ON FITTING CURVES TO OBSERVATIONAL SERIES BY THE METHOD OF DIFFERENCES

By

HARRY S. WILL

I. PRELIMINARY STATEMENT

Curve fitting may be technically described as the representation of a series of observations by a mathematical function. Given the observations and the function to be fitted, the problem is to determine the constants of the equation in such a way as to secure a valid representation. The method to be employed in the determination of these constants must take into account the object which the fitting process is intended to serve. If the object is to interpolate for undetermined items between specified ordinates of the series, any method which will give the constants of the equation will suffice, since the representation of the given ordinates is exact. In this case, questions of method will hinge on considerations of convenience. If, however, the object is to secure the representation of *all* the items of the series by means of a single function, questions of method will hinge on the validity of the representation, which, in this case, can only be approximate.

Functions used as approximate representations of observational series fall into two general classes: first, those which have the force of a law descriptive of a necessary sequence of events; and, second, those which depict a norm as a characteristic trend in growth. These two types of representation merit separate methodological consideration; and, in what is to follow, we shall make an analysis of the problems involved and develop a method, which, it is believed, will place in the hands of the statistician a new and serviceable instrument.

II. FUNDAMENTAL TYPES OF OBSERVATIONAL SERIES

For the purpose of fixing attention on certain characteristics of observational data, let us consider two distinctly different sorts of series. Let us suppose that the first series consists of a set of observations on a comet moving through space, and that the second consists of the record of gold production in the United States.

For the sake of simplicity, let us further suppose that the movement of both series is properly represented by the function $y=f(x)$. The two sets of observations may then be represented by an equation of the form

$$(1) \qquad Y \pm d \pm \epsilon = f(x) = Y \pm v$$

In this equation, the term ϵ represents an error of observation due to factors such as faulty judgment, clerical inaccuracies, and lack of precision in the use of instruments. The term d represents the deviation of the fitted function from the true magnitude of the phenomenon undergoing examination, after the series has been corrected for the errors ϵ. Taken together, d and ϵ make up the residuals

$$v = f(x) - Y$$

Now it is quite evident that, in the case of the first series, owing to the regularity of the path of the moving body, the deviations d will be negligibly small in comparison with the errors ϵ, and that, in the case of the second series, owing to the irregularity of production, the deviations d will be large in comparison with the errors ϵ. In fitting a curve to the first series, we assume that a true value exists and that the observational errors may be defined by the fitting process; while in fitting to the second, we assume a normal value merely, and seek to define the deviations of the observations from this norm.

These considerations suggest that the procedure which is applicable to the determination of constants in the one case may not be applicable in the other. Let us therefore inquire as to the solutions best suited to each case.

III. The Classical Solution

It was in 1806 that Legendre formulated his test of the validity attaching to the functional representation of an observational series. This formulation has become known as the principle of least squares and may be stated thus: *Where the constants of a mathematical function are to be determined from a set of empirical observations, that solution is best which makes the sum of the squares of the residual errors a minimum.*

So far as its mere statement is concerned, this principle is a rule of thumb which may be adopted or discarded at the discretion of the

individual. The principal has, however, been placed on a definite logical basis by Gauss and later writers, who have derived it from the normal law of error $p(x) = m \int_0^z e^{-z^2} dz$ Under the assumptions of this law, deviations from the most probable value are fortuitous in character, the term *fortuitous* implying that individual deviations are unanalytic in the sense that the forces operating to bring them about cannot be resolved into more elemental components. All that we can clam to know *a priori* about the values of such deviations is that they are as likely to be positive as negative and that they must remain within the bounds $\pm \infty$. The function $p(x)$ gives the probability for the occurrence of a deviation of magnitude $z = x/\sigma$.

Statisticians generally have accepted the principle of least squares as providing a sufficient theoretical basis for the fitting of curves to all sorts of series. Because of this, it becomes all the more important that certain limitations of the principle and its application to the analysis of statistical series should be carefully noted.

Considering again the case where the observations are made on a body moving through space, we see that the errors of observation committed may properly be regarded as fortuitous in character, for, on the basis of our assumption of precise motion in the path $y = f(x)$, the most probable value of the residuals is clearly defined as zero, so that the errors committed are as likely to be positive as negative; no finite bound can be set as to the possible magnitude of such random errors, and the forces determining their magnitudes cannot be resolved into their components. If our assumption as to the path of the moving body is valid, these errors conform to the normal law in the frequency of their occurrence, and their magnitudes may be accurately ascertained by a least squares determination of constants.

Returning now to the case where the observations consist of a record of gold production, can we claim to have the same basis for an application of least squares to the determination of our line of best fit? Two important considerations would lead us to think otherwise. The first of these is that the magnitude of deviations from trend is definitely restricted; for production is limited both by the capacity of the extractive industries and by the consumers' demand. The second is found in the highly analytic character of these deviations; for it is significant that whenever it becomes possible to resolve the forces determining the values of given deviations of a set into their elemental components, the prediction of the sign and magnitude of specified devia-

tions becomes in some measure possible; and when this occurs, such deviations are removed from the category of the fortuitous and unpredictable and placed in that of the analytic and predictable.

The arguments are supported by the use made of weighted deviations from trend in the forecasting of economic events. A rise in price or fall in production is not explained, in comparison with the normal trend, as a circumstance which is to be expected a certain number of times in a thousand, but rather because analysis shows the rise or fall to be the necessary result of known events. Obviously, a forecast based on unanalytic and purely fortuitous deviations could have no real significance whatever.

Granting that residuals may sometimes be obtained by least squares operations which may be regarded as a random sample of an approximately normal distribution, it must be clearly borne in mind that these residuals are brought into being by the creative act of curve fitting; and the mere marking off of a deviation does not justify our regarding it as being due to the working of forces distinct and different from those effective in producing the remaining part of the ordinate. In the case of the celestial observations which we have assumed, the act of fitting defines, but does not create, the errors.

The argument may be advanced at this point that it is not necessary to regard the principle of least squares as resting on the law of error; for we may obtain the normal equations from which our least squares determination is made by treating the solution as a simple problem in maxima and minima. But if we do, we cannot claim to have determined the *most probable* values of our constants; for this claim must rest on the derivation of the normal equations from the law of error.

The justification for the arbitrary use of the least squares technique that is most likely to be made is that it minimizes extreme deviations from the fitted line. This is unquestionably true; but it appears as a weakness of the method in the present connection rather than as an element of strength: for, in a least squares deduction of normal equations, we may regard each absolute deviation as being weighted with its own magnitude, deviations less than the mean deviation receiving weights less than the mean weight, and *vice versa*; and why, the query obtrudes, should we, in our determination of constants, overweight the observations most remote from what we term the norm

and underweight those which lie closest?

The argument that the least squares fit will avoid the commission of extreme errors in the projection of the curve beyond the limits of observation, or at least tend in that direction, is fallacious; for the fitting of a line to a given set of observations to secure the minimum sum of the squared residuals is unlikely to effect the same end when new observations are added. At least, we have no logical basis for the expectation of such a result unless we fall back on the position that the fitted curve describes a necessary sequence of events and that the residuals are fortuitous in character; and this is the very assumption we have found to be untenable for most economic and social series.

We may, then, say that fortuitous deviations are properly to be regarded as functions of the observations; while analytic deviations are to be regarded as functions of the hypothesis we set up with reference to the type of curve which is most appropriate to the data. In brief, our reasoning supplies a definite basis for the contention that, for data in which the errors of observation are small in comparison with the analytic deviations from trend, the least squares definitions do not lead to results which are to be regarded as necessarily best for all purposes.

IV. The Method of Differences

The method of curve fitting which is now to be presented was originated by the writer in the spring of 1925. Since that time, it has been put to a wide variety of practical tests and has been found to yield highly satisfactory results. The designation *method of differences* has been given to it because of the extensive and essential use made of the calculus of finite differences.

