BUILDING THE MEANING OF STATISTICAL ASSOCIATION THROUGH DATA ANALYSIS ACTIVITIES

Carmen Batanero, Juan D. Godino*, University of Granada, Spain*
& Antonio Estepa, *University of Jaén, Spain*

In A. Olivier & K. Newstead: Proceedings of the 22 Conference of the International Group for the Psychology of Mathematics Education (V.1, pp. 221-236 (Research Forum). University of Stellembosch, South Africa.

Abstract

In this research forum we present results from a research project concerning students' understanding of statistical association and its evolution after teaching experiments using computers. This research has been carried out at the Universities of Granada and Jaén over the years 1991-98. We have identified different incorrect preconceptions and strategies to assess statistical association and performed two different teaching experiments designed to overcome these difficulties and to identify the critical points, which arise in attempting to do this.

The concept of association has great relevance for mathematics education, because it extends functional dependence and it is essential for many statistical methods, allowing us to model numerous phenomena in different sciences. This topic has significant connections with functional thinking and other areas of mathematics education, such as probability and proportional reasoning. The main goal in studying association is to find causal explanations, which help us to understand our environment. However, association does not necessarily imply a causal relationship. "Spurious correlation" sometimes exists when a high coefficient of association among variables arises because of the influence of concurrent factors without there being any causal link.

Besides this epistemological difficulty, psychological research has shown that the ability to judge association well is not developed intuitively. Adults sometimes base their judgement on their previous beliefs about the type of association that ought to exist between the relevant variables rather than on the empirical contingencies between them presented in the data. The existence of such preconceptions in applied situations is another difficulty for the teaching of association. Despite these epistemological and psychological issues, mathematics education researchers have carried out very little research on this topic, and most psychological research has only concentrated on 2x2 contingency tables.

THEORETICAL FRAMEWORK

Our research has been based on a theoretical framework concerning the *meaning* and *understanding* of mathematical objects, for which we distinguish between its personal and institutional dimensions, and which has been described in Godino (1996), and Godino & Batanero (1998). The epistemological assumptions behind this theory are that mathematical objects emerge from the subjects' activity of solving problems, mediated by semiotic instruments provided by institutional contexts. Consequently, the meaning of mathematical objects is conceived as the system of practices linked to specific *problem fields*, where three different types of *elements* need to be considered:

(1) Extensional element of the meaning: The different prototype situations and problems where the object is used, that is, the *problem field* from which the object emerges.

(2) Instrumental/ representational elements of meaning: The different semiotic tools available to deal with or to represent the problems and objects involved.

(3) Intensional elements of meaning: Different characteristic properties and relationships with other entities: definitions, propositions, procedural descriptions, etc.

We postulate a relativity of the emergent mathematical objects, as regards the specific institutional contexts involved in the solution of the problems, as well as the tools and expressive forms available.

According to this model, understanding a mathematical concept will imply the appropriation of the different elements that compose the institutional meaning of the concept, and therefore, has a systemic nature.

Psychological research into association

The research design, as well as the interpretation of its results has been guided by the theoretical framework described above, and by previous research on statistical association, in the scope of Psychology, which we summarise below.

Research into reasoning about association has developed from the work by Inhelder & 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. They investigated the understanding of association with children aged 13 or older, posing the problem of association between two dichotomous variables. In Table 1 we describe the data in this type of problem, where a, b, c and d represent absolute frequencies.

Table 1: *Typical format for a 2x2 contingency table*

	B	Not B	Total
A	A	B	a+b
Not A	C	D	c+d
Total	a+c	b+d	a+b+c+d

Piaget and Inhelder found that some adolescents who are able to compute single probabilities only analyse the favourable positive cases in the association (cell [a] in Table 1). In other cases they only compare the cells two by two. This fact is explained by observing that understanding association requires considering quantities (a+d) as favourable to the association and (b+c) as opposed to it and also that it is necessary to consider the difference between cases confirming the association (a+d) and cases opposed to it (b+c) compared to all the possibilities. According to Piaget and Inhelder, recognition of this fact only happens at 15 years of age.

Following Piaget and Inhelder, several psychologists have studied the judgement of association in 2x2 contingency tables in adults, using various kinds of tasks and, as a consequence, it has been noted that subjects perform poorly when establishing judgements about association (see Crocker, 1981; Beyth-Maron, 1982 for a survey). Other researchers, such as Chapman & Chapman (1969) and Jennings, Amabile and Ross (1982), have studied the effect that previous theories about the context of a problem have on judging association. Their general conclusion is that when data do not coincide with these expectations there is a cognitive conflict which affects the accuracy in the perception of covariation.

