Batanero, C. (Ed.), Training Researchers in the Use of Statistics, 37-51.

Printed in Granada, Spain.

ANTONIO ESTEPA & FRANCISCO T. SÁNCHEZ COBO

EMPIRICAL RESEARCH ON THE UNDERSTANDING OF ASSOCIATION AND IMPLICATIONS FOR THE TRAINING OF RESEARCHERS

In this paper we summarise the main research findings on the understanding of association carried out in psychology and mathematics education and we present results from an assessment study on the understanding of correlation and regression by university students. We finally discuss the implications of these results for designing courses directed to train researchers in the use of statistics

1. INTRODUCTION

Association has great relevance for the training of researchers since it is essential for many statistical methods and techniques frequently used by researchers, such as simple and multiple regression, variance and covariance analysis, log-linear models and LISREL, in addition to the majority of multivariate methods. Association plays a main role in educational research (Blumberg, 2001), and its understanding is also needed to read research literature (Bangdiwala, 2001). On the other hand, there is a close connection between the concept of association and the idea of cause, on which scientific knowledge and decision making are based, since causal explanations allow us:

> *to explain the past, control the present and predict the future* (Crocker, 1981, p. 272).

This notion has been debated by mathematicians, statisticians and philosophers of science, and, even when research methodology is based to a great extent on studying the association between variables, causality and association do not always coincide.

Besides this epistemological difficulty, psychological and mathematical education research has shown that the ability to judge association well is not developed intuitively. Numerous difficulties and obstacles related to association might imply serious misinterpretations and misuses of statistics methods in research. Below, we summarise the main research in this area, and present the results of an assessment study on the meaning correlation and regression by undergraduates. We finally reflect on the implications of these studies for the training of researchers.

2. PREVIOUS RESEARCH ON THE DEVELOPMENT AND USE OF ASSOCIATION

2.1. RESEARCH FROM PSYCHOLOGY

The perception of covariation between stimuli, behaviour and outcomes is a main component of human adaptable behaviour, and this explains the interest towards association from clinical, social and development psychologists. Decision making under

uncertainty has been studied in these fields, where heuristics have been described (Kahneman, Slovic, & Tversky, 1982), using the metaphor of the intuitive statistician (Peterson & Beach, 1967).

Although the pioneering work by Inhelder and Piaget (1955) on the development of the idea of correlation was carried out with adolescents (12-15 year olds), most research in psychology has focussed on adults, such as undergraduates or professionals, and therefore its findings can be valid for the training of researchers. This research shows that adult's reasoning on association is, as a rule, very poor (Nisbett & Ross, 1980).

Strategies in the interpretation of association in contingency tables

Most psychological research on association studies adult's interpretation of association from 2x2 contingency tables, like table 1, by analysing the strategies employed, performance of association judgements and the factors that intervene in the same.

Table 1. Data in 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

A first set of papers study the strategies employed in solving association problems. Pérez Echeverría (1990) distinguishes seven types of strategies found in these papers, according to the cells employed in Table 1 and the form in which they are combined:

- Using only cell [a] (present - present) where the two variables simultaneously occur, estimating a positive, negative or null association if this value is greater than, smaller than or equal to the other three cells.
- Finding the difference [a-b] between cell a (present - present) and cell b (present - absent). This was the most widely used strategy in one experiment by Arkes and Harkness (1983).
- Using the difference between the absolute frequencies of cells a and c (absent - present).
- Comparing the differences found in cells [a, b, c], which, in terms of causal judgement, determine the need and sufficiency of causes.
- Comparing the differences [a-b] and [c-d] between the differences of the frequencies in two rows (or columns).
- Comparing [a+d] and [b+c], which are the cases favouring and contradicting the possible relationship. A variation of this strategy, which consists of computing the difference between the sum of the diagonals $\delta D = [(a+d)-(b+c)]$ was studied by Shaklee and Mims (1982), Arkes and Harkness (1983) and Allan and Jenkins (1983). Jenkins and Ward (1965) observed that this strategy is limited to the cases in which the marginal frequencies are equal.
- The last strategy consists of relating the absolute frequencies of the four cells using multiplication, that is, in comparing the conditional probabilities of a variable, given the alternative values of the other variable. Though this is the correct strategy, few students used this strategy in a spontaneous way.

