Students' Understanding of Test Statistics in Hypothesis Testing

Annie Burns-Childers
University of Arkansas Little Rock

Darryl Chamberlain Jr.
Georgia State University

Aubrey Kemp
Georgia State University

Leslie Meadows
Georgia State University

Harrison Stalvey
University of Colorado Boulder

Draga Vidakovic
Georgia State University

Hypothesis testing is a key concept included in many introductory statistics courses. Due to common misunderstandings of both scientists and students, the use of hypothesis testing to interpret experimental data has received criticism. This paper describes preliminary results obtained from a larger study designed to investigate introductory statistics students' understanding of one sample hypothesis testing. APOS theory is used as a guiding theoretical framework. Preliminary data analysis focused on two students' distinctions between test statistics when performing hypothesis tests on real world data. The results suggest a significant difference in these two students' understanding, one being identified having an action conception while the other had an object conception of hypothesis testing as situated in the study.

Key Words: Hypothesis Testing, Statistics, Test Statistics

Introduction

The use of statistics is crucial for numerous fields, such as business, medicine, education, and psychology. Due to its importance, statistics education has seen rapid growth over the past three decades (Vere-Jones, 1995). In the United States today, the Common Core State Standards for Mathematics calls for students to "understand statistics as a process for making inferences about population parameters based on a random sample from that population" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010, p. 81).

One method of making inferences (formulating a conclusion) about a population is hypothesis testing, which is widely used by researchers in the social sciences and is a key concept included in many introductory statistics courses. However, the use of hypothesis testing to interpret experimental data has received criticism (Nickerson, 2000) due to the common misunderstandings of both scientists and students when using this method (Batanero, 2000; Dolor & Noll, 2015; Vallecillos, 2000). LeMire (2010) defends the use of hypothesis testing and provides a framework that can be used to revise instructional content with the goal of further developing student understanding. This is an indication that on-going research should investigate students' understanding and curriculum effectiveness in light of the critiques surrounding methods of inference.

In this preliminary research report, we focus our attention on the following research question: What are students' understandings of hypothesis testing in two distinguished real world situations?

Theoretical Framework

The guiding theoretical framework for our larger study is APOS Theory (Asiala et al., 1996). APOS Theory is a framework which models an individual's mathematical conception using Actions, Processes, Objects, and Schema. An Action is an externally driven transformation of a

mathematical object (or objects). An Action can be described as an individual needing an external cue to follow, such as a step-by-step example. Once Actions are repeated and reflected on, an individual can start to interiorize them to become a Process. A Process no longer requires step-by-step external cues. An individual is now able to internally imagine the steps in a transformation, without having to actually perform them. When an individual is then able to see the Process as a totality, is aware that transformations can be applied to it, and can construct these transformations, then the Process has been encapsulated into an Object. The collection of all mental constructions of Actions, Processes, and Objects forms an individual's Schema of a particular mathematical concept.

Methodology

The focus of our study is on university students who are enrolled in an introductory statistics course based on the emporium model. The emporium model, originated at Virginia Tech, includes key components of "interactive computer software, personalized on-demand assistance, and mandatory student participation" (Twig, 2011, p. 26). For this particular institution, each week students were required to spend three academic hours in a computerized mathematics lab, as well as attend one academic hour class each week with an instructor. The time in the mathematics lab was spent actively learning using the mathematical software MyStatLab by Pearson. Students were also engaged in activities such as reading and discussing about the subject content with their peers, graduate and undergraduate lab assistants, and instructors.

Elementary Statistics Using Excel was the textbook used in the course, written through Pearson and adapted specifically for the university (Triola, 2014). The textbook describes a test statistic as "a value used in making decisions about the null hypothesis," (p. 415). While assuming the null hypothesis is true, a test statistic is found by converting a sample statistic, whether that is a sample proportion or a sample mean, to a standardized score. As students for the course are required to calculate a *p*-value for most hypothesis tests, the text describes the *p*-value as the "probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true," (Triola, 2014, p. 416). Although we discuss test statistics as one mathematical term or value, there is a distinction needed to be noted between test statistics calculated from sample proportions versus sample means. Specifically, in our study, the distinction needs to be made between the normal distribution and the Student's *t* distribution.

For proportions, students always assume a normal distribution, and thus calculate test statistics which are *z*-scores. When calculating a test statistic for a sample representing a population mean, we are referring to a standardized value that represents the extremeness of your sample in regards to what is expected. For means, students learn about test statistics in hypothesis testing for the normal distribution (*z*-scores) and the Student's *t* distribution (*t*-scores). Although these distributions appear similar, the distinction occurs depending on what we know about our sample. In particular, if we know the population standard deviation, then we know how the data is spread out and can use the normal distribution (*z*-scores). However, if we do not know the population standard deviation, but can estimate it with the sample standard deviation, we can use the Student's *t* distribution to estimate how the population is spread. For this reason, *t*-scores are greater than or equal to *z*-scores for the same value of *n* (equal as *n* approaches infinity) in order to overcompensate for the lack of knowledge of the distribution (see Figure 1).

