Students' understanding of the concepts involved in one-sample hypothesis testing

Harrison E. Stalvey
University of Colorado Boulder

Annie Burns-Childers
University of Arkansas, Little Rock

Darryl Chamberlain Jr.
University of Florida

Aubrey Kemp
Georgia State University

Leslie J. Meadows
Georgia State University

Draga Vidakovic
Georgia State University

Abstract

Hypothesis testing is a prevalent method of inference used to test a claim about a population parameter based on sample data, and it is a central concept in many introductory statistics courses. At the same time, the use of hypothesis testing to interpret experimental data has raised concerns due to common misunderstandings by both scientists and students. With statistics education reform on the rise, as well as an increasing number of students enrolling in introductory statistics courses each year, there is a need for research to investigate students' understanding of hypothesis testing. In this study we used APOS Theory to investigate twelve introductory statistics students' reasoning about one-sample population hypothesis testing while working two real-world problems. Data were analyzed and compared against a preliminary genetic decomposition, which is a conjecture for how an individual might construct an understanding of a concept. This report presents examples of Actions, Processes, and Objects in the context of one-sample hypothesis testing as exhibited through students' reasoning. Our results suggest that the concepts involved in hypothesis testing are related through the construction of higher-order, coordinated Processes operating on Objects. As a result of our data analysis, we propose refinements to our genetic decomposition and offer suggestions for instruction of one-sample population hypothesis testing. We conclude with appendices containing a comprehensive revised genetic decomposition along with a set of guided questions that are designed to help students make the constructions called for by the genetic decomposition.

Keywords: Hypothesis testing, Introductory statistics, APOS Theory

1 Introduction

The use of statistics is crucial for numerous fields, such as business, medicine, education, and psychology. According to the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report, more students are studying statistics due to its importance. As a result, the GAISE called for nine learning goals, one of which stated, "students should demonstrate an understanding of, and ability to use, basic ideas of statistical inference, both

hypothesis tests and interval estimation, in a variety of settings" (GAISE College Report ASA Revision Committee, 2016, p. 8). In other words, in an introductory statistics course, students should understand and be able to apply hypothesis testing in various situations.

Hypothesis testing, as explained by Davis and Mukamal (2006), is a procedure for testing a claim about a population parameter; it involves measuring the strength of evidence provided by sample data (see also Triola, 2014). In particular, hypothesis testing involves formulating opposing statements—the null hypothesis and alternative hypothesis—about the population parameter of interest. The goal of hypothesis testing is to determine whether or not to support the original claim, based on whether we reject the null hypothesis. To do so, a sample statistic is measured or observed, and a decision is made to reject or fail to reject the null hypothesis based on the extremity of the sample statistic. This decision is based on the probability, called the p-value, of observing the sample statistic, under the assumption that the null hypothesis is true. In particular, if the p-value is small enough, then we reject the null hypothesis.

While hypothesis testing is an important tool of statistical inference (Krishnan & Idris, 2015), its use to interpret experimental data has received criticism (Motulsky, 2014; Nickerson, 2000; Nuzzo, 2014) due to the common misunderstandings by both scientists and students when using this method (Batanero, 2000; Dolor & Noll, 2015; Vallecillos, 2000), such as the meaning behind the p-value. Rather than abandon this method of inference entirely, researchers have called for improving the education and understanding of hypothesis testing. For example, LeMire (2010) defended the use of hypothesis testing and provided a framework that can be used to develop instructional content, with the goal of fostering student understanding.

There are few studies on student understanding of hypothesis testing as a whole (Smith, 2008). With statistics education reform on the rise across the United States, as well as an increasing number of students enrolling in introductory statistics courses each year, there is a need for research that investigates students' understanding and curriculum effectiveness of hypothesis testing, a concept taught in almost every introductory statistics course (GAISE College Report ASA Revision Committee, 2016; Krishnan & Idris, 2015).

Our study investigated the understanding of one-sample population hypothesis testing[1] by university students enrolled in an introductory statistics course based on the emporium model (described in Section 4). This report aims to answer the following research question:

How do students reason about the concepts involved in one-sample population hypothesis testing? In particular, how do students reason about these concepts while working two problems involving real-world situations?

2 Literature review

Research has revealed that although students are able to perform the procedures surrounding hypothesis testing, they lack a strong understanding of the concepts and their use within the procedure (Smith, 2008). Providing a survey of research on students' understanding of statistical concepts, Batanero et al. (1994) stated that hypothesis testing "is probably the

[1]Throughout this paper, *hypothesis testing* is meant to specifically refer to one-sample hypothesis testing.

most misunderstood, confused and abused of all statistical topics" (p. 541). Students appear to experience a "symbol shock" (Schuyten, 1990), which provides an obstacle for students interpreting particular questions (Dolor & Noll, 2015; Liu & Thompson, 2005; Vallecillos, 2000). Vallecillos (2000) found that students have trouble with not only the symbols, but also with the formal language and meaning behind the concepts involved in hypothesis testing, including words such as "null" and "alternative" when referring to the hypotheses. Students interviewed were not able to accurately describe what these terms mean and how they impact the decision to either reject or fail to reject the null hypothesis (Vallecillos, 2000). Williams (1997) made a similar observation. She found that, due to the tedious process behind hypothesis testing, students were not able to connect the statistical concepts back to the context of the problem, stating, "the biggest hurdle is reaching a statistical conclusion, and the real meaning of the original question may be forgotten in the process" (p. 591).

Textbooks and instructors frequently give a specific step-by-step script to follow when performing hypothesis testing without explicitly summarizing the entire process afterwards, which does not provide students the opportunity to see the idea as a whole. Link (2002) described this practice as a six-part procedure, which leads many students to look for keywords and phrases as guides when solving hypothesis testing problems. He also found evidence that students can correctly substitute values into a formula selected from a formula sheet, yet they do not have an understanding of the logic behind the overall procedure of hypothesis testing.

Due to the rise of statistics education, calls for reform have led to a shift from an emphasis on procedural understanding to conceptual understanding (GAISE College Report ASA Revision Committee, 2016; Krishnan & Idris, 2015). Ways to teach for conceptual understanding have varied. For example, Hong and O'Neil (1992) suggested that to foster a conceptual understanding of hypothesis testing, instruction with an emphasis on concepts and diagrammatic problem representations should precede instruction regarding the procedures in hypothesis testing. Additionally, some studies revealed the successes of implementing statistical software to enhance students' conceptual understanding of hypothesis testing. Chandrakantha (2014) found Microsoft Excel to be an effective teaching tool, evidenced by the better performance by a class using the software in comparison to a traditional class. Yung and Paas (2015) also asserted that visual representation is beneficial to students. This is especially true when the students have the opportunity to experiment with real data and real world problems (Moore, 1997).

Below we describe the framework with which we analyzed student responses to two real-world hypothesis testing problems.

3 APOS Theory

Action–Process–Object–Schema (APOS) Theory is a constructivist framework used to describe how an individual might develop his or her understanding of a mathematical concept (Arnon et al., 2014; Asiala et al., 1996; Cottrill et al., 1996; Dubinsky & McDonald, 2001). It emphasizes the construction of cognitive structures called Actions, Processes, and Objects,

which make up a Schema.[2] These structures are the result of a mental mechanism called reflective abstraction, the notion of which comes from Piaget (Beth & Piaget, 1966). In short, APOS Theory uses the premise that reflective abstraction consists of reflection and reorganization of mental structures. That is, an individual reflects on the given problem-solving situation and constructs or reconstructs certain mental structures (Actions, Processes, Objects, and Schemas). Construction or reconstruction is achieved through the mechanisms of interiorization, reversal, coordination, encapsulation, and thematization (for more details about these mechanisms, see Arnon et al., 2014). The construction of Actions, Processes, and Objects that make up a Schema signify the stages of understanding a mathematical concept.

An Action is a transformation of mathematical objects (e.g., numbers, functions) in response to external cues. The primary characterization of an Action is the external cue, which could be keywords or a memorized procedure. For example, in the case of the concept of function, an Action could be inputting a value into an algebraic expression and simplifying to obtain the output. Having constructed an Action can be exhibited through an individual's ways of solving a problem. An individual who is limited to performing Actions is said to be at the Action stage or to possess an Action conception.

Reflection on a repeated Action can lead to its interiorization to a Process. While an Action is an external transformation of objects, a Process is an internal transformation of objects that enables an individual to think about the transformation without actually performing it. For example, a function Process could be the mental image of a function accepting inputs and transforming them into outputs. Having constructed a Process can be exhibited through an individual's ways of solving a problem. This signifies that the individual is at the Process stage or possesses a Process conception. Processes are constructed not only through the interiorization of Actions, but also through the reversal of an existing Process and through the coordination of two existing Processes.

Once a Process is conceived as a totality and the individual can perform transformations on it, the Process is said to have been encapsulated into an Object. For example, a function Object could be acted upon by another function or binary operation to obtain a new function. Having constructed an Object can be exhibited through an individual's ways of solving a problem. This signifies that the individual is at the Object stage or possesses an Object conception.

A collection of Actions, Processes, and Objects organized in a coherent manner is called a Schema.

3.1 Preliminary genetic decomposition

A crucial component of research informed by APOS Theory is the development of a genetic decomposition, which is a description of how an individual might develop an understanding of a mathematical concept. In particular, it is a description of the mental structures of Actions, Processes, and Objects that make up a Schema and how they might be constructed. In this section, we describe the prerequisite constructions that we suggest an individual should have

[2]In APOS Theory, the words Action, Process, Object, and Schema are capitalized to refer to a mental structure, to distinguish them from the colloquial use of these terms.

made prior to studying hypothesis testing. Then we describe the mental structures that correspond to the concepts involved in hypothesis testing. We developed the preliminary genetic decomposition prior to the data collection, and it is based on a literature review, the researchers' experiences, and the presentation of hypothesis testing in the textbook used in the introductory statistics course for our study.

Prerequisite constructions. Prior to the study of hypothesis testing, an individual should have developed Schemas for **representation**, **distribution**, **probability**, **function**, and **empirical evidence**.[3] A **representation** Schema should include Processes of representing a concept algebraically, graphically, and verbally. Furthermore, Processes in a **representation** Schema should be coordinated to transition between different representations of a concept. A **distribution** Schema enables the individual to describe a data set as being symmetric or not symmetric and to distinguish between the circumstances for a normal distribution and a Student's t-distribution. A **probability** Schema enables the individual to understand that a probability is a number between 0 and 1. In particular, the individual should understand that the closer the probability of an event is to 0, the lower the likelihood is of the event occurring, while the closer the probability of an event is to 1, the higher the likelihood is of the event occurring. A **function** Schema should include a Process that enables the individual to describe, in general, a function accepting an input and returning an output. An **empirical evidence** Schema should include statements as Objects, as well as a Process of determining and justifying the truth value of a statement.

The coordination (or interaction) of Schemas has been discussed previously in APOS Theory literature (e.g., Baker, Cooley, & Trigueros, 2000; Martínez-Planell & Trigueros Gaisman, 2012). By coordination of Schemas we mean that mental structures from two different Schemas have been utilized in some way, such as a Process in one Schema acting on an Object in another Schema. We consider the coordination of Schemas to be fundamental to understanding hypothesis testing. Coordinating **representation** and **distribution** Schemas enables an individual to conceptualize the normal distribution and Student's t-distribution graphically as a symmetric bell curve and to understand how different concepts in hypothesis testing correspond to the distribution curve. Further coordinating **representation** and **distribution** Schemas with a **probability** Schema enables the individual to understand that certain probabilities can be represented as areas under the distribution curve. Coordinating **probability** and **function** Schemas enables the individual to construct a **probability function** Process that returns values between 0 and 1. These coordinations result in a Schema that should be further coordinated with an **empirical evidence** Schema in order to form an argument using the concepts involved in hypothesis testing. Figure 1 demonstrates these coordinations.

We propose that **hypothesis testing** is a Schema, which is constructed through the aforementioned coordinations, resulting in mental structures called **hypotheses**, **test statistic**, **p-value**, **decision**, and **conclusion**. Figure 2 illustrates the development of these mental structures, particularly as Processes. Starting at the bottom of the figure, lower order

[3]We use **bold font** when referring to the primary mental structures that make up our genetic decomposition, to distinguish them from other uses of these terms. For simplicity, we do not use a different font to distinguish between the different stages (Action, Process, Object, Schema) corresponding to a concept.

