Journal of Glaciology, Vol. 11, No. 63, 1972

A FURTHER NOTE ON CORRELATION COEFFICIENTS DERIVED FROM CUMULATIVE DISTRIBUTIONS*

By DAVID P. ADAM†

(Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona 85721, U.S.A.)

ABSTRACT. This paper elaborates on the note by Andrews and others (1971). It demonstrates that one may obtain any arbitrary value of r between two series of observations by adjusting the mean values of the two series before cumulating them. A computer simulation is used to illustrate the behavior of random Normal series cumulated under varying conditions.

RÉSUMÉ. *Une note supplémentaire sur les coefficients de corrélation dérivés de distributions cumulatives.* Cet article s'appuie sur la note d'Andrews et autres (1971). Il démontre que l'on peut obtenir une valeur arbitraire de r entre deux séries d'observations en ajustant les valeurs moyennes des deux séries avant de des cumuler. Une simulation sur calculateur est utilisée pour illustrer le comportement d'une série aléatoire Normale cumulée sous diverses conditions.

ZUSAMMENFASSUNG. *Eine weitere Bemerkung zu den Korrelationskoeffizienten aus kumulativen Verteilungen.* Dieser Beitrag führt die Bermerkung von Andrews und anderen (1971) weiter. Er zeigt, dass man jeden beliebigen Wert für r zwischen zwei Beobachtungsreihen erhalten kann, wenn man die Mittelwerte der beiden Reihen vor ihrer Kumulierung entsprechend anpasst. Eine Computersimulation wird zur Illustration des Verhaltens von zufälligen Normalverteilungen, die unter varierenden Bedingungen kumuliert werden, benutzt.

ANDREWS and others (1971) have correctly noted that the product-moment correlation coefficient cannot be used on cumulative data. I wish to add that the value of r between two cumulative series is not independent of the scale used for measurement, and that it is in fact possible to obtain nearly any desired value of r by simply adjusting the mean values of the two series before cumulating them.

The product-moment correlation coefficient is designed to deal with data which follow a Normal distribution. Observations of such data may be expressed as

$$x_i = \bar{x} + e_{x_i} \tag{1}$$

where $\bar{x}$ is some mean value and e_x is $N(0, \sigma)$. When a series of observations of the Normal variate of Equation (1) is expressed in cumulative form, the nth observation becomes

$$c_n = n\bar{x} + \sum_{i=1}^{n} e_{x_i}. \tag{2}$$

The final term in Equation (2) introduces a serial correlation which destroys the independence of the observations and converts the series of random Normal observations into a one-dimensional random walk (Mitchell and others, 1966, p. 6). This effect is illustrated in Figure 1; 500 random Normal observations were generated on a CDC 6400 computer using the algorithm of Naylor and others (1966, p. 95), and these are plotted as a raw series (Fig. 1a) and as a cumulated series (Fig. 1b). It is clear that the cumulated series is far from random. The correlation between two random series is substantially altered by the transformation from raw to cumulated series, and this is shown in Table I. Ten pairs of random Normal series, e_x and e_y, were generated ($N = 500$) and the correlations between them were calculated for both raw (Equation (1)) and cumulated (Equation (2)) series, with $\bar{x} = \bar{y} = 0$. The correlations between the cumulated series give no hint of the basic lack of relationship between the raw observations.

Another potential source of error is that if the mean value of a series is different from zero, then the first term on the right side of Equation (2) will introduce a linear trend into the set of cumulative observations. The magnitude of the trend depends upon the absolute value of the mean and upon the length of the series, while the direction of the trend depends on the sign of the mean.

* Publication No. 39. Department of Geosciences, University of Arizona.

† Present address: U.S. Geological Survey, 345 Middlefield Road, Menlo Park, California 94025, U.S.A.

