[CONTRIBUTION FROM THE FRICK CHEMICAL LABORATORY, PRINCETON UNIVERSITY]

Statistical Procedures for Counter-current Distribution and Differential Spectroscopy

BY JAMES E. BACHER

A formula is presented for the distribution coefficient of a single solute, as a function of counter-current distribution data. The formula is a maximum likelihood estimator. Equations are presented for the quantitative analysis of mixtures, as derived from the data of counter-current distribution by the theory of least squares. The effect of operation at slight disequilibrium on a counter-current distribution is discussed.

The technique of counter-current distribution, which has been developed by Craig,[1] is well established as a method for the fractionation of mixtures and for the detection of heterogeneity. The mathematical procedures which have been developed for the interpretation of counter-current distribution data have been adequate for the needs, in that such procedures are rapidly executed and of sufficient accuracy for the above applications.[2,3] A pure solute should be distributed according to the binomial expansion, $\left(\frac{1}{1+K} + \frac{K}{1+K}\right)^n$, wherein the concentrations of solute in successive tubes are proportional to the successive terms of the expansion. K is the distribution coefficient and n is the number of transfers, or plates.

The development of analytical methods based on counter-current distribution requires certain restrictions. Sufficiently sensitive methods of detection of solute must be available so that the solutes of a group may be distributed in low enough concentrations to behave in a nearly ideal manner. This requirement is directly opposite to that of fractionation, for which high concentrations are desired.

A general equation is derived, valid for all values of K and n, which gives a formula, or estimator, for K as a function of all the data. In addition, the properties of the binomial distribution are used to define a range of possible values of the K of a given solute as a function of n and of the number of the tube which contains the maximum concentration of that solute in a given distribution. The mathematical procedure is extended to mixtures of known constituents present in unknown amounts, so that the unknown concentrations can be calculated from the counter-current distribution, as determined by measurements of optical densities at several wave lengths. The calculated values of the concentrations are best values by the criterion of least squares.

Equations for a Single Component.—A counter-current distribution has the properties of a discrete frequency distribution. Such distributions may be characterized by their moments.[4] For a counter-current distribution the first moment, or mean position, is given by the expression $\sum_i iD_i/\sum_i D_i$. The quantity D_i is any measure of the amount of solute in the ith tube. The first tube, into which the solute is introduced, should be given the number zero. The first moment of the binomial distribution is equal to $nK/(1 + K)$. Since observed distributions generally fit the expected binomial distributions quite well, the expressions for the observed and theoretical first moments may be equated

$$m = \sum_i iD_i/\sum_i D_i = nK/(1 + K) \qquad (1)$$

The two sums can be calculated simultaneously in two or three minutes with any computing machine which has provision for cumulative multiplication. The method applies for any value of n, just as long as the advancing limb of the distribution does not overtake the trailing limb during a multiple cycling of the machine. Equation (1) gives a maximum likelihood estimate of K. The binomial expansion in K can be computed for comparison with the observed values. If T_i is the ith term of the binomial expansion, with the first term given the number zero, then the expected value of D_i is $T_i\sum_i D_i$.

As a check on the value of K, an equation similar to (1) can be written for the second moment, otherwise known as the variance, or mean square deviation

$$\sigma^2 = \frac{\sum_i i^2D_i}{\sum_i D_i} - m^2 = \frac{nK}{(1 + K)^2} \qquad (2)$$

The solution of K can be obtained in less than five minutes. The right-hand side is the variance of the binomial distribution.

The higher the order of a moment, the greater is the inaccuracy of its measurement. Consequently it is not expected that the values of K computed from a distribution by equations (1) and (2) will agree exactly. Nonetheless, for a pure substance, distributed in an accurately made machine, the two values should be close together, the value from equation (1) being the better. Any experimental factor which causes deviations from ideal behavior during a countercurrent distribution will generally cause greater errors in the second and higher moments than it will in the first. It should be pointed out that the use of the variance alone for the estimation of K is particularly poor for the case of K approximately equal to one, because the derivative of $nK/(1 + K)^2$ with respect to K is zero for K equal to one.

Another property of the binomial distribution has value. Given that the maximum concentration of a substance occurs in tube p, with the first tube numbered zero, then the value of K is limited by the inequality

[1] L. C. Craig, *J. Biol. Chem.*, **155**, 519 (1944).

