Students' understanding of hypothesis testing: the case of the significance concept

Anne M. Williams
Centre for Mathematics and Science Education
Queensland University of Technology
Brisbane, Australia

Throughout introductory tertiary statistics subjects, students are introduced to a multitude of statistical concepts and procedures. One such term, *significance*, has been given considerable emphasis in the statistical literature with respect to the topic of hypothesis testing. However, systematic research regarding this concept is very limited. This paper investigates students' conceptual and procedural knowledge of this concept through the use of concept maps and standard hypothesis tests. Eighteen students completing a first course in university-level statistics were interviewed twice during a 14-week semester.

Significance is perceived as an important concept in hypothesis testing, given its attention in the related statistical literature. In this literature, statistical significance is said to be frequently misinterpreted by users of statistics (see Menon, 1993; West, 1990). The practice of characterising significance with descriptions (e.g., "very significant" or "clearly significant"), statements (e.g., there is more evidence against the null hypothesis at the 0.01 than the 0.05 significance level), or symbols (e.g., ** to indicate significance at the 0.01 level), has been criticised for reinforcing the apparent objectivity of significant results (Falk, 1986). Furthermore, the significance testing literature is critical of those who perceive that obtaining statistical significance is the main goal of statistical tests (e.g., Clements, 1993); that decision making is black-and-white, solely based on the significance level (e.g., Oakes, 1986); or that practical aspects are not worthy of consideration (e.g., Moore & McCabe, 1989). Given that even experts have been known to misinterpret significance (McCloskey, 1990), and that journal editors have at times advocated an emphasis on statistical, rather than practical, decision making in their publications (Menon, 1993), it is not surprising that errors persist.

The study described here begins to fill a gap in a vast literature in which almost no empirical research has been performed. This paper reports on a qualitative study investigating students' understanding of the significance concept in hypothesis testing, where hypothesis testing is confined to one- and two-sample t and z tests. More specifically, this study aims at investigating the conceptual and procedural knowledge of significance by students undertaking an introductory course in statistics at the tertiary level. Conceptual knowledge is defined here as the knowledge of concepts and their interrelationships. Knowledge of a concept involves being able to define it, as well as make other statements about it. The latter may entail being able to summarise the issues pertaining to it, provide an example, or describe its features, uses, or limitations. Procedural knowledge is the knowledge of symbols, rules, algorithms, and procedures. In particular, procedural knowledge involves the use of the concept in an applied situation such as an hypothesis test. Understanding the concept therefore means possessing both conceptual and procedural knowledge.

For the purposes of this investigation, significance is defined more succinctly, but aligned with discussion in the statistical literature. Significance has two aspects (i) the decision to reject the null hypothesis after a statistical test has been performed, and (ii) the evaluation of the decision. The first aspect is called statistical significance, the rejection of the null hypothesis when at least one of three conditions hold, namely: (a) the observed value arising from the test is more extreme than the critical value obtained from the appropriate tables, (b) the p-value is less than the level of significance, or (c) the p-value is sufficiently small. The second aspect of significance involves the evaluation of the result, the consideration of factors which are statistically relevant (e.g., sample size, power, effect size,

likelihood of replication), or practically relevant to the particular area being investigated. These factors may lead to practical significance, if the decision to reject is maintained. For first year students with limited knowledge, the evaluation step may be limited to considering the relevance of the statistical result to the question posed, and perhaps to considering the sample size.

Method

Data reported in this paper were collected through two clinical interviews conducted with 18 volunteer students enrolled in a first course in university-level statistics. Students were interviewed several weeks after the topic of hypothesis testing was introduced to them, and again after the final exam in the subject. They were asked to talk aloud as they completed a concept mapping task and standard hypothesis tests, and answer questions after their completion. Student responses were analysed in terms of the conceptual and procedural knowledge exhibited.

In the concept mapping task, concept names associated with hypothesis testing were typed on separate labels. Students were requested to place the labels on an A3 sheet of paper in such a way as to show the relationships between the concepts. Subsequent questioning drew discussion on these relationships, and on the concepts themselves. Given the above definition of conceptual knowledge, this task aimed mainly at investigating students' conceptual knowledge. The hypothesis tests were standard text book exercises with the question clearly defined and numerical information provided. The two main hypothesis tests, named the light bulb task (a two-tailed one sample z test), and the night shift task (a one-tailed two sample independent t test) aimed mainly at investigating students' procedural knowledge.

