DEVELOPING HYPOTHETICAL LEARNING TRAJECTORIES FOR TEACHERS' DEVELOPING KNOWLEDGE OF THE TEST STATISTIC IN HYPOTHESIS TESTING

Jason Dolor
Portland State University

In the past decade, educators and statisticians have made new suggestions for teaching undergraduate statistics. In light of these new recommendations it is important to (re)evaluate how individuals come to understand statistical concepts and how such research should impact curricular efforts. One concept that plays a major role in introductory statistics is hypothesis testing and the computation of the test statistic to draw conclusions in a hypothesis test. This proposal presents a theoretical approach through the development of a hypothetical learning trajectory of hypothesis testing by utilizing sampling distributions as the building block to understand statistical inference. In addition, this proposal presents how this hypothetical learning trajectory may support the development of research-based curricula that foster an understanding of the test statistic and its role in hypothesis testing.

Key words: Hypothesis tests, Test statistic, Sampling distributions, Guided reinvention

1. Introduction

Hypothesis testing has been used as a research tool in the fields of science, business, psychology, and education, among others. Because of the widespread use of hypothesis testing in research, it is no surprise that it has found its place in introductory statistics curriculum. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) report for collegiate curriculum was released outlining changes needed in introductory statistics courses (Aliaga et al, 2010). The report pushed for changes to instruction that fostered an understanding of statistics, including improving the education of hypothesis testing. If the goal of educators is to improve students understanding of hypothesis testing, then teacher's must help students view hypothesis testing as more than a procedure. Teachers must also have a robust understanding of the concepts of hypothesis testing because their knowledge has a direct impact on student learning. Robust knowledge of hypothesis testing includes understanding concepts like the level of significance, the p-value, the test statistic, and the sampling distribution.

In traditional hypothesis testing, the computation of the sample test statistic plays a crucial role. Surprisingly, introductory textbooks provide vague definitions of the test statistic. For example, Bluman (2012) defines the test statistic as "the numerical value obtained from a statistical test, computed from (observed value – expected value) / standard error" (p. 812). Levine and Stephan (2005) define the test statistic as "the statistic used to determine whether to reject the null hypothesis" (p. 274). These definitions do very little in outlining the importance of the test statistic or its relationship to other statistical concepts. In hypothesis testing, the test statistic is compared to a critical value or used to find the p-value to generate an inference. The test statistic is also crucial because the theoretical sampling distribution in traditional hypothesis testing is a result of a collection of many sample test statistics. If the goal of teachers is to foster student understanding of hypothesis testing then it is important that teachers understand the role of the test statistic and how test statistic formulas are generated. Thus, the goal of this research paper is to answer the following question: “What would a hypothetical learning trajectory look like for developing teacher understanding of test statistic formulas in hypothesis testing?”

To answer this question, the proposal begins with a review of the literature on hypothesis testing. Focusing on trends in the research and what factors are needed to improve the training of teachers. Then a hypothetical learning trajectory (HLT) designed to support in-service and pre-service teachers’ (IPST) learning of the role of a test statistic in hypothesis testing is proposed. Finally, the proposal ends with a discussion about the benefits and plans for future research in statistics education.

2. Literature Review

A review of the literature uncovered extensive research related to concepts of hypothesis testing (Batanero, 2000; Castro Sotos et al, 2007; Falk, 1986; Haller & Krauss, 2002; Thompson, Liu, and Saldahna, 2007; Vallecillos, 2002; Vallecillos & Batanero, 1997). Vallecillos and Batanero (1997) revealed that students have difficulties identifying the null and alternative hypotheses. Researchers have also found that students and teachers struggle with interpreting the level of significance and p-value (e.g. Batanero, 2000; Castro Sotos et al, 2007; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Falk, 1986; Liu, 2005; Liu & Thompson, 2005; Vallecillos, 2002; Vallecillos & Batanero, 1997) and fail to understand the role of a sampling distribution in hypothesis testing (e.g. Thompson, Liu, & Saldahna, 2007).

