Journal of Hydrology (NZ) 53 (2): 167-169, 2014

NOTE

Tendency toward negative correlations for positively-skewed independent random variables

Earl Bardsley

School of Science, University of Waikato, PO Box 3105, Hamilton. Corresponding author: e.bardsley@waikato.ac.nz

Abstract

Hydrological variables often display positive skewness, with means exceeding medians. Even when two such variables are independent, there is a more than 50% chance that a bivariate random sample will yield a negative value of the Pearson product-moment linear correlation coefficient *r*. Simulations from inverse Gaussian distributions suggest that this effect is small and unlikely to have any effect on significance tests of *r* if two skewed variables are in fact independent. However, a high frequency of negative sample correlations in a multi-site hydrological study could result in an incorrect deduction that some weak but widespread physical process is in operation.

Keywords

correlation coefficient, inverse Gaussian, data simulation, median

Introduction

Calculation of the sample Pearson product-moment linear correlation coefficient *r* is common practice for *x-y* scatter plots in hydrology. However, there may be estimation bias introduced for the positively-skewed data typical of hydrology (Habib *et al.*, 2001). This brief communication draws attention to the specific situation where positively skewed independent data will yield negative values of *r* more often than not, even though the true correlation is zero. In certain circumstances this effect could be misinterpreted as a discovery of a weak but real association between the variables concerned.

Scatter plots with skewed data

Suppose both X and Y are independent random variables (zero correlation) with positive skewness, such that their respective means exceed their medians. That is, $E(Y)$ is a constant independent of X. The left portion of an X-Y scatter plot will contain the greatest number of data points and therefore (most often) the largest values of Y. In contrast, at the right extreme of the data scatter, the few data points involved will have a tendency to be less than $E(Y)$. Suppose, for example, that Y is an exponential random variable and there is just a single value of Y within a given range of the x axis at the right side of the scatter plot. In this case $\text{Prob}[Y < E(Y)] = 1 - 1/e = 0.63$. However, if the same range of x at the left side of the scatter plot contains 10 values of Y, then the probability that the largest of these will be less than $E(Y)$ is just $0.63^{10} = 0.01$. The combination of left and right influences of this type will result in a greater frequency of negative values of r from random samples with positively skewed independent data.

Simulation example

The tendency toward negative correlation values is illustrated in the simulation results shown in Table 1. Both X and Y values in each simulation are independent random variables from a common two-parameter inverse Gaussian distribution. That is, for a given simulation, both X and Y values are generated from the same distribution. The selected sample sizes utilised were n = 10, 30, and 100 and the respective distribution skewness values were g_1 = 10, 5, 3, and 2, where g_1 is the Fisher-Pearson skewness measure. The simulations generated 100,000 random bivariate X-Y samples, over all combinations of skewness and sample size. The proportion of negative r values is shown for each simulation.

Table 1 – Simulation results: medians r' of the distribution of r and probabilities q of obtaining negative correlation coefficients when X and Y are independent random variables from common inverse Gaussian distributions (zero true correlation) for varying degrees of skewness g_1 and sample size n.

g_1	q	r'	n
10	0.66	-0.12	10
10	0.69	-0.06	30
10	0.68	-0.03	100
5	0.60	-0.09	10
5	0.61	-0.04	30
5	0.58	-0.02	100
3	0.56	-0.05	10
3	0.55	-0.03	30
3	0.54	-0.01	100
2	0.53	-0.03	10
2	0.53	-0.01	30
2	0.52	-0.01	100

The probability of obtaining a negative r value reaches almost 70% for very strong skewness and the sample size evidently has minimal effect over the range of n considered. This slow reduction in bias with increasing sample size has also been noted for lognormal distributions (Lai *et al.*, 1999; Habib *et al.*, 2001) and presumably applies to many positively skewed unimodal distributions of Y. In contrast, skewness is a dominant factor in the probability of generating a greater number of negative r values from simulated samples, with the tendency toward negative r almost disappearing for g_1 = 2.0.

Discussion and conclusion

Taking the inverse Gaussian results as indicative of hydrological data, the skewness effect in x-y plots will in fact have a negligible influence on producing values of r that are sufficiently negative so as to give statistically significant false negative correlations. This is because although the probability of obtaining a negative r may be high, the r values themselves are almost always near zero.

However, many hydrological studies do not evaluate just single x-y plots but may instead include a number of such plots derived over an extended land area, as might arise in spatial studies of rainfall. It could happen that some climatic variable X exhibits apparent negative correlation with seasonal rainfall Y in the sense of a preponderance of negative correlation coefficients over multiple sites. Although statistically significant correlation may not be evident at any one site, for a sufficiently large number of sites the application of a binomial test would establish that the proportion of negative correlation values exceeds 0.5 with a high level of significance. That is, the p-level involved could be much lower than 0.05 when viewed in terms of the frequency of negative values of r. This might be reported as the climatic variable concerned having a 'weak but strongly significant negative

regional association with rainfall'. In reality this would be a Type 1 error, though not obviously so, given a low probability p of a Type 1 error for the binomial test.

In the event of an evident spatial field of low but consistent negative correlations with positively skewed data, the best verification approach would be to carry out some transformation of r with improved properties of sampling distribution. The most well-known transformation of r is Fisher's z transformation (Fisher, 1921). However, as noted by Habib *et al.* (2001), this is not particularly useful for strongly skewed data. Habib *et al.* (2001) present an alternative transformation with respect to rainfall data application, using previous work by Stedinger (1981) for river discharge data. Withers and Nadarajah (2010) present general expressions which include application to improved estimators of the correlation coefficient in the presence of skewness of independent variables.

Acknowledgement

The author thanks two independent reviewers for useful comments on the original manuscript.

References

Fisher, R.A. 1921: On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron 1*: 3-32.

Habib, E.; Krajewski, W.F.; Ciach, G.J. 2001: Estimation of rainfall interstation correlation. *Journal of Hydrometeorology 2*: 621-629.

Lai, C.D.; Rayner, J.C.W.; Hutchinson, T.P. 1999: Robustness of the sample correlation - the bivariate lognormal case. *Journal of Applied Mathematics & Decision Sciences 3*: 7-19.

Stedinger, J.R. 1981: Estimating correlations in multivariate streamflow models. *Water Resources Research 17*: 200-208.

Withers, C.; Nadarajah, S. 2010: The bias and skewness of *M*-estimators in regression. *Electronic Journal of Statistics 4*: 1-14. Corrigendum: http://freepages.misc.rootsweb.com/~kitwithers/research/2010b.pdf

Manuscript received ???????; accepted for publication ???????