POST-SECONDARY TEACHERS' UNDERSTANDING OF *P*-VALUE

Jason Dolor & Dana Kirin
Portland State University, Portland, OR
jdolor@pdx.edu

The last three decades has seen a significant growth in the number of students taking introductory statistics at the post-secondary level. It is therefore imperative to ensure that those employed to teach statistics have the appropriate statistical knowledge to provide quality education in the classroom. This research is part of a larger study investigating the statistical knowledge of graduate teaching assistants and community college instructors on topics of probability in hypothesis testing. We will present work done by these two populations on surveys and task-based interviews related to the p-value. Through this research, we hope to provide information on the statistical knowledge of teachers employed at the post-secondary level that may in turn be used to support the professional development of teachers of statistics.

INTRODUCTION

A fundamental concept in post-secondary statistics is the concept of the *p*-value. The *p*-value is just one part of the larger process of hypothesis testing. Many researchers continue to use hypothesis testing as a primary method in their practice. Therefore, knowledge of the *p*-value is fundamental to having a robust understanding of hypothesis testing. Unfortunately, research has shown that developing an understanding of *p*-value is difficult for many students (e.g. Batanero, 2000; Garfield & Ben-Zvi, 2008). Garfield and Ben-Zvi (2008) highlight misconceptions regarding the *p*-value's interpretation, that include: (1) "The *p*-value is the probability that the null hypothesis is true; (2) The *p*-value is the probability that the null hypothesis is false; (3) A small *p*-value means the results have significance (statistical and practical significance are not distinguished); (4) *p*-value indicates the size of an effect (e.g., strong evidence means big effects); (5) Large *p*-value means the null hypothesis is true, or provides evidence to support the null hypothesis; (6) If the *p*-value is small enough, the null hypothesis must be false" (p. 270). What is striking is that some of these misconceptions extend to teachers of statistics (Haller & Krauss, 2002; Thompson, Liu & Saldahna, 2007).

In an effort to better support students' learning of statistics in general and the development of students' understanding of statistical concepts associated with inference in particular, researchers and policy documents have advocated for the use of technology, coupled with simulation, to teach statistical inference from an empirical perspective (c.f., Cobb, 2007; GAISE College Report ASA Revision Committee, 2016). This has led to the design of curriculum that integrates simulation methods and technology as a pedagogical tool to support the learning of statistical concepts (see for example Garfield, delMas, & Zieffler, 2012). With the evolution of new standards and curriculum, it is vital that teachers of statistics have the knowledge to adapt to these changes so they can integrate emerging pedagogical strategies and curriculum to better support the development of students' understanding of concepts like *p*-value using technology and simulation techniques.

This study focuses on the knowledge of two populations of teachers at the post-secondary level: community college instructors (CCIs) and graduate teaching assistants (GTAs). Little is known about the statistical knowledge of CCIs. The demographics of CCIs vary greatly in their educational background with some having minimal statistical experience (Mesa, Wladis, & Watkins, 2014). GTAs also possess a very broad educational background and are often still in the process of developing their knowledge of statistics or mathematics. Additionally, as Speer, Gutman, and Murphy (2010) point out many GTAs are first time teachers and therefore might have minimal knowledge of pedagogical strategies beyond what they have seen during their own educational experiences. Given the varied educational and teaching experiences of CCIs and GTAs, it is important for the research community to gain a better understanding of how these teachers understand

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

fundamental concepts like the *p*-value and how they might integrate their understanding with new pedagogical practices such as simulation. Therefore, we pose the following research question: *What strategies do GTAs and CCIs use to answer questions related to p-value when asked to compute a p-value that used computer simulations to conduct a hypothesis test?*

THEORETICAL PERSPECTIVE

Since the focus of this paper is on the strategies of post-secondary teachers it is important to consider the types of knowledge of these post-secondary statistics teachers. Groth (2007) introduces a framework that discusses important aspects of statistical knowledge for teaching. In his framework, mathematical knowledge involves concepts of statistics that rely heavily on mathematical ideas (e.g. computation of probability, etc.). In contrast, the nonmathematical side refers to concepts and processes that are unique to statistics such as creating sampling methods to account for variation. Groth further highlights that teachers themselves possess types of common knowledge and specialized knowledge. Common knowledge includes types of knowledge shared by teachers and students in the classroom. Specialized knowledge is knowledge that is used in teaching, but not necessarily taught to students. These two types of content knowledge interact and work together to help guide the pedagogical activities of teachers.