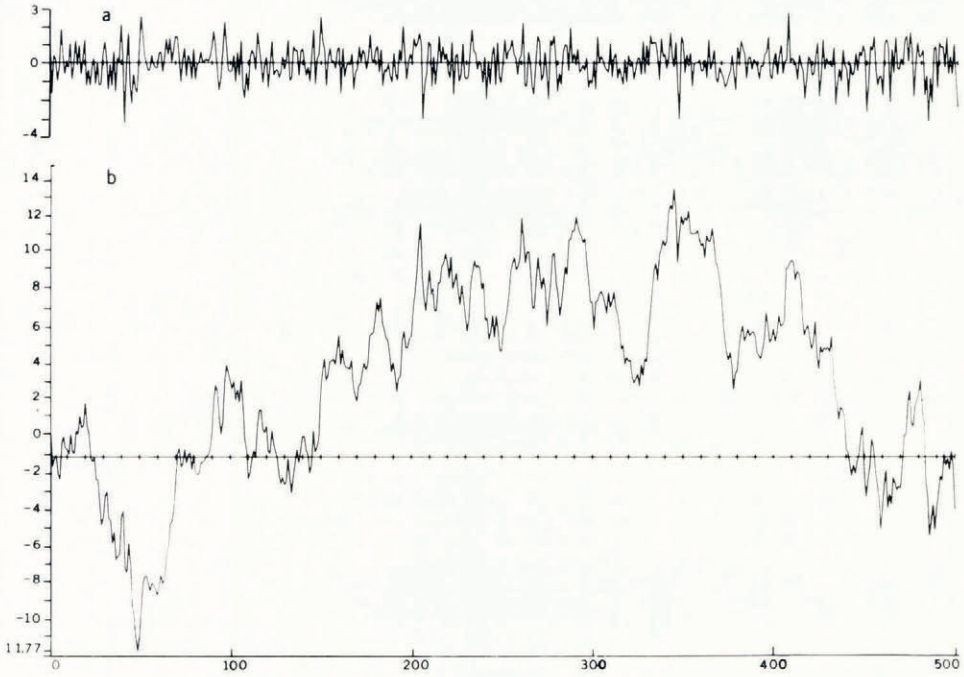

Fig. 1. A 500-observation random Normal series graphed (a) in raw form, and (b) in cumulated form.

TABLE I. SERIES MEANS ($\bar{x}$ AND $\bar{y}$) AND CORRELATIONS BETWEEN TWO SERIES OF RANDOM NORMAL VARIATES

r is the correlation between the two series in their raw form, and r_c is the correlation between the same two series, expressed in cumulative form. $N = 500$ observations.

Run	$\bar{x}$	$\bar{y}$	r	r_c
1	−0.0363	0.0657	−0.032	−0.397
2	0.0151	0.0080	−0.030	0.049
3	0.0618	0.0570	0.023	0.616
4	0.0241	0.0479	0.050	0.536
5	0.0068	−0.0807	0.043	0.389
6	0.0000	0.0094	−0.074	−0.185
7	−0.0587	−0.0596	−0.060	0.732
8	−0.0107	−0.0315	−0.031	0.440
9	−0.0696	−0.0056	0.001	−0.361
10	−0.0803	−0.0257	0.037	0.541
Means	−0.0148	−0.0015	−0.007	0.236

When the means of both series in a correlation analysis are different from zero, the introduced linear trends tend to dominate the relationship between the two sets of cumulative observations. A simulation model was designed to study the behavior of the correlation coefficient between two cumulated random Normal series, x and y, when different combinations of $\bar{x}$ and $\bar{y}$ were added to the series before cumulation. Two 500-observation series of random Normal variates corresponding to the e_x (or e_y) terms of Equation (1) were generated. Values of $\bar{x}$ and $\bar{y}$ were varied from −0.2 to +0.2 by steps of 0.04. For each possible combination of $\bar{x}$ and $\bar{y}$ the two series were converted to the cumulative form according to Equation (2), and the correlation between them was calculated.

https://doi.org/10.3189/S0022143000022401 Published online by Cambridge University Press

The results for one run of this model are shown in Table II. When $\bar{x}$ and $\bar{y}$ are of the same sign, the two series are strongly positively correlated, but when they are of opposite sign, strong negative correlations result. By choosing different mean values for two unrelated series of random observations and then expressing those observations in cumulative form, it is thus possible to obtain almost any desired value of r.

