Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 142 (2015) 204–209

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

New approach application of data transformation in mean centering of ratio spectra method

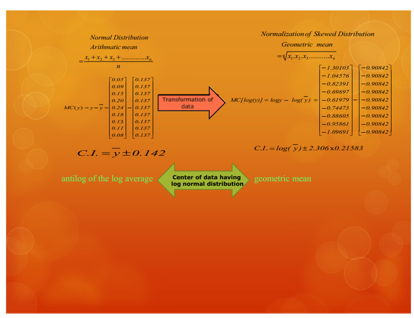
Mahmoud M. Issa [a,*], R'afat M. Nejem [b,*], Raluca Ioana Stefan Van Staden [c], Hassan Y. Aboul-Enein [d,*]

[a] Pharmaceutical Analytical Chemistry, Department of Chemistry, Al-Aqsa University, P.O. Box 4051, Gaza, Palestine
[b] Analytical Chemistry, Department of Chemistry, Al-Aqsa University, P.O. Box 4051, Gaza, Palestine
[c] National Institute of Research for Electrochemistry and Condensed Matter, Bucharest, Romania
[d] Pharmaceutical and Medicinal Chemistry Department, Pharmaceutical and Drug Industries Research Division, National Research Center, Dokki, Giza 12662, Egypt

HIGHLIGHTS

- Mean centering of ratio spectra (MCR) based on arithmetic mean.
- New approach of data transformation in MCR spectra based on geometric mean.
- Geometric mean is a better measure of central tendency than arithmetic mean.
- Logarithmic transformation and geometric mean were subjected to skewed data.
- Applied to resolve ASP, ATOR and CLOP in synthetic mixture and pharmaceuticals.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 August 2014
Received in revised form 4 November 2014
Accepted 25 January 2015
Available online 7 February 2015

Keywords:
Aspirin
Atorvastatin
Clopidogrel
Mean centering of ratio spectra – geometric mean
Arithmetic mean

ABSTRACT

Most of mean centering (MCR) methods are designed to be used with data sets whose values have a normal or nearly normal distribution. The errors associated with the values are also assumed to be independent and random. If the data are skewed, the results obtained may be doubtful. Most of the time, it was assumed a normal distribution and if a confidence interval includes a negative value, it was cut off at zero. However, it is possible to transform the data so that at least an approximately normal distribution is attained. Taking the logarithm of each data point is one transformation frequently used. As a result, the geometric mean is deliberated a better measure of central tendency than the arithmetic mean. The developed MCR method using the geometric mean has been successfully applied to the analysis of a ternary mixture of aspirin (ASP), atorvastatin (ATOR) and clopidogrel (CLOP) as a model. The results obtained were statistically compared with reported HPLC method.

Introduction

The arithmetic mean is the sum of the values divided by the n-measurements. It is the single number which best represent the center of a normally distributed group of measurements,

$$\text{Arithmetic mean} = \frac{x_1 + x_2 + x_3 + \cdots x_n}{n}$$

* Corresponding authors. Tel.: +20 1003678948; fax: +20 233370931 (H.Y. Aboul-Enein).

E-mail addresses: alaaash2002@yahoo.com (R.M. Nejem), haboulenein@yahoo.com (H.Y. Aboul-Enein).

http://dx.doi.org/10.1016/j.saa.2015.01.064
1386-1425/

The geometric mean may be calculated by multiplying each of the n measurements of data, together and taking the nth root of the product,

$$\text{Geometric mean} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

There are two types of distortion of particular interest to us. Much of the data collected by chemists have a skew distribution with a long tail and some have a normal distribution. Most of mean centering ratio spectra methods are designed to be used with data sets whose values have a normal or nearly normal distribution [1–6]. The errors associated with the values are assumed to be random. If the data are skewed, the results obtained by the mean centering ratio spectra methods may be questionable (e.g. *t*-tests and analysis of variance). Most of the time they assume normal distribution and if a confidence interval includes a negative value, it cut off at zero [7]. This renders analysis and distorts the sample arithmetic mean as a measure of central tendency. However, it may be possible to transform the data to a logarithmic scale to achieve at least an approximately normal distribution. Then the analyst will transform means and confidence intervals (C.I.) from the logarithmic scale back to original scale of measurements [8]. Statistical inferences in the logarithmic remain valid for the data. The result of the back transforming the mean of the logarithmic values to the original scale is the geometric mean. This statistic is less subject to distortion by the unusually large values in the tail of the skewed distribution of the data [8–13].

