The Non-Transitivity of Pearson's Correlation Coefficient: An Educational Perspective

Castro Sotos, Ana Elisa
Centre for Methodology of Educational Research, Katholieke Universiteit Leuven
Andreas Vesaliusstraat 2
B-3000 Leuven, Belgium
E-mail: anaelisa.castrosotos@ped.kuleuven.be

Vanhoof, Stijn
Centre for Methodology of Educational Research, Katholieke Universiteit Leuven
Andreas Vesaliusstraat 2
B-3000 Leuven, Belgium
E-mail: stijn.vanhoof@ped.kuleuven.be

Van den Noortgate, Wim
Centre for Methodology of Educational Research, Katholieke Universiteit Leuven
Andreas Vesaliusstraat 2
B-3000 Leuven, Belgium
E-mail: wim.vandennoortgate@kuleuven-kortrijk.be

Onghena, Patrick
Centre for Methodology of Educational Research, Katholieke Universiteit Leuven
Andreas Vesaliusstraat 2
B-3000 Leuven, Belgium
E-mail: patrick.onghena@ped.kuleuven.be

1. Introduction

Since its origins, research on statistics education has focused on misconceptions concerning main statistical concepts (see Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, in press, for a literature review on misconceptions of statistical inference). Based on their empirical studies, Batanero, Estepa, Godino, and Green (1996) and Estepa and Sánchez Cobo (2001, 2003) provided a classification for the most common students' misconceptions about correlation, namely the causal, deterministic, local, and unidirectional misconceptions. In our experience with students of statistics, we have repeatedly observed a misconception that has not been documented systematically and empirically from an educational perspective so far. It concerns the so-called non-transitivity property of Pearson's correlation. Unlike many common mathematical binary relations, correlation is not transitive. As already pointed out by McNemar in 1949, this non-transitivity property implies that, given three quantitative random variables *X*, *Y*, and *Z*, a positive correlation between *X* and *Y* and a positive correlation between *Y* and *Z* (in terms of Pearson's correlation coefficients, $\rho_{XY} > 0$ and $\rho_{YZ} > 0$), not necessarily mean that *X* and *Z* will be positively correlated. In fact, *X* and *Z* might be uncorrelated ($\rho_{XZ} = 0$) or even negatively correlated ($\rho_{XZ} < 0$). Langford, Schwertman, and Owens (2001) showed that transitivity for positive correlations only holds under very restrictive conditions, more specifically, when the sum of the squared correlations is larger than one $\left(\rho_{XY}^2 + \rho_{YZ}^2 > 1\right)$. In the social, educational, and behavioral sciences, correlations of a magnitude that satisfy this condition are very rare. The widely accepted standards established by Cohen (1988), for example, consider correlation coefficients of 0.1, 0.3, and 0.5 as small, medium, and large, and none of those values (even up to 0.7) would satisfy the necessary condition above. The mathematical boundary for transitivity is much higher than correlations that can be expected in practice.

An alternative way of looking at the non-transitivity property is by squaring the correlations. The square of a Pearson's coefficient is the amount of shared variance between variables. By representing the amounts of shared variance via Venn diagrams (as is done in Figure 1 for $\rho_{XY} = \rho_{YZ} = 0.5$), it is easier to see that even if both X and Z are positively correlated with a variable Y, X, and Z might be uncorrelated, as shown in the figure by an empty intersection of X and Z.

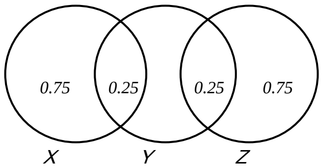

Figure 1: Non-Transitivity as Shared Variance for $\boldsymbol{\rho_{XY} = \rho_{YZ} = 0.5 \left(\rho_{XY}^2 = \rho_{YZ}^2 = 0.25\right)}$

2. Research Aim and Questions

The study presented here aimed to explore whether university students of statistics know that correlation is not transitive and how often the transitivity misconception appears. We also attempted to answer three additional questions. First, does the type of representation (Pearson's coefficients or shared variances) make a difference in the number of students who fall into the misconception? Second, is there an effect of the size of the correlations involved? Finally, does a context for the problem (vs. a theoretical wording) have an influence on the misconception? It can be argued that the perception of such property might be greatly influenced by the contextualization (e.g., correlations between school grades) and we wanted to identify the misconception independently from the context.

3. Methods and Procedures

A questionnaire including a multiple-choice problem that addressed the research questions was constructed and distributed in January 2007 to 279 university students (182 females, 97 males) from different disciplines (Mathematics, Psychology, Medicine…) who were following statistics courses at the Universidad Complutense de Madrid (Spain). Most participants were in their first year (70.61%) and had never followed a university statistics course before (64.87%). The questionnaires were completed during class time and there were several versions according to three factors: the way the transitivity problem was stated (in terms of Pearson's coefficients or shared variances), the values of the correlations for *XY* and *YZ* (0.3, 0.5, or 0.7), and whether the problem was contextualized or not (the context of the problem, when available, referred to school grades in Mathematics, English, and French). The combination of these factors resulted in 12 different versions of the problem; each of them was presented to 22 to 24 students. The problem presented the students with the (squared) correlations between *X* and *Y*, and *Y* and *Z*, and asked them to select all correct answers from five options about the correlation between *X* and *Z*: a) It will be positive, b) It can be zero, c) It will be higher than 0.5, d) It can be negative, and e) It will be between –1 and 1. A simple coding for the answers was used. Every time a student selected options a) or/and c), the transitivity misconception was considered to be "present". Any other combination of item options was coded as "no transitivity". "Perfectly correct answers" were those given by the combination b) + d) + e).