[2] A. Weissberger, "Technique of Organic Chemistry," Vol. III, Chap. 4 by L. C. Craig and D. Craig, Interscience Publishers, Inc., New York, N. Y., 1950.

[3] S. V. Lieberman, *J. Biol. Chem.*, **173**, 63 (1948).

[4] P. G. Hoel, "Introduction to Mathematical Statistics," John Wiley and Sons, Inc., New York, N. Y., 1947, Chap. III.

$$\frac{p+1}{n-p} > K \geqq \frac{p}{n+1-p} \qquad (3)$$

If two adjacent tubes have equal concentrations, p applies to the higher numbered tube, and K equals the expression on the right.

If an observed distribution appears to be that of a single component, K should be computed by equation (1) and checked either by equation (2) or by the calculation of the appropriate binomial expansion and comparison with the observed points. If there are appreciable discrepancies, a single adjustment of K will usually suffice to account for the impurity. The comparison of observed points with a normal distribution as a substitute for the binomial distribution is not recommended unless the observed distribution is highly symmetric, in which case the normal distribution should be assigned the mean and variance which are defined by equations (1) and (2), respectively. The preceding statement is based on a theorem of Laplace, which asserts that the distribution of a normalized binomial variable approaches the normal distribution as n increases without limit.

The equations are illustrated with data from the counter-current distributions of cytidine (Fig. 1) and uridine (Fig. 2). The D_i were obtained by measurements of optical density with a Beckman model DU spectrophotometer. The densities can be measured at several wave lengths. Such measurements at different wave lengths are dependent with respect to the operation of the machine, but are independent with respect to measurements of absorption. Consequently, the distributions at different wave lengths afford partially independent measures of K. The values of K for cytidine, from measurements at 260, 270 and 280 mμ, are 0.106, 0.107, and 0.108, respectively, as calculated from equation (1). By inspection, the comparison with the theoretical pattern is good; the value of K obtained from equation (2) for measurements at 270 mμ is 0.115. The values of K for uridine, from measurements at 250, 260 and 270 mμ, are 0.153, 0.154 and 0.154, as calculated from equation (1). By application of equation (2) to data obtained at 260 mμ, K is 0.175. The fit is not quite so good as it was for cytidine. The distribution for K equal to 0.157 fits the points a little better, except for the first three points. It is assumed that 1 or 2% of the material is uridylic acid.

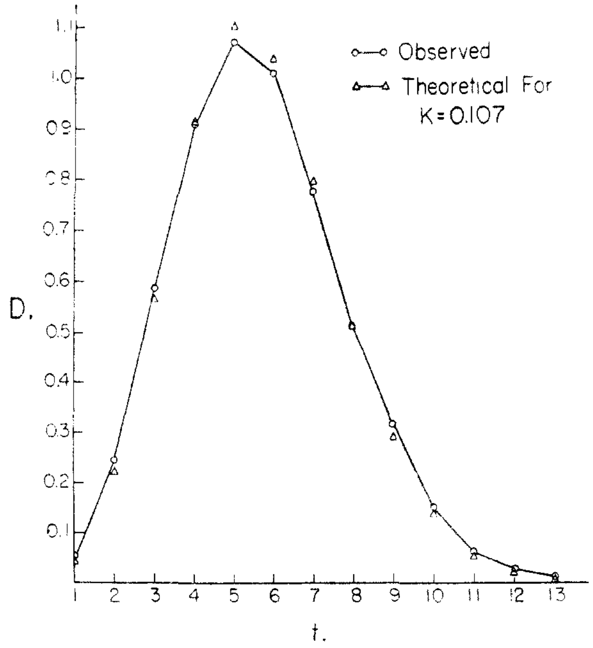

Fig. 1.—48-plate distribution of cytidine: in the figure t equals $i + 1$ of the text, *i.e.*, the solute was introduced into the tube $t = 1$.