In this investigation, logarithmic transformation and geometric mean were subjected to validate statistical interferences for skewed data. As a result, the geometric mean is a better measure of central tendency than the arithmetic mean. Therefore, the errors associated with results in mean centering ratio spectra method are not assumed to be random.

Theoretical background

Using the arithmetic mean

To explain the mean centering expression, let us consider a nine dimensional vector [14]:

$$y = \begin{bmatrix} 0.05 \\ 0.09 \\ 0.15 \\ 0.20 \\ 0.24 \\ 0.18 \\ 0.13 \\ 0.11 \\ 0.08 \end{bmatrix}$$

We center (MC) this column by subtracting the arithmetic mean of nine numbers,

$$MC(y) = y - \bar{y} \begin{bmatrix} 0.05 \\ 0.09 \\ 0.15 \\ 0.20 \\ 0.24 \\ 0.18 \\ 0.13 \\ 0.11 \\ 0.08 \end{bmatrix} - \begin{bmatrix} 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \\ 0.137 \end{bmatrix}$$

The estimated standard deviation, S, is 0.0616 with 8 degrees of freedom and the coefficient of variation (the standard deviation divided by the arithmetic mean and multiplied by 100) is 44.96%. The 95% confidence interval for a single determination would be $\bar{y} \pm t_{0.05,8} \times S$ or $\bar{y} \pm 2.306 \times 0.0616 = \bar{y} \pm 0.142$.

The confidence interval equal:

$$\text{C.I.} = \bar{y} \pm 0.142$$

At the $y = 0.137$ level, this interval includes negative values, which is physically impossible. A useful rule of thumb is: If the coefficient of variation is greater than 33%, the distribution is probably too skewed for accurate estimate. This rule is based on the fact that for a skewed distribution, the arithmetic mean ± 3S will be less than zero.

This means that the calculated value of the standard deviation may not be an accurate estimate because one of the assumptions which based on the calculation is not satisfied.

Calculation using the geometric mean

Let us consider a nine dimensional vector and the transformation of data may be used, then

$$y \xrightarrow{\text{logarithmic transformation}} \log(y) = \begin{bmatrix} -1.30103 \\ -1.04576 \\ -0.82391 \\ -0.69897 \\ -0.61979 \\ -0.74473 \\ -0.88605 \\ -0.95861 \\ -1.09691 \end{bmatrix}$$

We center the column (MC) by subtracting the arithmetic mean of nine numbers,

$$MC[\log(y)] = \log y - \log(\bar{y}) = \begin{bmatrix} -1.30103 \\ -1.04576 \\ -0.82391 \\ -0.69897 \\ -0.61979 \\ -0.74473 \\ -0.88605 \\ -0.95861 \\ -1.09691 \end{bmatrix} - \begin{bmatrix} -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \\ -0.90842 \end{bmatrix}$$

where $\log(\bar{y})$ is the logarithmic average (−0.90842). The estimated standard deviation, $S_{\log}$ is 0.21583 with 8 degrees of freedom and the coefficient of variation is 23.7%. The 95% confidence interval for a single determination would be $log(\bar{y}) \pm 2.306 \times 0.21583$. Therefore the log of the confidence interval is equal.

$$\begin{aligned} \log(\text{C.I.}) &= -0.90842 \pm 2.306 \times 0.21583 \\ &= -0.90842 \pm 0.49770 = -1.40612 \text{ to } -0.41072 \qquad (1) \end{aligned}$$

To obtain the confidence interval in terms of the original data, we consider the antilog

$$(\text{C.I.}) = 0.3925 \text{ to } 0.38840$$

This range is much more reasonable in that it does not include negative values. The coefficient of variation is less than 33% and therefore a log normal distribution is found.

Note that $\log(\bar{y})$ (−0.90843) is not the same as the log of the arithmetic mean (0.137). The antilog of the $\log(\bar{y})$ is the geometric mean, which better represents the center of data having log normal