

NOTE ON METHODS FOR INDICATING AND MEASURING CORRELATION, WITH EXAMPLES.

551.501

By H. W. CLOUGH.

[Weather Bureau, Washington, D. C., May, 1921.]

SYNOPSIS.

The conventional measure of correlation between two variables x' and y' is expressed by the ratio of the mean product of x, y to the product of the standard deviations of x' and y' , viz: $r = \frac{\sum(xy)}{n\sigma_x\sigma_y}$, in which x and y must be measured from the respective mean values of the variables. The present note indicates methods for securing approximate values of r with less labor of computation, also other methods of measuring both correlation and dispersion or scatter of data, and the analytical relations between them on the basis of a very large number of observations.

THE SIGNIFICANCE OF A CORRELATION COEFFICIENT.

The first and practically indispensable step to take in establishing the relation between the two variables is to make a dot chart of the data. If by inspection it appears that a straight line represents the arrangement of dots better than, or as well as, any other line it is practicable to employ, then we may proceed to determine the straight line of best fit by least square methods or otherwise, or we may follow the usual rules and compute at once the correlation coefficient.

Now perfect correlation means that all the dots fall exactly on the straight line. Ordinarily, however, there is considerable scatter or dispersion of the dots, and in such cases the coefficient is a measure of how closely the dots conform to the straight line of best fit. This coefficient is given by the expression

$$r = \frac{\sum xy}{n\sigma_x\sigma_y} \tag{1}$$

in which x and y are corresponding values of the variables measured in terms of the deviation from their respective mean values, n is the number of pairs and σ_x and σ_y are the respective standard deviations.

It is easy to show also that ¹

$$r = b \frac{\sigma_x}{\sigma_y} \tag{1a}$$

in which b is the tangent of the angle between the straight line of best fit and the X axis.

MEASURES OF DISPERSION OF DATA.

Up to the present time practically no use has been made in studies of correlation of any other measure of dispersion than the *standard deviation*, notwithstanding that other measures have long been known and sometimes used in other connections. In what follows it will be shown that any measure of dispersion may be substituted in the second member of equation (1a) for the standard deviation and with identical results, subject only to the limitations of paucity of data or of sampling.

There are several different indexes of the dispersion or scatter of data which vary in relative magnitude but nevertheless bear definite mathematical relations to each other when the number of observations is great and the distribution Gaussian.

The *standard deviation*, much employed in statistical investigations, is the square root of the mean of the squares of the departures from the true mean.

¹ Marvin, Chas. F.: Elementary Notes on Least Squares, etc. MO. WEATHER REV. 44: 567.

The *mean deviation* is the mean of the departures from the true mean, taken without regard to sign.²

The *standard variation*.—By analogy to the standard deviation, this term may be employed to designate the mean of the squares of the variations between consecutive values.

The *mean variation* is the mean of the differences between consecutive values of the variable at assumed equal intervals taken without regard to sign.

Relations between measures.³—In the case of numerous observations with a Gaussian distribution the mean deviation multiplied by 1.253 equals the standard deviation. The mean variation bears the same relation, 1 to 1.253, to the standard variation.

The mean deviation multiplied by $\sqrt{2} = 1.414$ = the mean variation where the order of succession of the data is fortuitous. Other relations are easily derived.

Since there are definite relations between the measures of dispersion just described, the ratios of the several measures of dispersion of x' and y' tend to approach equality with a large number of observations, hence a generalized form of equation (1a) may be written

$$r = b \frac{s_x}{s_y} \tag{2}$$

in which s_x and s_y are any measures of scatter of x' and y' respectively. It is obvious, therefore, that the coefficient of correlation should be sensibly the same whichever measure of dispersion is used in its derivation.

An example will show the differences which arise in a practical case.