The influence of previous theories

Association judgements are also influenced by previous theories or expectations,

which according to Shanks (1987) depend on the individual's experience of contingencies between actions and outcomes. Jennings, Amabile and Ross (1982) concluded that correlation is overestimated when previous theories exist. However, a strong correlation between the data is necessary to detect the association when previous theories do not exist, and in this case correlation is underestimated. Chapman and Chapman (1969) studied *illusory correlation*, which can be defined in the following words:

> *"When correlation is perceived on the basis of one's theories, but is not based on empirical facts"* (Murphy, & Medin, 1985, p. 301).

Tversky and Kahneman (1982b) argued that illusory correlation might be explained by the heuristic of accessibility, since:

> *"Lifelong experience has taught us that, in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur"* (Tversky, & Kahneman, 1982b, p. 14).

Another related term is the *"illusion of control"*, or

> "*expectancy of a personal success probability inappropriately higher than the objective probability would warrant"* (Langer, 1975, p. 232).

Accuracy of association judgements

Association judgements are influenced by several factors, such as the data format or the strength of correlation. Erlick and Mills (1967) studied the influence of positive or negative correlation and found that some subjects judged negative correlated variables to be independent. Lane, Anderson and Kellan (1985) concluded that graphical format induces judgements of stronger correlation than tabular format.

Finally, Beyth-Marom (1982) differentiated between symmetrical variables, where the values are given the same weight by an individual (e.g., being male – being female) and asymmetrical variables, when its values are not given the same weight by an individual, (e.g., being male – not being male) and found a better perception of correlation in symmetrical variables. According to Nisbett and Ross (1980) negative events have a lesser impact on people's attention than positive events, therefore in asymmetrical variables the two values are given different weights.

2.2. RESEARCH FROM STATISTICS EDUCATION

Within statistics education, research on association has been carried out with pre-University or University students. Of particular interest for researchers' training is the study of preconceptions, since these preconceptions might also be held by future researchers when starting their training in statistics, and therefore, they should be taken into account to organise the teaching of association. This research has focused on the subjects's interpretation of association in contingency tables, their perception of correlation between numerical variables, the comparison of numerical variables in two or more samples and the effect of instruction.

Association in contingency tables

Batanero, Estepa, Godino and Green (1996) studied pre-university students' preconceptions about statistical association by analysing students' strategies in judging the association from a mathematical point of view. They identified three misconceptions of statistical association:

- *Determinist conception of association*. Some students 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.
- *Unidirectional* conception of association. Sometimes students perceive dependence only when the sign is positive (direct association), and they consider an inverse association (negative sign) as independence.
- *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.

Correlation between numerical variables

The determinist and local conceptions were confirmed by Estepa and Batanero (1996) in their research on the intuitive strategies used by pre-university students (18 years old) upon evaluating correlation between numerical variables represented in a scatter plot. In addition, a new misconception was identified:

- *Causal* conception of association*:* Some students only considered the association between the variables if this could be attributed to a causal relationship between them.

Some of the above conceptions were also found by Morris (1997) in her Psychology students who deduced causality from correlation. She also found that some students believed that positive correlation was always stronger than negative correlation and that a negative correlation indicates independence between the variables.

Truran (1997) evaluated economics and business students' learning of regression. He studied the students' interpretation of the slope and intercept in the regression line, their interpretation of correlation and determination coefficients, and their predictions from the regression equation. Almost all the students in Truran's study identified moderate and negative correlation. Students' answers in making extrapolations were reasonable, as they took into account the strength of correlation and the sample size. However, the author observed a routine learning of the determination coefficient and found the determinist conception of association.

Sánchez (1999) studied the meaning of correlation and regression in undergraduates, and found misconceptions related to correlation and regression, difficulties in translating between the different representations of correlation (verbal description, table, scatter plot and correlation coefficient), difficulties in solving correlation problems and in computing and interpreting the two regression lines.

Comparison of two samples

Estepa, Batanero and Sanchez (1999) described the intuitive strategies employed by pre-University students in comparing two samples (independent and related samples). They found that the most commonly used strategies were different in related and

independent samples and also found the correct and incorrect strategies about association that we have described in the previous paragraphs.

