Errors and difficulties in understanding elementary statistical concepts

C. BATANERO, J. D. GODINO, A. VALLECILLOS
D.R. GREEN and P. HOLMES
International Journal of Mathematics Education in Science and Technology, 25(4), 527-547

This paper presents a survey of the reported research about students' errors, difficulties and conceptions concerning elementary statistical concepts. Information related to the learning processes is essential to curricular design in this branch of mathematics. In particular, the identification of errors and difficulties which students display is needed in order to organize statistical training programmes and to prepare didactical situations which allow the students to overcome their cognitive obstacles. This paper does not attempt to report on probability concepts, an area which has received much attention, but concentrates on other statistical concepts, which have received little attention hitherto.

1. Introduction

The teaching of statistics is currently increasing substantially in many countries, due to its widely recognized place in the general education of citizens. Some countries have dedicated much effort to design curricula and specific materials, for example those produced for the Schools Council Project on Statistical Education in England by Holmes *et al.* [I], the Quantitative Literacy Project in USA Landewehr and Watkins [2], Landewehr *et al.* [3], Gnanadesikan *et al.* [4]), and Azar y Probabilidad in Spain (Godino *et at.* [5]). The increasing interest in teaching statistics is also shown by the existence of specific journals *(Teaching Statistics; Induizioni; Stochastik in der Schule);* by regular international conferences (ICOTS I in Sheffield, 1982 [6]; ICOTS II in Victoria, 1986 [7]; ICOTS III in Otago, 1990 [8]) by an ongoing series of round table conferences promoted by the ISI (most recently at Lennoxville in 1992) and by the formation in 1992 of an international association, IASE (International Association for Statistical Education). This interest is also demonstrated by the setting up of Centres for Statistical Education in England, Italy and USA, the *Newsletter of the International Study Group for Research on Learning Probability and Statistics* from the University of Minnesota; the Edstat bulletin board and, very recently, the electronic *Journal/or Statistical Education* published by North Carolina State University-

The greater emphasis given to statistics in the different curricula, for example, the N.C.T.M. Standards in the USA [9], the Mathematics National Curriculum for England and Wales [10] and the new Spanish curricular proposal [11, 12] requires an intensive preparation of teachers, in order to allow them to successfully accomplish their educational goal. Many teachers need to increase their knowledge of both the subject matter of statistics and appropriate ways to teach the subject. This preparation should also include knowledge of the difficulties and errors that students experience during the learning of this topic.

The aim of this paper is to contribute to the diffusion of the results of research related to difficulties and errors, which are not sufficiently known by teachers. Some surveys of research in probability and statistics education have been reported: Hawkins and Kapadia [13], Garfield and Alhgren [14], Scholz [1 5] and Shaughnessy [16], but these works have been oriented to *researchers* rather than to *teachers* and their main goal has been to identify new research questions. Moreover, they have been focused primarily on probability, because the research in this area is far more extensive than that which relates to statistical concepts, although a relevant exception is the new book for statistics teachers by Hawkins *et al.* [17].

In this paper, we analyse research results concerning some of the main elementary statistical concepts and procedures, which have been included in many recent curricular proposals for teaching statistics at elementary levels. This analysis shows the complexity of some of these topics, and should provide the teacher with a deeper understanding of pupils' stochastical reasoning.

We consider it necessary to start this exposition by explaining the relevance of this type of research and by defining some related theoretical concepts. Nevertheless, we advise the reader that:

- statistics has received, to date, less attention than other mathematical topics;
- most research has been carried out in experimental situations and very little has been concerned with normal classroom practice;
- much has been with very young subjects or 18 + college students and rather little with the 11-16 age range;
- most early research was undertaken by psychologists rather than by statistical educators, although that has begun to change.

2. Research on errors, conceptions and obstacles: some theoretical concepts

An important part of the theoretical and experimental research which is, at present, being carried out in mathematics education, arises from the observed fact that students fail when they are asked to complete some tasks. Sometimes their responses are wrong, when compared with a standard accepted answer. Or, simply, they are unable to provide any answer at all. If this failure is not solely due to inattention, we say that the task is *too difficult* for the particular students. As Centeno ([18] p. 142) points out, 'a difficulty is something that inhibits the student in accomplishing correctly or in understanding quickly a given item. Difficulties may be due to several causes: related to the concept that is being learned, to the teaching method used by the teacher, to the student's previous knowledge, or to his ability'.

Moreover, errors and difficulties do not arise in a random, unpredictable way. Frequently it is possible to discover regularities in them, to find some association with other variables of the proposed tasks, of the subjects, or of the present and past circumstances. Didactic research is directed to characterize these regularities and to build explanatory models, in terms of relationships among the intervening variables. Some authors, such as Radatz [19], consider that the error analysis 'should be considered a promising research strategy for clarifying some fundamental questions of mathematical learning' (p. 16). In the same way, Borassi [20] presents the analysis of errors in mathematical education as 'a motivational device and as a starting point for creative mathematical exploration, involving valuable problem solving and problem posing activities' (p. 7).

A widely shared principle in educational psychology is stated by Ausubel *et al.* [21], 'the most important factor to influence learning is the student's previous knowledge. We ought to discover it and to teach consequently'. A recent interest within mathematics education research is exploring students' *conceptions* (Confrey [22]), which is a consequence of the aforementioned principle.

A difficulty is that some of these conceptions, which allow the student to solve correctly a given set of problems, are found inappropriate or inadequate when applied to more general situations. The student shows resistance to replacing these conceptions. In these circumstances, we speak of the existence of a *cognitive obstacle* which could explain the observed errors and difficulties. Brousseau [23] describes the following features of obstacles:

a. An obstacle is knowledge, not a lack of knowledge.
b. The student employs this knowledge to produce the correct answer in a given context, which he frequently meets.
c. When this knowledge is used outside this context, it generates mistakes. A universal answer requires a different point of view.
d. The student ignores the contradictions produced by the obstacle and resists establishing a deeper knowledge. It is essential to identify the obstacle and to replace it in the new

learning.

e. After the student has overcome the obstacle, recognising its inexactitude, nevertheless it recurs sporadically.

Brousseau has identified three kinds of obstacle:

a. *Ontogenic obstacles* (sometimes called psychogenetic obstacles) are due to features of child development. For example, proportional reasoning is required to understand the idea of probability.
b. *Didactical obstacles* arise from the didactical options chosen in teaching situations. For example, introducing new symbolism such as

$$(\sum x_i)/n$$

when the students need to work with concrete examples.

c. *Epistemological obstacles* are intrinsically related to the concept itself, and carry part of the meaning of the concept. For example, circularities which occur in the different definitions of the meaning of probability (classical, subjective, . . .) which show the necessity of an axiomatic definition. So, a necessary condition to build a relevant conception for a given concept is to identify and to overcome those obstacles, using historical and didactical analysis.

