Controversies around the Role of Statistical Tests in Experimental Research

Carmen Batanero, University of Granada

Mathematical Thinking and Learning, 2(1-2), 75-98.

Abstract

In spite of the widespread use of significance testing in empirical research, its interpretation and researchers' excessive confidence in its results have been criticized for years. In this paper, we first describe the logic of statistical testing in the Fisher and Neyman-Pearson approaches, review some common misinterpretations of basic concepts behind statistical tests, and analyse the philosophical and psychological issues that can contribute to these misinterpretations. We then revisit some frequent criticisms against statistical tests and conclude that most of them refer not to the tests themselves, but to the misuse of tests on the part of researchers. We agree with Levin (1998a) that statistical tests should be transformed into a more intelligent process that helps researchers in their work, and finally suggest possible ways in which statistical education might contribute to the better understanding and application of statistical inference.

Empirical sciences, in general, and particularly psychology and education, rely heavily on establishing the existence of effects using the statistical analysis of data. Statistical inference dates back almost 300 years, but statistical tests were systematized through the works of Fisher, Neyman, and Pearson and today most researchers implicitly use a mixture of the logics suggested by these authors. However, since the logic of statistical inference is difficult to grasp, its use and interpretation are not always adequate and have been criticized for nearly 50 years. For example, Yates (1951) suggested that scientists were paying too much attention to the tests' results and were forgetting the estimates of the magnitudes of the effects they were investigating. An early extensive review of criticisms against significance testing can be found in Morrison and Henkel (1970).

More recently, a comprehensive summary of these debates, and alternative approaches suggested, has been provided by Harlow, Mulaik, and Steiger (1997).

This controversy has recently increased within professional organizations (Ellerton, 1996; Levin, 1998a, 1998b; Levin & Robinson, 1999; Menon, 1993; Robinson & Levin, 1997; Thompson, 1996; Wilkinson, 1999) which are suggesting important shifts in their editorial policies regarding the use of statistical significance testing. For example, within the American Educational Research Association, Thompson (1996) recommended better use of statistical language in reported research, emphasizing effect-size interpretation and evaluating result replicability. Several organizations have established special committees to study the problem, and these committees have recommended that statistical tests should not be abandoned but rather supplemented by other statistical analyses (Levin, 1998b). For example, the American Psychological Association (APA), in a 1994 publication manual, noted that significance testing does not reflect the importance or

magnitude of an effect and encouraged researchers to provide effect-size information (American Psychological Association, 1994, p. 18). Subsequently, the Task Force on Statistical Inference established by the APA published an article to initiate discussion in the field prior to revising the APA publication manual (Wilkinson, 1999). Following a decision of the task force, this article covers methodological issues more general that null hypothesis significance testing. Among many recommendations are that exact p-values should be reported, together with effect-size estimates combined with interval estimates.

In spite of all the criticisms levelled, researchers persist in relying on statistical significance despite the arguments that statistical tests are not adequate to justify scientific knowledge. Some explanations for this persistence include inertia, conceptual confusion, lack of better alternative tools, and psychological mechanisms such as invalid generalization from deductive logic to inference under uncertainty (Falk & Greenbaum, 1995). In this paper we analyse these problems and finally suggest possible ways in which statistical education might contribute to the better understanding and application of statistical inference.

The Logic of Statistical Tests: An Example

In this section we first present a typical situation in which a researcher resorts to statistics to support a given hypothesis concerning her field of study. We then summarize the steps and logic of statistical testing. The example is used throughout the paper to contextualize the discussion.

> Example 1: According to some learning theories, representations contribute to the construction of meaning of mathematical objects, so that a richer context which facilitates change of representations would favor learning. A researcher who accepts this theory has good reasons to expect that using computers will reinforce the learning of statistics, because computers provide powerful tools and systems of representations for statistical concepts. To assess her conjecture, suppose the researcher selects at random a sample of 80 students among all the students entering a given university in a given year. Then, she randomly assigns the 80 students to equal groups and finds that the two groups are equivalent in their previous knowledge of statistics. An experiment is organized where the same lecturer, using the same materials, teaches an introductory statistics course to both groups for a semester. Group C (the control group) has no access to computers, while the teaching to Group E (the experimental group) is based on intensive use of computers.
>
> At the end of the period, the same test is given to both groups. If the learning with the two teaching methods is equally effective, there should be no difference between μ_e, the mean test score of the theoretical population of students taught with the help of computers and μ_c, the mean test score of the theoretical population of students with no access to computers. Groups E and C are considered to be representative samples from these populations, so if there is no differential effectiveness in learning with the two teaching methods, the difference between the two groups' mean scores will be about zero. If the researcher finds a positive difference between the two groups' mean scores, can she deduce that her conjecture is supported? It is important to remark that the researcher's interest goes beyond the two particular samples (groups C and E); she wants to assess the effect of computers on learning in the general population, and through this, to find empirical support for her hypothesis about the effect of context on learning.

The above example illustrates a prototypical situation where statistical hypothesis testing can be used to determine whether the experimental data (the students' scores in groups E and C) support

a *substantive hypothesis* (learning is influenced by the systems of representations available) or not. As we cannot directly test the substantive hypothesis, because it refers to theoretical entities, we instead organize an experiment to generate data to test a *derived research hypothesis* - that computers reinforce the learning of statistics. However, we still cannot directly confirm the research hypothesis, because learning is an unobservable construct that can not be directly assessed.

We then choose an instrument (the test) that is directly related to learning and produces some observable outcomes (the students' responses to the test). If the research hypothesis is true, we expect the scores to be higher in students taught with the help of computers than in students with no access to computers (*experimental hypothesis*). In any case, because of the multiple factors that affect learning in addition to computers, some students taught without computers will perform better than others taught with computers. Accordingly, we need a procedure that compares the whole distributions of scores in the two populations of students.

