JOURNAL OF COMPUTATIONAL PHYSICS 12, 143–149 (1973)

Note

A Nonfunctional Method for Reducing Cumulative or Histogram Data to a Smooth Distributional Form or for Constructing a Smooth Approximation to Experimental Data

This note is concerned with the solution of three types of problem which commonly arise in studies of a scientific or statistical nature. These problems are: (a) the estimation of a continuous frequency distribution from experimental data in cumulative form; (b) the estimation of a continuous frequency distribution from experimental data in histogram form; and (c) the estimation of a continuous curve which describes an array of experimental data points.

For situations in which the solution has some known analytic form the usual approach would be to use standard probability plotting or curve fitting methods. This note is, however, concerned with those examples in which the analytic form is not known or does not exist. The method of solution used is basically that due to Phillips [4] and Twomey [5]. The method is particularly suitable for use by the nonspecialist since the shape of the solution and the degree to which it fits the data is controlled by the value of a single smoothing parameter. In addition, the solution form may also be controlled by data weighting and facilities for presetting the solution values or gradients at any desired points along its range. This enables a priori information to be incorporated into the type of solution derived.

For a system described by the continuous frequency distribution $D(r)$, which is sampled experimentally to provide information on the distribution in cumulative or histogram form, the following integral relationship holds,

$$I_0 + \int_{r_i}^{r_j} D(r')\,dr' = F_j + \epsilon_j\,; \qquad j = 1, 2 \cdots n; \quad i = 1 \quad \text{or} \quad j - 1, \tag{1}$$

where F_j denotes the measured value, ϵ_j its associated error, and r_i and r_j the range values of the independent variable associated with the measurement. I_0 is equivalent to the integral $\int_{-\infty}^{r_1} D(r')\,dr'$ and is referred to as the zero integral.

The three problems posed above may now be described in terms of Eq. (1) as follows:

(a) For cumulative information, C_j,

$$\begin{aligned} &F_j \equiv C_j\,; \qquad i = 1 \quad \text{for all } j, \\ &I_0 \text{ variable} \end{aligned} \tag{2}$$

581/12/1-10

(b) For histogram information, H_j,

$$F_j \equiv H_j; \quad i = 1 \quad \text{for} \quad j = 1, \quad i = j - 1 \quad \text{for} \quad j = 2, 3 \cdots n,$$
$$I_0 = F_1 = \epsilon_1 = 0 \tag{3}$$

(c) An approximation, $F(r)$, of the information F_j is obtained by solving as for the cumulative case and then setting,

$$F(r) = I_0 + \int_{r_1}^{r} D(r')\, dr'. \tag{4}$$

When expressed in these forms all three problems then reduce to that of solving Eq. (1) for I_0 and $D(r)$.

Equation (1) may be solved by quadrature using the smoothing technique due to Phillips [4] and Twomey [5]. Thus, Eq. (1) is first written in the form

$$(\Delta r/3)\, Ad = f + \epsilon, \tag{5}$$

where f is an n-vector representing the values of F_j; ϵ is an n-vector representing the values of ϵ_j plus error terms introduced by the quadrature; d is an $(m + 1)$-vector in which the first m values denote $D(r)$ evaluated at $(m - 1)$ equispaced intervals, Δr, in the range r_1 to r_n, and the $(m + 1)$th value represents $I_0/\Delta r$; A is an n by $(m + 1)$ quadrature matrix operator which is chosen to numerically integrate d over the particular ranges specified by the problem. The factor $\Delta r/3$ is simply a scalar multiplier associated with A. The smoothing solution for Eq. (5) is then given by

$$d = (3/\Delta r)(\tilde{A}A - \alpha D)^{-1}\, \tilde{A}f, \tag{6}$$

where α is an undetermined Lagrange multiplier the value of which determines the amount of smoothing produced, $\tilde{A}$ denotes the transpose of A, and D is a square matrix of dimension $(m + 1)$ by $(m + 1)$, which arises from the smoothing process used. For the present problem adequate smoothing is obtained by minimizing the sums of the squares of the first or second differences, in which case D has essentially the form of either a second or fourth difference operator, respectively.