We believe this framework can be useful in understanding the strategies post-secondary teachers develop when making sense of the p-value in the context of a simulation task. The concept of a *p*-value is common knowledge for teachers and students for an introductory statistics classroom, while understanding the concept of p-value in the context of simulation may be better classified as specialized knowledge. It is therefore worthwhile to analyze the knowledge of post-secondary school teachers by analyzing their strategies of the *p*-value through the context of simulations.

METHODOLOGY

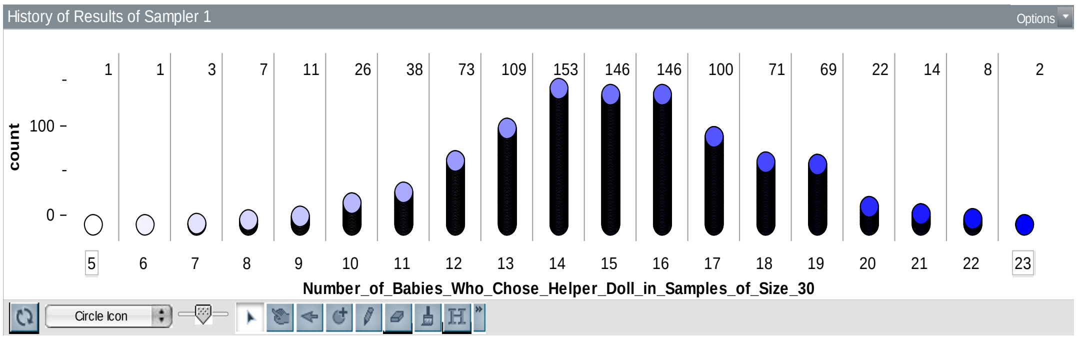
To answer the research question of interest, data were drawn from the first author's doctoral work on community college instructors and graduate teaching assistants understanding of the *p*-value (Dolor, 2017). The data consists of surveys collected from 55 participants who were given tasks that assessed their understanding of the *p*-value (see Dolor, 2017 for more detail). The participants were classified as community college teachers (CCI), four-year instructors (FYI), graduate teaching assistants (GTA), and graduate research assistants (GRA). The survey questions were the result of pilot research and feedback from experts in the statistics education community on concepts of the *p*-value. The survey focused on four main tasks that outlined important *p*-value concepts: (1) verbal interpretations of the *p*-value, (2) symbolic interpretation of the *p*-value, (3) the magnitude of the *p*-value, and (4) understanding of the *p*-value through simulations. The data also contained follow-up interviews with seven of the survey participants. The focus of this study is the fourth task that assesses post-secondary school teachers' understanding of *p*-value in the context of simulations.

The Helper-Hinder Task (Figure 1) describes a setting that begins with information about a study on infants who were asked to choose between a "good" or "bad" puppet. The task provides information about a hypothetical student who conducts a simulation to solve the task. The reader is asked to answer several questions about the task, which includes calculating an approximate *p*-value based on the information provided. Focusing primarily on the computation of the *p*-value, the participant was expected to use the empirical sampling distribution to determine a value for the p-value equal to 24/1000 or 2.4%.

Analysis of the data was conducted by both authors and focused on identifying and distinguishing between strategies used by the participants to compute the *p*-value. A coding scheme was generated and refined based on the methods used by the participants using a process of thematic analysis.

A sociology study was conducted to determine whether babies are able to recognize the difference between good and bad. In one experiment, 30 six-month old babies were randomly selected. Each baby was shown two possible puppets to play with, a ‘good’ puppet that helped and a ‘bad’ puppet that hindered. 21 out of 30 babies showed a strong preference for the helper puppet over the hinderer. In order to determine if this result provides strong statistical evidence that babies really do have a preference for the ‘good’ or helper puppet, James, a statistics student, conducted the following test procedure:

- James gets a coin and flips the coin 30 times.
    - If the coin lands on the “heads”, he records the baby as preferring the helper puppet.
    - If the coin lands on the “tails”, he records the baby as preferring the hinderer puppet.
- James then used a computer simulation to repeat the previous step 1000 times.
- James then plots the distribution for the number of times a baby chooses the helper puppet from each of the 1000 samples of size 30. This is shown in the graph below.



i.) James’ procedure is based on which assumption? Explain the reason for your choice.

a. A baby is more likely to choose the helper puppet.
b. A baby is equally likely to choose either the helper or hinderer puppet.
c. A baby is more likely to choose the hinderer puppet.

ii.) Suppose James wanted to conduct a right-tailed hypothesis test using the simulated data.