INTUITIVE STRATEGIES IN GENERAL CONTINGENCY TABLES, SCATTER PLOTS AND COMPARISON OF SAMPLES. PRECONCEPTIONS OF ASSOCIATION

Extensional meaning of association

Judging association in a 2x2 contingency table, such as that presented in item 1, is a particular case of the *problem field* from which statistical association has emerged. The processes of judging association in a contingency table require relating or operating with the different frequencies in the table.

Item 1: In a medical centre 250 people have been observed to determine whether the habit of smoking has some relationship with a bronchial disease. The following results were obtained:

	Bronchial disease	No bronchial disease	Total
Smoke	90	60	150
Don't smoke	60	40	100
Total	150	100	250

Using the information contained in this table, would you think that, for this sample of people, bronchial disease depends on smoking? Explain your answer.

Another related problem is assessing the existence of correlation between two quantitative variables (e.g., item 2), where we could compute the covariance or the correlation coefficient, or fit a line to the scatter plot. A third type of problem is inquiring whether a numerical variable has the same distribution in two different samples (item 3),where it is possible to compare the differences in means or medians, or the tabular or graphical representations of the two distributions. All these problems and activities are essential for progressively building up the concept of statistical association and form part of the *institutional meaning* of the concept within an introductory University course on data analysis. More specifically, the three types of problems described are prototypical *extensional elements of the meaning* of association within this institution.

Item 2. In a sociological study, data relative to daily consumption of animal protein and birth rate of different countries were collected (Figure 1). The data were represented in the attached scatter plot. Do you think that the relationship between daily consumption of animal protein and birth rate of these different countries is direct, inverse or that there is no relationship at all? (Explain your answer).

Item 3. The following data were obtained when measuring the blood pressure for a group of 10 women, before and after applying a medical treatment. Using the information contained in this table, do you think that the blood pressure in this sample depends on the time of measurement (before or after the treatment)? Explain your answer.

Blood pressure	Mrs. A	Mrs. B	Mrs.C	Mrs. D	Mrs. E	Mrs. F	Mrs. G	Mrs.H	Mrs. I	Mrs J
Before treatment	115	112	107	119	115	138	126	105	104	115
After treatment	128	115	106	128	122	145	132	109	102	117

Personal meaning of association

As a first step towards understanding how students develop this institutional meaning, our research project started by extending earlier research into the *personal intuitive meaning* given by students to the concept of association: What type of strategies do they use to solve the problems, and what conceptions about association can be deduced from their strategies. With this aim, an extensive study was carried out from 1992 to 1994. After some revisions with pilot samples, a questionnaire was developed and given to 213 students in their last year of secondary school (17-18 year-old students), which is the level at which association is introduced into the Spanish curriculum. The questionnaire was given to the students before the teaching of association began.

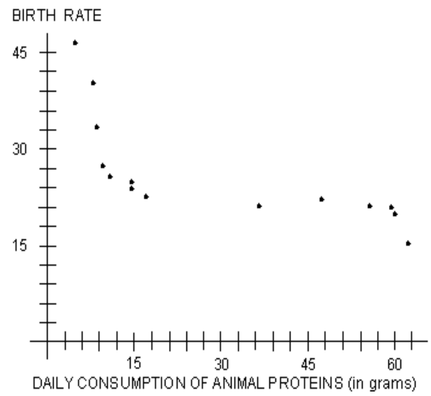

Figure 1

The 10 items in the questionnaire were similar to items 1 to 3 presented before and included 2x2, 2x3 and 3x3 contingency tables (e.g., item 1), scatter plots (e.g. item 3), and comparing a numerical variable in two samples (e.g., item 2). Sign and strength of the association and the relationship between the problem context and students' prior beliefs were taken into account in designing the questionnaire.