TABLE II. CORRELATIONS BETWEEN TWO CUMULATED RANDOM NORMAL SERIES AS A FUNCTION OF THE MEANS, $\bar{x}$ AND $\bar{y}$, OF THOSE SERIES. THE MEANS OF THE NON-CUMULATED SERIES ARE 0.0151 AND 0.0080, AND THE CORRELATION BETWEEN THEM IS −0.0295

						$\bar{y}$					
	Means	−0.20	−0.16	−0.12	−0.08	−0.04	0.00	0.04	0.08	0.12	0.16	0.20
	0.20	−0.972	−0.969	−0.960	−0.935	−0.811	0.199	0.855	0.936	0.957	0.965	0.969
	0.16	−0.963	−0.959	−0.951	−0.927	−0.804	0.194	0.845	0.926	0.946	0.955	0.959
	0.12	−0.946	−0.942	−0.935	−0.911	−0.793	0.185	0.827	0.907	0.928	0.936	0.940
	0.08	−0.910	−0.907	−0.900	−0.878	−0.767	0.169	0.790	0.869	0.890	0.898	0.902
	0.04	−0.818	−0.816	−0.811	−0.793	−0.698	0.135	0.701	0.775	0.795	0.803	0.808
$\bar{x}$	0.00	−0.532	−0.532	−0.531	−0.523	−0.472	0.049	0.433	0.489	0.505	0.512	0.516
	−0.04	0.170	0.167	0.161	0.148	0.102	−0.119	−0.198	−0.196	−0.193	−0.191	−0.190
	−0.08	0.691	0.686	0.676	0.651	0.543	−0.212	−0.649	−0.693	−0.702	−0.705	−0.705
	−0.12	0.862	0.857	0.847	0.819	0.693	−0.232	−0.790	−0.851	−0.865	−0.870	−0.872
	−0.16	0.923	0.918	0.907	0.879	0.748	−0.235	−0.838	−0.906	−0.922	−0.928	−0.930
	−0.20	0.949	0.944	0.934	0.906	0.773	−0.235	−0.858	−0.930	−0.947	−0.953	−0.956

Indeed, it is not necessary that the two series be unrelated in order to be able to select r at will. The two simple examples in Table III and Figure 2 show that it is quite easy to completely reverse the sense of a relationship by using cumulative series instead of raw data.

The high correlations between cumulative series reported by Andrews and others (1971) result from the fact that they used random numbers with a mean value of 50 for both sets of observations. By choosing different mean values for their initial series, they could have obtained any value they wanted.

Another quirk of the correlation coefficient between cumulated series is that it depends to a certain extent on the order in which the observations are cumulated. Only the final point in the cumulated series has a fixed value for a given set of points; the other points may assume different values depending on which point is chosen as the initial one and the sequence of the points which follow. When non-cumulated series are correlated, the order in which the pairs of observations are taken does not affect the correlation coefficient; in the case of cumulated series, however, variations in the magnitude of the coefficient do occur when different orders of accumulation are followed.

TABLE III. TWO SETS OF DATA WHICH SHOW REVERSAL OF THE CORRELATION COEFFICIENT WHEN THE DATA ARE TRANSFORMED FROM RAW TO CUMULATIVE SERIES. TOP, DATA FOR FIGURE 3A; BOTTOM, DATA FOR FIGURE 3B

	Raw data		*Cumulative data*	
Observation	x	y	x	y
1	−1	7	−1	7
2	−3	5	−4	12
3	−6	2	−10	14
4	−4	4	−14	18
5	−5	3	−19	21
6	−2	6	−21	27
	$r = +1.0$		$r = -0.969$	
1	1	7	1	7
2	6	2	7	9
3	3	5	10	14
4	5	3	15	17
5	2	6	17	23
6	4	4	21	27
	$r = -1.0$		$r = +0.968$	

