WHY SHOULD ONE TEACH PROBABILITY THEORY TO MAKE DECISIONS WHEN TESTING STATISTICAL HYPOTHESES?

Maria Manuel da Silva Nascimento
UTAD, Portugal
Nilson Luiz Castelucio Brito
UNIMONTES, Portugal
mmsn@utad.pt

Nowadays one shouldn't ignore the use of computers in statistics, the main reason being that they are faster and more reliable. Almost all statistical software computes the p-values; therefore students and researchers can take their decisions only based on its "usual" value, 0.05. If the p-value is lower than 0.05 then the null hypotheses for a statistical test is "simply" rejected by the computer. Do users ask about the meaning of this software output? If we are testing statistical hypotheses we have the null hypotheses tested against alternative hypotheses. Do users think about them? Since the decisions are based on sampling, the statistical tests decisions involve uncertainty and so two types of errors can be made. Do users think about them? A questionnaire was constructed and given to students and researchers in order to make a first approach about these subjects.

INTRODUCTION

At the present we can't ignore the use of computers to make statistical calculations and the main reason for their use is that computations become unavoidable since almost all the statistical software computes "everything" and the users of software should always care to ask about the statistical meaning of what they are doing. One of the simplest statistical calculations that there is software for, is testing statistical hypotheses nevertheless this is a difficult subject to deal with.

When we begin to speak about testing statistical hypotheses we have to say that in these tests we have two kinds of hypotheses: the null hypothesis that refers to the hypotheses that we wish to test (denoted by H_0) and an alternative hypotheses (denoted by H_1). In parametric tests, for instance, the null hypotheses refers to a population parameter will always be stated to specify an exact value of the parameter, whereas the alternative hypotheses allows for the possibility of several values or their ranges. The null hypotheses may be true or false, and, when testing it, we can either reject it or not. If we do not reject the null hypotheses, it means that the sample does not give enough evidence to refute it. On the other hand, rejection of the null hypotheses means that the sample evidence refutes it. Therefore, rejection means that there is a small probability of obtaining the sample information observed when, in fact, the hypotheses is true (Walpole and Myers, 1993). Sometimes either students or researchers find it difficult to establish either null or an alternative hypothesis, because both the statistical inference logical reasoning and interpretation are difficult (Batanero *et al*., 1994 and Vallecillos and Batanero, 1997). However, it is usual to see many researchers paying too much attention to the hypotheses tests computations and forgetting the magnitudes of the effects they are trying to investigate.

Since the decisions are based on sampling, the statistical tests involve uncertainty and so two types of errors that can be committed. Let us define α as the level of significance as well as the P (type I error) = P (reject H_0 | H_0 is true) and β as the P (type II error) = P (do not reject H_0 | H_0 is false) (Hawkins *et al.*, 1992). Using α and knowing the statistical test law, the critical region (CR, the region where H_0 is rejected) and the acceptance region (AR, the region where H_0 is not rejected) may be established. Computing the test statistic (TS) for a sample we do not reject H_0 if TS $\in$ AR and we reject H_0 if TS $\in$ CR. Usually the α value is pre-selected in an arbitrary way when statistical hypotheses are tested. On the other hand, the p-value is the lowest value of probability at which the observed value of the TS is significant, that is the probability at which H_0 may be rejected.

Using computers for hypotheses tests may produce an "automation effect" in the use of the p-value to make decisions. This is reported in some papers and degree projects (Pimenta and

Batanero, 2005). In many internet statistical applets (Duckworth *et al.*, 2005) we can do a hypotheses test and we can compute the p-value, for instance, about the mean of a variable that has a normal distribution with a known variance. At this point a question is raised: do statistical tests users ask about the meaning of the software or of the internet statistical applet output? When a linear regression adjustment is made, or a Student *t*-test, or a non-parametric test, or any other test, we are testing statistical hypotheses, and we can not forget it. However, some students (or even some researchers) only ask: is p-value lower than 0.05 (the α value usually considered)? If the answer is yes, they decide that they can reject H_0, based only in the predefined α value, and forgetting "how lowest" is this p-value.

The debate raised by those questions is included in a broader one; the controversy about the suitable use of statistics has recently increased among professional organizations such as the American Educational Research Association or the American Psychological Association. Those associations propose important shifts in their editorial policies regarding the use of statistics and are recommending better use of statistical language in published research. For instance, research journals in medicine such as the *British Medical Journal* or *Statistics in Medicine* have highlighted the poor quality of methodology and statistics in medical research. This debate is also reflected in research journals of psychology, and education (Batanero, 2001). What do they say about decisions taken based only in p-value? They are very poor, with no methodological and general questions, and if the conclusions do not include the meaning of the size of effects they do not have good acceptance in many cases. Also recently, the Task Force on Statistical Inference (Wilkinson, 1999) published a paper about those questions and the publishing editors decided that revisions must also refer to methodological questions. Among other aspects, they suggested to publish the exact p-value, the estimation of the effects and the confidence intervals.

