On the Distribution of the Error of an Interpolated Value, and on the Construction of Tables. By R. A. FISHER, Sc.D., Gonville and Caius College, and Mr J. WISHART.

[*Received* 8 August, *read* 24 October 1927.]

1. *Introduction.*

Before the introduction of interpolation formulae, beyond linear interpolation by proportional parts, the presentation of the numerical values of mathematical functions was much restricted, for the labour of computation and the cost of printing, to say nothing of the inconvenience of handling a bulky volume, had to be increased quite disproportionately with every increase in accuracy. A four-figure logarithm table occupies two small pages, Chambers's seven-figure table takes 150 pages, while Vega's ten-figure table requires 300 pages twelve inches long.

The modern tendency towards compact tables used with more or less high order interpolation formulae arises from two advances: (i) the introduction of simple and adequate interpolation formulae, and (ii) the development of the theory of the remainder term. The central difference formulae of Everett not only reduce the calculation of the interpolate to a few very simple operations, of a kind suitable for machine calculations, but since they require only even differences they enable adequate differences to be presented compactly with the table. Tables of the coefficients of the second, fourth and sixth differences have been prepared by A. J. Thompson (1, 1921). Even a moderate use of differences results in a very great saving. For example in A. J. Thompson's recent twenty-figure logarithm tables (2) the values from ·9 to 1·0 occupy only 100 pages, although differences beyond the second are not required, the fourth differences being negligible. The table could in fact have been made much more compact, by the use of higher differences, although in view of the special properties of the logarithm, it is doubtful if this would be desirable. Pearson has exhibited seven-figure logarithm and anti-logarithm tables which, using second differences, together occupy only a single page (3, 1920, p. 60).

Steffensen's recent book on interpolation (4, 1927) makes available in English the author's important work on the remainder term, and so makes it possible to gauge how far high order interpolation may be carried with advantage. The contributions to the interpolate of the tabular entries and of their second, fourth, etc., differences will usually form an asymptotic series, the magnitude of the terms falling to a minimum and then increasing. This minimum should be a negligible fraction in the last decimal place tabulated, and if so an interpolation formula will give the interpolate to the full accuracy of the table. A table should in fact always be regarded as giving not an isolated series of values of a function, but, by means of the interpolate function implicit in these, the whole continuum of its values. The possibility of adequate interpolation thus depends on the tabular intervals being sufficiently small for the remainder term to be negligible. This condition allows of the use of immensely more compact tables than the requirement that linear interpolation shall be valid, and brings with it the possibility of accurate tables of double and triple entry, without prohibitive labour either in their preparation or their use.

When the remainder term is negligible the error of the interpolate will depend only on the tabular errors, that is to say, on the differences between the exact value of the function and the corresponding value given in the table. When a table is "correct to the last figure" this error will be uniformly distributed between the limits $-0{\cdot}5$ and $+0{\cdot}5$ in the last place. Disproportionate efforts are often made to ensure this accuracy of the last place. Calculations are frequently carried to two or three figures more than are given in the final form, and the entries then cut down. Even so, it is by no means certain, owing to entries occurring that end in 50 or 500, that the final figure will in all cases be "correct"; and a lengthy series of calculations in such a doubtful case will seldom effect more improvement than to replace an error of say ·503 by one of ·497, a most trifling gain seeing that the standard error of a "correct" table is ·288. It has not been hitherto sufficiently realised that a table to a larger number of places, the last one or two of which are not necessarily correct, is capable of giving a more accurate interpolate than the same table cut down, provided that the standard error of the tabular entry is known. The expense of printing one or more figures would be more than compensated for by the increased accuracy obtainable. It is usually possible to ensure that positive and negative errors should be, in the long run, equally frequent, and that the distribution of errors shall follow approximately the normal law.

The purpose of this note is to establish the extent and nature of the errors to which the interpolate is liable, in so far as these depend only on tabular errors, in terms of those of the entries from which it is derived, whether in a table correct to the last figure or not. A statement of this error, which need not be given more than once on each page, would be a necessary and valuable addition to the many-place table that has been advocated above. It will be assumed in what follows that the standard deviations of tabular errors are all equal, within the range of the interpolation formulae, and that errors in adjacent values are independent of one another, i.e. no correlations exist between ϵ_1 and ϵ_2, ϵ_1 and ϵ_3, etc., where ϵ denotes the error of a single value. All the formulae of interpolation by differences are included in the following general reasoning, where Lagrange's formula has been used, and the intervals assumed, for simplicity, to be equal.

