STUDENTS' CONCEPTUAL UNDERSTANDING OF VARIABILITY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

in the Graduate School of The Ohio State University

By
Leigh Victoria Slauson, B.A., M.A.S.

*****

The Ohio State University
2008

Dissertation Committee

Dr. Patricia Brosnan, Adviser

Dr. Doug Owens

Dr. Dennis Pearl

Approved by

Patricia A. Brosnan

Adviser
College of Education

ABSTRACT

Research has shown that even if a student passes a standard introductory statistics course, they often still lack the ability to reason statistically. This is especially true when it comes to reasoning about variability. Variability is one the core concepts in statistics, yet many students come away from introductory course unable to discuss basic ideas of variability or make the connection between graphical displays of data and measures of variability.

This study investigated students' conceptual understanding of variability by focusing on two numerical measures of variability: standard deviation and standard error. Two sections of introductory statistics were taught at a small Midwestern liberal arts college. One section was taught with standard lecture methods for the topics of standard deviation, sampling distributions and standard error, and confidence intervals and the margin of error. The other section completed a hands-on active learning lab for each these topics. These labs were designed with a conceptual change framework. Students were asked to use their prior knowledge to make predictions, collect and analyze data to test their predictions, and then evaluate their predictions in light of their results. Assessment questions designed to test conceptual knowledge were included at the end of each lab.

Both classes completed the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) as a pretest and a posttest. The assessment questions from the active learning labs were analyzed and coded. And small number of students from each section also participated in twenty-minute interviews. These interviews consisted of statistical reasoning questions.

The analysis of the data showed students' conceptual understanding of ideas related to standard deviation improved in the active class, but not in the lecture class. There was no evidence of improvement on the topic of standard error in either class and some evidence that students had regressed in both sections. The analysis of the qualitative data suggests that understanding the connection between data distributions and measures of variability, and understanding the connection between probability concepts and variability is very important for students to successfully understand standard error. There is also evidence that students come to an introductory statistics course with more conceptual knowledge related to sampling distributions than was previously thought. There is evidence that the feedback portion of hands-on active labs is the most important feature of the conceptual change framework. Further research is needed to investigate how much prior knowledge students possess about sampling distributions and how important probability concepts are to understanding concepts of variability.

Dedicated to Lawrence Charles Weiss

I am my father’s daughter. And proud of it.

ACKNOWLEDGMENTS

First and foremost, I want to thank Dr. Patti Brosnan for her infinite patience and expertise. I could not have asked for a better adviser.

I want to thank my committee members, Dr. Dennis Pearl and Dr. Doug Owens for providing me with invaluable feedback and support throughout this process. I also want to thank Dr. Pearl for giving me so many opportunities to become a flourishing member of the statistics education community.

I want to thank Dr. Jackie Miller for being a mentor and a really good friend.

I want to thank the all the faculty and staff of the Academic Support Center and the Mathematical Sciences Department at Otterbein College, but most especially Ellen Kasulis and Susan Thompson. I am forever grateful for the opportunity to be a faculty member before I even earned my doctorate. And thank you so much for giving me the time, resources, encouragement and sugar to successfully finish my dissertation.

I want to thank my friends, Greg and Kelly Syferd, who kept me sane and helped me procrastinate with many, many, many games of Union Pacific.

I want to thank the Jeff and Becky Slauson, who have truly become my family. I am so, so lucky to have in-laws like you.

I want to thank my sisters, Anne and Debra, and little Xanis for reminding me there is much more to life than work. To Annie and Debra, best wishes on your own adventures in school.

To my mother, who has been waiting a long time for this document, thank you for helping me become the independent, intelligent woman I am today.

And finally to my husband Justin: there are no words to describe what you mean to me or how much I love you. I could not have done this without you. You are my lobster.

VITA

June 14, 1976	Born – Redbank, NJ
1998	B.A. Math & Classics, Vanderbilt University.
2000	M.A.S. Statistics, The Ohio State University.
1998	Teaching Assistant, Mathematics Department, Vanderbilt University.
1998 – 2000	Teaching Assistant, Department of Statistics, The Ohio State University.
2000 – 2001	Teaching Assistant, Department of Statistics, University of Illinois.
2001- 2006	Casual Teaching Assistant, Mathematics Department, The Ohio State University.
2001 – 2006	Adjunct Faculty, Mathematics Department, Columbus State Community College.
2005 – present	Visiting Instructor, Academic Support Center & Mathematical Sciences Department, Otterbein College.

PUBLICATIONS

1) Laughbaum, Ed. (2003). State of Ohio College Mathematics Placement Testing Program. (COMPASS Analyst – Leigh Slauson)

2) Elstak, I, Rice, G., Strayer, J. & Weiss, L. Dissertations in Mathematics Education Reported in 2000. In Reed, M.K., & Owens, D.T. (Eds.). (2003). *Research in Mathematics Education 2000.* Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

FIELDS OF STUDY

Major Field: Education
Area of Emphasis: Mathematics Education

Minor Field: Statistics

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

During the first few years that I taught introductory statistics, I always tried to follow my department guidelines for teaching statistics very closely. I had taken a job as an adjunct at a local community college while I started working on my doctoral degree. This particular mathematics department had many different instructors for introductory statistics. In an attempt to keep the introductory classes as similar as possible, they had laid out in great detail exactly what topics I should cover and when I should cover them. The course was taught in a bi-weekly class meeting, with a weekly computer lab period. They had a departmental final, weekly labs written by a course coordinator, and departmental guidelines on what should be taught with the graphing calculator. Initially, I thought this was great; I wouldn't have to do much planning. However, I discovered very quickly the course was tightly packed with material. I had very little time for any creative teaching innovations, such as demonstrations or fun activities. I was forced to lecture through most class periods just to get through the material on time. It was my first experience as a lead statistics instructor. I had been a teaching assistant for years, teaching in statistics recitation sections. Generally, however, this only entailed going over student questions or helping students work through pre-designed activities. I had never been in charge before of presenting the material for the whole course. It turned out to be quite the learning experience for me. One of the requirements at the end of the quarter was for me to cover a departmental final exam review sheet that included material from the entire ten-week course. In this review packet, there was always at least one question that asked students to calculate the standard deviation of a data set and then offer an explanation of what the standard deviation meant in the context of the data set. Generally, my students had little difficulty calculating the standard deviation using their required graphing calculators, except for perhaps forgetting under which menu in the calculator they were supposed to use. However more importantly, the explanation portion of that question proved incredibly difficult for my students. Some were able to give the general definition of standard deviation as a "measure of spread," but rarely were they able to elaborate further or explain any connection between the number they had just calculated and the dataset from which it came. And more often than I would like to admit, I would get responses such as the one I received from a student who said she had "just forgotten" what it was. As the instructor and a statistics education researcher, this was disappointing.

Rationale of the Study

The idea of variability and its associated measures such as standard deviation in a data set is one of the core concepts presented in an introductory course in statistics (Cobb, 1991; Moore, 1992; Snee, 1993). "Variation is at the heart of all statistical investigation. If there were no variation in data sets, there would be no need for statistics" (Watson & Kelly, 2002, p. 1). It is obvious, from both my own personal teaching experience and the literature in statistics education, that students struggle with understanding the concept of variability as it relates to data (Reading & Shaughnessy, 2004). Instructors struggle also with how to teach it effectively. In February of 2005, the American Statistical Association (ASA) released a report that presented Guidelines for Assessment and Instruction in Statistics Education (GAISE) at the college level. Those statistics educators who contributed to the report recognized that "many introductory courses contain too much material and students end up with a collection of ideas that are understood only at a surface level, are not well integrated and are quickly forgotten. If students don't understand the important concepts, there's little value in knowing a set of procedures" (ASA, 2005, p. 10). One of the arguments made for including technology (such as graphing calculators) in the statistics classroom is that it reduces the time needed to teach the procedures, such as the formula for calculating standard deviation, leaving instructors more class time to focus on the concepts (Collins & Mittag, 2005). Theoretically, this extra time spent on the conceptual nature of variation should result in better student understanding of this basic statistical idea. However, even in classes that utilize graphing calculators or other technology extensively, there is still overwhelming evidence that students do not leave an introductory statistics class with good understanding of the concept of variability (Hawkins, 1996). The solution to this problem then is much more complex than simply using technology to do the calculations. Understanding variability in a data set is a basic building block to all of the topics presented in an introductory class in statistics. If students don't understand this basic idea, this is a significant problem in statistics education.

There has certainly been a good deal of research in the last few years dealing with student learning in the elementary statistics classroom. As more and more researchers recognize that the problems of statistics education are separate from mathematics education (Moore & Cobb, 2000; Moore, 1988), the volume of literature in statistics education is slowly growing. Specifically, many of these studies focus on the development of students' *statistical reasoning* skills as the main goal of an introductory course in statistics. While there is disagreement about how statistical reasoning is defined[1], most researchers agree that it includes "the use of statistical tools and concepts to summarize, make predictions about and draw conclusions from data" (Lovett 2001, p. 350). Certainly, the act of making correct interpretations from the data is key to any definition of statistical reasoning. Because of the shift in focus to statistical reasoning in the introductory class (and away from a memorization of procedures), there have been an increased number of calls for change in the way statistics is taught (Ben-Zvi & Garfield, 2004). Several factors have been identified as to why this change is needed:

- Changes in the field of statistics, including new techniques of data exploration
- Changes and increases in the use of technology in the practice of statistics, and the growing availability of this technology in schools and at home
- Increased awareness of students' inability to think or reason statistically, despite good performance in statistics courses
- Concerns about the preparation of teachers of statistics at the K-12 and college level, many of whom have never studied applied statistics or engaged in data analysis activities (Ben-Zvi & Garfield, 2004, p. 5).

There has been research into statistics education topics since the seventies. Some of the first "statistics education" researchers attempted to identify and classify the types of statistical reasoning misconceptions people possess. Most notably in their research,

[1] The debate over the definition of statistical reasoning is discussed later in this chapter.

Kahneman, Slovic and Tversky (1982) documented many statistical reasoning fallacies[2] that people possess and classified these into heuristics (the processes by which people discover or learn). Their goal, however, was only to understand and classify how people reasoned statistically, not to develop methods to correct incorrect statistical reasoning. These studies were also not focused on students in statistics courses, but rather the general population.

In the eighties more researchers focused on whether students, who were enrolled in a statistics course, showed a higher level of statistical reasoning than the subjects of the previous research of the seventies. Fong, Krantz, and Nisbett (1986) showed that statistics students made only marginal gains in statistical reasoning from the beginning to the end of the course. Other researchers attempted to test different models of statistical reasoning to explain why students reasoned incorrectly (Konold, 1989; Pollatsek, Konold, Well, & Lima, 1984). These studies still did not provide suggestions for improving instruction, but mainly focused on trying to define how statistics students reasoned incorrectly.

In contrast, the nineties were marked by a rapid shift in the research towards providing instructional strategies that would help students learn to reason statistically (Garfield & delMas, 1991). This shift mirrored an effort in the mathematics education community to develop pedagogy that was aimed at helping students learn to reason better mathematically (Lovett, 2001; National Council of Teachers of Mathematics (NCTM), 2003; Shirley, 2000). Part of this shift was in response to the rapid shift in the practice of

[2] For example, the gambler's fallacy: the incorrect belief that the likelihood of a random event can be affected by or predicted from other independent events.

statistics. Software programs for not only doing statistics, but also for learning statistics were being developed and refined. A computer lab section where students could use these software applications was becoming more common in the introductory course (Lovett, 2001). The advent of graphing calculators in the late nineties that allowed students to have basic statistical capabilities right at their fingertips, drastically changed the possibilities for data analysis right in the classroom (Moore, 1997; NCTM, 2003).

The most recent research in statistics education is trying to pull all these different foci together: using the theories of how students learn to guide what kind of instructional strategies should be tested and used in the introductory classroom (Lovett, 2001). Certainly one pedagogical approach that has been recommended repeatedly by recent researchers is the use of active learning techniques instead of traditional lecturing (Gnanadesikan, Scheaffer, Watkins & Witmer, 1997; Moore, 1997; Shaughnessy, 1977). The use of active learning is based partly on the theories on constructivism and cooperative learning. Researchers argue that actually having students participate in activities where they can manipulate and analyze real data that they themselves have collected, leads to a much deeper and sophisticated understanding of the statistical concepts being presented (Ben-Zvi, 2004; Mackisack, 1994).

> Using active learning methods in class is a valuable way to promote collaborative learning, allowing students to learn from each other. Active learning allows students to discover, construct, and understand important statistical ideas and to model statistical thinking. Activities have an added benefit in that they often engage students in learning and make the learning process fun. (ASA, 2005, p.11)

However, the interpretation of what constitutes an active learning activity has been varied. Teachers and departments often identify labs completed in a weekly computer lab section as examples of active learning. This was certainly my experience at the previously mentioned community college where I was an adjunct. Students, either in groups or on their own, would work through these assigned labs on a computer using a statistical software package. Unfortunately, while these labs did provide students exposure to current statistical software and technology, many of these labs were frequently more of an exercise in following directions and did not provide students the opportunity to do real statistical investigation. Active learning techniques are "used in a variety of class settings to allow students to 'discover' concepts of statistics by working through a set of 'laboratory' exercises" (Scheaffer, Watkins, Gnanadesikan, & Witmer, 1996, p. vii). The most important feature of these active learning activities is that students are allowed to discover concepts, and unfortunately many computer lab activities don't allow that process of student discovery to occur. Several texts (e.g., *Workshop Statistics, Activity Based Statistics*, etc.) have been published in the last few years with a wealth of active learning activities that allow a process of student discovery that could be used in any type of statistics classroom. However, these textbooks are still considered experimental by many, and very few departments have adopted them as course textbooks. Regardless, they are invaluable references for the instructor who is looking to incorporate active learning activities into a "regular" introductory course.

One of the benefits of active learning techniques is that it often involves data collection. One of the main recommendations of the GAISE report is for teachers to "use real data" (ASA, 2005). Many times this comes in the form of either student run projects or experiments in the classroom. It has been noted that students are much more invested in understanding why a dataset behaves a certain way if it is their own (Mackisack, 1994; Hogg, 1991). Often, data that are presented in textbooks are contrived and neat. While contrived datasets may be necessary for teaching certain concepts, datasets in the real world are complex and students should be exposed to that complexity (ASA, 2005). Many different statistics educators have recommended student projects as an excellent way of providing this exposure to data analysis in the real world (Binnie, 2002; Field, 1985; Mackisack & Petocz, 2002; Sylvester & Mee, 1992). However, this can potentially take a great deal of time and effort for the instructor, who may have to mentor many students' projects individually.

