TEACHING HYPOTHESIS TESTING: A NECESSARY CHALLENGE

Wendy J. Post[1] and Marijtje A.J. van Duijn[2*]
[1]Department of Special Needs Education and Youth Care and [2]Department of Sociology
University of Groningen, the Netherlands
w.j.post@rug.nl

The last decades a debate has arisen about the use of hypothesis testing. This has led some teachers to think that confidence intervals and effect sizes need to be taught instead of formal hypothesis testing with p-values. Although we see shortcomings of the use of p-values in statistical inferences and the difficulties in really understanding hypothesis tests, we take a different view. We think that it is essential to understand what the fundamental principles are behind hypothesis testing in order to obtain correct statistical inference by interpreting confidence intervals (and p-values). In our course "Applied Statistics" for graduate students we designed course material in which we explain the three main approaches of hypothesis testing, Fisher, Neyman-Pearson and Bayesian, using a popular chance game as illustration. In this paper, we will shortly present the highlights of the course material, the results of the evaluation of our teaching, and suggestions for extensions.

INTRODUCTION

Most students and many researchers find the interpretation of statistical results, especially that of hypothesis testing difficult. Haller and Kraus (2002) showed that nearly 90% of psychologists and about 80% of methodology instructors had problems with the interpretation of p-values. Cumming (2011) stated in his book that confidence intervals (and meta-analysis) should be emphasized in teaching statistics, rather than p-values. If confidence intervals and effect sizes are the essence of statistical reasoning, one may even wonder whether traditional hypothesis testing needs to be taught at all.

As statisticians in a faculty of behavioural and social sciences, we frequently experience the problems of colleagues and students with statistics. As their primary interest is in psychology, education or sociology, doing statistical analyses is about how to get substantive results and conclusions, preferably as quickly and easily as possible. Although they may not completely understand the statistical models, they are well trained in making statistical computer programs produce p-values and confidence intervals. Understanding the results and drawing valid conclusions, however, is another cup of tea. P-values seem to be more a curse than a blessing. Does this mean that skipping p-values in teaching materials and publications will solve the problem? We do not think so. Confidence intervals and hypothesis tests are different sides of the *same* coin: Interpreting confidence intervals follows from understanding about hypothesis testing, and the other way round.

We have taken these experiences with students and colleagues into account when we designed a course titled "Applied Statistics", meant for research master students, a selected group of students of different countries with different statistical background (from very basic to more advanced). For this group of future academic researchers, it is particularly important to be educated about statistical modeling, and statistical reasoning. The purpose of the course is to teach students how to apply the principles of statistical design and analysis to empirical data, and how to interpret and report the results, following Wilkinson and the Task Force (1999). In addition to theoretical lectures, each student works on an individual research project and write a report in order to practice with real data.

We take the view that for valid statistical reasoning it is essential to pay attention to the philosophy of statistical reasoning. This means *more* instead of *less* theory about fundamentals of statistical testing. We therefore introduce in our course the basic principles of statistical inference which is contained in different approaches to hypothesis testing, namely Fisher's way of testing, Neyman-Pearson's approach, and Bayesian statistics. This view is in line with the pedagogical concept proposed by Haller and Kraus (2002) to teach Fisher's significance testing and Bayesian statistics together, warning against just replacing Fisher's testing by some other method.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

In this paper we will give a short outline of the design and content of the material for the hypothesis testing and statistical reasoning part of the course. We will discuss our experiences with the material in terms of students' ability to understand statistical reasoning, and motivate why and how we made adaptations to the components of the testing part.

DEVELOPMENT OF THE COURSE MATERIAL ON STATISTICAL REASONING

The design of the course material started in 2006, and was regularly adapted over the years. We distinguish four main development phases: 1. reading material, 2. attractive illustration, 3. extensions, and 4. exercises.

Reading Material

We selected the paper by Christensen (2005), which we found an accessible and clearly written paper that exactly served the purpose of the statistical reasoning module. In this paper the three main approaches of Fisher, Neyman-Pearson and Bayesian statistics are explained according to a simple, albeit rather abstract illustration. It provides the opportunity to repeat the basics of hypothesis testing, from p-value to likelihood ratio tests, from significance level to power. We were surprised that students found this paper hard to understand, and that they did not recognize the applicability of this paper to their own research project. Therefore the first adaption we needed to make was to work on an illustration of the theoretical concepts. We decided to develop an example that would appeal to the students through its familiarity as well as with which the different testing approaches could be demonstrated.