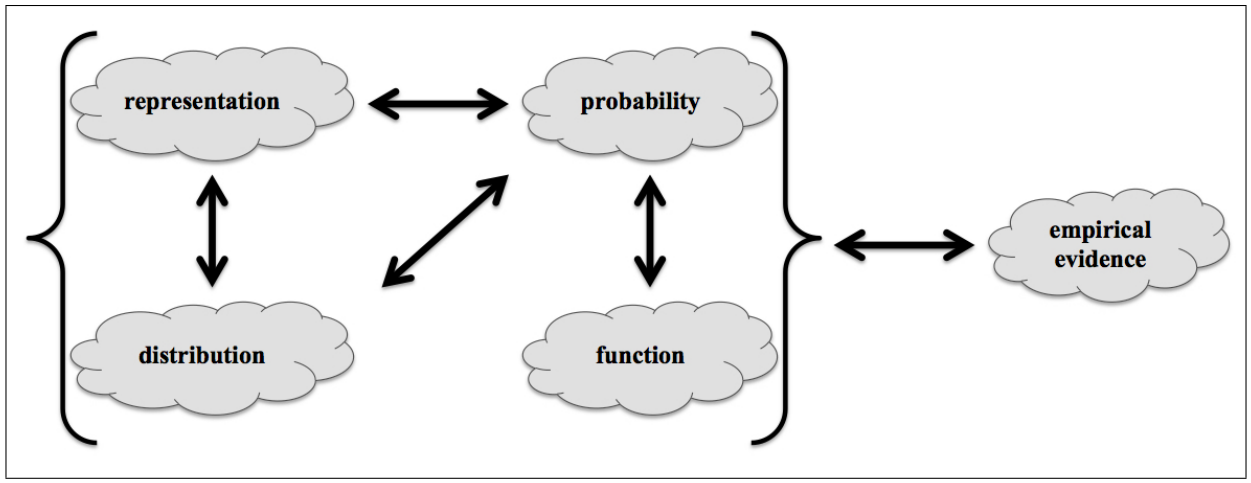

Figure 1: Coordination of Schemas.

Processes are constructed, and, moving up the figure, these Processes are coordinated into higher-order Processes, signifying the direction of the reflective abstraction. Figure 2 only makes up part of our preliminary genetic decomposition. In the passages that follow, we describe the stages of development of these mental structures in more detail.

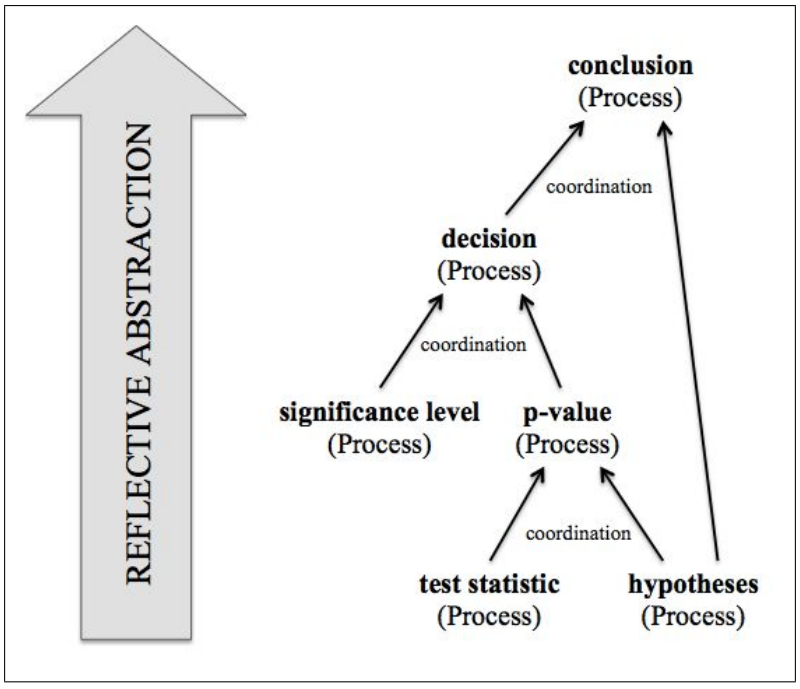

Figure 2: Interaction of Processes in the preliminary genetic decomposition.

Hypotheses. In hypothesis testing, the claim about the population parameter is used to formulate two hypotheses—the null hypothesis, H_0, and the alternative hypothesis, H_1. This can be thought of as a transformation, **hypotheses**, that acts on the claim as an

Object and returns two hypotheses statements. The following are descriptions of the stages of **hypotheses**.

Action: A **hypotheses** Action is a transformation in response to external cues that tell the individual how to formulate the hypotheses for a specific hypothesis test. An external cue could be a memorized procedure or keywords.

Process: The individual interiorizes into a Process the Actions of identifying the claim and formulating H_0 and H_1. In particular, he or she understands that for any hypothesis test the claim is used to formulate H_0 and H_1, and the individual can explain in his or her own words what H_0 and H_1 mean.

Object: The individual encapsulates a **hypotheses** Process into an Object when he or she can think of it as a totality and perform a transformation on it. For example, the individual can compare how he or she formulated the hypotheses between different situations. That is, the transformation being applied to this **hypotheses** Object is a comparison.

Test statistic. In hypothesis testing, the test statistic, as referenced in this paper, is a standardized value of the sample statistic represented by a score (such as z or t) and is used to determine whether to reject the null hypothesis. For the purposes of our study, the scores represent the relative distance of the sample statistic from the assumed value of the parameter. Calculating a test statistic can be thought of as a transformation, **test statistic**, which acts on various population parameters and sample statistics and returns a value, reasonably between -3 and 3, called the test statistic. The following are descriptions of the stages of **test statistic**.

Action: A **test statistic** Action is an external rule, such as a formula, that tells an individual how to calculate a test statistic's value. Through this Action an individual can input the required values of the population parameters and sample statistics into a formula and simplify the expression to obtain the test statistic's value.

Process: The individual interiorizes into a Process the Action of calculating the test statistic. This Process enables the individual to understand that, in general, the test statistic is the relative distance of the sample statistic from the assumed value of the parameter, and the individual can explain this in his or her own words.

Object: The individual encapsulates a **test statistic** Process into an Object by being able to distinguish the difference between extreme values of the test statistic and those that are not considered extreme. That is, the transformation that is being applied to this **test statistic** Object is a comparison of usual and unusual test statistic values.

P-value. In hypothesis testing, the p-value is the probability of obtaining a sample statistic at least as extreme as the one that was observed, under the assumption that the null hypothesis is true. Calculating the p-value can be thought of as a transformation, **p-**

value, which is a probability function. In particular, **p-value** acts on the output of **test statistic** and returns a number between 0 and 1, called the p-value. The following are descriptions of the stages of **p-value**.

Action: A **p-value** Action is an external transformation on a particular test statistic's value by following a step-by-step procedure to compute the p-value.

Process: The individual coordinates **hypotheses** and **test statistic** Processes to construct a **p-value** Process that acts on the output of **test statistic**. An individual who has coordinated his or her **probability** and **function** Schemas to construct a **probability function** Process can think of **p-value** as a function that returns the probability of obtaining the sample data or statistic that we observed or something more extreme, under the assumption that the null hypothesis is true. The individual can explain in his or her own words what the p-value represents.

Object: The individual encapsulates a **p-value** Process into an Object by comparing it with **significance level** as an Object. An individual with an Object conception of p-value views the p-value as a probability or area and compares it to the significance level, α, which is the probability of rejecting the null hypothesis given that it is true. That is, the transformation being applied to this **p-value** Object is a comparison with **significance level**.

Decision. In hypothesis testing, a decision about whether or not to reject the null hypothesis, H_0, is made based on the comparison of the p-value with the significance level, α, which is a probability that serves as an upper bound for the region of probabilities which would result in rejecting the null hypothesis. In particular, when the p-value is less than or equal to α, the null hypothesis is rejected. Forming the decision about H_0 can be thought of as a transformation, **decision**, that compares the p-value and α and returns the decision about whether or not to reject H_0. The following are descriptions of the stages of **decision**.

Action: A **decision** Action is an external transformation on a particular p-value and α by following a step-by-step procedure, such as the following:

Step 1. Compare the numerical values of the p-value and α.
Step 2. From the following, choose the option that applies to the above information:

"p-value $\leq \alpha$" $\Rightarrow$ "reject H_0."
"p-value $> \alpha$" $\Rightarrow$ "fail to reject H_0."

In particular, the individual performs the above Actions by relying on memorization or a formula sheet, without understanding the logic behind the steps.

Process: The individual interiorizes the above Actions into a Process (1) by understanding that the comparison of the p-value and α determines, in general, the decision about the null hypothesis, (2) by understanding that the p-value is the probability of

obtaining the sample data or statistic that we observed or something more extreme, under the assumption that the null hypothesis is true, and (3) by viewing α as the bound defining what is considered a low probability. This amounts to constructing a **significance level** Process, i.e., a Process of identifying α, and coordinating it with a **p-value** Process by comparing their outputs, namely, α and the p-value. That is, **decision** is a coordinated Process that acts on the outputs of **significance level** and **p-value** Processes and returns the decision about whether or not to reject H_0.

Object: The individual encapsulates a **decision** Process into an Object by viewing it as a totality and considering scenarios in which the conditions would yield the same decision or a different decision about H_0. That is, the transformation being applied to this **decision** Object is possibly a comparison of various situations in which the decision could have been different.

Conclusion. As stated previously, in hypothesis testing, the claim about the population parameter is used to formulate H_0 and H_1. Once a decision is made as to whether or not to reject H_0, a conclusion can be formed about the claim. Forming a conclusion about the claim can be thought of as a transformation, **conclusion**, that acts on the decision about the null hypothesis, namely, "reject H_0" or "fail to reject H_0," and returns "support claim" or "do not support claim." The following are descriptions of the stages of **conclusion**.

Action: A **conclusion** Action is an external transformation on the claim of a hypothesis test, the decision about H_0, and the hypothesis to which the claim corresponds by following a step-by-step procedure, such as the following:

Step 1. Identify which hypothesis corresponds to the claim.
Step 2. Review the decision about H_0.
Step 3. From the following, choose the option that applies to the above information:

"$H_0 =$ claim" and "reject H_0" $\Rightarrow$ "do not support claim."
"$H_0 =$ claim" and "fail to reject H_0" $\Rightarrow$ "support claim."
"$H_1 =$ claim" and "reject H_0" $\Rightarrow$ "support claim."
"$H_1 =$ claim" and "fail to reject H_0" $\Rightarrow$ "do not support claim."

In particular, the individual performs the above Actions by relying on memorization or a formula sheet, without understanding the logic behind the steps.

Process: The individual interiorizes the above Actions into a Process by describing the steps in general terms. This amounts to a **hypotheses** Process being reconstructed to act on the claim and put it in correspondence with the appropriate hypothesis. Then this reconstructed **hypotheses** Process is coordinated with a **decision** Process to construct a **conclusion** Process. In particular, a **conclusion** Process acts on the outputs of a **decision** Process and a reconstructed **hypotheses** Process and

returns the conclusion about whether or not to support the claim.

Object: The individual encapsulates a **conclusion** Process into an Object by considering that additional hypothesis tests can be performed on the same (or different) claim, understanding the different scenarios in which the above options would arise. That is, the transformation being performed on this **conclusion** Object is possibly a comparison of various scenarios in which the conclusion could have been different.

We must reiterate that this genetic decomposition is preliminary. In general, a goal of research involving APOS Theory is to revise, if necessary, the genetic decomposition, based on empirical results. Our particular goal was to investigate how, if at all, the above constructions emerged in students' reasoning and to determine if students made other constructions that we did not consider in our preliminary genetic decomposition. As we will demonstrate, we indeed found evidence suggesting the need to revise our genetic decomposition. This revision will be discussed after our presentation of the results.

4 Method

The participants of our study were students enrolled in an introductory statistics course based on the emporium model at a large public university in the southeastern United States. The emporium model, which originated at Virginia Polytechnic Institute and State University (Virginia Tech), is comprised of "interactive computer software, personalized on-demand assistance, and mandatory student participation," each of which are crucial components to this model (Twigg, 2011, p. 26).

At the institution where our study was conducted, students were required to spend three academic hours per week in an interactive mathematics computer lab, as well as attend a class for one academic hour per week with an instructor. The time in the mathematics lab was spent engaged in online assignments delivered through Pearson Education's MyStatLab. Additionally, graduate and undergraduate lab assistants, as well as instructors, were available to answer students' questions. The standard textbook for the course was *Elementary Statistics Using Excel* by Triola (2014), adapted by Pearson Education for this particular university.