However, simply incorporating active learning techniques into the introductory statistics classroom may not be enough to affect real change in student's conceptual understanding. In what became a multistage research project conducted at two different colleges, a study showed several students were still failing to develop correct understanding of sampling distributions after the topic was taught using active learning techniques (Chance, delMas & Garfield, 2004). This was a surprise to the researchers, given the numerous sources and studies that have proposed active learning helps to develop student's conceptual understanding. So the researchers decided to look outside of the research in mathematics and statistics education to science education. Statistics is considered in many circles to be more of a scientific discipline rather than a mathematical one, so looking into the research in science education to better understand how students might learn statistics certainly has merit. One learning theory that has a particularly strong research base in science education is conceptual change theory. The theory posits that students learn when an existing conception that they already possess is challenged.

“This model proposes that students who have misconceptions or misunderstandings need to experience an anomaly, or contradictory evidence, before they will change their current conceptions” (Chance, delMas & Garfield, 1999, p. 5).

While most research on the application of conceptual change theory occurred in science education, educators are now starting to see the potential applications in other fields, such as statistics education. “Teaching for conceptual change primarily involves 1) uncovering students' preconceptions about a particular topic or phenomenon and 2) using various techniques to help students change their conceptual framework” (Davis, 2001, para. 2.1). In a follow up study to the multistage research project described previously, Chance, delMas and Garfield (2004) designed computer simulation activities that followed a “predict/test/evaluate model” that allowed students to make predictions about graphical test questions, use a sampling software to create the distributions to test their predictions, and finally allowed students to evaluate their predictions with the computer results. Chance et al. (2004) found that these activities, designed from a conceptual change framework, significantly improved students’ reasoning about sampling distributions. This multistage research project, which was part of the impetus for this research, will be discussed in much greater detail in chapter two.

Problem Statement

It is important to understand how student’s come to reason about variability and specifically what kind of instructional techniques improve the conceptual understanding of variability. Research in the statistics education community has shown that traditional lecture techniques are not providing students with the statistical reasoning skills they should possess after a first course in statistics (Ben-Zvi & Garfield, 2004). Active learning techniques provide an alternative type of learning for students, but this covers a wide range of activities. What needs to be researched is how particular active learning techniques specifically improve students' conceptual understanding of variability.

Focus Questions

1) Do active learning labs, designed with a conceptual change framework, have a significant impact on students' conceptual understanding of variability in an introductory statistics course?
2) How do students articulate their understanding of variability?
3) What differences in understanding of variability are there between students whose introduction to variability measures occurs through a lecture/demonstration format and students whose introduction occurs through active learning labs?

Definitions

Variation vs. Variability

Many research studies use the terms variation and variability interchangeably, and assume these terms have self-evident definitions. Some recent research has attempted to be more specific about the difference between variation and variability. Reading and Shaughnessy (2004) distinguish the difference between the two terms as follows:

> The term variability will be taken to mean the [varying] characteristic of the entity that is observable, and the term variation to mean the describing or measuring of that characteristic. Consequently, the following discourse, relating to "reasoning about variation," will deal with the cognitive

> processes involved in describing the observed phenomenon in situations that exhibit variability, or the propensity for change (p. 202).

However, reasoning about variation and reasoning about variability are not always differentiated this way in the literature. It is generally accepted that variation or variability encompasses several things: the quantitative measures such as standard deviation and variance, how it is used as a tool and why it is used in certain contexts (Makar & Confrey, 2005). This particular study uses *variability* to encompass these ideas.

Statistical Literacy, Reasoning and Thinking

A distinction also needs to be made among several terms that refer to the goals of statistics education: statistical literacy (a component of quantitative literacy), statistical reasoning, and statistical thinking. In much of the statistics research, these terms often overlap one another in how they are defined and used. However, in recent years, there has been a concerted effort to come to an agreement on their definitions. They have been discussed and debated extensively at the International Research Forums on Statistical Reasoning, Thinking, and Literacy and the following definitions of each term is the product of several papers written on the subject by various statistics educators after the forums. Although these terms are being defined here, it should be noted that different researchers still use these terms interchangeably.

- *Statistical literacy* includes basic skills such as organizing data in graphs and tables, understanding the concepts, vocabulary, and symbols, and recognizing that probability is a measure of uncertainty. Statistical literacy goes beyond knowing statistical terminology to the "ability to understand and critically evaluate statistical results that permeate daily life…. and a questioning attitude one can assume when applying concepts to contradict claims made without proper statistical foundation" (Gal, 2002, p. 2).

- *Statistical reasoning* requires the ability to make interpretations about data based on graphs, tables or statistical summaries. It also includes the ability to make connections and explain the relationships between the different statistical processes "Reasoning means understanding and being able to explain statistical processes and being able to fully interpret statistical results" (Ben –Zvi & Garfield, 2004, p. 7).

- *Statistical thinking* requires that one gets the "big ideas" of statistics. Specifically, statistical thinking focuses on the how and why of statistical inference. Snee (1990) defines statistical thinking as "*thought processes*, which recognize that variation is all around us and present in everything we do, all work is a series of interconnected processes and identifying, characterizing, quantifying, controlling, and reducing provide opportunities for improvement" (p. 118). For example, it includes knowing when to use appropriate data analysis techniques and knowing how and why inferences can be made from samples to populations to understanding why designed experiments are necessary to show causation. Statistical thinkers are able to evaluate and critique statistical studies and can explain how models can be used to simulate random phenomena (Ben–Zvi & Garfield, 2004; Chance, 2002).

In much of the literature, the ideas of statistical literacy, statistical reasoning, and statistical thinking are hierarchal in terms of learning goals for an introductory course (statistical literacy being most basic to statistical thinking being most advanced). However, there is still a great deal of dissension in other research over the definitions of these terms. delMas (2002) points out that some researchers feel that these terms should cover separate ideas but recognize that there will be some degree of overlap. Other authors have posited that statistical thinking and reasoning are actually completely subsets of the overall idea of statistical literacy. Figure 1 shows a visual representation of these competing models of statistical literacy, reasoning and thinking.

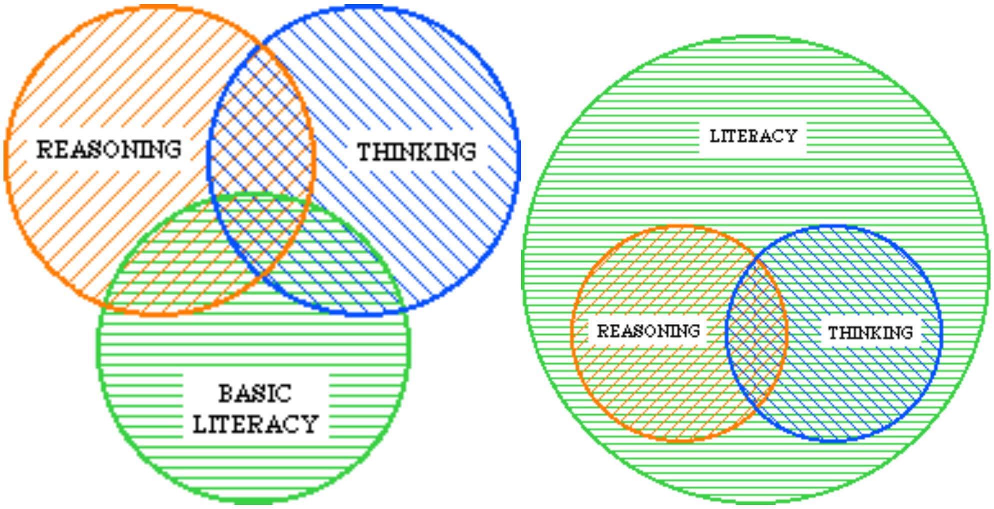


Figure 1. The overlap model versus the subset model of statistical literacy, reasoning, and thinking (delMas, 2002).

Although there is a disagreement in the literature about whether statistical reasoning is a unique goal onto itself, I feel that the goals of my research specifically target the idea of statistical reasoning. The main goals of my study revolve around students being able to see the connections between statistical measures of variability and how a dataset behaves. This goes beyond knowledge of how to calculate and define measures of variability as part of statistical literacy. It might be argued that the goals of my study, particularly related to students knowing when to apply standard deviation versus when to use standard error, might fall under the definition of statistical thinking. However, I still believe that this particular knowledge can also be defined as statistical reasoning, because it involves students "being able to explain statistical processes" (Ben–Zvi & Garfield, 2004, p.7).

Theoretical Framework

The main theoretical framework of this study is the theory of conceptual change. Conceptual change theory is based on Jean Piaget's ideas and writings on how students learn. Piaget used the idea of schemas to describe the mental skills that a student or a person possesses. Piaget defined two terms to describe how students process new information or ideas into their already existing schemas: *assimilation* and *accommodation*. Assimilation occurs when a student encounters a new phenomenon or idea and uses existing schemas to make sense of it. The new information is assimilated into the old schemas. However, if the students' existing schemas can't fully explain the new phenomenon, it creates a perturbation for the student. Then, a review of the phenomenon is initiated that may lead to an accommodation (Von Glasersfeld, 1997). During accommodation, a student must reorganize or replace his existing schemas to allow for understanding the new phenomenon.

> Assimilation and accommodation work like pendulum swings at advancing our understanding of the world and our competency in it. According to Piaget, they are directed at a balance between the structure of the mind and the environment, at a certain congruency between the two that would indicate that you have a good (or at least good-enough) model of the universe. This ideal state he calls equilibrium. (Boeree, 2006, para. 2)

His work eventually became the underlying foundations of constructivism, a theory that postulates, "the learner constructs his/her own understanding from the totality of the experiences which he/she sees as relevant to the content, belief, skill etc., being considered" (Gunstone, 1994, p. 132). Constructivism has been extensively researched and applied in mathematics education (Cobb, 1994).

The theory of conceptual change builds upon the ideas of assimilation and accommodation a bit further: that the process of learning or how a student comes to know a new concept is also guided by a student's existing schemas on how learning should take place. "Without such concepts it is impossible for the learner to ask a question about the (new) phenomenon, to know what would count as an answer to the question, or to distinguish relevant from irrelevant features of the phenomenon" (Posner, Strike, Hewson & Gertzog, 1982, p. 212). The central concepts of the conceptual change model are *status* and *conceptual ecology*.

> The *status* that an idea has for a person holding it is an indication of the degree to which he or she knows and accepts it: status is determined by its intelligibility, plausibility and fruitfulness to that person. The idea of a *conceptual ecology* deals with all the knowledge that a person holds, recognizes that it consists of different kinds, focuses attention on the interactions within this knowledge base, and identifies the role that these interactions play in defining niches that support some ideas (raise their status) and discourage others (reduce their status). Learning something, then, means that the learner has raised its status within the context of his or her conceptual ecology. (Hewson, Beeth & Thorley, 1998, p. 201)

Teaching for conceptual change can be difficult. Students may have to change their "fundamental assumptions about the world, about knowledge, and that such changes can be strenuous and potentially threatening" (Posner et al., 1982, p. 213). Many recent researchers have taken these concepts and conceived that conceptual change "involves the learner recognizing his/her ideas and beliefs, evaluating these ideas and beliefs (preferably in terms of what is to be learned and how this is to be learned), and then personally deciding whether or not to reconstruct these existing ideas and beliefs" (Gunstone 1994, p. 132). In short, this creates a "*recognize, evaluate, decide whether to reconstruct"* (Gunstone 1994, p. 132) formula of how conceptual change can occur.

In much of the literature, research on conceptual change and research on "misconceptions" are research on the same idea. In particular, some researchers define misconceptions as a "miscatergorized concept(s)", and that conceptual change entails re-categorizing these misconceptions correctly. There has been some dissension among conceptual change researchers as to what exactly sparks the process of conceptual change or recategorization for a learner. Vosniadou (2002) argues that learners begin with "naïve theory of how something works based on previous experience" (p. 65) and the process of conceptual change begins with instruction that is "inconsistent with their existing mental representations" (p. 65) and that creates a perturbation in the mind of the student. As students begin to assimilate the new knowledge, "they form synthetic meanings that lack coherence and stability" (Vosniadou, 2002, p. 65). Resolving these internal inconsistencies is a gradual process that can entail a progression of a series of mental models about the new knowledge. In other words, conceptual change is not an instantaneous reorganization and replacement of concepts (Mayer, 2002, p. 102), but a process that happens over time.

Chi and Roscoe (2002) argue that "to accomplish conceptual change, learners must become aware that they have miscategorized a concept" (Mayer, 2002, p. 104), but again they view this process as an incremental process that happens over time. Others stress the importance of sociocultural aspects in the process of conceptual change. Ivarsson, Schoultz and Saljo (2002) argue "human cognition is socialized through the participation in activities where tools are used for particular purposes" (p. 79). In particular, these authors ran an experiment that demonstrated children developed more sophisticated reasoning about the nature of earth when they had access to a world map during interviews. They argued that students' mental models depend heavily on the tools the students have access to in a particular learning situation.

Others stress that conceptual change and metacognition are irrevocably intertwined (Gunstone, 1994). Metacognition is the consciousness of our thoughts and the knowledge we have about our cognitive process. Gunstone argues learners must recognize and evaluate "with an understanding of learning goals, of relevant uses of the knowledge/skills/strategies/structures to be learned, of the purposes of particular cognitive strategies/structures to be learned, of the purposes of particular cognitive strategies appropriate to achieving these goals, of the processes of learning itself" (p. 133). For example, De Jong and Gunstone (1998) found in their study to affect conceptual change in physics students, that they were met with resistance because the students felt that learners needed "high intelligence" and "good memory" in order to be successful in physics and a student either had those things or they didn't. The researchers found that these students prevented themselves from learning because they believed they did not possess the necessary skills to learn.

Because of these differing viewpoints, defining all the aspects of the theory of conceptual change is difficult. And some researchers point out the limitations of the theory and the lack of "theoretical accountability concerning the nature of the mental entities involved in the process of conceptual change" (Limon & Mason, 2002, p. xvi). Since there is disagreement over how conceptual change operates, there is corresponding disagreement over how exactly conceptual change theory works in practical applications. However, common to all of the definitions and variations of the conceptual change model, is the importance given to students' knowledge prior to instruction (Hewson, Beeth, & Thorley, 1998). In one of the first articles published on conceptual change theory in science education, Posner, Strike, Hewson and Gertzog (1982) recommend the following teaching strategies for teaching toward conceptual change.