Before undertaking the task of deriving the formulas for the determination of constants, let us state the assumptions on which the method is based, as follows:

(a) The function to be fitted is logically appropriate to the data.

(b) The data are free of constant and systematic errors.

(c) Accidental errors of observation are relatively small and unimportant.

(d) Where a set of secular values is irregular and without sig-

nificant trend, the arithmetic mean is the best representation of the set.

The first of these assumptions is, in a general way, implied in any method of fitting. The effect of the second and third is to qualify the fitted deviations as analytic. The fourth is made use of constantly in the writing of formulas for the determination of parameters.

In the derivation of formulas, the essential steps are as follows: (1) equations defining each constant of the function fitted are developed by a process of differencing; (2) equations are formed from which approximations to the value of the given constant may be obtained; (3) the mean of the several approximations to the value of the given constant is taken as the most plausible value of the constant.

V. Notation

To avoid the possibility of misunderstanding, we shall explicitly define certain symbols made use of in this memoir.

The original observations are denoted by the symbol Y_i, $i = 0, 1, 2, \ldots n-1$; nd other capitals are used to designate empirical functions of the original observations; e. g., $U_i = Y_i - Y_o$. The symbol u_i denotes values of mathematical functions corresponding to the observations Y_i The argument is denoted by the symbol

$$x_i, \quad x_i = i\Delta_x$$

Summations within the definite bounds a and b is indicated by the symbol $\sum_a^b$; e. g., $\sum_{i=0}^{i=n-1} y_i = y_o + y_1 + \ldots + y_{n-1}$.

Finite differences of order r and rank k are defined by the symbol Δ_k^r; e. g., $\Delta_k^r y_i = \Delta_k^{r-1} y_{i+k} - \Delta_k^{r-1} y_i$, where the difference of zero order is taken as the quantity undifferenced. In particular, we have $\Delta_k^0 y_i = y_i$; $\Delta_k y_i = y_{i+k} - y_i$; $\Delta_k^2 y_i = y_{i+2k} - 2y_{i+k} + y_i$; $\Delta_k^3 y_i = y_{i+3k} - 3y_{i+2k} + 3y_{i+k} - y_i$.

In these relations, the values of k and r are integral. The value of y_{i+k} is precisely the value of the function $y = f(x)$ when $x = x_i + k\Delta x$. In the difference operations of the following sections, the usage $\Delta_x^2 = (\Delta x)^2 \neq \Delta(x^2)$ is adhered to. Note that $\Delta_k \log y_i = \log y_{i+k} - \log y_i$, and also that $\log \Delta_k y_i = \log (y_{i+k} - y_i)$.

Since, on taking logarithms, ratios resolve themselves into differences between logarithms, we have, analagous to the differences $\Delta_k^r y_i$, the ratios $p_k^r y_i = p_k^{r-1} y_{i+k} : p_k^{r-1} y_i$, where the ratio of order $r = 0$ denotes the specified quantity. In particular, we have $p_k^0 y_i = y_i$; $p_k y_i = y_{i+k} : y_i$; $p_k^2 y_i = y_{i+2k} \cdot y_i : y_{i+k}^2$; $p_k^3 y_i = (y_{i+3k} \cdot y_{i+k}^3) : (y_{i+2k}^3 \cdot y_i)$.

In forming the first differences $\Delta_k y_i$, where k is the increment in the y subscript corresponding to the increment $k\Delta x$ in x_i, it will be noted that the first k values of y_i are excluded as minuend; hence we can form but $n-k$ first differences of rank k, that is, when the increment in the y subscript is k. Similarly, in forming the second differences $\Delta_k^2 y_i = \Delta_k y_{i+k} - \Delta_k y_i$, the first k values of $\Delta_k y_i$ are excluded from appearance as minuend; hence we can form but $n-k-k = n-2k$ second differences from n values of y when the rank of differences is k. In general, when the rank of differences is k, we may form $n-rk$ differences of order r from a set of values of y. Evidently, the number of ratios which may be formed from a given number of observations follows the same rule as that which applies to differences.

VI. Linear Series

Let us write the equation of the linear series in the form

$$y_i = a + bx_i \tag{1}$$

Giving to x the increment $k\Delta x$, we get

$$y_{i+k} = a + b(x_i + k\Delta x) \tag{2}$$

Subtracting (1) from (2), we get

$$\Delta_k y_i = bk\Delta x \tag{3}$$

By making the substitution $\Delta_k Y_i$ for $\Delta_k y_i$ in equation (3), we may form $n-k$ approximations to the value of b, as follows:

$$\begin{aligned} b_0 &= \Delta_k Y_0 : k\Delta x \\ b_1 &= \Delta_k Y_1 : k\Delta x \\ &\cdots\cdots\cdots \\ b_{n-k-1} &= \Delta_k Y_{n-k-1} : k\Delta x \end{aligned} \tag{4}$$

Similarly, when b is determined, by substituting Y_i for y_i in equation (1), we are able to form n approximations to the value of a, as follows:

$$
\begin{aligned}
a_0 &= Y_0 - bx_0 \\
a_1 &= Y_1 - bx_1 \\
&\cdots\cdots \\
a_{n-1} &= Y_{n-1} - bx_{n-1}
\end{aligned}
\tag{5}
$$

By taking mean values of the approximations specified in equations (4) and (5), we arrive at the following formulas for determining the value of the parameters b and a:

$$
\begin{aligned}
b &= \left[\sum_{i=0}^{i=n-k-1} \Delta_k Y_i\right] : [k(n-k)\Delta x]. \\
a &= \left[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} x_i\right] : n.
\end{aligned}
\tag{6}
$$

$$2k = n \pm j \qquad j = 0 \text{ or } 1.$$

This arbitrary determination of k will be justified in a later section.

VII. Parabolic Series

The equation of the quadratic parabola is

$$y_i = a + bx_i + cx_i^2 . \tag{1}$$

Giving to x the increment $k\Delta x$, we have

$$y_{i+k} = a + b(x_i + k\Delta x) + c(x_i + k\Delta x)^2 . \tag{2}$$

Subtracting (1) from (2), we have

$$\Delta_k y_i = bk\Delta x + ck\Delta x(2x_i + k\Delta x). \tag{3}$$

Giving to x a second increment $k\Delta x$, we have

$$\Delta_k y_{i+k} = bk\Delta x + ck\Delta x\ (2x_i + 3k\Delta x) \tag{4}$$

Subtracting (3) from (4), we obtain

$$\Delta_k^2 y_i = 2ck^2\Delta x^2 \tag{5}$$

From equations (5), (3), and (1), we deduce the following approximations to parameters:

$$c_i = \Delta_k^2 Y_i : (2k^2\Delta x^2) \quad i = 0, 1, \cdots n-2k-1.$$

$$b_i = [\Delta_k Y_i - ck\Delta x\ (2x_i + k\Delta x)] : [k\Delta x], \quad i = 0, 1, \cdots n-k-1. \tag{6}$$

$$a_i = Y_i - bx_i - cx_i^2, \quad i = 0, 1, \cdots n-1$$

Taking mean values of the approximations indicated in equations (6), we have the following formulas for the determination of parameters:

$$c = \left[\sum_{i=0}^{i=n-2k-1} \Delta_k^2 Y_i\right] : [2k^2(n-2k)\Delta x^2].$$

$$b = \left[\sum_{i=0}^{i=n-k-1} \Delta_k Y_i - ck\Delta x \sum_{i=0}^{i=n-k-1} (2x_i + k\Delta x)\right] : [k(n-k)\Delta x]. \tag{7}$$

$$a = \left[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} x_i - c\sum_{i=0}^{i=n-1} x_i^2\right] : n.$$

$$3k = n \pm j, \quad j = 0, 1, \text{ or } 2$$

We shall next write the equation of the cubic parabola, which is

$$y_i = a + bx_i + cx_i^2 + dx_i^3 \tag{8}$$

Giving to x the increment $k\Delta x$, we have

$$y_{i+k} = a + b(x_i + k\Delta x) + c(x_i + k\Delta x)^2 + d(x_i + k\Delta x)^3. \tag{9}$$