In the statistics class in which the above students were enrolled, the term *significance* was not used very frequently by the lecturer. However, the lecturer gave the students advice on the significance concept in a number of ways. The need for assessing the evidence to support or reject a null hypothesis was emphasised, noting, for instance, that the smaller the p-value, the stronger the evidence was against the null hypothesis and in favour of the alternative hypothesis. Further, the lecture notes stated that a small p-value may mean (i) that the null hypothesis is false; (ii) that the null hypothesis is true and an unusual event has occurred; or (iii) that the model being tested is not applicable. In addition, the lecturer advised students that it was good practice to quote the p-value with the decision, so that the reader may also evaluate the conclusion; and that a result may be statistically significant but not practically important. In the Study Guide for the subject, *significance* was listed as one of the key words associated with the topic of hypothesis testing. However, the main text book for the course made little reference to the term, stating in only one example that the z value could be significant after the rejection of the null hypothesis at a particular level of significance.

Analysis

Analysis of students' responses during the first interview (beginning Week 11 of the semester) is followed by an analysis of students' responses during the second interview (after the final exam). For each interview, the concept map responses are examined prior to the hypothesis test responses. In this analysis, a student was considered to have demonstrated conceptual knowledge of the significance concept if the concept was defined or explained correctly, or linked with other concepts correctly. Demonstration of procedural knowledge of statistical significance was acknowledged when a student (i) reached a decision to reject or not reject the null hypothesis after performing a statistical test, and recognised significance or non-significance, (ii) discussed the steps used to obtain significance or non-significance (e.g., significance occurs when the p-value is less than the significance level), or (iii) evaluated the statistical decision. When at least one of these three steps was performed, yet the term *significance* was not mentioned, it was acknowledged as a demonstration of implicit procedural knowledge.

INTERVIEW 1 - WEEK 11

Concept Mapping Task: In Interview 1, 5 students used the significance label on their concept maps; 13 did not.

Definitional statements: Efforts to explain the significance concept during the concept map task fell into four main types: correct use, vague description, confusion with significance level, and incorrect use.

First was the correct use of the significance concept. Only Rhys was in this category. He suggested that rejection of the null hypothesis occurred with low values of the p-value, and that a p-value "closer to .5" was non-significant. His explanation, procedural as well as conceptual in nature, referred to the p-value method of hypothesis testing, and was the most advanced explanation offered by a student using the term.

Second, students used descriptions in the form of adjectives, phrases, or sentences, which gave a sense of where significance fitted into hypothesis testing. In general, these were vague or inadequate in demonstrating understanding. For instance, two students (Cheryl, Lisa) used *significance* to describe testing, decision, or results. Cheryl made fleeting reference to "significance in the testing" and "significance of decision". Lisa mentioned the need for "significant results." Neither provided additional information. Further, two students (Elvie and Koby) attempted more prolonged descriptions of significance. Elvie remarked, "I was just talking about like whether your results are statistically significant, because you can get results and like put in a probability and stuff and they look good but they might not be worthwhile kind of thing." Her interpretation of statistical significance hinted at finding a p-value. Koby, having stated that the z or t value had a certain significance, continued, "an amount of significance, yeah, and it affects your decision." Explanation of how significance affected the decision was not forthcoming. While the above students acknowledged the role of significance in testing, decision, results, or z or t values, none offered a clear explanation.

Third, students (e.g., Margaret, Phillip, Rhys, Kylie, Lorraine) interchanged the term with significance level, sometimes resulting in the interpretation of significance as a probability, or as a level for rejection. Several examples demonstrate such misinterpretations. Margaret stated:

> well significance is the, that's the probability that you want to be right, like if you want to be 99 percent sure that you're going to be correct, you need an alpha value of, you look up your little table and find nought point nought nine five or something like that.

Kylie explained, "significance is the, it's sort of interconnected with the significance level, it's the um level that you're prepared to reject it with, sort of the same thing." Consistent with his interpretation, Phillip wrote the alpha symbol α, next to the significance label on his concept map.

Fourth, significance was completely misinterpreted. As an illustration, Michael thought that significance referred to the significant figures in scientific notation for large and small numbers.