> 1) Develop lectures, demonstrations, problems, and labs that can be used to create cognitive conflict in students.
>
> 2) Organize instruction so that teachers can spend a substantial portion of their time in diagnosing errors in students' thinking and identifying defensive moves used by students to resist accommodation.
>
> 3) Develop the kinds of strategies that teachers could include in their repertoire to deal with student errors and moves that interfere with accommodation.
>
> 4) Help students make sense of science content by representing content in multiple modes (e.g. verbal, mathematical, concrete-practical, pictorial), and by helping students translate from one mode of representation to another.
>
> 5) Develop evaluation techniques to help the teacher track the process of conceptual change in students. (p. 225)

Researchers agree that teaching for conceptual change requires a great deal from teachers. There is much they need to know not only about the content of a particular course, but also the range of ideas students may have about a particular topic and the various pedagogical techniques that can be employed (Hewson, Beeth & Thorley, 1998).

Generally, researchers agree on the fundamental ideas of conceptual change and accommodation and assimilation. The disagreements discussed above relate to *how* this process actually occurs and *what* conditions must be present to stimulate the process.

Conceptual change theory has been a major theoretical framework for science education (especially physics) in the last thirty years or so. It has only been transferred to other fields in a limited capacity. However there are a few recent studies in mathematics and statistics that have used conceptual change theory successfully as their theoretical framework (Chance, delmas, & Garfield, 1999). In particular, Chance, delMas and Garfield (2004) used conceptual change theory to redesign an "activity to engage students in recognizing their misconceptions and to help them overcome the faulty intuitions that persisted in guiding their responses on assessment items" (p. 299). This was part of the multistage research project that was mentioned earlier in this chapter. The researchers in this case redesigned computer lab activities to follow a "predict/test/evaluate model" that followed closely the "recognize/evaluate/decide whether to reconstruct" model of conceptual change proposed by science education researchers. This study will be discussed in much more detail in chapter two of this document.

My personal rationale for this study

My intention for this study is to test how well conceptual change theory works in the introductory statistics classroom. Shortly before I began working on this research, I was hired as a full time instructor at a small, mid-western, liberal arts college. Teaching introductory statistics at this college turned out to be a much bigger challenge than I had encountered previously in my career. First, this course carried credit for the mathematics major and was taught from a very mathematical perspective. However, this course was taken and required by a wide variety of other majors, including business, nursing, and sports management. While all the students had the required college algebra prerequisite, there was still a great disparity in mathematical skills and reasoning abilities among the students in the class. Second, the college did not have a computer lab large enough to hold thirty students, which was the usual number of students in a particular introductory statistics section. This made the logistics of teaching introductory statistics with a computer component nearly impossible. And lastly, there was resistance within the department to see the need for change to the introductory statistics curriculum. It was my hope that by conducting my dissertation research at this college, I would be able to demonstrate the need to revise their introductory statistics curriculum. I was also hoping to demonstrate that the reforms proposed by the statistics education community would improve the course, as it already existed.

What to expect in the next chapters

Chapter 2 provides a literature review of studies done in statistics education and beyond that are relevant to the issues of developing students' conceptual understanding of variability in an introductory statistics course. Chapter 3 discusses the qualitative and quantitative methodologies that were employed to answer the research questions posed previously. Chapter 4 gives a detailed analysis of all the relevant data from assessment instruments, lab activities and student interviews. And chapter 5 discusses the issues related to this study, the conclusions I was able to draw from my data and the implications for further research on students' conceptual understanding of variability.

CHAPTER 2

LITERATURE REVIEW

Statistics Education

The number of statistics courses taught at college and universities around the United States and beyond has grown tremendously over the last twenty years. This is partly due to the increased demand by the marketplace that college graduates have at least a basic understanding of statistical concepts. Advancements in technology have created the ability to collect vast volumes of data in government, business, and other industries. Consequently, the ability to analyze data has become a much more valued commodity in the workplace. "Not surprisingly, the Curriculum and Evaluation Standards for School Mathematics (National Council of Teachers of Mathematics (NCTM), 1989) has echoed the increased attention to statistics and probability in society by recommending a prominent role for applications of data and chance in school mathematics" (Shaughnessy & Zawojewski, 1999, p. 713).

While the *Principles and Standards for School Mathematics*, published by National Council for Teachers of Mathematics, have increased the focus on data analysis at all grade levels, there is still a struggle to get statistics and probability instruction into K-12 classrooms. Many mathematics teachers, who end up becoming the statistics instructors in many schools, receive little education in statistics during their professional preparation (Fendel & Doyle, 1999). In most cases, only one statistics course is required for pre-service teachers and in some cases, the statistics course is optional. Certainly the college where this study took place requires only pre-service secondary mathematics teachers to take a statistics course. For pre-service elementary teachers, statistics is an optional course. Even so, there is evidence of an increase in the coverage of statistical topics in K-12 classrooms. The 1996 National Assessment of Education Progress (NAEP) showed that teachers are spending a "moderate" amount of time on statistical topics. This is opposed to surveys conducted in 1990 and 1992, which showed teachers spending "little" time on these topics (Shaughnessy & Zawojeski, 1999). The most current version of the report shows that 4th graders in many states are doing significantly better on the data analysis and probability sections of the assessment (United States Department of Education, 2007).

Even with this progress, for most students, their first formal experience with statistics occurs in college. Therefore, a significant amount of the literature in statistics education has focused on the college student and is the focus of this particular study. Many of the conclusions and recommendations cited here, however, can apply to both the college statistics student and the K-12 student. And I will discuss a few studies in this literature review, which are relevant to conceptual understanding of variability that took place in K-12 classrooms.

Statistical Reasoning: Early Studies

Advances in educational and statistical technology have made data analysis much more accessible to all grade levels. Before graphing calculators and computers, the introductory statistics curriculum had to focus on performing calculations and learning to manipulate complex statistical formulas. Now with the advancement and availability of technology, it is no longer necessary for statistics students to complete those complex calculations by hand. This has shifted the focus in statistics classrooms toward learning the underlying concepts of statistics and learning to reason statistically. Therefore, research in statistics education has shifted its focus to the unique educational issues that come with trying to teach and learn abstract statistical concepts (Ben-Zvi & Garfield, 2004). The goal is no longer mastery of statistical skills, but rather a mastery of statistical reasoning.

> If teachers were asked what they would really like students to know six months or one year after completing an introductory statistics course, …Many would indicate that they would like students to understand some basic statistical concepts and ideas, to become statistical thinkers, and to be able to evaluate quantitative information (Garfield, 1995, p. 26).

In many of these studies, there is discussion and debate over defining statistical reasoning versus statistical thinking. As I mentioned in Chapter 1, many authors use these two terms interchangeably or feel that the concepts they encompass overlap. Lovett (2001) defines statistical reasoning as "the use of statistical tools and concepts… to summarize, make predictions about, and draw conclusions from data" (p. 350). Wild and Pfannkuch (1999) define statistical thinking as "the complex thought processes involved in solving real world problems using statistics with a view to improving such problem solving” (p. 224). There does seem to be a consensus in the research that statistical thinking is higher level of cognitive activity than statistical reasoning. “Statistical reasoning may be defined as the *way* [italics added] people reason with statistical ideas and make sense of statistical information…. Statistical thinking involves an understanding of *why* and *how* [italics added] statistical investigations are conducted” (Ben-Zvi & Garfield, 2004, p. 7). Still, the definitions of statistical reasoning and statistical thinking are being developed and debated in the literature (delMas, 2004). Regardless of the exact definition, several studies have been conducted recently that focus on the issues of developing statistical reasoning.

Research into understanding how students reason in probability and statistics began with a few studies done in the late seventies and the early eighties. Two psychologists, Kahneman and Tversky (1982) did a series of studies trying to understand how people reason about probability and classify the different types of misconceptions people develop about probability. These studies were conducted with people who had no particular statistics training. Two of these heuristics that they developed are described below:

> 1) *Representativeness* – people tend to make decisions about the probability of an event happening depending on how close the event models what happens in the distribution from which it is drawn. For example, deciding that a particular sequence of coin flips, such as H H H T H is not likely because it does not contain 50% heads. (Shaughnessy, 1977)
>
> 2) *Availability* – people tend to assess the probability of a particular situation depending on how easily they can relate to it in their own mind. For example, a person might feel the probability of a middle-age person having a heart attack is high, if they have middle-age friends who have had heart attacks. (Kahneman & Tversky, 1982)

A number of studies were completed after Kahneman and Tversky published their work attempting to confirm their classifications and understand how to correct these misconceptions. Pollastek, Konold, Well, and Lima (1984) conducted two experiments in which they took the one particular type of assessment question Kahneman and Tversky had used in their work to identify use of the representativeness heuristic and posed them to current statistics students. All the students answered the questions as part of a bonus questionnaire that was passed out in class. Then some of those students were interviewed later. In the second experiment, the interviewers posed alternative solutions to the assessment questions to the interview subjects. These alternative suggestions were made in an effort to judge how strongly students held onto their original answer and to further probe what kind of reasoning the students were using. Pollastek et al. (1984) put forth their own model (active balancing) to classify what kind of reasoning students were using. However, they concluded that there was no evidence to support their model and students were indeed using reasoning consistent with Kahneman and Tversky's representativeness heuristic. One major implication of the study was that "since students' actual heuristic, representativeness, is so different in form from the appropriate mechanistic belief, it may not be easy…to effect any lasting change in students' beliefs" (Pollastek et al., 1984, p. 401). The researchers were very concerned that the methods students employed to reason about these probability assessment items were radically different from any correct reasoning methods.

Other studies focused on different aspects of Kahneman and Tversky's heuristics or used them as a framework for their own research. Fong, Krantz, and Nisbett (1986) performed a series of studies to determine whether the level of statistical training of subjects (Kahneman and Tversky's subjects had no statistical training) affected the statistical "quality" of responses to assessment questions. Fong et al. (1986) defined statistical "quality" as how much formal statistical reasoning was used to answer the assessment question. Although they were able to show that statistical training did have some effect on the quality for some assessment questions, they were surprised to find that statistical training had no effect on the statistical quality of the responses to two of their four assessment questions. Their results suggested that coursework and training in statistics would not necessarily guarantee a higher level of statistical reasoning skills (Lovett, 2001). Konold (1989) tested his own model of statistical reasoning, what he termed the "outcome approach", against the representativeness heuristic and found that there were "additional misconceptions that ought to be taken into account" (p. 93). He observed that students would predict the outcome to a single trial with a yes or no for any type of probability question and relied on casual, informal explanations of outcomes and variability. For example, some participants were given an irregularly shaped bone with each side labeled with different letters and asked which side was most likely to land upright. A composite "outcome approach" response from his interview transcripts is quoted below:

> S: Wow. If I were a math major, this would be easy. B is nice and flat, so if D fell down, B would be up. I'd go with B
> I: And how likely do you think B is to land upright?
> S: I wouldn't say it's much more likely. It depends on the roll.
> I: So what do you conclude, having rolled it once?
> S: Wrong again. [B] didn't come up.
> I = Interviewer S= Student (Konold, 1989, p. 67)

The first part of this composite response demonstrates what Konold meant by a casual, informal explanation: students not using any sort of a mathematical model to reason about the question. The second part of this composite response shows what he meant by a yes or no response: the student thought the probability of rolling a particular side of the bone (B) was determined to be either right or wrong after each roll. Konold used his results to demonstrate that Kahneman and Tversky's classifications were a start to understanding a student's statistical reasoning and misconceptions, but there was much more to the story.

Lecoutre (1992) added the idea of "equiprobability bias" to the previously mentioned heuristics. The equiprobability bias occurs when students feel that all results from random events are equally likely. For example, if a jar contains two red chips and one white chip, students will mistakenly feel that the probability of drawing two red chips is the same as the probability of drawing one red chip and one white chip. Lecoutre discovered this bias in a study of 87 students. She attempted to "trick" students into using the correct conceptual model by first asking students a probability question with the chance aspect masked and then following it with a more standard probability problem. Of the students that got the first question correct, only sixty percent were able to correctly transfer the reasoning to a second standard probability question. Another study confirmed that there are enduring misconceptions among students of statistics regardless of age or experience. Metz (1998) conducted a study of kindergarteners, third graders and undergraduates, attempting to trace the conceptual development of the statistical idea of randomness. Based on her videotaped analysis of students working three specific sampling tasks, she concluded that while students did develop a more sophisticated idea of randomness as they got older, there still remains “enduring challenges in the interpretation of random phenomena, manifested from kindergarten to adulthood” (Metz, p. 349).

Most of this work in the seventies and eighties (and some in the nineties) focused on reasoning in probability situations. The main conclusion of all of this research was that students have many enduring misconceptions in probability and statistics even after they complete an introductory course in statistics. Later research (Chance, delMas, & Garfield, 1999; Garfield, 1995) would expand upon these identified misconceptions and attempt to present pedagogy that would address and correct them.

Exploratory Data Analysis (EDA)

Much of the more recent research that has been done in the area of statistical reasoning has been done to ascertain how students develop these skills. One hypothesis set forth by researchers is that students understand much more conceptually if they are exposed to statistics through Exploratory Data Analysis (EDA). EDA is a statistics curriculum that asks students to analyze and explore data to answer open-ended questions with a heavy emphasis on interpreting graphical displays, often with the aid of technology (Ben-Zvi, 2004; Ben-Zvi & Friedlander, 1997; Burrill & Romberg, 1998). This parallels mathematics education paradigms such as problem-based learning and inquiry-based learning. Several researchers have done teaching experiments where EDA was the primary pedagogical method. Ben-Zvi (2004) observed the development of statistical reasoning in two students in a seventh grade EDA classroom in Tel-Aviv. This particular curriculum (*Statistics Curriculum)* was "based on the creation of small scenarios in which students can experience some of the processes involved in the experts' practice of data based enquiry" (Ben-Zvi, 2004, p. 125). For example, one task asked students to take data from the men's 100-meter Olympic races and look for trends and interesting phenomena. Then the students were asked to respond to an argument between two journalists as to whether there is a limit to how fast a human can run.

These students' reasoning development was tracked over the course of two and a half months through video recordings, classroom observations, interviews, student notebooks, and student projects. He found that in this controlled learning environment, students developed "their reasoning about data by meeting and working with, from the very beginning, ideas and dispositions....[that include] making hypotheses, formulating questions, handling samples and collecting data, summarizing data, recognizing trends, *identifying variability* [italics added] and handling data representations" (Ben-Zvi, 2004, p. 141).