Research conducted by Thompson, Liu, and Saldahna (2007) discovered that teachers have difficulty in seeing the role of sampling distributions in hypothesis testing and understanding the logic of hypothesis testing. They conducted a professional development seminar with eight teachers who had extensive coursework in statistics. The results of the research revealed “some of the teachers’ conceptions of probability were not grounded in the concept of distribution which hindered their thinking about distributions of sample statistics and the probability that a given statistic is within a given range of the center of the distribution” (p. 228). The fact that the relationship between distribution and probability is problematic for teachers is troubling because these ideas lie at the heart of statistical inference. Thompson et al. argued that instruction for teachers should focus on developing their understanding of sampling distributions. In the same study, Thompson et al. also investigated the teachers’ understanding of unusualness of samples in the context of statistical inference through sampling distributions. A sample statistic would be considered rare or unusual if it fell in a region of the sampling distribution that had small occurrences of other sample statistics. The importance of the sampling distributions in hypothesis testing (and statistical inference in general) is a view shared by many statistics education researchers (e.g. Garfield & Ben-Zvi, 2008; Lipson, 2003; Rubin et al, 1990; Saldanha & Thompson 2002; Thompson, 2004; Watson & Moritz, 2000).

There has been extensive research covering concepts of level of significance, p-value, sampling distributions and null hypothesis, but an exhaustive search of the literature revealed no research investigating teachers’ or students’ understanding of the test statistic. This lack of research testing is quite troubling because of the prominent role test statistics play in introductory statistics textbooks’ treatments of hypothesis testing. However, research pertaining to the concepts of unusualness and sampling distributions could play a key role in generating methods for developing IPSTs’ understanding of test statistics.

3. Theoretical Perspective

Many researchers recommend that sampling distributions be central in the teaching of statistical inference (Garfield & Ben-Zvi, 2008; Lipson, 2003; Rubin et al, 1990; Saldanha &

Thompson 2002; Thompson, 2004; Thompson, Liu, and Saldahna 2007; Watson & Moritz, 2000). For example, Thompson et al. (2007) state "we suspect that teachers who value distributional reasoning in probability and who imagine a statistic as having a distribution of values will be better positioned to help students reason probabilistically about statistical claims" (p. 229). Thus, if one uses suggestions by researchers to simply use a sampling distribution of sample proportions (or means), then making a statistical inference must rely on utilizing the sampling distribution in a hypothesis test. To illustrate this, consider the following problem:

> *Suppose a researcher wanted to determine whether a college population has more males than females. He surveys a group of people and finds that 70% of them are male. Is this sufficient evidence to claim there are more males than females?*

To perform a hypothesis test we begin by first assuming the population is equally proportioned between male and female. This identifies the null hypothesis as population proportion of males being 50% (i.e. H_0: p = 0.50) with an alternative hypothesis being that there are more males in the population (i.e. H_1: p > 0.50). This produces a hypothetical population distribution of 50% males. Using a computer simulation, one could generate multiple samples from the assumed population of 50% males. A sampling distribution of proportions could then be produced from the simulated samples. An individual could determine the unusualness of the *observed* sample proportion (i.e. $\hat{p} = 0.70$) by locating its position in the sampling distribution (Figure 1). If the observed sample fell in a region of the sampling distribution where other hypothetical null sample proportions are unlikely to fall, then the observed sample is considered to be unusual. A person can then claim with statistical significance that 50% is not likely to be the true population proportion of males in the college population.

