STUDENTS' UNDERSTANDING OF THE SIGNIFICANCE LEVEL CONCEPT

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

Throughout introductory tertiary statistics subjects, students are introduced to a multitude of new terms for statistical concepts and procedures. One such term, significance level, has been considered in the statistical literature. Three themes of discussion relate to this concept - the problem of interpretation (and misinterpretation), the selection of an appropriate level, and the evaluation of results based on significance level. However, empirical research regarding this concept is very limited. This paper reports on a qualitative study which used concept maps and standard hypothesis tests to investigate students' conceptual and procedural knowledge of the significance level concept. Eighteen students completing an introductory tertiary statistics subject were interviewed after their final exam in statistics. Results showed that many students did not have a good understanding of the concept.

In the statistical literature, the significance level concept has been discussed in three ways. First, it has been revealed that significance level is frequently misinterpreted (see Birnbaum, 1982; Falk, 1986). One common misinterpretation is that significance level is the probability of being wrong, that is the probability of the null hypothesis being true once the null hypothesis is rejected. Significance level really means the probability of rejecting the null hypothesis when the null hypothesis is true. The two probabilities are quite different in value. Second, the process of selecting a suitable level of significance can be a complex procedure, requiring reflection on a number of associated concepts and conditions, such as the balance of Type I and Type II errors, and the circumstances of the problem. Labovitz (1970) offered a list of 11 considerations. Third, evaluation of the results of hypothesis tests, independent of the conventional levels of 0.001, 0.01, and 0.05, has been advocated as an alternative to the black-and-white decision making processes of the past (see Clements, 1993; West, 1990).

While the statistical literature is vast, little of it is empirical. This paper reports on a qualitative study investigating students' understanding of the significance level concept in hypothesis testing (one- and two-sample *t* and *z* tests only), and is part of a larger doctoral study. More specifically, this paper aims at investigating students' conceptual and procedural knowledge of significance level. Conceptual knowledge is comprised of the knowledge of concepts (e.g., definition, example, issues, language, representations) and their interrelationships. Procedural knowledge is the knowledge of statistical symbols, rules, algorithms, and procedures, particularly in the context of performing hypothesis tests. Understanding means having both conceptual and procedural knowledge.

DATA COLLECTION

The data for this paper were collected through individual clinical interviews conducted with 18 volunteer students enrolled in an introductory subject in university-level statistics. The subject could be described as traditional in both teaching and topic. This meant that hypothesis testing was preceded by descriptive statistics, probability, and interval estimation, and introduced in lectures through the critical value method with examples worked by hand, then later by computer.

Interviews took place after the final exam in the subject. Students were asked to talk aloud as they completed a Concept Mapping [CM] task (used mainly for assessing conceptual knowledge) and two formal Hypothesis Testing [HT] tasks (used mainly for assessing procedural knowledge). In the CM task, concept names associated with hypothesis testing were typed on separate labels, and students were requested to place the labels on an A3 sheet of paper so as to show the relationships between the concepts. Subsequent questioning by the researcher drew discussion on these relationships, and on the concepts themselves. The HT tasks were two standard text book exercises with the question clearly defined and numerical information provided. The first was a two-tailed one-sample *z* test, the second was a one-tailed two-sample independent *t* test. This paper analyses student responses on these tasks in terms of the conceptual and procedural knowledge exhibited.

ANALYSIS

The following analysis inferred students' understanding of the significance level concept from their statements and actions during the performance of the tasks described above.

Conceptual Knowledge

Definitions: Definitions were provided by less than half of the students, and they were expressed in several ways - as a level for decision making, a percent or percentage area, a variance, or a measure of significance, confidence or error.

As a level for decision making, a typical comment was: “it's just the rejection levels, if the p-value is less than the significance level then you reject the null hypothesis, and if it's greater than you accept it”. This student had thus described the p-value method of hypothesis testing.

As a percent or percentage area, another student stated, “so the significance level is just that 5 percent area, that two and a half percent area if you’re using a two-sided graph, where you have said you can reject the null hypothesis I think.” He was remembering the distribution graph. An additional student’s less precise statement affirmed: “it’s a percent that you put on the graph thing.”

From a variance perspective, another student gave a lengthy explanation of his perception of significance level. He said:

> I would say the significance level is ultimately say a variance ... a standard deviation or something like that, say if we had a line ... on the x and y axis again, and we had a completely horizontal line which only ever passes through the y axis ... the significance level is how much we can believe that line which never crosses through the x axis actually falls within a place that we’re looking for, we believe that it’s going to be a value to 3 on the y axis, and the significance level would say determine whether or not our t value has actually fallen between a certain range.

Despite misinterpretation, one idea was correct - that the significance level determined a range of values in which the *t* value should or should not lie.

Another student tried to explain significance level in terms of significance, saying “it’s whether you can say it’s significance or not, the p-value, like is 1 percent much different from zero compared to 7 percent to zero, and 1 percent that’s practically zero and 7 percent that’s further away.” While the first part of this statement was correct but vague, the second appeared to be an attempt to explain the p-value method of hypothesis testing.

Significance level as error was illustrated in the following statement, “significance level is like for your alpha value, like the probability of making an error.” As this student could not define Type I error, it is unlikely that the reference was to Type I error. Another student remarked, “alpha level, I know that’s your mistake kind of thing.” These two quotes were vague, and reflected the common misinterpretation mentioned in the introductory paragraph of this paper.