Effect of instruction

Batanero, Estepa and Godino, (1997) and Batanero, Godino and Estepa, (1998) analysed the learning process of University students in some teaching experiences. They identified nine elements of the meaning of association that emerged in specific moments of the process of learning association. They also found that the determinist and local conceptions of association were overcome by the majority of students. The unidirectional conception improved only in some students and the causal conception hardly improved at all in the students taking part in the experiments.

4. THE UNDERSTANDING OF CORRELATION AND REGRESSION IN FUTURE RESEARCHERS AND PROFESSIONALS.

Below we describe an assessment study on the understanding of correlation and regression in future researchers, which is based on the theoretical framework described in Godino and Batanero (1997). In this theoretical framework the *meaning of a mathematical or a statistical object* (for example, association) is conceived as a complex entity, consisting of several interrelated *elements of meaning*. A distinction is also made between the *institutional meaning*, presented to a student in a given teaching institution, and the *personal meaning* of the object which is in fact acquired by the student. Understanding a mathematical object is conceived as the progressive matching of personal and institutional meanings.

The research described in Section 2 has only analysed partial aspects of the meaning of association, with special emphasis in the study of preconceptions before teaching. Our work aimed to evaluate the meaning that a sample of undergraduates give to correlation and regression, after an introductory statistics course, and to study interrelated elements of meaning. Since, in the specialities analysed, the students only receive this statistics course, we also characterise the knowledge about association in researchers entering doctorate and postgraduate courses as well as future professionals. This might be useful when predicting misinterpretations and misuses of association in these future researchers and professionals and when designing courses directed at postgraduates and future researchers.

4.1. METHODOLOGY

The questionnaire and the students' responses to the same are presented in the appendix. The questionnaire was made up of 11 items, each one comprising several subitems in which the students had to choose all the true responses. An additional task involves the ordering of some correlation coefficients (item 1).

Some of the items have been taken from previous research works, such as Tversky and Kahneman (1982 a) and Morris (1997); other items were taken from the book by Cruise, Dudley and Thayer (1984), and the remaining items were prepared by ourselves. The number of correct options in the items is variable in order to reduce random choices. A pilot sample was used to try the questionnaire, adjust wording problems, and study the questionnaire reliability.

The mean frequency (percentage) of correct answers in the items was 102.6 (53.2%),

s = 35.6, which indicates the difficulty that these concepts have for the students. Consequently, we can predict difficulties and errors in the use of correlation and regression in research projects, in analysis of real data, and in interpreting the results. The papers by Bishop and Talbot (2001), Godino, Batanero and Jáimez (2001), and Shimada (2001) also describe researchers' potential errors in applying statistics.

The population intended were the students majoring in business studies or in nursing at the University of Jaén (Spain). The sample of students who completed the questionnaire included 193 students, 104 (37 men and 67 women) from Business Studies and 89 (20 men and 69 women) from Nursing Studies. The average age of these students was 20. All the students had finished the introductory statistics course in the University that year.

In this paper association is studied only from a descriptive point of view and therefore, we do not discuss inferential problems. All our questions concerning association only refer to the specific sample data presented to the students and we never ask them to infer the existence of association in the population where the samples have been taken. In the course followed by the students who answered our questionnaire only a short time was devoted to the study of inference at the end of the year. Even then, the study was restricted to inference for means and proportions and the chi-square test of association. They were not introduced to inference for correlation or regression coefficients. This is an important point to understand our analysis of students' responses and strategies and how we classify them as correct or incorrect.

As regards previous statistics studies, we found that 117 students (60.6 %) had not studied statistics before entering the University; 62 students (32.1%) had studied some statistics in secondary education and/or pre-university studies, and 14 students (7.2 %) had studied statistics in other University courses, in the previous academic year.

Below we analyse the data, describing the meaning given by the subjects to different concepts that intervene in the overall meaning of association. These concepts are: dependence/independence, covariance, correlation and regression.

4.2. DEPENDENCE

We first analyse how the subjects conceived functional and random dependence and independence, as well as types of dependence.

Functional dependence. We analysed whether the students conceive random dependence as an extension of functional dependence, and whether they assign correct values to the correlation coefficient that corresponds to these types of relationships. In our questionnaire (item 7c), 112 students, 58%, asserted that if a functional relationship exists the correlation coefficient takes a value of ±1, and thus not all of them were conscious that a non-linear functional relationship (e.g., parabolic or exponential function) could give a different value to the correlation coefficient.