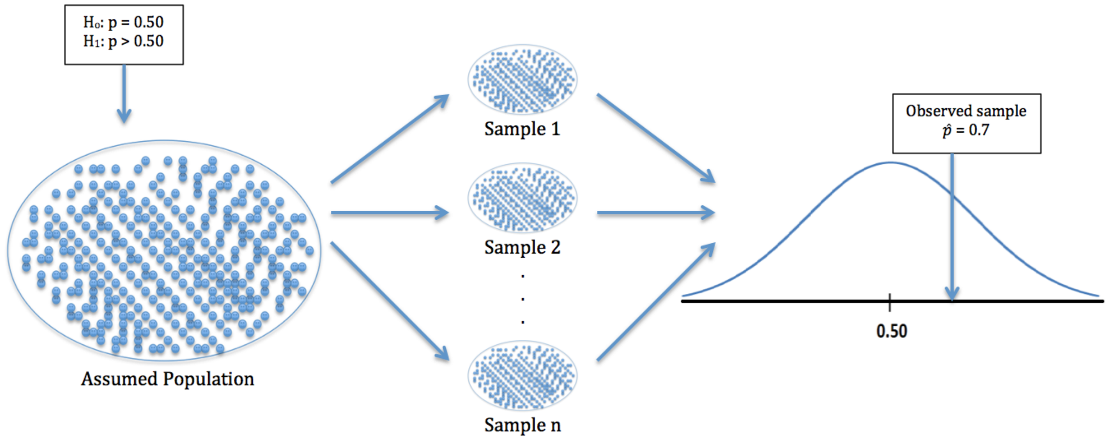


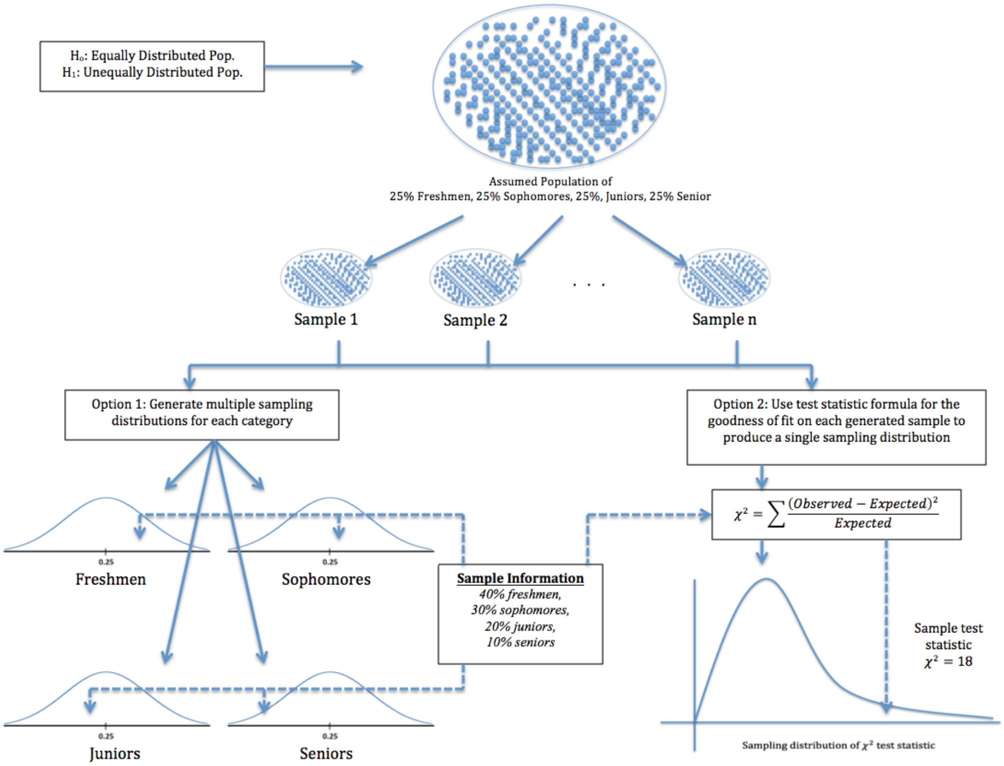
Figure 1. Hypothesis testing approach utilizing a sampling distribution.

This approach places the sampling distribution at the center of the hypothesis testing argument. Furthermore, the null hypothesis is prominent in this approach because it requires IPSTs to generate samples from an assumed null population. Finally, it allows IPSTs to realize that a sample's unusualness is a result of a *frequentist* approach to probability. That is, the probability of a sample proportion occurring is a result of a long-term stochastic process of sampling many times from the null population. The approach described above has many benefits, but problems may arise when dealing with complex situations.

Teaching hypothesis testing utilizing the above approach is valid if we study a single proportion. If the problem were to include multiple proportions, then one could speculate that the approach of hypothesis testing described above should easily transition to multiple proportions. Let us consider the following hypothesis test problem.