4.1 Data collection

Data collection took place during the Fall 2014 and Spring 2015 semesters through semi-structured interviews in which students were asked to elaborate on their thought processes after working on two real-life hypothesis testing problems. An invitation to be interviewed was extended to all students from a subset of classes taught by members of the research team. Twelve students volunteered, and there were ten interviews—eight interviews with one participant each and two interviews with two participants each.

Immediately prior to the interview, the participants worked alone for approximately thirty minutes on two questions involving hypothesis testing. They were allowed to use Microsoft Excel and/or a calculator to assist with obtaining their solutions. The interviews,

in which the participants discussed their solutions with a researcher, were approximately one hour long and video and audio recorded. If participants wished to write anything during the discussion, it was done using red ink to be distinguished from their original work. Conducting the interviews was divided among five members of the research team, who all followed the same protocol. When two participants were interviewed simultaneously, the interviewer prompted them to alternate turns on speaking first when discussing their answers to each part of the instrument questions, so that each student would have the opportunity to share their original thoughts. The relevant data from the interviews consisted of the participants' written work, Microsoft Excel files, and verbal discussion of their solutions.

This study was conducted with Institutional Review Board approval. Participants received credit toward the time required to spend in the lab. Prior to the data analysis, all student names were changed to pseudonyms.

4.2 Instrument

In this course, the parameter of interest is either the population proportion or the population mean (depending on the nature of the question). Thus, our instrument for the interview contained two questions, one pertaining to population proportion and the other pertaining to 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.

The participants had seen altered versions of these questions on previous assignments on MyStatLab. The questions were further broken down into multiple parts to which the participants were likely accustomed. These subquestions asked to (a) form the statements of the hypotheses, (b) calculate the test statistic, (c) calculate the p-value, and (d) form a conclusion about the null hypothesis and a conclusion about the claim. The only difference between the interview questions and the MyStatLab questions was that the MyStatLab questions had multiple choice options or drop down menus for the hypotheses statements, decision about the null hypothesis, and conclusion about the claim. The purpose of making these objectives free response during the interview was to encourage students to elaborate on their reasoning. Because the students had already taken an exam covering hypothesis testing, they were expected to know how to conduct and interpret hypothesis tests for both interview questions.

4.3 Method of data analysis

The video and audio recordings of the interviews were distributed among the six members of the research team, who transcribed the recordings. After the transcriptions were completed, the researchers divided themselves into three pairs, and the analysis of the transcriptions was organized in a way so that each interview transcription was reviewed by two pairs of researchers. The data set (transcriptions, written work, and Excel files) was analyzed and coded according to the stages of the mental structures in APOS Theory (described in Section 3). After each interview's codes were agreed upon by its corresponding pairs of researchers, the results were discussed with the entire research team. Following copious deliberation, the research team organized the data around the following five mental constructions, which made up our preliminary genetic decomposition: **hypotheses**, **test statistic**, **p-value**, **decision**, **conclusion**. The data and how they were coded were then used to develop individual learning trajectories[4] for each participant, focusing on the aforementioned constructions. With these individual learning trajectories, our analysis initially sought to classify each student's stage of understanding of each concept and of hypothesis testing overall. We found, however, that students exhibited only partial and, at times, inconsistent conceptions. Consequently, we revisited the analysis from a different perspective, particularly by looking for instances in which students provided evidence of an Action, Process, or Object and comparing them against our preliminary genetic decomposition. Isolated instances are not definitive evidence of a student's conception. Thus, the illustrations that we provide in Section 5 should be viewed as examples of Actions, Processes, and Objects, rather than characterizations of students' conceptions.

Below we present results obtained from the APOS-based analysis of student responses to the two given hypothesis testing problems.

5 Results

While performing a hypothesis test, it is conventional for an individual to formulate the hypotheses about a population parameter, evaluate the test statistic, find the p-value, compare the p-value to the significance level, form a decision about the null hypothesis, and form a conclusion about the claim. These objectives served as the rationale behind the constructions called for by our preliminary genetic decomposition, namely, **hypotheses**, **test statistic**, **p-value**, **decision**, and **conclusion**. The purpose of this section is to provide representative examples of how these constructions emerged as Actions, Processes, and Objects in the group of students we interviewed, as well as provide examples of how students made additional constructions not called for by our preliminary genetic decomposition. Additionally, in this section we highlight instances where students appeared to invoke or coordinate certain Schemas from the preliminary genetic decomposition. Our results are organized around the primary constructions of **hypotheses**, **test statistic**, **p-value**, **decision**, and **conclusion**.

[4]In APOS Theory, a genetic decomposition is a description of the mental constructions that should be made by an arbitrary student. By *individual learning trajectory*, we refer to the mental constructions that were made by a particular student. Our choice of the word "trajectory" should not be interpreted to mean linear. For more information about learning trajectories, see Weber, Walkington, and McGalliard (2015).

5.1 Hypotheses

Hypothesis testing involves two hypotheses, or statements about a population parameter. These statements are called the null hypothesis, H_0, and the alternative hypothesis, H_1. For the course in which our participants were enrolled, the null hypothesis is introduced as a statement that the value of a population parameter is equal to some particular value, while the alternative hypothesis is introduced as a statement that the value of a population parameter differs in some way from this particular value (Triola, 2014, p. 409).

Our preliminary genetic decomposition called for the construction of a mental structure, **hypotheses**, which, as a transformation, acts on the claim of the hypothesis test and returns the null and alternative hypotheses. Below, we illustrate how this construction emerged in the students that we interviewed.

5.1.1 Action—hypotheses

A **hypotheses** Action is formulating H_0 and H_1 in response to external cues, such as keywords or a memorized procedure.

Such an Action is illustrated in the following excerpt from Nicole, which contains her reasoning about Question 1.

I: So for the first one, we will have you answer, and we'll start with the null hypothesis.
Nicole: OK, I put equals 0.7 and does not equal 0.7 for my alternative.
I: So how did you come up with that?
Nicole: Um, well, my teacher taught me null is always equals. So, equals. Um, for the alternative, it said the claim that 70% of people, so it's like an exact number. It's not greater or less than, so I knew it had to be a two-tailed test.

Nicole explained that her teacher taught her the convention that the null hypothesis is a statement about the parameter being equal to some value. Based on her explanation, it seems as though this convention for writing the null hypothesis belonged to someone else (her teacher), and Nicole merely adhered to it. Thus, this convention was external, evidence of a **hypotheses** Action.

5.1.2 Process—hypotheses

As a Process, **hypotheses** is an internal transformation that acts on the claim as an Object and returns H_0 and H_1. A **hypotheses** Process is characterized by an awareness that, in general, the claim is used to formulate H_0 and H_1.

Such a Process is illustrated in the following excerpt from Steve, which contains his reasoning about the hypotheses for Question 1. Prior to this excerpt Steve had just finished giving an overview of the concepts involved in Question 1.

I: So here, so you've already stated the null and alternative hypotheses, and just to be on record, say what you listed as null and alternative hypotheses.

Steve: Um, well, when you're doing null and alternative you always focus on the claim they give you. Um, so 70%, and just to make things easier, uh, we do the null is equal to .7, and then the alternative would be whatever you're asking. In this case you're asking, is it 70%. So you use not equal to 70%.

⋮

I: Do you think that can be altered in any way, the 70%? It's fine. You did very well. Let's see, what kind of test is this in terms of left-tailed, right-tailed, two-tailed?

Steve: This is a two-tailed test.

I: OK, and, um, you knew that because ...

Steve: Because you're using not equals, because it didn't ask you, it's not asking the claim that it's greater than 70% or less than 70%. It's just asking if it's 70%.

Steve explicitly acknowledged, in general terms, that the claim is used to formulate H_0 and H_1. Also, when he said, "just to make things easier, uh, we do the null is equal to .7," he was acknowledging the convention for H_0 to be a statement about equality, while his use of the pronoun "we" suggests that he has adopted this convention to be his own. In addition, by referring to the alternative hypothesis as "whatever you're asking," he was able to explain in some way what the alternative hypothesis means. For this reason, we consider Steve to have provided evidence of a **hypotheses** Process. Furthermore, when Steve discussed the tails of the test, he was invoking his **distribution** Schema. Although Steve did not explicitly reference the parameter in his hypotheses, we consider this to be colloquial language and not negatively reflective of his understanding.

5.1.3 Object—hypotheses

An individual has encapsulated a **hypotheses** Process into an Object when he or she can think of it as a totality and perform additional transformations on it, such as comparing how to formulate the hypotheses between different situations.

Such an Object is illustrated in the following excerpt from Steve, which contains his reasoning about the hypotheses for Question 2. Prior to this excerpt, Steve provided an overview of the concepts involved in Question 2.

I: Can you read out your null and alternative hypothesis to me?

Steve: OK. I just did the same thing I did with proportion, and I said the null is equal to, um, 210, in this case, and, uh, the alternative is greater than 210. But the only reason I said that is because, um, this bottom line of the question says, test the claim that the sample is from a population, um, with a mean greater than 210.

Despite the fact that the questions on the instrument pertained to two different contexts, Steve said, "I just did the same thing I did with proportion." Steve's sense of "sameness" is further illustrative of a **hypotheses** Process, instead of isolated Actions that vary between different situations. Furthermore, he used the phrase, "in this case," to indicate that in his mind he distinguished his procedure for Question 2 from his procedure for Question 1, which he classified as pertaining to proportions. In order to be able to describe his procedures as

the same, while also distinguishing between them in the different situations in which they arose, he had to have compared them, which is evidence of a **hypotheses** Object. Note that although we previously stated that Steve provided evidence of a **hypotheses** Process, we are not contradicting ourselves. APOS Theory acknowledges that in mathematics it is cognitively necessary to view a concept as both a Process and an Object, depending on the situation in which it arises.

5.2 Test statistic

In hypothesis testing, a test statistic is a value corresponding to a sample statistic that is used to make a decision about the null hypothesis under the assumption that the null hypothesis is true (Triola, 2014, p. 411). For each of the questions in our instrument, one of the following formulas should be used to convert the sample statistic into a z-score or t-score called the test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}, \quad z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}, \quad t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}.$$

The first test statistic formula corresponds to a sample proportion, $\hat{p}$, and the second and third formulas correspond to a sample mean, $\overline{x}$. The choice of the formula corresponding to $\overline{x}$ depends on whether the population standard deviation, σ, is known.

Our preliminary genetic decomposition called for the construction of a mental structure, **test statistic**, which, as a transformation, acts on various population parameters and sample statistics and returns a standardized value, namely the test statistic, which is the relative distance of the sample statistic from the assumed value of the parameter. Below, we provide illustrations of how this construction emerged among the students we interviewed (see also Burns-Childers et al., 2017).

5.2.1 Action—test statistic

A **test statistic** Action is an external transformation on values of certain population parameters and sample statistics. In particular, this Action is inputing the required values of the population parameters and sample statistics into a formula and simplifying the resulting expression to obtain the test statistic. Evidence of a **test statistic** Action could be the inability to interpret a test statistic's value beyond how it was calculated.

The following excerpt illustrates how Shannon only described the concept of test statistic in terms of particular calculations. Prior to this excerpt, Shannon and the student whom she was interviewed alongside just finished discussing their hypotheses statements.

I: Alright, so how about the next one? You started us off, so you can start with the test statistic.

Shannon: OK, so I knew that the formula you had to use was the z equals p-hat minus p over square root of p times q over n, so I wrote each, like on Excel, like I plugged in the values on Excel, and I also did them on paper too.

I: OK.

Shannon: So I got p-hat was 0.784 minus 0.7 which is p, over the square root of ... This is your standard deviation right? And then 0.7 times 0.3 over 750, which is your n, which got me to 0.084 over 0.167, which got me to 5.02.

I: OK.

Shannon: Which seemed like a high value to me.

I: Why?

Shannon: I felt like it was extremely high. So, that's what I got for the z.

Shannon explained that she calculated the test statistic both on paper and by using Excel. In order to obtain the same result, she must have applied her **representation** Schema to consider the hand calculations and the equivalent Excel syntax. In addition, two things stand out in the above excerpt that are evidence of a **test statistic** Action. First, Shannon described the test statistic only in terms of the particular calculations that she had performed. Second, Shannon seemed to recognize the test statistic as a z-score and commented on how high it is. However, she was not able to explain why she believed it to be high. That is, she did not clearly state that the test statistic was high in comparison to other values, nor did she account for what led to this high value. Thus, we consider Shannon to have provided evidence of a **test statistic** Action. As we will see in the remaining sections of 5.2, a deeper understanding of the test statistic can be associated with coordinating more of the prerequisite Schemas from the preliminary genetic decomposition. With that said, it is possible that Shannon was unable to explain why she felt her test statistic was high because she did not effectively coordinate additional Schemas.