Linear dependence. Less than 10% of the students in Truran's research (1997) questioned the assumption of linearity. In response 6d, only 33.2% of our students replied that when the correlation coefficient is null the relationship might be non-linear. A related answer is choosing item 7a (independence), in which 138 (71.5%) students replied that if r=0, the variables are independent.

Consequently, many of these students might consider that there is dependence only if the correlation coefficient is different to zero, when analysing dependence between two variables in their professional life. If the correlation coefficient is zero, they might also consider that there is no dependence, and forget non-linear relationships or the possible influence of other variables (item 2b). There also was a low rate of correct

answers in item 2b, where only 11.4% of the students replied that if the covariance is positive the correlation could be non-linear. Although non-linear regression was taken into account in the teaching of these students, linear correlation received more attention, which probably can explain the incorrect answers in these items.

4.3. COVARIANCE

According to Wild and Pfannkuch (1999), recognising the role of variation is a basic element of statistical thinking. Understanding the covariance and related concepts is a first step in the study of correlation and regression. In the questionnaire we included an item (item 2) and two options (4d and 6b) to study the understanding of this concept. In the teaching of the theme to these students covariance was defined as the measure of the combined variability between X and Y, and the relation between the sign of covariance and the type of dependence was also studied. The correlation coefficient was defined by (1).

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (1).$$

M ore than a half of the students (54.3%) understood that a positive covariance corresponds to positive correlation (item 2) and to a positive correlation coefficient (item 2e). However, in spite of the fact that the slope of the regression line Y/X was defined as σ_{xy} / σ_x^2 in the instruction, only 45.1% of the students of this sample replied that when the covariance is greater than zero the slope of the regression line is positive. The remaining students did not relate the mathematical equations studied (item 2d).

We also emphasise the low rate of correct answers (43.5%) for option 6b, which is an immediate consequence of equation (1), as if $\rho = 0$, $\sigma_{xy} = 0$. Finally, only 17.1% of the students replied that when the strength of the relationship decreases the absolute value of covariance increases (item 4d), where again equation (1) is forgotten.

4.4. CORRELATION

With respect to the understanding of correlation, we analysed the following additional aspects:

The strength of the relationship between two variables depends on the scatter plot spread (item 3d and 4c). We can observe a substantial percentage of students that do not seem to understand this relationship.

The correlation coefficient is non-dimensional (item 3). 44.6 percent of the students in this sample did not recognise this property, consequently they thought that the correlation coefficient depends on the measurement units.

Positive correlation and variability (item 5). In item 5, options a) and c) are equivalent, however the students replied in a different way to both options. The students preferred positive cases to negative ones, a fact observed by Nisbett and Ross (1980) as we have described in section 2.

Strength of correlation (item 1). Ranking several values of the correlation coefficient makes the knowledge of the various types of dependence explicit and serves to assess the students' understanding of strength and sign of correlation. In Table 2 we show the frequencies and percentages of the different students' orderings of correlation coefficients. Morris (1997) gave this same item to her students, although she only asked them to find the correlation coefficient that represented the greatest strength of

correlation. We preferred to ask the students to rank all the values. Taking this difference into account, our results were very close to those of Morris. The percentage of correct answers was 46.1%. 137 students put zero in the last place (71%). These students recognised that r=0 indicates the smallest correlation between two variables.

Table 2. Frequency of Different Rankings of Correlation Coefficients in Item 1

Ranking of correlation coefficient	Frequency	Percentage
(I) -0.8, 0.5, -0.4, 0.2, 0	89	46.1
(II) 0.5, 0.2, 0, -0.4, -0.8	33	17.1
(III) 0.5, 0.2, -0.4, -0.8, 0	26	13.5
(IV) 0.5, 0.2, -0.8, -0.4, 0	9	4.7
(V) 0, 0.2, -0.4, 0.5, -0.8	3	1.6
(VI) -0.8, -0.4, 0.5, 0.2, 0	3	1.6
(VII) Other	23	11.9
No answer	7	3.6
Total	193	100.0

In response [II] the students classify the values applying the usual ordering in R, that is to say, they only take into account the numerical meaning of the correlation coefficient and not its statistical meaning. The knowledge of the order of negative numbers is an obstacle (Brousseau, 1997) when these students tried to rank the strength of correlation. Similar results were obtained by Batanero, Estepa and Godino (1997) and Batanero, Godino and Estepa (1998) when analysing the learning process by some students, and observing their comparison of two negative correlation coefficients. These students did not consider the greatest absolute value to represent the greatest correlation, and they also ordered negative correlation coefficients as if they were just negative numbers.