> *Suppose a researcher wanted to determine whether there was a difference between the proportion of freshmen, sophomore, juniors, and seniors in a college population. He surveys a sample of students from the college and the sample contains 40% freshmen, 30% sophomore, 20% juniors, and 10% seniors. Is this sufficient evidence to claim distribution of freshmen, sophomore, juniors, and seniors are not equal?*

In this example, we begin with the assumption that freshmen, sophomores, juniors, and seniors are equally distributed (i.e. 25% freshman, 25% sophomores, 25% juniors and 25% seniors). Following similar logic as above, this would generate an assumed population that is equally distributed between the different categories. Once again, an individual could use computerized simulations to produce an empirical sampling distribution based on many samples from the assumed population. This begs the question, “How do we create a sampling distribution to represent the null assumption and determine the rarity of an observed sample?” One option is to generate multiple sampling distributions for each category. That is, generate a sampling distribution for the percentage of freshmen, sophomores, juniors, and seniors with 0.25 as the population proportion (center) for each distribution. A second option is to use the chi-squared test statistic formula (i.e. $\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$) to generate a single sampling distribution (Figure 2).



H₀: Equally Distributed Pop.
H₁: Unequally Distributed Pop.
Assumed Population of
25% Freshmen, 25% Sophomores, 25%, Juniors, 25% Senior
Sample 1
Sample 2
. . .
Sample n
Option 1: Generate multiple sampling distributions for each category
Option 2: Use test statistic formula for the goodness of fit on each generated sample to produce a single sampling distribution
$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$
0.25
Freshmen
0.25
Sophomores
Sample Information
40% freshmen,
30% sophomores,
20% juniors,
10% seniors
Sample test statistic
$\chi^2 = 18$
0.25
Juniors
0.25
Seniors
Sampling distribution of χ^2 test statistic

Figure 2. Two approaches of hypothesis testing for multiple proportions.

Option 1, or the multiple sampling distribution approach (MSDA) may be a logical choice for someone who has already used sampling distributions as outlined in the first example (Figure 1) as a means for studying hypothesis testing. One could determine the likelihood of an observed sample by considering where the observed proportion from each category fell with respect to that category’s empirical sampling distribution. For instance, seeing where 40% freshmen fell in the sampling distribution of freshmen based on the 25% null assumption. This approach might seem intuitive at first, but an investigation of this approach uncovers significant problems. One such problem arises when considering the following question: What if two of the categories were considered unusual, but two were not? For example, sophomores and juniors are not unusual in their respective sampling distribution because 30% and 20% are not unusually far from the center of 25%, but freshmen (40%) and seniors (10%) are far from the center of their respective distributions. This would mean developing additional criteria to determine a sample’s unusualness. This problem increases in difficulty as additional categories are added. A second problem with this approach is that by generating multiple sampling distributions, the probability of generating an incorrect inference increases. Comparing an observed sample with multiple proportions across multiple sampling distributions compounds Type-I errors. If it is decided that unusual is a sample that has a 5% chance of occurring for each category, then this approach is not really comparing unusual at the level of 5%.

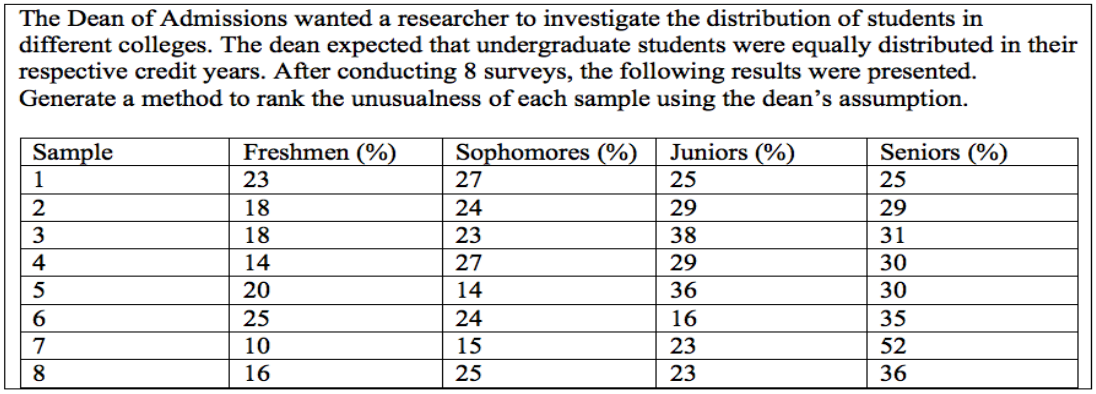
Option 2, or the single sampling distribution approach (SSDA), as the name implies, uses a single sampling distribution to discuss unusualness. The chi-squared test statistic formula generates a sampling distribution of sample chi-squared test statistics. Unusualness of a sample could once again be determined by locating where the observed sample’s chi-squared test statistic falls within the null sampling distribution. SSDA is the approach we want IPSTs to know when multiple categories are being investigated. SSDA is directly related to the traditional approach found in statistical textbooks, which is formally called the chi-squared goodness-of-fit test. Another example of a SSDA is IPSTs might generate a test statistic formula where they sum the absolute deviations (i.e. $\sum |Observed - Expected|$) to construct a sampling distribution. Thus, there are student-generated approaches (SGA) within SSDA and MSDA to testing multiple proportions. Currently we can only speculate what SGA of hypothesis testing might be, but it would be beneficial to analyze how SGA can be used to leverage IPSTs towards traditional hypothesis tests.