Finally, we note that other difficulties experienced by students are due to a lack of the basic knowledge needed for a correct understanding of a given concept or procedure. The purpose of the characterization of mathematical conceptions and obstacles is that this allows us to identify the different components which are implied in the understanding of a given concept. Recent research, for example, that concerning ‘acts of understanding’ in the context of the limit of a numerical sequence (Sierpinska [24]), shows the complexity of the meaning of mathematical objects.

3 Frequency tables and graphical representation of data

We start our exposition about errors and difficulties in learning statistics with those related to frequency tables and graphical representations.

Ability to critically read data is a component of numerical literacy and a necessity in our technological society. Curcio [25] has distinguished the three following different levels in the comprehension of data:

a. Reading the data in which interpretation is not needed. Only facts explicitly expressed in the graph or table are required.
b. Reading within the data, which requires comparisons and the use of mathematical concepts and skills.
c. Reading beyond the data where an extension, prediction or inference is needed.

For example if we analyse the tasks related to the interpretation of a scatter plot, ‘reading the data’ refers to questions about the labelling of the plot, interpretation of scales, or finding the value of one of the coordinates of a point given the other one. ‘Reading within the data’ refers, for example, to questions about the intensity of the co-variation, about whether this relationship could be represented or not by a linear function, or about whether the dependence is direct or inverse. Finally, if we require prediction of the y value for an x coordinate value not included explicitly in the graph, we would be working at the ‘reading beyond the data’ level.

Curcio assessed the effect of prior knowledge of the topic, mathematical content and graphical form on understanding the mathematical relationships expressed in graphs. He found that the main difficulties appear at the two high levels ('reading within the data' and -reading beyond the data') in the interpretation of the graphs and also he showed the effect of grade and age differences on these difficulties

Pereira-Mendoza and Mellor [26] undertook research into 9 to 11 year old students' conceptions in bar graphs. They reported errors in scales, lack of identification of patterns in graphs, errors in predictions, and inappropriate use of information.

Li and Shen [27] found many examples of incorrect choice of graph in statistical projects by secondary students. Some used a line graph with qualitative variables or a bar graph to represent the evolution of an index number through a sequence of years. The problem of incorrect choice of graphical representation is worsened by the increasing availability of graphical software. Li and Shen have remarked that through software restrictions and students' lack of knowledge about the software, very often the chosen scales are inadequate. Other common technical weaknesses reported by Li and Shen were:

- The scales of either or both the vertical and horizontal axes are omitted.
- The origin of coordinates is not specified.
- Insufficient divisions in scales on the axes are provided.
- The axes are not labelled.

At other times, inappropriate use of graphical software hides a misconception, as, for example, when in a pie chart the sectors are not proportional to the frequencies in the categories. The ease of making graphs with a computer carries with it the danger of not applying commonsense; for example, comparing in the same graph 30 chairs with 50 kg of meat. These problems can arise from too much choice provided in the software.

4. Summaries of the distribution: statistical measures

4.1. *The mean*

The mean is not only one of the most important concepts in statistics, but also it has many applications in everyday life. This concept is simple in appearance, but Pollatsek *et al.* [28] found errors in combining two weighted means as if they were simple means, as in the following item:

> There are ten people in an elevator, four women and six men. The average weight of the women is 120 pounds and the average weight of the men 180 pounds. What is the average of the weight of the ten people in the elevator?

There are a number of ways to go wrong when applying the computational rule: for example, using (120+180)/2 = 150, has been reported in the aforementioned research. Hawkins *et al.* [17] point out that this is not a sensible question, because of the two different distributions which are involved, but this problem could be set in terms of a student doing different numbers of hours at different vacation jobs.

The situations in which a weighted mean must be computed are not easily identified by students. Li and Shen [27] noted that, when using data grouped in intervals, students sometimes ignore the frequency of each one of the intervals when computing the mean.

Traditionally great emphasis has been placed on computational aspects of data analysis and this has been reflected in the research. Undeniably, the meaning of many statistical concepts depends upon a numerical context. An important question which has not yet been addressed is: 'In practice, can conceptual understanding be separated from computational competence?'

Another different item used in research by Pollatsek *el al.* [28] was designed to evaluate university students' conceptions about the expected value of an observation of a random variable for which the population mean is known.

You know that on average the verbal score of the population of high school seniors in a large school system is 400. You pick a sample of 5 students. The first 4 students in your sample have the following scores: 380, 400, 600, 400. What do you expect the fifth student's score to be?

The 'correct' answer to this item is 400, the expected value in the population. But, nevertheless, some students erroneously thought that they would get a better estimate of the last student's score by computing the number that would make all five-scores average to 400.

These results show that the difficulties with the mean are not only produced at computational level. Skemp [29] has drawn the distinction between instrumental and relational understanding of a concept. Instrumental understanding consists of having available a collection of isolated rules for arriving at the answer to specific problems. Relational understanding consists of having available appropriate schemes sufficient to solve a much broader class of problems. In a similar sense, Hiebert et *al.* [30] speak about conceptual knowledge and procedural knowledge.

Mevarech [31] observed that a possible explanation to the procedural errors found by Pollatsek *et al.* [28] is that students mistakenly assume that a set of numbers together with the operation of arithmetic mean constitutes a mathematical group satisfying the four axioms of closure, associativity, identity element and inverse element. This belief is obviously false, as can be seen in the following examples:

a. In calculating the overall mean of three given numbers one will obtain a different mean when averaging the first two numbers with the third number than when averaging the last two numbers with the first number.
b. An identity number does not exist, because the value of the mean is influenced by the value of every score in the distribution. Nevertheless, some students believe that adding a new value of zero to a distribution does not alter the value of the mean.

When the statistical measures mean and variance are first introduced to students, they constitute completely different operations, not extensions of those which are already known. However, novice students may unconsciously relate the properties which they know hold for the arithmetical operations to the mean and variance. So, the emphasis put into the learning of fundamental properties of arithmetical operations may constitute an obstacle to understand the computation of the mean value, because to this operation is given some non-existent properties.

The research to which we have referred above concerns the *computational* aspects of the mean. Concerning *conceptual* understanding, Strauss and Bichler [32] studied the development of children's understanding of the mean and they distinguished the following properties:

a. The average is located between the extreme values.
b. The sum of the deviations from the average is zero.
c. The average is influenced by all the values.
d. The average does not necessarily equal one of the values that were summed.
e. The average can be a fraction that has no counterpart in physical reality.
f. When one calculates the average, a value of zero, if it appears, must be taken into account.
g. The average is representative of the values hat are averaged.