Other researchers have focused on creating an EDA classroom with a heavier reliance on technology. Biehler (1997) conducted two studies in which students used a software tool for an entire course in data analysis, which culminated in an end of course project. Through in-depth interviews, Biehler was able to identify the problem areas and difficulties students and teachers encounter in elementary data analysis much more extensively; issues such as lacking the language to accurately describe graphical structure. All of these studies emphasize the importance as well, even with the increased use of technology, of the teacher in creating EDA environments that motivate and challenge students to really understand the complexities of a dataset (Ben-Zvi, 2004; Ben-Zvi & Friedlander, 1997; Groth, 2006).

In some introductory statistics classes that are geared for science majors, EDA can come in the form of scientific experiments. Because carrying out an experiment or project is an exercise in applying general principles to a specific task (Mackisack, 1994), these experiments can provide students with a way to examine aspects of data analysis that wouldn't be otherwise possible with traditional textbook problems. One of the main principles of EDA is to have students work with real data. "Learning about statistics is more meaningful – and fun – when students are analyzing real data that they themselves collected. Whenever possible, it is important to involve students in testing their own hypotheses" (Wolfe, 1993, p. 4). Issues such as critiquing experimental designs, giving detailed interpretations of results, exploring anomalous observations, questioning to elicit detailed information about somebody else's problem, and deciding what analysis is appropriate for a particular data set can be explored within the context of these experiments (Mackisack, 1994).

Reasoning about Distribution

The concept of distribution is one of the first topics introduced in an introductory statistics course. Researchers argue that it is fundamental for students to understand how data behaves graphically (Gal & Short, 2006) and is inextricably linked to conceptual understanding of variability (delMas & Liu, 2005; Leavy, 2006; Makar & Confrey, 2003; Reading & Reid, 2006).

> Variation is at the heart of statistical thinking but the reasoning about variation is enabled through diagrams or displays…such as graphs or frequency distributions of data. (Pfannkuch & Reading, 2006, p. 4)

Several researchers have researched student's conceptual understanding of distribution. Bakker and Gravemeijer (2004) observed the development of statistical reasoning about the shape of a distribution in students in four different seventh grade classrooms in the Netherlands over the course of a year. They used specifically designed series of interactive web applets (*Minitools*) that allowed students to first develop informal reasoning about concepts such as majority, outliers, chance, reliability and mean on given datasets. Later, students could use *Minitools* to invent their own dataset distribution in response to descriptions of the data presented by their teacher. For example, in an activity on battery life, students had to create a distribution that would show "brand A is bad but reliable; brand B is good but unreliable; brand C has about the same spread as brand A, but is the worst of all the brands" (Bakker & Gravemeijer, 2004, p. 154). During these lessons, the researchers conducted mini-interviews with students. By asking students to move back and forth between interpreting distributions and constructing distributions, the researchers were able to demonstrate that in this kind of classroom environment, students were able to represent many elements of distributions in their graphs. They also stressed that the assessments, which asked students to create graphs using given statistical information, was very effective in developing students' conceptual understanding of distribution.

Other researchers have also focused on the importance of understanding how teachers develop and talk about their own conceptions of distributions (Pfannkuch, 2006). One study in particular, Leavy (2006), explored the development of pre-service teachers' understanding of distribution in a mathematics methods course. The study consisted of three phrases: pretest/baseline data collection, instructional intervention, and a posttest. The pretest involved a bean sprout experiment in which the pre-service teachers were asked to present their conclusions. From these results, researchers were able to see that the teachers mostly relied on numerical methods instead of graphical displays to make data comparisons. Based on the pretest results, the pre-service teachers participated in a semester-long statistical investigation of the bean experiment with a specific focus on distribution. During the course of the semester, the pre-service teacher's work related to the ongoing investigations was used to frame discussion and instruction. They were also asked to keep a journal and participate in focus groups to discuss statistical strategies. The pre-service teachers, then, participated in a post-instruction investigation involving popcorn, but similar in structure to the pretest bean sprout experiment. The posttest results showed several improvements in the pre-service teachers' reasoning: they were more attuned to issues of sample size, their choice of descriptive statistics was much more careful, and the teachers were much more conscious of limitations of the choices they made for data presentation.

The results from this study had several important implications. The first was recognizing that before this course, pre-service teachers focused much more on summary of the data rather than exploration. Other researchers have noted this lack of statistical knowledge with pre-service teachers and encouraged degree programs to require content experience with data analysis and statistics (Lajoie & Romberg, 1998). Given that these teachers are a crucial component in bringing statistics education reforms into classrooms, it is important to realize that "their *lack of exposure* to statistical ideas and statistical inquiry lead to the blanket implementation to measures they are familiar with – the mean invariably" (Lajoie & Romberg, 1998). However, once participants' attention was drawn to variation, in concert with an emphasis on how variation is modeled in graphical representations, variation quickly became the central component of participants' understanding of distribution" (Leavy, 2006, p. 106). The study also highlights the connection between statistical inquiry, concepts of distribution and recognizing the importance of variation.

While some researchers and educators view distribution as the "lens" through which students learn and come to understand variation (Wild, 2006), other researchers view reasoning about variation as a prerequisite to understanding distribution. Reading and Reid (2006) conducted a study of statistical reasoning with 57 students in an introductory course. The students were asked to complete four minute-papers around four different themes: exploratory data analysis, probability, sampling distribution, and inference. The minute papers were short assessment questions that were given to students in the last five minutes of class and submitted to the instructor immediately. The papers were coded by the researchers initially with a consideration of student's reasoning about variation using a four level hierarchy (no consideration of variation, weak consideration of variation, developing consideration of variation, strong consideration of variation).

Later, the same minute papers were coded similarly but with a consideration towards reasoning about distribution. The researchers found that there was a very strong link between the code that a student received on variation and the code a student received on distribution. Students who were rated weak on variation rarely demonstrated strong understanding of distribution. Students who were rated developing consideration of variation did a much better job describing the different elements of a distribution and were more able to make the connection between variation and distribution. These results make the argument that conceptual understanding of variability and conceptual understanding of distribution are inextricably linked. However, it was noted by these researchers, “few responses were able to successfully apply their understanding of centre, spread, and density to the complex notion of the sampling distribution of the mean” (Reading & Reid, 2006, p. 57).

Reasoning about Sampling Distributions

Perhaps the most well known study of students’ conceptual understanding of sampling distributions, Chance, delMas, and Garfield (2004) conducted a seven-year, multi-stage project at two different American universities to track how college students were reasoning about sampling distributions. Eventually they had five different studies come out of this one larger project, each previous study driving the conceptual framework for the next study. Initially they were interested in evaluating the usefulness of simulation software for learning sampling distributions. Students participated in an instructional activity that utilized a computer simulation program developed by one of the researchers (*Sampling Distribution*). Students could change certain settings such as population shape and sample size for generating sampling distributions. They were then asked to summarize the results of what they had observed. The researchers used the results from pretest and posttest graphical based assessment items to measure change in reasoning about sampling distributions. The results from the assessment items showed some positive changes, but for the most part did not demonstrate that students were developing the correct reasoning about sampling distributions. From these results, the researchers hypothesized that the simulation activities needed to have more than just a visual component.

Thus, Chance, delMas, and Garfield (2004) attempted to redesign the simulation activities, using a conceptual change framework, i.e., "redesigning the activity to engage students in recognizing their misconceptions and to help them overcome the faulty intuitions that persisted in guiding their responses on assessment items" (p. 299). Students were asked to give their response to a graphical test item and then use the *Sampling Distribution* program to produce a sampling distribution in order to evaluate their response to the graphical test item. The "predict/test/evaluate" model for the redesigned activities seemed to produce better understanding of sampling distributions in students, but there were still misconceptions that persisted.

For the third study, Chance, delMas and Garfield (2004) looked at whether students' understanding of the knowledge prerequisite to understanding sampling distribution could explain students' persistent misconceptions about sampling distributions. Through analysis of the experiences and observations of the classroom researchers involved in the first two studies, plus a detailed analysis of student responses from the first two studies, the researchers were able to compile a list of the prerequisite knowledge necessary to understanding sampling distribution. These included the idea of variability, the idea of distribution, the normal distribution and the idea of sampling. The researchers then took those ideas and created a pretest designed to identify students' weaknesses and help guide instruction before students began the unit on sampling distributions.

The fourth study gathered more information about students' conceptions about the prerequisite knowledge and how they developed reasoning about sampling distributions though interviews with students in a graduate level introductory statistics class. Students were asked to answer open-ended questions about sampling variability while interacting with the *Sampling Distribution* software. From these interviews, a developmental model (Table 1) of how students come to reason about sampling distribution was proposed.

Level of Reasoning	Description
1: Idiosyncratic reasoning	The student knows some words and symbols related to sampling distributions, uses them without fully understanding them, often incorrectly, and may scramble these words with unrelated information.
2: Verbal reasoning	The student has a verbal understanding of sampling distributions and the Central Limit Theorem, but cannot apply this knowledge to actual behavior. For example, the student can select a correct definition or state the implications of the Central Limit Theorem as it describes sampling distributions, but does not understand how the key concepts such as variability, average, and shape are integrated.
3: Transitional reasoning	The student is able to correctly identify one or two dimensions of the sampling process without fully integrating these dimensions. For example, the student only understands the relationship between the sampling distribution shape and the population shape, the fact that large samples lead to more normal looking sampling distributions, or that larger sample size leads to a narrower sampling distribution (decreased variability among sample means).
4: Procedural reasoning	The student is able to correctly identify the dimensions of the sampling process but does not fully integrate them nor understand the process that generates sampling distributions. For example, the student can correctly predict which sampling distribution corresponds to the given parameters, but cannot explain the process and does not have confidence in his or her predictions.
5: Integrated process reasoning	The student has a complete understanding of the process of sampling and sampling distributions and is able to coordinate the rules (Central Limit Theorem) and behavior of the sampling process. The student can explain the process in her or his own words and makes correct predictions with confidence.

Table 1: Developmental model of statistical reasoning about sampling distributions (Garfield, 2002, p. 4).

This developmental model became the basis for the fifth study. Chance, delMas and Garfield (2004) developed a diagnostic assessment that was designed to identify students who were at potentially different levels of statistical reasoning. Three different populations of students took this diagnostic test after their unit on sampling distributions. Subsets of these students participated in interviews, consisting of four questions related to sampling distributions. The initial purpose of the interviews was to validate the level of reasoning each interviewee had been placed at by the diagnostic test. However, the researchers found it incredibly difficult to ascertain from the interviews the appropriate level of reasoning for particular student.

Each of these five studies contributed greatly to understanding why sampling distributions is such a difficult subject for students to learn. The researchers were able to identify some of the problem areas and document how "forcing students to confront the limitations of their knowledge…[helped] students…correct their misconceptions and …construct more lasting connections with their existing knowledge framework" (Chance, delMas, & Garfield, 2004, p. 312). The researchers, however, recognized that many students did not have the necessary prerequisite knowledge of variability and distribution and that research needs to focus on how "these early foundational concepts needs to be integrated throughout the course so students will be able to apply them and understand their use in the context of sampling distribution" (Chance, delMas, & Garfield, 2004, p. 314).

Models of Statistical Reasoning

Several other studies have proposed models of statistical reasoning, although these studies had different statistical topics as their focus. A longitudinal study (Watson 2004) of 22 Tasmanian students of varying grade levels classified students' reasoning about sampling into one of six categories. These categories were subsets of tiers (Table 2) of statistical literacy that had been developed by the researchers in an earlier study (Watson, 1997).

Tier 1	Understanding Terminology
Tier 2	Understanding Terminology in Context
Tier 3	Critical Questioning of Claims made without Justification

Table 2. Categories of developing concepts of sampling (Watson, 2004).

Watson (2004) used these categories and tiers to track students reasoning over the period of four years through student interviews. They were able to document students' responses to the assessment questions on sampling four years after their initial classification generally improved. However, it was impossible to ascertain whether "life experiences or the mathematics curriculum" (Watson, p. 290) were responsible for the improvement. Some of the older students referred to examples in their classes, which seemed to indicate that the mathematics curriculum was having an influence.

Several studies employed a version of the Structure of Observed Learning Outcomes (SOLO) taxonomy developed originally by Biggs and Collis (1982) to develop a hierarchy of statistical reasoning (Pfannkuch, 2005; Reading & Reid, 2006; Watson, 2004; Watson, Kelly, Callingham & Shaughnessy, 2003; Watson & Moritz, 2000). The SOLO taxonomy postulated the existence of five modes of functioning that span from birth to about 20 years old and five cognitive levels (prestructural, unistructural, multistructural, relational, and extended abstract) that "recycle during each mode and represent shifts in complexity of students' reasoning" (Jones, Langrall, Mooney, & Thorton, 2004, p. 99). Table 3 shows theses levels were modified and applied to describe levels of statistical reasoning.

Level of SOLO taxonomy	Description of application to statistical reasoning
Prestructural (P)	Does not refer to key elements of the concept.
Unistructural (U)	Focuses on one key element of the concept.
Multistructural (M)	Focuses on more than one key element of distribution.
Relational (R)	Develops relational links between various key elements of the concept.

Table 3. Interpretation of SOLO taxonomy for statistical reasoning (Reading & Reid, 2006).

Researchers have used this framework as the framework for their research into student statistical reasoning on a particular topic or concept. Reading and Reid (2006) used this to frame their study on reasoning about distribution, while Chance, delMas and Garfield (2004) used this model as the basis for determining "dimensions" of statistical reasoning behavior.

Reasoning about Variability

Despite recognizing the importance of variability in an introductory course, very little research "has been published about how people, particularly novices and statistics students, actually reason about variability, or how this type of reasoning develops in the course of statistics instruction" (Ben-Zvi & Garfield, 2004). Many of the studies already mentioned previously have recognized the importance of students' conceptual understanding of variability as crucial to students' understanding of other areas in statistics (Chance, delMas, & Garfield, 2004; Reading & Reid, 2006; Leavy, 2006). There have been some studies that have focused primarily on variability, but many within the context of other topics (distribution, sampling, etc). Reading and Shaughnessy (2004) examined students' conceptions of variability within the topic of sampling. Twelve students from both primary school and secondary school were interviewed and asked to answer a series of questions about the probabilities associated with two sampling situations. Students were presented with two mixtures of lollipops: the first had 50 red, 30 blue and 20 yellow lollipops; the second contained 70 red, 10 blue, and 20 yellow lollipops. Students were asked to predict how many red lollipops could be expected if a student drew 10 lollipops out of each mixture. Then they were asked if six people each drew 10 lollipops out of each mixture, and each person returned the lollipops to the bowl after each draw, how many red lollipops could be expected. Students were asked to explain their reasoning and given an opportunity to change their answer after they physically drew 10 lollipops out of the mixtures. The researchers were able to categorize most responses into one of two hierarchies:

> 1) Description hierarchy – based around students' descriptions of the variation occurring, developed from aspects of students' responses such as *it is more spread out, there's not the same number each time.*
>
> 2) Causation hierarchy - based around students' attempts to explain the source of the variation, developed from aspects of student responses such as *because there's heaps of red in there* (Reading & Shaughnessy, 2004).