After an exploratory analysis of the results, we constructed a log-linear model for the variable "falling into the transitivity misconception" (values 1 or 0) and factors: type of correlation (Pearson's coefficients or shared variances), size of correlation (0.3, 0.5, or 0.7), and contextualization (context or context-free problem).

4. Results

34.77% of the students fell into the transitivity misconception and only a small number of students answered the problem completely correct (see Table 1).

Table 1: Numbers and Percentages of Students for the Different Answers to the Transitivity Problem.

Missing	Transitivity Present	No Transitivity		Total
		Perfectly Correct	Other	
6 (2.15 %)	**97** (34.77 %)	**23** (8.24 %)	153 (54.84 %)	279 (100 %)

Table 2 shows the proportional distribution of the transitivity misconception for the three factors that were manipulated in our study. A log-linear analysis resulted in three main effects (see Table 3). First, students who were confronted with the problem stated in terms the Pearson's correlation coefficient fell into the misconception more often (54%) than those who had to answer the problem in terms of shared variances (only 18%). In addition, the larger the size of the correlation, the higher the percentage of students committing the transitivity misconception (23% for correlation equal to 0.3, 34% for 0.5, and 50% for 0.7). Finally, students who dealt with contextualized problems appeared to perform better (only 26% fell into transitivity) than those who solved a context-free problem (46%).

Table 2: Number of Students for the Different Answers to the Transitivity Problem According to Factors.

Type of Correlation	Transitivity Present	Transitivity Not Present	Size of Correlation	Transitivity Present	Transitivity Not Present	Context	Transitivity Present	Transitivity Not Present
Pearson	**73** (53.68%)	63 (46.32%)	0.3	**21** (22.58%)	72 (77.42%)	Context	**36** (25.90%)	103 (74.10%)
Shared	**24** (17.52%)	113 (82.48%)	0.5	**31** (34.44%)	59 (65.56%)	Context-free	**61** (45.52%)	73 (54.48%)
			0.7	**45** (50%)	45 (50%)			

Moreover, all two-way interaction effects were also statistically significant (see Table 3). The three-way interaction effect was not statistically significant $\left(\chi^2 = 3.43,\ \ p = 0.1796\right)$.

Table 3: Log-Linear Analysis

Source	DF	Chi-Square	Pr > ChiSq	Source	DF	Chi-Square	Pr > ChiSq
Type of Correlation	1	29.83	**<0.0001**	Type x Size	2	15.71	**0.0004**
Size of Correlation	2	15.05	**0.0005**	Type x Context	1	4.94	**0.0263**
Context	1	11.96	**0.0005**	Size x Context	2	10.82	**0.0045**

First, with regard to the interaction between the type and the size of the correlation, the difference between the amount of students who fall into the misconception for values 0.3 and 0.5 is much larger (and in the opposite direction) for items in terms of Pearson's correlation coefficients (28%-70%) than for shared variances (17%-0%). On the other hand, the difference between values 0.5 and 0.7 of the correlation is much smaller (and in the opposite direction) for Pearson's coefficients (70%-64%) than for shared variances (0%-36%). Second, the difference in the amount of students who make the transitivity error for contextualized and context-free problems is much smaller for problems with Pearson's coefficients (39%-69%) than for shared variances (13%-23%). Finally, the increase in the number of students who fall into the misconception depending on the size of the correlation is almost non-existent for context-free problems (44%-43%-49%) whereas for contextualized problems it is very obvious (2%-26%-51%).

5. Discussion and Conclusion

In our study we confirmed the existence of a misconception with respect to Pearson's correlation coefficient, which has not been systematically and empirically documented until now. A large number of students appears to think that, given three quantitative random variables X, Y, and Z, a positive correlation between X and Y, and a positive correlation between Y and Z, X and Z are always positively correlated.

This transitivity misconception was more evident when the problem was stated in terms of Pearson's correlation coefficient than when the problem was stated in terms of shared variances. It is possible that mental representations similar to Venn diagrams act as safeguards against falling into the misconception. An alternative explanation could be that the familiarity of Pearson’s coefficient makes it easier or more tempting to extrapolate popular and intuitive mathematical properties to this statistic. Furthermore, when the sizes for the two given correlations were manipulated, we observed that higher correlations provoked more transitivity which would fit with the idea that transitivity becomes more plausible for increasing correlations. Finally, students fell into the misconception more often when they were confronted with context-free problems as opposed to contextualized situations, which may be explained by students’ difficulties with statistical vocabulary as opposed to a better understanding of the question in contextualized situations.

These results suggest that an important step in the teaching of correlations might be missing in university statistics courses or that more stress would be needed on the issue of transitivity, since correlational studies are widely spread among all sciences as main tools for analysis and interpretation of real data. If students are expected to be able to understand and carry out this type of studies after a university statistics course, the instructor could better make sure that students understand that Pearson’s correlation is not a transitive relationship.

REFERENCES

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 2*, 151-169.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (in press). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Estepa, A. & Sánchez Cobo, F. T. (2001). Empirical research on the understanding of association and implications for the training of researchers. In C.Batanero (Ed.), *Training researchers in the use of statistics* (pp. 37-51). Granada, Spain: International Association for Statistical Education and International Statistical Institute.

Estepa, A. & Sánchez Cobo, F. T. (2003). Evaluación de la comprensión de la correlación y regresión a partir de la resolución de problemas [Evaluation of the comprehension of correlation and regression through problems' resolution]. *Statistics Education Research Journal, 2*, 54-68.

Langford, E., Schwertman, N., & Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician, 55*, 322-325.

McNemar, Q. (1949). *Psychological statistics*. New York: John Wiley and Sons.