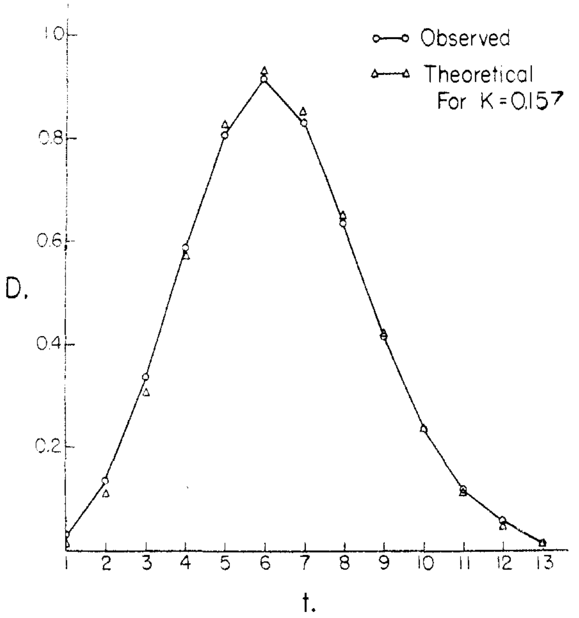

Fig. 2.—40-plate distribution of uridine; see explanation of Fig. 1.

The inequalities (3) yield a rough estimate of K very quickly. For cytidine the limits for K are 5/44 = 0.114 and 4/45 = 0.089. Since the concentration in the $(p + 1)$ tube is greater than that in the $(p - 1)$ tube, it is expected that K should be in the upper part of the range, in agreement with the value 0.107. For uridine the limits are 6/35 = 0.171 and 5/36 = 0.139. The $(p + 1)$ tube and the $(p - 1)$ tube have almost equal concentrations, so the value of K is almost in the center of the range.

It may be noted that in both figures the nature of the deviations is the same. The observed points are low in the region of the peak and high in the tail areas, relative to the calculated points. It is believed that there are two principal sources of error in the operation of the machine. The shearing plane of the machine may not exactly separate the pairs of phases. This case would be particularly true for a long run with the corresponding increased opportunity for leakage. The second source of error is operation at disequilibrium, to use the expression of Barry, Sato and Craig.[5] The essential feature of disequilibrium is that a migrating tube moving toward a peak carries less solute forward than is expected, and

[5] G. T. Barry, Y. Sato and L. C. Craig, *J. Biol. Chem.*, **174**, 209 (1948).

a migrating tube moving away from a peak carries more solute forward than is expected. Consequently a curve spreads faster, and the maximum decreases faster, than would be the case if equilibrium were established between each transfer. The symmetrical nature of operation at slight disequilibrium should permit the first moment to move forward with the theoretical rate, dm/dn, which equals $K/(1 + K)$. These conditions are precisely the ones obtained in the distributions of Figs. 1 and 2. Although the machine was rotated 50 times between each transfer,[5] perhaps only 99 per cent. of the theoretical exchange of solute was achieved in each case. Such a situation would explain the observed discrepancies. On the other hand, if K varies with concentration, the peak of a distribution will move either more rapidly or more slowly relative to the tail areas than would be expected for K independent of concentration.

The nature of disequilibrium may be considered in greater detail. The general practice is to rotate the machine a fixed number of times between plates. However, the approach to equilibrium in the mixing of two solutions is asymptotic[5]; consequently when the required exchange of solute to achieve equilibrium is relatively small, a fixed number of rotations of the machine will come closer to achieving equilibrium than if a relatively large exchange of solute is required. This point has an immediate application. At the beginning of a run, the migrating phase picks up a relatively large amount of solute from the bottom phase in the first tube of the machine, into which the solute was originally introduced. Toward the end of the run, adjacent tubes have nearly equal concentrations of solute, and relatively small amounts of solute must be exchanged at each plate to achieve equilibrium. Consequently, fewer rotations of the machine should be necessary at the end of a run than at the beginning of a run in order to achieve the same degree of approach to equilibrium at all plates.

