EMPIRICAL CONSTANTS AND FORMULÆ.

By Arthur G. Smith

It is the purpose of the writer, in the following brief sketch, merely to bring together certain of the elementary methods that may be used for obtaining some of the simpler empirical formulæ. No claim of originality is made for the methods employed but rather an adaptation. Any one desirous of making a thorough and systematic study of the subject should have at command at least the works referred to at the close of this article and must possess a fair battery of mathematics.

The determination of empirical constants and also of empirical formulæ is a matter of great interest to the engineer or to any one engaged in work introducing the physical properties of matter. The engineer in his study of the resistance of materials is every day brought face to face with the uncertainties of the substances with which he is dealing. The factor of ignorance stands ever at his elbow to mock when he wishes to be most exact.

Never shall we know, probably, why materials become fatigued and must have rest, or just exactly what are the laws of strength for long columns under pressure. Experiment proves the first and experiments combined with certain rational deductions must give us the latter.

The thousands of tests being made every year throughout the world are giving definite data from which to determine the laws or to derive approximate rules in accordance with which the physical and chemical constants of materials act and combine. It is many times difficult to determine the exact form of a function by rational methods and therefore assumption and hypothesis must be called in freely to aid in the work. If the form of the function is known and only the constants are in doubt then their most probable values become the solution of an ordinary problem in Least Squares and presents no difficulty in determining these empirical constants.

The problem however presents itself again as the determination of a function which may best coincide with observed values for purposes of interpolation.

Since the method of solving ordinary observation equation and of testing the results may not be familiar to all readers, the writer presents a short synopsis of the principles and forms used in the method of least squares outlining proofs and developments only, as they may be found in any of the works referred to at the close of this article.

The development of the law of error by Gauss was based upon the hypothesis, by some termed an axiom, that if several values are observed of any magnitude and each observation be equally precise, then the most probable value is the arithmetic mean of these observations.

The following brief defininitions may be taken to define terms that will be frequently used:

By an *observation* will be meant a recorded measurement upon the magnitude under consideration.

The most probable value will be taken as the *best* value to be obtained from all the data available.

An error is the difference between the observation and the true value of the observed quantity, and which never can be found.

A residual is the difference between an observation and the most probable value.

The weight of an observation is an arbitrary number asserting that a particular observation is better, or worse than another. If to any particular observation obtained under cer-conditions we ascribe the weight *one* or call it the standard observation, then to say that the weight of another observa-is p is equivalent to considering the second one as worth p single observations of weight one.

If $Z_1, Z_2, Z_3 \ldots\ldots Z_n$ be n observations all equally good, then the most probable value of Z the measured quantity is

$$1) \quad Z = \frac{Z_1 + Z_2 \;\ldots\ldots Z_n}{n} = \frac{[Z]}{n}$$

this last form is due to Gauss.

If $Z_1, Z_2 \ldots\ldots Z_n$ be n observations with the respective weights $p_1 \; p_2 \ldots . p_n$
then

$$2) \quad Z = \frac{p_1 Z_1 + p_2 Z_2 \ldots\ldots p_n Z_n}{p_1 + p_2 \ldots\ldots\ldots p_n} = \frac{[pZ]}{[p]}$$

Z here is called the *weighted mean.*

The principle of Least Squares may be simply stated as this: *That, that value of the unknown is the most probable, which makes the sum of the squares of the errors a minimum,* but since the true errors cannot be found but instead only the residuals, we may state the law of error as follows:

When a quantity has been observed with the greatest degree of accuracy possible and the results are vitiated only by accidental errors, having been freed from constant errors such as personal equation in the observer or the effects of temperature upon the instruments then the most probable value is that one which makes the sum of the squares of the residuals the least.

If we write

$$Z - Z_1 = \triangle_1 \;,\; Z - Z_2 = \triangle_2 \;,\; Z - Z_3 = \triangle_3 \ldots\ldots$$
$$\ldots\ldots Z - Z_n = \triangle_n$$

then $\quad \triangle_1^2 + \triangle_2^2 + \ldots\ldots \triangle_n^2 =$ a minimum

by *probable eror* is understood that error plus or minus within which it is an even chance that the probable value lies.

For example the expression for the length of a bar of metal as $\overset{\text{in}}{36.125} \pm \overset{\text{in}}{0.014}$ is interpreted to mean that it is an even chance that the value $\overset{\text{in}}{36.125}$ is within $\overset{\text{in}}{0.014}$ of the true value, but whether larger or smaller there is no method of knowing.

Instead of the probable error preferably the writer would use the *mean square error*, an error which is the mean of the true error. This is coming into more general use at the present time.

e the probable error of observation of weight unity.
e_p " " " " " " " p.
r_o " " " " arithmetic mean.

these quantities are given by the formulæ:

$$e = 0.6745 \left(\frac{[v\,v]}{n - 1} \right)^{\frac{1}{2}} \qquad (3)$$

$$e_p = \frac{e}{\sqrt{p}} = 0.6745 \left(\frac{[v\,v]}{p\,(n - 1)} \right)^{\frac{1}{2}} \qquad (4)$$

$$r_\circ = 0.6745 \left(\frac{[v\,v]}{n\,(n - 1)} \right)^{\frac{1}{2}} \qquad (5)$$

where $[v\,v] = v_1^2 + v_2^2 + v_3^2 + \ldots\ldots v_n^2$. $v_1\, v_2 \ldots\ldots v_n$ being the residuals found by subtracting the successive observations from the mean value and $n =$ the number of observations. (5) and (4) are seen to be the same and also they both show that the probable error decreases as the square root of the number of observation.

The simplest problem in the adjustment of observation and the determination of the probable values, is of course a series of direct observations upon the desired magnitude; in which case a simple determination of the arithmetic mean gives the desired quantity.

However many quantities are to be found as one among several united by some linear function with other unknowns. The general solution of such a series will be briefly sketched according to the algorithm given by Gauss. Who developed the principles of this method in his *Theoria Motus Corporum Cœlestum.*

Let the numbers be united in the observation equation.

$$\begin{array}{ll}
a_1\, x_1 + b_1\, x_2 + \ldots\ldots\ldots + k_1\, x_i - l_1 = o & \\
a_2\, x_1 + b_2\, x_2 + \ldots\ldots\ldots + k_2\, x_i - l_2 = o & \\
a_3\, x_1 + b_3\, x_2 + \ldots\ldots\ldots + k_3\, x_i - l_3 \ldots\ldots & \qquad (A) \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
a_n\, x_1 + b_n\, x_2 + \ldots\ldots\ldots\ldots k_n\, x_i - l_n = o &
\end{array}$$

where n here or the number of equations is greater than i the number of unknowns frequently many times greater.

Then if the n equation be solved in sets of i at a time a series of values may be obtained for $x_1, x_2, \ldots\ldots\ldots x_i$ and if every possible combination be made there will result

$$\frac{n\,(n-1)\;(n-2)\ldots\ldots(n-\mathrm{i}+1)}{1\;.\;2\;\;.3\ldots\ldots\ldots\ldots\ldots\ldots i}$$

values for each of the unknowns. The arithmetic mean of which would be the most probable values of the unknowns. This method is evidently impracticable because of the great labor involved, as say 30 observation equations in 5 unknowns would entail the solution of 23751 sets of five equations in 5 unknowns. The method to be used is the one devised by Gauss and which enables an absolute check upon the accuracy of the work to be easily applied.

It may readily be shown that the most probable values of the i unknowns can be obtained by the solving of i equations called the Normal Equations, found by the following method from the given n observation equation.

A normal equation say for x_k is found by multiplying each of the observation equations through by the coëfficient of x_k in that particular equation, and then forming the sums of all the equations so multiplied. A normal equation is in like manner found for each of the unknowns and the solution of these i normal equations gives the desired values of the unknowns.

Assume the observation equations

$$a_1\,x + b_1\,y + c_1 = o$$
$$a_2\,x + b_2\,y + c_2 = o$$
$$a_3\,x + b_3\,y + c_3 = o$$

If then we use Gauss' notation as follows:

$$[a\;a] = a_1^2 + a_2^2 + a_3^2$$
and $$[a\;b] = a_1\,b_1 + a_2\,b_2 + a_3\,b_3$$

then the normal equations become for x and y respectively

$$[a\;a]\,x + [a\;b]\,y + [a\;c] = o$$
$$[a\;b]\,x + [b\;b]\,y + [b\;c] = o$$

It will readily be seen that the r^{th} coefficient in the k^{th} column is the same as the k^{th} coëfficient in the r^{th} column.

