A Dreadful Secret of Pearson's *r*

Un secret épouvantable de *r* de Pearson: Un comportement aberrant du coefficient de la corrélation de Pearson quand les nombres des catégories des deux variables sont inégaux.

Kenpei SHIINA[1], Yoshihiro OUCHI[2], Saori KUBO[1], Takashi UEDA[1]

[1]Waseda University, [2]Josai International University

Key Words: Pearson's *r*, aberrant behavior

Pearson's correlation coefficient *r* is a standard scientific tool for measuring the correlation between any variables X and Y and its range is known to be [-1,1]. This paper reports an aberrant behavior of Pearson's correlation coefficient.

We found that *r* cannot attain -1 or 1

in ordinary situations in, but not limited to, social sciences that use rating scales, where

1) **variable X has $m \geq 2$ ordered categories and Y has $n \geq 2$ ordered categories,**
2) **the *m* and *n* categories are used at least once, and**
3) ***n* is different from *m*.**

The proof is simple.

Proof: Without loss of generality, we prove the case where *r* is positive. Let the number of data pairs be $D \geq \min(m, n)$. Suppose there exists a data set that satisfies the three conditions and $r = 1$. In order to attain $r = 1$, the *D* pairs should be on a straight line, *L*, with a positive slope in X-Y scatter diagram (*L* should not be perpendicular to X and Y axes because in such a case *r* cannot be computed due to zero variance.) If we consider the orthogonal projections of the *D* points on *L* onto X and Y axes, it is necessary that the number of projected points be the same for X and Y axes with a rule that if projected points overlap on an axis they should be counted as one. This contradicts condition 3) mentioned above, implying that when $m \neq n$ some of the *D* pairs do not lie on *L* and thus *r* cannot equal one. **Q.E.D.**

The case with negative *r* can be proved similarly. Note that this effect can be observed even when $m = n$ if the numbers of categories actually used are unequal. Strictly speaking, this phenomenon is not limited to the "coarse" situations that use rating scales, because, considering the number of significant figures of measurement devices, all measurements in science are analogous to category ratings with very large number of categories and thus can fall into this pitfall.

Karl Pearson, the inventor of *r*, did notice that when the categories are "broad," *r* is biased and this problem has been studied in sociology and psychology. Our finding is related to this problem, although it is different because in Pearson's case *r* can be 1 or -1 if $m = n$, irrespective of broadness of categories. Surprisingly, we did not find any clear statement in the literature indicating that *r* cannot equal -1 or 1 when $m \neq n$ regardless of the width of the category.

An important purpose of computing *r* is to estimate the population parameter ρ. Our proof implies that if $\rho = 1$ and the three conditions are satisfied, the estimate *r* should be smaller than 1 and thus is biased. We therefore numerically investigated the magnitude of the bias and the extent of its possible impact on the estimations when $\rho < 1$. To investigate this *unequal category bias*, we conducted a simulation. We generated 1024 pairs of two-dimensional Gaussian random numbers with specific values of ρ and *D*, categorized the numbers into *m* categories (on X axis) and *n* categories (on Y axis), and calculated *r* using integer category scores assigned to the categories. Table 1 shows mean *r*'s as a function of ρ, *m*, *n*, and *D*, averaged over 1024 categorized random number pairs. We can observe that a) the categorization lowers *r* substantially, and b) the unequal category bias is large especially when *m* and *n* are small and ρ is very high. We therefore recommend using as many categories as possible with $m = n$. If this is impossible, researchers will run the risk of having lowered estimates of ρ which may not pass statistical tests, and end up with unsuccessful research. Since millions of researchers and practitioners regularly use *r*, we hope such cases have been few in the past and will decrease in the future in the light of our findings.

Table 1 Estimated *r*'s. Italicized numbers show clear decreasing due to the unequal category bias. D = 1024.

	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.96$	$\rho = 0.98$	$\rho = 1.0$
$m = 3, n = 3$	0.615	0.728	0.828	0.878	1.00
$m = 3, n = 4$	0.629	*0.704*	*0.748*	*0.760*	*0.764*
$m = 3, n = 5$	0.658	0.743	*0.786*	*0.792*	*0.794*
$m = 3, n = 6$	0.669	0.755	*0.808*	*0.824*	*0.832*
$m = 4, n = 4$	0.729	0.835	0.904	0.933	1.00
$m = 4, n = 5$	*0.718*	*0.810*	*0.865*	*0.881*	*0.892*
$m = 4, n = 6$	0.732	*0.827*	*0.886*	*0.909*	*0.956*
$m = 5, n = 5$	0.733	0.828	0.893	0.925	1.00
$m = 5, n = 6$	0.745	0.840	0.897	*0.916*	*0.937*
$m = 6, n = 6$	0.758	0.854	0.915	0.940	1.00

A brief history

Pearson (1913) had noticed the "broad category" problem. Martin (1978) simulated the *broad category problem* and concluded: "The findings suggest that the amount of lost information is substantial." Bollen and Barb (1981) performed a similar simulation and arrived at a similar conclusion. Further, Bollen and Barb (1983, *American Sociological Review*, **48**, p.286) noticed that "more complex patterns occurring when the collapsed variables do not have the same number of categories."

Prior to the present paper, no study seems to have clearly pointed out the bias arising from unequal numbers of categories.