Take the relations between July rainfall and the yield of corn in four States.⁴ By least square methods the equation of the straight line of best fit is

$$y' = 24.07 + 2.027 x'$$

in which y' is the yield of corn per acre and x' the July rainfall. If the origin of coordinates is taken at the mean value of x' and y' the equation becomes for departures from the mean,

$$y = 2.027 x$$

that is, $b = 2.027$.

Deriving from the original data of this example the four measures of dispersion mentioned above we obtain the results in the table with corresponding values of r , derived by substituting in equation (2),

Measures of dispersion and correlation.

	Rainfall.	Acre yield of corn.	Correlation coefficient.
Standard deviation.....	<i>s.</i> 1.31	<i>s.</i> 4.45	<i>r.</i> 0.60
Mean deviation.....	1.07	3.46	.63
Standard variation.....	2.03	6.80	.60
Mean variation.....	1.63	6.52	.61

² For an abridged method of computing this, see Marvin, Theory and Use of the Periodocrite. MO. WEATHER REV., Mar., 1921, 49: 120.

³ For discussion of the various measures of dispersion and reference to original sources see Clough, A Statistical Comparison of Meteorological Data with Data of Random Occurrence. MO. WEATHER REV., Mar., 1921, 49: 125.

⁴ Marvin, C. F.: Elementary Notes on Least Squares. Loc. cit. p. 564.

The divergence between the values of r in the table is largely explained by the relatively small number of observations, 28. The coefficient of correlation computed by the method of variations is somewhat inferior, in dependability, to that by the usual method unless the data are sufficiently numerous.

SHORT METHOD.

It is easy to see that equation (2) furnishes a short method of computing a correlation coefficient with considerable accuracy, as follows:

By simple inspection, locate on the dot chart the straight line of approximately the best fit. For this purpose locate on the chart a master dot representing the mean values of the variables. The line of best fit must pass through this dot and be so inclined as to best represent all the dots. The equation of this line referred to the master dot as origin is $y' = bx'$ or $y' = a + bx'$ if referred to any parallel axes with origin not on the line. The value of b may be easily found by taking the ratio $y' \div x'$ for any point on the line more or less distant from the master dot.

This value of b together with the ratio of, say, the mean deviation of the variables, a quantity much more easily computed than the ratio of the standard deviations, gives at once by equation (2) a value of r which is increasingly accurate as r is greater.

In addition to its use in this short method of computing a coefficient of correlation, the mean deviation as a measure of dispersion will frequently suffice for other purposes and make unnecessary the tedious computations of standard deviations ordinarily resorted to.

Correlation by algebraic signs.—The writer has found the following still shorter method of deriving an index of correlation very useful when a large number of groups of observations must be examined and the employment of the usual methods would be impracticable. The method also serves as a preliminary test for determining quickly if sufficient correlation exists to justify computation by the more exact methods:

Count the number of times when the deviations of x' and y' from their respective means, or when the variations between consecutive values have the same sign. Divide by the number of observations. The ratio thus obtained is a rough index of the correlation. If there is a positive correlation, the percentage will range between 0.50 for absence of correlation to 1.00 for perfect correlation. If there is negative correlation it will range between 0.50 for absence of correlation to 0 for perfect correlation. To reduce this ratio to a type of correlation coefficient, subtract 0.50 from it and multiply by 2. In general the coefficient by this method is somewhat less than that computed by the usual method.

REMARKS ON THE PREPARATION OF DATA.

Correlation between two variables implies similarity of fluctuation due to a causal relation which one bears to the other or which both bear to a third variable. Fluctuations in a variable are usually measured by deviations from a mean. Meteorological variables are either daily observations or weekly, monthly, or yearly means. All meteorological data are characterized by fluctuations of varying length and amplitude. The fluctuations may be classed, in general, as short period or long period fluctuations, which latter may be clearly disclosed only when the minor fluctuations are eliminated. If it is desired to correlate two series of data it is necessary to discriminate carefully between the short period, which may be likened