The normal equations then become for (A)

$$\begin{array}{llr}
[a\,a]\,x_1 + [ab]\,x_2 + \ldots\ldots\ \ldots\ldots\ [a\,k]\ x_i & + [a\,l] = o & \\
[a\,b]\,x_1 + [b\,b]\,x_2 \ldots\ldots\ \ldots\ldots\ \ldots & + [b\,l] = o & \\
[a\,c]\,x_1 + [b\,c]\,x_2 \ldots\ldots\ \ldots\ldots\ \ldots & + [c\,l] = o & (B) \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\
[a\,k]\,x_1 + b\,k\,x_1 \ldots\ldots\ldots\ldots\ldots\ldots & [kl] = o &
\end{array}$$

A check upon the work may be carried out as follows, Write the observation equations with the coëfficients detached as follows, adding an extra column s where

$$a_1 + b_1 + c_1 \ldots . + l_1 = s_1$$

x_1	x_2	x_3	x_4..........			
a	b	c	d	....	l	s
a_1	b_1	c_1			l_1	s_1
a_2	b_2	....	....	....		s_2
a_3						
....	....	....	....	....		
....	....	....	....	....		
a_n	b_n	....	....	....	l_n	s_n

In multiplying by the coëfficient of x_k in forming the normal equation for x_k multiply s in every case, also by the coëfficient of x and we have at once the absolute check.

$$\begin{array}{l}
[a\,a] + [a\,b] + [a\,c] \ldots\ldots\ldots\ldots [a\,l] = [a\,s] \\
[a\,b] + [b\,b] + \ldots\ldots\ldots\ldots\ldots\ldots [b\,l] = [b\,s] \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ \ldots\ldots\ldots\ldots \\
[a\,l] + [b\,l] + \ldots\ldots\ldots\ \ldots\ldots [l\,l] = [l\,s]
\end{array}$$

this check is accurate and requires only the computation of $[a\ s]$, $[b\ s]$ .... $[l\ s]$ in addition to the necessary products.

SOLUTION OF THE NORMAL EQUATION.

Let us assume three normal equations and the solution may be finished, using the notation of Gauss very simply; the method is of course identical whatever be the number of normal equations.

The normal equations are

$$\left.\begin{array}{l}[a\,a]\,x+[a\,b]\,y+[a\,c]\,z+[a\,l]=0\\ [a\,b]\,x+[b\,b]\,y+[b\,c]\,z+[b\,l]=0\\ [a\,c]\,x+[b\,c]\,y+[c\,c]\,z+[c\,l]=0\end{array}\right. \qquad (C)$$

From the first equation

$$x=-\frac{[a\,b]}{[a\,a]}y-\frac{[a\,c]}{[a\,a]}z-\frac{[a\,l]}{[a\,a]} \qquad (D)$$

this substituted in the remaining equations gives, using the notation of Gauss,

$$\left.\begin{array}{l}[b\,b.\,1]\,y+[b\,c.\,1]\,z+[b\,l.\,1]=0\\ [b\,c.\,1]\,y+[c\,c.\,1]\,z+[c\,l.\,1]=0\end{array}\right. \qquad (E)$$

where

$$[b\,b.\,1]=[b\,b]-\frac{[a\,b]}{[a\,a]}[a\,b]$$

$$\left.\begin{array}{l}[b\,c.\,1]=[b\,c]-\dfrac{[a\,b]}{[a\,a]}[a\,c]\\[2ex] [b\,l.\,1]=[b\,l]-\dfrac{[a\,b]}{[a\,a]}[a\,l)\\[2ex] [c\,c.\,1]=[c\,c]-\dfrac{[a\,c]}{[a\,a]}[a\,c]\\[2ex] [c\,l.\,1]=[c\,l]-\dfrac{[a\,c]}{[a\,a]}[a\,l]\end{array}\right\}\ldots\ldots\ldots\ldots(F)$$

Now from the first of E

$$y=-\frac{[b\,c.\,1]}{[b\,b.\,1]}z-\frac{[b\,l.\,1]}{[b\,b.\,1]}\ldots\ldots\ldots \qquad (G)$$

this in the second of E gives

$$z=\frac{[c\,l.\,2]}{[c\,c.\,2]}\ldots\ldots\ldots\ldots \qquad (H)$$

where

$$\left.\begin{array}{l}[c\,l.\,2]=[c\,l.\,1]-\dfrac{[b\,c.\,1]}{[b\,b.\,1]}[b\,l.\,1]\\[2ex] [c\,c.\,2]=[c\,c.\,1]-\dfrac{[b\,c.\,1]}{[b\,b.\,1]}[b\,l.\,1]\end{array}\right\} \qquad (I)$$

the equations F and I are termed the auxiliary normal equations. It should be noticed that in every case in (E) the expression reduces to the corresponding coëfficient in (C) minus a fraction whose numerator is the product of the coëfficient first in the row times the one at the top of the column in which the desired coëfficient is found; while the denominator is the leading one in the first equation. For example from [C] to find the coëfficient [b c. 1] which in (E) replaces [b c] in C. We have

$$[b\,c - \frac{[a\,c]}{[a\,a]}[a\,b]$$

and the coëfficients of the successive auxiliary equations can necessarily be found by the same method.

Since in the determination of empirical constants and formulæ seldom is a large number of unknowns to be determined, no farther checks or suggestions for the solution of normal equations will be given here. Very complete checks as well as complete proofs of all the steps here briefly outlined may be found in the texts referred to at the close of this article by Wright, and Merriman.

II.

THE DETERMINATION OF ARBITRARY EMPIRICAL FORMULÆ

If we assume

(1) .... $y = \dfrac{m + x}{n + x}$ to be the relation connecting y and x, since from two equations two unknowns may be found it would require but two pairs of observations giving values corresponding of x and y to determine values of m and n. The problem may present itself in two ways. First, we may know that the form of the function given is correct, as for example, measurements to determine the equation of a straight line; here it is known the function must be $y = a + b\,x$ and we have only to determine the values of the arbitrary constants a and b; secondly, even the form of the function may be in doubt, in which case, the choosing of the form of the function is of great importance.

A large number of physical relations are of a form in which as one of the variables increases the other also continues to increase; and again, since any function of a variable may by Taylor's formula be developed into a series in increasing powers of the variable we may always assume as at least a possible form $y = a + b\,x + cx^2 + \ldots\ldots\ldots\ldots$

In practice, however, while giving a possible form, this particular form may be useless on account of insurmountable difficulties in obtaining terms sufficient to be exact without the form becoming unwieldly because of its great length. An empirical formula can not be considered as satisfactory until it will satisfy two imperative conditions. It must first be simple, convenient and not too cumbersome in application.

Secondly, it must upon the interpolation of any one of the observed values give the corresponding value of the other variable to a degree of accuracy at least equal to the exactness with which it can be measured. This measure of accuracy must hold true at least between the definite limits entering into the observations from which the formulæ has been derived.

From what has just been said it will be evident that the study of empirical formulæ involves two problems; first, the form to be chosen; second, the computation of the constants. It may be said here that the most important part is the choice of the best form, for while the choice of almost any form may give a formulæ of considerable exactness, yet the necessity of the form being *convenient* must never be lost sight of.

It frequently occurs that because of greater skill upon the part of one observer, or again because of more refined methods certain of the observed values have been obtained more accurately than others. Each pair of observed values serving to fix a point upon the actual observed curve; but since the formula adopted to represent the function does not in general give a curve exactly coinciding with any of the observed points, it is but natural we should desire a closer agreement for those values known to be most nearly correct. We wish then to emphasize the good observations and develop a function that shall coincide more closely with these better observed values. It is evident at once that this is equivalent to giving a greater weight to these particular values. If we consider one pair of observed values as being better than the rest, then this particular value may be looked upon as being repeated p times where p is the weight of the observation.

Let us assume then that we have a series of observation equations as follows:

$$\begin{array}{lrlll}
a_1 x_1 + b_1 x_2 + c_1 x_3 + \dots\dots & k_1 x_k + l_1 = o \;.\;. & \text{with weight} & & p_1 \\
a_2 x_1 + b_2 x_2 + \dots\dots\dots\dots & + l_2 = o & \text{“} & \text{“} & p_2 \\
a_3 x_1 + \dots\dots\dots\dots\dots\dots\dots & = o & \text{“} & \text{“} & p_3 \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots & & & & \\
a_n x_1 + b_n x_2 + \dots\dots\dots\dots\dots & + l_n = o & \text{“} & \text{“} & p_n
\end{array}$$

Now if we assume that the first observation equation is repeated p_1 times the second one p_2 times and so on to the last, and should actually rewrite each equation as many times as the weight would indicate, it is evident that the normal equations would become

$$\begin{array}{lr}
[p\,a\,a]\,x_1 + [p\,a\,b]\,x_2 + [p\,a\,c]\,x_3 + \dots\dots & + [p\,a\,l] = o \\
[p\,a\,b]\,x_1 + [p\,b\,b]\,x_2 + \dots\dots\dots\dots\dots & = o \\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots & \\
[p\,a\,k]\,x_1 + \dots\dots\dots\dots\dots\dots\dots\dots & + p\,[k\,l] = o
\end{array}$$

and this would be equivalent to multiplying each observation equation by the square root of its corresponding weight and then treating the weighted equations as a series of equations of equal weight.