This comparison is usually carried out by considering mean test scores for the two theoretical population and speculating about their difference. If the experimental hypothesis is true, we will expect that this difference is positive (*statistical alternative hypothesis*), although we cannot state a precise value for the difference. Therefore, we cannot directly work with the statistical alternative hypothesis, and we reason, instead, as if the two populations have the same performance, i.e. we assume that the *null hypothesis* (that the difference in mean scores in the populations is zero) is true. (In the example, in fact, we are dealing with a one-tailed test, because we have specified the direction of the departure from the null hypothesis. In this case, the null hypothesis is actually $\mu_e \quad \mu_c$ (the complement to the alternative hypothesis). From the mathematical point of view, however, we only need to consider the case where $\mu_e=\mu_c$, to compute the critical value for a one-tailed test, since whenever a result is significant for the hypothesis $\mu_e=\mu_c$, it will also be significant for the hypothesis $\mu_e < \mu_c$. Therefore, to simplify the exposition, we assume in the following that the null hypothesis is $\mu_e=\mu_c$).

To test this assumption, we compute a test statistic related to the parameter of interest from the data in our samples. Assuming that the null hypothesis is true determines the distribution of this statistic (a T distribution with 78 d.f.). This is used to compute the critical value and take a decision about whether or not we should accept our initial research hypothesis.

There are two different views of statistical tests: a) tests of significance that were introduced by Fisher and b) tests as rules for deciding between two hypotheses, which was the view of Neyman and Pearson. The difference does not lie in the calculations, but in the underlying reasoning. Following Moore (1995) we first describe the typical reasoning in a test of significance that would be adequate in the example 1. The following steps will typically be followed:

1. Describing the effect you are searching for in terms of one or several population parameters (the mean scores μ_e, in students taught with computers is higher than μ_c, the mean score in students with no access to computers). The effect we suspect is true is described by the alternative hypothesis: $H_1 \equiv \mu_e > \mu_c$.

2. Establishing the null hypothesis that this effect is not present: $H_0 \equiv \mu_e = \mu_c$, (there is no difference between the mean scores μ_e, in students taught with computers and μ_c, the mean score in students with no access to computers). The test of significance is designed to assess the strength of the evidence against the null hypothesis.

3. Computing a statistic from the sample results (calculated statistic). The distribution of this statistic is specified when we assume the null hypothesis to be true (in the case of the example, a T distribution with 78 df). Suppose that in Example 1 we obtain the following values for means in the

two samples: $\bar{x}_e = 115.10$; $\bar{x}_c = 101.78$ and that $s^2_e = 179.66$ and $s^2_c = 215.19$ are the unbiased estimators of the variances in the populations; the mean difference in scores in the two groups for these data is then $\bar{x}_e - \bar{x}_c = 13.32$, the pooled estimate of the variance $s^2 = 202.48$, and a value of t = 4.16 is obtained using the standard formula.

The question that significance testing is trying to answer is the following.

Suppose that the null hypothesis is true and that, on average, there is no difference in the mean scores of samples drawn from the two populations, is then the sample outcome t = 4.16 extremely large? Or could we easily get this value just because of chance fluctuation in sampling?

4. The probability of obtaining a t value as extreme or more extreme than the calculated t when the null hypothesis is true is called the p-value. For the example, the p-value is extremely small (less than .001). If the null hypothesis is true and the p-value is very small, the results are very unlikely and are called statistically significant. In this case, and if the data falls in the direction specified by the alternative hypothesis, we assume that our data provide evidence against the null hypothesis (this does not mean that we believe that the null hypothesis is impossible; it is only through a program of repeated experiments which replicate our results that a hypothesis would be accepted in the scientific community; science is built from cumulative findings).

5. Even when the statistical null hypothesis is true, some discrepancies between the mean scores in the experimental and control groups will be expected in Example 1, just because of the chance fluctuation in sampling. There is no fixed rule about how low the p-value should be for a result to be considered statistically significant. However, it has been conventional to adopt some fixed values with which we can compare the p-value to decide about statistical significance. This is the significance level α, or maximum p-value admissible to consider the data to be significant, which is used to compute the critical value. Suppose we take $\alpha = .05$ in example 1. The critical value is the maximum difference that will be expected in the two samples (groups E and C) with probability .05 (α), when the two populations have the same performance. This critical value is obtained from the theoretical distribution of the statistics in the case where the hypothesis null is true (T distribution with 78 d.f.).

Hypothesis Testing as a Decision Process

In example 1 we have used significance testing to assess the strength of the evidence against the null hypothesis. There are, however, other situations where inference is used to Make a decision between two possible actions.

> Example 2: Suppose a secondary school wants to assess the effectiveness of a new teaching method on their students' learning of statistics. The school randomly selects 80 students from all the students in the last school year and then randomly splits the 80 students into two groups of equal size. The same teacher teaches statistics to both groups for a semester. Group C (the control group) is taught with the usual method in that school, while the teaching in Group E (experimental group) is based on the new materials. At the end of the period, the same test is given to both groups (that were judged, using a pretest, to be initially comparable in ability). The school wants to change to the new system if the difference of mean scores in the two student populations is positive. Here the interest also goes beyond the two particular samples (groups E and C), as the method chosen would be applied to other students.

In Example 2, a statistical test would be used, with a different reasoning, as a procedure to take a decision. The following steps would be followed:

1. Establishing the null hypothesis $H_0 \equiv \mu_e = \mu_c$ and the alternative hypothesis: $H_1 \equiv \mu_e > \mu_c$, as in Example 1.
2. Computing the statistic from the sample data, as in Example 1.
3. Decision taking: The null hypothesis will either be rejected (and H_1 will be accepted) or H_0 will be not rejected.
4. The decision is made by comparing the p-value with the level of significance α, that is, by comparing the calculated t with the critical value. In Example 2 again (assuming the same numerical data) the calculated $t = 4.16$ is greater than the critical value $t_c = 1.665$, and therefore, the null hypothesis will be rejected.

It is worthwhile noting that rejecting the null hypothesis does not necessarily mean that it is false, as two kinds of error are possible in the decision taken as a result of the test. First, it is possible that there is no real difference in the test means in the students taught with the old and the new methods, and that, because of random variability in sampling we have obtained in our particular groups E and C a t-value that only happens with low probability. As we know, even when the probability of an event is low, this does not means that it is impossible. A Type I error happens if we reject a null hypothesis, when, in fact, it is true. The probability of a Type I error is numerically equal to the significance level, α.