Subtracting (8) from (9), we get

$$\Delta_k y_i = bk\Delta x + ck\Delta x\ (2x_i + k\Delta x) + dk\Delta x(3x_i^2 + 3x_i k\Delta x + k^2\Delta x^2). \tag{10}$$

TABLE I

Production of Gold in the United States

(in units of $100,000)

$y = 695.75 + 53.89x - 2.81x^2$

Year	Y	y	v	Year	Y	y	v
1900	792	696	96	1911	969	949	20
1901	787	747	40	1912	935	938	- 3
1902	800	792	8	1913	883	921	-38
1903	736	832	-96	1914	945	899	46
1904	805	866	-61	1915	1010	872	138
1905	882	895	-13	1916	926	839	87
1906	944	918	26	1917	838	800	38
1907	904	935	-31	1918	686	755	-69
1908	946	947	- 1	1919	603	705	-102
1909	997	953	44	1920	512	650	-138
1910	963	954	9				
Σ					17863	17863	0

Mean error of estimate 52.6

Again giving to x the increment $k\Delta x$, we have

$$(11)\quad \Delta_k y_{i+k} = bk\Delta x + ck\Delta x\,(2x_i + 3k\Delta x) + dk\Delta x\,[3(x_i + k\Delta x)^2 + 3k\Delta x\,(x_i + k\Delta x) + k^2\Delta x^2].$$

Subtracting (10) from (11), we obtain

$$(12)\quad \Delta_k^2 y_i = 2ck^2\Delta x^2 + 6dk^2\Delta x^2(x_i + k\Delta x).$$

Once more increasing x by $k\Delta x$, we have

$$(13)\quad \Delta_k^2 y_{i+k} = 2ck^2\Delta x^2 + 6dk^2\Delta x^2(x_i + 2k\Delta x).$$

Subtracting (12) from (13), we obtain finally

$$(14)\quad \Delta_k^3 y_i = 6dk^3\Delta x^3.$$

From equations (14), (12), (10), and (8), we deduce the follow-

ing parametric approximations:

$$d_i = \Delta_k^3 Y_i : (6k^3\Delta x^3), \quad i = 0, 1, \cdots n-3k-1.$$

$$c_i = [\Delta_k^2 Y_i - 6dk^2\Delta x^2(x_i + k\Delta x)] : [2k^2\Delta x^2], \quad i = 0, 1, \cdots n-2k-1.$$

$$(15) \quad b_i = [\Delta_k Y_i - ck\Delta x(2x_i + k\Delta x) - dk\Delta x(3x_i^2 + 3x_i k\Delta x + k^2\Delta x^2)] : [k\Delta x], \quad i = 0, 1, \cdots n-k-1.$$

$$a_i = Y_i - bx_i - cx_i^2 - dx_i^3, \quad i = 0, 1, \cdots n-1.$$

Taking mean values of the approximations indicated in equations (15), we have the following formulas for the determination of parameters:

$$d = \left[\sum_{i=0}^{i=n-3k-1} \Delta_k^3 Y_i\right] : [6k^3(n-3k)\Delta x^3].$$

$$c = \left[\sum_{i=0}^{i=n-2k-1} \Delta_k^2 Y_i - 6dk^2\Delta x^2 \sum_{i=0}^{i=n-2k-1}(x_i - k\Delta x)\right] : [2k^2(n-2k)\Delta x^2].$$

$$(16) \quad b = \left[\sum_{i=0}^{i=n-k-1} \Delta_k Y_i - ck\Delta x \sum_{i=0}^{i=n-k-1}(2x_i + k\Delta x) - dk\Delta x \sum_{i=0}^{i=n-k-1}(3x_i^2 + 3x_i - k\Delta x + k^2\Delta x^2)\right] : [k(n-k)\Delta x].$$

$$a = \left[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} x_i - c\sum_{i=0}^{i=n-1} x_i^2 - d\sum_{i=0}^{i=n-1} x_i^3\right] : n.$$

$$4k = n \pm j, \qquad j = 0, 1, 2, \text{ or } 3.$$

VIII. Hyperbolic Series

Let us write the hyperbolic series

$$(1) \quad y_i = a + bx_i + c : (x_i + 1)$$

Giving the increment $k\Delta x$ to x, we have

$$(2) \quad y_{i+k} = a + b(x_i + k\Delta x) + c : (x_i + 1 + k\Delta x)$$

By subtraction, we have

$$(3) \quad \Delta_k y_i = bk\Delta x - ck\Delta x : ((x_i + 1)(x_i + 1 + k\Delta x))$$

Giving a second increment $k\Delta x$ to x, we obtain

$$\Delta_k y_{i+k} = bk\Delta x - ck\Delta x : \left((x_i+1+k\Delta x)\ (x_i+1+2k\Delta x)\right). \tag{4}$$

By subtraction, we obtain

$$\Delta_k^2 y_i = 2ck^2\Delta x^2 : \left((x_i+1)\ (x_i+1+2k\Delta x)\right). \tag{5}$$

Making the substitutions $x_i' = (x_i+1)(x_i+1+k\Delta x)$, and $x_i'' = (x_i+1+2k\Delta x)$, we have, from equations (5), (3), and (1), the following parametric approximations:

$$\begin{aligned}
c_i &= (x_i''\ \Delta_k^2 Y_i) : (2k^2\Delta x^2), \quad i = 0, 1, \cdots n-2k-1. \\
b_i &= (\Delta_k Y_i + ck\Delta x : x_i') : (k\Delta x), \quad i = 0, 1, \cdots n-k-1. \\
a_i &= Y_i - bx_i - c : (x_i+1), \quad i = 0, 1, \cdots n-1.
\end{aligned} \tag{6}$$

By taking mean values of the approximations indicated in equations (6), we have the following formulas for determining parameters:

$$\begin{aligned}
c &= \left[\sum_{i=0}^{i=n-2k-1} x_i''\Delta_k^2 Y_i\right] : \left[2k^2(n-2k)\Delta x^2\right]. \\
b &= \left[\sum_{i=0}^{i=n-k-1} \Delta_k Y_i + ck\Delta x\left(\sum_{i=0}^{i=n-k-1} 1/x_i'\right)\right] : \left[k(n-k)\Delta x\right]. \\
a &= \left[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} x_i - c\sum_{i=0}^{i=n-1}(x_i+1)^{-1}\right] : n.
\end{aligned} \tag{7}$$

$$3k = n \pm j, \quad j = 0, 1, \text{ or } 2.$$

If the coefficient of x in (1) is zero, we have simply

$$y_i = a + b : (x_i+1). \tag{8}$$

Formulas (7) now reduce to

$$\begin{aligned}
b &= -\left[\sum_{i=0}^{i=n-k-1} x_i'\Delta_k Y_i\right] : \left[k(n-k)\Delta x\right]. \\
a &= \left[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1}(x_i-1)^{-1}\right] : n.
\end{aligned} \tag{9}$$

$$2k = n \pm j \quad j = 0 \text{ or } 1.$$

It is evident that a term in x^2 or x^3 could be added to equation (1) and a solution be obtained by a direct extension of the general method of analysis applied to equation (1).