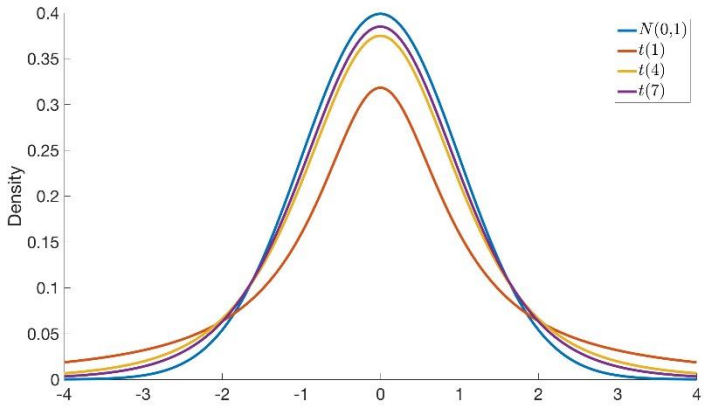

Figure 1. Graphical Representation of Normal Distribution and Student's *t* Distribution

Data was collected from these classes during the Fall 2014 and Spring 2015 semesters. Data consists of all students' work (more than 1,500 students) on hypothesis testing, including homework, quizzes, and tests. Semi-structured interviews took place with a targeted group of students of different capabilities. For this preliminary report, we will focus on two of the nine problem solving/interview sessions as the primary source of data. During the problem solving sessions, each participant worked alone on two hypothesis test questions. Following this, they participated in a semi-structured interview to further elaborate over the problems they solved. The first question asked the student to conduct and interpret a hypothesis test for a single population proportion. The second question asked the student to conduct and interpret a hypothesis test for a single population mean. The questions were as follows:

1. In a recent poll of 750 randomly selected adults, 588 said that it is morally wrong to not report all income on tax returns. Use a 0.05 significance level to test the claim that 70% of adults say that it is morally wrong to not report all income on tax returns. Use the *P*-value method. Use the normal distribution as an approximation of the binomial distribution.
2. Assume that a simple random sample has been selected from a normally distributed population and test the given claim. In a manual on how to have a number one song, it is stated that a song must be no longer than 210 seconds. A simple random sample of 40 current hit songs results in a mean length of 231.8 seconds and a standard deviation of 53.5 seconds. Use a 0.05 significance level to test the claim that the sample is from a population of songs with a mean greater than 210 seconds.

Students had seen these exact questions on their homework and quizzes when using the MyStatLab software. The only difference was that for the problem solving/interview sessions they did not have multiple choice/drop down menus as an option to answer the questions. Since there was active learning associated with these concepts in the mathematics lab and in class, students were expected to know how to conduct and interpret hypothesis tests for both questions, and in particular, they were expected to know to use the normal distribution to find the test statistic for Question 1 and the Student's *t* distribution to find the test statistic for Question 2.

Preliminary Results

Our preliminary results indicate that students are not always able to make distinctions between test statistics. Several students used the normal distribution to find the test statistic for not only Question 1, but also for Question 2. For those that did use the normal distribution when calculating the test statistic for Question 1 and the Student's *t* distribution for Question 2, some still could not articulate why they used two different test statistics or the relationship between the two test statistics. A common approach to finding the test statistic was to first look for key words, which suggests an Action conception of test statistic. This worked for some, however, others got confused with the language in the problems. For this preliminary report, we will describe the different approaches utilized by two students, Steve and Lana, and discuss the interpretations that helped them decide which test statistic to use.

Steve

For Question 1, Steve used the common approach of identifying key words to recognize which test statistic to use. He explained, "I immediately thought of these two formulas, and at first I wasn't sure which one to use, and then I was like, oh wait, there's no *x*-bar or mu or standard deviation. So that makes it pretty easy." He used a system of elimination to decide which formula not to use. Even though he correctly identified the test statistic, he went on to say that this is "just a formula that I've learned like any other" and that he "doesn't understand why we use that formula, other than we just use it". He concluded his explanation stating that Question 1 used a *z*-value because the problem was of proportions, what appears is likely a memorized rule applied to the problem.

For Question 2, Steve recognized that it was now a question of means, but mentioned that "it's basically the same problem" other than this discrepancy. Steve identified the distribution as normal because the question explicitly stated that "a simple random sample has been selected from a 'normally distributed' population". Based on the language in the problem, Steve associated the question with a normal distribution. He further explained that he would not know how to deal with a non-normally distributed population, that he could not recall learning anything other than a normal distribution, and that he did not know much about a Student's *t* distribution. Ironically, later in the interview, when asked if his answer was a *z*-value or *t*-value, he responded that it is "a *t*-value because you're testing means".