For each one of these properties they used different tasks, varying the material used in the testing (continuous, discrete) and the medium of presentation (story, concrete and numerical). No significant effects were found for the material or the medium of presentation. Their results also suggested that the children's understanding of the mean changes with age and that properties (a), (c), (d) are easier than (b), (f) and (g).

We have stated that the mean is a 'typical' or 'representative' value of the distribution. Campbell [33] noticed that, for this reason, there is a tendency to situate the mean in the centre of the range of the variation of the data. This is true when the distribution is symmetrical, but, if it is not, the mean is shifted towards one of the extremes and the median or mode would be a better representative of the data. Understanding the idea of 'typical value' implies, according to Russell and Mokros [34] three different types of abilities:

a. Given a data set, understanding the necessity of employing a central value and choosing the best one for the particular case.
b. Building a data set which has a given mean value.
c. Understanding the effect that a change in all or in a part of the data has on the averages (mean, median, mode).

Russell and Mokros [34] studied 4th to 8th grade students' conceptions about averages, using these kinds of tasks, and they found the second type was the most difficult. This type of task has also been proposed by Goodchild [35] who provided students with matchboxes upon which was printed 'average content 35 matches'. One of his questions required the pupil to make up an hypothetical distribution for the content of 100 boxes. The most notable feature of these distributions was their lack of form, because the graph did not look at all bell-shaped. Goodchild suggested that this is due to a lack of understanding of the average as a measure of location of the distribution which results from a stochastic process.

Russell and Mokros also found four general categories whereby to classify the students' misconceptions about averages:

a. the 'most frequent value' or mode;
b. the 'reasonable value';
c. the 'midpoint';
d. an 'algorithmic relationship'.

Each one of these aspects may be true in a given circumstance but may not be appropriate in another. They ended their paper by pointing out the necessity of the use of different contexts and representations in the teaching of a mathematical concept.

In summary, the inability of some students to solve problems is that they have yet to acquire a purely formal concept for the mean. Knowledge of a computational rule not only does not imply any deep understanding of the underlying concept, but may actually inhibit the acquisition of a more complete conceptual knowledge. Learning a computational formula is a poor substitute for gaining an understanding of the basic underlying concept. Most students know the rule by which the mean is calculated. If, however, students have only the computational knowledge of the mean, they are likely to make predictable kinds of error in all but the most transparent problems.

4.2. *Measures of spread*

The study of frequencies cannot be reduced to the study of averages; two different data sets with the same average may have different degrees of variability. Campbell [33] has pointed out that a frequent error is to ignore the spread of data. Lovie and Lovie [36] reported that when estimating means the variance is a factor, and that accuracy in estimating variance depends on the magnitudes involved.

The standard deviation measures how strongly data depart from the central tendency. Nevertheless, Loosen *et al.* [37] noticed that many textbooks put a stronger emphasis on the heterogeneity among the observations than on their deviations from the central tendency. As Loosen *et al.* note, the words used: variation, dispersion, diversity, spread, fluctuation, etc., are open to different interpretations. It is clear to the teacher whether these words refer to a relative or an absolute diversity. In one experiment, they took 154 psychology freshmen who had not received any lectures on variability. They showed the students two different sets of blocks, A and B. The lengths of blocks in set A were 10, 20, 30, 40, 50 and 60 cm. and the lengths of blocks inset B were 10, 10, 10, 60, 60 and 60 cm.

The students answered as follows: 50% said set A was more variable, 36% said B and 14% said they were equally variable. Loosen *el al.* interpreted this as showing that the intuitive concept of variability is concerned with 'unalikability', i.e. how much the values differ from each other (rather than from some fixed value such as the mean). In this sense, set A can certainly be considered more variable than set B. However, the standard deviation for set A is less than that for set B, indicating that standard deviation is a special measure of variability.

Mevarech [31] found, in university students, similar difficulties in the calculation of variance as when calculating the mean. In particular, the students tend to assume that the data set and the operation of variance together constitute a group structure. However, Mevarech's analysis is not entirely convincing.

One of the most common uses of the mean and standard deviation is the computation of z-scores. Most students have no difficulty in understanding this concept nor in computing z-scores for a particular data set. Nevertheless, Huck et al. [38] have noticed two widespread students' misconceptions concerning the range of variation of z-scores calculated from a finite sample or from a uniform distribution.

On the one hand, some students believe that z-scores will always range from -3 to +3. Other students think that there is no restriction on the maximum positive and negative values for z-scores. Each of those beliefs is linked to a misconception about the normal distribution. The students who think that z-scores always vary from-3 to + 3, have frequently used either a picture or a table of the standard normal curve, with this range of variation. In a similar way the students who believe that z-scores have no upper or lower limits, have learned that the tails of the normal curve are asymptotic to the abscissa and they make an incorrect generalization, because they do not notice that no finite distribution is exactly normal.

For example if you consider the number of girls out of ten newborn babies this is a random variable X which follows the binomial distribution with $n = 10$ and p=0.5. The mean of this variable is np=5 and the variance is $npq=2.5$. So the maximum z-score that could be obtained from this variable is $z_{max} = =(10-5)/\sqrt{2.5}=3.16$ and thus we have a finite limit, but it is greater than 3.

4.3. *Order statistics*

In recent years, the topic order statistics has received a great deal of attention for two different reasons:

- Exploratory data analysis, which started with Tukey [39], is based on order statistics, because they are *robust,* that is to say, they are not very sensitive to fluctuations in data or to outliers.
- Non-parametric methods are based on order statistics. These methods require fewer assumptions to be applied, and so can be more widely employed than parametric inference.

The study of order statistics presents computational as well as conceptual difficulties. First of all, the computation of median, quantiles and percentiles is taught with a different algorithm for data grouped in intervals than for non-grouped data. As we know, the decision of whether or not to group the data and the selection of the width of intervals is taken by the person who performs the analysis. Schuyten [40] has suggested that even university students find it difficult to accept two different algorithms for the same concept, and, moreover, different values for the same parameter depending on the chosen algorithm or on the width of the intervals.

If we work with non-grouped data, the graphical representation of the cumulative frequencies is a discontinuous function, which takes a constant value between two consecutive values of the variables. Estepa [41] observed students' difficulties when interpreting the cumulative frequency graph, because a value in the y-axis may have two different images, or several different values in the y-axis may have the same image.