The students using the rankings (III), (IV) (VI) classified all the values, except zero, according to different types of ordering. In (III) and (IV), the students wrote the positive numbers before the negative numbers. In (III) they classified each subset of numbers according to the usual numerical order, while in (IV) the students consider −0'8 to be greater than −0'4, which shows that these students have not mastered the order in R. The same difficulties are observed in (VI).

Correlation and proportionality. A frequent activity when teaching and learning proportionality is to ask students to discriminate between direct and inverse proportionality. The aim is providing students with tools for dealing with the different types of proportionality. When considering direct proporcionality in a functional setting, we usually characterise a *linear function f(x) = kx*, as a function that verifies the properties (2) and (3).

$$f(x+y) = f(x) + f(y) \quad (2)$$

$$f(\lambda x) = \lambda f(x) \quad (3)$$

With the aim of teaching students how to distinguish direct proportionality, it is frequent to tell them that, in equation (3) an increase (decrease) of a variable ($x \rightarrow \lambda x$) corresponds to a proportional increase (decrease) of the other ($f(x) \rightarrow \lambda f(x)$).

The analogy with correlation, and the use of the similar wording (direct and inverse correlation / direct and inverse proportionality), may lead some students to confuse proportionality and correlation. It is not surprising that some students compare the correlation coefficient r with a proportionality constant. This comparison can be done in

standardised variables, where the regression equation is reduced to a homogeneous linear function and r is the constant of proportionality that relates the value X to the average of the distributions Y conditioned to X, but not in the general case. However, 22.8 % of our students answered that *"if r = 0'6 the correlation between the variable X and Y is double that when r = 0'3"* (item 7b), and thus these students extended proportionality to the general case where it is not true. Our interpretation is reinforced by the number of correct answers in items 5a and 5c.

Confusion between r and r^2. The determination coefficient is a measure of the goodness of fit of the scatter plot to a straight line. It also expresses the reduction of variance when predicting the *y* value by using the regression line Y/X instead of the equation $y = \bar{y}$. Some of our students (10.4%) confused the correlation coefficient r with the determination coefficient r^2, since they chose item 7d, where it is said that r can be interpreted as a percentage of the variance.

Value of the correlation coefficient and its relationship with both regression lines (item 6, 11 and 12). Only half the students in the sample recognised the perpendicularity of the regression lines when the correlation coefficient is zero, while a seventh of the students (14.5%) replied that in the case of a null correlation coefficient, both regression lines have the same slope, that is to say, they are parallel (option a of items 6 and 11). Only 38.9% of the students recognised that if the two regression lines have the same slope then the correlation coefficient is +1 or -1, choosing the two correct options (options b and c, item 11). However, 37.3% of the students only accepted a correlation coefficient value of +1 in this case. We confirm here some resistance to accept the inverse correlation in agreement with Batanero, Estepa, Godino and Green, (1996) and Morris, (1997). Finally, 2/3 of the students recognised that when correlation is perfect the two regression lines are parallel (item 12).

4.5. CORRELATION AND PREDICTION

A main objective in both scientific research and decision making is to find causal relationships between variables through the analysis of association. However, there are different types of relationships that can explain correlation, such as unilateral dependence, interdependence, concordance, or spurious correlation. In the first case we need to distinguish between the dependent variable and the independent variables. Below we study the students' understanding of these questions, which are essential in many types of research (items 8, 9 and 10).

Correlation and causality (item 8). The causal conception of association (identifying association and causality) has been described in Estepa and Batanero (1996) and Morris (1997) and it might be influenced by the instruction that the students receive in mathematics, science and other areas, where all the phenomena are given a causal explanation. The students transfer their ideas from other topics to the study of association, and when finding a strong association between two variables they might infer there is a causal dependence between the variables.

In item 8c, a high percentage of students (89,1%) recognised that a strong relationship is expected when there is a high value of the correlation coefficient; however this percentage is considerably reduced (29.0%) when students have to recognise that other factors can influence the results (option a). There were also 22.3% of the students who accepted with certainty that double the surface sown corresponded to double the crop, expressing a causal conception of association. This fact is reinforced when we observe that 36 students (18.6%) chose simultaneously answers c) and d) in

item 8, which seems to indicate that these students reasoned that, as the correlation coefficient was high, it could be asserted with certainty that twice the planted surface will correspond to twice the crop.