On the other hand, if the result is not significant, that does not imply that the two populations perform equally well on the test. Even when the students who are taught with the new method perform, in general, better, we might fail in attaining a significant result in our particular samples, because the effect of the teaching is small or because there is too much variability in the data. A Type II error occurs when the researcher accepts the null hypothesis but the null hypothesis is, in fact, false. Since there are many different possibilities for the difference of means in the alternative hypothesis, the probability of Type II error, â is variable. We are usually interested in some particular values for this probability and compute this probability for the most unfavorable case.

The complement of â is called the power of the test. It is the probability of rejecting a null hypothesis when it is false and it is also variable, as it depends on the true value of the parameter (the difference of means in our example). It is important to emphasize the conditional nature of the probability of these two kinds of error, because it is in the interpretation of these conditional probabilities that most of the errors and misconceptions concerning statistical tests can be found.

Common Errors in Interpreting Significance Levels and p-values

Research into the understanding of inferential procedures has shown widespread misconceptions among both university students and scientists who use statistical inference in their daily work. These misconceptions refer mainly to the level of significance, α, which is defined as the probability of rejecting a null hypothesis, given that it is true. The most common misinterpretation of this concept consists of switching the two terms in the conditional probability, that is, interpreting the level of significance as the probability that the null hypothesis is true, once the decision to reject it has been taken. For example, Birnbaum (1982) reported that his students found the following definition reasonable: "A level of significance of 5% means that, on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong". Falk (1986) found that most of her students believed

that α was the probability of being wrong when rejecting the null hypothesis at a significance level α. Similar results were described in Pollard and Richardson (1987) in their study using researchers.

Vallecillos (1994) gave the following items to a sample of 436 University students from different backgrounds (statistics, medicine, psychology, engineering and business studies) who had previously been taught about statistical tests:

> Item 1: A level of significance of 5% means that, on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong (true /false). Justify your answer.
>
> Item 2: A level of significance of 5% means that, on average, 5 out of every 100 times the null hypothesis is true, we will reject it (true/ false). Justify your answer.

In item 2, a frequentist interpretation of the level of significance is presented (it is correct), while in item 1 the two events in the conditional probability have been exchanged (and it is incorrect). However, only 32% of the students in the research by Vallecillos (1994) gave the correct response to item 1 and 54% the correct response to item 2. From 135 students who justified their responses, 41% gave correct arguments for both items. A prevalent misconception in all the groups of students was exchanging the terms in the conditional probability, thereby judging item 1 to be true and item 2 to be false. Interviews with a subset of the students tested showed this belief in students who discriminated well between a conditional probability and its inverse (Vallecillos & Batanero, 1996). Other students did not distinguish between the two conditional probabilities, that is, considered both items to be correct.

That conditional probabilities with the terms switched are not, in general, equal is illustrated by the data in Table 1, relating to a school in which statistics is an optional subject. The probability of a randomly chosen pupil being a statistics student, given that pupil is a girl, and the probability of a randomly chosen statistics student being a girl, are different.

P (statistics student / girl) = 3/4; P (girl / statistics student) =3/8

Table 1. *Numbers of Girls and Boys in a Statistics Course*

	Girls	Boys	Total
Statistics	300	500	800
No Statistics	100	100	200
Total	400	600	1000

It is also important to remark that, even though we can fix the level of significance α, namely the probability of rejecting a null hypothesis (given that it is true) and we can compute the probability of obtaining a value of the test statistic lying beyond a particular value (given that the null hypothesis is true), the probability that the null hypothesis is true when we have rejected it, and the probability that the null hypothesis is true, given that we obtain a particular value of the test statistic, are not knowable.

The posterior probability of the null hypothesis, given a significant result, depends on the prior probability of the null hypothesis, as well as on the probabilities of having a significant result given the null and the alternative hypotheses. Unfortunately, these probabilities cannot be determined.

Moreover, a hypothesis is either true or false, and therefore it does not make much sense to compute its probability in a classical inferential paradigm (where we give a frequestist interpretation to objective probabilities). It is only within Bayesian inference that posterior probability of the hypotheses can be computed, although these are subjective probabilities. What we can do at best, and using Bayesian inference, is to revise our personal degree of belief in the hypothesis, in view of the result.

Other common misinterpretations concerning the significance level and the p-value are:

(a) Some people believe that the p-value is the probability that the result is due to chance. That this is a misconception can be deduced from the fact that, even when the null hypothesis is true (e.g., if there are no differences between the performance in the two student populations in Example 1), a significant result might be due to other factors, such as, for example, that the students in the experimental group worked harder than their counterparts to prepare for the test. Here we can see the relevance of experimental control to try to ensure that all the conditions (except the type of teaching) have been held constant in the two groups. The p-value is the probability of obtaining the particular result or one more extreme when the null hypothesis is true *and* there are no other possible factors influencing the result. What is rejected in a statistical test is the null hypothesis, and therefore we cannot infer the existence of a particular cause in an experiment from a significant result.

(b) Another common error is the belief in the conservation of the significance level value when successive tests are carried out on the same data set, which produces the problem of multiple comparisons. Sometimes many significance tests are applied to one body of data. The meaning of the definition of the significance level (see Item 2, above) is that if we carry out 100 comparisons on the same data set using in all of them a level of significance .05 it is expected that about 5 out of the 100 tests will be significant just by chance, even when the null hypothesis is true. This makes it difficult to interpret the results (Moses, 1992).

(c) The frequent use of .05 and .01 levels of significance is a matter of convention and is not justified by mathematical theory. When hypothesis testing is considered as a decision process (the view of Neyman and Pearson), the level of significance should be specified before the experiment is carried out and that choice determines the size of the critical and acceptance regions that will lead to the decision to reject the null hypothesis or not. Neyman and Pearson gave a frequentist interpretation to this probability: If the null hypothesis is true and the experiment is repeated many times with a .05 probability of Type I error, we will reject the null hypothesis 5% of the times then it is true.