Schuyten [40] has pointed out the large distance between the conceptual knowledge of the median and the algorithm employed to obtain its value. In going from the definition of the median as 'middle value of the distribution' or as the 'value such that exactly half the data are inferior to it' to its calculation, there are many steps which are not always sufficiently stated or not sufficiently understood. The final algorithm consists in solving the inequality: $F\backslash x) \leq n/2$, where n is the number of data items. It is necessary to solve this inequation in which $F(x)$ is the empirical distribution function and may not be given in an algebraic way but only by means of a table of numerical values. So one must use interpolation in order to approximate $F(x)$.

Barr [42] also noticed the lack of understanding of the median in a pilot study with students aged 1 7 to 21 years. About 50% of students gave an incorrect answer to the following question:

> The median in the following set of numbers: 1,5,1,6,1,6,8, is:
> *(a)* 1 (A) 4 (r) 5 (r/) 6 *(e)* other value (/) don't know-

Most students had grasped the idea that the median is a central value of something. Doubt as to what that something is, was evident. The students could interpret the- median as the middle point of the figures in the frequencies column, or as the middle point in the values of the variable column, or even as the middle point in the list of numbers before they have been ordered.

5. Association and regression

The idea of statistical association extends functional dependence, and is fundamental to many statistical methods which allow us to model numerous phenomena in different sciences. The term association is used to refer to the existence of statistical dependence between two random variables, whether they be quantitative or qualitative. The word correlation usually refers to the association between two quantitative variables. Both terms do not necessarily imply a cause- effect relationship, but merely the existence of a co-variation between the variables.

At an elementary level, there are several different concepts in this topic area: contingency tables, linear regression, and correlation between quantitative variables. Another topic which is related to the idea of association is experimental design by which we study how statistical tools make it possible to reach conclusions about a world in which a large number of variables influence any particular measurement (Rubin and Rosebery [43].

5.1. *Contingency tables*

A contingency table or cross-tabulation, is used to present, in a summarized way, the frequencies in a population or sample, classified by two statistical variables. In its simplest form, when the variables only involve two different categories, it has the format presented in Table 1.

Table 1. Typical format for the 2x2 contingency table

	A	No-A	Total
B	a	b	a+b
no-B	c	d	c+d
Total	a+c	b+d	a+b+c+d

We could propose to the students different problems concerning this type of table, for example:

- Interpreting the information contained in the table.
- Providing a judgement about the existence of association between the variables.
- Computing and interpreting a coefficient to measure the strength of the association.

Concerning the first category, we observe that this work is relatively complex, because from the absolute frequency contained in a cell, for example cell a, three different types of relative frequencies could be deduced; the unconditional relative frequency *[a/(a+b+c+d)]* and the conditional relative frequencies given the rows *[a/(a+b)]* or given the columns *[a/(a+c)]* of the table.

There is very little research into either students' interpretation of relative frequencies in the contingency table or of the correlation coefficient. Nevertheless, research on human Judgement of association has been the object of great interest within psychology. Psychologists have used 2x2 contingency tables, as in the following example.

> We are interested in assessing if a certain drug produces digestive troubles in old people. For a sufficient period, 25 old people have been followed, and these results have been obtained:

	Digestive troubles	No troubles	Total
Taking the drug	9	3	17
Not taking the drug	7	1	3
	16	9	25

> Using the data in the table, reason if for this sample, having digestive troubles depends or not on taking the drug.

If we analyse in detail the proposed task, it can be observed that, although apparently simple, it is a complex problem for the student. Its difficulty depends on certain data contained in the statement. In the example, there is an inverse association, because a smaller proportion of those taking the drug have digestive troubles. A direct association, an inverse association or no association ,s possible as a function of the values given in the four cells of the table.

Another fact that increases the difficulty of the task is the different number of old people in the two groups, that is to say, the marginal frequency of the independent variable (taking or not taking the drug) is not the same for its different values All these aspects which may influence the difficulty of the problem are called task variables of the problem (Kilpatrick [44]). Other possible task variables in this case are the strength of the relationship, and the concordance between the empirical association in the table and the student's beliefs about the expected association

The study of reasoning about statistical association started with Piaget and Inhelder [45], who considered the understanding of the idea of association as the last step in developing the idea of probability. So the evolutionary development of the concepts of association and probability are related, and understanding the idea of association has as prerequisites the concepts of proportionality and probability. For this reason, Inhelder and Piaget [46] only studied reasoning about association with children in their formal operation stages IIIa and IIIb. They found that stage IIIa children only analyse the relation between the favourable positive cases (cell *a* in the table 1) and the total number of cases. In our example concerning a drug, stage IIIa children would deduce, incorrectly, the existence of a positive association between the variables, because of the greater number of people that fulfil the two conditions Drug taken and Digestive troubles (i.e. cell *a)* when compared with the frequencies in the other three cells.

Adolescents at level IIIa only compare the cells two by two. Once they admit that the cases in cell *d* (absence-absence) are favourable to the existence of association they do not compute the relation between the cases confirming the association (a + *d) an*d the other cases *(b+c).* This is only produced at 15 years of age (stage IIlb)according to Piaget and Inhelder.

The same conclusions have been obtained in other research using adult students for example Smedslund [47]. Most adult students base their Judgement only using cell *a* or comparing *a* with *h,* that is, they use only the conditional distribution of having or not having digestive troubles in those who take the drug. This strategy would lead, m our case, to incorrectly conclude the existence of a direct relationship between the two variables, since in the group of people taking the drug there are more with digestive troubles than without.

The difficulty of this type of task is shown by the fact that, as Jenkins and Ward [48] pointed out, even the strategy of comparing the diagonals in the table considered as correct by Piaget and Inhelder, is only valid in the case of tables having equal marginal frequencies for the independent variable. Our example illustrates this difficulty. For the general case, Jenkins and Ward have proposed as the correct strategy examining the difference between the two conditional probabilities of *A* occurring when *B* is true and of *A* occurring when *B* is false:

$$\delta = a/(a+b) - c/(c+d)$$

so, in our case, comparing the ratios 9/17 with 7/8 (conditional frequencies) would be needed.

In addition to the difficulty of this topic, Chapman and Chapman [49] showed that there are common expectations and beliefs about the relationship between the variables that cause the impression of empirical contingencies. This phenomenon has been described as 'illusory correlation', because people maintain their beliefs and overestimate the association when they believe that a causation exists between the two variables (Jennings *et al.* [50]). Finally, as Scholz [15] has described, posterior studies have shown that for the same association problem structure, people adopt different strategies and even the same person may use different strategies in different contexts.

5.2. Lin*ear regression and correlation*

As we have pointed out, psychological research about association has been linked to the problem of taking decisions in uncertain environments (Scholz [51]). The interest of these studies is to analyse the way in which human beings take decisions, because of the profound implications in areas such as medical diagnosis, economy, and law. Teaching is not the aim, in general. This explains the little interest in this research for the understanding of concepts linked to linear regression and correlation, which are not so often employed in the area of decision taking.