Thus, while most students mentioned the term *significance*, little understanding of the concept was manifested in their explanations. Only 1 student offered a correct explanation. Alternatively, one of three things occurred: either inadequate and vague connections were made with testing, decisions, results, or z or t values; the term was interchanged, in both name and meaning, with significance level; or it was completely misinterpreted.

Other statements: In addition to the above explanatory statements, several statements made during the concept mapping task described the steps involved in concluding rejection of the null hypothesis. Kylie's and Lorraine's comments illustrate. Kylie noted, "if your p-value is less than .05 this means you can reject your null hypothesis with 95 percent significance level." Without mention of the word *significance*, and despite using a numerically incorrect value for significance level, she conveyed an understanding of one of the conditions leading to statistical significance. Referring to a p-value falling in the critical rejection region, Lorraine stated, "then if it's significant, go to decision, then you make a decision." Given Lorraine's responses on other tasks, her statement, though not as succinct as Kylie's, portrayed a similar meaning. Kylie's protocol described a comparison of the p-value and the significance level, whereas Lorraine's implicitly compared areas represented by p-value and critical region. Both students, as well as Rhys, displayed procedural knowledge of the significance concept. In the case of Kylie, this knowledge was implicit.

Hypothesis Tests: During the performance of the hypothesis tests, the term *significance* was not mentioned by a single student, and there was little overall evidence of procedural knowledge. Only Phillip, Kylie and Lorraine completed an hypothesis test, the light bulb task, during this interview. Phillip rejected the null hypothesis using the critical value and confidence interval methods, and Kylie used the critical value method, rather than the p-value method suggested by her previous quote. Lorraine used the p-value method. Only Phillip and Lorraine related their statistical conclusion back to the original question, and all 3 students had to be prompted for further discussion. Their procedural knowledge of significance was therefore implicit. Rhys, who had previously demonstrated procedural knowledge of significance, could not use it in the hypothesis tests. Thus, it appears that procedural knowledge of significance was evident in only a few students, and even then it was implicit.

The evidence above suggests that, at the time of the first interview (beginning Week 11), significance was not a well known term. Only 11 students mentioned the term during the course of the interview. In general, the concept was either linked vaguely to testing, decisions, results, or z or t values, or was used synonymously with significance level. Several students demonstrated implicit procedural understanding of significance when they described the conditions for decision making, or rejection of the null hypothesis. Only 2 students reached a conclusion to reject the null hypothesis during the performance of their hypothesis tests. This, too, was classed as demonstrating implicit procedural knowledge. Hence, conceptual knowledge and procedural knowledge were limited overall in both the explicit and implicit senses.

INTERVIEW 2 - AFTER FINAL EXAM

Concept Mapping Task: At the second interview, 9 students (compared to 5 previously) used the significance label on their concept maps; 9 did not.

Definitional statements: Explanations of the significance concept were classified as in Interview 2: correct; vague; interchanged with significance level; and misinterpreted. No statements fitted the first or fourth classifications.

Explanations of the second type (5 students) were illustrated through the brief or inadequate descriptions by Cheryl, Lorraine, and Koby. For example, Cheryl stated that significance related to significance testing. As before, she could elaborate no further. Lorraine reasoned that significance was "how significant the test is." When asked how that could be judged, she was unable to explain. Later she stated that, "if it lies in the accepted or rejected region and then you, if you want to test the significance of it there." While the meaning of both remarks remains unclear, the latter appears to imply that significance has something to do with acceptance or rejection regions. Koby said, "the statistic might be very significant if you have a high p-value, which means the null hypothesis is true." When asked to explain "very significant", she continued, "well if it's high, the p-value's high, there's a high significance level ... [like] .8, anything over .5 is reasonably high." Koby had interpreted the relationship with p-value in the wrong direction, but perceived p-value as

similar to significance level. Thus, explanations of this type, though generally connected to the final decision making concepts, were usually inadequate.

In explanations of the third type (3 students), the term *significance* was again used interchangeably with significance level, resulting in interpretations of significance as a level, a probability value or an area. Each interpretation is demonstrated respectively in the protocols of Dominic, Kylie and Phillip. For example, Dominic said, "it's [significance] the value or level I think it was, to determine the level at which you reject or keep your null hypothesis." Kylie explained, "that's [significance] just basically at 5 percent significance, it means you've got a 5 percent chance of a Type I or Type II error." More expansively, but similarly, Phillip stated:

> I think the significance area is just an area so much to either side of the bell-shaped curve, if you're using a two-sided z statistic then it's 5 percent significance or 10 percent significance, you wouldn't go past 20 percent.