For each item, we analysed the type of association perceived by the students (direct association, inverse association or independence) and a scheme was developed for classifying students' solution strategies from a mathematical point of view. This allowed us to identify the incorrect strategies previously reported by Inhelder and Piaget (1955) and later psychological research, as well as intuitively correct strategies pointing out correct or partially correct conceptions concerning statistical association (Estepa et al., 1994; Batanero et al., 1996; Estepa, & Batanero, 1996; Estepa, & Sánchez-Cobo, 1996).These are some examples:

(1) Using the increasing, decreasing or constant trend of points in the scatter plot to justify the type of association (negative, positive or null*): "Because when you increase the daily consumption of proteins the birth rate is going down"* (Item 3). As in the case of independence, there would be no joint variation of the two variables, this student is using a correct conception about association.

(2) Using means or totals for comparing the distribution of one variable in two different samples: *"Because the sum of all the values of the blood pressure before the treatment is lower than the sum of blood pressure values after treatment (*Item 2). Here, the student implicitly uses the correct idea that a difference in totals implies association between the variables.

(3) Comparing either a) the frequencies of cases in favour of and against each value of the response variable or b) the ratio of these frequencies in each value of the explanatory variable in 2xr contingency tables: *It is not dependent, because the odds for having bronchial desease in smokers is 3/2, and there are the same odds in non smokers"* (Item 1). This points to a correct conception, as the odds ratio can be used to assess association in a contingency table.

On the basis of the students' incorrect strategies, which lead them to incorrect judgements of association in the problems, we have described the following misconceptions concerning statistical association:

(1) *Determinist* conception of association. Some students do not admit exceptions to the existence of a relationship between the variables and expect a correspondence that assigns only a single value in the dependent variable for each value of the independent variable. When this is not so, they consider there is no dependency between the variables. That is, the correspondence between the variables must be, from the mathematical point of view, a function. An example is given in the following answer to item 2:

"There is not much influence from the treatment, because in some women the blood pressure increases, whereas in other cases it decreases" (item 3).

(2) *Unidirectional* conception of association. Sometimes students perceive the dependence only when the sign is positive (direct association), so that they consider an inverse association (negative sign) as independence. The following response illustrates the case of inverse association being interpreted as independence, as well as difficulties with proportional reasoning in item 1:

"I personally believe there is no dependence, because if you look at the table there is a higher proportion of people with bronchial disease in smokers".

(3) *Local* conception of association. Students form their judgements using only part of the data provided in the problem. If this partial information serves to confirm a given type of association, they adopt this type of association in their answer:

"*There is dependence on smoking in having bronchial disease, because if we observe the table, there are more smokers with bronchial disease than non-smokers: 90>60"* (Item 1).

Often this partial information is reduced to only one conditional distribution (as in the example, where the student only uses the distribution of people with bronchial disease) or even only one cell, frequently the cell for which the frequency is maximum. These strategies are similar to those reported by Inhelder and Piaget (1955). Konold et al. (1997) also described similar reasoning, which they attributed to lack of transition from thinking about individual cases to thinking about group propensities. The latter is of course essential in statistical reasoning.

(4) *Causal* conception of association*:* Some students only considered the association between the variables if this could be attributed to a causal relationship between them. This was particularly found in one problem concerning the ranking of a group of students by two different judges, even when there was a moderate correlation:

"Because one judge cannot influence the other. Each one has a preference and therefore there can not be much relation between the order given by them both".

RESEARCHING THE POTENTIAL OF COMPUTER-BASED ENVIRONMENTS FOR TEACHING AND LEARNING ASSOCIATION

After carrying out the study on students' initial conceptions about association, our research was aimed at assessing the impact of computer-based teaching experiments on the same. The use of computers in the teaching of Statistics is receiving increasing attention from teachers and researchers as is shown in Shaughnessy et al., (1996) and the IASE Round Table Conference on the impact of new technologies in teaching and learning statistics (Garfield, & Burrill, 1997). In this research it is suggested that computers may not only extend what statistics is taught, but may also affect how statistics is learnt, because technology provides students with powerful resources and multiple representations, which can help them to widen the meaning of statistical concepts. The aim of our research was to assess this impact for the specific case of statistical association.

Building instructional strategies

Biehler (1997) classified different types of statistical software according to their educational functions. *Tools* allow students to practice statistics as current statisticians do. In the particular case of exploratory data analysis, these tools should enable students to do interactive, exploratory and open-ended work, utilising flexible software, which is easy to use and to learn. *Microworlds* summarises interactive experiments, simulations and exploratory visualisations, which help students to conceptualise statistics. In our two experiments, we have considered c*omputer learning environments,* as an integrated instructional setting which allows the teacher and a group of students to work with tools, microworlds, data sets and related problems as well as with a selection of statistical concepts and procedures.

