CORRELATION AND REGRESSION IN SECONDARY SCHOOL TEXT BOOKS

Antonio Estepa Castro and Francisco T. Sánchez Cobo, University of Jaén. Spain

In this paper we present results from analysing the topic of correlation and regression in 11 Spanish textbooks, for secondary education. The study consisted of analysing the organisation and theoretical exposition of the topic, as well as the tasks proposed to the students. Some teaching implications are finally presented.

INTRODUCTION

Correlation and regression are two topics of special relevance for Secondary Education because of the following reasons: a) The study of correlation and regression can be started on a descriptive level, which is adequate for Secondary School; b) These topics extend functional dependence to the concept of random dependence, fundamental nowadays in empirical science; c) Many statistical concepts and procedures are based on correlation and regression. However, an adequate teaching of these topics at secondary level is necessary for a later understanding of statistics at university and in professional work. In spite of this relevance, this subject has received scant attention in Statistical Education. Most research related to correlation and regression has been done from a psychological point of view and only refers to two by two contingency tables, and very few research refers strictly to correlation and regression.

BACKGROUND

In 1990 an educational research project on statistics association was started at the University of Granada, Spain, and, as result, we described different students' conceptions and misconceptions concerning contingency tables, scatter plots, and comparison of two samples, intuitive strategies for solving correlation and regression problems and acts of understanding regarding these problems during a teaching experiment based on the use of computers (Estepa, 1994; Estepa et al., 1995; Batanero et al., 1996; Estepa and Batanero, 1996; Estepa and Sánchez, 1996; Batanero et al. 1997).

We are presently continuing our research into the study of correlation and regression in secondary textbooks, which are fundamental for teaching according to Robitaille and Travers (1992), *"Teacher decide what to teach, how to teach it, and what sorts of exercises to assign to their students largely on the basis of what is contained in the textbook authorised for their course"* (p. 706).

Sample

The study refers to the 3rd year of Bachillerato, (17 year-old students), where the following stochastic topics are included in the curriculum approved by the Spanish Government: Random variable. Binomial distribution. Bivariate distributions. Regression line. Correlation. The sample was chosen from a total of 28 textbooks edited from 1977 to 1990 years by the main book publishers in Spain. The following criteria were used to select a sample of 11 textbooks: year of publication, publisher, and way in which correlation and regression was presented in the book.

The books were analysed from two point of view: a) Theoretical presentation of correlation and regression, and, b) Characteristic of the tasks (290 exercises, activities and problems) proposed to the students in the topic. Below we present the main results.

THEORETICAL PRESENTATION OF THE TOPIC

Seven out of eleven textbooks introduce regression before introducing correlation, which does not facilitate learning. From the mathematical and didactic points of view it would be preferable starting with the study of correlation, in other words, studying first whether a relationship between two variables exits, and in case there is such relationship, attempting to predict a variable from the other.

All the concepts and properties related to the regression include an example. However with regards to correlation, there are 4 textbooks with no examples, 4 textbooks contain examples only for between 25% and 50% of the concepts and properties exposed, 2 textbooks contain examples for more than 50 % of the content and only 1 textbook include examples for all the concepts and properties. Moreover there is only one textbook presenting the examples before the theoretical presentation of the concepts; in 3 textbooks, some examples are exposed before than the concepts and other after the concepts; in the remaining books, all the examples are included after the concepts. According the NCTM (1982) *"Problems should be placed systematically throughout the textbooks"* N.C.T.M. (1982) which is true only in 2 of the books in our sample. In the remaining textbooks the tasks always appear at the end of the topic, which is a sign of a teaching approach "theory-practice", where mathematical concepts are taught first, and afterwards the students are asked to carry out different tasks to apply these concepts.

The learning is favoured when the students know the goals of the teaching. Only 2 of 11 studied textbooks specified these goals, and in one of them, the development of the topic do not matched the proposed goals.

We found three different ways of introducing the theoretical presentation of the topic: a) Two unidimensional statistical variables (4 textbooks), which was pointed by Hawkins et al., (1992): *"There is a school of thought which believes it to more reasonable to retain at the introductory level material that involves univariate distributional assumptions (regression) and to defer until a later part of the course, or text, material which requires joint bivariate distributional assumptions (correlation)"* (pp. 52); b) One bivariate statistical variable (6 textbooks). In this case bivariate frequencies distribution, where data are collected about two features of the same individuals are included. Some statistics of marginal distributions (such as covariance) are calculated and some new concepts are introduced implicitly or explicitly(double relative frequencies, marginal distributions, etc…); c) One bivariate random variable (1 textbook). In this case, there is a more formal and abstract presentation, including the notions of bivariate random variable, bivariate function of probability, partial derivatives, etc…). Obviously these concepts are not introduced in the first two approaches.

Only 3 textbooks discussed the difference between functional and random dependence. This might influence some students' belied in only a type of dependence (functional) producing a determinist conceptions of statistics association (Batanero et al., 1996). It can also prevent the students' understanding of the extension of functional dependence to random dependence.

Only one textbook discuss the different explanations of random dependence (unilateral causal dependence, interdependence, concordance, casual covariation, indirect dependence). The omission of this important aspect might contribute to some students' identification of dependence and causality, that is, to heir causal conception of correlation (Estepa and Batanero, 1996).

There were no discussion of the difference between direct and inverse correlation in 4 textbooks and less than half the textbooks not include examples or tasks about independence. These fact can induce confusion about the type and sign of random dependence.