5.2.2 Process—test statistic

A **test statistic** Process is an internal transformation on population parameters and sample statistics. Evidence of a **test statistic** Process could be interpreting the value of the test statistic, such as it being the distance from the assumed value of the parameter (based on the null hypothesis), without relying on the steps of how it was calculated.

As shown in following excerpt, Lana provided evidence of a **test statistic** Process through her graphical interpretation of the test statistic for Question 1.

I: Could you explain what the test statistic is?

Lana: Let me see if I can get this.

I: OK.

Lana: I think that, I'm picturing the big curve, the bell curve, and I'm picturing the test statistic is where the point that falls on there and anything ... OK, 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 am thinking.

I: So you started to talk about the bell curve, could you draw a picture?

Lana: OK.

I: And then?

Lana: So p-hat is like point, on this it's like, when I figured it out it was like 0.78 something, so I feel like it, and this is, I guess this is like the z-score away from this?

I: OK.

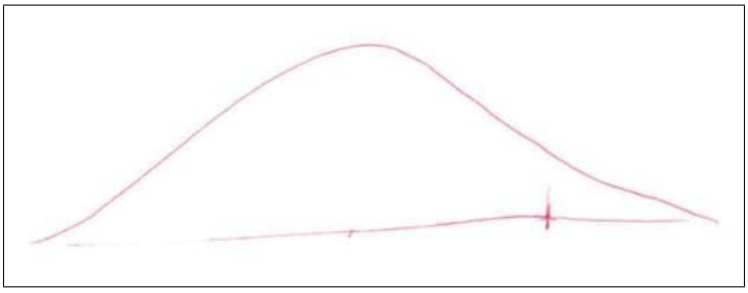

Figure 3: Lana's sketch when describing what a test statistic is.

When the interviewer prompted Lana to interpret the concept of test statistic further, she provided evidence of coordinating her **distribution** Schema with her **representation** Schema to "picture the big curve." Then she sketched a rough picture to show the relative placement of the test statistic on the normal distribution curve (see Figure 3). Lana explicitly stated that the test statistic is a distance from the center of the distribution.[5] We consider this to be evidence of a **test statistic** Process.

5.2.3 Object—test statistic

A **test statistic** Process has been encapsulated into an Object when it is viewed as a totality and transformations can be performed on it, such as comparing how extreme a test statistic value is to other test statistic values or describe a situation that would result in an extreme value of the test statistic.

The following excerpt is taken from the end of Steve's discussion of Question 1. He had just formed a decision to reject the null hypothesis before reflecting on how his decision could have been inferred from an earlier step. In particular, Steve described what accounted for an extreme value of the test statistic.

Steve: But going back on it, it makes sense, you know, if you've got a p-hat that, that's very, very different from your, from your p, you know, 78 is a whole 8% off of, uh, the 70%. And also your test statistic is very large. I'm not totally sure what a test stat is, but it reminds me of z-scores, and I remember when you have a z-score that gets above 3, it starts to get pretty, pretty crazy. So 5 is huge, which is also the reason that you're getting a bunch of zeros or very close to 1.

Steve appeared to have encapsulated into an Object the Process described by $z = (\hat{p} - p)/\sqrt{pq/n}$ in order to consider how it resulted in an extreme value of the test statistic. He explained that a large value of the test statistic resulted from having a value of the sample statistic that is very different from the value of the population parameter in the null

[5]Lana said that the center of the distribution corresponds to the mean. However, Question 1 does not pertain to means. Nevertheless, Lana understood that the test statistic is a distance from the center.

hypothesis. APOS Theory acknowledges, in general, that it is necessary to de-encapsulate an Object back into a Process, which appears to be the case with Steve. That is, he de-encapsulated his **test statistic** Object back into a Process to consider the difference between $\hat{p}$ and p. Furthermore, attempting to look for what led to an extreme value of the test statistic indicates that Steve applied his **empirical evidence** Schema.

We should note that based on Steve's statement, "I'm not totally sure what a test stat is, but it reminds me of z-scores," he appeared to have constructed isolated Processes for each test statistic, which he needed to further coordinate in order to construct a single **test statistic** Process. Despite this, Steve provided the clearest evidence of a **test statistic** Object. Our preliminary genetic decomposition did not consider the construction of separate Processes corresponding to each type of test statistic.

Evidence of a **test statistic** Object could also be distinguishing between various test statistic formulas and determining which one is appropriate for a given situation. The following excerpt comes from Haley's initial discussion of the concepts involved in Question 2. She distinguished between the different types of test statistics by explaining when a test statistic pertaining to means would be a t-score or a z-score.

I: OK, so what were your thoughts when you read this one?
Haley: Um, I knew it was going to be, um, a t because it said like, it gave you like, um, it said it was a simple random sample of 40 hit songs [*mumbling*] blah, blah, blah, and it gave me like the standard deviation and the mean of the sample. And then it said in a manual [*mumbling*], it said that a song must be no longer than 210 seconds. So since no longer than, I knew it was less for the alternative.
I: OK, and, um, so, let's see. Alright. Um. So what concepts are being used in this question?
Haley: Um, like what do you mean by concepts?
I: So we're talking about, um, I heard you say the mean and standard deviation from the sample and things like that, and we're doing hypothesis testing, right?
Haley: Like, what, like if it was a t-test or something? Like it was ...
I: Mmhm.
Haley: Well it was a t, and it was a left, like, it was a left-tail test.
I: So can a question like a hypothesis test about means, can it ever be a z-test? Like using the z-scores in a normal distribution?
Haley: Yeah.
I: When would that happen?
Haley: When they give you the population standard deviation.
I: OK, and so you knew that this was a sample standard deviation because of what?
Haley: Simple random sample [*reading part of the problem*].
I: OK, and, um, do you know what symbol we use to represent the sample standard deviation?
Haley: The s [*writes on sheet*].

Haley explained that she knew the problem dealt with t-scores by stating, "it said it was a simple random sample [...] and it gave me like the standard deviation and the mean of the

sample." After this, the interviewer asked what concepts were being used and clarified the question by summarizing what Haley already said about the "mean and standard deviation of the sample." Haley inferred that the interviewer was asking if it is a t-test. This part of the excerpt shows that Haley directly associated t-scores with sample statistics. When asked whether a hypothesis test about means could ever be a z-test, Haley explicitly stated that the test statistic would be a z-score if the population standard deviation is known. Also, she knew that the given standard deviation was from a sample by interpreting the phrase "simple random sample" instead of looking for the phrase "sample standard deviation" or the symbol s. Haley's awareness of the different types of test statistics suggests she has developed a **distribution** Schema. Furthermore, her ability to distinguish between the different types of test statistics in terms of the situations in which they would arise is evidence of a **test statistic** Object.

Our preliminary genetic decomposition did not consider a **test statistic** Object by distinguishing between various test statistic formulas in order to determine which one is appropriate for a given situation. Thus, our results suggest the need to revise our genetic decomposition.

Later, in Section 5.3.2, we will see another example of a **test statistic** Object. We hold off on this example for now because it is related to a **p-value** Process, yet to be discussed.

5.3 *P*-value

Triola (2014) defines 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" (p. 414). The students in this course calculated p-values by using Microsoft Excel DIST functions, which, in general, return the probability that something is less than a given value. In particular, NORM.S.DIST is used in the case of a normal distribution, while T.DIST is used in the case of a t-distribution. In some cases, it is necessary to subtract this value from 1 and/or multiply by 2, depending on the tail of the test.

Our preliminary genetic decomposition called for the construction of a mental structure, **p-value**, which, as a transformation, acts on the test statistic and returns a probability—a number between 0 and 1. Below, we provide examples of how this construction emerged in the group of students that we interviewed.

5.3.1 Action—*p*-value

A **p-value** Action is an external transformation of a particular value of a test statistic that results in the p-value. Evidence of a **p-value** Action could be the reliance on step-by-step calculations and the inability to interpret the result of those calculations.

In the following series of excerpts, we see that Shannon could perform the necessary steps for calculating a p-value, but she appeared to have little understanding of the p-value beyond her calculations. The first excerpt is from Shannon's discussion of the p-value for Question 1, which took place immediately after another student discussed the p-value.

Shannon: So I used Norm.S.Dist for my Excel function, because you don't have to put if it

is true or false, it just asks for the z-score. So I put Norm.S.Dist, I plugged in the z of 5.02 and I got 1 and subtracted that from 1 and I got 0.

I: OK.

Shannon: And I know for two tail you have to multiply by 2.

I: Mmhm.

Shannon: And I got 0 times 2 is 0.

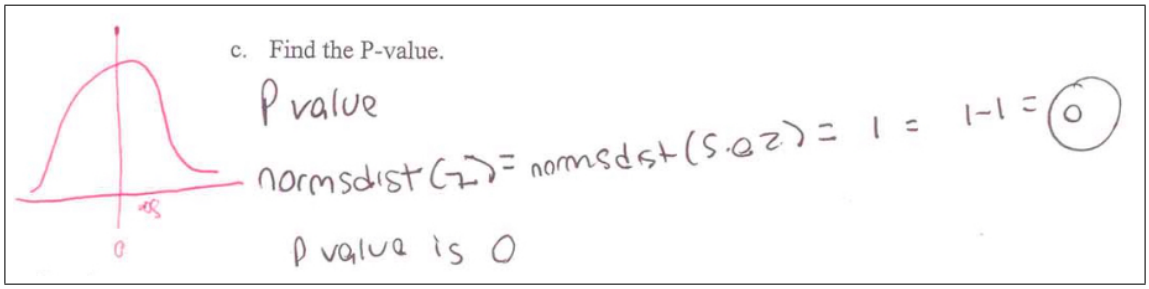

Figure 4: Shannon's calculation of the p-value along with her drawing for Question 1.

To calculate the p-value, Shannon appeared to apply her **function** Schema by explaining that she "plugged in" the z-score into an Excel function. What is notable about her description is that she did not provide any justification or interpretation of her solution. Instead, she only reiterated verbally the step-by-step calculations from her written work (see Figure 4) that resulted in her p-value of 0. Furthermore, Shannon's misuse of the equal symbol in her written work to say, "$1 = 1 - 1$," suggests that she was merely completing a sequence of memorized steps to calculate the p-value for a right-tailed test. In other words, Shannon knew to subtract the output of the Excel function from 1, but could not explain why. From her drawing we were not able to ascertain anything, such as where the sample statistic, p-value, and significance level fall in relation to the center of the distribution (see Figure 4).

For Question 2, however, Shannon's discussion of her drawing (see Figure 5) was more conclusive. The following excerpt shows that she was not able to interpret the meaning of the p-value.

I: Alright, you put more detail into this one than last time. So what is this tick mark down here?

Shannon: So this, I guess, would be ... this is 210, this one is 231, and you are trying to find the sample from songs greater than, so I colored it in past 210.

I: OK.

Shannon: And then I guessed ... I don't know, I just don't draw pictures.

To draw a curve and shade a region, Shannon applied her **representation** Schema. Also, it is possible that she utilized a **distribution** Schema, but it should be noted here that she mixed standardized scores with sample data on the horizontal axis. To justify the shading in her drawing, Shannon said, "you are trying to find the sample from songs greater than, so I colored it in past 210." Due to the incompleteness of this statement and the lack of identifying the p-value as a probability, it is likely that Shannon used the phrase "greater

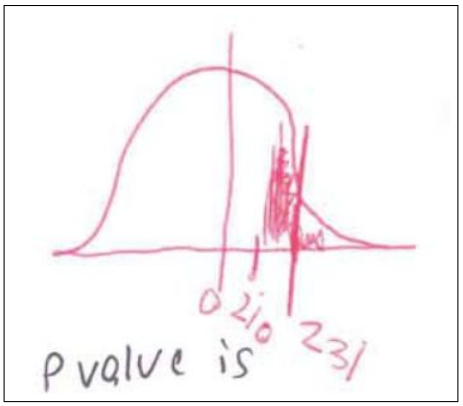

Figure 5: Shannon's drawing for Question 2.

than" as a cue, telling her in what direction to shade. Also, she never mentioned the p-value in relation to the region that she shaded. This is evidence of a **p-value** Action.