The content of the two courses included the fundamentals of descriptive statistics, using an exploratory data analysis approach. The specific statistical contents were the following:

(1) Random and deterministic experiments. Population and samples. Measurement scales. Type of variables. Collecting and organising data.

(2) Frequency, cumulative frequency, grouping data. Graphical representation: bar chart, pie charts, histograms, stem and leaf, graphical representation for cumulative frequencies.

(3) Parameters and statistics. Location; Spread. Order statistics. Skewness and kurtosis.

(4)Two-dimensional statistical variables: Contingency tables. Association in contingency tables. Statistical association in numerical variables. Covariance and correlation. Linear regression.

(5) In the second experiment students were also introduced to inference. The concepts of sampling, sampling distribution, interval confidence and hypotheses testing of means for one and two samples and Chi-square test were introduced.

The planning of the teaching involved the organisation of an instructional sequence to meet the learning goals and contents, the selection of appropriate data sets to contextualize statistical knowledge, and the design of a graded problem sequence including the main task variables relevant to understanding association (Godino, et. al, 1991).

All of this was planned to provide the students with a representative sample of extensional, intensional and representational elements for the meaning of association. We adopted a "multivariate perspective", even though only univariate or bivariate techniques were taught at a formal level. Therefore, students explored data files with the support of an interactive computer software package. The number of sessions (90 minutes long) were 21 in the first experiment and 40 in the second. In seven of these sessions in the first experiment and 20 sessions in the second, the students worked in the statistical laboratory, solving problems whose solutions required them to analyse different data sets provided by the lecturer or collected by themselves. In the remaining sessions students were introduced to statistical concepts and solved related problems.

First experiment: Identifying resistant conceptions and analysing the learning process

The first experimental sample consisted of nineteen 20 year-old University students in a first year course of exploratory data analysis and descriptive statistics. The students worked with the statistical software PRODEST, which had been developed by the research team some years before. Although this software had rather limited capability compared with modern statistical packages, it included all the different tools needed in a course of exploratory data analysis at undergraduate level: frequency tables and graphical representation for grouped and non grouped data; computation of statistics; stem and leaf plots; box and whiskers plots; cross tabulation; linear regression and correlation. In addition, data file facilities and possibilities of selecting part of the data set were available.

To assess the changes in students' conceptions, two equivalent versions of a questionnaire were given to the students as a pre-test and post-test. We found general improvement in students' strategies as well as the persistence of unidirectional and causal misconceptions concerning statistical association for some students after the teaching process. All these results have been described in greater detail, see Batanero et al. (1997).

One pair of students was observed throughout their work in the laboratory sessions to trace their learning process. As stated by Biehler (1994), when working with the computer, an adequate solution to statistical problems is only found through a feedback process with the specific problem and data. The students do not just choose an algorithm, but have more freedom, because they have a system of options available that they can combine and select according to their strategies and partial solutions when solving the problem.

A member of the research team observed the students' work, gathering their written responses to the different problems. This observation also included the recording of their discussions, of their interaction with the lecturer and the computer. These students were also interviewed at the beginning and at the end of the experiment. When we studied in detail the observations made on these students, some recurring difficulties relating to the idea of association were identified. Some of them were finally solved, either by the students themselves, when discussing and looking at the results of several computer programs, or with the lecturer's help, although they reappeared from time to time. At other times the difficulty was not solved, in spite of the lecturers' explanations. Occasionally, the lecturer did not realise the students' confusion.

Intensional elements of meaning of association

In the following, we describe the learning process for these two students, commenting on nine key intensional elements of the institutional meaning of association (Godino & Batanero, 1998). We found evidence in our data that students' understanding of these elements of meanings seemed to develop at specific moments in time throughout the understanding process.

1. To study the association between two variables, the comparison of two or more samples has to be made in terms of relative frequencies. However, in the first session the students compared absolute frequencies of the same variables in two samples. Although the lecturer commented on this mistake at the end of that session, the same incorrect procedure appeared again in sessions 2, 3 and 5. Afterwards, the students seemed to overcome this difficulty.

2. *The complete distribution in the different samples should be used to assess the differences in the same variable between two or more samples.* Finding local differences is not sufficient, but rather the association should be deduced from the complete data set. In spite of this, the students started solving the problems by comparing isolated values in the two samples. For example, they only compared the values with maximum and minimum frequencies for both samples in the first session. Although these differences pointed to a possible association, they were not sufficient to quantify its intensity. This difficulty reappeared in Sessions 2 and 3 and finally disappeared.