The researchers used these hierarchies to analyze and code the students' responses. In these descriptions, students emphasized two main concepts of variation. One concept was on how spread out the numbers were: students gave responses that indicated they were considering variation when they were dealing with extreme values. The other concept was on what was happening with the numbers within the range of possibilities. The researchers also discovered that "students demonstrated more of their developing notions of variation in those tasks that they found more difficult" (Reading & Shaughnessy, 2004, p. 221). The challenge of dealing with the 70% red setup versus the 50% setup and increasing the number of repetitions provided the researchers more insight into students' conceptual understanding as students struggled with the complexity of the situation. Researchers concluded that instructors should not shy away from more challenging data sets, as they allow "more opportunity for discovering" (Reading & Shaughnessy, 2004, p. 223).

Liu and delMas (2005) explored students' conceptual understanding of the standard deviation, one of the main statistical measures of variability, as students interacted with a designed computer environment and an interviewer. The computer environment was a Java program that allowed students to compare and manipulate distributions and standard deviations. Students from four sections of an introductory statistics course at a large Midwest research university participated in a three-phrase interview process: introduction, exploration and testing. The introduction allowed the students time to become familiar with the computer program. During the exploration phase, students were presented with five different sets of bars and asked to arrange them so that they produced the largest possible standard deviation. The computer program then allowed students to check their answers and receive feedback.

During the testing phase, students were presented with 10 pairs of histograms. The mean and the standard deviation of the left histogram were presented but only the mean of the right histogram were presented. Students were asked to decide whether the standard deviation of the right histogram was larger, smaller or the same as the histogram on the left. After each pair, students were asked to explain their reasoning and then were allowed to check their answers. Figure 2 shows an example pair of histograms from a test item from this study.

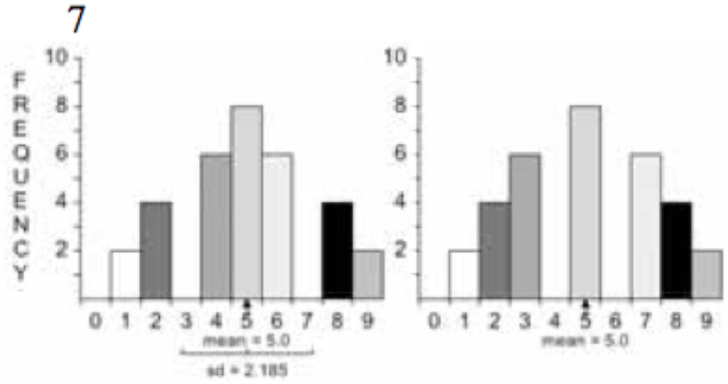


Figure 2: Sample test item from Liu and delMas' (2005) study

The purpose of this study was exploratory in nature and was meant to address the lack of research describing how students think about variability. The researchers were able to identify eleven types of justifications students used to explain their responses to the test items. These justification types included bell-shaped, big mean, and equally spread out. In addition, the researchers were able to demonstrate that the interaction with the computer and the interviewer did help students form the correct conceptualization of how standard deviation relates to a histogram.[3]

Several authors have researched how teachers understand variability. Makar and Confrey (2005) looked at the types of language pre-service teachers use when engaged in data analysis tasks. Interviews were conducted with 17 pre-service teachers at both the beginning and end of a pre-service course on assessment, which included a unit on exploratory data analysis. During the interviews, participants were given data from an education project and asked to use the provided data and dotplots to comment on the effectiveness of a remediation program. From the transcripts of these interviews, the researchers were able to document the types of statistical and nonstandard language that the participants used in comparing distributions and how much that improved over the course. The use of statistical language (proportion, mean, maximum, sample size, outliers, range, shape, standard deviation) did increase from the beginning of the course

[3] . It should be noted that several features of Liu & delMas' research study were used in my own research study (although I did not realize it at the time). The teaching unit that was presented to these classes before the completion of the interviews was identical to the standard deviation distribution worksheet that I included in my standard deviation lab (see Appendix A). Liu and delMas' study also utilized a conceptual change framework and focused on conceptual understanding of standard deviation, as did I in the first part of my own research study, although I did not use a computer interface.

to the end. The nonstandard language (evenly spread out, the bulk of them, bunched up, clustered, etc) that participants used also increased from beginning to end of the course.

The researchers found that even though the majority of participants were using at least some form of non-standard language to compare the distributions, they still demonstrated a “keen awareness of variation in the data” (Makar & Confrey, 2005, p.47). In one example, prospective teachers split the distribution into parts and examined a chunk of the distribution. This was a similar finding in a study completed by Hammerman and Rubin (2004), in which prospective teachers broke up the distribution into bins and compared slices of the data. “Although simple, this is a more variation-oriented perspective than responses that took into account only the proportion or number of students who improved on the test” (Makar & Confrey, 2005, p. 48). The researchers argue that learning to recognize nonstandard language when students are describing statistical concepts is incredibly important. Not recognizing non-standard language can mean that teachers miss understanding their students’ reasoning and send a tacit message to students that statistics can only be understood if it is communicated in the “proper” language (Makar & Confrey, 2005). Use of nonstandard language by both students and teachers allows individuals to convey their conceptions of variability using words that hold meaning for them, which is incredibly important from a constructivist perspective. It also allows statistics to be more accessible to a wider population of students and may eventually lead students and teachers down the path toward using more standard statistical language.

Other studies have explored how teachers reason about variability. Hammerman and Rubin (2004) used the software program *TinkerPlots* with current teachers during professional development workshops and then observed them in their own classrooms. Their goal was describe to how teacher's conceptions of variability developed and changed as they went from learner to teacher using *TinkerPlots.* The researchers found that the visual features of the software program helped deepen the teachers' understanding of variability and the analyses they completed were much more sophisticated. This was a cyclical process as the teachers were faced with more complex variability questions from their students as a result of their students' interactions with *TinkerPlots.*

Active Learning in Statistics

The use of active learning techniques in the introductory statistics classroom is not a new idea. Shaughnessy (1977) conducted a study in 1976 that demonstrated the benefits of active learning for a probability classroom. He compared two groups of students: one group took a traditional lecture-based course in elementary probability, while the other group took an experimental, activity-based course. The purpose of his study was to "test the relative effectiveness of a lecture-based approach in overcoming certain misconceptions that college students have about probability… that arise from reliance upon the heuristics of representativeness and availability[4]" (Shaughnessy, 1977, p. 289). The experimental course consisted of nine activities in probability,

[4] The representativeness and availability heuristics (Kahnman & Tversky (1982)) are described in more detail at the beginning of chapter two.

combinatorics, game theory, expected value, and elementary statistics. Each activity required the participants to complete experiments, gather data, organize and analyze the data, and finally summarize their findings and hopefully discover some mathematical principle. Students were very much encouraged to work together and told to keep a log of all of their work both in and out of the classroom. The instructor's role in the experimental section was to circulate throughout the room and clarify students' questions. Both groups of students, lecture and experimental, were given a pretest and a posttest developed by the researcher and specifically designed to test "for knowledge of some probability concepts and for reliance upon representativeness and availability in estimating the likelihood of events" (Shaughnessy, 1977, p.308). Some example items from that assessment are shown in Figure 3 below:

A Representativeness Item

R1: The probability of having a baby boy is about ½. Which of the following sequences is more likely to occur for having six children?
(a) BGGBGB (b) BBBBGB (c) about the same choice for each

Give a reason for your answer

An Availability Item

A2: A man must select committees from a group of 10 people. Would there be:

(a) More distinct possible committees of 8
(b) More distinct possible committees of 2
(c) About the same number of committees of 8 as committees of 2

Give a reason for your answer (Shaughnessy, 1977, p. 309- 313)

Figure 3. A sample of assessment items used to detect the use of an availability or representativeness heuristic.

Shaughnessy found that the experimental students were able to discover many of the probability concepts he presented in the lecture sections and that there was evidence that students in the experimental course were more successful in overcoming the representative and availability heuristics. He also found that the activities had a positive effect on student attitudes.

In the mid 1990s, as part of a NSF funded project, Scheaffer, Gnanadeskian, Watkins, & Witmer (1996) published a textbook called *Activity Based Statistics*.

> So that students can acquire a conceptual understanding of basic statistical concepts, the orientation of the introductory statistics course must change from a lecture-and-listen format to one that engages students in active learning. This is the premise underlying an effort of the authors to produce and use a collection of hands-on activities that illustrate the basic concepts of statistics covered in most introductory college courses. Such activities promote the teaching of statistics more as an experimental science and less as a traditional course in mathematics. An activity-based approach enhances learning by improving the students' attention, motivation, and understanding. (Gnanadesikan et al, 1997, ¶1)

Another project called Workshop Statistics focused on

> the development and implementation of curricular materials, which guide students to learn fundamental statistical ideas through self-discovery. Using the workshop approach, the lecture-format was completely abandoned. Classes are held in microcomputer-equipped classrooms in which students spend class-time working collaboratively on activities carefully designed to enable them to discover statistical concepts, explore statistical principles, and apply statistical techniques. (Rossman, 1997, ¶ 1)

The publication of these textbooks highlights the shift in statistics education toward emphasizing a more activity-based classroom. Few studies have documented, other than Shaughnessy (1977), the effectiveness of a complete activity-based classroom in an introductory statistics course. There have been several studies on various types of active learning in the literature: computer simulations (Burrill, 2002; Lipson, 2002; Mills, 2002;), exploratory data analysis (Ben-Zvi, 2004), experiments (Martinez-Dawson, 2003), and semester long projects (Binnie, 2002). The research of this dissertation is focused more on hands-on, in-class small experiments and activities. There are a few studies that have specifically addressed this type of active learning.

Jun and Pereira-Mendoza (2002) conducted a study that used focused teaching interventions with small group activities. Two grade 8 classes participated in six 40-minute lessons twice a week outside regular class time. These lessons were designed to overcome misconceptions that had been identified in a previous study. One of the classes had access to a computer for simulations. The other components of the class (activities, workbook, problems for class discussion and the teacher) were the same for both of the classes. The difference between the classes was in how data was given to students after the activities. In the computer class, data from extended versions of the activities completed in class were simulated in front of the students, whereas the other class was presented data and told it was computer generated. Based on pre- and posttests, their results from the teaching interventions did help students overcome some misconceptions, although it is unclear whether the hands-on activities were the reason or simply the instruction specifically targeted at overcoming misconceptions.

Researchers do recognize however that activities don’t always work. Mickelson and Haseman (1998) found that activity-based learning did not promote a deeper conceptual understanding of sampling distributions for all students. They concluded there was “very complex interaction concerning the epistemology the student brings to the class, the connection between the students' epistemology and the epistemology inherent in constructivist instructional methods, the content of the activity, prior educational experiences, and the social/academic atmosphere of the class/institution” (CAUSE website literature index). Other researchers recognize that the logistics and time commitment necessary to do activities may simply be too much in an introductory class, especially in larger lecture sections. Activities must be carefully planned and the point of the activity needs to be explicit to students (Gnanadesikan, Scheaffer, Watkins, & Witmer, 1997).

Learning Theory Underpinning Active Learning

“The literature that describes misconceptions about statistics and probability is much more extensive than literature on what can practically be done to ameliorate them” (Garfield & Ahlgren, 1988, p. 53). That statement was published in 1988. Fortunately in the last twenty years, however, that has changed. There is now a body of research that attempts to address how students learn basic statistical concepts and how to implement instructional techniques to accomplish this. This has led to a reform in statistics education that has been motivated and supported by current cognitive theory in education. Most researchers agree that students learn statistics by doing statistics (Smith, 1998). Lovett and Greenhouse (2000) delineated five principles of learning that can be applied to statistics education.

1) Students learn best what they practice and perform on their own.
2) Knowledge tends to be specific to the context in which it is learned.

3) Learning is more efficient when students receive real-time feedback on errors.
4) Learning involves integrating new knowledge with existing knowledge.
5) Learning becomes less efficient as the mental load student must carry increases. (p. 196)

These principles of learning are derived from the learning theories of constructivism and situated learning. Constructivism postulates that students actively "construct" their knowledge. They integrate new information with pre-existing knowledge. Garfield (1995) concluded "regardless of how clearly a teacher or book tells them something, students will understand the material only after they have constructed their own meaning for what they are learning. Moreover, ignoring, dismissing, or merely 'disapproving' the student's current ideas will leave them intact – and they will outlast the thin veneer of course content" (p. 26). Situated learning postulates that learning is inextricably linked to the activity in which the knowledge was developed (Brown, Collins, & Duguid, 1989). "Situated thinking exploits as much as possible the specific quality, relations, and elements of the respective situation" (Ben-Zvi, 2000, p. 142). Other authors have noted that "activity in which knowledge is developed is not separable from or ancillary to learning … Rather, it is integral part of what is learned" (Bradstreet, 1996, p. 73).

Conceptual change theory has been used more extensively in science education, than in mathematics or statistics education, but it builds on the premises of constructivism and situated learning. Primarily, it focuses on a "shift or restructuring of existing knowledge and beliefs is what distinguishes conceptual change from other types of learning. Learning for conceptual change is not merely accumulating new facts or learning a new skill. In conceptual change, an existing conception is fundamentally changed or even replaced" (Davis, 2001, para. 2). Various studies in science education have shown that when students come to conceptual understanding of a topic through conceptual change instruction, their conceptual skills are much stronger because they understand *how* they come to their conceptual understanding (Vosniadou, 1994).

In addition to hands-on activities, use of cooperative learning groups in statistics has been promoted in the literature in the last few years. "Cooperative learning is a technique that requires students to work together in small fixed groups to accomplish some task" (Magel, 1996, p. 52). These tasks can range from completing worksheets and assignments in class (Keeler & Steinhorst, 1995; Magel, 1996) to completing semester-long experiments or projects (Mackisack, 1994; Martinez-Dawson, 2003). In many cases activity-based learning and cooperative learning go hand in hand. Shaughnessy (1977) made use of both in his study.