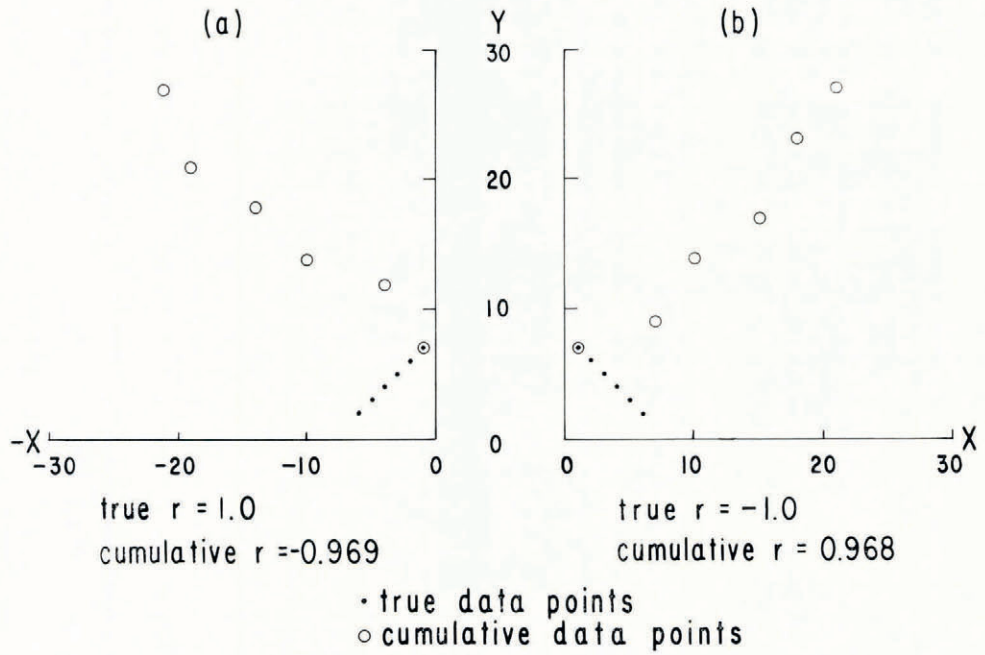

Fig. 2. Two examples of the reversal of the correlation coefficient between two variables when the data are transformed from raw to cumulated series.

In summary, the transformation of a set of observations to cumulative form destroys the independence of the observations and makes the correlation coefficient strongly dependent on the scale used for measurement and on the length of the series. Correlating cumulated series is thus a procedure whose use should be restricted to special circumstances or completely eliminated.

ACKNOWLEDGEMENTS

A portion of this work was supported by NSF Grant GA-4128 to Valmore C. LaMarche. Computer time was supplied by the University of Arizona Computer Center. I thank John Sims for helpful criticism of the manuscript.

MS. received 10 December 1971

REFERENCES

Andrews, J. T., *and others*. 1971. Note on correlation coefficients derived from cumulative distributions with reference to glaciological studies, by J. T. Andrews, B. D. Fahey and D. Alford. *Journal of Glaciology*, Vol. 10, No. 58, p. 145–47.

Mitchell, J. M., *jr., and others*. 1966. Climatic change. Report of a working group of the Commission for Climatology prepared by J. M. Mitchell, Jr., chairman, B. Dzerdzeevskii, H. Flohn, W. L. Hofmeyr, H. H. Lamb, K. N. Rao, C. C. Wallén. *World Meteorological Organization. Technical Note* No. 79. (WMO No. 195. TP.100.)

Naylor, T. H., *and others*. 1966. *Computer simulation techniques, by T. H. Naylor, J. L. Balintfy, D. S. Burdick and K. Chu*. New York, John Wiley and Sons, Inc.