Let us consider the following observed values taken from *Steinhauser*, for the liquefaction of ammonia gas; the observed quantities being respectively y the pressure in atmospheres and x the temperature centigrade the observations are

$$\begin{array}{llll}
y_1 = 2.48 & y_2 = 5.00 & y_3 = 7.00 & y_4 = 8\,00 \\
x_1 = -17.8 & x_2 = 4.2 & x_3 = 16.3 & x_4 = 20.3
\end{array}$$

We will assume the form

$$y = a + b\,x + cx^2 \;\dots\dots\dots\dots\dots (2)$$

The observation equations then become

$$\begin{array}{lll}
2.48 = a - 17.8\,b + 316.84\,c & \text{weight} & 2 \\
5.00 = a + 4.2\,b + 17.64\,c & \text{``} & 3 \\
7.00 = a + 16.3\,b + 265.69\,c & \text{``} & 4 \\
8.00 = a + 20.3\,b + 412.09\,c & \text{``} & 5
\end{array}$$

The writer has arbitrarily assumed the individual weights. Multiplying each equation by the square root of its weight we have

$$\sqrt{(2)}\Big(a - (17.8)\ b + (316.84)\ c - (2.48)\Big) = o$$

$$\sqrt{(3)}\Big(a + (\ 4.2\ )\ b + (\ 17.64\ )\ c - (5.00)\Big) = o$$

$$\sqrt{(4)}\Big(a + (16.3)\ b + (265.64)\ c - (5.00)\ \Big) = o$$

$$\sqrt{(5)}\Big(a + (20.3)\ b + (413.09)\ c - (8.00)\Big) = o$$

The normal equations become

$$\begin{array}{l}
14.00\,a + 143.70\,b + 3809.81\,c - 87.9600 = o \\
143.70\,a + 3809.810\,b + 48092.88\,c - 1243.1120 = o \\
3809.81\,a + 48092.883\,b + 1333082.98\,c - 9274.0464 = o
\end{array}$$

The solution of these equations gives

$$a = 18.9894\ ,\ b\ =\ .3808\ ,\ c\ =\ -\ .006106$$

the formula becoming

$$y = 18.9894 + 0.3808\,x - 0{,}06106\,x^2 \ldots\ldots\ldots\ldots (3)$$

substituting the observed values of x in the formula we derive for

$$\begin{array}{llll}
x = -17.8 & y = -\ 7.135 & \text{instead of} & y = 2.48 \\
x = \ \ 4.2 & y = 19.5117 & \text{``}\quad\text{``} & y = 5.00 \\
x = \ \ 1.63 & y = \ \ 7.9734 & \text{``}\quad\text{``} & y = 7.00 \\
x = \ \ 20.3 & y = \ \ 1.5574 & \text{``}\quad\text{``} & y = 8.00
\end{array}$$

these values give us the residuals

$$\delta_1 = -\ 9.62,\ \delta_2 = 14.51,\ \delta_3 = 0.97,\ \delta_4 = -\ 6.44$$

$$[\delta] = -\ 0.58$$

It may be desired in a simple case to so determine the formula that it shall pass through all of the observed points. This result may be accomplished by assuming the formula with n constants to be determined n being the number of observations given. Thus for the observations given upon the pendulum, l being the length and t the time of vibration of the pendulum.

$$\left\{\begin{array}{l} l_1=37\ 85 \\ t_1=\ 1.24 \end{array}\right., \left\{\begin{array}{l} l_2=63.55 \\ t_2=1.645 \end{array}\right., \left\{\begin{array}{l} l_3=84.05 \\ t_3=\ 1.83 \end{array}\right., \left\{\begin{array}{l} l_4=99.35 \text{ centimeters} \\ t_4=\ 2.00 \text{ seconds} \end{array}\right.$$

the formula should be assumed

$$l = a + b\,t + c\,t^2 + d\,t^3$$

or

$$t = m + n\,l + p\,l^2 + q\,l^3$$

The equations may be solved of course by any method for the treatment of an ordinary set of simultaneous linear equations. The interpolation formulæ of Lagrange, however offer a simple and very satisfactory method in this instance.

Let $\left\{\begin{array}{l} u_1 \\ v_1 \end{array}\right., \left\{\begin{array}{l} u_2 \\ v_2 \end{array}\right., \left\{\begin{array}{l} u_3 \\ v_3 \end{array}\right. \ldots\ldots\ldots\ldots\ldots\ldots, \left\{\begin{array}{l} u_n, \\ v_n \end{array}\right.$

be the pairs of observed values.

And let x and y be two variables corresponding to u and v respectively and let us assume that the relation obtains

$$y = v_1 F_1(x, u_1 u_2 \ldots u_n) + v_2 F_2(x, u_1, u_2 \ldots u_n) + v_3 F_3\,(x, u_1, u_2, \ldots u_n)$$
$$+ \ldots\ldots\ldots\ldots\ldots\ldots v_n F_n\,(x, u_1, u_2, \ldots u_n) \ldots\ldots (1)$$

where $F_1\,(x_1, u_1, u_2 \ldots)$, $F_2(x_1, u_1, \ldots) \ldots F_n\,(x_1\,u_1 \ldots)$ are functions of the independent variable x.

Now bearing in mind the assumed condition that the function must, upon substitution of one of the observed quantities, give the exact value of the corresponding other observed quantity, the following conditions must be satisfied.

When

x assumes the value u_1 , y must equal v_1
x " " " u_2 , y " " v_2
. . . .
. . . .
x " " " u_n , y " " v_n

These conditions will be satisfied when, writing F to represent the entire function $F(x, u_1\,u_2 \ldots u_n)$, we have for

$x = u_1$, $F_1 = 1$, while $F_2 = F_3 = F_4 = \ldots\ldots F_n = o$
$x = u_2$, $F_2 = 1$ " $F_1 = F_3 = F_4 = \ldots\ldots F_n = o$
$x = u_n$, $F_n = 1$ " $F_1 = F_2 = F_3 = \ldots\ldots F_{n-1} = o$

Taking the general term we see that when $x = u_r$ $F_r = 1$

while for every other observed value the function must vanish. This property would naturally suggest from its similarity to the vanishing of an algebraic equation for one of its roots, that F_r may be written in the form

$$F_{(r)} = K_r (x - u_1)(x - u_2) \,..\, (x - u_{r-1})(x - u_r + 1) \,..\, (x - u_n)$$

The factor $(x - u_r)$ is of course omitted and is the only one. K_r is a constant to be determined. The form just given satisfies the condition of vanishing for every observed value of x except $x = u_r$. There is, however, the farther condition that $F_{(r)}$ must equal 1 when $x = u_r$ that is

$$1 = K_r (u_r - u_1)(u_r - u_2) \,..\, (u_r - u_{r-1})(u_r - u_r + 1) \,..\, (u_r - u_n)$$

This at once determines the value of K_n

$$K_n = \frac{1}{(u_r - u_1)(u_r - u_2) \,..\, (u_r - u_{r-1})(u_r - u_n + 1) \,..\, (u_r - u_n)}$$

Replacing K_r in F by the value just found we have

$$F_r = \frac{(x - u_1)\ (x - u_2) \ldots (x - u_{r-1})(x - u_{r+1}) \,..\, (x - u_n)}{(u_r - u_1)\ (u_r - u_2) \ldots (u_r - u_{r-1})\ (u_r - u_{r+1}) \,..\, (u_r - u_n)}$$

The desired function of y can now be written by simple substitution, it becomes

$$\left.\begin{aligned} y = v_1 & \frac{(x - u_2)\ (x - u_3)\ \ldots\ldots (x - u_n)}{u_1 - u_2\ (u_1 - u_3) \ldots\ldots (u_1 - u_n)} + \\ + v_2 & \frac{(x - u_1)\ (x - u_3) \ldots\ldots (x - u_n)}{(u_2 - u_1)\ (u_2 - u_3) \ldots\ldots u_2 - u_n} + \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ + v_n & \frac{(x - u_1)\ (x - u_2) \ldots\ldots (x - u_{n-1})}{(u_n - u_1)\ (u_n - u_2) \ldots (u_n - u_{n-1})} \end{aligned}\right\} \ldots\ldots(2)$$