In his book “Design of Experiments”, Fisher (1935) suggested selecting a significance level of 5% as a convention to recognize significant results in experiments. In later writings, however, Fisher considered that every researcher should select the significance level according to the circumstances, stating that "in fact, no scientific worker has a fixed level of significance at which from year to year and in all circumstances, he rejects hypotheses" (Fisher, 1956, p. 42). Instead, Fisher suggested publishing the exact p-value obtained in each particular experiment which, in fact, implies establishing the significance level after the experiment.

In spite of these recommendations, research literature shows that the common arbitrary levels of .05, .01, .001 are almost universally selected for all type of research problems. Skipper, Guenter, and Nass (1970) suggested that this has the consequence of differentiating research findings that will be published or not and warned us to choose level of significance with full awareness of its

implications for the problem under investigation. Sometimes, when the power of the test is low and Type II error is important, a higher probability of Type I error might be preferable.

(d) Misinterpretations of the significance level are linked to misinterpreting significant results, about which there was another disagreement between Fisher and Neyman and Pearson. A significant result, for Fisher, implied that the data provided evidence against the null hypothesis, while for Neyman and Pearson it just stated the relative frequency of times that we would reject a true null hypothesis (the Type I error) in the long run. On the other hand, we should distinguish between statistical and practical significance. In Example 1 we obtained a difference in mean scores between the two groups of 13.32, which was significant. However, we might have obtained a higher level of significance with a smaller experimental effect and a larger sample size. Practical significance involves statistical significance plus a sufficiently large experimental effect.

The Different Levels of Hypotheses in Research

The level of significance is not the only concept misunderstood in significance testing. Some research papers have also shown confusion between the roles of the null and alternative hypotheses (Vallecillos, 1994, 1995) as well as between the statistical alternative hypothesis and the research hypothesis (Chow, 1996). Chow distinguishes diverse hypotheses implicated at different levels of abstraction in experimental research directed to confirm theories, such as that outlined in Example 1, as follows.

(a) *Substantive hypothesis* (which, in Example 1, is that learning is affected by the semiotic tools available to deal with a concept). A substantive hypothesis is a speculative account for a given phenomenon. Usually it is not possible to investigate this hypothesis directly, because it refers to an unobservable construct or mechanism. To investigate the substantive hypothesis some observable implication from the substantive hypothesis must be deduced.

(b) *Research hypothesis* (that computers will improve the learning of statistics). This is an observable implication of the substantive hypothesis. If we do not obtain support for the research hypothesis, the substantive hypothesis will not be supported.

(c) Often the research hypothesis is not specific enough for conducting empirical research. Therefore, it is necessary to devise a well-defined dependent variable (in Example 1, total score) obtained from an experimental task (the test) given to some subjects (students) in a specific context (experimental and control groups after the teaching experiment). On this basis, an *experimental hypothesis* can be constructed (that performance in the test will be better for the experimental group).

(d) An implication of the experimental hypothesis will be stated to carry out the statistical analysis (that the mean score in the test will be higher in the students taught with the use of computers than in the students who have no access to computers). This implication is the *alternative statistical hypothesis,* $H_1 \equiv \mu_e > \mu_c$, which is not identical to the experimental hypothesis, but a consequence of it at the statistical level.

(e) Finally, the logical complement of the alternative statistical hypothesis is that the mean scores in the two student populations will be the same, $H_o \equiv \mu_e = \mu_c$. Stating the null hypothesis serves to specify the sampling distribution of the test statistic. Then we can start a chain of reasoning (Table 3) that will lead us to accept or reject the series of hypotheses that we have described and that are shown in Table 2.

Table 2. *Different Levels of Hypotheses in Experimental Research*

Hypothesis involved	Example
Substantive hypothesis	Learning is affected by the representations available
Research Hypothesis	Computers will favor the learning of statistics
Experimental hypothesis	Test scores are higher in students who use computers
Alternative statistical hypothesis	$H_1 \equiv \mu_e > \mu_c$
Null hypothesis	$H_0 \equiv \mu_e = \mu_c$

Table 3. Chain of Reasoning Involved in Getting Support for a Substantive Hypothesis

Implication 1	If learning is affected by the representations available, then computers will favor the learning of statistics
Implication 2	If computers favor the learning of statistics, then test scores will be higher in students who use computers
Implication 3	If test scores are higher in students who use computers, then $\mu_e > \mu_c$
Implication 4	If it is not the case that $\mu_e > \mu_c$, then $\mu_e = \mu_c$
Implication 5	If $\mu_e = \mu_c$, then a significant value of $\bar{x}_e - \bar{x}_c$ is highly unlikely
Observation	$\bar{x}_e - \bar{x}_c$ is significant,
Conclusion 5	$\bar{x}_e - \bar{x}_c$ is significant, therefore we reject $\mu_e = \mu_c$
Conclusion 4	We reject $\mu_e = \mu_c$, therefore we assume that $\mu_e > \mu_c$
Conclusion 3	$\mu_e > \mu_c$; so provided that there was adequate experimental control, we assume than test scores are higher in students who use computers
Conclusion 2	Test scores are higher in students who use computers; provided that the test is a reliable measure of learning, then computers favor the learning of statistics
Conclusion 1	Computers favor the learning of statistics; provided the only difference in the two teaching methods is the representations available, then the substantive hypothesis is supported by our data.

Following Chow (1996) we present in Table 3 the series of embedded implicative deductions from which only the innermost one (Implication 5) is related to the process of testing for significance. This is the central core of the whole series of implications 1 to 5 that together constitute the process of scientific inference to support a substantive hypothesis. Therefore, the view of statistical tests as significance testing fits naturally with this type of research, whereas the view of statistical tests as a decision process would be preferable in a practical situation where a decision should be taken, such as in Example 2 or in quality control.