The study of the relationship between two quantitative variables includes two different problems: correlation and regression. In the study of correlation the two variables play a symmetrical role. The aim is to determine whether the two variables co-vary or not, whether this occurs in the same direction (positive correlation) or in the opposite direction (negative correlation) and to measure the strength of this association.

If a relatively strong association is perceived, the problem arises of finding a function y=f(x) (regression line) that could be used approximately to predict the values of y, from the values of *x.* Since the dependence is not functional, this prediction refers to the mean value of y for given *x.* As we know, this problem has no unique solution, and a sequence of personal choices is needed:

- To select the family of functions from which the line of regression will be selected (linear, exponential, . . .). The decision would be based on previous knowledge about the phenomenon, as well as from analysis of the shape of the scatter plot.
- To choose the decision criteria. When we have decided, for example, to take a linear

approximation, we could use as a criterion the least square approach or to use the Tukey line. The student's understanding of the chosen criterion would allow him to correctly interpret the line of regression and the relation of the line with the data (goodness of fit).

Finally, when the line of regression has been determined it is still possible to commit errors in its interpretation or in its employment to make predictions. Campbell [33] has pointed out that some students consider that extrapolation is impossible.

5.3. *Experimental design*

Rubin and Rosebery [43] have planned and observed a teaching experiment in order to study teachers' difficulties with stochastic ideas. They reported that both the students and their teacher had misinterpreted some of the basic ideas that underlie experimental design.

One of the lessons in the aforementioned teaching experience used a basketball shooting activity whose purpose was to determine the effect of distance (distances were varied regularly from 1 to 9 feet) and positional angle (for angles of 0, 45 and 90 degrees). Each student shot one ball from each combination of distance and angle. The aim of the lesson was to explore the separate effects and the interaction of these two variables.

The observation of the discussion between the teacher and the students, concerning the idea of independent, dependent and extraneous variables in the- shooting experiment, has shown the confusion between these concepts. Some students suggested as possible independent variables individual characteristics, such as the height or ability of each student. Given the height of the basket which was fixed during the experiment was considered as an independent variable by some of the students.

Some other students suggested that lighting could differ for different combinations of angle and distance, so the students and the teacher were left with the impression that the presence of such influences made conclusions about angle and distance impossible.

Finally, Rubin and Rosebery have remarked on the difficulty of distinguishing between characteristics of individuals that have no effect on the outcome of the experiment from other variables that may have an influence. Understanding the role of random allocation in disregarding individual differences was also found to be a cause of difficulty.

6. Inference

6.1. *Sampling*

The key idea in inference is that a sample provides 'some' information about the population and in this way increases our knowledge about the population. It does not furnish complete information, but an approximation. As Moses [52] points out, 'one can think of statistical inference as a collection of methods for learning from experience'. Rubin *et al.* [53] noted that, in practice, this implies the possibility of finding the values of the parameters of interest in the population, that is to say, obtaining confidence intervals for these parameters.

Understanding this fundamental idea supposes a balance between two apparently antagonistic ideas: sample representativeness and variability. The first one suggests that, when the process of selecting the sample has been performed properly, the sample will often have characteristics similar to those of the population. The second one, shows us that not all the samples are identical; so not all of them can resemble the population from which they have been selected. Finding the balance point between total information and null information about the population is complex, because of the dependence on three factors: population variability, sample size and level of confidence.

Research into errors concerning the idea of sampling has been important in psychology in the context of decision taking, notably the work of Kahneman and Tversky. A summary of their work can be found in Kahneman *et al.* [54] who attribute these errors to the use of certain judgemental heuristics in statistically naïve people. The term heuristics is employed in psychology, artificial intelligence and problem solving (Groner *et al.* [55]). Although there is no general consensus about the meaning of the term heuristics, it is normally used to refer to cognitive processes or mechanisms that are employed to reduce the complexity of a problem, during the solution process.

Kahneman and Tversky have defined three fundamental heuristics in probabilistic judgements: representativeness, availability, and adjustment and anchoring. Also they have studied the associated biases and the theoretical and practical implications.

According to the representativeness heuristic, the likelihood for samples are estimated in accordance with how well they represent some aspects of the parent population. In consequence, there is insensitivity to the sample size and over-confidence in small samples. This phenomenon is known as 'belief in the law of small numbers'. For, example, let us consider the following problem:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of a year each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?
> *(a)* The larger hospital *(b)* the smaller hospital (f) about the same.

Many people believe that (c) must be the correct answer, due to the fact that they consider that in both hospitals the proportion of boys is the same, and believe this is the only fact of importance in order to determine the probability of the required events. They do not pay attention to the sample size, although probability theory shows us that there is more fluctuation in the values of the proportion in small samples than in large samples.

According to Kahneman and Tversky, this over-confidence in small samples has serious consequences in the application of statistics, especially in research. The 'believer in the law of small numbers' underestimates the size of confidence intervals, overestimates the significance in tests of hypothesis, and is over-confident in obtaining the same results in figure replications of the experiment.

Another consequence of applying the representativeness heuristic is the 'gambler's fallacy'. For example, many people believe that after a run of heads, tails is more likely to come up.

When comparing the biases in probabilistic judgements with the expert's conception of random sampling, Pollatsek *et al.* [56] noticed that experts widely use 'urn-drawing' as the model for random sampling. In this model, random sampling is viewed as isomorphic to the process of drawing a number of balls from an urn, replacing them and then drawing again. Naive subjects might have no mechanistic way of thinking about this process or might have an erroneous process model of random samples, from which representativeness of even small samples would follow. Since these people may never have drawn a ball out of an urn, this model is itself theoretical and not genuinely practical. As Steinbring [57] has noted, the idea of independence has also a theoretical character and it is difficult to be sure of its applicability in a practical context. For this reason independence provides an example of the distance between the conceptual understanding of a concept and the ability to apply this concept in problem solving (Heitele [58]).

Another problem related to sampling is the different levels of concretion of the same statistical concept in descriptive statistics and in inference (Schuyten [40]). In descriptive statistics the *case* is the unit of analysis (e.g. a person, an object) and we compute the mean x of a sample of such units. In inference, we are interested in obtaining information about the theoretical mean or expectation $E(X)$ in the population from which the sample has been taken. We consider the particular sample as a unit from another different population—the set of all the possible samples with the given size that could be obtained from the parent population. We have changed the analysis unit; it is now the *sample* and we speak about the mean of the sample as a random variable. So, we study the distribution of the mean in the sampling process and the expectation *E(X)* of this random variable. It is necessary to distinguish between the theoretical mean in the population, which is an unknown constant, the particular mean obtained in our sample, the possible values of the different means that would be obtained with the different possible samples of size *n* in the population (a random variable), and the theoretical mean of this random variable, which coincides with the population mean for a random sampling process. This is very difficult conceptually.

