Correlation of Pearson using datasets of different dates

Ananda Banerjee

B.C. Roy Road, Near Shyamkhola More
PO South Jagaddal, South 24 Parganas, West Bengal, India, Pin 700151

***Abstract*: Many natural phenomena are dependent upon rainfall. Such as, flowering of trees or fruit body formation of mushrooms or egg laying of frogs. But the time gap between the cause and the effect is not fixed. In some cases it is just few hours, and in other cases it may be few days. In a particular area two species of trees suppose A and B start blooming after rainfall. Species A starts blooming within few hours of rainfall and species B starts blooming after one day or more. In human perception both have positive correlation with rainfall. But it is very difficult to prove statistically, especially using statistical software. By modifying the Pearson's correlation coefficient slightly it can be solved uniquely. Though it is a very small modification in the formula, such thing has not been reported earlier and this calculation is available in no readymade computer software. In Ecological Studies this formula will be very useful in solving such issues.**

***Keywords*: Ecology, Pearson's correlation**

Introduction: The Pearson's Correlation Coefficient (Stigler 1989, Pearson 1895, Pearson 1920) is widely used in the different areas of sciences. It measures linear correlation between two variables *X* and *Y*. Its value ranges between +1 and −1, where 1 is total positive linear correlation, 0 is no correlation, and −1 is total negative linear correlation. It was first invented by Francis Galton in the 1880s and later developed by Karl Pearson.

In agricultural science, accountancy, ecology, environmental sciences, medical sciences, psychology, risk analysis, social sciences etc., Correlation of Pearson is one of the most highly used statistical tool. Though it is very easy to calculate, in some situations this tool becomes inoperable. Rigidity of the method is that the data of the dependent and independent variables must be arranged in the same sequence. Correlation of Pearson of dependent variable cannot be calculated if the independent variable is arranged in a different sequence; though situation often demands such requirement.

In nature many phenomena are dependent upon rainfall. It may be flowering of trees or fruit body formation of mushrooms or egg laying of frogs. Time gap between the cause and effect is not fixed. In some cases it is just few hours, and in some other cases it may be few days.

Suppose, in a particular area two species of trees A and B start blooming after rainfall. Species A starts blooming within few hours of rain fall and species B starts blooming after one day. In human perception both have positive correlation with rainfall, but it is very difficult to prove statistically, especially using statistical software.

(Case I): Number of flowers in tree (A) is seen to be positively correlated with the amount of rainfall of that day. Immediately after rainfall flowers start blooming.

	Column X	Column Y
Dates	Rainfall in mm.	Number of flowers bloomed in tree (A)
1	1	1
2	2	2
3	1	1
4	2	2
5	25	24
6	2	2
7	1	1
8	2	2
9	18	17
10	1	1

Here, Pearson's Correlational coefficient r = 0.999

(Case II): Number of flowers in tree (B) is seen to be positively correlated with the amount of rainfall. But in this case it takes some time to bloom the flowers. So, here the same amount of flower blooms, but one day after the rainfall.

	Column X	Column Z
Dates	Rainfall in mm.	Number of flowers bloomed in tree (B)
1	1	1
2	2	1
3	1	2
4	2	1
5	25	2
6	2	24
7	1	2
8	2	1
9	18	2
10	1	17

Here, Pearson's Correlational coefficient r = -0.204

In case of tree (B), in naked eye though it is visible that flowering is totally dependent upon rainfall, but using pearson's correlation formula it is not possible to prove.

Methodology:

A ten day study period was considered for analysis. In case (I) in column X, amount of rainfall and in column Y, number of 'A' flowers bloomed on the corresponding dates are expressed.

In case (II) in column X, amount of rainfall and in column Z, number of 'B' flowers bloomed on the corresponding dates are expressed.

xi= Rainfall of first day.

yi= Number of flowers bloomed in tree (A) on the first day

zi= Number of flowers bloomed in tree (B) on the first day

zii= Number of flowers bloomed in tree (B) on the second day

In simple correlation or Pearson's correlation

$$r = \frac{\mathrm{cov}(X,Y)}{s_x s_y}$$

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

To solve the problem, in case II, the formula is slightly modified and Xi is compared with Zii, instead of Zi.

$$r = \frac{\sum x_i z_{ii} - n\bar{x}\bar{z}}{(n-1)s_x s_z}$$

Result: Using the new modified Pearson's formula, for Case II also r is found as 0.999.

Though it is a very small adjustment in the formula, its implication is big. This tool has not been reported earlier and in no readymade computer software this calculation is available. In Ecological Studies this formula can be very useful in solving such issues. One can determine the maximum effect of rainfall taking place actually on which day by comparing data of Xi with Zi,

Zii, Zii, Ziv.....and so on. Though here the problem is solved using an ecological incident, it can be used in any other fields ranging from economics to sociology.

Acknowledgement: I thank Nazma Banerjee for encouragement.

References:

[1] Stigler, S.M. 1989. Francis Galton's Account of the Invention of Correlation. Statistical Science, 4, 73-86.
[2] Pearson, K. 1895. Notes on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London, 58, 240-242. https://doi.org/10.1098/rspl.1895.0041
[3] Pearson, K. 1920. Notes on the History of Correlation. Biometrika, 13, 25-45.
Stable URL: http://www.jstor.org/stable/2331722