Naturally, the concern is in regard to the last few tenths of one per cent. of the expected exchange of solute. Even though the error is too small for direct measurement, it may measurably affect the dispersion of a counter-current distribution. This possibility has been investigated with the following model. Let K^* be the partition of solute actually achieved between the two phases in a tube after a certain number of rotations of the machine. From the discussion above, the average value of K^* achieved by the trailing half of the solute in a distribution is $K(1 - x)$, and the average K^* for the advancing half is $K(1 + x)$, where x is some small fraction, *e.g.*, 0.01. Then K^* averaged over the whole distribution is just K. However the variance of the distribution is increased over the theoretically expected value $nK/(1 + K)^2$. An approximate analysis indicates that the relative increase in the dispersion of a distribution, as measured by the square root of the variance, is indeed considerably larger than x under most circumstances. It is inferred that the dispersion of a counter-current distribution is very sensitive to operation at disequilibrium.

Equations for Mixtures.—The previous discussion concerned the behavior of a single solute during counter-current distribution. It is desirable to obtain equations for use in the quantitative analysis of mixtures. It will be assumed that the components in the mixture are known, but that their concentrations are unknown. It has been shown, primarily by the splendid work of Craig and his co-workers, that observed distributions are generally in good agreement with the expected binomial distributions. Estimators for the unknown concentrations can be obtained, based upon the supposition that the solutes behave ideally and are distributed exactly according to the binomial distribution. However, such estimators have disadvantages. The formulas are complicated and computations based on them would be long and tedious in application. Furthermore, it was shown that operation at slight disequilibrium produces appreciable and systematic deviations from ideal behavior. Simpler estimators can be obtained which are independent of small deviations from ideal behavior. Instead of the assumption that the exact distribution of each solute is known, it is assumed only that the first moment of the distribution for each solute is known and is reproducible with accuracy. Measurements must be made, under standardized conditions, of the first moment of pure samples of each solute and of the molar extinction coefficients at several wave lengths.

A consideration of the equations for mixtures may start with the general equation of differential spectroscopy, as given by the theory of least squares. Let

F_j = optical density at the jth wave length, as measured on the unknown solution.

ϵ_{jk} = extinction coefficient of the kth substance at the jth wave length; all ϵ_{jk} are assumed to be known.

C_k = concentration of the kth substance in the unknown solution.

$E(F_j)$ = expected value of F_j.

Then

$$E(F_j) = \sum_k \epsilon_{jk} C_k \qquad (4)$$

If there are u different solutes, *i.e.*, $k = 1, 2, \ldots, u$, then, by the principle of least squares, there are u equations of the form

$$\sum_j \epsilon_{jk'} \sum_k \epsilon_{jk} C_k = \sum_j \epsilon_{jk'} F_j \qquad (5)$$

from which the C_k can be solved if at least u different wave lengths are used. For two components, one has

$$\Big(\sum_j \epsilon_{j1}^2\Big) C_1 + \Big(\sum_j \epsilon_{j1}\epsilon_{j2}\Big) C_2 = \sum_j \epsilon_{j1} F_j$$

$$\Big(\sum_j \epsilon_{j1}\epsilon_{j2}\Big) C_1 + \Big(\sum_j \epsilon_{j2}^2\Big) C_2 = \sum_j \epsilon_{j2} F_j$$

Once the ϵ_{jk} have been determined under standardized conditions, the coefficients of the C_k can be calculated. The calculations for any particular experiment are then readily made.

Next, the equation for the counter-current distribution of a mixture may be presented. If D_{ij} is the optical density at the j^{th} wave length of the solution in the ith tube of the counter-current distribution, then $F_j = \sum_i D_{ij}$. Let C_k be the concentration of the kth substance when all of that substance is present in the first tube of the machine, *i.e.*, before any plates are applied. Let A_{ik} be the fraction of the kth substance which is in the ith tube of the machine after the distribution has been completed. The A_{ik} will be closely related to the terms of a binomial expansion, but this information is immaterial for the present purpose. Then

$$E(D_{ij}) = \sum_k \epsilon_{jk} A_{ik} C_k$$

and

$$E(iD_{ij}) = \sum_k \epsilon_{jk} C_k i A_{ik}$$

$$E\left(\sum_i iD_{ij}\right) = \sum_k \epsilon_{jk} C_k \sum_i iA_{ik} \qquad (6)$$

$\sum_i iA_{ik}$ is equal to m_k, the first moment of the kth substance. The left side of equation (1), which is the estimated first moment of a substance when it is the only absorbing substance present, may be written $\sum_i i(D_i/\sum_i D_i)$. In this case the factor, $D_i/\sum_i D_i$, is simply the measured value of A_i. Let $G_j = \sum_i iD_{ij}$. Then equation (6) may be written

$$E(G_j) = \sum_k \epsilon_{jk} m_k C_k \qquad (7)$$

Equation (7) has the same form as equation (4), so in correspondence with equations (5)

$$\sum_j \epsilon_{jk'} m_{k'} \sum_k \epsilon_{jk} m_k C_k = \sum_j \epsilon_{jk'} m_{k'} G_j$$

where the ϵ_{jk} and the m_k are assumed to be known constants.