Implications 1 to 4 in Table 3 are not supported by statistical theory, but by theoretical considerations from the field under study, and by an adequate experimental control that ensures that all the relevant concomitant variables have been held constant and that the test given to the students is a reliable and valid measure of the construct studied (learning). According to Chow (1996), many of the criticisms against statistical testing are misdirected, as they refer, not to the statistical process (implication 5 in Table 3), but to the other components of the inferential procedure (implications 1 to 4). Replacing or supplementing statistical tests by other statistical methods, such as confidence intervals or power analysis, will not solve the problem of adequate experimental control or lack of adequate theoretical background in a particular research study.

Some Philosophical Issues

We have now identified several reasons for difficulties in understanding statistical tests. On the one hand, statistical tests involve a series of concepts such as null and alternative hypotheses, Type I and Type II errors, probability of errors, significant and nonsignificant results, population and sample, parameter and statistics, sampling distribution. Some of these concepts are misunderstood or confused by students and experimental researchers. Moreover, the formal structure of statistical tests is superficially similar to that of proof by contradiction; however, there are fundamental differences between these two types of reasoning that are not always well understood.

In proof by contradiction we reason in the following way:

> If A implies B cannot happen
>
> Then, if B happens, we deduce A is false

In statistical testing, it is tempting to apply similar reasoning as follows:

> If A implies B is very unlikely to happen
>
> Then, if B happens, we deduce A is very unlikely to be true

However, this would not be a valid conclusion, and herein lies the confusion.

In addition to these difficulties, we have seen that the controversy surrounding statistical inference involves the philosophy of inference and the logical relations between theories and facts. Science is built from empirical observations and it is not possible to take data from whole populations but only from samples. We expect from statistical testing more than it can provide us, and underlying this expectation is the philosophical problem of finding scientific criteria to justify inductive reasoning, as stated by Hume. Until now, the contributions made by statistical inference in this direction have not achieved a complete solution to this problem (Black, 1979; Burks, 1977; Hacking, 1975; Seidenfeld, 1979).

On the other hand, there are two different views about statistical tests that sometimes are confused or mixed. Fisher saw the aims of significance testing as confronting a null hypothesis with observations and for him a p-value indicated the strength of the evidence against the hypothesis (Fisher, 1958). However, Fisher did not believe than statistical tests provided inductive inferences from samples to population, but, rather, a deductive inference from the population of possible samples to the particular sample obtained in each case.

For Neyman (1950), the problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action. To accept a hypothesis means only to decide to take one action rather than another. This does not mean that one necessarily believes that the hypothesis is true. For Neyman and Pearson, a statistical test is a rule of inductive behaviour; a criterion for decision-making, which allows us to accept or reject a hypothesis by assuming some risks.

Today, many researchers employ the statistical tools, methods, and concepts of the Neyman-Pearson theory with a different aim, namely, to measure the evidence in favour of a given hypothesis (Royal, 1997). The inner reasoning in Table 3 (Implication 5, observation and conclusion 5) can serve to describe the usual reasoning in statistical tests today, which consists of:

(a) A binary decision - deciding whether the result is significant or not. This decision is made by comparing the p-value with the level of significance, which is stated before collecting the data.

(b) An inferential procedure involving a conditional syllogism (implication 5 in Table 3): If $\mu_e = \mu_c$, then a significant value of $\bar{x}_e - \bar{x}_c$ is highly unlikely; $\bar{x}_e - \bar{x}_c$ is significant, therefore we reject $H_0 \equiv \mu_e = \mu_c$.

(c) Another inferential procedure involving a disjunctive syllogism. Either $\mu_e > \mu_c$, or $\mu_e = \mu_c$; if we reject $\mu_e = \mu_c$, then $\mu_e > \mu_c$.

Therefore, the current practice of statistical tests contains elements from Neyman-Pearson (it is a decision procedure) and from Fisher (it is an inferential procedure, whereby data are used to provide evidence in favor of the hypothesis), which apply at different stages of the process.

Other features taken from Neyman-Pearson are that H_0 is the hypothesis of no difference, that the α level must be chosen before data analysis and it must remain unchanged, and that there are two types of error. From Fisher we preserve the suggestion that the inference is based on a conditional probability; the probability of obtaining the data given that H_0 is true, and that H_0 and H_1 are mutually exclusive and complementary. We should also add that some researchers often give a Bayesian interpretation to the result of (classical) hypothesis tests, in spite of the fact that the view from Bayesian statistics is very different from the theories of either Fisher or Neyman and Pearson.

Psychological Factors Contributing to the Prevalence of Common Errors

The above practice of statistical tests has been called the framework of Neyman-Pearson orthodoxy (Oakes, 1986) or Neyman-Pearson hybrid logic (Gingerenzer, 1993), who think it can explain the belief that statistical inference provides an algorithmic solution to the problem of inductive inference, and the consequent mechanistic behavior that is frequently displayed in relation to statistical tests.

As we have described, Fisher and Neyman/ Pearson had different interpretations of statistical tests, which include the way in which we should determine the level of significance, as well as the meaning of a significant result. According to Gingerenzer et al. (1989), the dispute between these authors has been hidden in applications of statistical inference in psychology and other experimental sciences, where it has been assumed that there is only one statistical solution to inference. Textbooks, such as that by Guilford (1942), have contributed to spreading a mixture of Fisher's logic of significance tests with some Neyman-Pearson components, where Bayesian interpretations were given to the level of significance and related concepts. As a result, these books have also helped to spread the misinterpretation of statistical tests.

Using an insightful analogy, Gingerenzer et al. (1989) compare the Neyman-Pearson features within the current practice of statistical testing with the superego of statistical reasoning, because they prescribe what should be done and do not leave freedom to researchers. They require the specification of precise hypotheses, significance levels, and power before the data are collected, and that the probability of errors should be interpreted in the context of repeated sampling. Fisher's features are compared to the ego of statistical reasoning. It is convenient for researchers who want to carry out their dissertations and get their papers published, if they can determine the level of significance after the experiment, establish a diffuse or no alternative hypothesis before collecting the data, and interprete the probability of error as the probability of error in their own experiment.

