

BUILDING HIGH SCHOOL PRE-SERVICE TEACHERS' KNOWLEDGE TO TEACH CORRELATION AND REGRESSION

Carmen Batanero¹, Maria M. Gea¹, Carmen DÃ-az² and Gustavo R. CaÃ±adas¹

¹University of Granada, Spain

²University of Huelva, Spain

batanero@ugr.es

In this paper we describe a workshop which was aimed to develop the knowledge needed to teach correlation and regression in high school teachers. The workshop was based on a formative cycle model used in our previous research and directed to simultaneously increase the teachers' statistical and pedagogical knowledge. To develop the teachers' knowledge of correlation and regression we proposed them to complete a statistical project based on real data taken from the UNESCO web server, followed by a collective discussion of their solutions to the tasks proposed in the project. A didactical analysis of the project helped them to increase their pedagogical content knowledge. Some results in a sample of prospective Spanish teachers will be briefly reported.

INTRODUCTION

New curricular reforms that emphasize statistical reasoning and its role in decision making and professional work are recently being introduced in many countries (e.g., NCTM, 2000). Main content in these curricula for high school in Spain (MEC, 2007) and other countries are correlation and regression, which are fundamental statistical ideas that expand the previous knowledge about univariate distributions and mathematical functions.

Research related to understanding correlation

In spite of the relevance of the topic, previous research suggests poor results in people's estimation and understanding of correlation. For example, Erlick and Mills (1967) found that negative correlation is estimated as close to zero. Other authors studied the influence of previous theories about the context of the problem on the accuracy in estimating correlation. Chapman (1967, pp. 151) described "illusory correlation" as "the report by observers of a correlation between two classes of events which, in reality, (a) are not correlated, (b) are correlated to a lesser extent than reported, or (c) are correlated in the opposite direction from that which is reported". The estimates are more accurate if people have no theories about the type of association in the data. If the subject's previous theories agree with the type of association reflected by the empirical data, there is a tendency to overestimate the association coefficient. But when the data do not reflect the results expected by these theories, the subjects are often guided by their theories, rather than by data (Jennings, Amabile, & Ross, 1982).

According to Barbancho (1992), a correlation between variables may be explained by the existence of a unilateral cause-effect relationship (one variable produces the other), but also to interdependence (each variable affects the other), indirect dependence (there is a third variable affecting both variables), concordance (matching in preference by two judges in the same data set) and spurious correlation. In addition to the estimate accuracy, understanding correlation involve the discrimination of these types of relationships between variables.

Estepa (1994) studied the conception of correlation in a sample of 213 high school students and the accuracy in the estimation of the correlation coefficient in a subsample of 51 of these students. The author defined the *causal conception* according to which the subject only considers correlation between variables, when it can be explained by the presence of a cause - effect relationship. He also described the *unidirectional conception*, where the student does not accept an inverse association, considering the strength of the association, but not its sign and assuming independence where there is an inverse association. Our research is aimed to complement Estepa's study with other analyses; moreover we focus on prospective teachers, while Estepa's research was carried out with high-school students. We base on our previous analysis (Gea, Batanero, CaÃ±adas, & Contreras, 2013) of the presentation of correlation and regression in Spanish high school textbooks.

Teacher knowledge to teach statistics

The success in the teaching of statistics will depend on the extent to which we can prepare teachers adequately (Batanero & D  az, 2012). Consequently, it is important to assess teachers' conceptions and to find suitable activities where teachers work with meaningful problems and are confronted with potential difficulties. At the same time we should develop the pedagogical knowledge related to the teaching of particular statistical topics. The aim of this paper is describing our experience in a workshop directed to empower prospective high school mathematics teachers in relation to correlation and regression and its teaching. The education of teachers is an important area of research, given the complexity of teaching and the variety of professional knowledge and competence required by the teacher, which have been described in different theoretical models.

One such model (Ball, Lubienski, & Mewborn, 2001; Hill, Ball, & Schilling, 2008) describes the *mathematical knowledge for teaching* (MKT) as the knowledge needed to teach mathematics, and considers different categories in this knowledge, which includes common and specialised knowledge of content, knowledge of content and students, teaching and curriculum. In order to adapt the MKT framework to the teaching of statistics, Burgess (2011) crossed the categories in the mathematical knowledge for teaching with each fundamental mode of statistical thinking in Wild and Pfannkuch's (1999) model (need for data, transnumeration, variation, reasoning with models, integration of statistical and contextual knowledge); in this way he described with precision the specific knowledge used by the teacher to teach statistics. The workshop described below is directed to increase these components of teacher' knowledge.