Since measurements are made at several wave lengths, equations (4) and (7) may be considered together to give two independent relations for each wave length at which measurements are made. The sum of squares to be minimized with respect to C_k is

$$\sum_j \left[W\left(F_j - \sum_k \epsilon_{jk} C_k\right)^2 + \left(G_j - \sum_k \epsilon_{jk} m_k C_k\right)^2\right]$$

W is a statistical weighting factor. If it is assumed that the errors in the observed quantities, D_{ij}, are normally distributed with mean zero and constant variance, and if there are t tubes in the machine, with the first tube numbered zero, then W is equal to $(2t^2 - t)/6$. The large value of W results from the fact that the G_j are much larger numbers than the F_j. Multiplication by W merely adjusts the two square terms to a common numerical level. It may be found experimentally that some adjustment is necessary in the value of W, but the above expression gives the correct order of magnitude.

The resulting equations for two components are

$$\begin{aligned} \left\{\sum_j \epsilon_{j1}^2 (W + m_1^2)\right\} C_1 + \left\{\sum_j \epsilon_{j1}\epsilon_{j2} (W + m_1 m_2)\right\} C_2 &= \sum_j \epsilon_{j1}(F_j W + G_j m_1) \\ \left\{\sum_j \epsilon_{j2}\epsilon_{j1} (W + m_1 m_2)\right\} C_1 + \left\{\sum_j \epsilon_{j2}^2 \left(W + m_2^2\right)\right\} C_2 &= \sum_j \epsilon_{j2}(F_j W + G_j m_2) \end{aligned} \qquad (8)$$

As with the equations for differential spectroscopy, the coefficients of the C_k in equations (8) need only be calculated once for a standardized method. The quantities on the right sides of equations (8) are given by the expression $\sum_j \epsilon_{jk}\left(W\sum_i D_{ij} + m_k \sum_i iD_{ij}\right)$, which equals $\sum_j \epsilon_{jk} \sum_i D_{ij}(W + im_k)$. The quantities $(W + im_k)$ are functions of i and k, and are known numbers once the procedure has been standardized. Let $q_{ik} = (W + im_k)$. Then the right sides of equations (8) are $\sum_j \epsilon_{jk} \sum_i D_{ij} q_{ik}$. This form simplifies the computations.

There is a possibility that there may be interactions among the members of a group of solutes, so that the values of the m_k may vary, depending on the composition of a mixture. Of course, before any procedure is applied to unknown mixtures, it should be tested on mixtures of known composition. Data obtained from the counter-current distributions of known mixtures would permit a critical test for interactions. Equation (7) is symmetrical with respect to m_k and C_k. It is only necessary to apply the method of least squares to equation (7), treating the m_k as unknown parameters and the C_k as known numbers. The sum of squares $\sum_j \left(G_j - \sum_k \epsilon_{jk} m_k C_k\right)^2$ is minimized with respect to m_k. For two components the equations are

$$\left\{\sum_j (\epsilon_{j1} C_1)^2\right\} m_1 + \left\{\sum_j (\epsilon_{j1} C_1 \epsilon_{j2} C_2)\right\} m_2 = \sum_j \epsilon_{j1} C_1 G_j$$

$$\left\{\sum_j (\epsilon_{j1} C_1 \epsilon_{j2} C_2)\right\} m_1 + \left\{\sum_j (\epsilon_{j2} C_2)^2\right\} m_2 = \sum_j \epsilon_{j2} C_2 G_j$$

Analyses of a graded series of known mixtures by the standardized procedure would permit a characterization of the behavior of each solute for all conditions which are expected during the analysis of unknown mixtures of the given group of substances. To be rigorous, the effect of interactions on the values of the extinction coefficients should also be examined.