6.2. *Test of hypothesis*

In some countries, one of the topics introduced in the last years of secondary education is the test of hypothesis. The scope of application for hypothesis testing is wide indeed, but, as Brewer [59] has commented, this area of inference is probably the most misunderstood, confused and abused of all statistical topics. For this reason, and because this topic can be considered complex for the student, it is a widespread and exciting area for mathematical education researchers to explore.

The term test of hypothesis could be applied to a great number of statistical procedures: test of differences between means, analysis of variance, non-parametric procedures, multivariate tests .... All these procedures share a common nucleus: a set of basic concepts (null hypothesis and alternative hypothesis, level of significance, power function, etc.), and some general procedures which are modified for particular cases. The correct application of these procedures involves many kinds of choices, including: the sample size, the level of significance a and the appropriate statistic. In particular, Peskun [60] has pointed out students' difficulties with the following aspects:

a. the determination of the null hypothesis H_0 and the alternative hypothesis H_1;
b. the distinction between type I and type II errors;
c. understanding the purpose, use and availability of operating characteristic curves or power curves; and
d. understanding the terminology used in stating the decision.

Similar difficulties have been described by Reeves and Brewer [61], Johnson [62] and Shoesmith [63].

One of the key aspects in the correct application of a test of hypothesis is understanding the concept of level of significance, which is defined as the 'probability of rejecting the null hypothesis, when it is true'. This definition is expressed in the following identity:

$$(1)\ \alpha = P(\text{Rejecting } H_0 \mid H_0 \text{ true})$$

Falk [64] has pointed out the interchange of the conditional event and the condition a frequent error in this definition and the mistaken interpretation of the level significance as 'the probability that the null hypothesis is true, once the decision to reject it has been taken', that is:

$$(2)\ P(Ho \text{ true we have rejected } Ho)$$

She suggests, as one of the possible causes of this error, the language used in the definition of the level of significance, which is also referred to as 'the probability of committing a type I error'. In this expression, it is not explicitly established that we are dealing with a conditional probability, so, it is supposed by the student that it is possible to define a 'conditional event'. In consequence, the distinction between the two conditional probabilities (1) and (2) is not being made.

The correct definition of the level of significance a would be expressed in the following statement:

(i) A significance level of 5% means that, on average 5 times out of every 100 times that the null hypothesis is true, we will reject it.

Research by Birnbaum [65], and others, has shown that some students consider correct the following incorrect definition of a:

(ii) A significance level of 5",, means that, on average, 5 times out of every 100 times we reject the null hypothesis, we will be wrong.

Statements (i) and (ii) were presented to university students by Vallecillos [66] who asked them whether each was true or false, and then she analysed the students' reasoning. She also analysed the concept of level of significance and its relation with other concepts that intervene in a test of hypothesis. She distinguished four different aspects in the understanding of this concept and she identified misconceptions related to each one of these aspects:

(a) *The test of hypothesis as a decision problem.* The test of hypothesis can be considered as a decision problem between two exclusive and complementary hypotheses, with the possible consequences of committing or not committing one of the two types of error, that are disjoint but not complementary events. Nevertheless, some students consider the type I and 11 errors to be complementary events, so that in their view the probability of committing one or the other of these errors would be unity.

(b) Probabilities of error and relation between them. The two types of error have probabilities α (type I) and β (type II). It is necessary to understand the conditional probabilities that occur in the definition of α and β*,* of the dependence *of* β in terms of the unknown value of the parameter, and the relation between α and β. Besides the error pointed out by Falk [64], mentioned above, other misleading interpretations of the level of significance have been found: suppressing the condition in the conditional probability which used to define a; interpreting a as the probability of any kind or error (both types I and II) in the decision taken.

(c) l*evel of significance as the risk of the decision maker.* The values α and β determine the risks that the decision maker is willing to assume m his decision and, along with the hypothesis, will serve in establishing the decision criterion. Some students believe that changing the level of significance does not change the risk of Type II error.

(d) Interpretation of a statistically significant result. Obtaining a statistically significant result leads to the rejection of the null hypothesis. However, obtaining a statistically significant result does not necessarily imply any practical significance, but some students confuse these. Also, some students believe that a statistically significant result corroborates the null hypothesis rather than rejecting it.

White [67] has mentioned the case of misinterpretation of a statistically non-significant result. He also considered a different aspect concerning peoples understanding of statistical significance which is the problem of multiple comparison that occurs when many significance tests are applied to one set of data. For example an epidemiological investigation might measure 150 variables on each person in a group of healthy people and also measure the same variables on each person in a group of people with a certain disease. Now, if we choose a significance level as 0.05 then since 150 x 0.05 =7.5, 7.5 'statistically significant results' are to be expected, on average, even if the disease has no relationship at all with any of the variables studied (Moses [52]).

Finally we refer the reader to the book edited by Morrison and Henkel [68] which documents the considerable amount of indiscriminate use of significance tests in research, which is a clear indication of the difficulty, misinterpretation and misuse of significance tests.

7. Final remarks

In a survey of research literature Garfield and Ahlgren [141 enunciated reasons for some of the difficulties which have arisen in teaching statistical concepts at the college level:

- Many of the statistical concepts, such as probability, correlation, per cent, require proportional reasoning, known to be a difficult concept in mathematics. . .
- The existence of false intuitions which students bring to the statistics classroom Many of these intuitions are well known in the case of probability (Piaget and Inhelder [45]), (Fischbein [69]), but very few have been studied for statistical concepts.
- Sometimes students have developed a distaste for statistics, because they have been exposed to the study of probability and statistics in a highly abstract and formal way.

There are two more reasons that possibly influence the difficulty of the subject. First probability and statistics have developed recently. Although there is, at present, a well established axiomatic system for probability, due to the work of Kolmogorov the controversy about the meaning of the term 'probability- has not been Spoiled and there are different traditions: empiricists, subjectivists, etc. (Fine [70]). 1 his controversy shows itself in statistical inference-, there is a debate as to whether it is possible or not to attribute a probability to a hypothesis in both the classical and Bayesian approaches (Rivadulla [71]). Numerous research works have shown that, during the learning process, the student must frequently overcome the same epistemological difficulties that have been found in the historical development of knowledge.

Second, many statistical concepts have arisen outside mathematics. Statistics has been from its very beginning an interdisciplinary science. The most important periods in its development have been marked by contributions from different fields in which there was the necessity of solving specific problems. In the classroom, the concepts are presented in isolation from their original applications which contributed to their global meaning (Steinbring [72]). For example, the concept of mean has special different meanings when applied to the centre of gravity, to life expectation, or to an index number.