Remaining definitional statements, usually expressed numerically, made reference to significance level as a measure of confidence. A typical statement was: “the significance level is whether you’re gonna prove it to 90 percent or 95 percent, 95 percent confidence interval”. Evident in explanations of this type was the common use of large “complementary” percentage values, which demonstrated an intuitive sense of a relationship between significance level and confidence, with larger percentage values representing greater confidence. Alternatively, as students generally did not define confidence interval correctly, they may have remembered the distribution graph, misinterpreting the middle area as the confidence interval, and the remaining area as the significance level.

Acknowledgement of numerical values: Most students acknowledged the conventional significance levels of 1, 5 or 10 percent (or their “complementary” larger levels of 99, 95 or 90 percent), with the 5 percent level being mentioned more frequently.

Recognition of alternative representations: One student used the symbol during the CM task, but he wrote it next to the significance label and called it sigma, and five acknowledged it in the other tasks. Four students referred to the significance level as *alpha*. Four students discussed the distribution tail percentages for one- or two-tailed tests.

Relationships with other concepts: In students’ statements, significance level was mainly linked to the critical region, rejection/non-rejection, and p-value concepts, but these links were each made by a small number of students. Despite the large numerical values and the difficulties of expression, most conveyed the correct ideas about the link between significance level and critical region. These included ideas that: the significance level determined the *z* or *t* values for the critical rejection regions; these values could be obtained from the tables and represented diagrammatically; and a change in the value of the significance level leads to a change in the rejection areas.

The link with rejection or non-rejection, demonstrated in several of the above protocols, was expressed in terms of the critical regions or the values denoting them. Generally, the correct idea was conveyed. However, the links with p-value were not always precise. For example, one student stated, “it’s from your significance level that you say whether that p-value is good or not ... from p-value you get significance level, and you see if that p-value is significant or not to base your decision on.” Like several other students, he only partially explained the p-value method. Sometimes the comparison was applied inconsistently. Relationships between significance level and other concepts such as confidence interval, Type I error, and Type II error were either poorly expressed or incorrect. In general, students who could describe the role of significance level in the decision-making process, as well as provide a correct definition, performed better procedurally.

Procedural Knowledge

Representation of percentage values on distribution diagram: Only one student drew a distribution graph during the HT tasks. On each, she acknowledged the correct percentage values in the tails for the 0.05 level.

Use of numerical value: Nine students used (or implied) numerical values such as 0.01 or 0.05 on the HT tasks.

Rejection of an hypothesis at a particular significance level: One student used the significance level in her final statement to reject the null hypothesis at the 5% significance level.

Use of statistical tables to obtain critical values: Students experienced different degrees of success when they attempted to use statistical tables to find critical values or use confidence intervals. Three were successful; one had difficulty reading the *z* tables but could read the chi square tables (used inappropriately on the second HT task); and two were unsuccessful, one of them guessing a critical value.

p-value method: One student incorporated the p-value method in her first HT task solution. Another student, who could explain the p-value method, tried unsuccessfully to remember how to obtain a p-value. The others used different methods in their solutions.

SUMMARY AND DISCUSSION

Several important points can be made about students’ learning of the significance level concept from the study results. First, the tasks in this study elicited comment on only one of the three discussion themes from the statistical literature, that of interpretation. It seems that standard hypothesis tests do not promote the consideration of the other two. Second, analysis of the task responses helped to determine some constituents of students’ knowledge of the concept. These were evident in the categories used in the Conceptual Knowledge and Procedural Knowledge sections. Third, given the low number of students contributing to the categories, it must be concluded that most student’s knowledge of the significance level was limited. However, correct definition and accurate description of its role in the decision-making process seemed to be associated with better procedural performance. Fourth, the study highlighted several problems associated with students’ knowledge of the concept. It was clear that students had difficulty defining it. In particular, it was misinterpreted as variance, significance, confidence, or error. The common use of the "complementary" percentage values demonstrated an intuitive misunderstanding, or a confusion of meanings. Some students possessed a vague notion of where the significance level fitted in relation to other concepts (e.g., p-value) but lacked a precise knowledge of the relationship. Moreover, alternative representations were not well known, and few students demonstrated an ability to use statistical tables. One overriding problem was the conveying of statistical ideas. Many ideas were vague, incorrect, or clumsily expressed, often due to a lack of knowledge.

Lecturers in statistics need to be mindful of the difficulties students experience in learning statistical concepts such as significance level. The very nature (complex, abstract) of the concepts exacerbates the process, and what seems simple to an experienced user may be a massive hurdle for a novice. To assist the learning process, lecturers may need to refine their own understanding of each basic concept, highlight its most important aspects, and provide activities which develop sound understanding and lessen the chances of misinterpretation. As shown above, standard hypothesis tests do not seem to do this. For example, well-designed and interesting projects may encourage students to consider all areas addressed in the literature. Group discussion and oral presentations may force students to crystallise their ideas and improve their statistical expression. Computer-worked hypothesis tests may facilitate the exploration of ideas through reflection and evaluation, and promote the development of conceptual knowledge. The more time-consuming hand-worked tests do not encourage these.

REFERENCES

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics, 4(1),* 24-26.

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.

Labovitz, S. (1970). Criteria for selecting a significance level: A note on the sacredness of .05. In D. E. Morrison, and R. E. Henkel (Eds.), *The significance test controversy* (pp. 166-171). Chicago: Aldine.

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