IX. Logarithmic Series

Let us write the logarithmic equation

$$(1) \qquad y_i = a + b\,x_i + c \cdot \log\,(x_i + 1).$$

Giving x the increment $k\Delta x$, we have

$$(2) \qquad y_{i+k} = a + b\,(x_i + k\Delta x) + c \cdot \log\,(x_i + 1 + k\Delta x).$$

Subtracting (1) from (2), we get

$$(3) \qquad \Delta_k y_i = bk\Delta x + c\Delta_k \log\,(x_i + 1).$$

Giving to x a second increment $k\Delta x$, we get

$$(4) \qquad \Delta_k y_{i+k} = bk\Delta x + c\Delta_k \log(x_i + 1 + k\Delta x).$$

TABLE II

Deaths from Typhoid Fever in Greater City of New York

(Number of deaths per 1,000,000 inhabitants)

$y = 143.899 + 8.695\,x - 206.652\,\Gamma x$

Year	Y	y	v	Year	Y	y	v
1911	111.7	152.6	−40.9	1919	21.8	25.6	− 3.8
1912	100.5	99.3	1.2	1920	24.2	24.8	− 0.6
1913	72.0	71.7	0.3	1921	21.3	25.1	− 3.8
1914	65.0	54.7	10.3	1922	22.1	26.0	− 3.9
1915	63.5	43.4	20.1	1923	23.6	27.4	− 3.8
1916	40.6	35.8	4.8	1924	30.0	29.6	0.4
1917	42.4	30.7	11.7	1925	31.9	32.1	− 0.2
1918	35.6	27.4	8.2				
Σ					706.2	706.2	0.0

Mean error of estimate 7.6

Subtracting (3) from (4), we obtain

$$\Delta_k^2 y_i = c\,\Delta_k^2 \log(x_i+1). \tag{5}$$

From equations (5), (3), and (1), we have the following approximations to parameters:

$$c_i = \Delta_k^2 Y_i : \Delta_k^2 \log(x_i+1), \quad i=0, 1, \cdots n-2k-1.$$

$$b_i = \Delta_k Y_i - c\Delta_k \log(x_i+1), \quad i=0, 1, \cdots n-k-1. \tag{6}$$

$$a_i = Y_i - bx_i - c\cdot\log(x_i+1), \quad i=0, 1, \cdots n-1.$$

Taking mean values of the approximations indicated in equations (6), we have the following formulas for the determination of parameters:

$$c = \Big[\sum_{i=0}^{i=n-2k-1}(\Delta_k^2 Y_i : \Delta_k^2 \log(x_i+1))\Big] : [n-2k].$$

$$b = \Big[\sum_{i=0}^{i=n-k-1}\Delta_k Y_i - c\sum_{i=0}^{i=n-k-1}\Delta_k \log(x_i+1)\Big] : [k(n-k)\Delta x].$$

$$a = \Big[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} x_i - c\sum_{i=0}^{i=n-1}\log(x_i+1)\Big] : n. \tag{7}$$

$$3k = n\pm j, \quad j=0, 1, \text{ or } 2$$

If the coefficient of x in equation (1) is zero, we have

$$y_i = a + b\cdot\log(x_i+1). \tag{8}$$

Formulas (7) then reduce to

$$b = \Big[\sum_{i=0}^{i=n-k-1}(\Delta_k Y_i : \Delta_k \log(x_i+1))\Big] : [n-k].$$

$$a = \Big[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1}\log(x_i+1)\Big] : n. \tag{9}$$

$$2k = n\pm j, \quad j=0 \text{ or } 1$$

X. General Polynomial Series

The solutions of polynomials presented in the preceding sections, while best for the series considered, are too specialized in mode of analysis for application to polynomials generally. We shall now develop a solution which is applicable to any polynomial in $z=f(x)$, $f(x)$ being a function of x whose value is known, as for example, x^{-1}, $\log x$, $\tan x$, etc.

We write

(1) $y_i = d+cz_i+bz_i^2+az_i^3$.

Giving to z_i the increment $\Delta_k z_i$, we have

(2) $y_{i+k} = d+cz_{i+k}+bz_{i+k}^2+az_{i+k}^3$.

Subtracting (1) from (2), we get

(3) $\Delta_k y_i = c\Delta_k z_i + b\Delta_k z_i^2 + a\Delta_k z_i^3$.

This is the i^{th} equation $\Delta_k y$. We write the $i+k^{th}$ equation as

(4) $\Delta_k y_{i+k} = c\Delta_k z_{i+k} + b\Delta_k z_{i+k}^2 + a\Delta_k z_{i+k}^3$.

Multiplying (3) by $\Delta_k z_{i+k}$ and (4) by $\Delta_k z_i$ and then subtracting (4) from (3), we obtain

(5) $\Delta_k' y_i = b\Delta_k' z_i^2 + a\Delta_k' z_i^3$,

where $\Delta_k' y_i = \Delta_k y_i \cdot \Delta_k z_{i+k} - \Delta_k y_{i+k} \cdot \Delta_k z_i$; $\Delta_k' z_i^2 = \Delta_k z_i^2 \cdot \Delta_k z_{i+k} - \Delta_k z_{i+k}^2 \cdot \Delta_k z_i$;

and $\Delta_k' z_i^3 = \Delta_k z_{i+}^3 \cdot \Delta_k z_{i+k} - \Delta_k z_{i+k}^3 \cdot \Delta_k z_i$.

In (5), we have the i^{th} equation $\Delta_k' y$. We write the $i+k^{th}$ equation as

(6) $\Delta_k' y_{i+k} = b\Delta_k' z_{i+k}^2 + a\Delta_k' z_{i+k}^3$.

Multiplying (5) by $\Delta_k' z_{i+k}^2$ and (6) by $\Delta_k' z_i^2$ and subtract-

ing the latter result from the former, we obtain

$$\Delta''_k y_i = a\,\Delta''_k z_i^3 , \tag{7}$$

where $\Delta''_k y_i = \Delta'_k y_i \cdot \Delta'_k z_{i+k}^2 - \Delta'_k y_{i+k} \cdot \Delta'_k z_i^2$;

and $\Delta''_k z_i^3 = \Delta'_k z_i^3 \cdot \Delta'_k z_{i+k}^2 - \Delta'_k z_{i+k}^3 \cdot \Delta'_k z_i^2$.

From equations (7), (5), (3), and (1), we are now able to write the following parametric approximations:

$$\begin{aligned}
a_i &= \left[\Delta''_k Y_i\right] : \left[\Delta''_k z_i^3\right], \quad i=0,\ 1,\ \cdots\ n-3k-1. \\
b_i &= \left[\Delta'_k Y_i - a\Delta'_k z_i^3\right] : \left[\Delta'_k z_i^2\right], \quad i=0,\ 1,\ \cdots\ n-2k-1. \\
c_i &= \left[\Delta_k Y_i - b\Delta_k z_i^2 - c\Delta_k z_i^3\right] : \left[\Delta_k z_i\right], \quad i=0,\ 1,\cdots n-k-1. \\
d_i &= Y_i - cz_i - bz_i^2 - az_i^3, \quad i=0,\ 1,\ \cdots\ n-1
\end{aligned} \tag{8}$$

When $z_0 = \pm\infty$, i takes the values 1, 2, ... $n-rk-1$, r being the number of reductions essential to the approximation.