to accidental fluctuations, and the long period or systematic fluctuations, otherwise the correlation may be spurious. There may be high correlation between the larger fluctuations and low correlation if the smaller fluctuations are considered, or vice versa. For example, take the daily temperature normals for the month of March for two 20-year periods. If the two series of normals be plotted as two variables, x' being the normal value for any day of March of the first period and y' the normal value for the corresponding day of March of the second period, the chart will show a pronounced tendency for high values of one period to be associated with high values of the second period. This is a correlation, however, due to the annual change common to both periods. If this seasonal variation be eliminated by plotting deviations from the true normal, derived from the whole 40-year series by harmonic analysis, the resulting plot shows that there is no correlation between the residuals for corresponding dates of the two 20-year periods.

Another illustration is furnished by daily maxima and minima of temperatures, at any locality. Suppose we have daily maxima and minima for April. If the annual variation common to both variables be not eliminated, a correlation coefficient manifestly too high will be the result. By eliminating the annual variation the coefficients at different places are comparable. Thus the correlation is found to be very high on a mountain peak and low in a valley having a large daily range.

The correlation between the daily values of the vapor pressure at the surface and the total amount of the water vapor in the whole atmosphere vertically above the station as determined at Mount Wilson, Calif., by spectroscopic methods further illustrates this principle. Obviously there is correlation between the two variables. Both variables are high in summer and low in winter. On the other hand if one variable shows an increase from one day to the next, the other may show a decrease and in fact there is small correlation between their day-to-day fluctuations. Taking a large number of observations, the vapor pressure at the surface multiplied by a constant gives approximately the total vapor content of the whole atmosphere, but there are wide variations in this ratio from day to day. We may desire a correlation which gives an indication of the reliability of the spectroscopic method, assuming that taking the averages of the spectroscopic determination and the surface values over a long period of time, there should be substantial agreement between their variations. By combining the data into 5 to 10 day means the minor fluctuations which show little similarity are smoothed out, and there is then disclosed a high correlation between the two variables. At Calama, Chile, a correlation of +0.95 was found from the weekly means. On the other hand if we seek the correlation between the small day to day fluctuations, the seasonal variation must be first eliminated. Then the irregular short-period fluctuations common to both variables may be eliminated by taking deviations from consecutive seven-day means. These residuals are the data to be employed in computing the correlation coefficient.

The relation between two variables may be such that the value of one depends largely upon the value of the other but one variable is subject to an influence not affecting the other. The minor fluctuations synchronize but the major fluctuations are unlike. This is illustrated by the relation between the atmospheric transmission coefficient and the solar intensity at air-mass

zero. Both quantities are evaluated in a single operation which consists in fitting a straight line to plotted pyrheliometric observations of the solar intensity at varying air masses, and prolonging it to the zero of abscissas or air mass zero. The data show pronounced correlation in respect to the minor day-to-day fluctuations, as shown by the method of correlation by variations, while by the usual method of correlation by deviations from the mean, little correlation results, owing to the existence of long-period changes in the transmission coefficient without corresponding changes in the solar intensity.

Variation in the mean.—The mean may vary either systematically through long-period fluctuations or accidentally, as by a change of hours of observation, or exposure of instruments, rendering the data nonhomogeneous. A gradual or abrupt change in the regimen of

a river or an increase in the number of crop reporting points in a State over a term of years or an increase in yields due to cultivation are further examples of such actual or accidental changes in the mean. Frequently it is difficult or impossible to distinguish between an actual change in the mean and a true secular change, both of which may cause deviations which are not present in the variations of another variable with which a relation is sought. The usual method of correlation by deviations yields results more or less spurious. In all such cases the method of correlation by variations should be employed, since it is quite independent of nonsimultaneous changes in the means of the two variables, either accidental or systematic. The usual method by deviations is, however, appropriate in the case of systematic changes if the secular change be first eliminated by taking deviations from means varying with the general trend.

THE TEXAS FLOODS OF SEPTEMBER, 1921.