5.3.2 Process—*p*-value

A **p-value** Process is an internal transformation of the test statistic, which results in a probability called the p-value. Evidence of a **p-value** Process could be the ability to describe, in general terms, the transformations on the test statistic that result in the p-value, while recognizing that the p-value is a probability (represented graphically as an area).

The following excerpt contains Nicole's reasoning about the p-value for Question 1, which she represented graphically in Figure 6. Nicole had just finished explaining that the number she found is a z-score.

Nicole: Yeah, because we are finding z equals. OK. And so the tails, which is what are on the right and what's on the left of that line is, well, this is the z-score, and we are going to find the area of this later.

I: Right, so let's go ahead and we'll start with that. So it's your turn for p-value.

Nicole: Um, well I knew it was a two-tailed because it's not equal to.

I: OK.

Nicole: Its alternative. And I actually got, I did on Excel, I used Norm.S.Dist.

I: Mmhm.

Nicole: Um, and you plug in z, which is, we already found the test statistic and then 1. Anyways, um, I did 1 minus that because that's only going to show me what's to the left of the 5.

I: OK.

Nicole: And I need to know what's on the right of that. And it's a two-tailed, so I multiplied that one by two, the one I already subtracted by 1.

I: OK.

Nicole: Which gave me zero.

I: So what was it that we were finding here?

Nicole: The area of the tails.

I: Perfect. Alright.

Figure 6: Nicole's drawing for Question 1.

Several things stand out in the above excerpt. First, when Nicole said, "you plug in z," she was describing a z-score as an input of a transformation, evidence that she applied her **function** Schema. Second, when she continued, "which is, we already found the test statistic," she indicated that, to her, z is a concept, not just a number. In particular, she was able to think about **test statistic** as a totality to which some transformation could be applied, meaning that she encapsulated her **test statistic** Process into an Object. This provides another example of a **test statistic** Object, in addition to those illustrated in Section 5.2.3. Third, when Nicole said, "I did 1 minus that because that's only going to show me what's to the left of the 5," she acknowledged that the total area under the curve is 1. This is evidence that she applied her **probability** Schema. Now, these observations together suggest that Nicole constructed a **probability function** Process and coordinated it with another Process related to the alternative hypothesis and the tail of the test. Mathematically, this second Process appears to be closely related to the **hypotheses** Process. Thus, we consider Nicole to have coordinated her **hypotheses** Process with her **probability function** Process to construct her **p-value** Process. Generally speaking, Nicole's **p-value** Process acted on **test statistic** as an Object and returned the p-value. Throughout her reasoning, when Nicole referenced the "tails," she was invoking her **distribution** Schema, and she was able to represent the p-value graphically in Figure 6 by using her **representation** Schema.

Recall that Nicole provided evidence of a **hypotheses** Action in Section 5.1.1, while here we are suggesting that she utilized a **hypotheses** Process. To explain this discrepancy, it is possible for an individual provide evidence of a Process related to a concept in one instance but not in others. This is perhaps because a Process was not cognitively necessary for Nicole to complete the earlier task.

Our preliminary genetic decomposition did not explicitly state that the encapsulation of a **test statistic** Process into an Object is the result of a **p-value** Process acting on it. This suggests a refinement to our preliminary genetic decomposition.

5.3.3 Object—*p*-value

A **p-value** Process has been encapsulated into an Object when it is viewed as a totality and transformations can be performed on it, such as comparing the p-value to the significance level, particularly as probabilities or areas. While we indeed found evidence of this, we hold off on this discussion until Section 5.4.2, because it is related to a **decision** Process, yet to be discussed.

Other evidence of **p-value** Object could be distinguishing between various procedures for calculating the p-value and determining which one is appropriate for a given situation. To illustrate this, we consider the following two excerpts from Steve's discussion of the p-value for Question 1, in which he explained various procedures for calculating the p-value, depending on the situation. Prior to this excerpt, Steve struggled to graphically represent the p-value, but this struggle was resolved shortly after this except (as we will see in Section 5.4.2).

I: And why did you choose that Excel formula?

Steve: Well, whenever you're finding a p-value you're doing a DIST function, and when you're doing proportions, it's NORM, and when you're doing means, it's T. So in this case we used NORM.S.DIST 'cause I think the other formula is silly. But, uh, since it's a two-tailed test I couldn't just stop there. I had to 1 minus that and then double it.

I: OK, OK. And you did the 1 minus, why?

Steve: Um, because if you don't do 1 minus, it ends up being something very, very close to 1. So a bunch of .9999 ..., and you can't double that. Whenever I got stumbled, I was like, oh wait, do I, uh, do I double the 1 minus or it by itself. Well, you can't go over 1. It can't go over 1.

I: OK.

Steve explained that an Excel DIST function is used "whenever you're finding a p-value," while also acknowledging that a p-value "can't go over 1." This suggests that he coordinated his **function** and **probability** Schemas to construct a **probability function** Process. By describing the general steps of his calculations and by referencing the tail of the test, Steve, like Nicole, appeared to have coordinated his **probability function** Process with his **hypotheses** Process to construct a **p-value** Process. Furthermore, when Steve described situations in which NORM.S.DIST and T.DIST are used, he provided evidence of having encapsulated his **p-value** Process into an Object. Although Steve was not completely correct in stating that T.DIST is always used in the context of means, he clearly compared different procedures for calculating the p-value and considered situations in which these procedures would arise.

The next excerpt, which is a continuation of the previous one, contains Steve's description of the additional transformations performed on the result of the .DIST function, depending on the tail of the test.

Steve: Of course, you usually end up doubling the right-tailed test anyway.

I: You end up doubling the right-tailed test?

Steve: Yeah, so it's like, um, layers within itself, like a Russian nesting doll. So your NORM.S.DIST is left-tailed test, and if you want to go to right-tailed, you do 1 minus, and if you want to go to two-tailed, you just double that.

I: OK. What if this had been a -5.02?

Steve: Then, um, the NORM.S.DIST would give you a, would give you a .0000 ... like very close to 0. Basically it would flip the right-tailed and left-tailed. Oh, that makes so

much more sense now. Sorry.

Steve explained that the procedure for calculating the p-value depends on the tail of the test. However, instead of only explaining the procedure for this particular problem, he explained what procedure would be performed for the different situations in which the test was left-tailed, right-tailed, or two-tailed. Also, he was able to describe how the p-value would change for a test statistic of -5.02, indicating that he could consider multiple scenarios and how they affect the p-value. We consider this to be further evidence of a **p-value** Object.

Our preliminary genetic decomposition did not consider a **p-value** Object by distinguishing between various procedures or Excel functions for calculating the p-value in order to determine which one is appropriate for a given situation. Thus, our results suggest the need to revise our genetic decomposition.

5.4 Decision and conclusion

In hypothesis testing, we make a decision about whether or not to reject the null hypothesis, H_0, by comparing the p-value to the significance level. Triola (2014) defines the significance level as follows:

> The significance level (denoted α) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. If the test statistic falls in the critical region, we reject the null hypothesis, so α is the probability of making the mistake of rejecting the null hypothesis when it is true (p. 413).

The students in this course, as with most introductory level statistics courses, used a predefined significance level (identified within the question) as an upper bound for the p-value when trying to decide whether or not to reject H_0. In particular, we reject H_0 if and only if the p-value is less than or equal to α. This decision is then used to form a conclusion about the claim. If the null hypothesis is the claim, and if the null hypothesis is rejected, then we conclude that there is not sufficient evidence to support the claim. If the alternative hypothesis is the claim, and if the null hypothesis is rejected, then we conclude that there is sufficient evidence to support the claim.

Our preliminary genetic decomposition called for the construction of mental structures called **decision** and **conclusion**. Although independent from each other, these constructions were often intertwined in our results. For this reason, we illustrate them together in this section.

5.4.1 Action—decision and conclusion

A **decision** Action is characterized by forming a decision about H_0 in accordance to an external procedure involving the comparison of the p-value and the significance level, α. Meanwhile, **conclusion** as an Action is characterized by forming a conclusion about the claim in accordance to an external procedure involving the decision about H_0. Evidence of such Actions could be forming a decision about H_0 and a conclusion about the claim without being able to explain the logic behind the decision and conclusion.

The following excerpt, taken from Shannon's discussion of Question 1, illustrates a **decision** Action, as well as a **conclusion** Action. The dialogue took place right after Shannon explained her calculation of the p-value and sketched her drawing (see Section 5.3.1).

Shannon: Um, what I looked at in the examples and what I remember is like, if the p-value is less than the significance level, then you do reject? I think so, I think that's what I remember.

I: You don't look very confident about it.

Shannon: So I wrote the p-value is less than the significance level, therefore we reject the null hypothesis. And then this is where I get confused, if it is sufficient or insufficient, and I put there is sufficient evidence to support the original claim.

I: OK, so you memorized, you don't know why?

Shannon: I just don't, yeah.

Shannon explicitly stated that she was operating based on what she remembered from previous examples. In particular, she rejected H_0 based on a memorized rule, indicative of a **decision** Action. Furthermore, she was unsure of her conclusion about the claim, but she concluded that there is sufficient evidence to support the claim. However, in Question 1, the claim corresponded to H_0. Thus, based on her rejection of H_0, Shannon should have concluded that there is insufficient evidence to support the claim. This suggests that Shannon was relying on her memory of previous examples instead of reflecting on the logic behind determining the conclusion about the claim. We consider this to evidence of a **conclusion** Action.

5.4.2 Process—decision and conclusion

A **decision** Process is characterized by an internally driven transformation that involves comparing the p-value and significance level, α, in order to form a decision about H_0. Evidence of a **decision** Process could be viewing the p-value and α as probabilities or areas and explaining, in general, how their comparison leads to a decision about H_0 without needing to perform all of the steps. Meanwhile, a **conclusion** Process is characterized by an internally driven transformation that involves forming a conclusion about the claim based on the decision about H_0. Evidence of a **conclusion** Process could be explaining, in general, how to form a conclusion about the claim without needing to perform all of the steps.

To illustrate a **decision** Process, we first consider the following excerpt from Steve, which provides evidence that he compared the p-value and significance level as areas. Shortly before this excerpt, Steve struggled to graphically represent the p-value. After explaining his calculations (see Section 5.3.3), he was able to resolve his struggle.

Steve: Oh wait! Wasn't the p-value supposed to be from the edge? So wasn't the p-value supposed to be like this ... [*draws on paper*] ... the stuff on the outside? I remember now. It was um ... I don't see how that relates to those, but I know it relates to the significance level 'cause your .05 is going to be outside of that.

Throughout this excerpt, Steve described aspects of the distribution curve while drawing

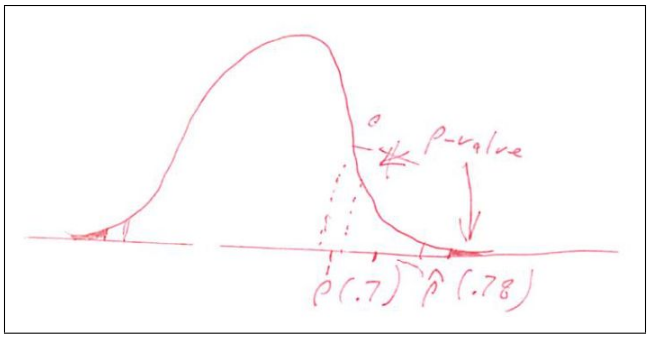

Figure 7: Steve's graph of the p-value for Question 1.

a graph. Thus, he was invoking his **distribution** and **representation** Schemas. At first glance, when Steve asked, "wasn't the p-value supposed to be from the edge," he could have been describing the test statistic as forming the edge of the region whose area is the p-value. It is clear, however, based on Steve's drawing in Figure 7 that he was not aware of a graphical relationship between the test statistic and the p-value, which is likely what he was referring to when he said, "I don't see how that relates to those." Nevertheless, he still viewed the p-value and significance level as areas and compared them by saying, "I know [*the p-value*] relates to the significance level 'cause your .05 is going to be outside of that." To clarify, when Steve said, ".05 is going to be outside of that," we interpret it to mean that the rejection region is not strictly contained in the region whose area is the p-value.[6] We consider this example to be evidence of a **p-value** Object, in addition to the examples discussed previously in Section 5.3.3.