Expanding the numerator of $F(r)$ and multiplying out the product in the denominator we see that $F(r)$ is a function of the $n - 1$ degree and may be written

$$F(r) = \frac{a_1 + b_1 x + c_1 x^2 + \ldots\ldots\ldots\ldots k_1 x^{n-1}}{D_r}$$

using similar notation we may write the desired formula

$$\left.\begin{aligned} y = v_1 & \frac{a_1 + b_1 x + c_1 x^2 \ldots\ldots\ldots\ldots k_1 x^{n-1}}{D_1} \\ + v_2 & \frac{a_2 + b_2 x + c_2 x^2 + \ldots\ldots\ldots\ldots k_2 x^{n-1}}{D_2} \\ + & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ + v_n & \frac{a_n + b_n x + c_n x^2 + \ldots\ldots \ldots\ldots k_n x^{n-1}}{D_u} \end{aligned}\right\} \ldots\ldots(3)$$

clearing of fractions and collecting like powers this gives

$$y = A + Bx + Cx^2 + \ldots\ldots\ldots\ldots Kx^{n-1} \ldots\ldots\ldots\ldots(4)$$

the desired formula satisfying the observed values and contains the n required constants.

As an application of the above again take the observed values

$$\begin{cases} l_1 = 37.85 \\ t_1 = \;\; 1.24 \end{cases}, \begin{cases} l_2 = 63.55 \\ t_2 = 1.645 \end{cases} \begin{cases} l_3 = 84.05 \\ t_3 = \;\; 1.83 \end{cases} \begin{cases} l_4 = 99.35 \text{ cent.} \\ t_4 = \;\; 2.00 \text{ sec.} \end{cases}$$

assume the formula

$$l = a + b\,t + c\,t^2 + dt^3$$

a form introducing four constants.

Using (2) we have

$$\begin{aligned} l = 37.85 & \frac{(t - 1.645)\ (t - 1.83)\ (t - 2.00)}{(1.24 - 1.645)\ (.24 - 1.83)\ (1.24 - 2.00)} \\ + 63.55 & \frac{(t - 1.24)\ (t - 1.83)\ (t - 2.00)}{(1.645 - 1.24)\ (1.645 - 1.83)\ (1.645 + 2.00)} \\ + 84.05 & \frac{(t - 1.24)\ (t - 1.645)\ (t - 2.00)}{(1.83 - 1.24)\ (1.83 - 1.645)\ (1.83 -- 2.00)} \\ + 99.35 & \frac{(t - 1.24)\ (t - 1.645)\ (t - 1.83)}{(2.00 - 1.24)\ (2.00 - 1.645)\ (2\ 00 - 1.83)} \end{aligned}$$

This becomes

$$\begin{aligned} l = - \; & 208.423\ (t - 1.645)\ (t - 1.83)\ (t - 2.00) \\ + \; & 2389.245\ (t - 1.24)\ (t - 1.83)\ (t - 2.00) \\ - \; & 4529.653\ (t - 1.24)\ (t - 1.645)\ (t - 2.00) \\ + \; & 2166.09\ (t - 1.24)\ (t - 1.645)\ (t - 1.83) \end{aligned}$$

Expanding we have

$$\begin{aligned} l = &- 208.422\ (t^3 - 5.475\,t^2 + 9.96035\,t - 6.0207) \\ &+ 2389.245\ (t^3 - 5.07\,t^2 + 8.4092\,t - 4.5384) \\ &- 4529.653\ (t^3 - 4.885\,t^2 + 7.8098\,t - 4.0796) \\ &+ 2166.09\ (t^3 - 4.715\,t^2 + 7.31935\,t - 3.732834) \end{aligned}$$

This gives for the final formula,

$$l = - 182.741\,t^4 + 941.884\,t^2 - 1505.64\,t + 805.022 \ldots\ldots (4)$$

For testing formula we have the additional observation values

$$l_1 = 46.65 \qquad l_2 = 72.75 \qquad l_3 = 88.35$$
$$t_1 = 1.37 \qquad t_2 = 1.72 \qquad t_3 = 1.885$$

Substituting the values of t_1, t_2, t_3, we get for l_1, l_2, l_3, respectively the values

$$40.29 \quad , \quad 71.91 \quad , \quad 90{,}14$$

instead of the observed values, giving the residuals

$$\delta_1 = 6.36 \qquad \delta_2 = 0.84 \qquad \delta_3 = - 1.79$$
$$[\delta] = 5.41.$$

Empirical formnlæ that shall exactly satisfy one or more of the observed values, or some known value determined by theoretical conditions.

For example the following formula for the elastic force of water vapor at any temperature

$$(5) \ldots\ldots\ldots \mu = 760 + b(t - 100) + c(t - 100)^2 + \ldots\ldots$$

gives $\mu = 760$ when $t = 100$ where μ is in millimeters and t the temperature centigrade.

In general then

the formula $y = y_1 + b(x - x_1) + c(x - x_1)^2 + \ldots\ldots\ldots\ldots$ will give the exact value $y = y_1$ when $x = x_1$. The equations of condition then become for finding the cöefficients

$$\begin{aligned} y_1 - y_1 &= b\ (x_1 - x_1) + c\ (x_1 - x_1)^2 + d\ (x_1 - x_1)^3 + \ldots \\ y_2 - y_1 &= b\ (x_2 - x_1) + c\ (x_2 - x_1)^2 + d\ (x_2 - x_1)^3 + \ldots \\ y_3 - y_1 &= b\ (x_3 - x_1) + c\ (x_3 - x_1)^2 + \ldots\ldots\ldots\ldots \\ y_4 - y_1 &= b\ (x_4 - x_1) + \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots & \\ y_n - y_1 &= b\ (x_n - x_1) + c\ (x_n - x_1)^2 + d\ (x_n - x_1)^3 + \ldots \end{aligned}$$

the normal equations become

$$(6)\ldots\ldots\quad \begin{aligned}&[(y-y_1)(x-x_1)] = b[(x-x_1)^2] + c[(x-x_1)^3] \\ &+ d[(x-x_1)^4] + \ldots\ldots \\ &[(y-y_1)(x-x_1)^2] = b[(x-x_1)^3] + c[(x-x_1)^4] \\ &+ d[(x-x_1)^5] + \ldots\ldots \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\end{aligned}$$

In substituting in the above every value of y and x must be used from y_1, x_1 to y_n, x_n.

The solution of the above normal equations gives the complete formula, which may be expanded and written in the simple form,

$$y = a_0 + b_0 x + c_0 x^2 + d_0 x^3 + \ldots\ldots\ \ldots\ \ldots\ldots\ldots\ldots \quad (7)$$

Again suppose it is desired to develop a parabolic form that shall exactly satisfy *two* given points.

We must then have the two condition equations.

$$y_1 = a + b\,x_1 + c\,x_1^2 + d\,x_1^3 + \ldots\ldots\ldots\ldots$$

$$y_2 = a + b\,x_2 + c\,x_2^2 + d\,x_2^3 + \ldots\ldots\ldots\ldots$$

Subtract and solve for b finding

$$b = \frac{y_1 - y_2}{x_1 - x_2} - c\,(x_1 + x_2) - d\,(x_1^2 + x_1\,x_2 + x_2^2) \ldots\ldots$$

also

$$a = \frac{x_1\,y_2 - y_1\,x_2}{x_1 - x_2} + c\,x_1\,x_2 + d\,(x_1^2\,x_2 + x_1\,x_2^2) + \ldots\ldots$$

Insert these values in the assumed equation

$$y = a + b\,x + c\,x^2 + d\,x^3 + \ldots\ldots\ldots\ldots$$

and we have

$$(8)\ldots y = \frac{y_1\,(x - x_2) - y_2\,(x - x_1)}{x_1 - x_2} + c\,(x - x_1)\,(x - x_2)$$
$$+ d\,(x - x_1)\,(\mathrm{X} - x_2)\,(x - x_1 + x_2)$$
$$+ \ldots\ldots\ldots\ldots$$

This gives at once for $x = x_1$, $y = y_1$ and for $x = x_2$, $y = y_2$.