By modifying Eq. (5) it is possible to derive solutions for d in which any particular element or elements have preset values. Suppose, for example, the elements d_k are preset for given values of k. Equation (5) is then rewritten

$$(\Delta r/3)\, A'd' = f' + \epsilon, \tag{5'}$$

where the prime denotes a modified form such that $A' \equiv A$ modified to remove its kth columns A_k, $d' \equiv d$ modified to remove its kth elements d_k, and

$$f' = f - \sum_k (\Delta r/3)\, A_k d_k .$$

Proceeding as for the previous case, a smoothing solution is then obtained of the form

$$d' = \left(\frac{1}{\alpha}\bar{A}'A' - D'\right)^{-1}\left(\frac{3}{\alpha\,\Delta r}\bar{A}'f' + \sum_k (D_k)'\, d_k\right), \tag{6'}$$

where $D' \equiv D$ modified to remove its kth columns and rows and $(D_k)'$ denotes the kth columns of D modified to remove their kth elements.

In addition to presetting particular elements of the solution it is also possible to incorporate weighting of the data points. This can be done with the aid of the diagonal weighting matrix W of dimension n by n, which contains the required weighting values along its diagonal. Upon incorporating this it can be shown that the solution then becomes

$$d' = \left(\frac{1}{\alpha}\bar{A}'WA' - D'\right)^{-1}\left(\frac{3}{\alpha\,\Delta r}\bar{A}'Wf' + \sum_k (D_k)'\, d_k\right). \tag{7}$$

It should be noted that the practical consequencies of presetting the elements of d depend upon the type of problem being considered. Thus, for problems (a) and (b) (see Eqs. (2) and (3)) the effect is clearly to force the distributional solutions to have given values at the selected points. On the other hand, for problem (c) (see Eq. (4)) d describes the gradient behavior of the solution. Consequently, presetting elements of d in this case produces a solution with given local gradient values. However, the actual solution values may also be controlled, if required, by supplying highly weighted data values at the selected points.

A Fortran IV computer program called SMUVIT has been written to execute the foregoing procedure. The program has been written as a multipurpose package which can analyze the input data either as cumulative, histogram or curve fitting information. Facilities for data weighting and presetting solution values as discussed above are also incorporated. The results of all the following examples were computed using this program. These examples were chosen purely to illustrate the applications of this smoothing technique. No attempt has been made to derive 'best' solutions; the results have simply been chosen to illustrate the types of solution which are produced.

In the first example exact data corresponding to a normal frequency distribution were analyzed. These data provided cumulative values for unit intervals of the independent variables in the range -3 to $+3$. The data values were all given unit

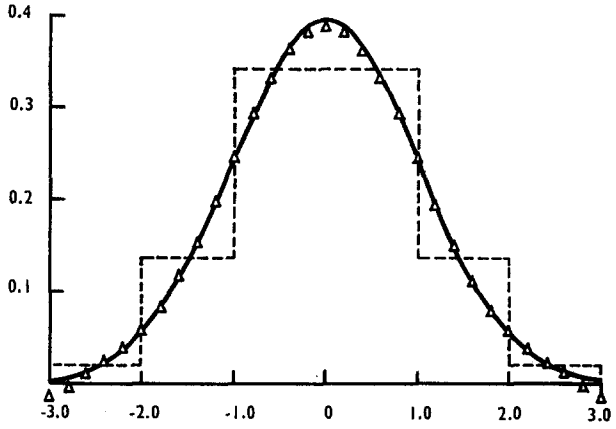

FIG. 1. Reproduction of a normal distribution from exact cumulative data. ——— true curve; - - - - - input data in histogram form; △ smoothing result for α equals 10^2.