The previous excerpt established that Steve was able to compare the p-value and significance levels as areas, which we consider to be a necessary characterization of a **decision** Process. To further illustrate a **decision** Process, we consider the following excerpt containing Steve's reasoning about whether or not to reject H_0 for Question 1. Note that part of this excerpt was discussed previously in Section 5.2.3.

I: OK, so, and how did you arrive at your conclusion? What did you arrive at?

Steve: I just remembered anytime the p-value is less than the, uh, significance level you reject the null, uh, I think [*laughs*]. But going back on it, it makes sense, you know, if you've got a p-hat that, that's very, very different from your, from your p, you know, 78 is a whole 8% off of, uh, the 70%. And also your test statistic is very large. I'm not totally sure what a test stat is, but it reminds me of z-scores, and I remember when you have a z-score that gets above 3, it starts to get pretty, pretty crazy. So 5 is huge, which is also the reason that you're getting a bunch of zeros or very close to 1.

⋮

Steve: So it's interesting, we always go all the way out to the p-value, but you can pretty much tell from your test statistic if it's correct or not.

[6] According to Triola (2014), the rejection region is a region under the extremities of the distribution curve and has an area equal to the significance level, α. If the test statistic falls within the rejection region, it implies that the p-value associated with this sample is less than α.

Initially, Steve rejected H_0 based on a memorized rule, suggestive of a **decision** Action. However, he reflected on this Action when he said, "but going back on it, it makes sense." Referring to the test statistic, Steve said, "so 5 is huge, which is also the reason that you're getting a bunch of zeros." When Steve described "getting a bunch of zeros," he seemed to be referring to calculating a p-value that has a zero in the tenth, hundredth, thousandth, etc., places. In other words, he related an extreme test statistic to a small p-value. This suggests that he applied his **function** Schema to describe the relationship between the test statistic and the p-value. Also, the contrast in his phrase, "getting a bunch of zeros or very close to 1," indicates he was invoking his **probability** Schema. In the last line of the above excerpt, Steve explained that depending on the magnitude of the test statistic, you can potentially form a decision about the null hypothesis without comparing the p-value to the significance level. The ability to describe the result of a transformation without needing to perform all of its steps is evidence of a Process. Thus, we consider Steve to have provided evidence of a **decision** Process.

To illustrate a **conclusion** Process, we consider the following excerpt from Steve's discussion of the conclusion about the claim for Question 2.

I: In terms of your final conclusion, take me through it.
Steve: Well, you get the same conclusion as you did in the first question because the p-value is also less than the significance level. So you reject the null hypothesis cause there's sufficient evidence against it.
I: Against?
Steve: Against it equaling 210.
I: And what does that mean about your claim?
Steve: Well, you reject it. You know that the population mean is far more likely to be greater than 210 than ...
I: And which one was actually, do you know which one was your claim?
Steve: The null is your claim, right?
I: Well, what does it say there? [*Points to problem.*]
Steve: Yeah, test the claim ... OH! The claim is your alternative. Oh well. Either way, it doesn't really change anything. You know that it's greater than.

Steve said that he rejected the null hypothesis because there is sufficient evidence against μ equaling 210 seconds. When the interviewer prompted him to explain his decision about the null hypothesis in terms of the claim, he invoked his **empirical evidence** Schema by saying, "you know that the population mean is far more likely to be greater than 210," which was reasonable considering that this hypothesis test was right-tailed. The remainder of the excerpt shows that Steve mistakenly believed the null hypothesis to be the claim. When he realized that the alternative hypothesis was the claim, he said, "it doesn't really change anything. You know that it's greater than." That is, he knew that his inference remained accurate regardless of which hypothesis was the claim. Because Steve did not find it necessary to go through the step of determining which hypothesis corresponds to the claim in order to make an accurate inference, we consider this to be evidence of a **conclusion** Process.

5.4.3 Object—decision and conclusion

The Processes of **decision** and **conclusion** have been encapsulated into Objects when the individual views them as totalities and can perform transformations on them. A **decision** Object could be characterized by considering scenarios in which the conditions would yield the same decision or a different decision about H_0. A **conclusion** Object could be characterized by considering additional hypothesis tests on the same (or different) claim and understand the different scenarios in which various conclusions about the claim would arise.

None of the students we interviewed provided evidence of a **decision** or **conclusion** Object. One possible explanation for this is that our instrument and protocol did not probe for such reasoning. This suggests the need to update our protocol and suggest instructional strategies to foster the construction of **decision** and **conclusion** as Objects.

5.4.4 Process—significance level

Before we conclude this section, we revisit the concept of significance level. Our preliminary genetic decomposition did not consider **significance level** as a Process other than identifying the significance level in the statement of the problem. However, we found evidence of a deeper understanding of the significance level.

The following excerpt is taken from the beginning of Steve's interview when the interviewer asked him to provide an overview of Question 1. This excerpt contains his description of the significance level, which he gave while sketching his graph (see Figure 7, discussed previously).

I: So what are the statistical concepts that you believe are a part of this?

Steve: Well, I think it has to do with normal distributions and figuring out, um, and just comparing the proportion you get from a sample test to, um, a population proportion that has been reported to you outside the, in this case, you're comparing 588/750 to the 70% claim. But, uh, yeah.

I: OK. Very good.

Steve: Go like that, which isn't perfect, but whatever. I guess the idea is that your p is, um, .7 and, um, I remember from the Excel sheet 588/750 came out to be .78. So p-hat, well that's too high up, say it's like there, .78, if that makes sense. I don't know if that's clear. And using the significance level there's basically a realm around .7, like an area that if your p-hat fell into, then you can accept that .7 is an actual legitimate claim, but if it falls outside of that, then it's not.

When Steve described the significance level as forming a "realm" or "area" around p, which will or will not contain $\hat{p}$, he was describing a rejection region. This suggests that he coordinated his **distribution** and **probability** Schemas to view the significance level as an area under the distribution curve. Because his description references p and $\hat{p}$ instead of z-scores, Steve seemed to be alluding to the notion of confidence interval. To calculate confidence intervals, the significance level is used to calculate a critical value, which is eventually used in finding a lower and upper bound for usual sample statistics. We believe this upper and

lower bound is what Steve referred to as "a realm around .7." It is then possible to determine from this interval if a sample statistic is unusual. Thus, for Steve, the significance level was a transformation on the parameter, p, to create an interval centered at p that will or will not contain $\hat{p}$. Since, in hypothesis testing, the significance level is also used to form a rejection region and determine if sample statistics are unusual, the two are similar processes. Steve exhibiting an understanding of this relationship is evidence of a **significance level** Process.

5.5 Summary

Our preliminary genetic decomposition called for the construction of the following mental structures: **hypotheses**, **test statistic**, **p-value**, **decision**, and **conclusion**. In our presentation of the results, we illustrated how these mental structures emerged in the group of students that we interviewed. Also, we provided evidence that some students made constructions different from how we predicted, such as with **test statistic**, **p-value**, and **significance level**. Below, we provide a summary of our findings.

Students who provided evidence of a **hypotheses** Action appeared to follow an external convention for formulating the statements H_0 and H_1. Students who provided evidence of a **hypotheses** Process acknowledged that, in general, the claim is used to formulate the hypotheses. Students who provided evidence of a **hypotheses** Object were able to compare their procedures for formulating hypotheses between different problems.

Our results pertaining to the concept of test statistic indicate that students might construct isolated **test statistic** Actions, each corresponding to the different types of test statistics. Our preliminary genetic decomposition did not consider different constructions corresponding to each test statistic. Students who provided evidence of a **test statistic** Action did so by focusing on particular calculations, without further interpretation of the result of their calculation. Students who provided evidence of a **test statistic** Process described a test statistic graphically as the distance from the center of the distribution. Evidence of a **test statistic** Object appeared in three ways. The first way was by distinguishing between the various test statistic formulas in light of the situations in which they would arise, which we did not consider in our preliminary genetic decomposition. The second way was by considering what accounted for an extreme value of the test statistic in a situation. The third way was by explaining that the test statistic determines the p-value, suggesting that a **p-value** Process induces the encapsulation of a **test statistic** Process into an Object.

Students who provided evidence of a **p-value** Action did so by only being able to state the step-by-step calculation for a particular p-value and could not represent it graphically. Students who provided evidence of a **p-value** Process tended to explain the logic behind the steps of their calculation or interpret the result of their calculation, either verbally or graphically. Our results suggest the need to refine our description of **p-value** as a Process, which appeared to be the result of coordinating the prerequisite **probability function** Process with a **hypotheses** Process. A **test statistic** Object appeared in two ways. The first way was by distinguishing between the various procedures for calculating the p-value and the situations in which they would arise, which we did not consider in our preliminary genetic decomposition. The second way was by comparing the p-value to the significance level, particularly as areas, suggesting that a **decision** Process induces the encapsulation of a **p-value** Process into an Object.

Students who provided evidence of **decision** and **conclusion** Actions did so by relying on memorized procedures to form a decision about H_0 and a conclusion about the claim, without understanding the logic behind the steps. Students who provided evidence of a **decision** Process were able to explain their decision about H_0 without needing to go through the step of comparing the p-value to the significance level. Instead, the extremity of the test statistic could, at times, be used to form a decision about H_0. Similarly, students who provided evidence of a **conclusion** Process did so by explaining their conclusion about the claim without needing to go through all of the steps. In particular, such students did not need to determine which hypothesis corresponds to the claim in order to state an accurate inference based on the decision about H_0. We did not find evidence of **decision** or **conclusion** Objects, suggesting the need to revise our instrument and interview protocol to probe for such reasoning. This also suggests the need to develop curriculum that fosters such reasoning.

As a Process, **significance level** is constructed through the coordination of **distribution** and **probability** Schemas. A **significance level** Process is a transformation on a parameter (or z-score or t-score of 0) to form a rejection region. While our preliminary genetic decomposition did not consider **significance level** as a Process other than identifying the significance level in the statement of the problem, we found evidence of a deeper meaning by students. We will discuss this further in our revisions to the genetic decomposition in Section 6.1.

6 Discussion

Our results suggest that the concepts involved in hypothesis testing are related through the construction of higher-order, coordinated Processes operating on Objects. It has been widely recognized in APOS Theory literature that encapsulation of a Process into an Object is difficult to achieve (Arnon et al., 2014). This provides a possible explanation for why hypothesis testing is such a challenging topic for students, as reported by other researchers (Dolor & Noll, 2015; Liu & Thompson, 2005; Vallecillos, 2000; Williams, 1997).

Many of the students in our study appeared to follow a "script," phrasing their explanations based on wording they remembered from MyStatLab or their instructor. For example, when explaining how to formulate the hypotheses, Nicole said,"my teacher taught me null is always equals" (see Section 5.1.1). Other students used phrases like, "I remember," without reflecting on what they remembered in order to develop a deeper understanding, such as Shannon when she made a decision about the null hypothesis (see Section 5.4.1). When it came to reasoning about the test statistic and p-value, students without a deeper understanding resorted to merely reiterating their calculations instead of justifying or interpreting their calculations (see, for example, Shannon in Sections 5.2.1 and 5.3.1). These students, thereby, exhibited Actions pertaining to one or more of the concepts involved in hypothesis testing. Our results indicate that being limited to performing Actions can be, in part, attributed to a focus on what Link (2002) referred to as a six-part procedure, in which each step serves as an external cue for the next step.

Some of the students that we interviewed exhibited a deeper understanding of the concepts involved in hypothesis testing. One particular student, Steve, received a substantial amount of attention in this report, due to the fact that he elaborated on his solutions using

his own words, thereby providing a considerable amount of rich data. In several instances during the interview, Steve reflected on his understanding of a concept in order to explore its meaning beyond what he remembered from a step-by-step procedure (see, for example, Section 5.4.2). As a result, he consistently provided evidence of a Process and/or Object pertaining to the concepts in hypothesis testing by explaining them in general terms that would apply across various problem situations. We should note that this characterization of Steve's reasoning is contrary to our preliminary analysis (Burns-Childers et al., 2017), in which we used excerpts from Steve to illustrate a **test statistic** Action. Further analysis found that Steve's construction of Processes and Objects was different from how we predicted in our genetic decomposition (in some instances), but he made these constructions nevertheless.

Throughout our analysis of the interviews, we did not find evidence of the construction of a **decision** Object or a **conclusion** Object. One possible reason for this is that our instrument and protocol did not probe for such an understanding. Also, it is possible that instruction did not lead students to encapsulate **decision** and **conclusion** Processes into Objects. Later, in Section 7, we offer suggestions to facilitate the construction of these mental structures.

