to be within one standard deviation from the mean).

If we now apply the Y-test, we will obtain: $Y_{max} = 0.15/0.12 = 1.25$ and $Y_{min} = 0.09/0.12 = 0.75$, which gives $d = Y_{max} - 1 = 1 - Y_{min} = 25\%$. Hence some of our data are outside the desired confidence band. Let us assume that x_1 and x_{20} are outside.

Then, applying the Y-test again, we obtain: $Y_{max} = 0.13/0.12 = 1.08$ and $Y_{min} = 0.11/0.12 = 0.92$, which gives $d = Y_{max} - 1 = 1 - Y_{min} = 8\%$

If we were to carry out the usual computations of standard deviation, etc. we would find: $N = 20$, $X = 0.12$, $s = 0.01$, $d = 8\%$ and for the narrower band: $N = 18$, $X = 0.12$, $d = 7\%$.

A systematic procedure for a statistical analysis of this kind is the following:
1. Sort the R_F values in ascending or descending order of magnitude*.
2. Compute the mean R_F value.
3. Compute all $| X—x_k |$ and, correspondingly, all $| X—x_k |^2$.
4. From these calculations determine s and d.

The above procedure is followed in Tables I and II, and then comparisons are drawn by applying the criterion developed in Theory. The two examples show that, if the Y-test is applied, the above procedure is reduced to steps 1 and 2 only:
1. Sort the R_F values in ascending or descending order of magnitude*.
2. Compute the mean R_F value.
3. Obtain an estimate of the coefficient of dispersion by applying the Y-test (for example, $d =$ (greatest R_F)/(mean R_F) $- 1$), or see if the data falls within established confidence limits (for a desired d).

TABLE II

ANALYSIS OF THE R_F VALUES OF VITAMIN B_6-AMINE-5PO_4 FROM MOUSE BRAIN

k	x_k	$\|X - x_k\|$	$\|X - x_k\|^2 \cdot 10^{-6}$
1	0.10	0.04	1600
2	0.11	0.03	900
3	0.12	0.02	400
4	0.13	0.01	100
5	0.13	0.01	100
6	0.13	0.01	100
7	0.13	0.01	100
8	0.13	0.01	100
9	0.13	0.01	100
10	0.13	0.01	100
11	0.13	0.01	100
12	0.13	0.01	100
13	0.13	0.01	100
14	0.14	0.00	0
15	0.14	0.00	0
16	0.14	0.00	0
17	0.14	0.00	0
18	0.15	0.01	100
19	0.15	0.01	100
20	0.15	0.01	100
21	0.16	0.02	400
22	0.18	0.04	1600

Here $N = 22$, $X = 0.14$. Let us again set $d = 0.10 = 10\%$.

If we now apply the Y-test, we will obtain: $Y_{max} = 0.18/0.14 = 1.29$ and $Y_{min} = 0.10/0.14 = 0.71$, which gives $d = Y_{max} - 1 = 1 - Y_{min} = 29\%$. Hence some of our data are outside the desired confidence band. Let us assume that x_1, x_2, x_3, x_{21}, and x_{22} are outside. Then, applying the Y-test again, we obtain: $Y_{max} = 0.15/0.14 = 1.07$ and $Y_{min} = 0.13/0.14 = 0.93$ which gives $d = Y_{max} - 1 = 1 - Y_{min} = 7\%$.

If we were to carry out the usual computations of standard deviation, etc. we would find: $N = 22$, $X = 0.14$, $s = 0.02$, $d = 14\%$ and for the narrower band, $N = 17$, $X = 0.14$, $d = 6\%$.

---

* This step is not necessary. Its convenience lies in the facts that the distribution of the R_F values around the mean will be appreciated by a glance, and x_{max} and x_{min} will be easiest to pick out.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to the United States Air Force (Wright-Patterson Air Force Base, Ohio) for the financial support under which this work was carried out.

SUMMARY

Testing the reproducibility of chromatographic data with the usual statistical tests (*t*-Student's, Chi-Square, etc.) is in most cases a time-consuming error-inviting procedure. The criterion developed in this paper allows the experimenter to obtain a good estimate of the confidence bounds by applying an extremely simple test. Two examples involving the R_F values of pyridoxamine-5PO_4 illustrate the method.

REFERENCES

1 H. VINK, *J. Chromatog.*, 15 (1964) 488 and 18 (1965) 25.
2 D. A. MCQUARRIE, *J. Chem. Phys.*, 38 (1963) 437.
3 W. EDWARDS DEMING, *Statistical Adjustment of Data*, Dover Publications, New York, 1964.