The mean values of equations (8) give the following determinations:

$$\begin{aligned}
a &= \left[a_0 + a_1 + \cdots + a_{n-3k-1}\right] : \left[n-3k\right]. \\
b &= \left[b_0 + b_1 + \cdots + b_{n-2k-1}\right] : \left[n-2k\right]. \\
c &= \left[c_0 + c_1 + \cdots + c_{n-k-1}\right] : \left[n-k\right]. \\
d &= \left[\sum_{i=0}^{i=n-1} Y_i - c\sum_{i=0}^{i=n-1} z_i - b\sum_{i=0}^{i=n-1} z_i^2 - a\sum_{i=0}^{i=n-1} z_i^3\right] : n .
\end{aligned} \tag{9}$$

If equation (1) is simplified to

$$y_i = c + bz_i + az_i^2 , \tag{10}$$

the parametric approximations become the following:

$$a_i = [\Delta'_k Y_i] : [\Delta'_k z_i^2], \quad i = 0, 1, \cdots n-2k-1.$$

$$(11) \qquad b_i = [\Delta_k Y_i - a\Delta_k z_i^2] : [\Delta_k z_i], \quad i = 0, 1, \cdots n-k-1.$$

$$c_i = Y_i - bx_i - az_i^2, \quad i = 0, 1, \cdots n-1.$$

The mean values of these approximations give the parameters sought.

XI. Exponential Series

We shall now write the equation of the exponential series

$$(1) \qquad y_i = d + cx_i + be^{ax_i}.$$

Giving to x the increment $k\Delta x$, we have

$$(2) \qquad y_{i+k} = d + c(x_i + k\Delta x) + be^{a(x_i + k\Delta x)}$$

Subtracting (1) from (2), we get

$$(3) \qquad \Delta_k y_i = ck\Delta x + bhe^{ax_i},$$

where $h = e^{ak\Delta x} - 1$.

Giving to x a second increment $k\Delta x$, we obtain

$$(4) \qquad \Delta_k y_{i+k} = ck\Delta x + bhe^{a(x_i + k\Delta x)}.$$

Subtracting (3) from (4), we obtain

$$(5) \qquad \Delta_k^2 y_i = bh^2 e^{ax_i}$$

Taking logarithms, we have

$$(6) \qquad \log \Delta_k^2 y_i = \log(bh^2) + ax_i$$

Again giving x the increment $k\Delta x$, we have

$$(7) \qquad \log \Delta_k^2 y_{i+k} = \log(bh^2) + a(x_i + k\Delta x).$$

Subtracting (6) from (7), we obtain

$$\Delta_k \log \Delta_k^2 y_i = a k \Delta x \tag{8}$$

From equations (8), (5), (3), and (1), we form the following parametric approximations:

$$\begin{aligned}
&a_i=(\Delta_k \log \Delta_k^2 Y_i) : (k\Delta x), \quad i=0, 1, \cdots n-3k-1.\\
&b_i=(\Delta_k^2 Y_i) : (h^2 e^{ax_i}), \quad i=0, 1, \cdots n-2k-1.\\
&c_i=(\Delta_k Y_i - bhe^{ax_i}) : (k\Delta x), \quad i=0, 1, \cdots n-k-1.\\
&d_i=Y_i - cx_i - be^{ax_i}, \quad i=0, 1, \cdots n-1.
\end{aligned} \tag{9}$$

Taking mean values of the approximations indicated in equations (9), we have the following formulas for determining parameters:

$$\begin{aligned}
&a=\Big[\sum_{i=0}^{i=n-3k-1} \Delta_k \log \Delta_k^2 Y_i\Big] : \big[k(n-3k)\Delta x\big].\\
&b=\Big[\sum_{i=0}^{i=n-2k-1} (\Delta_k^2 Y_i : e^{ax_i})\Big] : \big[h^2(n-2k)\big].\\
&c=\Big[\sum_{i=0}^{i=n-k-1} \Delta_k Y_i - bh\sum_{i=0}^{i=n-k-1} e^{ax_i}\Big] : \big[k(n-k)\Delta x\big].\\
&d=\Big[\sum_{i=0}^{i=n-1} Y_i - c\sum_{i=0}^{i=n-1} x_i - b\sum_{i=0}^{i=n-1} e^{ax_i}\Big] : n
\end{aligned} \tag{10}$$

If, in equation (1), the coefficient of x is zero, we have

$$y_i = c + be^{ax_i} \tag{11}$$

and the formulas for determining parameters become

$$\begin{aligned}
&a=\Big[\sum_{i=0}^{i=n-2k-1} \Delta_k \log \Delta_k Y_i\Big] : \big[k(n-2k)\Delta x\big].\\
&b=\Big[\sum_{i=0}^{i=n-k-1} (\Delta_k Y_i : e^{ax_i})\Big] : \big[h(n-k)\big].\\
&c=\Big[\sum_{i=0}^{i=n-1} Y_i - b\sum_{i=0}^{i=n-1} e^{ax_i}\Big] : n.
\end{aligned} \tag{12}$$

XII. Logistic Series

Let us write the equation of the logistic series

$$(1) \qquad y_i = [d + cx_i] : [1 + be^{ax_i}].$$

Multiplying by the denominator on the right and transposing, we get

$$(2) \qquad y_i = d + cx_i - by_i e^{ax_i}.$$

Giving to x the increment $k\Delta x$, we have

$$(3) \qquad y_{i+k} = d + cx_{i+k} - by_{i+k} e^{ax_{i+k}}$$

Subtracting (2) from (3), we get

$$(4) \qquad \Delta_k y_i = ck\Delta x - by_{i+k} e^{ax_{i+k}} + by_i e^{ax_i}.$$

Again giving to x the increment $k\Delta x$, we obtain

$$(5) \qquad \Delta_k y_{i+k} = ck\Delta x - by_{i+2k} e^{ax_{i+2k}} + by_{i+k} e^{ax_{i+k}}$$

Subtracting (4) from (5), we have

$$(6) \qquad \begin{aligned} \Delta_k^2 y_i &= -by_{i+2k} e^{ax_{i+2k}} + 2by_{i+k} e^{ax_{i+k}} - by_i e^{ax_i} \\ &= -be^{ax_i}(y_{i+2k} e^{a2k\Delta x} - 2y_{i+k} e^{ak\Delta x} + y_i). \end{aligned}$$

If, in (6), we give to x the increment $k\Delta x$, we get

$$(7) \qquad \begin{aligned} \Delta_k^2 y_{i+k} &= -by_{i+3k} e^{ax_{i+3k}} + 2by_{i+2k} e^{ax_{i+2k}} - by_{i+k} e^{ax_{i+k}} \\ &= -be^{ax_i}(y_{i+3k} e^{a3k\Delta x} - 2y_{i+2k} e^{a2k\Delta x} + y_{i+k} e^{ak\Delta x}) \end{aligned}$$

On dividing (7) by (6) and multiplying the quotient on the right by the parenthetical expression of (3), we have

$$(8) \qquad \begin{aligned} & p_k \Delta_k^2 y_i (y_{i+2k} e^{a2k\Delta x} - 2y_{i+k} e^{ak\Delta x} y_i) \\ &= y_{i+3k} e^{a3k\Delta x} - 2y_{i+2k} e^{a2k\Delta x} + y_{i+k} e^{ak\Delta x}. \end{aligned}$$

TABLE III

Population of Ohio

U. S. Census Count Interpolated to January 1; Unit, 1.000 persons

$$y = 91.8\ (1 + e^{4.026607 + 29.0267 \log \sin x})$$

Year	Y	y	v	Year	Y	y	v
1800	41.2	91.8	+ 50.6	1870	2651.8	2751.3	+ 99.5
1810	219.7	319.8	+100.1	1880	3175.9	3237.2	+ 61.3
1820	560.3	639.3	+ 79.0	1890	3652.7	3738.1	+ 85.4
1830	922.9	1005.6	+ 82.7	1900	4137.4	4252.5	+115.1
1840	1495.3	1405.6	− 89.7	1910	4749.3	4778.9	+ 29.6
1850	1961.3	1832.8	−128.5	1920	5759.4	5316.0	−443.4
1860	2324.5	2282.3	− 42.2				
Σ					31651.7	31651.2	− 0.5

Mean error of estimate 108.2

Predicted Population

Year	1930	1940	1950	1960	1970	1980
Population	6306	6862	7425	7996	8573	9156

Simplifying (8), we have

$$y_{i+3k}\, e^{3\kappa\Delta x} - y_{i+2k}\,(2 + p_k \Delta_k^2 y_i)\, e^{2\kappa\Delta x} + y_{i+k}\,(1 + 2p_k\Delta_k^2 y_i) e^{\kappa\Delta x} - y_i\, p_k \Delta_k^2 y_i = 0 \tag{9}$$

Equation (9) is evidently cubic in $e^{\kappa\Delta x}$, and its roots are to be found by conventional methods, care being taken to select the root which will give the parametric approximation most consistent with the hypotheses under which the function is being fitted.

