A Comparison of Rank-Difference and Product-Moment Correlation of Precipitation Data

JAMES E. MCDONALD AND CHRISTINE R. GREEN

University of Arizona
Tucson, Arizona

Abstract. A large sample (4650 cases) of comparisons of the Spearman rank-difference correlation coefficient r' and the Pearson product-moment correlation coefficient r is presented. The correlation data are 50-year records of winter and summer half-year precipitation totals for 220 stations, well distributed throughout the United States. The agreement between corresponding values of r' and r is very close; about 72 per cent of all cases differ by less than one-half the standard error of r, and about 94 per cent differ by less than one standard error of r. It is concluded that the distributions of r' and r are so close that for most geophysical correlation applications r' is as adequate a statistic as r, and in certain cited cases is preferable to r; hence r' deserves much wider use in climatology and hydrology than it has received in the past.

Introduction. In problems in climatology and hydrology where correlation techniques are needed it has become common practice to make almost sole use of the Pearson product-moment correlation coefficient. When it is desired to obtain regression equations for prediction purposes (for example, in certain rainfall-runoff studies), then the product-moment coefficient r is obtained almost as a by-product of the regression calculations. But in a much larger class of problems, one seeks only a reliable and objective measure of degree of association between two sets of data. For example, in pilot studies of climatological homogeneity of a region, one is chiefly interested in depicting the spatial pattern of correlation between one or more base stations and a large number of surrounding stations. In such problems, another type of correlation coefficient, the Spearman rank-difference correlation coefficient r', deserves much wider use by geophysicists than it has received in the past. A particularly strong argument in favor of use of r' is found in its nonparametric nature, which makes it conceptually preferable to r for correlating non-normally distributed data, of which most hydrologic and climatic data are examples. But, in addition, ranking procedures are inherently faster than the procedures required to compute r, at least for sample sizes of less than 40 [*Snedecor*, 1946], particularly when manual methods rather than electronic computer methods are to be used.

When r' is to be used merely to detect the *existence* of correlation between two sets of variates (that is, when one seeks only to answer the question of whether a sample value of r' differs from zero by an amount too great to be attributed simply to chance), then reference may be made to published significance tables [for example, *Dixon and Massey*, 1951]. But equally frequently, or perhaps even more frequently, the investigator wishes to draw conclusions from the magnitude of the coefficient itself (as in the aforementioned example of homogeneity studies), and here an obstacle is encountered. The sampling distribution of r' for parent bivariate populations of *non-zero* correlation has never been theoretically deduced. The geophysicist who seeks advice on this point in the literature of statistics will find that some statisticians suggest that the familiar standard error σ_r for the product-moment correlation coefficient will be quite trustworthy if applied to r', whereas others say that sampling errors of r' simply cannot be assessed pending further theoretical developments for the non-zero case.

Because theoretical knowledge of the sampling distribution of such a statistic as r' is lacking, we must resort to empirical determinations. One published effort of this kind that has come to our attention is a very limited sampling experi-

ment described by *Snedecor* [1946, p. 166], involving only 10 trial samples. Its results indicate encouraging agreement between the distributions of r and of r'. The only other such experiment of which we are aware was done by *McDonald* [1957*a*], using only 14 precipitation correlations. In the latter study, 11 of the 14 cases gave r' values falling within one standard error of r, which is also fairly encouraging evidence that r and r' have closely similar distributions; but so small a sample is far from conclusive.

In the course of an extensive precipitation-correlation analysis that we have recently carried out on an automatic computer, we have had an excellent opportunity to secure a very much larger sample of comparison values of r and r'. The purpose of the present note is to summarize the results to show how very close an estimator of r the rank-difference coefficient r' is when applied to actual geophysical data.

Data and analysis. The present empirical comparisons were obtained as an incidental part of a study in which 50-year seasonal (winter and summer) precipitation records for 220 U. S. Weather Bureau stations were correlated, in groups of the order of 100 secondary stations each, against various base stations. Fifteen base stations, well distributed over the entire United States, yielded a total of 2325 summer and 2325 winter correlations, or a total sample of 4650 independent correlations. Each comprised time series of 50 seasonal-precipitation totals. These results will be reported elsewhere, so no listing of stations will be given here. 'Winter' was the period from November through April, 'summer' the balance of each year. All data were taken from published Weather Bureau records. The 50-year period was from 1906 to 1955. In a very small percentage of cases, missing monthly totals were estimated by the method described by *McDonald* [1957*b*]. Selection of the 15 base stations was done on the basis of double-mass analysis in order to be sure that all the base station records were homogeneous. The records of the secondary stations were not tested by double-mass methods; they were taken as published.