While designing an experiment it is very important to think about, for instance, the size of the random sample, the type of variable and the statistical methods that should be used. For example, we may use U-test or Mann-Whitney-Wilcoxon (M-W-W) test. Some of the parametric hypotheses tests can be used if we may suppose that the populations have a Gaussian distribution, their variances are known or they have equal variances and the samples are independent. However, in many situations, these conditions can not be met and, instead, we may use non-parametric tests, alternative processes based in less restrictive assumptions. For instance, the M-W-W test is a non-parametric alternative to the Student *t-test* for two samples, and if any of the parametrical assumptions is not verified, and its goal is to test if two samples come from identical populations, that is if the two populations have the same median. We can test the null hypotheses $\eta_1 = \eta_2$, against the alternative hypotheses that may be $\eta_1 \neq \eta_2$. Using this test we have no need to be sure that the original populations that we are comparing have the same Gaussian distribution, as need in a Student t-test. M-W-W test is based on ranks (ordering of data) and not on the actual values of the two random samples. To perform this test, we may use the same methodology of a hypotheses parametric test (Freund *et al.*, 2000):

- Step 1: To establish the two hypotheses, H_0 and H_1;
- Step 2: To fix the significance level, α;
- Step 3: To define the critical region and the acceptance region;
- Step 4: To compute the test statistic;
- Step 5: To compare to the *U* distribution;
- Step 6: To decide between to reject or not reject H_0.

Many times when we are teaching non-parametric tests, or other tests of statistical hypotheses, students are from other fields other then Mathematics or Statistics. Almost every time, students only want to be introduced to the use of software: input the data, watch its output and make a decision only based on the "usual" p-value. They do not begin to ask if the test is suitable for this kind of data or not, or if those tests require any kind of assumptions or even if those are met. These are some of the reasons that have encouraged us to begin a research about the ideas that the software users - student and researchers - have about its use for testing statistical hypotheses. Based on the results of this first inquiry, in order to prepare the students (or researchers) to use those statistical methods we will try to establish some guidelines for future

studies, as well some others to answer the question: why should we teach the use of probability theory to make decisions when testing statistical hypotheses?

RESEARCH AND RESULTS

The research is the analysis of a survey was given to students and researchers in order to make a first approach about some topics in testing statistical hypotheses. It has been carried out on a sample of students and researchers of different scientific areas (other then Mathematics or Statistics). The questionnaire was given to 35 people: researcher workers, Master and Graduate students in Montes Claros, MG, Brazil and in Vila Real, Portugal.

The questions were:

1. What kind of scientific work did you publish in last two years?
2. In what way did you participate in the work (author, co-author or other)?
3. Did you do any statistical treatment or not?
4. If you have used any statistical treatment, select the treatment you have applied.
5. Did you formulate the statistical hypotheses of your statistical tests and did you know if it was the correct test to be used?
6. Did you base your decision criterion for rejecting the null hypotheses only using p-value?
7. If you have answered p-value, in what did you based its use?
8. Did you ask someone to help you to confirm your statistical decisions in view of your scientific work?
9. In affirmative case, who has helped you?

Referring to the first question, we had 85 answers. Twenty were "scientific papers" (23,5%); but considering only that we had 35 replies, 57,1% wrote "scientific papers" in the last two years. Twenty four answers referred as a "scientific work" a M.Sc. (11) or a Ph.D. (13), that is 68% (or 28,2% if we consider all the 85 multiple answers). Other 24 answers also referred as a "scientific work" a poster or an abstract in a scientific meeting, that is 68% (or 28,2% if we consider all the 85 several answers). The remaining 18 answers (21,1% of the 85 answers, were about other "scientific work" (graduation reports, experimental designs reports, textbooks, seminars and book chapters).

Referring to the second question, 34.3% of the answers stated the condition of "author." About the third question, only 5.7% of the answers stated that the scientific work had statistic treatment and between those answers (fourth question) 23.5% stated that Analysis of Variance was the statistical treatment used. Here a remark must be made, we suspect that most of the students and researchers were related to forestry, animal and agricultural sciences in view of the profiles of our universities research areas and degrees.

Lastly the fifth question, 82.9% of the answers stated "yes," meaning that they had knowledge about formulation of the statistical hypotheses and they knew that it was the correct test to be used. Those answers made us very suspicious of the nature of the knowledge about the hypotheses formulation. The truth is that we should have been more careful in the exploring the concepts implicit in those hypotheses formulations.