The observation equations then become

$$y_3 = \frac{y_1 (x_3 - x_2) - y_2 (x_3 - x_1)}{x_1 - x_2} + c (x_3 - x_1) (x_3 - x_2)$$
$$+ d (x_3 - x_1) (x_3 - x_2) (x_3 + x_1 + x_2) + \dots.$$

$$y_4 = \frac{y_1 (x_4 - x_2) - y_2 (x_4 - x_1)}{x_1 - x_2} + c (x_4 - x_1) (x_4 - x_2)$$
$$+ d (x_4 - x_1) (x_4 - x_2) (x_4 + x_1 + x\ ) + \dots.$$

. . . .

We will for simplification replace the first terms upon the right side by M_3 , M_4 . . . .
that is

$$M_3 = \frac{y_1 (x_3 - x_2) - y_2 (x_3 - x_1)}{x_1 - x_2}$$

Replacing the cöefficients of *c* by
N_3 , N_4 , ...
and of *d* by P_3 , P_4 . . . .
the equations become

$$y_3 = M_3 + c\, N_3 + d\, P_3 + \dots\dots\dots$$
$$y_4 = M_4 + c\, N_4 + d\, P_4 + \dots\dots\dots$$

. . . .

The normal equations become

$$[(y - M) N] = c [N^2] + d [N P]$$
$$[(y - M) P] = c [N P] + d [P^2]$$

. . . .

The solution may be readily completed and by elementary substitutions the complete formulæ may be obtained.

The above methods may be extended to the determination of formulæ which shall exactly satisfy for three or more definite values. There will be in each case as many equations of condition as the number of values desired, and these equations will determine an equal number of the assumed cöefficients in terms of the exact values of the observed quantities, which it is desired the formula shall satisfy.

The determination of formulæ that shall give some of the observed values more accurately than others is only a problem in weighting and involves no difficulty when once the proper form has been chosen.

Formulæ in which one variable increases while the other decreases.

If the observations indicate approximately an inverse ratio to exist between the variables the form $x\,y = c$ may be assumed. Since however the curve may not be symmetrical with respect to the center of coördinates it may be better to assume the center of the hyperbola as not being at the origin.

Taking the center at (m,n) we may write the equation

$$(y-n)(x-m)=\mathrm{K}$$

and this may be changed to

$$y=n+\frac{\mathrm{K}}{x-m}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

the signs of the constants here may be plus or minus, and, taking all possible combinations we have eight cases. Four hyperbolas with branches in first and third quadrants, and four hyperbolas with branches in the second and fourth quadrants.

MISCELLANEOUS FORMS.

1st. *Both variables proceed in arithmetic progressions of the first order.*

Representing the common differences by d and $\triangle$ we may write

$$x_n=x_1+(n-1)\,d$$
$$y_n=y_1+(n-1)\,\triangle$$

Eliminating $n-1$ we find

$$\frac{y_n-y_1}{\triangle}=\frac{x_n-x_1}{d}$$

or

$$y_n=\frac{\triangle}{d}x_n-\frac{\triangle}{d}x_1+y_1$$

Writing $\mathrm{A}=y_1-\frac{\triangle}{d}x_1$, $\mathrm{B}=\frac{\triangle}{d}$

this form becomes at once

$$y=\mathrm{A}+\mathrm{B}x\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

The following observations made by C. C. Foster and H. P. Burgum upon the torsion of a small pine beam, give series of values that would suggest the above form at once, p is the weight applied at end of lever arm, while θ is the angle in radians through which the beam was twisted.

$p_1=1120$	$p_5=1962$	$\theta_1=.02575$	$\theta_5=.1375$
$p_2=1360$	$p_6=2160$	$\theta_2=.0550$	$\theta_6=.1650$
$p_3=1532$	$p_7=2360$	$\theta_3=.0825$	$\theta_7=.1925$
	$p_8=2580$	$\theta_4=.1100$	$\theta_8=.2200$

2nd. *Again if y varies in an arithmetic series of the third order, while the independent variable proceeds by one of the first order.*

By the above statement one is merely to understand that the dependent variable y in the third order of differences *tends* to approach a constant, were it the case exactly any ordinary interpolation form would be exact.

We have the following equations:

$$y_n=y_1+(n-1)\triangle_1+\frac{(n-1)(n-2)}{1\,.\,2}\triangle_2+\frac{(n-1)(n-2)(n-3)}{1.2.3.}\triangle_3+..$$

and

$$x_n=x_1+(n-1)d_1$$

Eliminating the n from these equations we again obtain after simplifying,

$$y=a+bx+cx^2+dx^4 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11)$$

We see then that the condition given above for the variables is to be developed in this parabolic form, and the same is true, and proven in exactly the same manner when the dependent variable proceeds by an arithmetic series of the fourth order.

3rd. *The y proceeds by a geometrical series while the x is in arithmetical progression.*

We must have then

1) $y_n=y_1r^{n-1}$

2) $x_n=x_1+(n-1)d$

or $$n-1=\frac{x_n-x_1}{d}$$

this in 1) gives

$$y_n=y_1r^{\frac{x_n-x_1}{d}}$$

taking logarithms

$$log\, y_n = log\, y_1 + \frac{x_n-x_1}{d}\, log\; r$$

reducing

$$log\; y_n=\left(log\; y_1-\frac{log\;\, r}{d}\,x_1\right)+\left(\frac{log\;\, r}{d}\right)x^n$$

write this

$$log\; y_n=\mathrm{M}+\mathrm{N}x_n$$

introducing M and N for simplification, dropping the subscripts the formula is

$$log\; y=\mathrm{M}+\mathrm{N}x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

We have then the series of observation equations

$$log\; y_1=\mathrm{M}+\mathrm{N}x_1$$
$$log\; y_2=\mathrm{M}+\mathrm{N}x_2$$
$$\ldots\ldots\ldots\ldots$$
$$\ldots\ldots\ldots\ldots$$

and from these we find M and N. It should be noticed here that we are not in strict accord with the method of least squares making the sum of the squares of the errors in the variables but of the errors in the *logarithms* of the variables a minimum. The results will however be in very close accord.

The form of an empirical formula may often be determined by theoretical considerations and then the constants formed by experiment. Hodgkinson developed an empirical formula for the crushing load in the case of long cast iron columns as follows: The load which a column will support varies no doubt directly with *some* power of the diameter and also just as clearly varies inversely as some power of the length. We may then write for the crushing load P of a column the expression

$$p=a\,\frac{d^x}{l^y} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (13)$$

where p=load in tons necessary to crush

d=diameter of columns in inches

l=length in feet and a, x, y are constants to be determined by experiment.

Passing to logarithms we have

$$\log p + y \log l = \log a + x \log d$$

This may be written by substituting

$$\log p = \text{P} \qquad \log d = \text{D}$$
$$\log l = -\text{L} \qquad \log a = z$$
$$\text{P} = z + \text{L}y + \text{D}x \ldots\ldots\ldots\ldots\ldots \quad (14)$$

This gives a series of observation equations of the form

$$\text{P}_1 = z + \text{L}_1 y + \text{D}_1 x$$
$$\text{P}_2 = z + \text{L}_2 y + \text{D}_2 x$$
$$\ldots\ldots\ldots\ldots\ldots$$

These give the normal equations

$$[\text{P}] = z + [\text{L}]y + [\text{D}]x$$
$$[\text{PL}] = [\text{L}]z + [\text{L}^2]y + [\text{L,D}]x$$
$$[\text{PD}] = [\text{D}]z + [\text{LD}]y + [\text{D}^2]x$$

the solution of these will give the most probable value of x and y and the log a, and the problem is completed.

Again the case when y varies in a series where *first differeuces* are in geometric ratio, while x varies in arithmetic progression. We have then for x

$$x_n = x_1 + (n-1)d \qquad (1)$$

$$\text{or} \quad n - 1 = \frac{x_n}{d} - \frac{x_1}{d} \qquad (2)$$

calling r the ratio in the differences of y

and $\triangle_1$, $\triangle_2$............the differences

then $\triangle_2 = \triangle_1 r$............

also $y_3 = y_2 + \triangle_2 = y_1 + \triangle_1 + \wedge_1 r$

$y_4 = y_3 + \triangle_3 = y_1 + \triangle_1 + \triangle_1 r + \triangle_1 r^2$

....