The application of equations (8) requires a procedure which is considerably more complicated than the procedure of differential spectroscopy alone. However, the application of countercurrent distribution combined with differential spectroscopy has two advantages over the use of the latter procedure alone. The number of independent relations is doubled and the precision should be increased thereby. More important, traces of contaminants are much more likely to be detected in the counter-current distribution patterns than in the absorption spectra alone. The presence of contaminants which are absorbing in the applied spectral range would invalidate equations (4) and (7), but an analysis could still be obtained from the counter-current distribution patterns. Up to this point it has not been necessary to consider the shape of a distribution. However, it has been shown that the distributions correspond closely to the appropriate binomial expansion. Thus, once the concentrations in a mixture have been calculated, the expected composite distribution can be calculated. If the calculated and observed distributions are in essential agreement, one has quite convincing evidence for accuracy and for the absence of substances other than those which were expected.

Partial consideration of the shapes of the individual distributions can be made by the addition of terms based on the second moments into the general sum of squares. Thus equation (2) may be written

$$\frac{\sum_i i^2 D_i}{\sum_i D_i} = m^2\left(\frac{nK + 1}{nK}\right)$$

From a development similar to the development of equation (6)

$$E(\sum_i i^2 D_{ij}) = \sum_k \epsilon_{jk} C_k m_k^2 \left(\frac{nK_k + 1}{nK_k}\right)$$

K_k is the distribution coefficient of the kth substance. The incorporation of the additional set of independent, but less accurate, square terms into the general sum of squares should increase the precision of the analysis. It may be an improvement to use an empirical value for the variance, rather than the theoretical value above.

The cyclic nature of the machine may give rise to circumstances which may be misinterpreted. Let it be supposed that one has a machine with t tubes, numbered from 0 to $t - 1$. If a substance for which the distribution coefficient is 1.0 is distributed with $3t$ plates and if equation (1) is used to calculate K, the summation over i must be from 0 to $3t$, and the tubes from t to $2t - 1$ must be considered to be the tubes which contain the solute, with the other tubes empty. This decision could be made from the shape and variance of the distribution, if an approximate value of K were not known. On the other hand, in the equations for the analysis of mixtures, the summation over i is always from 0 to $t - 1$, regardless of the number of plates which are applied in the standard procedure. Thus m_k may have no meaning in terms of the distribution coefficient of the kth substance. For example, let it be supposed that for the kth substance the distribution coefficient is equal to 1.0, and the standard procedure requires $2t$ plates. The distribution of the kth substance will be centered in the tube which is numbered zero with the advancing limb in the lowest numbered tubes, and the trailing limb in the highest numbered tubes. The value of m_k is $(t - 1)/2$, which is directly opposite on the machine from the actual position of the substance. Yet $(t - 1)/2$ is the correct value of m_k to use in equations (8). In this connection, it is important that the standard conditions separate the values of m_k as much as possible. As a guide in the selection of the best value for the number of plates in the standard procedure, it may be pointed out that a plot of m_k, as a variable, *versus* the number of plates is the plot of a damped oscillator.

It is clear that if a pair of solvents is available which is known to separate readily the components which occur in a mixture to be analyzed, then there is no need to apply equations (8). However, the equations were developed for the analysis of mixtures of the pyrimidine ribosides, which can be prepared in a quantitative manner from nucleic acid,[6,7] and for which a good pair of solvents is yet to be found. Small amounts of free purine substances, which may be found in the mixtures of pyrimidine ribosides, are readily separated from the ribosides by counter-current distribution.

The author wishes to express his appreciation to Dr. Elizabeth L. Scott, of the Statistical Laboratory, University of California, and to Mr. Carl Bennett, associated with the Department of Mathematics, Princeton University, for their kind advice and many valuable suggestions.

[6] J. E. Bacher and F. W. Allen, "The Dephosphorylation of Nucleotides; Their Analysis by Counter-Current Distribution," In press.

[7] S. E. Kerr, K. Seraidarian and M. Wargon, *J. Biol. Chem.*, **181**, 761 (1949).

PRINCETON, NEW JERSEY RECEIVED AUGUST 21, 1950