2. *The Mean Variance of the Interpolate.*

Let p be the fraction of the interval (taken as 0 to 1) for which it is desired to obtain an interpolated value. Let $q = 1 - p$. Then we shall examine the contribution of any one error to the variance of the interpolate for one-, two-, three-, four-, ... point formulae and average this over the interval. The total contribution will be expressed as a polynomial in pq and we shall use the result that

$$\int_0^1 (pq)^r \, dp = \frac{(r\,!)^2}{(2r+1)\,!}.$$

1-*point Interpolation.*

The interpolate is here equal to the tabular entry, and so the ratio of the variance of the interpolate to that of the entry is unity.

2-*point Interpolation.*

We have contributions $q\epsilon$ at 0 and $p\epsilon$ at 1. The variance of the interpolate is then the average value of

$$(p^2 + q^2)\,\sigma^2 = (1 - 2pq)\,\sigma^2,$$

and our required ratio is $\frac{2}{3}$.

3-*point Interpolation.*

Contributions to the error are

$$\frac{(p-\frac{1}{2})(p-\frac{3}{2})}{2} \text{ at } -\tfrac{1}{2}, \quad (p+\tfrac{1}{2})(p-\tfrac{3}{2}) \text{ at } +\tfrac{1}{2},$$

and

$$\frac{(p+\frac{1}{2})(p-\frac{1}{2})}{2} \text{ at } +\tfrac{3}{2}.$$

We then have the total variance

$$= \left[\frac{(2p-1)^2}{64}\{(2p-3)^2 + (2p+1)^2\} + \frac{(2p+1)^2(2p-3)^2}{16}\right]\sigma^2$$

$$= \frac{\sigma^2}{32}\,[(1-4pq)(5-4pq) + 2\,(3+4pq)^2],$$

and the average value of the variance of the interpolate is

$$\frac{429}{480}\,\sigma^2,$$

the ratio of variances being ·89375.

A similar procedure gives, for 4-point interpolation, ·77566, 5-point ·89574, 6-point ·82244, 7-point ·90291, and 8-point ·84942.

3. *The Limiting Condition for High Order Interpolation.*

It is easy to show that the limit to the ratio of variances as the number of points is indefinitely increased must be unity. For the function

$$\frac{(x-a_1)(x-a_2)\ldots\ldots(x-a_{p-1})(x-a_{p+1})\ldots\ldots(x-a_n)}{(a_p-a_1)(a_p-a_2)\ldots\ldots(a_p-a_{p-1})(a_p-a_{p+1})\ldots\ldots(a_p-a_n)}$$

vanishes at all points a_1, a_2, ... a_n except a_p, at which its value is unity. When n is large, and the intervals equal to unity, the function tends to resemble

$$\frac{\sin\pi(x-a_p)}{\pi(x-a_p)}.$$

The average contribution of any one error to the variance of the interpolate is then

$$\int_{-a_p}^{1-a_p}\frac{\sin^2\theta}{\theta^2}\,dx, \qquad \text{where } x=\frac{\theta}{\pi},$$

$$=\frac{1}{\pi}\int_{-\theta_p}^{\pi-\theta_p}\frac{\sin^2\theta}{\theta^2}\,d\theta,$$

and the average value of the total variance, expressed as a ratio to that of the tabular entry, will be

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin^2\theta}{\theta^2}\,d\theta=-\frac{1}{\pi}\left[\frac{\sin^2\theta}{\theta}\right]_{-\infty}^{\infty}+\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{2\sin\theta\cos\theta}{\theta}\,d\theta,$$

of which the first term vanishes, while the second is

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin\phi}{\phi}\,d\phi=1. \qquad \ldots\ldots(3{\cdot}1)$$

We have, then, the result that the average standard error of the interpolate is always somewhat less than that of the tabular entry, being least for linear interpolation, and the advantage of a system of tabulation which is not concerned with the accuracy of the last figure recorded, but which states the standard error of entry, is apparent.

For certain functions it is possible to tabulate better values for the even differences than those derived from the tabular values. If exact values for the differences are employed in the interpolation formula, errors will only be introduced by the part calculated from the two adjacent tabular values, and the distribution of errors will be exactly the same as for 2-point interpolation. This condition is nearly realised when even differences are given correct to the last digit of the tabular values.