The third component in the researcher's behavior described by Gingerenzer et al. (1989) is the Bayesian wish to assign probabilities to the hypothesis on the basis of research data (the id of the hybrid logic). When we find a significant result, we ask ourselves whether this result may be due to chance or whether in fact it was a result of our experimental manipulation. Falk (1986) finds it natural that one interprets the level of significance as the posterior probability of error, once we have rejected the hypothesis, because this is in fact the probability in which the researcher is interested. Gingerenzer et al. (1989) suggest that the conflict among these three psychological components is what explains our misuses of statistical inference, and the institutionalization of the level of significance as a measure of research quality in scientific journals and statistics textbooks.

On the other hand, biases in inferential reasoning can be seen simply as examples of adults' poor reasoning in probabilistic problems, which has been extensively studied by psychologists in relation to other concepts, such as randomness, probability, and correlation (Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). In the specific case of misinterpreting statistical inference results, Falk and Greenbaum (1995) suggest the existence of a profound psychological mechanism leading people to believe that they eliminate chance and minimize uncertainty when they obtain a significant result. They describe *the illusion of probabilistic proof by contradiction*, or *the illusion of attaining improbability*, which consists of the erroneous belief that one has rendered the null hypothesis improbable by obtaining a significant result, based on a misleading generalization from logical reasoning to statistical inference (Birnbaum, 1982; Lindley, 1993). While a contradiction definitely disproves the premise from which it is drawn, the belief that obtaining data whose conditional probability under a given hypothesis is low implies that the conditioning hypothesis is improbable is a fallacy. The illusion of probabilistic proof by contradiction is, however, apparently difficult to eradicate, in spite of clarification in many statistics textbooks. In other cases, this misconception is implicit in textbooks, as shown by Falk and Greenbaum (1995).

According to Falk (1986), misconceptions around the significance level are also related to difficulties in discriminating between the two directions of conditional probabilities, otherwise known as *the fallacy of the transposed conditional* (Diaconis and Friedman, 1981), which have been long recognized as pervasive among students and even professionals. In addition, Falk (1986) suggests that the verbal ambiguity in presenting α as P(Type I Error) may provoke confusion between the two opposite directions of conditional probabilities amongst students, who seem to believe that they are dealing with the probability of a single event. Falk suggests that "Type I error" is an unfortunate expression and should not be used on its own. A "conditional event" is not a legitimate concept, and only conditional probabilities are unequivocally defined, even though this confusion sometimes appear in textbooks.

Although α is a well defined conditional probability, the expression "Type I error" is not conditionally phrased, and does not spell out to which combination of the two events it refers. Now, when H_0 is rejected and we wish to ask ourselves which kind of error can be committed, the concept of "Type I error" comes immediately to mind, as the crucial distinction between the two

opposite directions of the conditional probabilities is blurred. This leads us to interpret the significance level as the conjunction of the two events "the null hypothesis is true" and "the null hypothesis is rejected" (Menon, 1993).

For many years, criticisms have been raised against statistical testing, and many suggestions have been made to eliminate this procedure from academic research. However, significant results continue to be published in research journals, and errors around statistical tests continue to be spread throughout statistics courses and books, as well as in published research. Falk (1986) suggests that researchers experience an illusory confidence in statistical tests because of the sophistication of mathematical terms and formulas, which contributes to our feeling that statistical significance guarantees objectivity. An additional problem is that other statistical procedures suggested to replace or complement statistical tests (such as confidence intervals, measuring the magnitude of experimental effects, power analysis, and Bayesian inference) do not solve the philosophical and psychological problems we have described.

Some Common Criticisms against Hypothesis Testing Revisited

We have analysed in detail the logic of statistical testing, its role in scientific inference, and the philosophical and psychological factors that contribute misunderstanding and misuses of statistical tests. In this section, we revisit some frequent criticisms against statistical testing.

1. What is asserted in the null hypothesis in example 1 is that there is no difference between the two populations' means. It is evident to many critics that the null hypothesis is never true and therefore statistical tests are invalid, as they are based on a false premise (that the null hypothesis is true).

That this criticism is not pertinent can be deduced from the fact that, even when the null hypothesis is not true, the logic of statistical testing is not invalid. This logic is not affected by whether the null hypothesis is true or false, because what is asserted in a test is that a significant result is improbable, given that the null hypothesis is true. This is a mathematical property of the sampling distribution that has nothing to do with the truth or falsity of the null hypothesis.

2. In practice we identify the hypothesis of interest with the statistical alternative hypothesis. However, the alternative hypothesis says nothing about the exact magnitude of the difference between the population means. Statistical significance is not informative about the practical significance of the data.

When this criticism is applied to significance testing (Example 1) it might be due to confusion between the different levels of hypotheses implied in the inferential procedure shown in Tables 2 and 3. The aim of experimental research directed towards theory confirmation is providing support for the substantive hypothesis. As we saw in example 1, the magnitude of the difference between the population means has nothing to do with the substantive hypothesis. This difference refers to the sampling distribution of the statistics, that is, to the alternative statistical hypothesis. There is not a unique correspondence between the substantive hypothesis and the statistical alternative hypothesis, which is derived from a particular experiment and a particular test instrument. Theories should be assessed with careful thinking and sound judgment (Harlow, 1977).

In the context of taking a decision (Example 2), however, the magnitude of the effect could be relevant to the decision. In these cases, statistical tests are still useful in making the decision, though

they should be complemented with power analysis and/ or estimates of the magnitude of the effects as subordinated to research questions of interest (Levin, 1998a).

3. The choice of the level of significance is arbitrary; therefore some data could be significant at a given level and not at another different level.

It is true that the researcher chooses the level of significance. This arbitrariness does not, however, mean that the procedure is invalid or unuseful. Moreover, it is also possible, following the approach of Fisher, to use the exact p-value to reject the null hypothesis at different levels, though in the current practice of statistical testing it is advisable to chose the significance level before taking the data to give more objectivity to the decision.