In general

$$y_n = y_1 + \triangle_1 (1 + r + r^2 \ldots\ldots\ldots\ldots\ldots r^{n-2})$$

$$y_n = y_1 + \triangle_1 \frac{r^{n-1} - 1}{r - 1} \ldots\ldots\ldots\ldots\ldots (3)$$

Substituting (2) in (3) we have

$$y_n = y_1 + \triangle_1 \frac{r^{\frac{x_n}{d}}}{(r -- 1)\, r^{\frac{x_1}{d}}} - \frac{\triangle_1}{r-1} \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Letting $y_1 - \frac{\triangle_1}{r-1} = \mathrm{M}$

$$\frac{\Delta_1}{(r-1)\, r^{\frac{x_1}{d}}} = \mathrm{N}$$

$$r^{\frac{1}{d}} = \mathrm{P}$$

we have

$$y = \mathrm{M} + \mathrm{N}\,\mathrm{P}^x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

transforming to parallel axes through (*o*, N,) we have

$$y = \mathrm{N}\,\mathrm{P}^x$$

Taking logarithms this gives

$$log\ y = log\ \mathrm{N} + x\ log\ \mathrm{P}.$$

Using new constants this may be written

$$y = \mathrm{K} + \mathrm{H}x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

a logarithmic line.

Resuming equation (5) we may look upon this as an approximate formula only, considering it as written

$$y^1 = \mathrm{M} + \mathrm{N}\,\mathrm{P}^x$$

while the exact formula is

$$y = (\mathrm{M} + \mu) + (\mathrm{N} + \nu)(\mathrm{P} + \rho)^x \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where μ, ν, and ρ are small corrections to be determined. Then using Taylor's Theorem and dropping second powers

$$y = y_1 + \frac{\delta y^1}{\delta \mathrm{M}} \mu + \frac{\delta y^1}{\delta \mathrm{N}} \nu + \frac{\delta y^1}{\delta \mathrm{P}} \rho \ \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

since $y^1 = \mathrm{M} + \mathrm{N}\,\mathrm{P}^x$

then $\frac{\delta y^1}{\delta \mathrm{M}} = 1$; $\frac{\delta y^1}{\delta \mathrm{N}} = \mathrm{P}^x$; $\frac{\delta y^1}{\delta \mathrm{P}} = \mathrm{N}x\,\mathrm{C}^{x-1}$

this gives in (8)

$$y = y^1 + \mu + \nu\,\mathrm{P}^x + \rho\,\mathrm{N}x\,\mathrm{P}^{x-1} \ldots\ldots\ldots\ldots$$

rewrite this $y - y^1 = \mu + P^x v + N x\, P^{x-1} \rho \ldots\ldots\ldots\ldots\ldots (9)$ and substituting the observed values for y & x in this equation and for y^1 use the value found by substituting the observed values in form (5).

We can now form the normal equations for the determination of μ, ν and ρ.

The normal equations are

$$[(y - y^1)] = [\mu] + \nu [P^x] + \rho [N\, P^{x-1}]$$
$$[(y - y^1)] = \mu [P^x] + \rho [(P^x)(N\, P^{x-1})]$$
$$[(y - y^1)\, N\, P^{x-1}] = \mu [N_x P^{x-1}] + \rho [(N\, P^{x-1})^2]$$

The accuraey of this result depends largely upou the values of μ, ν and ρ for this corrected formula assumes them to be so small that their squares may be neglected.

Owing to the limits to this article the writer has been obliged to omit the treatment of periodic functions.

The reader who desires a short and concise treatment of such functions will find the same in T. W. Wright's *Treatise on the Adjustment of Observations.*

The readers attention is now called to a direct method of treating this subject without the use of the method of Least Squares.

Professor Karl Pearson has taken up the matter of empirical formulæ[1] and treated this subject from an entirely different standpoint. His results offer a systematic method of treating this snbject which yields sensibly as accurate results as does the method of least squares, with, in most cases, less labor, and which treats in a simple and direct manner forms that are prohibitive practically, on account of the difficulty of reducing to linear form.

In his paper upon this subject no attention is given to the form of the curve to be chosen, further than the calling attention to the unfortunate assumption so often made, that the parabola

$$y = a + b\,x + c\,x^2 + d\,x^3 + \ldots\ldots\ldots\ldots$$

is always a good form to choose.

[1] On the Systematic Fitting of Curves, Biometrika, Vol. I, Parts III and IV, C. J. Clay & Sons, Pub's., Cambridge, Eng.

Professor Pearson shows that for a function representing the fecundity of brood mares the formula

$$y = k\left(1 + \frac{x}{a}\right)^{p} e^{-\frac{px}{a}}$$

with only three constants, p, a and k at his disposal gives a better *fit* to the actual curve than does *seven* constants disposed in the form

$$y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6.$$

This should be kept well in mind when one considers the often made statement that the simple increase of arbitrary constants will be sufficient to give any degree of accuracy desired. It is not so much the number as the form in which they are combined.

I shall now attempt to briefly outline Professor Pearson's method, together with an application made by him to one particular instance. The following is much shortened and any one desirous of doing much in this work should have the full text of the articles cited.

By an application of the calculus of variations the following result is obtained, that

"*To fit a good theoretical curve* $y = \phi(x, c_1, c_2, c_3, \ldots\ldots c_n)$ *to an observed curve, express the area and moments of the curve for the given range of observation in terms of* $c_1, c_2, c_3, \ldots\ldots c_n$ *and equate these to the like quantities for the observations.*"

In forming these moments they may be taken about any line parallel to the axis y and thus they may be in some cases much simplified.

To make use of the above theorem one must know how to find the moments of any system of observed quantities and also to find the moments of the theoretical curve in terms of $c_1, c_2, c_3, \ldots\ldots c_n$.

The common case in physical work will be when a series of actual measurements have been made of some number, r say, of the ordinates of the actual curve.

Often however, the actual observations may represent the areas of small base elements r in number.

An example of this would be the number of deaths from smallpox in each year. If we increase the number of elements then the latter case above becomes practically the same as the former.

In the application of this method it becomes necessary to find the area and moments of the observed curve in which in the first case r ordinates are measured.

We will represent by Am_n the n^{th} moment of the area, where A is the area. Then what is wanted is

$$Am_n = \sum_{x_0}^{x_r} y\, x^n dx$$

the solution of this requires a good quadrature formula and a selection of these may be found in Boole's Calculus of Finite Differences, pp. 46, 54 and 98.

We may consider the area in two ways: as divided by the observed ordinates into small trapezia with the ordinates entire forming the bounding parallel sides of these, or as forming the ordinates half way between the sides, that is mid ordinates.

In the first case the observed or boundary ordinates y_0, y_1, $y_2, \ldots\ldots. y_r$ and in the second case $y_{\frac{1}{2}}\ y_{\frac{3}{2}} \ldots . y_{r-\frac{3}{2}}\ ,\ y_{r-\frac{1}{2}}$ Let these ordinates be taken at equal distances b this will in general be the base unit and therefore unity.

The following quadrature formula is due to Simpson, where $2\,r$ = number of elements:

$$\sum_{y_0}^{y_r} y\, dx = \tfrac{1}{3} b \left\{ \begin{array}{l} y_0 + 2(y_2 + y_4 + \ldots . y_{2r-2}) + y_{2r} \\ \qquad + 4(y_1 + y_3 + \ldots . y_{2r-1}) \end{array} \right\} \ldots\ldots (1)$$

Weddle's Rule 6 r elements

$$\sum_{y_0}^{y_{6r}} y\, d\, x = \tfrac{3}{10} b \left\{ \begin{array}{l} y_0 + y_2 + y_4 + y_8 + y_{10} + \ldots y_{6r-2} + y_{6r} \\ \qquad + 2(y_6 + y_{12} + \ldots . y_{6r-6}) \\ \qquad + 5(y_1 + y_5 + y_7 + \ldots . y_{6r-1}) \\ \qquad + 6(y_3 + y_9 + y_{15} + \ldots . y_{6r-3}) \end{array} \right\} \ldots\ldots (2)$$

The following two quadrature formulæ are due to Mr. W. F. Sheppard, and their proofs are given in *L. Math. Proc.*, Vol. 32, p. 270.

In each case they use few differences and apply whatever be the number of elements used.

Case 1. When the boundary ordinates are known, making use of two differences:

$$Area = \mathrm{A}c + \frac{1}{120}\frac{r(15r-26)}{(r-1)(r-2)}\left\{(y-y_0)-(y_r-y_{r-1})\right\}b$$

$$-\frac{1}{120}\frac{r(5r-6)}{(r-2)(r-3)}\left\{(y_2-y_1)-(y_{r-1}-y_{r-2})\right\}b \qquad \ldots\ldots\ldots(3)$$

Case 2. Mid ordinates known, two differences:

$$Area = \mathrm{A}t - \frac{1}{960}\frac{(80r-177)}{(r-2)(r-3)}\left\{(y_{\frac{3}{2}}-y_{\frac{1}{2}})-(y_{r-\frac{1}{2}}-y_{r-\frac{3}{2}})\right\}b$$

$$+\frac{1}{960}\frac{r(40r-57)}{(r-3)(r-4)}\left\{(y_{\frac{5}{2}}-y_{\frac{3}{2}})-(y_{r-\frac{3}{2}}-y_{r-\frac{5}{2}})\right\}b$$

Here $\mathrm{A}c = b(\frac{1}{2}y_1+y_2+y_3+\ldots\ldots\ldots y_{r-1}+\frac{1}{2}y_r)$,

$\mathrm{A}t = b(y_{\frac{1}{2}}+y_{\frac{3}{2}}+\ldots\ldots\ldots\ldots\ldots y_{r-\frac{3}{2}}+y_{r-\frac{1}{2}})$,

The above formulæ are in general quite accurate enough, and with them the area and moments may all readily be expressed.