For Steve, it appeared as though he was basing his ideas of whether to use the test statistic associated with the normal distribution or the test statistic associated with the Student's *t* distribution off of key words found in the problems. When the problem for Question 2 included the phrase of "normally distributed", he used this to identify the question as a normal distribution. What could have been a result of memorized rules, led to a misconception of when to use the *z*-test and *t*-test. Steve appears to exhibit an Action conception of the test statistic as he relied on external cues, such as the key words in the problem, to identify which test statistic to use.

Lana

Lana, for Question 1, used the normal distribution to find the test statistic based on the fact that the problem was about proportions. To explain the test statistic, she initially explained how she was picturing the "big curve, the bell curve, and I'm picturing the test statistic is where the point that falls on there … so this is the mean right in the middle, and the test statistic is one side of it, saying this is how far away from what they are saying is the mean, this is what the mean of this, I guess that's what I'm thinking". Comparison of the mean and test statistic possibly indicated an

Object conception of test statistic. After her explanation, she then drew a picture to illustrate her thinking (Figure 2) of a graphical representation of the normal distribution.

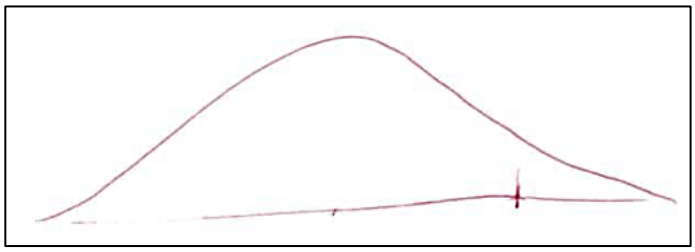

Figure 2. Lana's Graphical Representation of the Normal Distribution

For Question 2, initially, Lana in her written work used the normal distribution to find the test statistic. During the interview she became worried when asked if the question was a *z*-test. She mentioned that she remembered from high school to use a *t*-score if the sample is of 30 or less. After prompting, she realized she should have used a *t*-score, however, she also immediately recognized that her *z*-score would "probably not" be very different from the *t*-score. She knew that in the relationship between the two test statistics, "they are really close together".

Lana was able to picture in her head a graphical interpretation of the test statistic, and was also able to identify a relationship between the two test statistics, suggesting an Object conception. This was more than what most other students were capable of interpreting.

Concluding Remarks

Our data suggests that one area students have trouble with when conducting hypothesis tests on real world data is correctly identifying and interpreting which test statistic to use. Students with an Action level of understanding of hypothesis testing relied on memorizing facts and identifying key words. Without a deeper understanding, misconceptions emerged. Students who appear to base their understanding off of ideas and concepts, not memorized rules, seem to have a better grasp of the relationship between the two test statistics, and when and why to use each one. It is also suggested that not having the MyStatLab multiple choice/drop down solutions to choose from could have possibly influenced the students' responses in the analysis. The next step in analysis will be to further identify misconceptions and understanding related to hypothesis testing as a whole, not just the test statistic in particular.

Questions for the Audience

1. We have observed that the way students use certain mathematics software can influence students in developing a list of 'hints and shortcuts' for how to approach and solve certain types of problems. Does anyone in the audience have similar observations? How do we lower or eliminate this happening in our classes?
2. Does having multiple choice affect your students understanding of concepts? Do you observe your students trying to 'manipulate the system' to get correct answers instead of putting in the same amount of effort to understand the concept? How do you combat this issue?
3. Does the use of Excel versus other types of technology (calculators, statistical programs, etc) make a difference in the learning of students?

References

Asiala, M., Brown, A., DeVries, D. J., Dubinsky, E., Mathews, D., & Thomas, K. (1996). A framework for research and development in undergraduate mathematics education. In J. Kaput, E. Dubinsky, & A. H. Schoenfeld (Eds.), *Research in collegiate mathematics education II* (pp. 1-32). Providence: American Mathematical Society.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1-2), 75-98.

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal, 14*(1), 60-89.

LeMire, S.D. (2010). An argument framework for the application of null hypothesis statistical testing in support of research. *Journal of Statistics Education, 18*(2).

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core standards for mathematics.* Washington, DC: Authors.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Triola, Mario. (2014). *Elementary statistics using Excel* (5th ed.). Upper Saddle River, NJ: Pearson.

Vallecillos, A. (2000). Understanding the logic of hypothesis testing amongst university students. *JMD, 21*, 101-123.

Vere-Jones, D. (1995). The coming of age of statistical education. *International Statistical Review, 63*(1), 3-23.