Borresen (1990) compared three types of statistics classes, two of which were activity based and utilized cooperative learning, over a period of three years. The first type of classroom was traditional, the second type was a classroom where student groups assigned by instructor worked on their assignments together in class, and the third type utilized small groups formed by students voluntarily. Borresen was mainly interested in building on a previous study of cooperative learning in teaching statistics to see if how the groups were formed made a difference in the test scores of his students. Borresen found, as the previous study did, that the total semester point distribution of his cooperative learning classes were significantly higher than his traditional class, but there did not seem to be a significant difference between the test scores of the two cooperative learning classrooms. However, he also found that “the subject of the questions (from students) changed from computing procedure to statistical rationale” (Borresen, p. 27) in the cooperative groups as the semester progressed, indicating that the cooperative learning groups moved students towards discussions of statistical concepts.

In another study, Smith (1998) researched the use of out of class statistical projects throughout a semester-long course. Students, in teams of three, were assigned biweekly projects. Each member of the three-person team did two written reports and two oral presentations of the project results. “The use of teams fosters cooperative learning, develops team-working skills, and often builds considerable camaraderie” (Smith, 1998, ¶ 11). And the projects did seem to help students learn. Although Smith did not conduct a formal study, his exam scores have improved dramatically from an average of 80% before the implementation of the projects to an average 92% in classes that completed projects.

Even if cooperative learning groups are not used for the entirety of a course, they still have benefits to students. Dietz (1993) reported that she employed cooperative learning techniques on the unit dedicated to methods of selecting a sample. She spent four classes (50 minutes in length) discussing and carrying out the activity. Although not a formal study, Dietz observed that students were able to discover the different sampling methods through their group discussions and this seemed to build students’ confidence immensely.

The use of experiments is another way to create the possibility of deeper understanding of statistical knowledge. Because carrying out an experiment or project is an exercise in applying general principles to a specific task (Mackisack, 1994), these experiments can provide students with a way to examine aspects of data analysis that wouldn't be otherwise possible with traditional textbook problems. One of the main goals of many statistics educators is to have students work with real data. "Learning about statistics is more meaningful – and fun – when students are analyzing real data that they themselves collected. "Whenever possible, it is important to involve students in testing their own hypotheses" (Wolfe, 1993, p. 4). Issues such as critiquing experimental designs, giving detailed interpretations of results, exploring anomalous observations, questioning to elicit detailed information about somebody else's problem, and deciding what analysis is appropriate for a particular data set can be explored within the context of these experiments (Mackisack, 1994).

Student Attitudes towards Statistics

One major concern of statistics instructors is the pervasive negative attitudes towards statistics that most students take with them into an elementary statistics course. Frustrations, boredom, dislike, and, at times, fears of the course are some of the common attitudes observed in students (Lee, Zeleke, Meletiou, & Watchel, 2002). "Consequently, students enter the statistics course with high anxiety and low expectation of learning" (Lan, Bradley, & Parr, 1993, p. 38). Overcoming these negative attitudes can be a major obstacle for statistics instructors. Garfield (1995) stated the following attitude goals should be set for students in statistics:

1) It is important to learn some fundamentals of statistics in order to better understand and evaluate information in the world.

> 2) Anyone can learn important ideas of statistics by working hard at it, using good study habits, and working together with others.
> 3) Learning statistics means learning to communicate using the statistical language, solving statistical problems, drawing conclusions, and supporting conclusions by explaining the reasoning behind them.
> 4) There are often different ways to solve a statistical problem.
> 5) People may come to different conclusions based on the same data if they have different assumptions and use different methods of analysis. (p. 26)

Student attitudes are normally measured by student surveys. Many studies in statistics education have included some sort of class evaluation at the end of the project to assess student attitudes (Borresen, 1990; Magel, 1998; Rossman, 1997). Magel (1998) kept track of the percentage of students who dropped his class and the percentage of students who attended class as a measure of students' attitudes about the class in general. Most of the studies that utilize activity-based learning, cooperative-based learning, and technology have claimed improved student attitudes even if there was no formal evidence. One study in particular, Mvududu (2003), examined the relationship between students' attitudes toward statistics and the use of constructivist strategies in the introductory course in statistics. Two samples of students were selected, one from an undergraduate institution in the United States (U.S.) and one from an undergraduate institution in Zimbabwe. Each sample of students completed two assessment instruments, the Constructivist Learning Environment Survey (CLES) and the Attitude Toward Statistics (ATS) scale. The CLES is a 25-question scale that measures students' preferences towards constructivist strategies in their learning environment for constructivist strategies and whether constructivist strategies are actually present in their classroom. The ATS (Wise, 1985) is an "29-item attitudinal scale (that) consists of two subscales: attitude toward the field of statistics (20 items) and attitude toward the course (9 items)" (ARTIST). Through a principal-components analysis, Mvududu was able to show that the personal relevance of what they are learning was very important to the U.S. students' success in introductory statistics. There was also a significant preferred social component for the U.S. students. This suggests that use of cooperative learning groups in introductory statistics classes is important. The social aspects of learning, combined with real-world data, are components of active learning.

Assessment

In statistics education, as in all of education, there is a current debate as to how to best assess student learning, and more importantly, assess statistical reasoning. New questions have arisen about standard testing techniques in a statistics classroom and whether it authentically measures mathematical and statistical reasoning. Many studies still use traditional tests and quizzes or end of the term point totals to measure statistical understanding. Other studies have relied on pretest/posttests that they themselves have authored in order to measure learning gains (Shaughnessy, 1977). However suggestions have been made that in order to accurately measure students' conceptual understanding, alternative methods of assessment are necessary (Garfield, 1995). "An adequate assessment of student outcomes is not possible using only multiple choice or short answer questions" (Gal & Garfield, 1997, p. 6). Certainly, many researchers have discovered much more about student statistical reasoning through interviews and other open-ended questioning techniques (Makar & Confrey, 2005).

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) report recommends an educator “uses assessment to improve and evaluate student learning” (p. 13). Several alternative assessments, in addition to the more traditional assessments such as quizzes and tests, have been suggested to better assess conceptual understanding of statistical ideas. Course-long projects that allow students to grapple with the issues of data collection and analysis in a real-world way have become more popular in introductory courses (Smith, 1998). One curriculum that has been successful with middle school students is the Authentic Statistics Program (ASP). Students are “transformed into small groups of design teams, where each team constructs a research question, collects data to the question, represents the data graphically, analyzes the data and presents its project to the classroom” (Lajoie, Lavigne, Munsie, & Wilkie, 1998, p. 200). The statistical knowledge needed, such as descriptive statistics, was presented to students through an interactive computer tutorial before students completed the project. In order to validate the different forms of assessment that this classroom used, the researchers followed two of the design teams as case studies through multiple mediums of assessment (pretests, posttests, regular tests, homework, journals, presentation questions) over time. The researchers were interested to find that while both groups improved significantly on the posttest, by analyzing the other forms of assessment, they found that the two groups had made gains in different areas. By identifying how individuals differ in terms of knowledge through these multiple means of assessment, then instruction can be designed based on students’ strengths and weaknesses (Lajoie et al, 1998).

Summary

This chapter was intended as a review of the literature that is relevant to “Students Conceptual Understanding of Variability”. Some early studies that attempted to classify different student misconceptions in probability were reviewed. Studies on reasoning about specific topics in an introductory statistics course, such as distributions and variability were also reviewed. In particular, a multistage research project that studied students’ conceptual understanding of sampling distributions was discussed in detail, as that study was part of the inspiration for this particular research paper. Research which has incorporated different pedagogical approaches into an introductory statistics course, such as active learning and exploratory data analysis, were highlighted. And a discussion of different statistical reasoning models that other researchers have developed based on their work was presented.

What follows is a description of the methods used in my research in Chapter 3, the analysis of the different data that were collected in Chapter 4 and a review of the conclusions, limitations and implications of this research in Chapter 5.

CHAPTER 3

METHODS

Introduction

This study was a combination of qualitative and quantitative methods. The quantitative results will hopefully provide a foundation and background for a more in-depth analysis of the qualitative results. In recent years, mathematics educators have moved away from quantitative methods and relied heavily on qualitative methods. However, a recent report form the American Statistical Association (ASA) recommends finding middle ground between quantitative and qualitative methods (ASA, 2007).

For the qualitative portion of my study, I will be conducting an action research project. Classroom research, also often called action research is defined as a "form of inductive, practical research that focuses on gaining a better understanding of a practice problem or achieving a real change or improvement in the practice context" (Quigley & Kuhne, 1997). Often this term *action research* is used interchangeably with *teacher research* or *classroom-based research* in the literature. Generally, this practice involves the teacher in a classroom being both the practitioner (teacher of the class) and the researcher. In recent years, this form of research has been used more frequently (Ball, 2000). The benefit of this type of research uses "the position of the teacher to ground questions, structure analysis, and represent interpretation" (Ball, p. 365). As an insider, the teacher has insights to her classroom that outsiders do not. One might argue one loses objectivity and transferability in the research (Chance, delMas, & Garfield, 1999), however with the appropriate structure and rigor, “such work can be seen as a means to legitimate perspectives of practice in the construction of knowledge about teaching” (Ball, p. 375).

Classroom research, although initially conducted in elementary and secondary classrooms has, in recent years, become recommended in postsecondary research (Cross & Steadman, 1996). Several statistics education researchers have used this model of research in their own work. Chance, delMas and Garfield (1999) used the following questions to guide them through their classroom research on sampling distributions:

> 1) **What is the problem?** What is not working in the class? What difficulties are students having learning a particular topic or learning from a particular type of instructional activity? The identification of the problem emerges from experience in the classroom, as the teacher observes students, reviews student work, and reflects on this information. As a clearer understanding of the problem emerges, the teacher may also refer to published research to better understand the problem, to see what has already been learned and what is suggested regarding this situation, and to understand what might be causing the difficulty.
>
> 2) **What technique can be used to address the learning problem?** A new instructional technique may be designed and implemented in class, a modification may be made to an existing technique, or alternative materials may be used, to help eliminate the learning problem.
>
> 3) **What type of evidence can be gathered to show whether the implementation is effective?** How will the teacher know if the new technique or materials are successful? What type of assessment data will be gathered? How will it be used and evaluated?
>
> 4) **What should be done next, based on what was learned?** Once a change has been made, and data have been gathered and used to evaluate the impact of the change, the situation is again appraised. Is there still a problem? Is there a need for further change? How might the technique or

> materials be further modified to improve student learning? How should new data be gathered and evaluated? (para. 1)

One of the most important guiding principles one should follow in conducting action research is that "its primary purpose is to simultaneously study and generate knowledge about the very practice that it seeks to change" (Doerr & Tinto, 2000, p. 408). This is a cyclic process in which I, as the teacher, planned and executed the methods of research discussed below; but as I collected data, I analyzed the information and adjusted my methods as necessary. This was especially true in terms of the treatment labs I created. The labs that addressed sampling distributions and confidence intervals were constructed during the quarter of research and were adjusted based on what I had perceived was happening in my class. I also reformulated the questions I sought to answer. "Action research is. …characterized by spiraling cycles of problem identification, systematic data collection, reflection, analysis, data-driven action taken, and finally, problem redefinition" (Johnson, 1993, p. 1).The nature of my research questions originally broadly focused on students' conceptual understanding of variability in general. However, as the quarter progressed and through my initial data analysis, this focus narrowed to specifically how students understand standard deviation and standard error. I had picked those measures originally as my method for assessing students' conceptual understanding of variability in general. However, there is so much prerequisite and co-requisite knowledge that is essential to understanding how these two measures (standard deviation and standard error) relate to general variability in a dataset, and understanding each measure is closely tied to the concepts of probability distribution, sampling distribution and the central limit theorem. Therefore I felt simply looking at how students' understood these measures was not enough to make a definite conclusion about how students conceptually understood variability all together. This is one of the issues that statistics education researchers wrestle with in conducting any research on statistical reasoning on any topic: each concept is inextricably linked to several other concepts (Chance, delMas, & Garfield, 1999; Reading & Reid, 2006). Trying to narrow the focus of one's research is incredibly difficult.

Overview of Methods Used

I compared the results from two sections of an introductory course in statistics that I taught in the Winter Quarter of 2007. One class was the control section and the other was the treatment section. The control section (lecture) was taught with more traditional methods of lecture, discussion, demonstration and some activities. The treatment section (active) was taught with the same methods with the exception of three research labs that addressed the topics of standard deviation, sampling distributions, and confidence intervals. The last two labs were designed to address, indirectly, standard error. Students in both sections completed a pretest and posttest for overall comparisons. The responses of the active class on the research labs were collected and analyzed. A small sample of students from both sections participated in interviews during the eighth week of the quarter.

Location of Study

This study took place a small, Midwestern, four year, liberal arts college. While this institution attracts students from all over the United States, the majority of students are from the state of Ohio. They enroll approximately 3000 students, approximately 600 of which are continuing education adult learners. They have a student-to-faculty ratio of 12 to 1 and offer 56 majors and 41 minors, as well as individualized fields of study. Master's degree programs are offered in education, nursing, and business administration.

This college offers an introductory statistics course at the sophomore level through the mathematics department. It has a college algebra perquisite. This course caters to a wide variety of students and majors and is the only introductory statistics course offered at the college. In addition, this college has a very successful continuing studies program so often there are several non-traditional students enrolled in this statistics course. This wide variety of student and mathematics skill in this course makes it very difficult to teach. The syllabus covers the topics typically taught in an introductory course: descriptive statistics, graphical displays, probability, binomial & normal probabilities, regression, correlation, central limit theorem, and hypothesis testing. It does so with a very strong bent toward the mathematical theory. This is partly because this course does count toward a major in mathematics.

I taught two sections of this course in the winter quarter of 2007. These sections met back to back on Mondays, Wednesdays and Fridays; the first class ran from 9:30 to 10:50 am and the second ran from 11:00 am to 12:20 pm. Because the classes were so close together in time, this allowed me to minimize the differences in teaching schedule between the two sections. The college does not have a computer classroom that is large enough for the number of students in an introductory statistics section, so there was very little opportunity to use any computer tools during class meetings. The regular classroom was outfitted with an instructor computer station and projector, which did allow for computer simulation demonstrations. However, any data analysis that was done by students in the classroom was done with their required graphing calculators.