3. *From the same absolute frequency in a contingency table cell two different relative conditional frequencies may be computed, depending on which one is the conditioning variable. The role of condition and conditioned in the conditional relative frequency is not interchangeable.* Falk (1986) and other authors have pointed out that students have difficulties in the interpretation of conditional probabilities, because they do not discriminate between the probabilities P(A/B) and P(B/A). Many students in our sample showed a similar confusion referring to the conditional relative frequencies in the pre-test and throughout the experimental sessions. This confusion was noticed in the students observed during Session 5, although they solved it with the lecturer's help. They did not show this confusion during the rest of the sessions.

4. *Two variables are independent if the distribution of one of these variables does not change when conditioning by values of the other variable.* Until Session 5, the students did not discover that a condition for independence is the invariance of the conditional relative frequency distribution when varying the value of the conditioning variable.

5. *The decision about what size of differences should be considered to admit the existence of association is, to some extent, subjective. It is difficult to obtain either perfect association or independence. The problem of association should be set in terms of intensity instead of in terms of existence.* Though the students had not studied hypothesis testing, in Session 5 they discovered that judging association implies taking a decision about whether to attribute small differences to sampling fluctuations or to real association between the variables. They also realised that there are different grades of association, from perfect independence to functional relationship.

6. *When studying association both variables play a symmetrical role. However, when studying regression the role played by the variables is not symmetrical.* The fact that correlation ignores the distinction between explanatory and response variables, whilst in regression this difference is essential (Moore, 1995), caused a great deal of confusion for the students. When they had to select the explanatory variable for computing the regression line, in Sessions 5, 6 and 7, they did not know which variable ought to be chosen.

For example, when computing the regression line between height and weight, the students were misled by the fact that there was a mutual dependence of the two variables. A great amount of discussion follows in which the students were not capable of solving this confusion. The lecturer did not notice the problem and finally, the students computed the regression lines by choosing the explanatory variable at random. At the end of the teaching period these students had not discovered that two different regression lines can be computed.

7. *A positive correlation points to a direct association between the variables.* Although, in Session 6, students could interpret the size of the correlation coefficient, they did not discuss the type of association (direct or inverse). At the end of the session, they noticed that when the correlation coefficient is positive, and there is a linear relationship, the variables are positively associated and above-average values for one tend to accompany above-average values for the other. However, they did not explicitly use the term "direct association."

8. *A negative correlation points to an inverse association between the variables.* When, in Session 6, students came across a negative correlation coefficient for the first time, they were so surprised that they asked their lecturer if this was possible. They also had trouble when comparing two negative correlation coefficients.

The students knew that a negative number with a high absolute value is smaller than a negative number with a low absolute value. However, a negative correlation coefficient with a high absolute value points to a higher degree of dependence than a negative correlation coefficient with a lower absolute value. This fact caused much misinterpretation in the problems in which a negative correlation occurred. Therefore, the knowledge of the properties of negative number ordering acted as an obstacle to dealing with negative correlation.

Although, with the lecturers' assistance, they observed that a negative correlation coefficient corresponded to a negative slope of the regression line and that this meant that the *y* value decreased when the *x* value increased, they did not explicitly use the term "inverse association", neither did they differentiate between the two types of association at the end of their learning.

9. The absolute value of the correlation coefficient shows the intensity of association. Although the students related the absolute value of the correlation coefficient with the intensity of association, they did not relate this idea to the spread of the scatter plot around the regression line.

Second experiment: Improving instrumental/representational elements of meaning

In the second experiment, we concentrate on analysing the use that students make of different statistical tools to solve the problems following teaching. 36 students took part in the second experiment, working with some procedures in the statistics package *Statgraphics*, where new statistical instrumental/ representational procedures, as well as a more user-friendly and dynamic environment were available to students for studying association. These statistical tools allow students to operate or represent bivariate data in different ways, leading them to progressively build a more complete meaning of association. We could classify these tools by their level of data reduction, numerical or graphic nature, and their analytical approach (descriptive or inferential):

(1) Numerical representations at a descriptive level: Contingency tables (e.g., Item 1), and different associated frequencies. Unidimensional frequency tables of conditional distributions and their statistics, correlation and determination coefficients, and parameters of the regression line.