The following is an application of this method to fit Makeham's Curve to mortality statistics made by Prof. Pearson in the article cited above.

The problem is as follows: Given a mortality table which gives the number of survivors from n people born the same year at each year of the age of the group.

Then if l_x denote the number who live to the age of x the table will, between say the ages 25 and 85, be closely represented by

$$l_x = \mathrm{K}\, s^x\, (g)^{c^x} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

where K, s, g, and c are constants to be determined from the table.

Taking logarithms and representing them by the large letters primed we have,

$$L'_x = K' + S'x + G'c^x \text{ .... } (6)$$

a form which not being linear is not in form for treatment by least squares. Because of the number of terms some sixty taking from 25 to 85 the ordinary methods of treating this problem are very unsatisfactory.

Take l for the range of the table with the origin of x' at the middle of the range. Let x_0 be the age corresponding to the origin and write

$$lx' = K' s' \frac{x'}{l} (g) c^{\frac{x'}{l}} \text{ .... } (7)$$

$$\text{that is} \quad \left.\begin{array}{l} s' = sl \\ c' = cl \\ g = g c^{x_0} \\ k' = k s^{x_0} \end{array}\right\} \text{ .... } (8)$$

Taking logarithms the formula may be written

$$L = K + S \frac{x}{l} + G e^{\frac{2nx}{l}} \text{ .... } (9)$$

$$\text{where } e^{2n} = c'$$

Since we have four constants to be found it will be necessary to find the area and the first *three* moments of (9) about the middle of the range; that is we want A , Am_1 , Am_2 and Am_3. These will then be equated to the same quantities found by quadrature formulæ from the mortality table.

This will give direct equations for finding K, S, G and n. Now

$$A = \Sigma_{-\frac{l}{2}}^{+\frac{l}{2}} L\, dx = K\, l + \frac{G\, l}{2\, n} (e^n - e^{-n})$$

$$\text{let} \quad a_0 = \frac{A}{l}$$

$$\text{then} \quad a_0 = K + \frac{G \sinh n}{n} \text{ .... } (10)$$

$$\mathrm{A}m_1 = \Sigma_{-\frac{l}{2}}^{\frac{l}{2}} \mathrm{L}\, x\, d_{re}$$

and if $a_1 = \dfrac{12\ \mathrm{A}m_1}{l^2}$

then $a_1 = \mathrm{S} + 6\,\mathrm{G} \left\{ \dfrac{cosh}{n} - \dfrac{sinh\ n}{n^2} \right\}$ ....(11)

$$\mathrm{A}m_2 = \Sigma_{-\frac{l}{2}}^{+\frac{l}{2}} \mathrm{L}\, x^2\, dx$$

and if $a_2 =. \dfrac{12\, a\, m_2}{l^3}$

then $a_2 = \mathrm{K} + 3\,\mathrm{G} \left\{ \dfrac{sinh\, n}{n} - \dfrac{2\ cosh\ n}{n^2} + \dfrac{6\ sinh\ n}{n^3} \right\}$ (12)

$$\mathrm{A}m_3 = \Sigma_{-\frac{l}{2}}^{+\frac{l}{2}} \mathrm{L}\, x^3\, dx$$

and if $a_3 = \dfrac{80\ \mathrm{A}m_3}{l^4}$

then

$$a_3 = \mathrm{S} + 10\mathrm{G}. \left\{ \frac{cosh\ n}{n} - \frac{3\ sinh\ n}{n^2} + \frac{6\ cosh\ n}{n^3} - \frac{6\ sinh\ n}{n^4} \right\}$$

........(13

Subtracting (11) from (13) and (10) from (12) we may find

$$(14) \ldots\ \frac{a_3 - a_1}{a_2 - a_o} = \frac{4\left(\dfrac{20\, m_3}{l^3} - \dfrac{3\, m_1}{l} \right)}{\dfrac{12 m^{\,2}}{l^2} - 1}$$

$$= 10 \left\{ \frac{\dfrac{cosh\ n}{5n} - \dfrac{6}{5}\,\dfrac{sinh\ n}{n^2} + \dfrac{3\ cosh\ n}{n^3} - \dfrac{3\ sinh\ n}{n^4}}{\dfrac{sinh\ n}{n} - \dfrac{3\ cosh\ n}{n^2} + \dfrac{3\ sinh}{n^3}} \right\}$$

write this equal to H and it readily reduces to

$$tanh\ n = \frac{2\, n^3 + 30\, n + 3\, \mathrm{H}\, n^2}{\mathrm{H}\, n^3 + 12\, n^2 + 3\, \mathrm{H}\, n + 30},$$

and substituting for the hyperbolic tangent its value we have

$$(15)\ldots\ldots e^{2n}=\frac{(\mathrm{H}+2)n^3+3(\mathrm{H}+4)n^2+3(\mathrm{H}+10)n+30}{(\mathrm{H}-2)n^3-3(\mathrm{H}-4)n^2+3(\mathrm{H}-10)n+30}$$

Equation (12)— (10) gives the value of G
Equation (11) gives S
Equation (12) gives K.
The constants then may be all found.

For solving (15) use Newton's method of approximations. See *Burnside & Panton's Th. of Alg. Equations.*

An approximate value of n may be readily found which we will call $n_\circ$.

It is known that the *log* $c=e^{\frac{2n}{\iota}}$ from experience is not far from .04, so use this value for n_0.

Substitute $n_\circ+h$ for n in (15) discarding h^2 and higher powers.

Putting N for numerator and D for denominator and writing $e^{2n}=\frac{\mathrm{N}}{\mathrm{D}}$ we find at once since $h=\frac{f\ (n_0)}{f'(n_0)}$

$$h=\frac{e^{2n}-\frac{\mathrm{N}_0}{\mathrm{D}_0}}{\frac{1}{\mathrm{D}_\circ}\left(\frac{d\,\mathrm{N}}{dn}\right)_0-\frac{\mathrm{N}_\circ}{\mathrm{D}_\circ}\left(\frac{d\,\mathrm{D}}{d\,n}\right)-2\left(\frac{\mathrm{N}_0}{\mathrm{D}_0}\right)-2\left(e^{2n}-\frac{\mathrm{N}_0}{\mathrm{D}_0}\right)}$$

The subscript indicates the value after n is replaced by $n_\circ$. If we put

$$\left.\begin{aligned}\mathrm{Z}&=2\,n^3+3\,\mathrm{H}\,n^2+30\,n\\ \mathrm{Y}&=\mathrm{H}\,n^3+12\,n^2+3\,\mathrm{H}\,n+30\end{aligned}\right\}\ \ldots.(16)$$

then

$$e^{2n}=\frac{\mathrm{Y}+\mathrm{Z}}{\mathrm{Y}-\mathrm{Z}}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(17)$$

$$\mathrm{N}=\mathrm{Y}+\mathrm{Z}$$
$$\mathrm{D}=\mathrm{Y}-\mathrm{Z}$$

and $\frac{d\,\mathrm{N}}{dn}=\frac{dy}{dn}-\frac{d\,\mathrm{Z}}{dn}$

$$\frac{d\,\mathrm{D}}{dn}=\frac{dy}{dn}+\frac{d\,\mathrm{Z}}{dn}$$

forming the differentials in (16) we readily find e^{2n} and all the quantities are now entirely determined.

Prof. Pearson using the mortality tables given in the Textbook for Actuaries taking the years from 25 to 85 and using Weddle's Quadrature rule (2) above found,

$$\begin{aligned} A &= 221.843235 \\ Am_1 &= 275.103222 \\ Am_2 &= 64464.355986 \\ Am_3 &= 162062.316564 \end{aligned}$$

$$a_\circ = 3.69738725\dot{0}, \qquad a_2 = 3.581353110\dot{3}$$

$$a_1 = -\ .91701074\dot{0}, \qquad a_3 = -\ 1,00038467\dot{0}\dot{1}4\dot{8}.$$

these give $H = .718529308595$

using the method indicated, as the eighth approximation he found the value

$$n = 2.807343873$$

which is correct to the last figure.