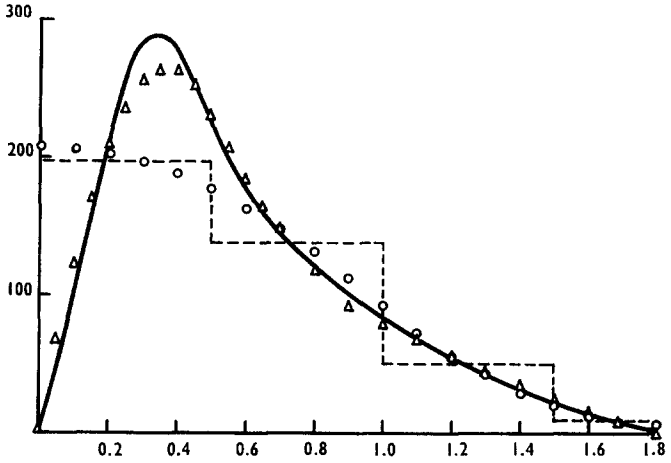

FIG. 2. Reproduction of a skewed distribution from exact histogram data. ——— true curve; - - - - - input histogram data; △ and ○ smoothing results with and without preset tail values.

weighting and no attempt was made to preset any of the distributional elements. The result obtained using second difference smoothing with α equals 10^2 is presented in Fig. 1 together with the true normal distribution and the histogram corresponding to the input data. Reproduction of the true distribution is seen to be good except for the tail regions which tend to go negative. The latter could, however, be prevented if required by presetting the tail values to zero or some small positive value.

The value of the presetting facility is demonstrated by the second example which considers data relating to the frequency size distribution of an atmospheric aerosol [2]. This data was supplied in histogram form as illustrated in Fig. 2 and distributional solutions were derived using first difference smoothing with α equals 10. With no presetting the solution (given by circles) is monotonic and bears little resemblance to the true distribution. This result is understandable from the shape of the input histogram and is, in fact, the point Fuchs [2] makes concerning the misleading nature of the data. However, if, acting on a priori information the initial distributional value is preset zero, a solution (given by triangles) is then derived which resembles fairly accurately the true distribution.

The application of the technique to approximation problems is well illustrated by considering an example used by Guest [3] when discussing methods of polynomial regression. This data together with the results obtained using second difference smoothing with α equals 1 and 10 are presented in Fig. 3. Polynomials

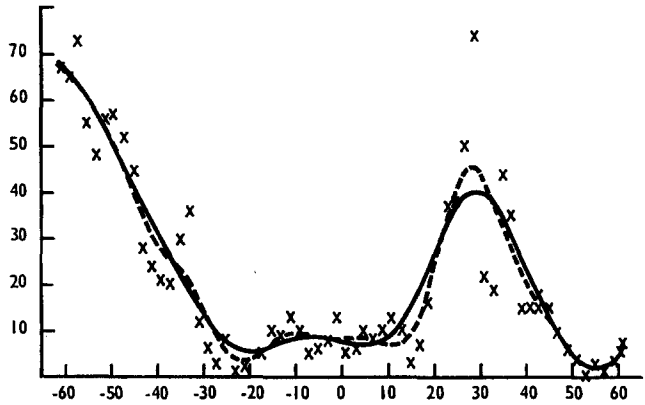

FIG. 3. Production of a smoothed approximation to scattered data. × input data; – – – – – smoothing result for α equals 1; ——— smoothing result for α equals 10.

of up to degree eight were also fitted to these data, but it was found that the detailed structure exhibited by the smoothing results could not be reproduced by these polynomials. Whether or not this structure is meaningful will, of course, depend on the data. The main point of the example is, however, to illustrate that the smoothing technique is the more suitable of the two for approximating data containing such a structure.

The final examples, Figs. 4 and 5, illustrate some of the effects of scatter on reproducing distributions from inexact histogram data. This data was obtained by drawing 50 observations at random from the distributions under consideration and was due originally to Fryer [1]. The least smoothing results are seen to follow to some extent the variations in the data, but these are sufficiently smoothed out in

the large smoothing results to provide a reasonably good representation of the true distributions.

It is evident from the previous examples that the smoothing technique described herein provides a useful means of interpreting cumulative or histogram data in terms of a smooth frequency distribution. Although potentially misleading data such as that illustrated by Fig. 2 may produce incongruous results, the application of a priori information in conjunction with the presetting facility provides a means of avoiding this.