Distinction between independent and dependent variables (item 10). A previous step in a prediction problem is to distinguish between the independent and the dependent variables. This distinction is essential to discriminate the two regression lines, which according to Sánchez (1999) are only correctly distinguished by 32.0% of undergraduates. The low percentage of correct answers (36.3%) in this item confirms Sánchez's (1999) results as 73.7% of the students in our sample did not discriminate between the variables and 34.2% confused the two regression lines. This fact will have a negative influence in the use of the regression equations for predictions in research projects.

Interdependence (item 9). In functional dependence independent and dependent variables are univocally determined. Association extends the idea of functional dependence although, in this case, the variables are not always univocally determined, since in situations of interdependence, both variables can play either the role of dependent or independent variable. This situation is presented in item 9, where almost 60% of the students accepted interdependence, although 41.4% preferred to fix one variable as independent and the other as dependent. We note that while height was chosen as an independent variable by 34.7% of the undergraduates, weight was chosen as an independent variable only by 6.7% of students. This possibly can be due to the fact that height does not usually vary in an adult, while weight normally does.

5. IMPLICATIONS FOR THE TRAINING OF RESEARCHERS

In this paper we have summarised the main research results on the understanding of association, including our empirical study carried out with undergraduate students, whose results can be applied to future researchers and professionals. Many research projects are intended to find related variables and to establish causal relationships between them, in such a way that a response variable can be explained or predicted by one or several explanatory variables. Consequently, a correct understanding of statistical association and of all its interrelated elements of meaning is basic in research methodology. In spite of this, undergraduates do not always acquire a correct understanding of association and show misconceptions, some of which are not always overcome after instruction. Association judgements are influenced by previous theories and misconceptions, and incorrect strategies are sometimes used to carry out association judgements.

Our research results suggest that some elements of the meaning of association are only acquired by a few students. The difficulty of the concepts related to association (such as correlation, covariance, and regression line) was greater than predicted, since the mean rate of correct answers in the questions dealing with these concepts is slightly superior to fifty percent (53.2%).

The undergraduate's difficulties in relating linear regression, correlation coefficient and covariance should warn us about the need to emphasise these fundamental relations in the teaching of association and base this teaching on the understanding of covariance, as a measure of the joint variation. Few undergraduates in our study showed a correct knowledge of the properties of the correlation coefficient. In particular non-dimensionality, strength of correlation, and negative correlation were scarcely understood. The existence of possible obstacles associated to these mistakes suggests

that these difficulties could possibly be overcome through didactical situations that produce cognitive conflicts.

We have also found confusion between association and causality; lack of distinction between interdependence and unilateral dependence, problems in adequately choosing the dependent and independent variables; and excessive emphasis on linear dependence. Linear dependence is sometimes an oversimplification, and this should be emphasised in teaching, where students should be encouraged to explore other models, which nowadays is easy with the help of computers.

Without a full integration of the elements of meaning, the conceptions of correlation and regression acquired by future researchers will be biased and incomplete, and will produce improper uses of statistics and, consequently, incorrect research conclusions. Therefore, these points should be taken into account in the researcher's statistical instruction. The planning of the researchers' training in statistics should take into account the elements of meaning of correlation and regression and psychological and didactic research results, in order to improve the results of the instruction.

ACKNOWLEDGEMENTS:

This research was supported by the grant PB97-0851, Secretaría de Estado de Universidades, Investigación y Desarrollo, Spain.

APPENDIX: QUESTIONNAIRE AND RESULTS

(Correct answers are marked with x)

Item 1. Order the 5 following correlation coefficients from the one that indicates the highest amount of correlation to that which indicates the lowest amount of correlation or no correlation: 0'5, -0'8, 0'2, -0'4, 0.