To prepare for the computation of the rank-difference correlation, the winter records and summer records of each of the 220 stations were separately ranked by punchcard sorting techniques. Then, in the main part of the computing program, the computer not only performed the steps required to obtain the product-moment correlation, but concurrently took rank-differences and from these computed r' according to the equation

$$r' = 1 - 6 \sum_{i=1}^{N} (D_i^2)/N(N^2 - 1) \qquad (1)$$

where D_i is the difference in ranks of the ith pair of variates and N is the total number of pairs, here 50 in every instance. The 'tied-ranks' problem, about which the statistics literature contains a variety of discussions, was simply ignored because it was found that treating the ties by the method specified in standard statistical references could not be expected to change numerical values of r' by more than about 1 per cent in the least favorable cases. The tied ranks were simply ranked in the *chronological* order in which they inevitably came through the ranking operation.

With so large an empirical distribution-sample as was involved in this study, it was indispensable to have some scheme by which the computer itself would do virtually all the comparison analysis, so the following simple method was employed. For each of the 4650 pairs of r and r' values, we calculated a quantity q defined according to

$$q = (r' - r)/\sigma_r \qquad (2)$$

where σ_r is the conventional standard error of r given by

$$\sigma_r = (1 - r^2)/(N - 1)^{1/2} \qquad (3)$$

Thus q measures the amount by which each r' exceeds its corresponding r, expressed in units of the standard error of r itself. As a final step in processing the data, the entire sample of 4650 correlations was sorted with respect to r, without regard to station or season, in order to group the results by r intervals of 0.20; and then within each r interval the result cards were ranked by value of r' to permit the final tallying, which is summarized below. This last procedure was employed to permit detection of any trends in the q distribution as r ranged from positive values near 1.0 (nearby station

TABLE 1. Percentage Distributions of q Values

r interval	q interval: 1.1 to 1.5	0.6 to 1.0	0.1 to 0.5	−0.4 to 0.0	−0.9 to −0.5	−1.4 to −1.0	−1.9 to −1.5	−2.4 to −2.0	−2.9 to −2.5	Number of cases
.70 to .89	2.1	4.1	9.3	24.8	32.0	17.6	6.2	3.1		97
.50 to .69		7.2	20.8	31.3	19.1	10.6	6.6	1.0		303
.30 to .49	0.6	6.0	26.9	37.2	19.4	8.5	1.5			862
.10 to .29	0.8	7.7	33.8	39.2	14.7	2.8	0.8			1402
−.09 to .09	0.8	9.4	39.0	40.6	8.2	1.5				1315
−.29 to −.10	0.9	8.6	42.8	38.9	8.4					581
−.49 to −.30	1.1	11.6	47.2	29.1	9.3					86
−.49 to 1.0	0.7	8.0	34.0	38.2	13.5	3.9	1.2	0.3	0.1	4646

pairs) to the maximum negative values of almost −0.5. This proved useful, for an interesting trend did appear, as will be noted below.

Results. In Table 1 are presented the percentage distribution of q values, grouped here in r intervals of 0.2 to facilitate examination. The merely four cases wherein r exceeded 0.90 are omitted, so the total sample comprises 4646 cases in Table 1. The number of cases falling within each r interval varied widely, since the original selection of stations was, of course, made without knowledge of how the r values would come out. That so many values of r lie near zero resulted from our extension of the correlation-field analyses out to distances often exceeding a thousand miles from the base stations. Even in the two least populous r intervals of Table 1, however, the subsamples are large enough to permit reliable conclusions to be drawn.

Discussion and conclusions. The results of principal interest in Table 1 are found in the bottom line of the table. For the entire sample, 72.2 per cent of the q values were of absolute magnitude equal to or less than 0.5; that is, this high percentage of r' values did not depart from their corresponding r values by more than one-half a standard error of r itself. And in 93.7 per cent of all cases r' lay within one standard error of r.

It seems justifiable to conclude that in almost all applications in which the geophysicist seeks a measure of correlation, r' will serve him just as well as r. Only in instances where regression methods are to be employed or where it is established that the investigator deals with truly bivariate normal distributions could it be claimed that r will clearly be a better correlation statistic than r'. But the latter instances are almost unheard of in geophysical work, since nonnormality is the rule, not the exception. Hence these results indicate that much wider use should be made of r' as an easily computed measure of correlation that is an impressively accurate predictor of r itself.

It will be noted that the maxima of the individual q distributions for the several class intervals of r in Table 1 exhibit a quite systematic trend, such that the mean absolute magnitude of r' tends to be somewhat less than that of r. Only for the class interval involving values of r between −0.09 and 0.09 is the q distribution symmetric about zero. Thus, it is generally true that r' is *slightly conservative* as an estimator of r. This property is also one that recommends use of r', since, most of the time, one will not be misled into inferring an unduly large degree of association between variates when one employs r' rather than r.