Participant Selection

Each section of the course started with approximately thirty students. Students were informed of the study on the second day of the quarter and informed consent forms (see Appendix A) were distributed at that time. All students in both sections returned the form and only one student in the lecture section chose not to participate in the study. I discovered in the third week of the quarter, that I had not explicitly asked students for permission to access their Otterbein records for the purposes of demographics in my research[5]. A supplemental informed consent form (see Appendix B) was distributed at that time and all the students still enrolled in the course at that time completed and returned this second informed consent. An additional two students in the lecture section choose not to have their academic records included in the demographics for this study. Regardless of participation in this research study, all students participated in the labs and completion of the pretest Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) and a posttest CAOS. There was no way for any of the students in the class to know who was participating in the study and who was not. The data (labs and test scores) for the students who choose not to participate were simply not analyzed for this

[5] I have permission to access participants' records as a faculty member and had failed to realize I would need separate permission from the participants as the researcher. To fix this oversight, I resubmitted the paper work to the Institutional Review Board and handed out the supplemental informed consent.

research study. In addition, three students in each class did not finish the course, and were not included in any of the analyses.

Interview Subject Selection

As part of the original informed consent, I asked students to indicate if they would be willing to participate in a short interview toward the end of the quarter. Approximately 10 students from each section indicated they would be willing to participate. Later in the quarter, I emailed these students in order to coordinate the interview schedule. As a result of the limited time slots for the interviews, five students from each section were selected for interviews. These students were selected in such a way to ensure equal selection from the lecture and active classes and to ensure that a variety of majors and mathematics backgrounds were represented.

Class Instruction

I was the sole instructor for the statistics classes that were used for this study. As part of my research, I kept a daily log of my own reflections on my instruction, student questions and feedback in class, my perceptions of student attitudes. My goal with this journal was to have a written record of the course's progress that I could refer back to later during the data analysis portion of my study.

I taught both sections of this course with a mixture of lecture methods and active learning methods. Since my study is analyzing specifically whether specific active learning techniques affects a student's conceptual understanding of variability, I wanted especially the participants in the active class to be comfortable and familiar with active learning techniques by the time I got to the labs that are specific to this study. Therefore there were other labs, which both classes completed, that were unrelated to the purposes of my research.

Researcher Perspective

I have now taught statistics courses nearly ten years, in several different settings. My personal experience as a student in my undergraduate introductory statistics course guides what I believe today as an instructor and researcher in statistics education. I am someone who learns by doing. Even in my own mathematics and statistics courses, I always learned by mimicking example problems. Only after I understood the methods and applications, was I able to understand the general theory. The course I took in college was taught entirely through lecture. Fortunately, I had a very good instructor who did lots of examples and his teaching methods matched well with my learning style. However, I did not fully understand the Central Limit Theorem and the connection to sampling distributions until my second course in statistics. I had made it through my first course understanding only that in certain situations that asked about "the probability that mean was …", I needed to use standard error in my calculations instead of standard deviation. I had no idea what a sampling distribution was or how the Central Limit theorem was connected to these problems. It wasn't until I got to the second course that had a more applied focus, analyzing real data sets, that everything started to make sense. I literally had an "a – ha" moment.

Because of my own experience learning statistics and the experiences of students I have observed over the last ten years, I am very much aware as the teacher that many students pass through a statistics course in the same way I did, learning just enough to do the problems but without really understanding concepts. This is why I am so interested in statistics education and the movement toward a more hands-on approach in the introductory course. My regular teaching methods include a mixture of contexts for any particular concept. For example, when I introduce quantitative measures of center, I discuss the formula and the notation for mean and median, I do a sample problem with help from the class, and then I hand out a worksheet that has three datasets (one left skewed, one right skewed, and one symmetric) on it and ask students in groups to calculate the mean and median for each. I also ask them to discuss in groups any relationship they see between the distribution shape and where the mean and median fall. Usually a good number of groups will "discover" the relationship and I always bring the class back together and discuss the results. I believe just the simple act of asking students to do some quick data analysis in class helps them become more engaged with the topics and ideas I am trying to present.

Even with my focus on providing multiple contexts for a concept, I still tend to control the presentation and discussion of the concepts when I teach, and that continued to be true for the control section of this study. The active learning labs in the treatment section were different from my normal pedagogical approach in that I let labs and the activities present the material. Conceptual change theory is predicated on the idea of creating cognitive conflict in students and I designed the labs to help students diagnose their own misconceptions and to help correct them. Students were much more in control of their learning. I was there simply to provide assistance or clarification when it was needed during these labs.

Research Labs

For the purposes of my research I conducted three active-learning labs that focused specifically on aspects of learning variability: one to specifically address the concept of standard deviation and its connection to distribution (Appendix C), one to specifically address the topic of sampling distributions and standard error (Appendix D), and one to specifically address the difference between standard error and standard deviation and how that relates to confidence intervals and hypothesis testing (Appendix E). These three labs had a specific structure to ensure a conceptual change framework. I first asked students to make certain predictions and conjectures. I then asked them to collect data that allowed them to test their conjectures. I asked them to analyze their results and discuss how their results compared to their original conjectures. Finally, I asked some extension questions that were meant to ascertain whether students had learned the concepts presented in the lab and could transfer their conceptions to problems in a different context. These extension questions were obtained from the assessment builder section of the Assessment Resources Tools for Improving Statistical Thinking (ARTIST) website.

> The assessment item database (consists of)…items… pulled from exams of the project staff (co-investigators and advisory board members) and also solicited from the statistics community through a posting on the ARTIST website….The co-investigators and some of the advisors reviewed items to organize them by topic and learning outcome….The database currently consists of more than 1100 items, with new items being added periodically as they are submitted by ARTIST users and reviewed (ARTIST).

This structure of the labs was designed to challenge or reinforce students' prior conceptions of variability, one of the major underpinnings of conceptual change theory.

Pretest/Posttest

As part of the class, I asked students to take the Comprehensive Assessment of Outcomes for a first course in Statistics (CAOS) test at the beginning of the quarter and again at the end of the quarter. Garfield, delMas, Chance & Oooms (2006), originally developed this test over a three-year period as part of a National Science Foundation (NSF)-funded ARTIST project. The test consisted of multiple-choice questions that were specifically designed to assess student's statistical reasoning skills.

> The CAOS test was designed to provide an instrument that would assess students' statistical reasoning after any first course in statistics. Rather than focus on computation and procedures, the CAOS test focuses on statistical literacy and conceptual understanding, with a focus on reasoning about variability. The test was developed through a three-year process of acquiring and writing items, testing and revising items, and gathering evidence of reliability and validity. (ARTIST)

The test has gone through several rounds of feedback and validity testing. Initially, the ARTIST advisory board conducted these revisions. The second round of feedback came from high school AP Statistics teachers who used the CAOS test as part of their course. The third round of feedback and revisions came from 30 statistics instructors who were faculty readers of the Advanced Placement Statistics exam in June of 2005.

Administration of the test was done online and during class time. Because of the lack of a computer classroom that was large enough to accommodate the entire class, a few students had to take the test in a different building. Since the test was accessed by a unique pass-code and was restricted to a specific time period, I was at least able to make sure all students were still taking the test in the same window of time. Completing the pretest CAOS and the posttest CAOS counted as a 10-point lab score each; however, their actual score on the test was not factored into their grade. After the test was administered, I was able to download a copy of the test, with the percentage breakdowns for my class overall, as well as a spreadsheet with the percentage correct for each individual student. Later during the data analysis portion of my study, I contacted the test administrators for the raw scores for all of my students. This was necessary because there were three students who completed the pretest who did not complete the posttest in each section and I needed to eliminate their information from the summaries that ARTIST provided me. I also needed to remove the one student who did not wish to participate in the research study. Since I was selecting specific CAOS test questions for analysis, I needed more detailed information than the regular CAOS test reports provided.

As a postscript to my study, I emailed a small sample of students at the end of the spring quarter to retake the CAOS test. These students had indicated previously that they would be willing to do this on the supplemental informed consent. This was to check for long-term retention of the concepts in the introductory course. Unfortunately, only five students completed this test and they were all from the lecture section. These scores provided me no useful data.

Interviews

Interviews were conducted with eight students, four from each section of the course. Ten interviews were scheduled, but two students failed to show up. The interviews were designed to be approximately half an hour long and took place in the children's area of the college library. Because the children's area is in the basement of library, it is a fairly quiet location. The interviews took place during the eighth week of the quarter on the afternoons of February 21st and February 23rd of 2007. These interviews were be conducted by a fellow Ohio State education graduate student. This graduate student is the head of the mathematics department at a local community college and is also conducting research in the field of statistics education. Since I was concerned about participants feeling pressure or uncomfortable being interviewed by their current statistics instructor, this graduate student became an ideal candidate to conduct these interviews[6]. The interview participants were given a ten-dollar gift certificate to the bookstore as a thank you for their time and to help recruitment of interview subjects. This graduate student recorded each interview with a digital voice recorder. Interview recordings were labeled with a code number.

The individual interviews had two main purposes. First, I wanted to collect more detailed information about these specific students' backgrounds and their mathematics histories before entering my class. Secondly, I asked them three assessment questions that dealt with variability concepts and asked them to respond out loud. One question dealt with standard deviation, one dealt with standard error, and one question dealt with the difference between standard deviation and standard error. The interview protocol is attached in Appendix G. This allowed me to analyze with much more detail how they reasoned through a variability problem and how well they conceptually understood these measures of variability. These questions, as well as the extension questions on the labs,

[6] To return the favor, I conducted the interviews that this graduate student required for his dissertation research.

were pulled from the ARTIST online assessment builder database. The interviewer used a digital recorder and each interview file was named only with a code number. At the end of the interview sessions, the interviewer returned the digital recorder to me and it was stored in my locked office desk until the end of the quarter. I did not listen to the tapes until after the end of the quarter and grades had been assigned. Although there were no identifiers on each interview recording, it should be noted that I knew these students very well, and could recognize each student's voice. In the transcriptions, on the recordings and in all other written work produced as part of this research study, interview subjects were identified only by their interview code number.

Demographic Data

As a member of the faculty at the college, I had access to background information about the student enrolled in my statistics sections. With permission from my students, I retrieved their records for the following information: class standing, major, ACT mathematics score, ACT comprehensive score, college GPA, college mathematics GPA, and score on the college's mathematics placement exam. The main purpose of this was to provide a background picture of the students in these two sections of the course and to demonstrate the wide variety in the type of student enrolled in this course. I also wanted to ensure there were no significant differences between the two sections in terms of mathematics skills or major.

Data Analysis

Quantitative Data

The data analysis began with the CAOS results. I chose to use a paired t-test on each section separately first. The paired t-test is designed to compare means on the same or related subject over time or in different situations. Often this entails subjects being tested in a before-and-after situation. In this case, the pretest/posttest results on the CAOS test qualified as a before-and-after situation. I used this test to see if there was any significant improvement in each section on its own.

I then used a two-sample t-test to compare the average improvement between the lecture class and the active learning class on the entire test, and then on only the questions that directly related to my study (there were 12 questions out of 40 that directly assessed either standard deviation or standard error). A two-sample t-test is generally used to compare the means of two independent groups that have been randomly assigned. Since I was not able to randomly assign students to each section, I also used two-sample test to test the demographic information collected from each class for significant differences in academic ability (college GPA, college mathematics GPA, ACT mathematics score, or college mathematics placement score). So while students were not randomly assigned to each section, I was reasonably confident there were minimal differences in the make-up of the two sections.

Qualitative Data

The qualitative data consisted of written responses to lab questions in the active learning class and the interview transcriptions. Over the course of the quarter, I scanned copies of all the labs (both research and otherwise) for all of the participants in each section. I began the coding of this data by looking through each of the scans of the research labs and simply recording how many students answered each question correctly or incorrectly. Once this was done, I started with the assessment questions on each lab and started making notes about what jumped out me or about things that seemed to be repeated in student responses over and over again. Given my experience teaching this course over the years, I did have some preconceived notions of what I was expecting to find in the data analysis. For example, one of the assessment questions on the standard deviation lab (Appendix C) was a two-part question that should have required similar reasoning. I expected students would either get both parts correct or get both parts incorrect. Also, during the analysis of the central limit theorem lab, I was expecting most students that got the first assessment question incorrect did so because they used the standard deviation instead of the standard error. I found that in a few cases, I was correct; but for the most part analyzing student responses was far more complicated.

After the initial few passes through the research labs, I went back and re-read Makar and Confrey's (2005) article on use of language in statistics. The authors had conducted a study with pre-service teachers and had taken an in-depth look at the type of language these teachers used when describing the concepts of distribution and variation. I looked especially at the categories of non-standard language they had identified and used this as a lens for my own data. I found that I had identified several of the same types of nonstandard phrases that Makar and Confrey had. For example, the word *range* was used liberally by students in their responses, but in a way that was not statistically correct.

I also attempted to use Garfield's model of statistical reasoning (see Table 1 in Chapter 2) as a rubric for "scoring" my students responses. However, I discovered that this was incredibly difficult. My students had written very short responses to the lab questions and I found it nearly impossible to classify a student's thinking neatly into a single category. I simply did not have enough information. I also felt that for the particular purpose of my research, trying to classify these responses into categories did not provide me any further insight into my student's thinking or understanding of variability. So I went back to the data and I simply continued to develop my own classifications grounded in the data.

The interview audio files were also transferred to my computer and transcribed near the end of the spring quarter in 2007. I discovered that the interviews would become my richest source of data. I listened to them once, just to get a sense of how they went, and I was truly shocked. The student responses to the three assessment questions in the interviews were very different from what I was expecting. Most significantly, the students' responses to the interview questions were mostly wrong, even though these were mostly students who were doing well in the class. So I spent a great deal of time transcribing and taking detailed notes on the interviews. This process was very similar to the process I employed for analyzing the research labs. I started by making a note of how many students got each question correct and incorrect. I then went back and tried to find patterns in the responses. It was surprising how much a student's reasoning changed from question to question, even though I felt these questions were connected. Because of this, I ended up mostly analyzing each interview question individually. Later, I went back and tried to find connections across questions. I also matched up the interviews to some of the lab work, to see if that would give me a clearer picture of why students could progress through the research labs successfully, yet still retain incorrect reasoning that would persist through the course and beyond.

I collected these multiple forms of qualitative data in an effort to triangulate the data. "The use of multiple methods, or triangulation, reflects an attempt to secure an in-depth understanding of the phenomenon in question" (Denzin & Lincoln, 2000, p. 5). The multiple methods I employed included interview transcriptions, the CAOS test results and the student written responses to the lab questions.