The examples above provide motivation towards developing a hypothetical learning trajectory (HLT) from which to study teachers’ development of test statistics for more complicated hypothesis tests. I conjecture that an approach to understanding a test statistic must encompass a relationship between the observed sample information, unusualness, null hypothesis, and sampling distribution. If the new approach towards hypothesis testing is to base decisions on the sampling distributions, it is important for teachers to also understand the meaning of the points used to generate the sampling distribution. These points are the direct results of the test statistic formula being applied to samples. Therefore, the test statistic formula provides a numerical summarization of sample information. The motivation for developing a test statistic that generates a single numerical value is to generate a single sampling distribution rather than multiple sampling distributions. Furthermore, the sampling distribution we wish to generate must express the unusualness of samples in light of the null hypothesis. In other words, developing a test statistic formula should be viewed as a way to quantify unusualness of an observed sample under the null assumption. Viewing the test statistic formula through this perspective could be useful in generating tasks where IPSTs reinvent the test statistic formula. One such approach to teaching where IPSTs reinvent mathematical concepts is through *guided reinvention* (Gravemeijer, 2004).

Guided reinvention is part of the theoretical framework of realistic mathematical education (RME). The basis of RME is that mathematics should be learned naturally through discovery and discussion, as students are involved in solving mathematical problems realistic to his/her perspective. In short, mathematical knowledge is developed by an individual through experiences. Rather than a traditional lecture, students learn through instructional tasks and discussion. Students' shared ideas play a central role of the learning while the teacher serves as a mediator to ensure discussions are directed toward a learning goal. The goal is outlined through a *hypothetical learning trajectory* (HLT). "The notion of a hypothetical learning trajectory entails that the teacher has to envision how the thinking and learning, in which the students might engage as they participate in certain instructional activities, relate to the chosen learning goal" (Gravemeijer, 2004, p. 8). The HLT consists of three components: (1) establishing learning goals, (2) envisioning students mental process, (3) instructional design (Gravemeijer, 2004).

In order to generate activities with the goal of building understanding of the test statistic, a careful description of the HLT is needed. Prior to working with tasks on developing a test statistic formula for multiple proportions, IPSTs should already have an understanding of hypothesis testing using single proportions (i.e. Table 1). This way, when IPSTs are presented the task of multiple proportions they are already motivated to generate a sampling distribution(s) to make decisions about the null assumption. The goal of the HLT for test statistic activities is to move IPSTs towards SSDA above, where they begin to understand the role of the test statistic as a numerical quantification of unusualness leading towards the development of a single sampling distribution.

When generating a task around the goodness-of-fit test to develop a test statistic formula, one approach is to have IPSTs compare unusualness of samples against other samples in light of an assumption. One such task could be the ranking task below (Figure 3). The goal of the task is for IPSTs to generate a method to numerically measure unusualness of a sample in light of the null assumption.



The Dean of Admissions wanted a researcher to investigate the distribution of students in different colleges. The dean expected that undergraduate students were equally distributed in their respective credit years. After conducting 8 surveys, the following results were presented. Generate a method to rank the unusualness of each sample using the dean's assumption.