From equations (9), (6), (4), and (1), we are able to form the

following approximations to parameters:

$$a_i=[\log e^{\theta k\Delta x}]:[k\Delta x] \quad , \quad i=0,\ 1,\ \cdots n-3k-1 .$$

$$b_i=[\Delta_k^2 Y_i]:[-Y_{i+2k}e^{\theta x_{i+2k}}+2Y_{i+k}e^{\theta x_{i+k}}-Y_i e^{\theta x_i}],\ i=0,1,\cdots n-2k-1.$$

(10)

$$c_i=[\Delta_k Y_i+bY_{i+k}e^{\theta x_{i+k}}-bY_i e^{\theta x_i}]:[k\Delta x],\ i=0,\ 1,\ \cdots n-k-1.$$

$$d_i=Y_i+bY_i e^{\theta x_i}-cx_i, \qquad i=0,\ 1,\ \cdots n-1 .$$

The mean values of the indicated approximations give the best values of the parameters sought.

If, in equation (1), the coefficient of x is zero, we have the Verhulst logistic,

(11) $y_i=c:(1+be^{\theta x_i})$.

The solution of this equation by the method of analysis applied to equation (1) leads eventually to the following:

(12) $y_{i+2k}e^{\theta 2k\Delta x}-y_{i+k}(1+p_k\Delta_k y_i)e^{\theta k\Delta x}+y_i p_k\Delta_k y_i=0.$

which is evidently quadratic in $e^{\theta k\Delta x}$.

The parametric approximations take the form

$$a_i=[\log e^{\theta k\Delta x}]:[k\Delta x] \quad , \quad i=0,\ 1,\ \cdots n-2k-1.$$

(13) $$b_i=-[\Delta_k Y_i]:[Y_{i+k}e^{\theta x_{i+k}}-Y_i e^{\theta x_i}],\ i=0,\ 1,\ \cdots n-k-1.$$

$$c_i=Y_i+bY_i e^{\theta x_i} \quad , \quad i=0,\ 1,\ \cdots n-1 .$$

The mean values of these approximations give the values of parameters.

The Verhulst logistic may also be solved by applying formulas (12), section XI, to the ordinates $1/Y_i$, the solution being for $1/c$,

b/c, and a. Similarly, a solution for the serial equation

$$y_i = d : [1 + cx_i + be^{ax_i}] \tag{14}$$

may be had by applying formulas (10), section XI, to the ordinates $1/Y_i$, the solution giving the values of $1/d$, c/d, b/d, and a.

To solve for certain other series which are of interest, we write

$$y_i = m e^{be^{ax_i}} \tag{15}$$

The solution is obtained by applying formulas (12), section XI, to the ordinates log Y_i.

We have also

$$y_i = y_0 (1 + e^{b + ax_i}) = y_0 + Be^{ax_i} \tag{16}$$

where $B = y_0 e^b$, and the argument $x = \phi(x)$ is chosen so that $ax_0 = -\infty$. This condition is met when $f(x)$ takes the form $1/x$, log x, cot x, log sin x, etc., the sign of a being sometimes plus and sometimes minus.

From (16), by forming the function $u_i = y_i - y_0$, we get

$$\log u_i = B + ax_i . \tag{17}$$

On taking a first difference of rank k, this becomes

$$\Delta_k \log u_i = a \Delta_k x_i . \tag{18}$$

From equations (18), (17), and (16), we deduce the following parametric approximations:

$$a_i = \Delta_k \log U_i : \Delta_k x_i , \qquad i = 1, 2, \cdots n-k-1 .$$

$$\log B_i = \log U_i - ax_i , \qquad i = 1, 2, \cdots n-1 . \tag{19}$$

$$Y_{0i} = Y_i - Be^{ax_i} , \qquad i = 0, 1, \cdots n-1 .$$

The mean values of a_i, B_i, and y_0 give the values of the parameters sought.

If a term in x is added to the exponent of equation (16), we have

$$(20) \qquad y_i = y_0 \left(1+e^{c+bx_i+gx_i^2}\right).$$

The solution for these is obtained by carrying the analysis applied to equation (16) to second order differences and applying formulas (7), section IX, to the ordinates $\log u_i$.

If equation (20) is rewritten as

$$(21) \qquad y = y_0 \left(1+e^{c+bx_i+gx_i^2}\right),$$

The solution is obtained by applying formulas (11), section X, to the ordinates $\log U_i$. This solution holds, it will be noted, only when the signs of b and a are such that $bz_0+gz_0^2=-\infty$

XIII. Determination of the Rank of Differences

In the writing of formulas for the determination of parameters, the rank of differences has been fixed in a purely arbitrary manner. We shall now give a rational justification for the rank assigned.

In what follows, we shall speak of the process by which one of the parameters is eliminated from the equation of the function $y=f(x)$ as a *reduction*; and the definition shall be understood to hold whether the reduction takes place through a simple difference Δy, a logarithmic difference $\Delta \log y$, a product difference $\Delta' y$, or a ratio py, the rank of the reduction being the same as the rank of the difference or ratio involved. In this, we interpret $\Delta_k^s \log \Delta_k^r y$ and $\Delta_k^s p_k^r y$ as determining reductions of order $s+r$

The process by which a first parameter is eliminated we shall call a first reduction; that by which a second is eliminated, a second reduction; and so on to the r th reduction. The process by which the last but one of the parameters of the original function is eliminated we shall term the *ultimate reduction*; and the parameter definied by the ultimate reduction we shall term the *ultimate parameter*.

Now, a little thought or experimentation will quickly reveal that, for any ultimate parametric approximation, the value of the approx-

imation will vary with the rank of the reduction from which it results; for, regarding a parameter as a statistical characteristic of a series of observations, when the rank of a parametric approximation is at a minimum, or when $k = 1$, the given approximation, viewed as a single instance of a number of possible approximations, is least characteristic of the complete series; and when the rank of the given approximation is at a maximum, or when $k = n - r$, the approximation, again viewed as a single instance of a number of possible approximations, is most characteristic of the complete series.

All this, of course, assumes, as we have always done in writing the equations of parametric approximations, that approximations are written in terms of the observational ordinates Y_i; for, if approximations are written in terms of the functional ordinates y_i, the value of the parameter is independent of the rank of the reduction, a fact which follows from the manner of deriving the equation defining the ultimate parameter.

Since, then, the value and representative character of an ultimate parameter varies with the rank of the reduction by which it is defined, we may, when the ultimate reduction is of the first order, express the weight of an approximation by the relation

$$(1) \qquad w_k\ (p_i) = k \cdot$$

Here $w_k\ p_i$ is used as the arbitrary symbol for the weight of a parametric approximation defined of first order and rank k.

Suppose, now, that the given ultimate parameter is arrived at by two reductions, the first of rank k and the second of rank h. Clearly, the value of the approximation will, in this case, vary with h as well as k. Under these conditions, the weight of the approximation is expressed by the relation

$$(2) \qquad w^2_{k,h}\ (p_i) = k \cdot h$$

Here, the symbol $w^2_{k,h}\ (p_i)$ denotes the weight of a parametric approximation involving a second reduction and the ranks k and h.