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

*(3)Unidimensional representations of conditional distributions: S*team and leaf plots, bar graphs, box plots (Fig. 2), pie charts, histograms (Fig.4), quantile plots (Fig 3), density traces.

(4) Bidimensional graphics: Three-dimensional histograms, mosaic graphics (Fig. 5) and scatter plots (Fig. 6).

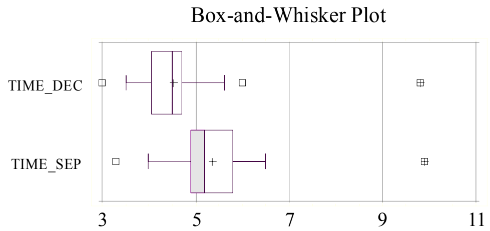

Figure 2

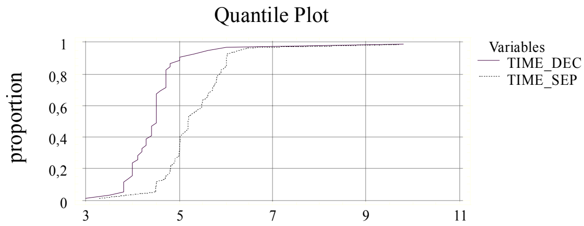

Figure 3

As shown in the figures, each one of these representations provides different data summaries and therefore different elements of *meaning* for association. For example, in Fig. 2 and 3 the two distributions of the time, in seconds, taken to run 30 meters by a group of students in September and December are represented. However, in Fig. 2 the outliers and slight asymmetry in the distributions, as well as the interval of 50% central values are more clearly shown, whereas in Fig. 3 it is easier to notice that the probability of obtaining a given value of the time is always greater in one distribution (September), which means that the time in September is stochastically larger than the time in December.

As well as the usual possibilities for selecting data and variables, it was also possible with Statgraphics to display several tabular and graphical representations on the screen at once, at the same time as the student handled features like width, format scale of graphics, etc. The *Statfolio* file also allowed the research report to be written and partial results to be included in the text at the same time as the analysis was being carried out.

Personal meaning of association following teaching

The students' previous statistical knowledge was wide-ranging, because the course was optional and students from different backgrounds, such as Education, Psychology and Business, took the course. During the course, generative and interpretative skills (Gal, 1997) were evaluated from students' solutions to data analysis activities, paper and pencil tests, and from their individual projects. In addition, at the end of the course a test was given to the students concerning the analysis of a new data set to assess the final meaning given to association by students. Each student worked alone with the computer and his/her solutions were recorded individually on a disk file, using the "Statfolio", which included the calculations and graphics used, together with his/her comments and solutions. The test 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. Below we analyse the solutions to four association problems.

Problem 1. Do you think that, in this data set, practising sport depends on a person's sex.
Problem 2. Is there any relationship between practising sport and number of heartbeats after 30 press-up?
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 took to run 30 meters between September and December?
Problem 4. Do you believe that the number of heartbeats after 30 press-up depends on the pupils' heartbeat when resting?

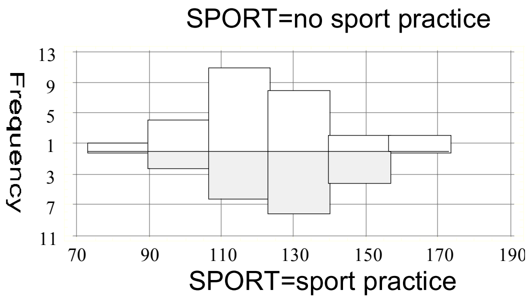

Figure 4

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 (problem 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).

The students should identify these differences (they need to discriminate between different extensional elements of meaning). Then they should select from the different tools available (instrumental/representational elements of meaning) those that are adequate to solve the problem, such as contingency tables and related statistics, mosaic chart (Fig. 5), or comparing bar graphs in problem 1, comparing histograms (Figure 4), quantile plots (Fig.3), density traces, box plots (Fig.2) or different statistics in problem 2 and 3, and studying the correlation coefficient or the scatter plot (Fig. 6) in problem 4. Finally, intensional elements of meaning should be used to interpret the outputs of the different programs and to make an accurate judgement of association.