4. Statistical significance is not informative as to the probability of the hypothesis being true. Nor is statistical significance informative of the true value of the parameter. For this reason many researchers suggest replacing tests with confidence intervals.

It is true that tests are not informative of the probability of the hypothesis being true. However, confidence intervals are not informative of this probability either. Confidence intervals give an interval within which the true value of the parameter should be in a given percentage of samples, though they do not ensure in which interval the parameter lies for our particular experiment. Therefore, they do not substitute for, but rather complement tests of significance, and are subject to similar controversies and misconceptions.

4. Type I error and Type II errors are inversely related. To critics, researchers seem to ignore Type II errors while paying undue attention to Type I error.

Though the probabilities of the two types of errors are inversely related, there is a fundamental difference between them. While the probability of Type I error α is a constant that can be chosen before the experiment is done, the probability of Type II error is a function of the true value of the parameter which is unknown. To solve this problem, power analysis assumes different possible values for the parameter and computes the probability of Type II error for these different values. This practice is useful for some applications of inference, such as decision taking (Example 2) and quality control. However, we can apply the same objections as in point 2 as regards experiments oriented towards supporting a given theoretical substantive hypothesis (Example 1) in which there is no indication about the particular size of particular parameters of the experimental dependent variables. That is, the two types of errors do not play the same role in the corroboration of scientific theories, altough they could be equally important in other applications of inference, such as quality control.

5. It is not clear what the meaning of a non-significant result is. For some critics this may be due to the fact that the test used is not of sufficient power.

We can apply here the same reasoning as in point 4. It is clear that the null and alternative hypothesis, rejecting and accepting the null hypothesis, significant and non-significant results do not play a symmetric role in significant testing. While a significant result contradicts the null hypothesis because of its low probability, a non-significant result is highly likely when the null hypothesis is true, but could also have been produced by other factors. This can also happen in proof by contradiction, where we reason in the following way:

If A is true implies B is false

Then, if B is true, A is false

If B happens, we conclude A is not possible, but when B does not happen, we can not deduce that A is necessarily true; here there is also asymmetry between the consequences of B and not-B.

Teaching and Learning Inference Concepts

In this paper we have described the logic of significance tests, their role in experimental research, the conceptual, psychological and philosophical difficulties related to them and, finally, we have revisited some frequent criticisms against statistical significance testing. These criticisms cannot be applied to the mathematical procedure in statistical testing, where there are no contradictions. On the contrary, they are related to the misuses of significance testing and are the consequence of conceptual misunderstandings, and the philosophical and psychological problems that we have analysed throughout the paper.

Statistical educators are not indifferent to these problems, as it is shown by the Invited Paper Meeting on Statistical Education and the Significance Tests Controversy at the International Statistical Institute's Fifty-second Session and by the International Association for Statistical Education's Round Table on "Training Researchers in the Use of Statistics". As described by Ito (1999), there are three different levels in the statistical tests controversy:

(a) The dispute within statistics itself, where different methods and various interpretations for the same methods are recommended in the Fisher, Neyman-Pearson and Bayesian approaches.

(b) The controversy in the applications of statistics, where, in practice, the significance test is an informal blending of Fisher's original pure significance test and Neyman-Pearson theory with concepts and interpretation which are not part of the latter. Moreover, journal editors and professional societies are suggesting changes in research and publication policies as regards statistical methods (Lecoutre, 1999).

(c) The teaching controversy about when, how, and to what extent we should teach about statistical inference.

We agree with Ito that these three different levels are in fact interrelated, because our conceptions about statistical theory also affect our applications and teaching of statistics. This is particularly important, as with the increase of research in the teaching and learning of statistics, data handling is increasingly being introduced at school level (Shaughnessy, Garfield, & Greer, 1997) including also, in many countries, the rudiments of inference (Dahl, 1999). Our view is that hypothesis testing should not be abandoned in social sciences and education, but rather its teaching and practice should be changed to lead to a "meaningful process" (Levin, 1998b), which includes independent replication of studies, choosing optimal sample sizes, combining hypothesis testing with confidence intervals and/or effect-size estimation, and specifying criteria of "success" prior to the experiment.

It is clear, from our analysis, that there is a conceptual complexity in statistical tests, and that particular attention should be paid to the teaching of inference if we want to prevent future misunderstanding in our students similar to those described by Vallecillos (1999). Revision of the teaching methodology in introductory statistical courses has been suggested (Moore, 1997 and related discussion) to change towards a constructivist model of learning wherein the teacher guides his/her students to move towards specific statistical competencies and knowledge. New texts that

change the students' role from listening towards a more active participation in structured activities (e.g., Rossman, 1996) facilitate such an approach.

Since computers make a variety of computations and graphical displays possible, Moore (1997) recommends giving students the opportunity to experiment with real data and problems. In particular, computer simulations could contribute to improving students' understandings of the ideas of sample variability, sample statistic and its distribution, about which there are many misconceptions (Rubin, Bruce, & Tenney, 1991; Well, Pollatsek, & Boyce, 1990) and that are essential to understanding the logic of significance testing. For example, delMas, Garfield and Chance (1999) describe the Sampling Distribution software and instructional activities designed to guide students in their exploration of sampling distributions. In their experiments, the students were able to change the shape of the theoretical distribution in the population (normal, skewed, bimondal, uniform, U-shaped) and simulate sampling distributions of different statistics for various sample sizes. The activities were aimed to focus the students' attention towards the Central Limit Theorem.

However, and even when results demonstrated a significant positive change in the students as a consequence of the activities, delMas et al. (1999) warn that the use of technology and activities based on research results did not always produce effective understanding of sampling distributions. The authors suggest that the new activities and the learning of the software might be too demanding for some students, and the amount of new information about the sofware might interfere with the students' learning of sampling distributions, whose understanding require students to integrate ideas of distribution, average, spread, sample, and randomness. It is clear that we need more research that helps us to understand how technology may be used to help students in their learning process (see Ben-Zvi, this issue). In particular we need to find good didactical situations in which students can be confronted with their psychological misconceptions, such as the confusion between a conditional probability and its inverse or their belief in the possibility of computing the probability of a hypothesis (within a objective conception of probability).