TRAINING TEACHERS TO TEACH CORRELATION AND REGRESSION: THE LIFE EXPECTANCY PROJECT

The workshop analyzed in this presentation was directed to prospective Spanish high school teachers; the sample included 23 prospective teachers in a Masters' course in Mathematics Education, which is compulsory in Spain for those who want to apply for a teaching position in secondary or high school. All the students have previously finished a Bachelor in mathematics (10 students), Statistics (2), or Sciences/Engineering (12) and, consequently, their mathematical knowledge was high. However their statistical preparation was in general reduced to one statistics course with theoretical orientation (apart the 2 students with a Bachelor in statistics). None of them had previously studied statistics education contents.

The workshop design followed a formative cycle proposed by Godino, Ortiz, Roa and Wilhelmi (2011) and consisted of four 2- hour long sessions, where students had access to computers and Internet. In the first session the participants were given the statistical project: "Life expectation in different countries", where some data taken from the UNESCO web server (<http://hdrstats.undp.org/en/tables/index.html>) were used to investigate the relationship between the life expectation in different countries and different international indicators of human development. In the first session, participants were given an Excel file with the data set (9 variables in 195 countries that contained the values collected in 2009). After the lecturer presented the origin and relevance of the data, discussed with the students the meaning of each variable and described the way they were collected, students were given a set of scatter plots displaying each independent variables (in the X axis) against the Life expectancy at birth (X axis) and a related questionnaire to be filled individually. Life expectancy was the dependent variable, as the aim of the project was to study what factors influence life expectation at birth. In the second session, the students' written solutions to the questionnaire were discussed in the classroom and a debate was organised by the lecturer, in case of conflict of opinions.

In the third and four sessions a didactical analysis, was carried out in order to analyze the statistical knowledge needed to solve the project and the pedagogical content knowledge involved in teaching statistics through projects work and with this particular project; potential students' difficulties; use of materials and software, availability of Internet applets that could be used to support the students 'learning; ways to conduct the lesson, were also debated. In this paper we limit our analysis to the project activities related to understanding correlation and regression, although there were additional activities related to univariate analysis (statistical graphs, distribution, centres and spread).

Variables analyzed in the project

To start the project, the lecturer provided the students with an Excel file with the following statistical variables (the definition of these variables are reproduced from the UNESCO web site):

- *Life expectancy at birth (years)*: Number of years a newborn could expect to live if prevailing patterns of age-specific mortality rates at the time of birth stay the same throughout the infant's life.
- *Human Development Index (HDI) value*: A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living.
- *GDP per capita (2005)*: Sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products, expressed in international dollars using purchasing power parity rates and divided by total population during the same period.
- *Adolescent fertility rate*: Number of births to women ages 15–19 per 1,000 women of these ages.
- *Under-five mortality*: Probability of dying between birth and exactly age 5, expressed per 1,000 live births.
- *Expenditure on health, public*: Current spending from government (central and local) budgets, external borrowings and grants (including donations from international agencies and nongovernmental organizations), and social (or compulsory) health insurance funds, expressed as a percentage of GDP.
- *HDI education index*: takes into account distribution of years of schooling.
- *Population, total* both sexes (in thousands).
- *Population, urban*: percentage of total population living in areas classified as urban according to the criteria used by each area or country.

Table 1. Criteria used in selecting the independent variables

Variable	R value	Fitting model	Explaining the relationship	Subjects' expectation Versus data
Human Dev. Index	0,91	Lineal	Interdependence	Coincidence
GDP per capita	0,61	Logarithm	Indirect dependence	Coincidence
Adolescent fertility rate	-0,73	Lineal	Indirect dependence	No expectation
Under-five mortality	-0,92	Exponential	Cause-effect	Coincidence
Expenditure on health	0,38	Polynomial	Cause- effect	Weaker than expected
HDI education index	0,78	Lineal	Indirect dependence	No expectation
Population, total	0	Independence	Independence	No expectation
Population urban	0,62	No model	Indirect dependence	Contrary