Similarly, we have

$$(3) \qquad w^3_{k,h,f}\ (p_i) = k \cdot h \cdot f\ .$$

Evidently, by a direct extension of our method of induction, we arrive at the general relation

$$(4) \qquad w^{r}_{k^1, k^2, \cdots k^r}(p_i) = k_1 \cdot k_2 \cdot \cdots \cdot k_r .$$

In the derivation of all formulas, it has been assumed that k is constant for all reductions; hence, equations (2), (3), and (4) become

$$(5) \qquad w^{2}_{k}(p_i) = k \cdot k = k^2$$

$$(6) \qquad w^{3}_{k}(p_i) = k \cdot k \cdot k = k^3$$

$$(7) \qquad w^{r}_{k}(p_i) = k^r$$

Giving verbal expression to the relation (7), we say that the weight of a parametric approximation involving a reduction of the r th order and k th rank is equal to the r th power of k.

We have already shown, section V, that the number of differences of order r and rank k which can be formed from n observations is $n - rk$; likewise, the number of parametric approximations which can be formed when reductions are of the r th order, is $n - rk$; and, since the reliability of a parameter as determined from a formula must vary with the number as well as the weight of the several approximations, we may write the following equation, conditioning the reliability of the ultimate parameter p:

$$(8) \qquad \psi(p) = k^r(n - rk).$$

Regarding k as a continuous variable, we may obtain the condition for $\psi(p) =$ a maximum by differentiating ψ with respect to k, thus:

$$(9) \qquad D_k \psi(p) = nrk^{r-1} - r(r+1)k^r .$$

Setting (9) equal to zero and solving, we have

$$(10) \qquad k = n : (r-1) ,$$

That $\psi(p)$ is a maximum and not a minimum when k is determined from equation (10) is shown by taking the second derivative

of ψ and substituting for k, thus:

$$(11)\qquad \begin{aligned} D_k^2\psi(\rho) &= r(r-1)nk^{r-2} - r^2(r+1)k^{r-1} = \\ & rk^{r-2}\left((r-1)n - r(r+1)n : (r+1)\right) = -rnk^{r-2}, \end{aligned}$$

which is negative, since r, n, and k are positive.

Equation (10) may give fractional values of k; but, in practice, k is always integral; hence, we write (10) in the form

$$(12)\qquad k = (n \pm j) : (r+1) .$$

This is the relation from which we have determined the value of k in the writing of formulas.

We may now formulate the following rule for the determination of the rank of differences: *When the equation defining an ultimate parameter involves a reduction of order r and rank k, the value of k is to be obtained from the relation $k = (n \pm j) : (r+1)$, j being assigned the smallest integral value that will make $n \pm j$ an exact multiple of k.* In case $n+j = n-j$, that value of k is taken which gives the highest value for $\psi(\rho)$ when k is substituted in equation (8).

XIV. Numerical Computations

In carrying through the numerical computations prescribed by formulas developed in this memoir, the following abridgments are useful in the summation of differences:

$$(1a)\qquad \sum_{i=0}^{i=n-k-1} \Delta_k Y_i = \sum_{i=0}^{i=n-k-1} (Y_{i+k} - Y_i)$$

$$(1b)\qquad = \sum_{i=0}^{i=n-k-1} Y_{i+k} - \sum_{i=0}^{i=n-k-1} Y_i$$

$$(1c)\qquad = \sum_{i=k}^{i=n-1} Y_i - \sum_{i=0}^{i=n-k-1} Y_i$$

$$(1d)\qquad = \sum_{i=n-k}^{i=n-1} Y_i - \sum_{i=0}^{i=k-1} Y_i ,$$

$$(2a)\qquad \sum_{i=0}^{i=n-2k-1} \Delta_k^2 Y_i = \sum_{i=0}^{i=n-2k-1} (Y_{i+2k} - 2Y_{i+k} + Y_i)$$

$$(2b)\qquad = \sum_{i=0}^{i=n-2k-1} Y_{i+2k} - 2\sum_{i=0}^{i=n-2k-1} Y_{i+k} + \sum_{i=0}^{i=n-2k-1} Y_i$$

$$(2c)\qquad = \sum_{i=2k}^{i=n-1} Y_i - 2\sum_{i=k}^{i=n-k-1} Y_i + \sum_{i=0}^{i=n-2k-1} Y_i$$

(2d) $$= \sum_{i=n-k}^{i=n-1} Y_i - 2\sum_{i=n-2k}^{i=n-k-1} Y_i + \sum_{i=0}^{i=k-1} Y_i \ .$$

(3a) $$\sum_{i=0}^{i=n-3k-1} \Delta_k^3 Y_i = \sum_{i=0}^{i=n-3k-1} Y_i (Y_{i+3k} - 3Y_{i+2k} + 3_{i+k} + Y_i)$$

(3b) $$= \sum_{i=0}^{i=n-3k-1} Y_{i+3k} - 3\sum_{i=0}^{i=n-3k-1} Y_{i+2k} + 3\sum_{i=0}^{i=n-3k-1} Y_{i+k} - \sum_{i=0}^{i=n-3k-1} Y_i$$

(3c) $$= \sum_{i=3k}^{i=n-1} Y_i - 3\sum_{i=2k}^{i=n-k-1} Y_i + 3\sum_{i=k}^{i=n-2k-1} Y_i - \sum_{i=0}^{i=n-3k-1} Y_i$$

(3d) $$= \sum_{i=n-k}^{i=n-1} Y_i - 3\sum_{i=n-2k}^{i=3k-1} Y_i + 3\sum_{i=n-3k}^{i=2k-1} Y_i - \sum_{i=0}^{i=k-1} Y_i \ .$$

These relations evidently apply quite generally to the summation of differences. They may also be used to check the accuracy of differences formed. When j is positive in the relation $(r+1)k = n \pm j$, equations (c) are most convenient; when j is negative, equations (d) are most convenient.

A useful check on the product difference Δ' employed in section X is obtained as follows:

(4) $\Delta_k Y_i + \Delta_k z_i + \Delta_k z_i^2 + \Delta_k z_i^3 = S_i$.

(5) $\Delta_k Y_{i+k} + \Delta_k z_{i+k} + \Delta_k z_{i+k}^2 + \Delta_k z_{i+k}^3 = S_{i+k}$.

Multiplying (4) by $\Delta_k z_{i+k}$ and (5) by $\Delta_k z_i$, we have

(6) $\Delta_k Y_i \cdot \Delta_k z_{i+k} + \Delta_k z_i \cdot \Delta_k z_{i+k} + \Delta_k z_i^2 \cdot \Delta_k z_{i+k} + \Delta_k z_i^3 \cdot \Delta_k z_{i+k} = S_i \Delta_k z_{i+k}$.

(7) $\Delta_k Y_{i+k} \cdot \Delta_k z_i + \Delta_k z_{i+k} \cdot \Delta_k z_i + \Delta_k z_{i+k}^2 \cdot \Delta_k z_i + \Delta_k z_{i+k}^3 \cdot \Delta_k z_i = S_{i+k} \cdot \Delta_k z_i$.

Subtracting (6) from (7), we get

(8) $\Delta_k' Y_i + \Delta_k' z_i^2 + \Delta_k' z_i^3 = \Delta_k' S_i$.

Evidently, similar relations hold for the product differences Δ'', etc.

Another check that is constantly useful in the computations is the well known relation $\Sigma a f_i = a \Sigma f_i$.