By printing out each student's statfolio, we categorised the students' procedures and solutions. We summarise the results below, which will be presented in detail in Batanero, & Godino (in press). As a rule, the students reached a correct association judgement, yet some correct solutions were obtained through a procedure unsuitable for the type of problem, because students did not always correctly link the association extensional and instrumental/ representational elements of meaning.

Moreover, among the tools suitable for 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 that gives a significant result. Seven students, however, computed the correlation coefficient instead, which provides a value of 0.28. As this value is very small, they interpreted that there was no relationship between the variables. Another example is problem 3, where a student tried to display the relationship using a scatter plot that is not adequate.

This selection of a correct though non optimum procedure points to the students' lack of flexibility for changing representations of association and the greater facility to interpret the correlation coefficient as compared to contingency tables. For example, students S11, S23, S26, S35, S36 solved 3 problems using the correlation coefficient; student S12 solved 3 problems by comparing bar graphs, student S10 solved all 4 problems by comparing graphical representations of marginal distributions and S27 solved all the problems by comparing double frequencies in contingency tables. Another example is not taking as an explanatory variable the one which allows the simplest interpretation of the analysis, like 3 students who reached an incorrect conclusion when computing conditional relative frequencies for practising sports with regards to the number of heartbeats, a strategy that makes it difficult to display the relationships.

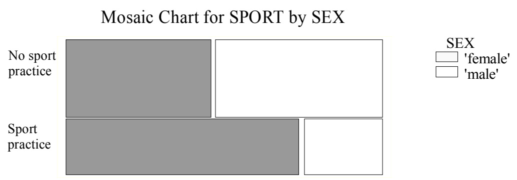

Figure 5

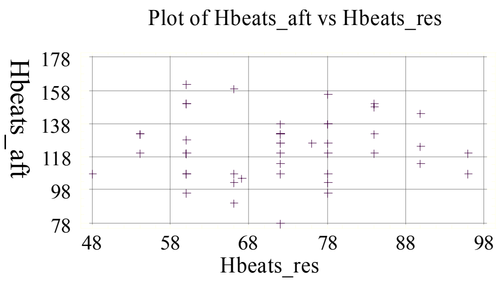

Figure 6

Other students misinterpreted results of correctly selected and carried out analysis due to failures in grasping some intensional elements in the meaning of association. 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 a marginal distribution (1 student in problem 1) or comparing the marginal distributions of the two variables involved (1 student in problem 4);

c) comparing the correlation coefficient for each variable implied with a different variable, because of a 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 the dependence of heartbeats at the time when they are registered would imply a constant value of the number of heartbeat at each given time, that is to say, interpreting association in a deterministic way (1 student in problem 3);

f) using 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 frequencies in bar charts, instead of using relative frequencies (1 student in problem 2).

As a rule, we observe that students preferred numerical to graphical representations of association, especially in problems 1 to 3. This is possibly due to the fact that each available graph requires its own interpretation that the students do not always master. They also preferred numerical summaries, because of the difficulty with 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 their problems.

FINAL REFLECTIONS

Our research results reveal the complexity of the *meaning* and *understanding* of association, which should be conceived in a systemic and complex way (Godino, 1996; Godino & Batanero, 1998). In constructing this meaning, three different types of *elements* need to be grasped and linked for students to master the concept of association and use it in problem solving:

(1) *Extensional element of the meaning*: The different situations and problems whose solution requires the study of association and that have been described in this paper.

(2) *Instrumental/ Representational elements of meaning*: The use of different tools to deal with or represent the concept, such as mosaic charts, scatter plots, two-way tables or a series of parallel box-plot, cumulative plots, histograms or bar charts.

(3) *Intensional elements of meaning*: such as the difference between statistical dependence and functional dependence, the different relative frequencies that can be deduced from a contingency table, the role of independent and dependent variables, and the parameters in the regression equation, the interpretation of the correlation coefficient, and the difference between correlation and causality.

A second consequence of our research is the distinction between the *personal* (subjective) and *institutional* (mathematical) dimensions of knowledge, meaning and understanding in mathematics. This is particularly shown when comparing the meaning of association that was presented in the teaching experiments (institutional meaning within the particular institution of a course on exploratory data analysis) and the personal meaning that the students have finally acquired, where they have only built part of the intended meaning, and some of the incorrect conceptions on association still remain.