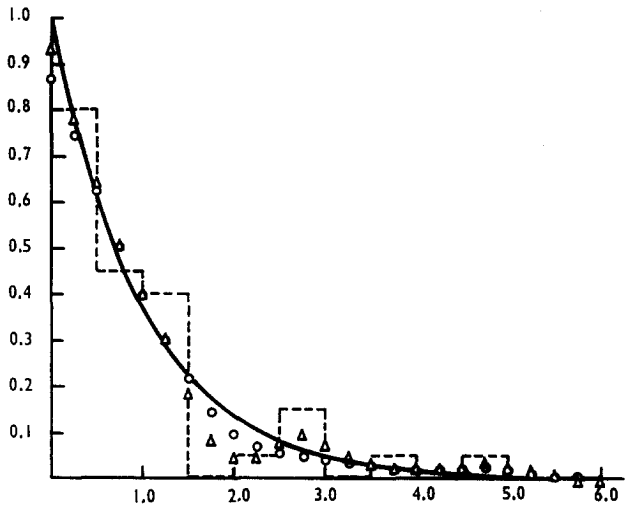

FIG. 4. Reproduction of an exponential distribution from scattered data. ——— true curve; - - - - - input histogram data; △ and ○ smoothing results for α equals 10 and 10^2, respectively.

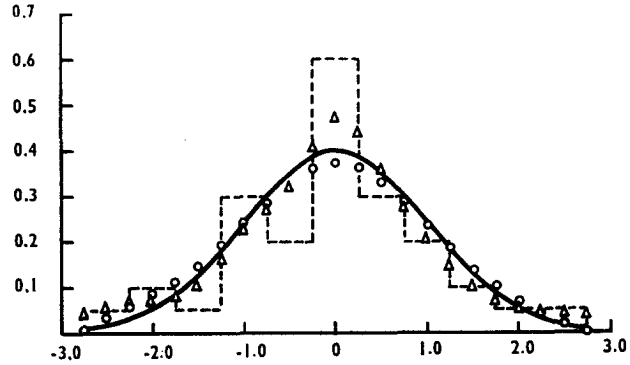

FIG. 5. Reproduction of a normal distribution from scattered data. ——— true curve; - - - - - input histogram data; △ and ○ smoothing results for α equals 10 and 10^2, respectively.

The smoothing technique has, in addition, useful applications in the field of data approximation. By simply varying the value of the smoothing parameter, the nature of the approximation can be changed from a relatively simple shape to one of much greater detail, greater, in fact than normal polynomial approximations will provide. Facilities also exist by which the solution gradient may be preset at any required points. Although this had no application in the examples considered, it would clearly be useful for that type of problem in which the asymptotic behavior of the solution was known.

A further facility which was not demonstrated by example is that of data weighting. Increasing the weighting of a data point both decreases the difference between the measured and estimated values at that point and also decreases the level of smoothing in the neighborhood of the point. By this means, therefore, the level of smoothing over the entire solution may be varied as required.

ACKNOWLEDGMENT

The work was carried out at the Central Electricity Research Laboratories, Leatherhead, England, and the paper is published by the permission of the Central Electricity Generating Board.

REFERENCES

1. M. J. FRYER, *IMA Bulletin* **7** (1971), 323.
2. N. A. FUCHS, "The Mechanics of Aerosols," p. 9, Pergamon, Oxford/London/Edinburgh/New York/Paris/Frankfurt, 1964.
3. P. G. GUEST, "Numerical Methods of Curve Fitting," pp. 194–195, Cambridge Univ. Press, London/New York, 1961.
4. D. L. PHILLIPS, *J. Assoc. Comput. Mach.* **9** (1962), 84.
5. S. TWOMEY, *J. Assoc. Comput. Mach.* **10** (1963), 97.

RECEIVED: May 8, 1972

P. M. FOSTER

Central Electricity Research Laboratories
Kelvin Avenue, Leatherhead
Surrey, England