TABLE IV

Auxiliary Functions Computed in Fitting to the Ohio Population

U	$\log U$	x	$\log \sin x$	$a \log \sin x$	B_i	$B10^{ax}$
0.0	$-\infty$	0	$-\infty$	$-\infty$		
178.5	2.25164	1	− 1.75814	− 2.22212	29769	228.0
519.1	2.71525	2	− 1.45718	− 1.84173	36056	547.5
881.7	2.94532	3	− 1.28120	− 1.61931	36697	913.8
1454.1	3.16259	4	− 1.15642	− 1.46160	42091	1313.8
1920.1	3.28332	5	− 1.05970	− 1.33935	41944	1741.0
2283.3	3.35856	6	− 0.98077	− 1.23960	39642	2190.5
2610.6	3.41674	7	− 0.91411	− 1.15534	37332	2659.5
3134.7	3.49620	8	− 0.85644	− 1.08246	37902	3145.4
3611.5	3.55769	9	− 0.80567	− 1.01829	37669	3646.3
4096.2	3.61238	10	− 0.76033	− 0.96097	37441	4160.7
4708.1	3.67285	11	− 0.71940	− 0.90924	38202	4687.1
5718.2	3.75726	12	− 0.68212	− 0.86212	41627	5224.2
31116.1	39.22980		−12.43148	−15.71213	456372	30457.8

$\Delta_6 \log U$	$\Delta_6 \log \sin x$	a_i
1.16510	0.84403	1.3805
0.78095	0.60074	1.3000
0.61237	0.47553	1.2877
0.44979	0.39609	1.1356
0.38953	0.34030	1.1447
0.39870	0.29865	1.3350
3.79644	2.95534	7.5835

$a = \frac{1}{6}\Sigma a_i = 1.2639$; $B = \frac{1}{12}\Sigma B_i = 38031$;

$y_0 = \frac{1}{13}(\Sigma Y - \Sigma B 10^{ax}) = 91.8$; $b = \log B - \log y_0 = 2.617301$.

In computing the ordinates

$$y_i = a + bx_i + cx_i^2 \tag{9}$$

the following formulas are useful:

$$\Delta y_o = (b + c\Delta x)\Delta x . \tag{10}$$

$$\Delta y_{i+1} = \Delta y_i + 2c\Delta x . \tag{11}$$

$$y_o = a . \tag{12}$$

$$y_{i+1} = y_i + \Delta y_i . \tag{13}$$

The formulas for Σx, Σx^2, and Σx^3 are to be found in any standard reference work on statistical computations.

As illustrations of the quantities to be obtained in actual computations, we give, in Table IV, the auxiliaries computed in fitting the curve $y = y_o(1 + e^{b+a \cdot \log \sin x})$ to the population of Ohio.

XV. Critical Review

We have now presented, at some length, the technique of fitting curves by the method of differences. The term, "method of differences" is doubtless sufficiently descriptive for general purposes; but the designation *method of mean difference functions* would better convey an idea of the chief features of the technique elaborated, namely, the dependence on functions of finite differences in the derivation of equations defining parameters and the determination of the best value of a given parameter by taking the mean of the several approximations.

The fundamental requirement of this method is that, under the procedure followed, the reliability of the parameters determined shall be a maximum. This requirement results in a sum of absolute residuals which is less than that to be obtained by the Gaussian method of least squares or the Pearsonian method of moments. Rigorous adherence to the Edgeworthian requirement that the sum of the absolute residuals shall be a minimum is, it will be observed, not a demand of the present method. It can be shown that Lipka's method of averages will give the same residuals for a linear series as the method of differ-

ences; but it does not, however, give the same results in general.

The following claims to merit may be advanced for the method of differences:

(1) The computations involved in the determination of parameters are simple and easily checked.

(2) The method permits of fitting to a wide variety of functions by the direct application of its fundamental principles.

(3) The general technique developed may be adapted to special solutions in particular cases; e. g., the solutions of parabolic series given in section VII are special cases of the solution for the general polynomial series given in section X.

(4) The parametric approximations or some function involved in their determination give a convenient test of fit. If these approximations are nearly constant or fluctuate irregularly about a central value, the implication is that the test function is appropriate to the data; if the approximations show a systematic change or trend in their values, the implication is that the test function is inappropriate.

(5) The method yields satisfactory results in practice.

That the residuals do register our failure to predict the values of the observations is undeniable; but it does not follow that the least squares definition of residuals leads to the equation of greatest value for predictive purposes; for we can scarcely hope to establish that a set of residuals determined from a small number of observations constitutes a system of normally distributed variates.

Let us now consider the logistic series

$$y = m e^{-be^{-ax}} \tag{2a}$$

$$y = y_o \left(1 + e^{b+ax}\right)^{-1}, \tag{3a}$$

and

$$y = m : \left[1 + be^{-ax}\right]. \tag{4a}$$

This last is the Verhulst logistic.

The origin, maximum, and point of inflection of these three functions are determined by the following relations:

$$y_o = me^{-b},$$

(2b) $$dy = abme^{-ax}e^{-be^{-ax}}dx,$$

$$d^2y = a^2bme^{-be^{-ax}}(be^{-2ax} - e^{-ax})dx^2.$$

$$y = y_o(1+e^{b-\alpha x}),$$

(3b) $$dy = \alpha\beta e^{\alpha z}dz,$$

$$d^2y = \alpha^2\beta e^{\alpha z}dz^2 + \alpha\beta e^{\alpha z}d^2z.$$

$$y_o = m : [1+b],$$

(4b) $$dy = abm[e^{-ax} : (1+be^{-ax})^2]dx,$$

$$d^2y = 2a^2b^2me^{-2ax}(1+be^{-ax})dx^2 - 2a^2bme^{-2ax}(1+be^{-ax})^2dx^2 - (1+be^{-ax})^4dx^2.$$

With the origin at $f(x_o)$ and the maximum at $f(x_\infty)$, these curves show essentially the same properties and, therefore, negate the claim of Professors Pearl and Reed to have discovered in the Verhulst type logistic the unique mathematical expression for the growth of populations. This assertion, of course, makes no statement as to the type of population which is best represented by each curve.

In fitting type (2) to the population of Ohio, we have obtained, while not an ideal fit, certainly one much better than can be obtained by fitting type (3). These results, however, serve to enhance rather than diminish the general usefulness of the logistic hypothesis as an empirical generalization of the growth of populations.

As the writer conceives it, this hypothesis may be stated as follows: *When the growth of a population is not known to be correlated with events whose sequence is definitely known, it is best represented by a*

curve which proceeds from one horizontal straight line as asymptotic origin, passes through a point of inflection, and approaches a second horizontal straight line as asymptotic terminus. The rate of growth of such a curve may be characterized as proceeding from a minimum to a maximum and then decreasing toward zero as limit, a characterization which is in full accord with our decrease in knowledge concerning the rate of growth as time goes on. It will be noted, in our statement of the logistic hypothesis, that it is not necessary to place any restriction on the chronological direction of growth, interpretation of growth as proceeding forward or backward being equally permissible.

It is, of course, true that the particular function fitted to the Ohio population does not conform rigorously to the logistic type; for at $x = 90$, the ordinates begin to decline in value. But this is no detriment in the application of the function in the particular case, since no one would place any reliance on a forecast of such date when projected several centuries into the future. The use of such a function as we have employed seems far preferable to fitting the Verhulst function to subpopulations on the ground that the sum of logistics cannot be executed itself to be a logistic, a procedure which is strictly valid only when the growth changes of the subpopulations are mutually independent.[1]

When the growth of a population is known to be definitely correlated with an observed sequence of events, the logistic hypothesis must be modified accordingly. In a region where the population could not be recruited from without, an abrupt increase in the death rate, a decrease in the birth rate, or an emigration to regions outside would necessitate a modification of the growth formula.

NOTE:—In presenting this memoir to the public, the writer desires to make grateful acknowledgement of the invaluable assistance given by his wife, Hazel J. Will, in the preparation of the manuscript.

1. cf. Pearl and Reed: "The Population of an Area Around Chicago and the Logistic Curve." J. A. A. S., March, 1929.

Harry S. Will