UNDERSTANDING GRAPHICAL AND NUMERICAL REPRESENTATIONS OF STATISTICAL ASSOCIATION IN A COMPUTER ENVIRONMENT

Carmen Batanero and Juan D. Godino, University of Granada, Spain

Exploratory data analysis activities may help to reinforce the students' understanding of association, since most exploratory data analysis activities rely on this concept to a certain extent. Furthermore, technological availability provides students with powerful tools and multiple representations of association, which could help students to widen the meaning of the concept. Students, however, do not always possess sufficient conceptual knowledge about the concepts underlying these representations or lack interpretation abilities. The limitations of the available software and the variety in the students' basic background sometimes contribute to the persistence of students' difficulties after instruction. These problems are discussed below by presenting the results of assessing students' final capacity to judge association after a basic data analysis course.

MEANING OF ASSOCIATION IN A COMPUTER-BASED TEACHING EXPERIMENT

The concept of association is highly relevant, since it extends functional dependence to random situations and intervenes in many statistical procedures, such as simple and multiple regression, analysis of variance and most multivariate methods. Its understanding rests on an understanding of both randomness and independence and incorporates other elementary statistical ideas like variation and measures of central tendency as well. Research into the understanding of association started with Inhelder and Piaget (1955), who considered that the evolutionary development of the concepts of association and probability are related, and that understanding association requires prior comprehension of proportionality, probability, and combinatorics. Later didactic research results show students' misconceptions and incorrect strategies in judging association, both on an intuitive level, and after teaching (Batanero et al., 1996, 1997; Estepa and Batanero,1995).

The meaning of mathematical concepts is closely linked to the problem situations proposed, their task variables, and the tools available for their solution (Godino and Batanero, 1997). In our course (80 hours) the students worked in the computer laboratory twice a week (1 or 2 students per computer) using some procedures of the statistical package Statgraphics. The 32 students had, in general, basic previous knowledge of statistics, though they had never worked with a statistical package before. The students'

range of knowledge was large, because the course was optional and students came from different backgrounds, such as Education, Business studies, and trainee teachers.

Description of the course

The lessons were based on the analysis of real data sets provided by the teacher or collected by the students from different sources. In addition to individual projects, the course was complemented with a series of structured data analysis activities that, by setting up research questions on the available files, served as a basis for discussing and clarifying the following basic concepts: Data, type of data, collecting data; data structure and variables; Distributions and graphics; Central position, dispersion and symmetry; Association: contingency tables, correlation and regression; Intuitive bases of probability and probability distributions; Basic principles on sampling; Intuitive introduction to confidence intervals and hypothesis testing.

Tools available for studying association

In addition to raw data, different representations of association were available to the students, which can be classified according to the level of data reduction, their numerical or graphic nature and the analysis approach (descriptive or inferential):

a) Numerical representations on a descriptive level: Contingency tables, and different associated frequencies. Unidimensional frequency tables for conditional distributions and their statistics, which involve a new data reduction and measure the central values position, dispersion and shape of the distributions. Correlation and determination coefficients, which summarise the joint distribution and directly measure sign and strength of association. Parameters of the regression line, in particular the slope, whose sign indicates the type of relationship for the variables (direct or inverse). Using regression or conditional distributions implies the problem of selecting the independent (explanatory variable) and dependent variables (response variable).

b) Numerical representations on an inferential level: Confidence intervals for means or for the difference between means. Hypothesis tests on the means or medians. Chi – squared test of association between variables.

c) Unidimensional graphical representations of conditional distributions: The stem and leaf plot does not summarise the distribution, but allows its shape, mode and atypical values to be seen. Bar graphs display part – part comparisons and have different

formats: stacked, clustered, percentual, which supply different information with regards to the association between variables. Pie charts display the frequencies as part – whole relationships. Histograms imply a second data reduction level and their visual characteristics vary with the interval width, though for students it is not always easy to choose the appropiate width. Box plots summarise five order statistics of the distribution, shows symmetry and atypical values, and can be used on an "inferential level" by adding notches.

d) Bidimensional graphics: Mosaic graphics vary as a function of the explanatory variable. Scatter plots do not involve any reduction of the data or vary according to the variables role (explanatory or response). They serve to visually display the type of relation and it is a natural extension of functional representations.