Many of the students in our study provided evidence of coordinating Schemas in order to either reason through or complete a sequence of steps with hypothesis tests. In some cases, the successful coordination of Schemas seemed to illuminate additional information for the student, like with Steve who appeared to use all five of the prerequisite Schemas proposed in our preliminary genetic decomposition (see, for example, Section 5.4.2). On the other hand, the inability to effectively coordinate Schemas may lead to an obstacle in understanding a problem. For example, we believe that as a result of not successfully coordinating her **function** and **distribution** Schemas, Shannon was unable to explain why she felt her test statistic was very high (see Section 5.2.1). Activities designed with the goal of coordinating such Schemas may be beneficial for students studying hypothesis testing.

6.1 Revisions to the genetic decomposition

Based on our results in Section 5, we propose several revisions to our genetic decomposition. Due to the length of our genetic decomposition, we provide here only modifications. For the reader's convenience, we include the revised genetic decomposition in its entirety in Appendix A. Recall that Figure 2 in Section 3.1 illustrated how the Processes in the genetic decomposition interact with each other. In this section, we offer figures that show what these Processes look like microscopically.

Our first revision pertains to the concept of test statistic. Instead of constructing a single **test statistic** Action, our results in Section 5.2.3 suggest that distinct Actions be constructed corresponding to each type of test statistic, referred to as **z-score proportions**, **z-score means**, and **t-score**. Once these Actions are interiorized into separate Processes (by the same name as their Actions), they should be pairwise coordinated to construct a single Process, namely **test statistic**. This **test statistic** Process is illustrated in Figure 8. One way for a **test statistic** Process to be encapsulated into an Object, that we did not initially consider, would be by distinguishing between the various test statistic formulas and the situations in which they would arise, as illustrated in Section 5.2.3.

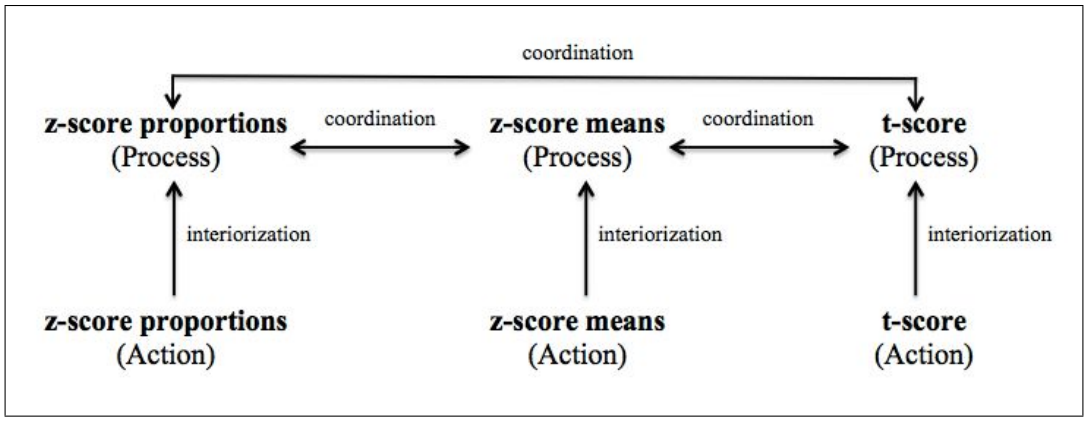

Figure 8: The **test statistic** Process.

Our second revision pertains to the concept of p-value. Our preliminary genetic decomposition proposed that a **p-value** Process is the result of coordinating **hypotheses** and **test statistic** Processes. However, we found in Section 5.3.2 that a prerequisite **probability function** Process should be coordinated with a **hypotheses** Process to construct a **p-value** Process. This **p-value** Process then induces the encapsulation of a **test statistic** Process into an Object. Figure 9 illustrates this **p-value** Process. One way for a **p-value** Process to be encapsulated into an Object, that we did not initially consider, would be by distinguishing between the various procedures for calculating the p-value and the situations in which they would arise, as illustrated in Section 5.3.3.

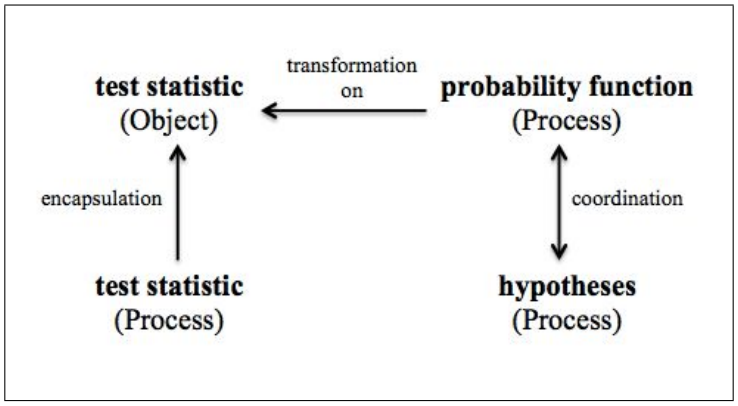

Figure 9: The **p-value** Process.

Our third revision pertains to forming a decision about H_0. Our preliminary genetic decomposition said that a **decision** Process is the result of coordinating **p-value** and **significance level** Processes. Our results in Section 5.4.2 suggest that this coordination involves encapsulating **p-value** and **significance level** Processes into Objects, which can then be compared in order to form a decision about H_0. Figure 10 illustrates this **decision** Process.

Our fourth revision pertains to the concept of significance level. Instead of a **significance level** Process merely involving the identification of the significance level in the statement of the problem, our results in Section 5.4.4 suggest that it is a transformation on existing

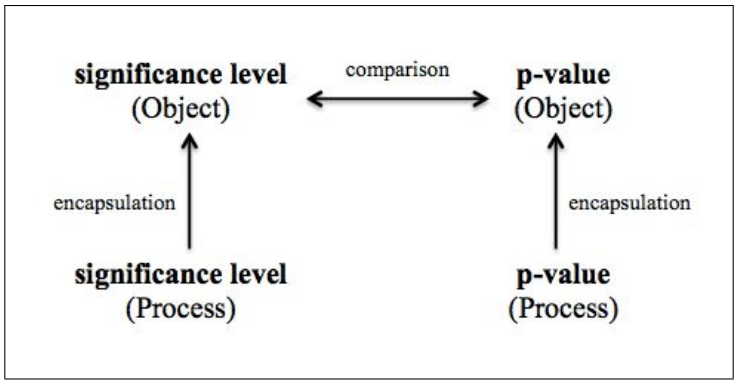

Figure 10: The **decision** Process.

Objects, namely a z-score or t-score of 0, to form a rejection region. Such a Process enables the individual to understand that the significance level can also be used to form a confidence interval centered at the population parameter.

7 Conclusion

In this study, we investigated students' reasoning about hypothesis testing while solving two real-world problems. Our results agree with existing literature acknowledging that hypothesis testing is a challenging topic for students (Dolor & Noll, 2015; Liu & Thompson, 2005; Vallecillos, 2000; Williams, 1997). With statistics education reform on the rise (GAISE College Report ASA Revision Committee, 2016) and with much criticism of hypothesis testing (Motulsky, 2014; Nickerson, 2000; Nuzzo, 2014), it is crucial that curriculum be developed to address students' challenges. The results of our study offer a direction for how this can be achieved.

By using APOS Theory to explore the cognitive aspects of learning hypothesis testing, we found that its difficulty can be attributed to not constructing higher-order mental structures to deal with the associated concepts. In particular, we found that an emphasis on the sequence of steps in the hypothesis testing procedure, as opposed to the concepts, can prevent students from interiorizing Actions into Processes. In order for students to develop a deeper understanding of the concepts, we suggest that instruction emphasize the importance of each step, in its own right, as well as the relations between steps. Regarding the test statistic, for example, students could be asked the following:

- Calculate the test statistic for this problem and identify the sample statistic that it represents.
- Is the test statistic a z-score or a t-score? Why?
- Does the test statistic that you calculated seem extreme? Justify your response.

A comprehensive list of guided questions about the concepts in hypothesis testing can be found in Appendix B. These questions were written based on the results of our study to help students make the constructions called for by the genetic decomposition.

There are several limitations to our study.

1. We restricted our study to a single university. Our results could vary with a different population. On a related note, our genetic decomposition followed the presentation of hypothesis testing by Triola (2014), as it likely had an influence on how our population made (or did not make) constructions in the genetic decomposition. We did not consider other approaches to teaching and conducting hypothesis testing, such as randomization, simulation, and resampling (Cobb, 2007; Lock et al., 2017).
2. Our sample size was twelve, which is small relative to the size of our population. A larger sample could offer a wider range of reasoning about the concepts in hypothesis testing.
3. Our instrument contains only two hypothesis testing questions, while three different kinds of situations could arise in hypothesis testing (proportions, means with a normal distribution, and means with a t-distribution) for the students in our study. Moreover, both questions in our instrument resulted in clear decisions to reject the null hypothesis because the p-values were extremely small. More might be revealed by students' reasoning when the p-value is closer to α, or when the p-value is greater than α.
4. Probability did not play a large role in students' reasoning. The p-value was treated by most students as merely a number between 0 and 1. Further probing into students' understanding of probability could have been insightful.
5. We can only state what we observe. It is possible for Actions, Processes, and Objects involved in hypothesis testing to emerge in additional ways from what we illustrated in this report, and it is possible for students to possess a different conception than what they exhibited.

As a basis for further research, the guided questions in Appendix B, which are applicable to a variety of hypothesis testing questions, can be administered as an instructional method. Then a research instrument and interview protocol should be developed to test if our guided questions can help students make the constructions called for by the genetic decomposition. It is also possible for such an instructional method to result in students making these mental constructions in different ways than those exhibited in the current study, thereby leading to another refinement in the genetic decomposition.

Our preliminary genetic decomposition included several prerequisite constructions, which we assumed had already been made by the students in our study. Thus, we did not investigate how students might make these prerequisite constructions. Future research could potentially benefit from such an investigation. In particular, we suggest that research investigate students' development of their **representation**, **distribution**, **probability**, and **function** Schemas, as well as their understanding of concepts such as probability functions. Knowledge of how students make these constructions could offer additional refinements to the genetic decomposition.

References

Arnon, I., Cottrill, J., Dubinsky, E., Oktaç, A., Fuentes, S. R., Trigueros, M., et al. (2014). *APOS theory: A framework for research and curriculum development in mathematics education*. New York, NY: Springer.

Asiala, M., Brown, A., DeVries, D. J., Dubinsky, E., Mathews, D., & Thomas, K. (1996). A framework for research and curriculum development in undergraduate mathematics

education. In J. Kaput, A. H. Schoenfeld, & E. Dubinsky (Eds.), *CBMS issues in mathematics education, Volume 6, Research in collegiate mathematics education II* (pp. 1–32). Providence, RI: American Mathematical Society.

Baker, B., Cooley, L., & Trigueros, M. (2000). A calculus graphing schema. *Journal for Research in Mathematics Education*, *31*(5), 557–578.

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

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527–547.

Beth, E. W., & Piaget, J. (1966). *Mathematical epistemology and psychology.* Dordrecht, The Netherlands: D. Reidel.

Burns-Childers, A., Chamberlain, D., Jr., Kemp, A., Meadows, L., Stalvey, H., & Vidakovic, D. (2017). Students' understanding of test statistics in hypothesis testing. In A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, & S. Brown (Eds.), *Proceedings of the 20th annual conference on Research in Undergraduate Mathematics Education* (pp. 82–92). San Diego, CA: SIGMAA on RUME.

Chandrakantha, L. (2014). Visualizing and understanding confidence intervals and hypothesis testing using Excel simulation. *The Electronic Journal of Mathematics and Technology*, *8*(3), 212–221.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1).

Cottrill, J., Dubinsky, E., Nichols, D., Schwingendorf, K., Thomas, K., & Vidakovic, D. (1996). Understanding the limit concept: Beginning with a coordinated process scheme. *Journal of Mathematical Behavior*, *15*, 167–192.

Davis, R. B., & Mukamal, K. J. (2006). Statistical primer for cardiovascular research. *Circulation*, *114*, 1078–1082.

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

Dubinsky, E., & McDonald, M. A. (2001). APOS: A constructivist theory of learning in undergraduate mathematics education research. In D. Holton (Ed.), *The teaching and learning of mathematics at university level: An ICMI study* (pp. 275–282). Netherlands: Kluwer Academic Publishers.

