A paper of great scientific and practical importance has been published under the above title by S. Krichewsky, Technical Assistant, Ministry of Public Works, Egypt, Physical Department Paper No. 23, 1937.

Prior to 1921, students employing correlation coefficients in the investigation of scientific questions were accustomed to gauge the significance of a coefficient by its magnitude and probable error. Mr. W. H. Dines 1 showed that—

If there is a cause A and a result M with a correlation r between them, then in the long run A is responsible for r2 of the variation of M.

Krichewsky points out that, although the validity of the r2 or Dines’ law was known to be based on the assumption that the cause A must be entirely independent of other contributory causes, B, C, etc., to variations of M, and therefore limited in its application, nevertheless this limitation has been so lightly emphasized that research workers may easily disregard its fundamental assumptions. In fact, he quotes a sentence in the article as an actual example of the misapplication of the r2 law.

In discussing a table giving, among others, the correlation coefficient +0.69 between E0, the b holographic solar constant and A0, the pyrheliometer readings extrapolated by straight lines to zero air mass, I stated that “the coefficient +0.69, interpreted by Dines’ law, means that 48 per cent of day-to-day variation in these two values of the solar constant, which are derived from the same parent data, occur in synchronism.” This application of the r2 law is of course a technical error, because E0 and A0, drawn from the same parent data, are obviously affected by covariation, due to possible changes in solar intensity, plus interrelated causes. The effect of this error on my analysis was to cause me to assign only 48 per cent as the measure of synchronism of E0 and A0, due to a common cause, whereas Krichewsky now claims that “more than 69 per cent of the variation of E0 occurs in synchronism with less than 69 per cent of the variation of A0 as a result of the common factor I0 + a, provided the individual errors of these two values of the solar constant are mutually uncorrelated.”

Asking the question, is Krichewsky not himself in error in the application of his extended r2 law to the particular case under consideration, I wish to make it plain that while writing my paper I was fully aware that a high relationship must exist between E0 and A0, as shown by the following words which immediately precede those quoted by Krichewsky:

Errorless values of E0 and A0 should show a high correlation, unless the fortuitous differences between them due to polychro-
absorption do not cause more or less the same systematic effects at both stations, such as shown by annual periodicity in values of $E_0$ or the negative correlation of $E_0$ and air transparency $a$.

(3) Homogeneous values of $E_0$ at a single station, but separated by a sufficiently long interval of time, as a fortnight, a month, or otherwise, might be used as values of $M_1$ and $M_2$ provided the covariation due to atmospheric transparency at the separate intervals were entirely independent, thus satisfying the fundamental assumptions.

It is very doubtful if any existing values of $E_0$ at separate stations are sufficiently free from artificial as well as physical correlations due to terrestrial cause to justify the labor of analyzing them by means of the relationships Krichewsky has developed.

The full significance of the simple relations presented by equation (3) is so well stated by its author that we can not do better than quote him in full:

(a) The two unknown coefficients $r_1$ and $r_2$ cannot be determined from $r_{12}$ unless some additional information exists about their mutual relation.

(b) Neither of the two is smaller than $r_{12}$, otherwise one of them would be greater than unity, which is impossible. Hence the values of $r_1$ and $r_2$ lie between $r_{12}$ and unity.

(c) The equation (3) may be written

$$r_{12} = r_1 r_2$$

which means that the correlation coefficient between $M_1$ and $M_2$ is equal to the geometric mean of the actual variations occurring in both owing to the action of $A$. [Same as $I_0$, c. f. M.]

Two important corollaries may be drawn from (4).

(i) If the values of $r_1$ and $r_2$ are unequal, say $r_1 < r_2$ then

$$r_{12}^2 < r_1^2 r_2^2$$

(ii) If $r_1 = r_2 = r$ then

$$r_{12} = r$$

So it appears that in this particular case the coefficient of the correlation itself and not its square is the true measure of the percentage of covariation occurring in the two variables owing to a third controlling factor. It may be of interest to point out that the relation (4b) might be used to calculate $r = \sqrt{r_{12}}$ in order to estimate $A$ from the data given by two instruments $M_1$ and $M_2$ or two observers working simultaneously and known to be of equal precision.