Sample	Freshmen (%)	Sophomores (%)	Juniors (%)	Seniors (%)
1	23	27	25	25
2	18	24	29	29
3	18	23	38	31
4	14	27	29	30
5	20	14	36	30
6	25	24	16	35
7	10	15	23	52
8	16	25	23	36

Figure 3. Ranking task to developing a test statistic formula for multiple proportions

I conjecture that IPSTs would intuitively see how an observed sample differs from the expectation for each category. The goal of the task is to lead IPSTs towards developing a chi-squared test statistic formula in order to apply a SSDA for a hypothesis test problem. Ideally, we would want IPSTs to generate the chi-squared test statistic. It is possible that IPSTs will not generate this test statistic at first. For instance, IPSTs might develop a formula where they sum the absolute deviations. In this case, moving towards the chi-squared test statistic would require additional tasks and discussions. This task also helps IPSTs connect the relationship of the null assumption with the test statistic formula and build an understanding that unusualness is based on comparing samples. Following this task, IPSTs can attempt a hypothesis test utilizing their constructed test statistic formula to generate a sampling distribution. Discussion regarding properties of the test statistic formula and its role in the hypothesis testing procedure can follow.

4. Discussions

A review of the literature has uncovered a lack of research on student or teacher understanding of the test statistic. This paper presents a methodology for developing tasks that would foster an understanding of the test statistic formula. By utilizing the suggestions of researchers, the methodology offered here supports the importance of sampling distributions as a major part of instruction on hypothesis testing by extending sampling distributions to encompass complex samples. I also offer a perspective of the test statistic formula by viewing it as a tool to quantify the unusualness of a sample in light of the null hypothesis. Further, I present a possible HLT that could be utilized in order to build understanding of a test statistic for a goodness-of-fit test. Plans for future research will focus on actual implementation of a teaching experiment using the prescribed HLT. During the teaching experiment, it would be worthwhile to also examine other approaches that might differ from the MSDA and SSDA described above. Finally, this research focuses on samples with multiple proportions but it would be worthwhile to consider tasks where IPSTs develop test statistics formula for the various hypothesis tests. The goal of this paper was to produce a methodology towards understanding the test statistic formula. In the process, I have also uncovered a new view of hypothesis testing that could be useful for the future of statistic education research.

References

Aliaga, M., Cuff, C., Garfield, J., Lock, R., Utts, J., & Witmer, J. (2010). *Guidelines for Assessment and Instruction in Statistics Education College Report*.

Batanero, C. (2000). Controversies around the Role of Statistical Tests in Experimental Research. *Mathematical Thinking and Learning*, *2*(1-2), 75–98.

Bluman, A. G. (2012). *Elementary statistics: A step by step approach* (8th ed.). New York, NY: McGraw Hill.

Castro Sotos, A. E., Vanhoof, S., Noortgate, W. V. den, & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*, 98 – 113.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83 – 96.

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Practice*. Springer.

Gravemeijer, K. (2004). Creating opportunities for students to reinvent mathematics. Presented at the ICME 10, Denmark.

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

Levine, D. M., & Stephan, D. F. (2005). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, NJ: Pearson Prentice Hall.

Lipson, K. (2003). The Role of the Sampling Distribution in Understanding Statistical Inference. *Mathematics Education Research Journal*, *15*(3), 270–287.

Liu, Y. (2005, August). *Teachers' Understandings of Probability and Statistical Inference and their Implications for Professional Development* (Dissertation). Vanderbilt University, Nashville, Tennesse.

Liu, Y., & Thompson, P. (2005). Teachers' Understandings of Hypothesis Testing. *27th annual* *meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Presented at the PME-NA.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning About Sampling: Trouble at the Core of Statistics. Presented at the American Educational Research Association.

Saldanha, L., & Thompson, P. (2002). Conceptions of Sample and their Relationship to Statistical Inference. *Educational Studies in Mathematics*, *51*, 257 – 270.

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.

Thompson, Pat, Saldanha, L., & Liu, Y. (2004). Why statistical inference is hard to understand. Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypothesis testing by university students. *Themes in Education*, *3*(2), 183 – 198.

Vallecillos, A., & Batanero, C. (1997). Activated concepts in the statistical hypothesis contrast and their understanding by unversity students. *Reserchers en Didactique des Mathematiques*, *17*, 29 – 48.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, *31*, 44 – 70.