- What would you estimate for the *p*-value?
- Explain how you found the *p*-value and interpret it in the context of James’ research.

iii.) Based on your estimated *p*-value, what do you think should be James’ conclusion? Explain the reason for your choice.

a. There is statistically significant evidence that babies are more likely to choose helper puppets.
b. There is statistically significant evidence that babies are more likely to choose hinderer puppets.
c. There is statistically significant evidence that babies are equally likely to choose helper or hinderer puppets.

Figure 1. Excerpt from the Helper-Hinderer task

RESULTS

A total of seven strategies were coded as ways teachers computed the *p*-value. Table 1 shows a description of each of the codes and examples taken from the data illustrating each strategy that was identified.

Table 1. Categorization of strategies for computing a p-value

Category	Description	Example
Computation using relative frequency (CRF)	Participant only describe computing the p-value by counting the number of observations in the empirical sampling distributions that occurred for the outcomes of 21 and higher.	"It is the area passed the expected results. So (14+8+2)/1000."
Computation using relative frequency assuming null hypothesis (CRFNH)	Participant uses a relative frequency approach to compute the *p*-value using the simulated data and states the importance of a null assumption.	"If I'm to assume that there were no samples in which more than 23 of the 30 babies chose the helper doll, then a good estimate would be (14 + 8 + 2)/ 1000, or about 2.4%."
Computation ignores observed sample (CIOS)	Participant computes the *p*-value using a relative frequency, but ignores the observed sample in their computation.	"It is bootstrapping method. one side test. P(T>21|p=0.5) = 10/1000=0.01."
Computation using regions (CUR)	Participant computes the *p*-value using a relative frequency, but incorrectly assumes properties of theoretical sampling distributions.	"I just took the heights corresponding to 16,17,18,19,20 and added them up (which was 408) which I, then, subtracted from 500 (total of the right half). / 0.09 is still not a very small p-value, it is not a strong evidence in the favor of alternative hypothesis which would be the validity of the claim that the babies do in fact differentiate between good and bad."
Computation using level of significance (CLS)	Participant makes a reference to the level of significance and how the *p*-value is related to it (i.e. greater or less than the level of significance), but provides no actual computation.	"Level of significance = .05, split into .025 in each tail."
Computation using theoretical probability (CTP)	Participant computes the *p*-value using knowledge of theoretical probability or counting technique (e.g. binomial distribution).	"1/(2^30)*[30nCr16 + 30nCr17 + ... + 30nCr30] - We could interpret the p-value as being the probability that James' obtained 16 or more heads. The calculation is simply a binomial coefficient."
Computation using hypothesis testing methods (CHTM)	Participant computes the *p-value* by using traditional hypothesis testing method (i.e. one-proportion test).	"With a null hypothesis that the population proportion is .5 and an alternative that the proportion is greater than .5, we can find a z-score by taking the difference between the sample proportion of .7 and null proportion of .5 and dividing by the standard error of sqrt(.5*.5/30) and with the z-score got the associated area to the right to represent the p-value."

Table 2 shows the results of the various participants and the number of individuals who fell within each category. These categorizations illustrate that post-secondary school teachers do not share a similar way of thinking about the *p*-value when asked to compute one in the context of a simulation approach of hypothesis testing. Two of the categories (CTP and CHTM) showed teachers who preferred computing a *p*-value theoretical even with the simulated data present. Even some teachers who correctly used a relative frequency approach to computing the *p*-value showed varying approaches with some containing different misconceptions in their calculations (CIOS and CUR).