This gives

$$\sinh n = 8.252746794593,$$

$$\cosh n = 8.313111911675,$$

Eq (12) — (10) gives

$$G = -\ .064875005350$$

(11) gives $\dfrac{S}{l} = -\ .002866074767$

and (10) gives $K = 3,888100258$

and $c = e^{\frac{2n}{l}} = 1.098096393273$

a result that Professor Pearson believes correct to the last figure. The formula then for L_x the number of survivors of age $55 + x$ years is:

$$L_x = 3.888100258 - .002866074767\,x$$
$$- .064875005350(1.098096393273)^x \quad \ldots\ldots\ldots (17)$$

The mean difference between the values of the computed value and the observed value of L_x as taken from the mortality table is only .00116, and the method gives as this solution shows a definite method of treating this complex mathematical form. Many functions yield themselves much more readily than does this. The method of moments may be looked upon as giving great aid in the work of determining empirical formulæ.

The improvement of empirical formulæ that have been approximately determined. When a formula has been obtained that seems to be of the right form and yet does not for interpolated values give satisfactorily accurate results, the formula may often be bettered by the use of one of the three following methods:

1st. By bettering or correcting the constants.

2nd. By increasing the number of terms.

3rd. By the substitution of a new function of x for x or some of the constants.

Let $y = F(a, b, c \ldots . x)$ be the desired formula and suppose $\psi = F(A, B, C, \ldots . x)$ to be an approximate formula, which has been derived from the given conditions and observations. Let $\alpha, \beta, \gamma, \ldots .$ be the corrections to A, B, C,.... which would give us the true form that is

$$y = F(A + \alpha\ , B + \beta, , C + \gamma \ldots \ldots . x)$$

we shall suppose in the following develonment that α, β, γ.. are so small that the second powers may be discarded, although in many cases in practice this will be found far from true.

Developing the above expression by Taylor's theorem and discarding higher powers we have

$$y = F(A + \alpha\ , B + \beta + \ldots \ldots \ldots \ldots)$$

$$= F(A, B, C \ldots . x) = \frac{\delta F}{\delta A}\alpha + \frac{\delta F}{\delta B}\beta + \frac{\delta F}{\delta C}\gamma + \ldots \ldots$$

or this may be written

$$y = F + \frac{\delta F}{\delta A}\alpha + \frac{\delta F}{\delta B}\beta + \frac{\delta F}{\delta C}\gamma + \ldots \ldots .$$

Consider the previous form given

$$y = n + \frac{K}{m - x}$$

Suppose the approximate form has been found

$$F = A + \frac{B}{C - x}$$

the corrected formula becomes at once

$$y = F + \frac{\delta F}{\delta A}\alpha + \frac{\delta F}{\delta B}\beta + \frac{\delta F}{\delta C}\gamma$$

where

$$\frac{\delta F}{\delta A} = 1 \ ; \ \frac{\delta F}{\delta B} = \frac{1}{C - x} : \frac{\delta F}{\delta C} = \frac{-B}{(C - x)^2}$$

then

$$y = A + \frac{B}{C - x} + \alpha\frac{B}{C - x} - \frac{B\gamma}{(C - x)^2}$$

Now substituting the successive observed values of y and x the quantities α, β and γ may be found.

This process may be repeated any number of times until the corrections become so small as to cease repaying the considerable amount of labor which each application requires. This method may be applied very successfully in many cases. The particular form $y = A + \frac{B}{C - x}$ is to be avoided when possible, owing to the great difficulty in treating the term involving $\frac{1}{(C - x)^2}$ by the method of least squares, owing to the function not being linear.

A formula may frequently be improved by the addition of another term as the addition of a new constant increases the number of fundamental conditions which the curve must satisfy and thus it is more definitely defined. When the new term is looked upon merely as a correction to the form already determined it may readily be found by considering the values of the constants already found as the true values and writing the *one normal* equation from which the new constant may be found. For example, let the formula already found be

$y = A + Bx + Cx^2$ and let Vx^3 be the new term that is to be added, writing the equation then

$$y = A + Bx + Cx^2 + Vx^3$$

and introducing the observed values of x and y we form the normal equation for V

$$[x^3]\,y = A\,[x^3] + B\,[x^4] + C\,[x^5] + V\,[x^6]\,,$$

which gives the value of V at once.

This is an easy method of adding a small correction which may render a form satisfactory that did not give results as accurate as may have been desired.

An empirical form may again be bettered by the substitution of a new function of x in place of adding additional terms.

We will suppose that the approximate form $y = F(x)$ has been obtained and interpolating values of the observed x we find a series of values y_1' , y_2' , y_3' .... instead of the observed values y_1 , y_2 , y_3........ or we have the series of differences

$$y_1 - y_1' = d_1$$
$$y_2 - y_2' = d_2$$
$$y_3 - y_3' = d_3$$
$$\dots\dots\dots\dots$$
$$\dots\dots\dots\dots$$

Then there is required an additive function in the nature of a corection that may be added to F(x) and which we will represent by $f(x)$, such that when $x = x_n$ $\qquad f(\mathrm{x}) = d_n$

The form is now

$$y = F(x) + f(x)$$

Now as to the form of $f(x)$ it is entirely arbitrary. We have simply to determine an empirical formula considering $x_1\ x_2\ x_3$ ....x_n and d_1, d_2, d_3....d_n as the observed quantities and everything that has been said before this concerning such formulæ will apply with full force in this case.

The above cases are of such simple application that no concrete case will be given because of lack of space. Other methods such as the breaking up of one series into two arithmetic series of the first or higher orders can not be here considered although in some cases this offers a neat solution of the difficulty.

As a simple case of bettering a desired formula by the addition of another term we will consider the formula already established connecting the time of vibration of a pendulum and its length.

The approximate formula has already been found

$$l = -182.741\,t^3 + 941.884\,t^2 - 1505.64\,t + 805.022 \ldots\ldots (1)$$

and having given the observed values

l	t
37.88	1.24
46.65	1.37
72.75	1.72
84.05	1.83
88.35	1.885
99.35	2.00

Assuming the approximate formula above and introducing literal cöefficients we have the observation equations

$$\left.\begin{array}{l} l = A + B\,t_1 + C\,t_1^2 + D\,t_1^3 + V\,t_1^4 \\ l = A + B\,t_2 + C\,t_2^2 + \ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ l = A + B\,t_6 + C\,t_6^2 + D\,t_6^3 + V\,t_6^4 \end{array}\right\}\ldots\ldots(2)$$

where V is the new cöefficient of the correctional term, and A, B, C, D, are the cöefficients in (1) above. The normal equation for V will be

$$[l\,t^4] = A[t^4] + B[t^5] + C[t^6] + D[t^7] + V[t^8] \ldots\ldots\ldots\ldots (3)$$

substituting the observed values for l and t in the equation (2) and then forming the multiplications and summations indicated in (3) we find.

$$4550.02183625 = 4570.081354983 + V\,605.706756$$

$$\text{or} \qquad V = -.033117$$

and the corrected form now becomes

$$l = 805.022 - 1505.64t + 941.884t^2 - 182.741t^3 - .033117t^4$$

Substituting the observed values we find

for	$t = 1.24$	$l = 37.218$	instead of	$l = 37.85$
"	$t = 1.37$	$l = 40.273$	" "	$l = 46.65$
"	$t = 1.51$	$l = 52.292$	" "	$l = 56.10$
"	$t = 1.65$	$l = 64.714$	" "	$l = 63.55$
"	$t = 1.72$	$l = 72.85$	" "	$l = 72.75$
"	$t = 1.77$	$l = 76.243$	" "	$l = 78.05$

for	$t = 1.83$	$l = 83.68$	instead	of	$l = 84.05$
“	$t = 1.89$	$l = 89.71$	“	“	$l = 88.35$
“	$t = 1.95$	$l = 95.55$	“	“	$l = 94.95$
“	$t = 2.00$	$l = 98.82$	“	“	$l = 99.35$

In conclusion, it may, I think, be said that the most important thing in this work is the choice of the best form. That the choice of a parabola even of high order may not be as accurate as some other form using fewer constants. That the method of least squares is very long and tedious and in many cases is because of its being so cumbersome, practically useless. That the development of some method like the method of Moments by Professor Pearson is very desirable.

For further developments of this work the reader is referred to

Steinhauser's Aufstellung Empirishes Formeln; T. W. Wright's Treatise on Adjustment of Observations; Merriman's Method of Least Squares; Karl Pearson, Vol. I, Part 3, and Vol. II, Part 1, of Biometrika. Also numerous articles in other magazines.