CONF-8510151--6

CONF-8510151--6

DE86 001335

DETECTING ISOTOPIC RATIO OUTLIERS

Charles K. Bayne

Computing and Telecommunications Division
Oak Ridge National Laboratory*
Oak Ridge, Tennessee 37831

David H. Smith

Analytical Chemistry Division
Oak Ridge National Laboratory*
Oak Ridge, Tennessee 37831

*Operated by Martin Marietta Energy Systems, Inc. under contract No. DE-AC05-84OR2140 for the U.S. Department of Energy.

DETECTING ISOTOPIC RATIO OUTLIERS

Charles K. Bayne[1] and David H. Smith[2]

[1]Computing and Telecommunications Division, Oak Ridge National Laboratory*, Oak Ridge, Tennessee 37831

[2]Analytical Chemistry Division, Oak Ridge National Laboratory*, Oak Ridge, Tennessee 37831

ABSTRACT

An alternative method is proposed for improving isotopic ratio estimates. This method mathematically models pulse-count data and uses iterative reweighted Poisson regression to estimate model parameters to calculate the isotopic ratios. This computer-oriented approach provides theoretically better methods than conventional techniques to establish error limits and to identify outliers.

INTRODUCTION

The most sensitive detection technique in mass spectrometry is pulse-counting, where each ion is registered as a count by the detection system. This requires an electron multiplier detector operated at high gain (>10^6) and fast amplification electronics to allow reasonable ion counting rates. Most pulse-counting systems have dead times on the order of 10-30 nsec; count rates are generally limited to 5 x 10^5/sec or less to reduce count-loss correction due to detector system dead time. Counts are accumulated for each isotope, and calculations of isotopic abundance are made on that basis. Although pulse-counting introduces a few problems not associated with analogue methods, its ability to provide reliable analyses on small samples (1 ng or less for uranium) makes it essential when sample size is limited due to lack of source material or radioactivity.

Calculation of isotopic ratios and abundances has previously been executed as if the data were entirely analogous to those generated by current integration techniques. This work presents evidence that, due to their integer nature, such treatment is not proper and introduces a model through which the problem may be addressed.

*Operated by Martin Marietta Energy Systems, Inc. under contract No. DE-AC05-84OR2140 for the U.S. Department of Energy.

approximated by a normal distribution when the expected values are large ($P(\mu)$, $\mu > 30$). However, probability distributions for differences and ratios of Poisson counts are not well known.

The problem of deriving the probability distribution for the ratio of count data is very difficult. We are not aware of any known mathematical expression for this distribution. The range of values that the ratio of counts can have is a set of rational numbers with many count ratios producing the same rational numbers (e.g., 0.5 = 1/2 = 100/200 = 1000/2000 =,...). To visualize the probability distribution of the ratio of two Poisson variates, an example for the ratio of counts from P(5) to counts from P(25) was constructed for all values occurring within $\pm$ 3 standard deviations of the mean for the two distributions. The results are given in Fig. 1. From Fig. 1, we note that the distribution of count ratios is non-symmetric, skewed to the right, and has more than one mode. In addition, values close to the mean, 0.20, can have a lower probability of occurring than values further from the mean. Values with low probability of occurring that are in the middle of the distribution are termed "inliers". This situation means that we can encounter "inliers" as well as "outliers" for our isotopic ratio estimates. There are no methods presently available to detect "inliers". As the means of the Poisson distributions increase, we would hope the situation improves. However, for expected value ratios of 100:1,000,000, the distribution is still non-symmetric, skewed to the right, and the number of modes increases.