	Item 2. When the covariance between X and Y is greater than 0, then:	Frequency	Percentage
x	a) The correlation between X and Y is positive	127	65.8
x	b) The relationships between X and Y might be non-linear	22	11.4
	c) The variables might be not interrelated	17	8.8
x	d) The regression line has a positive sign slope	87	45.1
x	e) The correlation coefficient is positive	115	59.6

	Item 3. John correlated height and weight of graduate male students. He used meters and kilograms as his measures. Angela also correlated height and weight on the same group of subjects using centimetres and grams as her measures. John and Angela computed correlation coefficients between their two sets of measures.	Frequency	Percentage
	a) Angela's correlation coefficient will tend to be greater than John's	23	11.9
x	b) The two correlation coefficients will be approximately equal	107	55.4
	c) John's correlation coefficient will tend to be greater than Angela's	17	8.8
x	d) The value of the correlation coefficient depends on the data spread	102	52.8

	Item 4. When the strength of the relationship between two variables decreases:	Frequency	Percentage
	a) The slope of the regression line of Y/X increases	23	11.9
	b) The slope of the regression line of X/Y increases	32	16.6
x	c) There is greater spread in the scatter plot	124	64.2
	d) The absolute value of the covariance increases	33	17.1

	Item 5. If two variables are positively correlated:	Frequency	Percentage
x	a) As one increases, the other increases	152	78.8
	b) As one decreases, the other increases	8	4.1
x	c) As one decreases, the other decreases	90	46.6
	d) There is a linear relationships between the two variables	91	47.2

	Item 6. If the correlation coefficient between two variables is zero:	Frequency	Percentage
	a) Both regression lines Y/X and X/Y are parallel	28	14.5
x	b) The value of the covariance is zero	84	43.5
	c) Both regression lines Y/X and X/Y coincide	19	9.8
x	d) The variables may have a non-linear relationship	64	33.2
x	e) Both regression lines Y/X and X/Y are perpendicular	96	49.7

	Item 7. If *r* is the correlation coefficient of two variables:	Frequency	Percentage
	a) If $r = 0$ the variables are independent	138	71.5
	b) When $r = 0.6$ there is twice as much correlation between the variables X and Y than when r = 0.3	44	22.8
x	c) When there is a perfect lineal relationship between the variables, *r* is +1 or - 1	112	58.0
	d) The correlation coefficient can be interpreted as a percentage of the variance	20	10.4

	Item 8. A farmer studied the wheat surface sown in thousands of hectares and the crops obtained in millions of metric quintals, over five consecutive years, and he obtained a correlation coefficient of 0.91. Consequently,	Frequency	Percentage
x	a) There may be other factors that make the results vary	56	29.0
	b) We should have to take a larger sample to be able to express the relationship between the planted surface and the obtained crop	6	3.1
x	c) There is a strong correlation between the crop and the planted surface	172	89.1
	d) If we plant double the surface, we will double the crop with absolute certainty	43	22.3

	Item 9. In which forecast do you have more confidence?	Frequency	Percentage
	a) Estimating the height of a man from his weight	13	6.7
	b) Estimating the weight of a man from his height	67	34.7
x	c) The two estimations are equally reliable	114	59.1

	Item 10. In a study incomes are used to predict savings. Both variables are measured in thousands of pesetas. Which of the following statements is true if y = 1000 + 0.1x is the regression equation?	Frequency	Percentage
	a) y is the income, x is the saving, the income is the independent variable	26	13.5
	b) y is the income, x is the saving, the saving is the independent variable	66	34.2
	c) y is the saving, x is the income, the saving is the independent variable	56	29.0
x	d) y is the saving, x is the income, the income is the independent variable	70	36.3

	Item 11. If both regression lines have the same slope, what is the value of r ?	Frequency	Percentage
	a) 0	29	15.0
x	b) 1	149	77.2
x	c) –1	80	41.5
	d) 0.5	7	3.6

	Item 12. If X and Y have perfect correlation (r=1 or r=-1), the angle between the two regression lines is:	Frequency	Percentage
	a) 120º	5	2.6
	b) 90º	32	16.6
	c) 45º	24	12.4
x	d) 0º	129	66.8

REFERENCES

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgement of influence. *Learning and Motivation*, 14, 381-405.

Arkes H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General*, 112(1), 117-135.

Batanero, C., Estepa, A. Godino, J. D., & Green, D. (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. Burril (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 191-205). Voorburg, The Netherlands: International Association for Statistical education and International Statistical Institute.

Batanero, C., Godino, J., & Estepa, A. (1998). Building the meaning of statistical association through data analysis activities. In A. Olivier, & K. Newstead (Eds.), *Proceedings of the 22nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 221-236). Stellenbosh, South Africa: University of Stellenbosh.