4. *The Error Distribution for High Order Interpolation.*

When we turn to examine not merely the variance but the form of the distribution of the error in the interpolate, a normally distributed tabular error is seen to have an added advantage beyond that due to the increased accuracy attainable when the table is not cut down to be "correct to the last place." In what follows n, the number of points, will be assumed to be large, and the theory put forward may be regarded as that to which we tend as the number of differences used in interpolation is increased. It will first be necessary to find the average values of the powers of the several errors. For the average value of the 4th powers of the tabular errors we require

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin^4\theta}{\theta^4}\,d\theta=\frac{1}{3!\,\pi}\int_{-\infty}^{\infty}(2^3-4)\frac{\sin\theta}{\theta}\,d\theta=\frac{2}{3}. \quad \text{...(4·1)}$$

Similarly $$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin^6\theta}{\theta^6}\,d\theta=\frac{1}{5!}(3^5-6.2^5+15)=\frac{11}{20}, \quad \text{......(4·2)}$$

while for the eighth, tenth, and twelfth powers we have the average values

$$\frac{151}{315},\ \frac{15619}{36288}\ \text{and}\ \frac{655177}{1663200}. \qquad \text{......(4·3)}$$

The general term for this average can be written

$$\frac{1}{2}\cdot\frac{1}{(2r-1)!}\,\delta^{2r}u_0,$$

where $u_n=|\,n\,|^{2r-1}$.

Alternatively the actual values for particular points of interpolation, and not merely the averages over all points, of the mean powers of the errors may be reached by noting that the series of coefficients of the independent tabular errors, in the error of the interpolate, is

$$\frac{\sin\theta}{\theta},\ \frac{\sin(\theta\pm\pi)}{\theta\pm\pi},\ \ldots\ldots,$$

so that the sum of their squares is 1 if $\theta=0$, and

$$\frac{1}{\pi^2}\sin^2\theta\ \underset{r=-\infty}{\overset{\infty}{S}}\left(\frac{1}{x+r}\right)^2;\ \theta\neq 0,$$

where $$x=\frac{\theta}{\pi},$$

and r is an integer. We have

$$\overset{\infty}{\underset{r=0}{S}} \frac{1}{(x+r)^2} = \frac{\partial^2}{\partial x^2} \log (x-1)\,!.$$

Also
$$\overset{-1}{\underset{r=-\infty}{S}} \frac{1}{(x+r)^2} = \frac{\partial^2}{\partial x^2} \log (-x)\,!.$$

[$(x-1)\,!$ is written for $\Gamma(x)$ even though x is not an integer.] Adding, we have

$$\overset{\infty}{\underset{-\infty}{S}} \frac{1}{(x+r)^2} = \frac{\partial^2}{\partial x^2} \log \frac{\pi}{\sin \pi x} = \pi^2 \operatorname{cosec}^2 \theta.$$

Thus the sum of squares is unity for all values of θ. The sum of the fourth powers is, similarly,

$$\frac{1}{3\,!} \sin^4 \theta \frac{\partial^2}{\partial \theta^2} (\operatorname{cosec}^2 \theta) = 1 - \tfrac{2}{3} \sin^2 \theta. \qquad \text{......(4·4)}$$

For the sixth powers we have

$$1 - \sin^2 \theta + \tfrac{2}{15} \sin^4 \theta, \qquad \text{......(4·5)}$$

and for the eighth powers

$$1 - \tfrac{4}{3} \sin^2 \theta + \tfrac{2}{5} \sin^4 \theta - \tfrac{4}{315} \sin^6 \theta. \qquad \text{......(4·6)}$$

These expressions are the coefficients of h^2, h^4, etc., in the expansion of

$$hx \cot (\theta - hx),$$

where $x = \sin \theta$.

The assumption is made that the error distribution of the interpolate will be symmetrical, so that only even moments need be considered. The two cases that arise are (a) normal distribution of tabular errors, which occur when the table is used as constructed, and not cut down, (b) rectangular distribution of errors, as in the case of all ordinary tables, which are correct to the last figure. So that in (b) if dp represent the small element of frequency ($=f(x)\,dx$), we have $dp = dx$ for all values of x between $-\frac{1}{2}$ and $+\frac{1}{2}$, say, and $dp = 0$ outside these limits.