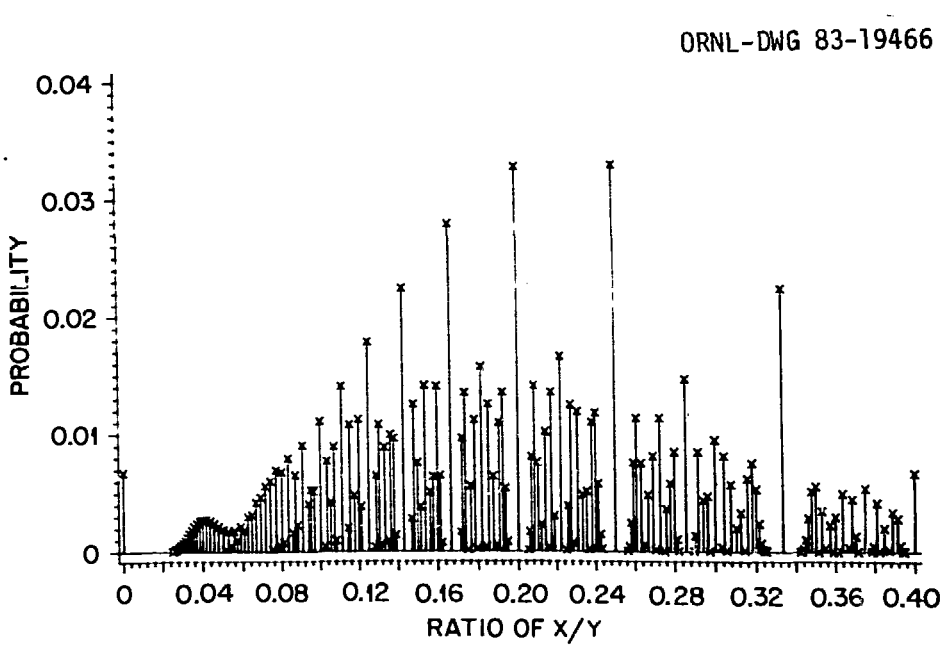


Fig. 1. Probability density function for the ratio of two Poisson variables (X/Y) with expected values of E(X) = 5 and E(Y) = 25.

The greatest handicap with the distribution of ratios of Poisson counts is that the distribution cannot be accurately approximated by normal or other well known probability distributions. This handicap prevents the evaluation of the statistical properties of the isotopic ratio estimates. Chapman [2] showed that the ratio of the expected values from two Poisson probability distributions cannot have an unbiased estimate with a finite variance, although for large counts the bias of the usual estimator is negligible. What is more important is that estimates of precision and percentile points needed to establish confidence intervals of the isotopic estimates cannot be easily obtained. The poor statistical properties of the isotopic ratio estimates necessitate investigating new methods for estimating these ratios.

ISOTOPIC RATIO ESTIMATION BY POISSON REGRESSION

An alternative method for estimating isotopic ratios, called Poisson regression [3], is based on mathematical models that account for the sources of pulse-count data. These mathematical models contain parameters that represent the isotopic ratios and are estimated by a weighted nonlinear least squares algorithm.

The pulse-count data are first adjusted for dead time and sweep factors. The adjusted data are assumed to have a Poisson distribution. That is, if Y(i,j) is the adjusted pulse-count datum in the i-th run and j-th column, its distribution is:

$$Y(i,j) \sim P[F(\underline{\theta})] \quad ,$$

where $F(\underline{\theta})$ represents the expected value of the pulse-count data as a function of the parameter vector $\underline{\theta}$. This expected value has the following form:

$$F(\underline{\theta}) = \text{background} + (\text{bias})^{-1}(\text{run-effect})\ (\text{isotopic-effect}).$$

background = exp[B(i)], represents background counts.

bias = bias factors.

run-effect = exp[R(i)], represents effects that vary from run to run, such as filament-source geometry, sample-filament chemistry, and filament temperature.

isotopic-effect = exp[I(k)], the effect due to the k-th minor isotope where the major isotope has an effect defined to be 1 and the minor isotopes are normalized to the major isotope. In other words, the minor isotopic effect is quantitatively a fraction of the abundance of the major isotope.