Our results finally show that analysing data is a highly skilled activity even at an exploratory level, requiring a wide knowledge about the problems and concepts underlying graphical, numerical, descriptive and inferential procedures to deal with association. It requires selecting the best data instruments and representations, flexibility in changing the selected procedure, adequate interpretation of results (intensional elements), the ability to relate them to the problem (extensional elements) and to assess the validity and reliability of the conclusions drawn. Even when many of our students achieved 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 different prior knowledge and capacities is not a simple task and certainly requires greater time and experience that what is possible to provide in an introductory statistics course. The research carried out supports the view of mathematical objects as signs of cultural units, whose systemic and complex nature cannot be described merely by formal definitions when the perspective taken is that of the study of teaching and learning processes. Based on this viewpoint, we might explain some learning misconceptions and difficulties, not only in terms of mental processes, but by recognising the complexity of meaning of mathematical concepts and the necessarily incomplete teaching processes found in teaching institutions.

Acknowledgement: This report has been funded by Project PB96-1411 (Dirección General de Enseñanza Superior, M.E.C., Madrid).

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,* 27(2), 151-169.

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

Batanero, C., & Godino, J.D. (In press). Understanding graphical and numerical representations of statistical association in a computer environment. To be published into the *Proceedings of the V International Conference on Teaching Statistics.* Singapore, 1998.

Beyth-Marom, R. (1982). Perception of correlation re-examined. *Memory and Cognition,* 10(6), 511-519.

Biehler, R. (1994). Software tools and mathematics education: the case of statistics. In C. Keitel and K. Ruthnen (Eds.), *Learning form computers: Mathematics education & technology* (pp. 68-100). Berlin: Springer Verlag.

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

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology,* 74, 271-280.

Crocker, J. (1981). Judgement of covariation by social perceivers. *Psychological Bulletin*, 90 (2), 272-292.

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

Estepa, A., & Sánchez-Cobo, F. (1996). Association judgments in the comparison of two samples. In L. Puig, & 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.

Estepa, A., Green, D. R., Batanero, C., & Godino, J. D. (1994). Judgements of association in contingency tables. An empirical study of students' strategies and preconceptions. In J. P. Ponte, & J. F. Matos (Eds.), *Proceedings of the XVIII International Conference on the Psychology of Mathematics Education* (v.2, pp. 312-319). University of Lisbon.

Falk, R. (1986). Conditional probabilities: Insights and difficulties. In R. Davidson, & J. Swift (Eds.), *Proceedings of the II International Conference on Teaching Sstatistics* (pp. 292-297). Voorburg: International Statistical Institute.

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). Hawthorne: Swiburne University of Technology.

Garfield, J.B., & Burrill, G. (1997) (Eds.), *Research on the role of technology in teaching and learning statistics. Proceeding of the IASE 1996 Roundtable Conference.* Voorburg, The Netherlands: International Statistical Institute.

Godino, J.D. (1996). Mathematical concepts, their meaning and understanding. In L. Puig, & A. Gutiérrez (Eds.), *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education* (v.2, pp.417-424). Univerity of Valencia.

Godino, J. D., & Batanero, C. (1998). Clarifying the meaning of mathematical objects as a priority area of research in mathematics education. In A. Sierpinska, & J. Kilpatrick (Eds.), *Mathematics Education as a research domain: A search for identity* (pp. 177-195). Dordrecht: Kluwer A. P.

Godino, J. D., Batanero, C., & Estepa, A. (1991). Task variables in statistical problem solving using computers. In J. P. Ponte et al. (Eds.), *Mathematical problem solving and new information technologies. Research in contexts of practice* (pp. 193-203). Berlin: Springer Verlag.

Inhelder, B. & Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent [The growth of logical thinking from childhood to adolescence].* Paris: Presses Universitaires de France.

Jennings, D. L., Amabile, M. T. & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgements. In D. Kahneman, P. Slovic & A. Tversky (Eds*), Judgement under uncertainty: Heuristics and biases* (pp. 211-230). New York: Cambridge University Press.

Konold, C., Pollatsek, A., Well, A., & Gagnon, A. (1997). Students analysing data: Research of critical barriers. In J. B. Garfield, & G. Burrill (Eds.), *The impact of new technologies in learning and teaching statistics. IASE 1996 Round Table Conference* (pp. 151-167). Voorburg: International Association for Statistical Education.

Moore, D. S. (1995). *The basic practice of statistics.* New York: Freeman.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. Bishop et al. (Eds.), *International handbook of mathematics education* (v.1, pp. 205-237). Dordrecht: Kluwer, A. P.