GENERAL DISCUSSION.

By B. BUNNEMEYER, Meteorologist.

[Weather Bureau, Houston, Tex., Oct. 10, 1921.]

627.41 (764)

Torrential rains in southern and central Texas from September 8 to 10, inclusive, 1921, resulted in phenomenally rapid floods in streams and lowlands, especially in Bexar, Travis, Williamson, Bell, and Milam Counties, and caused the death, so far as is known, of 215 persons and property loss estimated at over \$19,000,000. This exceeds the havoc wrought by the record-breaking floods of December, 1913, when 177 persons lost their lives and property valued at nearly \$9,000,000 was destroyed. But in December, 1913, there were practically no crops in the fields.

The heaviest precipitation was reported from Taylor, Williamson County, where 23.11 inches occurred in 24 consecutive hours, September 9–10, which is the greatest 24-hour rainfall of record for the State of Texas, the previous record being 20.60 inches at Montell, Uvalde County, on June 28–29, 1913.

Throughout the stricken area traffic by railroad, street car, or other conveyances was interrupted by washouts, loss of bridges, and accumulation of débris; telegraph, telephone, electric light, and other public services were crippled, and numerous small houses and other structures were carried off by the currents that swept through cities and rural districts, resulting in the loss of many lives. Much other damage was done, largely to crops, mostly corn and cotton. Considerable damage was also caused by violent thunderstorms and squalls occurring in various localities during the downpour, although it was overshadowed by the havoc due to the flood.

While creeks and other tributaries rose to appalling heights, the trunk streams were much less seriously affected than was anticipated from the deluge, the redeeming features being a previously dry soil and low streamflow. The run-off was swift and much of the back water did not return to the streams, resulting in a rapid diminution of the volume of water rushing toward the Gulf of Mexico. The subsidence of the flood wave on the Brazos River was so rapid that flood stage was not attained in the lower reaches of that stream at or below Rosenberg, while at Valley Junction, where the water poured in from the Little River, the stream was 14.2 feet above flood stage and only 0.8 foot below the record high watermark of the December, 1913, flood.

Cause of the rains.—Evidence is strong that the precipitation was the result of the breaking-up in Texas of the disturbance that moved westward toward the Mexican coast south of Tampico on September 7, 1921. Although the distribution of the pressure was such that the storm could not be charted, the shifting winds, the progressive northeastward extension of the rainfall area, and the profound agitation of the atmosphere as evidenced by violent squalls and thunderstorms over the stricken sections, can hardly be ascribed to any other cause. The storm apparently moved in from Mexico over Webb County and passed in a northeasterly direction over Bexar, Comal, Hays, and Travis Counties into Williamson, Bell, and Milam Counties where it abruptly dissipated. Milam County borders on the west bank of the Brazos River, and there was very little precipitation along the east bank of that stream. An area of high pressure of apparently feeble energy backing in over eastern Texas was probably a contributory cause of the record breaking rains and their abrupt termination near the Brazos River. While the rains were disastrous to life and property over a large area, there were many localities in southern Texas where they proved beneficial by relieving the drought, reviving ranges, and providing stock water.

Warnings.—The flood waters accumulated so rapidly in creeks and lowlands that residents were taken completely by surprise. Warnings of impending rises were issued, however, immediately upon the receipt of rainfall reports to the main streams on which river stations are maintained. On the morning of September 9, to the lower Rio Grande from Rio Grande City to Brownsville; and on September 10 to the Colorado below Austin, and to the Brazos from Valley Junction to Richmond, with injunctions to keep live stock from the lowlands and protect other interests. Warnings were repeated on September 11 to residents along the Colorado and Brazos, and extended along the latter stream from Richmond to Freeport. Thereafter residents were kept informed daily of the progress of the floods until danger was over. Similar warnings and injunctions were issued September 11 to the lower Guadalupe. Earlier warnings were not advisable as there were no data available to justify them.