All the procedures mentioned are implemented in Statgraphics, which, in addition, has the usual possibilities for data and variables selection. The system provides a multiple window environment, with automatic resizing capabilities, where several graphs, tables and results can be displayed selectively and simultaneous for comparative purposes, as recommended by Biehler (1997).

In addition to these representations supplied by the available software, psychological research has shown that subjects use their previous ideas about the type of association between the variables (Chapman and Chapman, 1967). We can consider these previous theories as well as knowledge about the context as another "representation" of association for the student. Finally, the student can use the direct comparison of the data, for example, comparing the variation in each pair of values to study the association between a quantitative and a qualitative variable (comparing two related samples).

ASSESSING STUDENTS' LEARNING

Throughout the course, learning was evaluated from data analysis activities solutions, paper and pencil tests, and students' individual projects. In addition, at the end of the course the test developed by Estepa (1994) was given to the students. This consisted of the analysis of a new data set concerning the scores of 48 pupils in a physical education course, for which the students were given some related questions. Each student worked alone with the computer and his(her) solutions were recorded individually in a disk, using the "statfolio", which included the calculations and graphics used, together

with his/her comments and solutions. Below we analyse the solutions to four association problems.

Problem 1. Reason whether in this data set, practising sport depends on a person's sex.

Problem 2. Is there any relationship between practising sport and the number of heartbeats after 30 press-ups?

Problem 3. The teacher wants to assess the improvement in the physical preparation of his pupils. Do you think there has been any improvement in the time pupils spend running 30 metres between September and December?

Problem 4. Do you believe that the number of heartbeats after 30 press-ups depends on the number on heartbeats for pupils when resting?

The main difference between the problems is the type of variables involved: two qualitative variables (problem 1), two quantitative variables (problem 4) and a variable of each type (problems 2 and 3). Another important variable is the strength of association: independence (problem 4), weak dependency (problem 2), moderate association (highly significant test t value in problem 3), (significant value of the chi – squared test in problem 1). From the printout of the statfolios provided by the students we classified the students procedures and solutions.

Gal (1997) distinguishes two types of questions to be posed when asking students to interpret statistical information. For literal reading questions, answers can unambiguously be classified as either right or wrong. In contrast, to evaluate questions aim at eliciting students' ideas about overall data patterns, we need information about the evidential basis for the students' judgements, their reasoning processes and the strategies used to relate data elements to each other.

As general rule, the students achieved a correct association judgement, yet some correct solutions were obtained through a procedure inadequate to the type of problem. Moreover, among the tools adapted to the problems, the students did not always choose what a statistician analysing the data would have chosen. Consequently, the students' solutions did not always coincide with the "standard" solution.

For example, the best solution to problem 1 would have been using the chi-squared test to compare the proportions of males and females practising sport, which gives a significant result. Seven students computed the correlation coefficient instead,

which provides a value of 0.28, which is very small. Therefore they interpreted it as if there were no relationship between the variables. Another example is problem 3, where a student tried to represent the relationship using a scatter plot, which is not appropiate.

This selection of a correct though non optimum procedure points to the students' lack of flexibility for changing representations of association and their greater facility for interpreting the correlation coefficient as compared to contingency tables. For example, students A11, A23, A26, A35, A36 solved 3 problems using the correlation coefficient; student A12 solved 3 problems by comparing bar graphs, student A10 solved all 4 problems by comparing graphical representations of marginal distributions and A27 solved all the problems by comparing double frequencies in contingency tables.