Phillip's explanation went further than Dominic's or Kylie's, by including a connection with the regions at the ends of the distribution curve, which Phillip had previously referred to as rejection regions. Again, with consistency, he wrote the alpha symbol, α, next to the significance label. Summarising students' explanatory statements about significance, none clearly showed a full understanding of the concept. Students continued to provide definitional statements which were vague or inadequate, or confused the terminology and meaning of significance with significance level. However, several did make the connection with tests, values, or rejection regions.

Example statements: In Interview 3, one attempt was made to illustrate significance by an example. Intending to explain the link between significance and probability, Rhys stated:

> significant or not, if you're looking at the male female ratio and seeing if it's dependent on AIDS or something like that, you can test to see if it's significant or not, making the stratified allocation thing to see if it'd be worthwhile.

His meaning was unclear, and the example failed to achieve its purpose.

Other statements: During the concept mapping task, in addition to those statements summarised above, several described the steps involved in obtaining rejection or non-rejection of the null hypothesis. As before, this was acknowledged as demonstrating implicit procedural knowledge of the significance concept, as no mention was made of the term *significance*. Seven students explained how to reject the null hypothesis using the critical value method, and eight offered explanations via the p-value method. Examples of these statements are reflected in the following extracts. Pointing to the tails on the distribution graph, Margaret stated that, "you reject your null hypothesis if it [value from test statistic] falls inside the critical region [defined by the significance level]." Karl remarked, "reject it if the p-value is less than the significance level of .01 or .05." In the hypothesis tests, six of these students could not apply this knowledge. A summary of responses on these tasks follows.

Hypothesis Tests: During the performance of the hypothesis tests, Kylie was the only student to mention the word *significance*. While completing the light bulb task, she referred to "5 percent significance", "significance of 5" and "35 percent significance level", thus maintaining consistency in her exchange of the two terms, *significance* and *significance level*. Furthermore, after standard statistical procedures, 7 students (compared to 3 previously) reached a statistical conclusion. Five used only the critical value method, one only the p-value procedure, while the seventh student eventually reached his conclusion via the confidence interval method on the light bulb task, and via the critical value method on the night shift task. These conclusions were recognised as implicitly establishing statistical significance. In addition, only 5 of the 7 students continued on to relate their conclusion (rejection or non-rejection) back to the original question. Invariably, further evaluation of the decision was forthcoming only after prompting by the researcher, implying that practical

significance was not a normal consideration for these students. Thus, for the hypothesis tests, students' procedural knowledge of the significance concept was recognised only implicitly in the few students who reached a statistical decision. Little emphasis was placed on practical considerations without prompts from the researcher.

Summarising the findings of Interview 3, it is clear that even at the end of the semester, students' understanding of the concept of significance remained poorly developed in the explicit sense. Several did not even recognise the concept name, and use of the significance label on the concept maps was not an indication of understanding the concept. Several students still thought of significance as significance level.

The major findings in these interviews were first, that several students did not recognise the concept name, and few students could adequately explain significance, or its role in hypothesis testing. Students' conceptual knowledge was confined to brief, vague, inaccurate, or inadequate descriptions, or loose connections with rejection or non-rejection. Through misinterpretation, incorrect links were sometimes made to significance level. Second, most students (12 of 18), given their omission of the *significance* term, exhibited implicit procedural knowledge of the significance concept, either through their correct explanations of the p-value or critical value methods or through their statistical conclusion at the end of a standard hypothesis test. However, many who were able to describe the process to obtain a statistical conclusion could not actually perform the process. Third, students placed more emphasis on statistical measures (statistical significance) than practical ones (practical significance). Even the statistical conclusion was not always related back to the original question that was posed. Therefore, in general, evidence of explicit conceptual and procedural knowledge appeared to be limited by the widespread lack of familiarity with the term itself, and the inability to perform hypothesis tests.

Discussion

Each of the major findings described above will be discussed in order.