Attractive Illustration

The example was inspired by a board game called “Settlers of Catan”, very popular among students in the Netherlands. In this board game which is more than just a chance game, players throw a pair of dice in every round. We set up the game to consist of four students playing against each other. Three students are related to the godfathers of the three main statistical hypothesis testing approaches (i.e., we named them Ronald (Fisher), Jerzy-Egon (Neyman-Pearson), and Thomas (Bayes)). The fourth student is called Victor after Victor Lustig, the most famous fraudster of Europe in the twentieth century. As his names predicts, Victor wins the game. The other three students become suspicious, and want to test the fairness of the dice. This question is formulated as a simple null hypothesis and a simple alternative, for the testing of which a simple experiment with a single throw of the pair of dice is proposed. All three approaches are applied to this testing problem, rehearsing the theory as explained by Christensen. In the setup of the example, the three approaches lead to different results and conclusions.

As a preparation, students were asked to read the paper by Christensen, and to play the board game (at home). The students not only enjoyed playing the game, they also enjoyed the lecture with the story of the four students and the game setting. The attention in the class room during the lecture was very high. Students were intrigued and curious how to come to conclusions, and they were very active. They even had critical comments and questions about the simple alternative, and the single throw, motivating us to extend the illustration to more complex situations. However, the exam results showed that more than half of the students found the theory still hard to reproduce, and were still not able to apply it well. On the other hand, some students did show a good understanding of the theory. These are the more advanced students with a firmer statistical background, whereas most other students definitely had improved statistical reasoning skills by their efforts to understand the rather tough theory, convinced that the theory really matters.

Extensions

We were encouraged to extend the illustration to a more realistic test situation, considering more throws in the experiment, and formulating a composite instead of a simple alternative hypothesis. The complete illustration and its application (including the extensions) are published in Post et. al (2012). Due to time restrictions and due to the fact that for most students the simple illustration suffices, we do not discuss the extensions in detail during the lecture, but refer to this paper in order to satisfy the needs of the more advanced students.

Exercises

The final adaption we made was to ask students to come up with a particular game setting of their own and to formulate a specific alternative hypothesis. Students are encouraged to work together, and to apply the three main approaches to the setting specified by themselves. In a response lecture, we organize an open discussion among all students and lecturers about the chosen game settings and specifications of alternative hypotheses, as well as ways of interpreting the results and the corresponding conclusions. Thus, all elements of hypothesis testing, from research question, experiment, test results, to interpretations are repeated several times. The emphasis during the discussion is on different ways of interpreting the test results. Various ways to formulate conclusions for each of the main approaches are suggested. Students enjoy the response lecture and are actively participating, ready to learn more. Moreover, the approach with students defining their own research problems turns out to be rather successful. In the exam assignments, students show more ability to apply the different approaches. Consequently, we have the feeling that the material and approach to teaching hypothesis testing is in a more or less final state. That is not to say that we see no further possible developments.

CONCLUSIONS

The main aim of teaching different approaches to hypothesis testing was to increase students understanding of statistical reasoning in order to draw valid conclusions in a more nuanced way. This approach was successful: students have become more aware that there are different approaches, that all approaches are based on (slightly) different principles, with pros and cons, and that there is no best approach. Therefore, students are more capable of interpreting statistical significance in a valid way. What we did not achieve so far is that they are more aware of practical significance or relevance. In all research reports, significance (p-values) are reported and discussed. Although confidence intervals are reported, they are not fully interpreted. Therefore, we feel the need to consider extending the course material to estimation principles within the testing approaches. A natural way to do this would be by extending the illustration with estimation.

[*] both authors contributed equally to this work

REFERENCES

Christensen, R. (2005), Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121-126.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, Confidence intervals, and Meta-analysis*. New York: Routledge.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*(1). http://www.mpr-online.de

Post, W. J., van Duijn, M.A.J., & Boomsma, A. (2012). A tutorial on teaching hypothesis testing, *Model Assisted Statistics and Applications, 7,* 143-157.

Wilkinson, L., & the Task Force on Statistical Inferences (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist, 54*(8), 594-604.