In summary, as Green [73] has pointed out: 'Statistical concepts provide a fascinating area to explore. What the statistician regards as straightforward and obvious (terms such as average, variability, distribution, correlation, bias, randomness, . . .) are the distilled wisdom of several generations of the ablest minds. It is too much to expect that there will not be a struggle to pass on this inheritance'.

References

[1] HOLMES, P. *et al.,* 1980, *Statistics in Your World* (Slough: Foulsham Educational).
[2] LANDEWEHR, J**.,** and WATKINS, A. E., 1986, *Exploring data* (Palo Alto: Dale Seymour).
[3] LANDEWEHR, J., WATKINS, A. E., and SWIFT, J., 1987, *Exploring Surveys: Information from Samples* (Palo Alto: Dale Seymour).
[4] GNANADESIKAN, N., SCHEAFFER, R. L., and SWIFT, J., 1987, *The Art and Technique of Simulation* (Palo Alto: Dale Seymour).
[5] GODINO, J. D., BATANERO, M. C., and CAÑIZARES, M. J., 1987, *Azar y Probabilidad Fundamentos didacticos y propuestas curriculares* (Madrid: Sintesis).

[6] GREY, D. R., HOLMES, P., BARNETT, V., and CONSTABLE, G. M., (eds), 1982, *Proceedings of the First International Conference on Teaching Statistics* (Sheffield: University ot Sheffield).
[7] DAVIDSON, R., and SWIFT, J., (eds), 1987, *Proceedings of the Second International Conference on Teaching Statistics* (Voorburg, The Netherlands: International Statistical Institute).
[8] VERE-JONES, D. (ed.), 1991, *Proceedings of the Third International Conference on Teaching Statistics* (University of Otago, Dunedin, (Voorburg, The Netherlands: International Statistical Institute).
[9] NCTM, 1989, *Curriculum and Evaluation Standards/or School Mathematics* (Reston, VA: NCTM).
[10] DES, 1991, *Mathematics in the National Curriculum* (London: Department of Education and Science and the Welsh Office).
[11] MEC, 1988a., *Diseño curricular base para la enseñanza primaria.* (Madrid: Ministerio de Educacion y Ciencia).
[12] MEC, 1988b, *Diseño curricular base para la enseñanza secundaria obligatoria* (Madrid Ministerio de Educacion y Clencia).
[13] HAWKINS, A., and KAPADIA, R., 1984, Children's conceptions of probability- a psychological and pedagogical review. *Educ. Studies in Math.,* 15, 349 377.
[14] GARFIELI), J., and ALHGREN, A., 1988, Difficulties in learning basic concepts in probability and statistics: implications for research. *J. Res. Math. Educ.,* 19(1), 44-63.
[15] SCHOI.Z, R., 1 991, Psychological research on the probability concept and its acquisition. In R. Kapadia (ed.) *Chance Encounters: Probability in Education* (Amsterdam: Reidel), pp. 213-249.
[16] SHAUGHNESSY, J. M., 1992, Research in probability and statistics: reflections and directions. In D. A. Grouws (ed.) *Handbook on Research in Mathematics Education* (New York: McMillan), pp. 465-94.
[17] HAWKINS, A., JOLLIFFE, F., and GI.ICKMAN, L... 1992, *Teaching Statistical Concepts* (London: Longman).
[18] CENTENO, J., 1988, *Numeros decimates ¿por que? ¿.para que?* (Madrid: Sintesis).
[19] RADATZ, H. C., 1980, Students' errors in mathematical learning: a survey. *For the Learning of Mathematics,* 1(1), 16-20.
[20] BORASSI, R., 1987, Exploring mathematics through the analysis of errors. *For the Learning of Mathematics, 7* (3), 2-8.
[21] AUSUBEL, D. I., NOVAK, J. D., and HANESIAN, H., 1983, *Psicología educativa. Un punto de vista cognoscitivo.* (Mexico: Trillas).
[22] CONFREY, J., 1990, A review of the research in students conceptions in mathematics, science and programming. *Rev. Res. Educ.,* 16, 3-35.
[23] BROUSSEAU, G., 1983, Les obstacles epistémologiques et les problémes en mathématiques. *Recherches en Didactique des Mathématiques,* 4 (2), 164-198.
[24] SIERPINSKA, A., 1991, Some remarks on understanding in mathematics. *For the Learning of Mathematics,* 10 (3), 24-36.
[25] CURCIO, *F.* R., 1987, Comprehension of mathematical relationships expressed in graphs. *J. Res. Math. Educ.,* 18 (5), 382-393.
[26] PEREIRA-MENDOZA, L., and MELLOR, J., 1991, Students' concepts of bar graphs-some preliminary findings. In D. Vere-Jones (ed.) *Proceedings of the Third International Conference on Teaching Statistics,* (Voorburg, The Netherlands: International Statistical Institute), pp. 150 157.
[27] LI, K. Y., and SHI-:N, S. M., 1992, Students' weaknesses in statistical projects. *Teaching Statistics,* 14(1), 2-8.
[28] POLLATSEK, A. LIMA, S., and WELL, A. D., 1981, Concept or computation: students' understanding of the mean. *Educ. Studies in Math.,* 12, 191-204.

[29] SKEMP, R., 1978, Relational understanding and instrumental understanding. *Arithmetic Teacher,* November, 9-15.

[30] HIEBERT, J, and LEFEBRE, P., 1986, Conceptual and procedural knowledge in mathematics: an introductory analysis. In J. Hiebert (ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics* (Hillsdale, NJ: Lawrence Eribaum), pp. 1-27.

[31] MEVARECH, Z. R., 1983, A deep structure model of students' statistical misconceptions. *Educ. Studies in Math.,* 14, 415-429.

[32] STRAUSS, S., and BICHLER, E., 1988, The development of children's concepts of the arithmetic average. *J. Res. Math. Educ.,* 19 (1), 64-80.

[33] CAMPBELL, S. K., 1974, *Flaws and Fallacies in Statistical Thinking* (New Jersey: Prentice-Hall).

[34] RUSSELL, S. J., and MOKROS, J. R., 1991, What's typical?: children's ideas about average. In D. Vere-Jones (ed.) *Proceedings of the Third International Conference on Teaching Statistics* (Voorburg, The Netherlands: International Statistical Institute), pp. 307-313.

[35] GOODCHILD, S., 1988. School pupils' understanding of average. *Teaching Statistics,* 10 (3), 77-81.