In this connection the following formula should be added.

Squaring and adding up each of the equations (1) we obtain the relations.

$$\sigma_t^2 = r_1^2 \sigma_1^2 + \sigma_2^2$$

which allow of estimating the relative magnitudes of $r_1$ and $r_2$ or even their exact values in case the standard deviations occurring in both owing to the action of $A$. [Same as $I_0$, c. f. M.]

The equations (5) may be interpreted that in the long run, $A$ is responsible for $r_2$ of the scatter occurring in $M$ as measured by the square of its standard deviation. This fact is nothing else than Dines's law.

(d) Lastly, let us add the useful interpretation of the formula (3) that the correlation coefficient $r_1$ between $M_1$ and its true controlling factor $A$ is reduced by $r_2$ per cent, and becomes $r_{12}$ in case $M_2$ is substituted for $A$ to represent it with a degree of precision measured by $r_2$.

We now come to the most important aspect of the whole question of interpretation.

Equation (3) and all the relations and deductions that precede are based upon the pure assumption that $I_0$ is really an independent variable. However, the magnitude and nature of possible variations in $I_0$ have not as yet been conclusively disclosed and evaluated. Accordingly, we are fully justified in making the assumption that $I_0$ is a constant within limits of the precision of measurements of $M_1$ and $M_2$. How must the correlation coefficient $r_{12}$ be interpreted under this assumption? Obviously $r_1$ and $r_2$ are simply nonexistent and $r_{12}$ is simply the covariation due to $B$ and $C$, which in the case of solar measurements at a single station not only comprise independent instrumental errors of different kinds but also errors in common and effects due to atmospheric transparency. I believe no one has computed actual values of $r_1$ so we can hardly say what will be found even if data satisfying the basic assumptions were available.

If $I_0$ is constant then errorless values $E_0$ would have to be strictly a constant and $A_0$, a variable depending upon the effects which arise from extrapolation of pyrheliometer readings to zero air mass by straight lines. The coefficient +0.69 must therefore be interpreted to mean that a comparatively large part of the pyrheliometer and bolograph fluctuations which originate in the initial observations and measurements are extrapolated to zero air mass.

If we assume solar variability, then equations like (3) will seem to support solar variability. On the other hand, if interpreted on the assumption that $I_0$ is constant, the same correlation coefficients like $r_{12}$ represent nothing whatever but local terrestrial and instrumental effects. This is peculiarly the case in the analysis of solar data because the total fluctuations are quantitatively very small.

In my earlier paper I set up three simultaneous equations (10), (11), (12), pages 289 and 290, in MONTHLY WEATHER REVIEW, July, 1925, by which the solar variability could be computed from independent observations at two stations. Owing to the lack of suitable observations up to the present time it has never been possible to apply these equations in any practical way. I am now impressed, however, with the importance of repeating a word of caution I expressed in the earlier article and which must always govern the interpretation we put upon results secured from equations based upon certain hypothetical assumptions which may not in fact be justified. The quotation reads:

The mathematician recognizes, of course, that securing a seemingly rational and finite value of $\sigma_t$ (solar variation in the solution of the three equations for a group of simultaneous observations is no proof of solar variability. Having assumed solar variability, a solution of the equations simply apportions to solar variation such part of the total variation as best satisfies the observations at the two stations under the assumed conditions. Some sets of observations may give imaginary roots, and it is obvious that errors of observation can be neither zero nor imaginary.

Solar variation can be shown by these equations only when the results are based on several groups of data from wholly independent stations. As pointed out above, equations of the type of (9) are valid only if $\sigma_t$ is unrelated to $\sigma_r$ or $\sigma_d$ in magnitude.

In addition to the comparatively simple and elementary portion of Krichewsky's paper discussed in the foregoing, he has extended his analysis to a general investigation of Dines's law. This important addition to the statisticians' facilities for the interpretation of the results of their investigations is discussed in the following paper by Mr. Woolard.