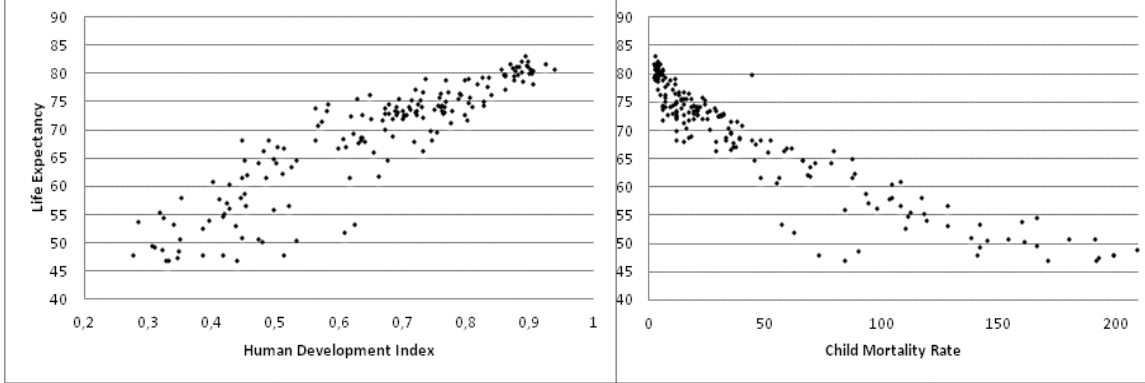
The aim of using real data was leading the prospective teachers to reasoning with evidence and multivariate data, where the relationships between the variables are not limited to linear regression. We also tried to provide them the opportunity to observe different signs and strengths of correlation in real-world situations. We hoped that this activity will help prospective teachers to increase their statistical literacy capacity to be informed citizens, and make important personal decisions in particular, related to education (Ridgway, Nicholson, & McCusker, 2006). The following criteria were taken into account (Table 1).

- *Strength of correlation*: ranging from very strong correlation to independence.
- *Sign of correlation*: including both positive and negative correlations, which are harder to be perceived by the students according to Estepa (1994).
- *Model fitting to the data*: including lineal and no lineal models.
- *Explanation of correlation*: we used relationships that may be explained by cause- and effect, as well as interdependence or indirect dependence.
- *Agreement between students' previous expectation and correlation in the data*: that may coincide or not.

Tasks given to participants

In Figure 1 we summarize the questionnaire given to the participants in the workshop. Only two scatter plots are included in Figure 1 (although similar scatter plots were displayed in the questionnaire for the 6 remaining independent variables. In questions 1 and 2 we ask the prospective teachers to estimate the correlation coefficient absolute value and sign; in question 3 they should identify the real correlation coefficient (all the data correlations as well as two additional values were included in the list). Question 4 asks the participants to reflect on the difference between correlation and causation and to find potential explanations of correlation. In question 5 we assess the prospective teacher's competence to relate the scatter plot spread with the reliability of estimation and in Question 6 we assess their competence to deduce a mathematical model that fits the data.

Variables influencing Life expectancy. In the following graphs the Life expectancy is displayed as a function of different variables¹



Questions:

1. Assign to each graph a score according your perception of the strength of the relationship; 0 means there is no relationship and 1 is the maximum strength.
2. In each graph assign a sign (+ or -) depending on whether you think the relationship is direct or inverse.
3. Below we provide some correlation coefficient values. Find out the coefficient that corresponds to the relationship between the life expectancy and each independent variable.

Lineal correlation coefficient $r =$

-0,40	1	0,78	0,91	-0,92	0,5	0,62	0,2	-0,73	0	0,38	0,61
-------	---	------	------	-------	-----	------	-----	-------	---	------	------

4. *Explaining the correlation:* Which variables have a cause-effect relationship with life expectancy? Why?
5. Order the variables according to your perception of their better utility to predict the value of life expectancy
6. Do you think it is possible to use a mathematical model (function) to estimate the value of the life expectancy, given the value of some of these variables? What type of function?

¹Similar scatter plots were provided for all the different independent variables in the file

Figure 1. Questionnaire given to participants

SUMMARY OF RESULTS

In Table 2 we summarize the responses of participants to the questionnaire that were collected in the first session of the workshop. The average estimated and assigned absolute values of correlation were in general close to its true value; with higher precision for strong correlations, and still more for strong positive correlations. Contrary to Estepa's (1994) results the estimated correlation value where there was independence in the data (correlation of Total population with Life expectancy) was close to zero. This result suggests a better perception of independence in prospective teachers as compared with high school students. There was some influence of participant's previous theories (urban population) and of negative correlations (adolescent fertility rate; under five mortality) on the accuracy of estimates; this might imply the existence of illusory correlation (Chapman y Chapman, 1967) or unidirectional conception of correlation (Estepa, 1994)

in prospective teachers. Participants were consistent in estimating and assigning correlation (providing close values in both cases) for the variables: index of human development, expenditure on health, education index and urban population (there was a significant correlation between both responses in each of these variables; $p < 0,001$).

Table 2. Summary of results

Variable	R value	Mean value			% correct		
		Estimated R abs. value	Assigned R	Order of prediction	Sign assignment	Assuming cause-effect	Fitting model
Human Dev. Index	0,91	,90	,90	26	100	87,0	91,3
GDP per capita	0,61	,72	,56	35	100	52,2	91,3
Fertility rate	-0,73	,47	-,51	61	100	26,1	78,3
Child mortality	-0,92	,80	-,86	44	100	65,2	65,2
Exp. on health	0,38	,22	,30	13	61	43,5	0,0
HDI educ. index	0,78	,65	,67	57	100	30,4	100,0
Population, total	0	,01	,03	83	21,7	0,0	0,0
Population urban	0,62	,36	,44	13	100	4,3	34,8

The sign of correlation was correctly perceived in most variables; the only doubts arose where there was independence in the data (although no sign is applicable, many participants assigned either a positive or a negative sign) and for expenditure on health, where the relationship is polynomic. The fitting model was also recognised for the majority of plots; the main difficulty arose in the recognition of a polynomial function as a model to fit the relationship between the life expectancy and expenditure on health.