Another example is not taking the one which permits the simplest interpretation of the analysis as an explanatory variable. For example, 3 students reached a wrong conclusion when computing conditional relative frequencies of practising sports with regards to the number of heartbeats, which makes visualisation of the relationships difficult.

Other students misinterpreted results of correctly selected and carried out data analysis procedures. The main difficulties were the following:

a) Confusing relative conditional with double frequencies (9 students) or with marginal frequencies (4 students) in contingency tables;
b) Using only one marginal distribution (1 student in problem 1) or comparing marginal distributions of the two variables involved (1 student in problem 4);
c) Comparing the correlation coefficient of each variable implied with a different variable, because of lack of understanding of entry parameters in the programs (1 student in problem 4);
d) Using previous theories without taking the data into account (problem 4);
e) Believing that dependence of heartbeats on the time in which they are recorded would imply a constant value of the number of heartbeats at each given time, that is to say, interpreting association in a deterministic way (1 student in problem 3);
f) Using a histogram, instead of a bar graph to represent a qualitative variable (1 student in problem 1);
g) Using the student in the problem 2); i) Misinterpreting steam and leaf plots (1 student in the correlation coefficient between the same variable in two related

samples to study the differences in the two samples (1 student in problem 3) or the coefficient of determination (1 student in problem 3)

h) Comparing absolute frequency bar charts, instead of using relative frequencies (1 problem 3).

As a rule, we observe that students preferred numerical to graphical representations, especially in problems 1 to 3. This is possibly due to the fact that each available graph requires its own interpretation that students do not always master. They also have mainly employed numerical summaries, because of the difficulty of the idea of distribution, which was also shown in the confusion about the different types of frequency. We finally point out the scant use of inferential procedures. Possibly students require a longer period of study to understand these concepts before deciding to employ them in solving the problems.

CONCLUSIONS

Our results show that analysing data even on an exploratory level is a high skill activity and requires knowledge about many concepts underlying graphical, numerical, descriptive and inferential representations of association. It also implies selecting the optimum representation depending on the problem data, flexibility in changing representation systems, proper interpretation of results and relating them to the research questions. Even when many of our students got correct solutions to the problems, we could observe their difficulties in each step of the process described.

Being able to master this complex activity, beyond routine or elementary tasks, or being capable to teach it to a group of students with various knowledge and prior capacities is not a simple task and certainly requires more time and experience than it is possible to provide in a introductory statistics course. Neither software with graphical representation and data transformation availability nor the teachers' didactic knowledge is enough. More research is needed to design feasible goals for these courses that should be directed to make future professionals aware of the relevance and difficulty of statistics practice and the need to cooperate with expert statisticians for solving complex data analysis problems.

ACKNOWLEDGEMENT

This research has been founded by the Spanish Ministry of Education (Grant . PB96-1411).

REFERENCES

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* (pp. 169-190). International Statistical Institute.

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

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review, 65(2),* 167-190.

Chapman, L. J., and Chapman, J. P. (1967). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271–280.

Estepa, A. (1994). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores* (Preconceptions on statistical association and its evolution after a computer-based teaching experiment). Unpublished Ph. D. University of Granada.

Gal, I (1997). Assessing students' interpretations of data: Conceptual and pragmatic issues. In B. Phillips (Ed.), *Papers on Statistical Education presented at ICME-8* (pp. 49-58). Swinburne University of Technology.

Estepa, A., and Batanero, C. (1996). Judgements of correlation in scatter plots: students' intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education, 4*, 21-41.

Godino, J. D., and Batanero, C. (1996). Clarifying the meaning of mathematical objects as a priority area of research in mathematics education. In A. Sierpinska, and J. Kilpatrick (Eds.), *Mathematics education as a research domain: The search of an identity.* Dordrecht: Kluwer.

Piaget, J. and Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant.* Translated by L. Leake, jr, P.D. Burrell, and H.D. Fischbein (1975), *The origin of the idea of chance in children* London: Routledge and Kegan Paul.