The mathematical model for the Es example with parameter vector $\underline{\theta}' = (B(1),\ldots,B(10),R(1),\ldots,R(10),I)$ and row $i = 1,2,\ldots,10$, is given by:

$$F[\underline{\theta}] = \begin{cases} \exp[B(i)], \text{ (background).} \\ \exp[B(i)] + (\text{bias})^{-1} \exp[R(i)], \text{ (major isotope),} \\ \exp[B(i)] + (\text{bias})^{-1} \exp[R(i) + I], \text{ (minor isotope).} \end{cases}$$

This model is nonlinear in the parameters. An iterative nonlinear least squares algorithm is used to solve for the parameter values. Because the variance of a Poisson distribution is equal to its expected value, we weight the data inversely to their variances or

$$W(\underline{\theta}) = 1/\text{var}(Y(i,j)) = 1/F(\underline{\theta}) \quad .$$

The estimation criterion is to minimize the weighted sums of squares with respect to the estimated parameter vector, $\hat{\underline{\theta}}$.

$$S(\hat{\underline{\theta}}) = \sum_{i,j} W(\hat{\underline{\theta}})[Y(i,j) - F[\hat{\underline{\theta}}]]^2 \quad .$$

To solve for the minimum of the weighted sums of squares, we first linearize the function using a Taylor series approximation. An initial estimate of the parameter vector is made, and a minimum solution can be found using ordinary weighted least squares. The initial solution is updated and the process repeated. Iteration continues until an insignificant change in the model parameter vector occurs. The solution algorithm is shown schematically in Fig. 2. The algorithm used for the solution is a modified algorithm by Frome [4] that is equivalent to finding the maximum likelihood estimates for the parameter vector.

The estimated parameter vector has several asymptotic properties:

1. $\hat{\underline{\theta}}$ is asymptotically normal.
2. $\hat{\underline{\theta}}$ is asymptotically unbiased.
3. $\hat{\underline{\theta}}$ has a lower variance than any other parameter estimate with properties 1 and 2.

These asymptotic properties give a theoretical basis for estimating isotopic ratios and establishing standard deviations and confidence intervals for these estimates.

Two diagnostic tools are available to identify outliers and judge the suitability of the model. The first tool is the Freeman-Tukey (F-T) [5] residuals that are used to identify individual pulse-count values as suspected outlier values. The F-T residuals are standardized to have an approximate normal distribution with zero mean and standard deviation of one. The second tool is the chi-square statistic that is used to judge the appropriateness of removing a suspected outlier or the fit of the model.

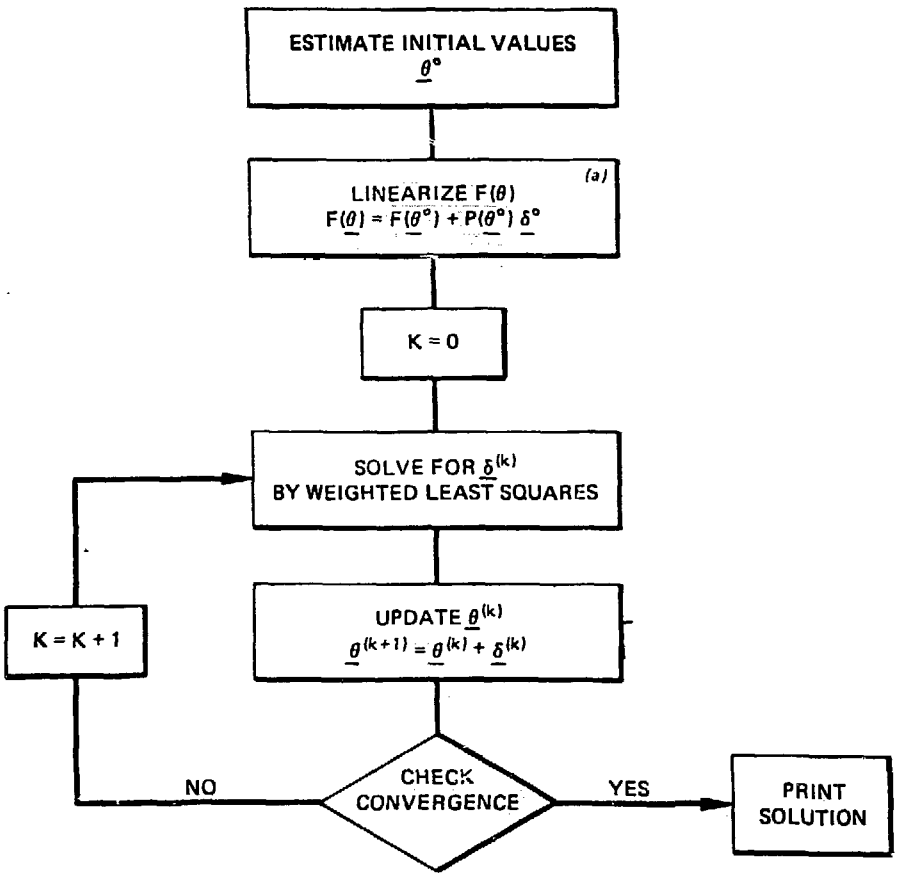
Freeman-Tukey Residual (F-T):