The nature of the distributions is best exhibited by studying their characteristic functions. If, in the usual notation,

$$\mu_r' = \int_{-\infty}^{\infty} x^r \, dp$$

is the rth moment of the distribution (taken as unlimited in range) about an arbitrary origin, and μ_r is the corresponding moment about the mean, then the Moment Generating Function is defined as

$$M = \int_{-\infty}^{\infty} e^{xt} dp = 1 + \mu_1' t + \mu_2' \frac{t^2}{2\,!} + \mu_3' \frac{t^3}{3\,!} + \ldots\ldots$$

$$= e^{\mu_1' t}\left(1 + \mu_2 \frac{t^2}{2\,!} + \mu_3 \frac{t^3}{3\,!} + \ldots\ldots\right). \quad \ldots\ldots(4{\cdot}7)$$

The logarithm of M may be called the Kappa Generating Function, and we have

$$\mathrm{K} = \log M = \kappa_1 t + \kappa_2 \frac{t^2}{2\,!} + \kappa_3 \frac{t^3}{3\,!} + \ldots\ldots, \quad \ldots\ldots(4{\cdot}8)$$

where the κ's are the semi-invariants of Thiele. Thus

$$\kappa_1 = \mu_1', \quad \kappa_2 = \mu_2, \quad \kappa_3 = \mu_3, \quad \kappa_4 = \mu_4 - 3\mu_2^2, \quad \kappa_5 = \mu_5 - 10\mu_2\mu_3, \text{ etc.}$$

An important property of the K-functions is that if the distribution of ϵ has for its characteristic function

$$\mathrm{K} = \kappa_1 t + \kappa_2 \frac{t^2}{2\,!} + \kappa_3 \frac{t^3}{3\,!} + \ldots\ldots,$$

then that of $p\epsilon$, where p is constant, has

$$\mathrm{K}_1 = p\kappa_1 t + p^2 \kappa_2 \frac{t^2}{2\,!} + p^3 \kappa_3 \frac{t^3}{3\,!} + \ldots\ldots.$$

Also for any linear compound of independent variates the K-functions are added, i.e. for

$$X = p_1 \epsilon_1 + p_2 \epsilon_2 + p_3 \epsilon_3 + \ldots\ldots,$$

where $p_1, p_2, p_3, \ldots$ are constant and $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$ are independent variates of a distribution given by K, then for the distribution X

$$\mathrm{K}_2 = \kappa_1 S(p)\, t + \kappa_2 S(p^2) \frac{t^2}{2\,!} + \kappa_3 S(p^3) \frac{t^3}{3\,!} + \ldots. \quad \ldots(4{\cdot}9)$$

This property enables us to reach very quickly the form of the distribution of the error of the interpolate for a given distribution of tabular errors. Thus for the rectangular distribution of common occurrence

$$M' = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{t'x} dx = \frac{2}{t'} \sinh \frac{t'}{2}$$

$$= 1 + \frac{1}{3\,.\,2^2} \cdot \frac{t'^2}{2\,!} + \frac{1}{5\,.\,2^4} \cdot \frac{t'^4}{4\,!} + \frac{1}{7\,.\,2^6} \cdot \frac{t'^6}{6\,!} + \ldots\ldots.$$

Adjusting so as to obtain a unit standard error by writing

$$t' = t\sqrt{12},$$

we have

$$M = 1 + \frac{t^2}{2\,!} + \frac{3^2}{5} \cdot \frac{t^4}{4\,!} + \frac{3^3}{7} \cdot \frac{t^6}{6\,!} + \ldots\ldots. \quad \ldots(4{\cdot}10)$$

Then $$K=\frac{t^2}{2\,!}-\frac{6}{5}\cdot\frac{t^4}{4\,!}+\frac{48}{7}\cdot\frac{t^6}{6\,!}-\frac{432}{5}\cdot\frac{t^8}{8\,!}+\ldots\ldots\quad\ldots(4\cdot11).$$

So for the interpolate, using the results (4·4), (4·5), (4·6), we have

$$K=\frac{t^2}{2\,!}-\frac{6}{5}\left(1-\frac{2}{3}\sin^2\theta\right)\frac{t^4}{4\,!}+\frac{48}{7}\left(1-\sin^2\theta+\frac{2}{15}\sin^4\theta\right)\frac{t^6}{6\,!}$$
$$-\frac{432}{5}\left(1-\frac{4}{3}\sin^2\theta+\frac{2}{5}\sin^4\theta-\frac{4}{315}\sin^6\theta\right)\frac{t^8}{8\,!}+\ldots\quad\ldots(4\cdot12)$$