Bangdiwala S. I. (2001). Training of statisticians worldwide to collaborate as co-investigators within country clinical epidemiology units: The experience of the International Clinical Epidemiology Network (INCLEN). In Batanero, C. (Ed.), *Training researchers in the use of statistics* (pp. 265-275). Granada: International Association for Statistical Education and International Statistical Institute.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory and Cognition*, 10(6),

511-519.

Bishop, G, & Talbot, M. (2001). Statistical thinking for novice researchers in biological sciences. In Batanero, C. (Ed.), *Training researchers in the use of statistics*. (pp. 215-226). Granada: International Association for Statistical Education and International Statistical Institute.

Blumberg, C. J. (2001). Training regular education and special education teachers in the use of research methodology and statistics. In Batanero, C. (Ed.), *Training researchers in the use of statistics*. (pp. 231-244). Granada: International Association for Statistical Education and International Statistical Institute.

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Dordrecht: Kluwer.

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

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

Cruise, R. J., Dudley, R. L., & Thayer, J. D. (1984). *A resource guide for introductory statistics.* Dubuque, IO: Kendall / Hunt.

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology*, 73(1), 9-14.

Estepa, A., & Batanero, 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., Batanero, C., & Sanchez, F. T. (1999). Students' intuitive strategies in judging association when comparing two samples. *Hiroshima Journal of Mathematics Education*, 7, 17-30

Godino, J. D., & Batanero, C. (1997). Clarifying the meaning of mathematical objects as a priority area for research in mathematics education. In A. Sierpinska, & J. Kilpatrick (Eds.), *Mathematics education as a research domain: A search for identity* (pp. 177-195). London: Kluwer.

Godino, J. D., Batanero, C., & Gutiérrez Jáimez, R. (2001). The statistical consultancy workshop as a pedagogical tool. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics* (pp. 339-353). Granada: International Association for Statistical Education and International Statistical Institute.

Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant á la logique de l'adolescent* [From the child's logic to the adolescent's logic]. Paris: Presses Universitaires de France.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of de contingency between responses and outcomes. *Psychological Monographs,* 79, 1-17.

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.

Kahneman, P. Slovic, & A. Tversky (Ed.), (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Lane, D. M., Anderson, C. A., & Kellam, K. L.(1985). Judging the relatedness of variables: The psychophisics of covariation detection. *Journal of Experimental Psychology. Perception and Performance*, 11(5), 640-649.

Langer, E. J. (1975). The illusion of control. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 231-238). New York: Cambridge University Press.

Morris, E. J. (1997). An investigation of students' conceptions and procedural stills in the statistical topic correlation. *Centre for Information Technology in Education*, Report n. 230. Milton Keynes, U.K: The Open University.

Murphy G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.

Nisbett, R., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment.* New Jersey: Prentice Hall.

Pérez Echeverría, M. P. (1990). *Psicología del razonamiento probabilístico.* [Psychology of probabilistic reasoning]. Madrid: Ediciones de la Universidad Autónoma.

Peterson C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68(1), 29-46.

Sanchez, F. T. (1999). *Significado de la correlación y regresión para los estudiantes universitarios* [Meanings of correlation and regression for undergraduates]. Unpublished PhD. University of Granada.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 8 (3), 208-224.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation*, 18, 147-166.

Shimada T. (2001) Precaution against errors in using stochastic software. In Batanero, C. (Ed.), *Training researchers in the use of statistics.* (pp. 129-137). Granada: International Association for Statistical Education and International Statistical Institute.

Truran, J. M. (1997). Understanding of association and regression by first year economics students from two different countries as revealed in responses to the same examination questions. In J. Garfield, & J. M. Truran, (Eds.), *Research papers on stochastics educations from 1997* (pp. 205-212). Minneapolis, MN: University of Minnesota.

Tversky, A., & Kahneman, D. (1982). Causal schemas in judgments under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 117-128). New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1982b). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*, (pp. 3-20). New York: Cambridge University Press.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-248.

Antonio Estepa,
Facultad de Humanidades y Ciencias de la Educación
Universidad de Jaén, Spain
E-mail: aestepa@ujaen.es

Francisco Tomás Sánchez Cobo
Escuela Politécnica Superior
Universidad de Jaén, Spain
E-mail: fsanchez@ujaen.