$$F\text{-}T = \sqrt{y(i,j)} + \sqrt{y(i,j) + 1} - \sqrt{4F[\hat{\theta}] + 1}$$

Chi-Square Statistic:

$$\chi^2 = \sum_{i,j} \frac{[Y(i,j) - F[\hat{\theta}]]^2}{F[\hat{\theta}]}$$

ORNL–DWG 83-19520A



(a) $P(\theta)$ = MATRIX OF PARTIAL DERIVATIVES OF $F(\underline{\theta})$ WITH RESPECT TO $\underline{\theta}$. $\underline{\delta}^\circ = \underline{\theta} - \underline{\theta}^\circ$.

Fig. 2. Schematic diagram of the iterature reweighted least squares algorithm.

An example of these two statistics is given in Table 2 for the einsteinium data. We note that a large value of the F-T residuals occurs in the first row for $^{254}Es^+$, indicating a potential outlier. By removing this data point and refitting the data, the probability of the chi-square statistic is increased from 0.03 to 0.59. Using the revised data, the 95% confidence interval for the Poisson regression estimate ($^{254}Es^+/^{253}Es^+$ = 0.01965) is (0.01943,0.01987).

Table 2. Freeman-Tukey residuals and chi-square statistics for einsteinium data.

Background	$^{253}Es^+$	$^{254}Es^+$
1.1	0.8	-3.0[a]
-0.0	-0.0	0.1
0.1	0.1	-0.4
-0.4	-0.4	1.5
0.3	0.3	-1.0
-0.1	-0.2	0.6
0.0	0.0	-0.1
0.3	-0.2	0.9
-0.5	-0.3	1.2
-0.2	-0.1	0.4

[a]Suspected outlier value

Chi-square statistics: (18.3 and 6.5)

1. $Pr[X^2(9) \geq 18.3] = 0.03$ (with outlier)
2. $Pr[X^2(8) \geq 6.5] = 0.59$ (without outliers)

The pattern of the F-T residuals can also be used as an indicator for potential physical problems with the sample. For example, variable fractionation of a sample occurs when one or more of the isotopes has a fluctuating volatility rate when compared to other isotopes in the sample. Table 3 gives the usual estimates for a uranium sample of the isotopic ratios $^{234}U^+/^{238}U^+$, $^{235}U^+/^{238}U^+$, and $^{236}U^+/^{238}U^+$. The ten values of each isotopic ratio were checked for outliers using the common outlier tests: Dixon's criterion, studentized deviates, and studentized ranges [6]. Each test indicated that the minimum and maximum isotopic values were not outliers at the 5% significance level. However, an F-T residual plot for the $^{235}U^+$ and $^{238}U^+$ counts in Fig. 3 shows large residual for both isotopes for the first three runs and suspected values for $^{238}U^+$ counts at run numbers four and nine. The residual pattern shows the residuals for the two isotopic counts are about the same magnitude for each run but opposite in sign. This residual pattern has been observed in other samples when variable uranium fractionation was suspected. The value of the chi-square statistic was large (X^2 = 151.11, df = 28) indicating a significant lack of fit of the model ($P < 0.001$). This result gave additional evidence of physical problems with the sample and indicates it should be reanalyzed.

Table 3. Isotopic ratios for a uranium sample.