The order in the power of prediction was harder to be recognized, although many students gave partly correct responses (providing a number of order close to the right response). There was also more difficulties in discriminating cause and correlation, since cause was assumed to explain the covariation for 5 out of the 8 plots by the majority of subjects; some of them argued that the linear tendency of the graph was a proof of a cause- and effect relationship.

CONCLUSION

These results suggest the prospective teachers' high level of mathematical knowledge and competence to estimate correlation and identify a model of fit. Yet, there was less understanding of the difference between correlation and causation. Some participants identified linear relationship with cause and effect; while others did not discriminate cause-effect from interdependence or existence of other variables that influence the correlation.

All these difficulties arose in the debate organised by the lecturer along session 2, where the prospective teachers could sustain their arguments in favour or against the different responses to the questionnaire. A didactic analysis was carried out individually in a written protocol by each prospective teacher in Session 3 and a general debate of this analysis was organised by the lecturer on Session 4. The following points were considered in the didactic analysis:

- *Statistical knowledge:* what statistical concepts, properties, representations, procedures and type of reasoning learn high school students when working with this project? Are the problems proposed useful to contextualise this knowledge?
- *Students' potential difficulties:* do you think that high school students possess the initial knowledge needed to work with this project? What are the expected main difficulties in dealing with the projects? How would you help students solve these difficulties?
- *Materials needed:* What resources (material and technological resources, time, etc.) are needed to work with this project? Do you think these resources are widely available in high school centres?
- *Links to other topics:* How the project relates to other areas of mathematics? How it relates to other curricular areas?

All these activities were useful to develop the teacher's knowledge to teach correlation and regression and to empower them for their future work in activities such as "figuring out what students know; choosing and managing representations of mathematical ideas; selecting and modifying textbooks; deciding among alternative courses of action" (Ball, Lubienski, & Mewborn, 2001, p. 453). Data from the didactic analyses carried out by the prospective teachers on Session 3 are currently being analyzed, and preliminary results suggest that the activity was also useful to increase their didactic knowledge. A final thought is the need to continue research that helps improve the statistics education these future teachers are receiving during their training.

ACKNOWLEDGEMENT

Project EDU2010-14947 (MICINN-FEDER) and group FQM126 (Junta de Andalucía).

REFERENCES

- Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 433-456). Washington, DC: American Educational Research Association.
- Batanero, C. & Díaz, C. (2010). Training teachers to teach statistics: What can we learn from research? *Statistique et Enseignement, 1*. www.statistique-et-enseignement.fr/ojs/
- Barbancho, A. G. (1992). *Estadística elemental moderna* (Modern elementary statistics). Barcelona: Ariel.
- Burgess, T. A. (2011). Teacher knowledge of and for statistical investigations. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 259-270). Springer Netherlands.
- Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 151-155.
- Erlick, D.E., & Mills, R.G. (1967). Perceptual quantification of conditional dependency, *Journal of Experimental Psychology, 73*, 1, 9-14.
- Estepa, A. (1993). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores* (Preconceptions on association and its evolution with computer-based teaching). Unpublished Ph.D. University of Granada.
- Gea, M. M., Batanero, C., Cañadas, G. R., & Contreras, J. M. (2013). Un estudio empírico de las situaciones-problema de correlación y regresión en libros de texto de bachillerato (Empirical study of problem-situations of correlation and regression in high school textbooks). In A. Berciano, G. Gutiérrez, A. Estepa, & N. Climent (Eds.), *Investigación en Educación Matemática XVII* (pp. 293-300). Bilbao: SEIEM.
- Godino, J. D., Ortiz, J. J., Roa, R., & Wilhelmi, M. R. (2011). Models for statistical pedagogical knowledge. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 271-282). Springer Netherlands.
- Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*, 372-400.
- Jennings, D.L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211-230). New York: Cambridge University Press.
- MEC (2007). Real Decreto 1467/2007, de 2 de noviembre, por el que se establece la estructura del bachillerato y se fijan sus enseñanzas mínimas. Madrid: Author.
- Ridgway, J., Nicholson, J., & McCusker, S. (2006). Reasoning with evidence—New opportunities in assessment. Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil: The Netherlands: International Statistical Institute.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67* (3), 223-265.