Table 2. Frequency counts for categorizations of strategies by instructor classifications

	Instructor Classification						
Category	GTA	GRA	GTA & GRA	CCI	FYI	CCI & FYI	Total
CRF	8	1	0	4	2	0	15
CRFNH	7	4	2	7	4	0	24
CIOS	1	0	0	0	1	0	2
CUR	1	0	0	1	1	0	3
CLS	3	0	0	2	0	0	5
CTP	1	0	1	0	0	0	2
CHTM	1	0	0	0	1	1	3
No Response	0	0	0	1	0	0	1
Total	22	5	3	15	9	1	55

The results of the coding scheme showed a variety of approaches made by post-secondary teachers when computing a *p*-value in the context of simulations. Even during the interviews, we noticed that participants' strategies were tied very closely to the pedagogical choices they would make in the classroom. Angie, a GTA in statistics who's strategy was categorized as a CTP, could correctly articulate how to compute a *p*-value using a simulation versus a theoretical approach during the interview, but preferred teaching from a theoretical perspective (as seen in the excerpt below).

> *Angie:* Personally I would teach the normal approximation because we don't get into the binomial distribution and I would probably prefer something theoretical over simulation as well.
>
> *Interviewer:* Why would you say that?
>
> *Angie:* I mean simulation works if that is all you can use, but if you have theory that's probably more sound than simulation unless you have some reason to believe your assumptions are wrong but if you're simulating under the same assumptions your theory is just as sound.

In contrast, Phil, a CCI whose strategy was categorized as CRFNH, expressed a preference for using simulations to teach statistical inference in the classroom.

> *Phil:* I personally think this should be way we should teach it on day one. This is my personal opinion and I think Allan (Rossman) agrees with me, but there is no reason to force parametric test if you can simulate using non-parametric methods like this. These are...kids get these.

These excerpts not only show contrasting strategies of computing the *p*-value but differing pedagogical views on the use of simulation when teaching hypothesis tests. Additionally, these excerpts offer some insight into why some secondary teachers might opt to use a particular strategy over another.

CONCLUSIONS

The results illustrate that the types of approaches used by post-secondary teachers vary greatly. The fact that there is such a varied amount of strategies illustrates that some of our post-secondary teachers do not all share a similar understanding of simulations and its role in statistical inference. While many of the teachers could compute the *p*-value correctly, a few of them chose to either ignore the simulated data or had misconceptions regarding the use of the simulated data. For example, Angie was a GTA who mentioned in her interview that she preferred to compute the *p*-value theoretically instead of using simulated data. This suggests that the strategies post-secondary teachers use when computing the p-value in the context of a simulation task may be mediated by their beliefs regarding the role of simulation and theoretical methods in the teaching and learning of hypothesis testing.

These results lead to greater implications for the statistics research community. In terms of professional development, these strategies can be used as a starting point to think about how we might move teachers towards correct ways of discovering the power of using simulations in the statistics classroom. Cobb (2007) highlights that statistics education should move towards using simulations in the classroom as an alternative to traditional approaches because it more aligns with the practice of statisticians today. Current work by the statistics education community has shown promise in developing curriculum that uses simulations to develop deeper understanding of statistics. For example, work done by Garfield, delMas, & Zieffler (2012) highlight a potential curriculum that stresses the importance of simulations in developing students' understanding of sampling distributions which is a crucial component in understanding statistical inference. Dolor & Noll (2015) have seen success in using simulations in classrooms with teachers in helping them conceptualize concepts of hypothesis testing. This work is just a small part of a much richer area of researcher into how we come to understand teachers of statistics and how we as a research community can support their development.

REFERENCES

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

Cobb, G. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum. *Technology Innovations in Statistics Education*, *1*(1). Retrieved from https://escholarship.org/uc/item/6hb3k0nz

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

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016*. Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM - The International Journal on Mathematics Education*, *44*(7), 883–898.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, *38*(5), 427–437.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research Online*, *7*(1), 1-20.

Mesa, V., Wladis, C., & Watkins, L. (2014). Research problems in community college mathematics education: Testing the boundaries of K-12 research. *Journal for Research in Mathematics Education*, *45*, 173–193.

Speer, N., Gutman, T., & Murphy, T. J. (2010). Mathematics teaching assistant preparation and development. *College Teaching*, *53*(2), 75–80.

Thompson, P., Liu, Y., & Saldanha, L. (2007). Intricacies of statistical inference and teachers' understandings of them. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (1st ed., pp. 207–231). Mahwah, NJ: Psychology Press.