Run	$^{234}U^{+}/^{238}U^{+}$	$^{235}U^{+}/^{238}U^{+}$	$^{236}U^{+}/^{238}U^{+}$
1	0.009130	0.9316	0.000983
2	0.009116	0.9342	0.000986
3	0.009144	0.9343	0.001002
4	0.009350	0.9409	0.000964
5	0.009344	0.9395	0.000981
6	0.009356	0.9393	0.000952
7	0.009369	0.9405	0.000977
8	0.009224	0.9394	0.000971
9	0.009257	0.9415	0.001013
10	0.009334	0.9401	0.000964
Average	0.009262	0.9381	0.000979
95% Confidence Interval	±0.000072	±0.0025	±0.000014
Poisson Regression Estimates	0.009291	0.9392	0.000977
95% Confidence Interval	(0.009251, 0.009329)	(0.9387, 0.9396)	(0.000965, 0.000990)

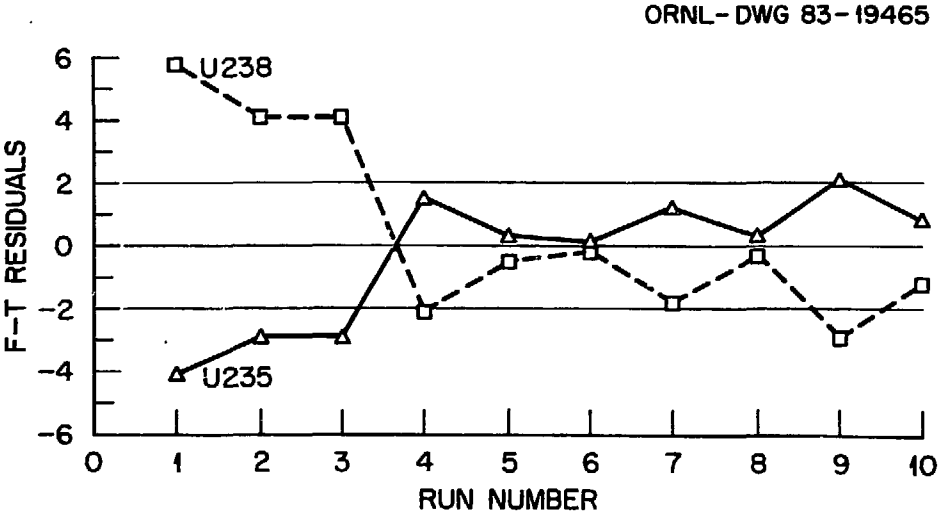


Fig. 3. Freeman-Tukey residuals for $^{238}U^{+}$ counts (squares) and $^{235}U^{+}$ counts (triangles) from a Poisson regression anaylsis.

CONCLUSIONS

We have demonstrated that the probability distribution of the ratio of two Poisson variables, such as occurs in pulse-counting mass spectrometry, is ill-defined and certainly not normal. For this reason, it is not appropriate to apply normal statistical treatment to the data. We have developed a model that successfully describes the single-element case. We are currently working on a model to describe more complex cases (isotopic interferences, etc.). The long range goal is to obtain more information concerning the quality of the analysis and thus arrive at better estimates of isotopic ratios.

It is clear that this approach transcends mass spectrometry and is applicable to many analytical techniques where data are collected as individual counts.

REFERENCES

1. E. Serge, Nuclei and Particles, W. A. Benjamin, Inc., New York, 1965.
2. D. G. Chapman, Biometrika, 49 (1952) 45-49.
3. E. L. Frome, M. H. Kutner, and J. J. Beauchamp, J. Am. Stat. Assoc., 68 (1973) 935-940.
4. E. L. Frome, "PREG: A Computer Program for Poisson Regression Analysis", ORAU-178, 1978.
5. Y.M.M. Bishop, S. E. Fienberg, and P. W. Holland, Discrete Multivariate Analysis: Theory and Practice, The MIT Press, Cambridge, Massachusetts, 1975, 130-141.
6. V. Barnett and T. Lewis, Outliers in Statistical Data, John Wiley and Sons, New York, 1978.