On the other hand statistical testing is just a part of the more general process of scientific inference, as indicated in Tables 2 and 3. However, we frequently find that statistical inference is taught in isolation without connecting it with a more general framework of research methodology and experimental design. From our point of view, it is necessary to discuss the role of statistics within experimental research with the students and make them conscious of the possibilities and limitations of statistics in experimental work. Moroever, we agree with Wood's (1998) suggestion to focus introductory statistical courses around statistical thinking, that is, around the Plan-Do-Check-Act learning cycle. Statistical data analysis is not a mechanical process, and therefore should neither be taught nor applied in this way. Since statistics is not a way of doing, but a way of thinking that helps us to solve problems in science and everyday life, teaching statistics should begin with real problems through which students develop their ideas, working through the different stages of solving a real problem (planning a solution, collecting and analyzing data, checking initial hypotheses, and taking appropriate decisions).

Finally, we recommend that researchers should recognize the complexity of applying inference to solving real problems and realize that they need the collaboration of professional statisticians, in addition to using their professional knowledge to judge the extent to which their research questions may be answered by statistical analyses.

References

American Psychological Association (1994). *Publication Manual of the American Psychological Association* (Fourth edition.). Washington, DC: American Psychological Association.

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics, 4*, 24–27.

Black, M. (1979). InducciÛn y probabilidad [Induction and probability]. Madrid, Spain: C· tedra.

Burks, A. W. (1977*). Chance, cause, reason: An inquiry into the nature of scientific evidence*. Chicago: University of Chicago Press.

Chow, L. S. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

delMas, R. C., Garfield, J. B., & Chance, B. (1999, April). *Exploring the role of computer simulations in developing understanding of sampling distributions*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Dahl, H. (1999, August). Teaching hypothesis testing. Can it still be useful? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 197-200). Helsinki, Finland: International Statistical Institute.

Diaconis, P., & Freedman, D. (1981). The persistence of cognitive illusions. *Behavioral and Brain Sciences, 4*, 378-399.

Ellerton, N. (1996). Statistical significance testing and this journal. *Mathematics Education Research Journal, 8*(2), 97–100.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*, 83–96.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5* (1), 75-98.

Fisher, R. A. (1935). *The design of experiments*. Edimburgh: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1958). *Statistical methods for research workers* (Thirteenth edition). New York: Hafner.

Gingerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren & C. Lewis (Eds*.), A handbook for data analysis in the behavioral sciences: Methdological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gingerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kr¸ger, L. (1989). *The empire of chance. How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Guilford, J. P. (1942*). Fundamentals of statistics in psychology and education*. New York: Basic Books.

Hacking, I. (1975). *The logic of statistical inference*. Cambridge: Cambridge University Press.

Harlow, L. L. (1997). Significance testing: Introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds*.), What if there were no significance tests?* (pp. 1-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997*). What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Ito, P. K. (1999, August). *Reaction to invited papers on statistical education and the significance tests controversy*. Invited paper presented at the Fifty-second International Statistical Institute Session, Helsinki, Finland.

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

Lecoutre, B. (1999). Beyond the significance test controversy: Prime time for Bayes? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 205-208). Helsinki, Finland: International Statistical Institute.

Levin, J. R. (1998 a). To test or not to test H_0? *Educational and Psychological Measurement, 58,* 313-333.

Levin, J. R. (1998 b). What if there were no more bickering about statistical significance tests? *Research in the Schools, 2,* 45-53.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review, 11,* 143-155.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15,* 22-25.

Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5*(1), 4–18.

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

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123–155.

Moses, L. E. (1992). The reasoning of statistical inference. In D. C. Hoaglin & D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107-122). Washington, DC: Mathematical Association of America.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970*). The significance tests controversy. A reader*. Chicago: Aldine.

Neyman, J. (1950). *First course in probability and statistics*. New York: Henry Holt.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgments.* Englewood Cliffs, NJ: Prentice Hall.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10,* 159-163.

Robinson, D. H., & Levin, J. T. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rossman, A. J. (1996). *Workshop statistics: Discovery with data.* New York: Springer.

Royal, R. (1997). *Statistical evidence. A likelihood paradigm*. London: Chapman & Hall.

Rubin, A., Bruce, B., & 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* (pp. 314–319). Voorburg, The Netherlands: International Statistical Institute.

Seidenfeld, T. (1979). *Philosophical problems of staistical inference: Learning from R. A. Fisher*. Dordrecht, The Netherlands: Reidel.

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (Volume 1) (pp. 205-237). Dordrecht, The Netherlands: Kluwer.

Skipper, J. K., Guenter, A. L., & Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social sciences. In D. E. Morrison & R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 155-160). Chicago: Aldine.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Vallecillos, A. (1994). Estudio teÛrico experimental de errores y concepciones sobre el contraste de hipÛtesis en estudiantes universitarios [Theoretical-experimental study of errors and conceptions concerning statistical tests in university students]. Unpublished doctoral dissertation, University of Granada, Spain.

Vallecillos, A. (1995). ComprensiÛn de la lÛgica del contraste de hipÛtesis en estudiantes universitarios. [University students' understanding of the logic of statistical tests]. *Recherches en Didactique des MathËmatiques, 15*(3), 53–81.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 201-204). Helsinki, Finland: International Statistical Institute.

Vallecillos, A., & Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. In L. Puig & A. GutiÈrrez (Eds.), *Proceedings of the Twentieth Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4) (pp. 271–378). Valencia, Spain: University of Valencia.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of the sample size on the variability of the means. *Organizational Behavior and Human Decision Processes, 47*, 289–312.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wood, G. R. (1998). Transforming first year university statistics teaching. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (Volume 1) (pp. 167-172). Singapore: International Statistical Institute.

Yates, F. (1951). The influence of "Statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

Author Notes: The author is indebted to Joel R. Levin and Paul K. Ito for their comments and helpful suggestions on an earlier version of this paper.

This research has been supported by the grant PB96-1411 (M.E.C, Madrid).