In what concerns the sixth question, 37.1% of the answers stated that have used "the test statistic (TS)" and the same percentage answered "the TS and p-value." Considering the answer "TS and p-value," 92.3% of the answers justified that the choice was based in "easiness to work with p-value" and the remaining 7.7% justified that "only the p-value has an interpretation." This remaining 7.7% of the answers "only the p-value has an interpretation" seemed to us that students and researchers have misunderstanding some p-value interpretation concepts. On the other hand, the answer "easiness to work with p-value" lead us to suspect that students and researchers from other areas (other then Mathematics or Statistics) are using this "easiness" to immediately conclude their works, forgetting methodological questions, determination of the size of effects and its confidence intervals, as well as sample sizes.

Referring to the eighth question, 71.4% of the answers stated that students or researchers "asked someone to help them to confirm their statistical decisions of that scientific work." Referring to the ninth question, from the 71.4% of the previous answers, 36% answered "a teacher of the same area has helped me" and 40% answered "a teacher of statistics has helped me." An independence chi-square test performed by crossing the variables "The decision criterion

was based in TS or in p-value only" and "In affirmative case, who has helped you?" showed that there is a relationship among the decision criterion and the kind of help received by the researcher (chi-squared TS = 19.393; d.f. = 24; p-value = 0.733).

DISCUSSION AND CONCLUDING REMARKS

During the presentation of the results, we have already referred that some answers made us very suspicious of the nature of the knowledge about the hypotheses formulation. As we have already acknowledged we should have been more careful in the exploring the concepts implicit in the hypotheses formulations. We have also assumed that everyone used computer software to make statistical calculations, but we did not remember to ask it directly, and which was the software used, and what kind of help they could provide to students or researchers. The suspicions aroused during the presentation of these results lead us to conclude that, perhaps the main points that need to be explained and explored in the future are those related with the concepts involved in the use of probability theory to make decisions when testing statistical hypotheses. In a brief revue of topics of research referred in the statistical didactics those concepts are mainly related with: *a. the determination of the null hypotheses H_0 and the alternative hypotheses 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.* (Batanero *et al*., 1994 and Vallecillos and Batanero, 1997) This review is emphasized by the remarks related with teaching hypotheses testing: *a. The test of hypotheses as a decision problem; b. Probabilities of error and relation between them; c. Level of significance as the risk of the decision maker; d. Interpretation of a statistically significant result*. (Vallecillos and Batanero, 1997)

At the beginning of this work we suspected that "why should we teach the use of probabilities to make decisions when testing statistical hypotheses?" was not a simple question to ask but at the end we have got sure that the work that needs to be done is a much bigger task than we could have imagine. Nevertheless, we are willing to change in order to promote a chain reaction among students and researchers. As regards future research work we need to rethink the questionnaire, including questions in order to assess the conceptual topics in testing statistical hypotheses. At the same time we should try a different approach facing the challenges when teaching this subject and when talking with researchers about it. May be in the future we should be concerned on how to teach statistical concepts using software and/or internet statistical applets, since students and researchers seem be encouraged by their use.

REFERENCES

Batanero, C., Godino, J. D., Green, D. R., Holmes, P. and Vallecillos, A. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology*, 25(4), 527-547.

Batanero, C. (Ed.) (2001). *Training Researchers in the Use of Statistics*. Granada: International Association for Statistics Education e International Statistical Institute.

Batanero, C. and Pimenta, R. (2005). Raciocínio estatístico: Avaliação a partir de projectos em Ciências da Saúde. *Actas do V Congresso Ibero-Americano em Ensino da Matemática*, 18.

Batanero, C. and Vallecillos, A. (1997). Análisis del aprendizaje de conceptos clave en el contraste de hipótesis estadísticas mediante el estudio de casos. *Recherches en Didactique des Mathématiques*, 17(1), 29-48

Duckworth, W., McCabe, G., Moore, D. and Sclove S. (2005). Statistical applets: P-Value of a test of significance, http://bcs.whfreeman.com/pbs/cat_050/pbs/pvalue_pbs.html.

Freund, J. E. and Simon, G. (2000). *A Estatística Aplicada: Economia, Administração e Contabilidade* (9th edition). Porto Alegre: Bookman.

Glickman, L., Hawkins, A. and Jolliffe, F. (1992). *Teaching Statistical Concepts*. London, Longman.

Myers, R. H. and Walpole, R. E. (1993). *Probability and Statistics for Engineers and Scientists* (5th edition). Upper Saddle River, NJ: Prentice Hall International Inc.

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