[36] LOVIE, P., and LOVIE, A. A., 1976. Teaching intuitive Statistics: estimating means and variances. *Int. J. Math. Educ. Sci. Technol.,* 7(1), 29-39.

[37] LOOSEN, F., LIOEN, M., and LACANTE, M., 1985, The standard deviation: some drawbacks of an intuitive approach. *Teaching Statistics,* 7(1), 2-5.

[38] HUCK, S., CROSS, and CI.ARK, S. B., 1986, Overcoming misconceptions about z-scores. *Teaching Statistics,* 8 (2), 38—40.

[39] TUKEY, J. W., 1977, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley).

[40] SCHUYTEN, G., 1991, Statistical thinking in psychology and education. In D. Vere-Jones (ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Voorburg, The Netherlands: International Statistical Institute), pp. 486-490.

[41] ESTEPA CASTRO, A., 1990, Enseñanza de la Estadistica basada en el uso de ordenadores: Un Estudio exploratorio. Memoria de Tercer Ciclo. (Universidad de Granada: Departamento de Didactica de la Matematica).

[42] BARR, G. V., 1980, Some student ideas on the median and the mode. *Teaching Statistics, 2* (2), 38 41.

[43] RUBIN, A., and ROSEBERY, A. S., 1990, Teachers' misunderstandings in statistical reasoning; evidence from a field test of innovative materials. In A. Hawkins (ed.), *Training Teachers to Teach Statistics* (Voorburg, The Netherlands: ISI), 72-89.

[44] KILPATRICK, J., 1978, Variables and methodologies in research on problem solving. In L. L. Hartfield and D. A. Bradbard (eds), *Mathematical Problem Solving. Papers from a Research Workshop* (Columbus, Ohio: ERIC/SMEAC).

[45] PIAGET, J., and INHELDER, B., 1951, *La génese de I'idée de hasard chez I'enfant* (Paris: Presses Umversitaires de France).

[46] INHELDER, B., and PIAGET, J., 1955, *De la logique de l'enfant a la logique de l'adolescent* (Paris: P.U.K.).

[47] SMEDSI.UND, J., 1963, The concept of correlation in adults. *Scand. J. Psvchol.,* 4, 165-173.

[48] JENKINS, H. M., and WARD, W. C., 1965, Judgement of contingency between responses and outcomes. *Psychological Monographs,* 79.

[49] CHAPMAN, I.. J., and CHAPMAN, J. P., 1967, Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *J. Abnormal Psychol.,* 74, 271-280.

[50] JENNINGS, D. L., AMABII.E, T. M., and Ross, L., 1982, Informal covariation assessment: data based versus theory based judgements. In D. Kahneman; P. Slovic and A. Tversky (eds.), *Judgement under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press), pp. 211-30.

[51] SCHOLZ, R., 1987, *Decision Making Under Uncertainty (Amsterdam:* North Holland).
[52] MOSES, L. E., 1992, The reasoning of statistical inference. In D. C. Hoaglin, and D. S. Moore (eds), *Perspectives on Contemporary Statistics* (U.S.A.: Mathematical Association of America), pp. 107-121.
[53] RUBIN, A., BRUCE, B., and TENNEY, Y., 1991, Learning about sampling: trouble at the core of statistics. In D. Vere-Jones (ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Voorburg, The Netherlands: International Statistical Institute), pp. 314 319.
[54] KAHNEMAN, D., SLOVIC, P., and TVERSKY, A., 1982, *Judgment Under Uncertainly: Heuristics and Biases* (New York: Cambridge University Press).
[55] GRONER, R., GRONER, M., and BISCHOF, W. F., 1983, The role of heuristics in models of decision. In R. W. Scholz (ed.): *Decision Making Under Uncertainty* (Amsterdam: North Holland), pp. 87 108.
[56] POLLATSEK, A., KONOLD, C. K., WEI.I., A. D., and LIMA, S., 1991, Beliefs underlying random sampling. *Memory and Cognition,* 12, 395—401.
[57] STEINBRING;, H., 1986, L'independence stochastique. *Recherches en Didactiuie des Mathematiques, 7* (3), 99-1 18.
[58] HEITELE, D., 1975, An epistemological view on fundamental stochastic ideas. *Educ. Studies in Math.,* 6, 187-205.
[59] BREWER, J. K., 1986, Behavioral statistics textbooks: source of myths and misconceptions? In R. Davidson and J. Swift (eds), *Proceedings of the Second International Conference on Teaching Statistics* (Voorburg, The Netherlands: International Statist ical Institute), pp. 127-131.
[60] PESKUN, P. H., 1987, Constructing symmetric tests of hypotheses. *Teaching Statistics, 9* (1), 19 23.
[61] REEVES, C. A., and BREWER, J. K., 1980, Hypothesis testing and proof by contradiction: an analogy. *Teaching Statistics, 2* (2), 57-59.
[62] JOHNSON, L. M., 1981, Teaching statistics as a six step process. *Teaching Statistics,* 3(2) 47-49.
[63] SHOESMITH, E., 1983, Simple power curve constructions. *Teaching Statistics,* 5 (3), 78-83.
[64] FALK, R., 1986, Misconceptions of statistical significance. *J. Structural [.earning* 9, 83 96.
[65] BIRNBAUM, I., 1982, Interpreting statistical significance. *Teaching Statistics,* 4(1), 24-27.
[66] VALLECILLOS, A., 1992, Nivel de significación en un contraste estadistico de hipotesis. Un estudio teorico-experimental de errores en estudiantes universitarios. Memoria de Tercer Ciclo. (Universidad de Granada: Departamento de Didactica de la Matematica).
[67] WHITE L., 1980, Avoiding errors in educational research In R. J. Shumway (ed.), Research in Mathematics Education (Reston, VA. N.C.T.M.) pp. 49- 65.
[68] MORRISON, D. E., and HENKEL, R. E. (eds), 1970, *The Significance Test Controversy* (Chicao: Aldine).
[69] FISCHBEIN, E., *The Intuitive Sources of Probabilistic Thinking in Chidren* (Dordrecht: Reidel).
[70] FINE, T. L., 1973. *Theories of Probability. An Examination of Foundations.* (New York: Academic Press).
[71] RIVADULLA, A., 1991, *Probabihdad e Inferencia científica* (Barcelona: Anthropos).

[72] STEINBRING, H., 1990, The nature of stochastical knowledge and the traditional mathematics curriculum. some experience with in-service training and developing materials. In A. Hawkins (ed.), *Training Teachers to Teach Statitics* (Voorburg, ISI).

[73] GREEN, D. R., 1992, Data analysis: what research do we need? Paper presented at the *Eight ISl Round Table Conference on Teaching Statistics,* Lennoxville, 1992.