The first finding consisted of three parts. First, several students did not recognise the concept name, and few could explain the concept. This finding is not of major importance, because the actual use of the term is not essential to obtaining or interpreting a statistical result. It is the ideas behind obtaining significance that are important, not necessarily the term itself. In any case, those who attempted an explanation appeared to have had some idea of the concepts with which significance was related. Furthermore, in this study, the students' text books barely mentioned the term, and the lecturer did not emphasise it often. Second, students' conceptual knowledge was limited to vague descriptions, or loose connections. Throughout both interviews, students appeared to have difficulty expressing themselves statistically. Either they lacked the statistical language, or the statistical knowledge to convey the intended meaning. Most probably the cause was a combination of the two. Third, students confused significance with significance level. Three reasons are offered for this occurrence: the concept names are similar; the interpretation of statistical significance through the p-value and critical value methods involves the significance level; and the low emphasis given to this concept in lectures may explain students' inability to distinguish between them.

The second finding was that students' procedural knowledge was implicit, given their omission of the term *significance*. As indicated above, more students could actually describe the process leading to rejection or non-rejection than could actually successfully perform it. Hence, it must be concluded that reaching a statistical decision is difficult for introductory students. Yet, hypothesis testing is a topic in most introductory statistics subjects, and these students had completed a first course in statistics. Several factors may provide a rationale for this difficulty. First, hypothesis testing is an involved procedure. The translation of data into statistical terms, interpretation of formulas or symbols, and the actual techniques are possible sources of confusion. Second, hypothesis tests performed by hand are tedious and time-consuming, and many mistakes of a mathematical or statistical nature

may be made along the way. Third, given the timing of the first interview, a few weeks after the introduction of hypothesis testing, students' inability to perform a test could be explained by a lack of practice. This would support the finding that several students could explain the process, yet not actually do it. Given the timing of the second interview, after the final exam in the subject, lack of performance could be explained by the fact that there was time to forget the procedure (if it was ever known). Forgetting was aided by the fact that later in the semester, emphasis was on the interpretation of computer printouts using the p-value (e.g., multiple regression). Hand, rather than computer, calculations were necessary for the completion of the hypothesis tests in this study.

The third finding was that statistical conclusions were apparently more important than other considerations. Evaluation of the statistical decision is an important step in hypothesis testing, and several reasons are offered for the failure of students to either link the statistical conclusion back to the question posed, or evaluate the decision. These reasons follow the previous rationale, namely that the complexity of the process and the tedium of hand calculations override additional considerations. The biggest hurdle is reaching a statistical conclusion, and the real meaning of the original question may be forgotten in the process. In addition, typical text book examples do not always stimulate deep reflective and interpretive thinking. Furthermore, many lecturers do not encourage it.

Teaching should attempt to overcome the problems outlined above. If the term itself is used, then it must be mentioned frequently by lecturers in conjunction with rejection and interpretation of statistical results. The encouragement of group discussions, or oral presentations, may foster students' use of statistical language. If the term is not emphasised, then it is the integral ideas of significance that must be accentuated. First, with respect to the difficulties in reaching a statistical conclusion, there are many good statistical computer packages available, which perform speedy numerical calculations, deemphasising the use of confusing statistical symbols and formulas. Once an understanding of the process is formed, students may be curious to learn the statistical rationale and techniques behind it. Computers also facilitate the exploration of statistical conclusions under different conditions, for example sample size. However, the success of computer packages always depends on the lecturer's ability to integrate the important ideas. Second, with respect to the evaluation of statistical results, interpretive skills can be developed through well-designed projects and examples that are applicable to students' fields of interest. Typical text book questions limit students' statistical experience. The ICOTS conferences, in particular, provide ideas for classroom use.

Many of the above comments have been made before by others. However, this study is the first of its kind to provide empirical, rather than anecdotal, evidence of the problems students have with the significance concept.

References

Clements, M. A. (1993). Statistical significance testing: Providing historical perspective for Menon's paper. *Mathematics Education Research Journal*, *5*(1), 23-27.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 93-96.

McCloskey, D. M. (1990). Formalism in the social sciences, rhetorically speaking. *The American Sociologist*, *21*(1), 3-19.

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

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. New York: W. H. Freeman and Company.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chichester: John Wiley & Sons.

West, L. J. (1990). Distinguishing between statistical and practical significance. *Delta Pi Epsilon Journal*, *32*(1), 1-4.