If the successive terms are averaged, by the use of (4·1), (4·2), (4·3), we have what may be called the synthetic distribution of the interpolate, given by the function

$$K=\frac{t^2}{2\,!}-\frac{4}{5}\cdot\frac{t^4}{4\,!}+\frac{132}{35}\cdot\frac{t^6}{6\,!}-\frac{7248}{175}\cdot\frac{t^8}{8\,!}+\ldots,\quad\ldots(4\cdot13)$$

in which the terms beyond the first indicate the departure from normality. It will be seen that the distribution does not tend to normality, even though the number of tabular errors involved is increased without limit. Pearson's β_2 tends to 2·2 instead of 3·0. If the moments are wanted the procedure is reversed and we obtain

$$M=1+\frac{t^2}{2\,!}+\frac{11}{5}\cdot\frac{t^4}{4\,!}+\frac{237}{35}\cdot\frac{t^6}{6\,!}+\frac{4127}{175}\cdot\frac{t^8}{8\,!}+\ldots\quad\ldots(4\cdot14)$$

Alternatively we may consider the distribution compounded of the several distributions of the successive tabular errors. Thus if the frequency distributions are

$$dp=\phi_1dx,\quad \phi_2dx,\quad \phi_3dx,$$

and so on, we consider the distribution $dp=\bar{\phi}dx$, $\bar{\phi}$ being the mean of ϕ_1, ϕ_2, ϕ_3, ... The corresponding Moment Generating Function is

$$M=\int_{-\infty}^{\infty}e^{xt}\bar{\phi}\,dx=\bar{M}.$$

This procedure is equivalent to reversing equation (4·12) and then averaging the terms. Thus

$$M=1+\frac{t^2}{2\,!}+\frac{9+4\sin^2\theta}{5}\cdot\frac{t^4}{4\,!}+\frac{135+180\sin^2\theta+32\sin^4\theta}{35}\cdot\frac{t^6}{6\,!}$$
$$+\frac{1575+4200\sin^2\theta+2352\sin^4\theta+192\sin^6\theta}{175}\cdot\frac{t^8}{8\,!}+\ldots\ldots$$
$$\ldots\ldots(4\cdot15)$$

$$=1+\frac{t^2}{2\,!}+\frac{11}{5}\cdot\frac{t^4}{4\,!}+\frac{237}{35}\cdot\frac{t^6}{6\,!}+\frac{4617}{175}\cdot\frac{t^8}{8\,!}+\ldots\ldots\quad\ldots(4\cdot16)$$

The synthetic and compounded distributions agree, as was to be expected from the constancy of the variance, up to the sixth moment.

5. The point we now make is that if the distribution of tabular errors is normal, as may be reasonably assumed for a table not cut down, the only term left in the K-function is $\frac{t^2}{2!}$. It follows that K for the error in the interpolate is always $\frac{t^2}{2!}$ for all values of θ, hence the distribution of error at any point of interpolation is exactly the same as that of the tabular error. The distinction between the synthetic and the compounded distributions now disappears.

A distribution of error such as that given by (4·12) or (4·15) is by no means easy to deal with. Variation will take place as the interval between the tabular entries is traversed, and even when this is averaged over the interval, and very high order interpolation formulae are used, the distribution is still of a type which has not been investigated, and for which probability integral tables are not available. The best that could be done would perhaps be to use the Pearsonian Type II

$$df = \frac{\frac{n-2}{2}!}{\sqrt{n\pi}\,\frac{n-3}{2}!}\left(1-\frac{x^2}{n}\right)^{\frac{n-3}{2}} dx, \quad n = \frac{11}{2},$$

which agrees with it as far as the fourth moment, and does not differ widely in the neighbouring higher moments. It is, however, a distribution of finite range, whereas the true range must tend logarithmically to infinity as the order of interpolation is increased.

On the other hand, when the distribution of tabular errors is normal, that of the interpolate is also normal, and is constant throughout the interval. The restriction still applies, however, that this result has only been demonstrated for high order interpolation. With formulae using fourth or sixth differences the distribution is still normal but with a slightly inconstant variance; the synthetic distribution formed by averaging the variances is normal, and may be used for most purposes. The convenience of doing so should be borne in mind when new tables are in contemplation, and it is suggested in particular that, before a manuscript table is cut down to ensure the accuracy of the last figure, the effect of this procedure on the accuracy of the interpolate should receive very careful consideration.

REFERENCES.

(1) A. J. Thompson (1921). *Table of the Coefficients of Everett's Central Difference Interpolation Formula. Tracts for Computers,* V. Cambridge University Press.

(2) —— (1924). *Logarithmetica Britannica,* Part IX. Issued by the Biometric Laboratory.

(3) Karl Pearson (1920). *On the Construction of Tables and on Interpolation. Tracts for Computers,* II. Part I. Cambridge University Press.

(4) J. F. Steffensen (1927). *Interpolation.* Baltimore: The Williams and Wilkins Company.