The results described above fill in, at least on an empirical basis, the one significant gap in the arguments that can be advanced in support of viewing the Spearman rank-difference coefficient as the *preferred* measure of correlation in most geophysical application. These arguments have been discussed in some detail by *McDonald* [1957*a*], so here they need only be summarized very briefly:

1. The fact that r' is a distribution-free (nonparametric) statistic, plus the well-known fact that most rainfall, streamflow, and other geophysical data are not normally distributed, makes r' conceptually preferable to r. The vexing (but probably overemphasized) normality question simply disappears when an investigator measures correlation with r' rather than with r.

2. *Hotelling and Pabst* [1936] showed that r' has very high statistical efficiency even in the case of bivariate normal distributions. In the large-sample limit, and for normal bivariate populations characterized by *zero* correlation (the one theoretically soluble case), r' has an efficiency of 0.91. That is, in order to obtain rank correlations as sensitive (in the sense of equal sampling variance) as product-moment correlation coefficients, one must employ sample-sizes about 10 per cent larger for r' than for r, a consequence of information lost through converting from continuous variates to ranks. For non-normal cases, Hotelling and Pabst concluded that r' would have efficiency even higher than 0.91.

3. Not infrequently an investigator wishes to correlate quantities which are themselves nonlinearly related to the independent variables governing them. Relative humidities, vapor pressures, and kinetic energies of flow are familiar examples. When product-moment correlation of such quantities is carried out, there is reason to doubt whether some linearizing transformation of the quantities should precede correlation analysis. But, inasmuch as ranks are invariant under any kind of transformation that only stretches scales, this question entirely disappears when rank-correlation techniques are used, a point called to our attention by R. A. Bryson.

4. When correlation is to be done manually or by desk calculator, and when the sample size is less than about 40, r' can be computed more quickly than r. This advantage becomes particularly noticeable with sample sizes of only 20 or so, for then the ranking step proceeds very rapidly, since the squares of the rank-differences may be simply written from memory.

5. The large-sample comparison reported here strongly supports the conclusion that the sampling distribution of r' for the case of non-zero correlation (the case for which there remains no theoretical solution) must be so nearly identical with that for r that, in most geophysical studies, entirely acceptable estimates of sampling error of r' can be obtained simply by using σ_r, the familiar standard error of r.

In the three most recently published works on the use of statistics in climatology and meteorology, there is so little information on r' that one is led to believe that this statistic has little to recommend it. *Conrad and Pollak* [1950] and *Panofsky and Brier* [1958] do not mention r', and *Brooks and Carruthers* [1953, p. 235] go so far as to say that correlation by ranks is not likely to have much application in meteorology. We strongly disagree with the latter view, and recommend that r' be regarded as the preferred correlation statistic in all cases of moderate sample size, where (as is almost invariably true) the data are not known to be normally distributed, and where regression anaysis is not the fundamental objective.

Acknowledgments. We wish to express our appreciation to Robert W. Mitchell of the University of Arizona Numerical Analysis Laboratory for his assistance in computational matters. The work reported here was supported by the Office of Naval Research under Contract NR 082-164.

References

Brooks, C. E. P., and N. Carruthers, *Handbook of Statistical Methods in Meteorology,* M. O. 538, Air Ministry, H. M. Stationery Off., London, 412 pp., 1953.

Conrad, V., and L. W. Pollak, *Methods in Climatology,* 2d ed., Harvard Univ. Press, Cambridge, 459 pp., 1950.

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis,* McGraw-Hill, New York, 370 pp., 1951.

Hotelling, H., and M. R. Pabst, Rank correlation and tests of significance involving no assumptions of normality, *Ann. Math. Statist.,* 7, 29–43, 1936.

McDonald, J. E., A critical evaluation of correlation methods in climatology and hydrology, *Sci. Rept. 4, Inst. Atm. Physics,* Univ. Arizona, 36 pp., 1957*a*.

McDonald, J. E., A note on the precision of estimation of missing precipitation data, *Trans. Am. Geophys. Union, 38,* 657–661, 1957*b*.

Panofsky, H. A., and G. W. Brier, *Some Applications of Statistics to Meteorology,* Pennsylvania State Univ., University Park, Pa., 224 pp., 1958.

Snedecor, G. W., *Statistical Methods,* 4th ed., Iowa State Coll. Press, Ames, Iowa, 485 pp., 1946.

(Manuscript received August 29, 1959.)