All the textbooks include some proof as well as the procedures to obtain regression lines. Although two-columns reporting of proofs is an useful teaching tool

when students are first introduced to writing proofs (N.C.T.M., 1991), we found not two-columns reporting of proofs in our text books.

According to Skemp (1979), two different kinds of understanding: relational and instrumental can be distinguished. We found 60% of relational definition of concepts, 34% of instrumental definitions, and 6% instrument-relational definitions.

Scatter plots were present in all the text, though 70% of them show positive correlation, 20% present negative correlation, and only 10% were examples of independence. This might contribute to the persistence of students' unidirectional conception of correlation (Batanero et al., 1996). Furthermore, there were scarce discussion in these books about the lack of proportionality of the correlation coefficient, the effect of outliers on the correlation coefficient or the interpretation of the coefficient of determination.

CHARACTERISTIC OF THE TASKS

Concerning the 290 tasks presented in the textbooks, many of them (31.7%) only require the students to perform computations from a data matrix, without neither relevant research questions nor reference to the context or variables from which the data had been taken. Only one textbook presented a research project asking the students to collect their own data to answer several questions. Other tasks used physical-chemical contexts, people's weight and height, etc....

Tasks mathematical content included regression (44.5%), correlation (32%), representing scatter plots (12%); and computing variance, averages, or typical deviation, (11.4%). Tasks of calculus were more frequent than task of interpretation, and (41.7%) of task only demanded the calculus of regression line, without previous calculus of covariance, correlation coefficient or representation of scatter plots. The distribution of task was 49.3% of computation; 18.3% of interpretation; 13.8% of graph representation; 10.3% of prediction; 2.8% of collecting and analysing data; and 5.5% comparison and

verification tasks. Calculus, therefore prevails over collecting, representing, interpreting, and processing data, which are activities of major importance in contemporary society.

According to the type of dependence 31% of tasks presented interdependence; 27.2% unilateral causative dependence; 4.1% indirect dependence; 0.3% concordance; and 37.2% had no context. The greatest percentage corresponds to lack of context, where

the students are only able to do long, routine and tedious calculations, to obtain no meaningful numbers. The high percentage of unilateral causative dependence might contribute to the students' confusion between dependence and causality, that is, to the causal conception of correlation (Estepa et al., 1996).

The kind of dependence found in 62.4% of the tasks was direct; 21.4% of tasks showed negative dependence, and 15.2% of them independence (the correlation coefficient was less than 0.03); in 1.0% there were no data in the tasks. These percentages might lead the students to acquire biases, like believing that only direct or inverse dependence exists and that independence is an exception.

The absolute value of the correlation coefficient of the tasks was less than 0.2 in 12.5% of tasks; less than 0.8 in 58.8 % of tasks; less than 0.99 in 75.6% of tasks and the remaining cases showed correlation coefficients higher than 0.99 or even equal to 1. There was a strong trend to high and correlation in the tasks, which might induce the students to acquire the belief correlation coefficient is always high in case of dependence, which is not true in social sciences.

CONCLUSION

Our study suggest that some misconceptions the students acquire about correlation and regression might be spread or induced by some text books, since we have found indicators suggesting a teaching that support biased learning and misconceptions: studying regression before than correlation; formal methodological approach (theory-practice); lack or inappropriate treatment of fundamental concepts; strong biases when presenting several concepts (positive dependency versus negative dependency and independence, high correlation coefficients, types of dependence, too strong emphasis on calculation, ...). Consequently, we believe that writing a text book would require a deep epistemological analysis of the concepts; taking into account educational research results on the topic; favouring the construction of knowledge by the student by using adequate tasks; taking into account age, and previous knowledge. This is, according Freudenthal (1991) the way in which text books can be the motor of change in teaching/learning processes.

REFERENCES

Batanero, M. C., Estepa, A., Godino, J. D. , and Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27(2),* 151-169.

Batanero, C., Estepa, A., and Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and learning Statistics. 1996 IASE Round Table Conference Papers,* (pp. 191-205). Voorburg: International Statistical Institute.

Estepa, A. (1.994). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores* [Preconceptions of statistical association and their evolution as result of computer-based teaching] Ph. D. University of Granada. Barcelona: ETD Micropublicaciones, S.L.

Estepa, A., and Batanero, M.C. (1995). Concepciones iniciales sobre la asociación estadística. (Preconceptions about statistical association). *Enseñanza de las Ciencias, 13(2),* 155 - 170.

Estepa, A., and Batanero, M. C., (1996). Judgments of correlation in scatter plots: An empirical study of students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education, 4,* 25-41.

Estepa, A., and Sánchez, F.T. (1996). Association judgements in the comparison of two samples. In L. Puig y A. Gutiérrez (Eds.) *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education* (v. 2, pp. 337-344). University of Valencia.

Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht: Kluwer

Hawkins, A. Jolliffe, F., and Glickman, L. (1992). *Teaching statisitical concepts*. London: Longman

N.C.T.M. (1989). Curriculum and evaluation STANDARDS for school mathematics. Reston (VA): N.C.T.M.

N.C.T.M. (1982). How to evaluate mathematics textbooks. Reston (VA): N.C.T.M. Robitaille, D. F., and Travers, K. J. (1992). International studies of achievement in Mathematics. In D. A. Grows (Ed.*). Handbook of research on mathematics teaching and learning* (pp. 687-709). New York: MacMillan Publishing Company

Skemp, R. (1979). *Intelligence, learning and action*. Chichester: Willey