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education college report.* Alexandria, VA: American Statistical Association. Available from `http://www.amstat.org/education/gaise`

Hong, E., & O'Neil, H. F. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology*, *82*(2), 150–159.

Krishnan, S., & Idris, N. (2015). An overview of students' learning problems in hypothesis testing. *Jurnal Pendidikan Malaysia*, *40*(2), 193–196.

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).

Link, W. C. (2002). An examination of student mistakes in setting up hypothesis testing problems. *Proceedings of the Louisiana-Mississippi section of the Mathematical Association of America.* Available from

http://sections.maa.org/lams/proceedings/spring2002/conway.link.pdf
Liu, Y., & Thompson, P. (2005). Teachers' understanding of hypothesis testing. In S. Wilson (Ed.), *Proceedings of the twenty-seventh annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education.* Roanoke, VA: Virginia Polytechnic Institute and State University.
Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2017). *Statistics: Unlocking the power of data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
Martínez-Planell, R., & Trigueros Gaisman, M. (2012). Students' understanding of the general notion of a function of two variables. *Educational Studies in Mathematics*, *81*, 365–384.
Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–155.
Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*, *3*(1), 1–8.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.
Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*, 150–152.
Schuyten, G. (1990). Statistical thinking in psychology and education. *ICOTS*, *3*, 486–489.
Smith, T. M. (2008). *An investigation into student understanding of statistical hypothesis testing.* Unpublished doctoral dissertation, University of Maryland, College Park, MD.
Triola, M. (2014). *Elementary statistics using Excel* (5th ed.). Upper Saddle River, NJ: Pearson.
Twigg, C. (2011, May–June). The math emporium: A silver bullet for higher education. *Change Magazine*, 25–34.
Vallecillos, A. (2000). Understanding of the logic of hypothesis testing amongst university students. *JMD*, *21*, 101–123.
Weber, E., Walkington, C., & McGalliard, W. (2015). Expanding notions of "learning trajectories" in mathematics education. *Mathematical Thinking and Learning*, *17*, 253–272.
Williams, A. M. (1997). Students' understanding of hypothesis testing: The case of the significance concept. *MERGA*, *20*, 585–591.
Yung, H. I., & Paas, F. (2015). Effects of computer-based representation on mathematics learning and cognitive load. *Educational Technology & Society*, *18*(4), 70–77.

Appendix A Genetic decomposition for one-sample hypothesis testing

The following is the comprehensive version of our revised genetic decomposition.

Hypotheses. In hypothesis testing, the claim about the population parameter is used to formulate two hypotheses—the null hypothesis, H_0, and the alternative hypothesis, H_1. This can be thought of as a transformation, **hypotheses**, that acts on the claim as an

Object and returns two hypotheses statements. The following are descriptions of the stages of **hypotheses**.

Action: A **hypotheses** Action is a transformation in response to external cues that tell the individual how to formulate the hypotheses for a specific hypothesis test. An external cue could be a memorized procedure or keywords.

Process: The individual interiorizes into a Process the Actions of identifying the claim and formulating H_0 and H_1. In particular, he or she understands that for any hypothesis test, the claim is used to formulate H_0 and H_1, and the individual can explain in his or her own words what H_0 and H_1 mean.

Object: The individual encapsulates a **hypotheses** Process into an Object when he or she can think of it as a totality and perform a transformation on it. For example, the individual can compare how he or she formulated the hypotheses between different situations. That is, the transformation being applied to this **hypotheses** Object is a comparison.

Test statistic. In hypothesis testing, the test statistic, as referenced in this paper, is a standardized value of the sample statistic represented by a score (such as z or t) and is used to determine whether to reject the null hypothesis. For the purposes of our study, the scores represent the relative distance of the sample statistic from the assumed value of the parameter. Calculating a test statistic can be thought of as a transformation, **test statistic**, which acts on various population parameters and sample statistics and returns a value, reasonably between -3 and 3, called the test statistic. The following are descriptions of the stages of **test statistic**.

Action: A **test statistic** Action is an external rule, such as a formula, that tells an individual how to calculate a test statistic's value. Through this Action an individual can input the required values of the population parameters and sample statistics into a formula and simplify the expression to obtain the test statistic's value.

Process: The individual interiorizes into Processes the Actions of calculating various types of test statistics. This Process enables the individual to understand that, in general, a test statistic is the relative distance of the sample statistic from the assumed value of the parameter, and the individual can explain this in his or her own words. These isolated Processes corresponding to each type of test statistic should then be coordinated to construct a more general **test statistic** Process.

Object: The following are ways in which a **test statistic** Process can be encapsulated into an Object:

1. Explain what accounts for an extreme value of a test statistic in comparison to other test statistic values. The transformation being applied to **test statistic** is a comparison of usual and unusual test statistic values.
2. Distinguish between the various test statistic formulas in order to determine which one is appropriate for a given situation. The transformation being applied to **test statistic** is a comparison of different ways to compute a test statistic.

3. Describe the test statistic, in general, as determining the p-value. The transformation being applied to **test statistic** is a **p-value** Process.

P-value. In hypothesis testing, the p-value is the probability of obtaining the sample data or statistic that we observed or something more extreme, under the assumption that the null hypothesis is true. Calculating the p-value can be thought of as a transformation, **p-value**, which is a probability function. In particular, **p-value** acts on the output of **test statistic** and returns a number between 0 and 1, called the p-value. The following are descriptions of the stages of **p-value**.

Action: A **p-value** Action is an external transformation on a particular test statistic's value by following a step-by-step procedure to compute the p-value.

Process: The individual constructs a **probability function** Process (through coordinating **probability** and **function** Schemas), which is then coordinated with a **hypotheses** Process to construct a **p-value** Process. This **p-value** Process acts on **test statistic** as an Object and returns the probability of obtaining the sample data or statistic that we observed or something more extreme, under the assumption that the null hypothesis is true. The individual can explain in his or her own words what the p-value represents.

Object: The following are ways in which a **p-value** Process can be encapsulated into an Object:

1. Distinguish between the various procedures for calculating a p-value in order to determine which one is appropriate for a given situation. The transformation being applied to **p-value** is a comparison of different ways to compute a p-value.
2. View the p-value as an area or probability and compare it to the significance level, α, which is the probability of rejecting the null hypothesis given that it is true. The transformation being applied to **p-value** is a comparison with **significance level**.

Decision. In hypothesis testing, a decision about whether or not to reject the null hypothesis, H_0, is made based on the comparison of the p-value with the significance level, α, which is a probability that serves as an upper bound for the region of probabilities which would result in rejecting the null hypothesis. In particular, when the p-value is less than or equal to α, the null hypothesis is rejected. Forming the decision about H_0 can be thought of as a transformation, **decision**, that compares the p-value and α and returns the decision about whether or not to reject H_0. The following are descriptions of the stages of **decision**.

Action: A **decision** Action is an external transformation on a particular p-value and α by following a step-by-step procedure, such as the following:

Step 1. Compare the numerical values of the p-value and α.

Step 2. From the following, choose the option that applies to the above information:

"p-value $\leq \alpha$" $\Rightarrow$ "reject H_0."
"p-value $> \alpha$" $\Rightarrow$ "fail to reject H_0."

In particular, the individual performs the above Actions by relying on memorization or a formula sheet, without understanding the logic behind the steps.

Process: The individual interiorizes the above Actions into a Process (1) by understanding that the comparison of the p-value and α determines, in general, the decision about the null hypothesis, (2) by understanding that the p-value is the probability of obtaining the sample data or statistic that we observed or something more extreme, under the assumption that the null hypothesis is true, and (3) by viewing α as the bound defining what is considered a low probability. This amounts to constructing a **significance level** Process (through the coordination of **distribution** and **probability** Schemas) and encapsulating **significance level** and **p-value** Processes into Objects by comparing them. In particular, a **decision** Process acts on **significance level** and **p-value** and returns the decision about whether or not to reject H_0.

Object: The individual encapsulates a **decision** Process into an Object by viewing it as a totality and considering scenarios in which the conditions would yield the same decision or a different decision about H_0. That is, the transformation being applied to this **decision** Object is possibly a comparison of various situations in which the decision could have been different.

Conclusion. As stated previously, in hypothesis testing, the claim about the population parameter is used to formulate H_0 and H_1. Once a decision is made as to whether or not to reject H_0, a conclusion can be formed about the claim. Forming a conclusion about the claim can be thought of as a transformation, **conclusion**, that acts on the decision about the null hypothesis, namely, "reject H_0" or "fail to reject H_0," and returns "support claim" or "do not support claim." The following are descriptions of the stages of **conclusion**.

Action: A **conclusion** Action is an external transformation on the claim of a hypothesis test, the decision about H_0, and the hypothesis to which the claim corresponds by following a step-by-step procedure, such as the following:

Step 1. Identify which hypothesis corresponds to the claim.
Step 2. Review the decision about H_0.
Step 3. From the following, choose the option that applies to the above information:

"$H_0 =$ claim" and "reject H_0" $\Rightarrow$ "do not support claim."
"$H_0 =$ claim" and "fail to reject H_0" $\Rightarrow$ "support claim."

"H_1 = claim" and "reject H_0" $\Rightarrow$ "support claim."
"H_1 = claim" and "fail to reject H_0" $\Rightarrow$ "do not support claim."

In particular, the individual performs the above Actions by relying on memorization or a formula sheet, without understanding the logic behind the steps.

Process: The individual interiorizes the above Actions into a Process by describing the steps in general terms. This amounts to a **hypotheses** Process being reconstructed to act on the claim and put it in correspondence with the appropriate hypothesis. Then this reconstructed **hypotheses** Process is coordinated with a **decision** Process to construct a **conclusion** Process. In particular, a **conclusion** Process acts on the outputs of a **decision** Process and reconstructed **hypotheses** Process and returns the conclusion about whether or not to support the claim.

Object: The individual encapsulates a **conclusion** Process into an Object by considering that additional hypothesis tests can be performed on the same (or different) claim, understanding the different scenarios in which the above options would arise. That is, the transformation being performed on this conclusion Object is possibly a comparison of various scenarios in which the conclusion could have been different.

Appendix B Guided questions

The following questions, in general, can be applied to hypothesis testing problems involving a population proportion or a population mean.

1. **Hypotheses.**
   (a) What parameter is this hypothesis test concerning? What is the symbol for this parameter?
   (b) What is the claim of the test?
   (c) Formulate your null and alternative hypotheses.
   (d) What is the relationship between the null and/or alternative hypotheses and the claim?
   (e) Is this a left, right, or two-tailed test?

2. **Test statistic.**
   (a) Sketch the relevant sampling distribution for the sample. You will be adding to this drawing throughout the rest of this worksheet.
   (b) What type of distribution is this? Why?
   (c) What value would be located in the middle of your distribution?
   (d) Calculate the test statistic for this problem. Then identify the sample statistic that it represents.
   (e) Is the test statistic a z-score or a t-score? Why?
   (f) Add the test statistic to your drawing.
   (g) Does the test statistic that you calculated seem extreme? Justify your response.

(h) Based on the value of the test statistic in part (d) and your answer in part (g), describe what this may mean in terms of the p-value (which has not been calculated yet). That is, do you think the p-value will be large or small?

3. **P-value.**

(a) Prior to calculating the p-value, shade the region that represents the p-value on your drawing of the sampling distribution.
(b) Explain how you determined what region to shade.
(c) Calculate the p-value. Then explain why you performed each step in your calculation. That is, what information did you use to determine the steps of your calculation?
(d) Does the p-value that you calculated make sense in relation to the test statistic? Why or why not?
(e) What is the meaning of the p-value in the context of the problem?

4. **Significance Level.**

(a) What is the significance level for this hypothesis test?
(b) What does the significance level represent?
(c) How does the significance level compare to your p-value?
(d) Based on your answer to part (c), estimate where the significance level would be on your drawing. Label this as the rejection region.
(e) Does your test statistic fall within this region?

5. **Decision and Conclusion.**

(a) Does your answer to question 4(e) imply that the test statistic is extreme?
(b) Based on your comparison in question 4(c) of the p-value and significance level, we reject/fail to reject (circle one) the null hypothesis, which was _____. Write a complete sentence describing what this means in the context of the problem.
(c) What does part (b) imply about the alternative hypothesis?
(d) What do parts (b) and (c) imply about the claim in the problem? It may help to refer to your answer to question 1(d).
(e) Write a few sentences to summarize the result of the hypothesis test in the context of the problem.