Timeline

Dates	**Action**
January 5, 2007	CAOS Pretest
January 12, 2007	Research Lab: Standard Deviation
February 16, 2007	Research Lab: Sampling Distribution
February 21, 2007	Research Lab: Confidence Intervals
February 21 & 23, 2007	Student Interviews
March 9, 2007	CAOS Posttest
March 26, 2007 – June 6, 2007	Qualitative Analysis Course Transcription of Interviews Initial Coding and Analysis of Research Labs
May 27, 2007 – June 3, 2007	Window for CAOS deferred posttest
June – September 2007	Data analysis and writing a first draft

Table 4. Timeline of Research Methods

Summary

In this chapter, I described the mixed methods research methodology I employed in conducting my research. Two introductory statistics classes were taught in the winter of 2007 at a small liberal arts college, one in a traditional lecture format, the other in an experimental format with active-learning research labs. I collected pretest and posttest scores on the CAOS test, a statistical reasoning assessment instrument. I also collected written responses to statistical reasoning questions that were placed at the end of each research lab. Interviews were conducted with four students from each section late in the quarter of instruction. This chapter enumerated in detail these different forms of data and explained the rationale behind collecting each form of data.

CHAPTER 4

DATA ANALYSIS

Introduction

In this chapter, I present the data analysis. I will begin with the quantitative results of the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) test that give an overview of the conceptual understanding of the two separate classes. I will then qualitatively analyze written responses in student lab work. Finally, I will analyze the interviews. Both the labs and interviews will be analyzed in the context of the three measures of variability that were addressed in this study: standard deviation, standard error and the difference between standard deviation and standard error. This will provide a more detailed and complete analysis of the conceptual understanding of variability for a sample of students from each of the two different classes. The goal of this study was to answer the following research questions:

> 1) Do labs designed with a conceptual change framework (a make and test conjecture model) have a significant impact on students' conceptual understanding of variability, specifically the measures of variability, in an introductory statistics course?
>
> 2) How do students articulate their understanding of variability?
>
> 3) What differences in understanding of variability are there between students whose introduction to variability measures occurs mainly through

standard lecture/demonstration methods and students whose introduction occurs through active learning/discovery methods?

Demographics

At the end of the quarter, there were twenty-seven students remaining in the lecture class and there were twenty-six students remaining in the active class. For the purposes of the data analysis, only these participants were included. Table 5 provides a comparison of the demographics between the lecture class (control section) and the active class (treatment class). The p-values for two sample t-tests testing the difference between the quantitative demographic measures are included. A Mann-Whitney statistic was used to test for a difference between class standing for each class.

	Lecture Class	**Active Class**	**P-Value**
Class Standing	Freshman 37.9% Sophomore 37.9% Junior 13.8% Senior 10.3%	Freshman 10.3% Sophomore 37.9% Junior 27.8% Senior 20.7% Special Undergraduate[7] 3.4%	.0012
Average ACT Mathematics	24.8	24.1	.455
Average College GPA	3.331	3.348	.907
Average College Mathematics GPA	3.045	3.101	.796

Table 5. Demographic summary of each class section

[7] A special undergraduate is a student that has temporarily transferred to take classes, but will not be at the college permanently.

The p-values in Table 5 from the two sample t-tests showed that there was no statistically significant differences between the lecture and active classes in terms of college GPA, college mathematics GPA, and ACT mathematics score, that needed to be taken into account for the quantitative analyses. There was a significant difference between class standing. This is important to note as the active class contained more upperclassmen. Upperclassmen are more experienced in college and may have better study skills.

Both sections of the course had a wide variety of majors represented. Figure 4 shows separate bar charts of the distribution of majors in each class section.

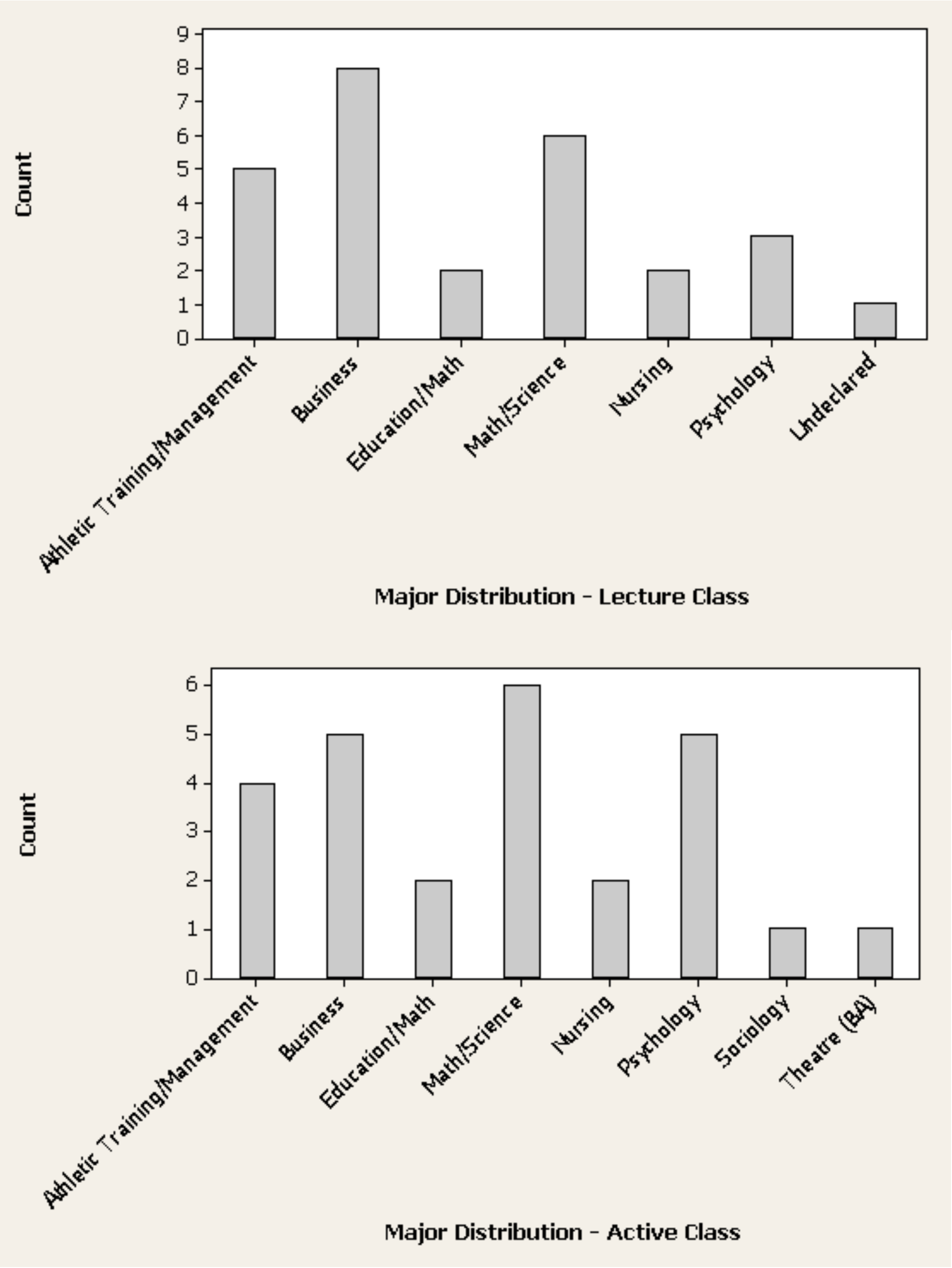
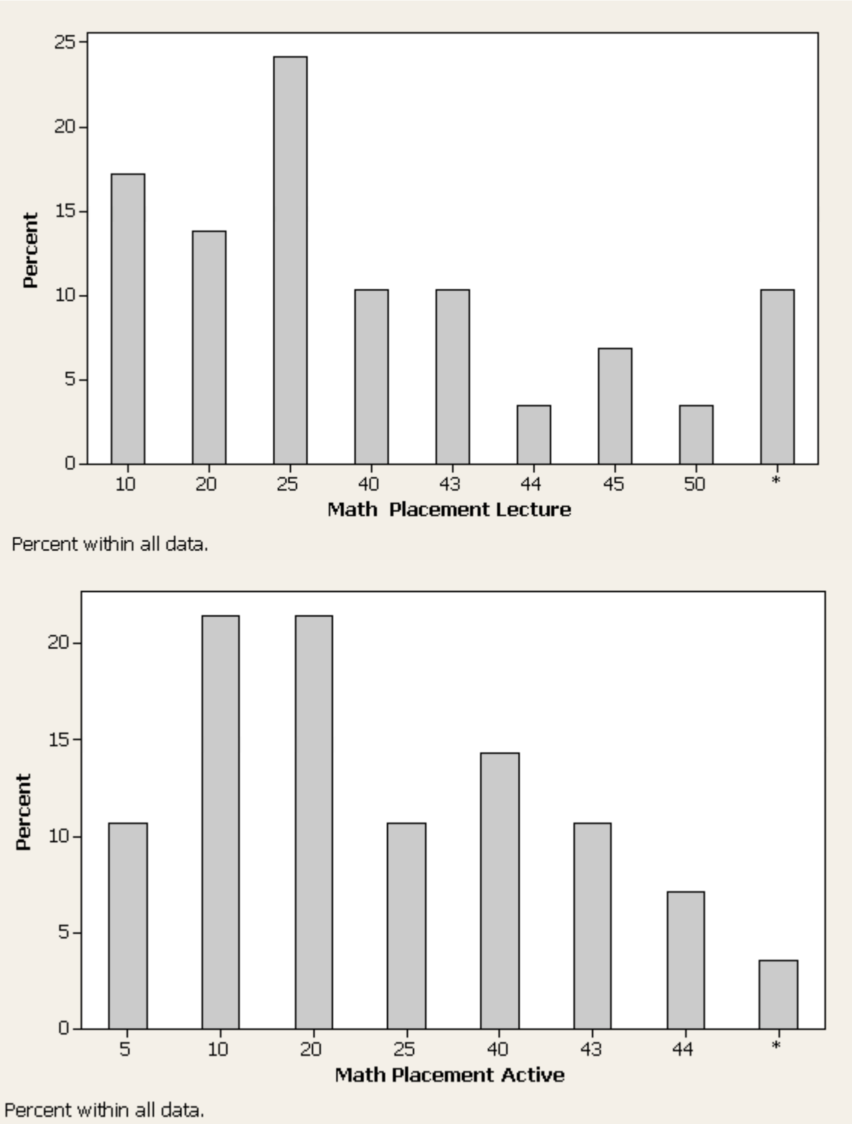


Figure 4. Distribution of majors for each class section

In larger undergraduate institutions, there are different introductory statistics course for different types of majors. Having vastly different majors, such as Business, Mathematics, Education and Athletic Training/Management, all in the same introductory course is unusual and can make teaching the course difficult. It is assumed that mathematics majors will come to the course with a different set of mathematical reasoning skills than a sports management major. The different mathematical skill sets can be indirectly seen through the distribution of mathematics placement scores for each class in Figure 5.



Placement Score Key	5	10	20	25	40	43	44	*
Initial Course Placement	Remedial Algebra	Pre-College Algebra	College Algebra	Inter-mediate Algebra	Pre-Calc	Calc 1	Calc 2	Un-know n

Figure 5. Mathematics placement scores distribution for each class section

The college requires each incoming student to take the Mathematics department's placement test. The distributions in Figure 5 show the initial mathematics course placement of students when they arrived at this college. Since this is a liberal arts college, all students are required to take at least one college level mathematics course. The college algebra course is a prerequisite for the introductory statistics course, so all students theoretically have this minimum mathematical knowledge when they enter the course. Although, a placement score is not necessarily a perfect indicator of mathematics skills, the above distributions do show a great deal of variation and that there are many students who have much more than the required algebra background (such as calculus). Both classes, however, had this wide range of mathematics placement scores. A Mann-Whitney test compared the placement scores and showed no evidence of a significant difference between the two classes (p-value = .2525).

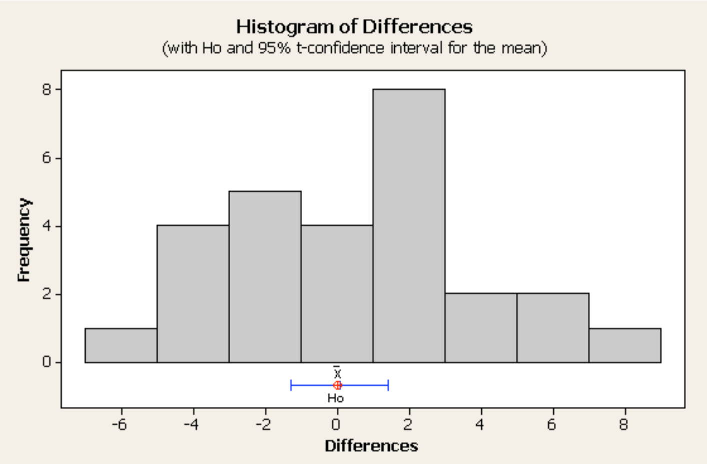
Quantitative Results – the CAOS test

Analysis of complete CAOS test results

The CAOS test is an assessment tool that measures statistical reasoning skills, with an emphasis on reasoning about variability. However, the test measures more than just reasoning on the three measures of variability that were the focus of this study.

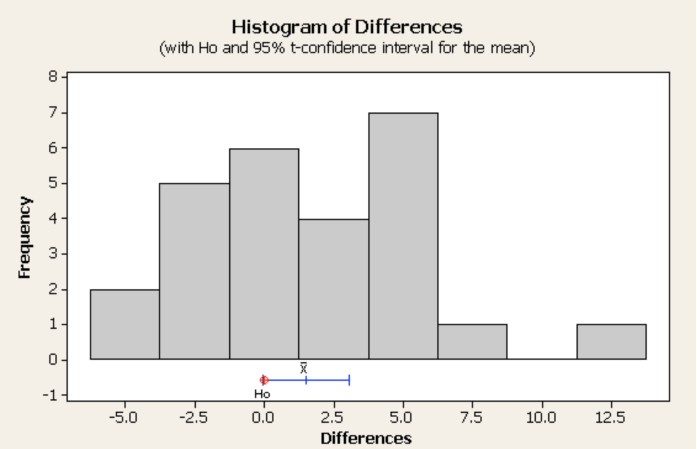
> The CAOS test was designed to provide an instrument that would assess students' statistical reasoning after any first course in statistics. Rather than focus on computation and procedures, the CAOS test focuses on statistical literacy and conceptual understanding, with a focus on reasoning about variability. The test was developed through a three-year process of acquiring and writing items, testing and revising items, and gathering evidence of reliability and validity. (ARTIST)

So while the CAOS test focused on reasoning on variability, there were many test items that did not directly address my research questions. These test items covered topics, such as measures of center and hypothesis testing, that I had introduced in both classes using the same teaching method. Therefore, although students took the entire CAOS test both at the beginning of the quarter and the end of the quarter, the overall scores from the entire test were not useful in answering any of my research questions. However, as part of the process of analyzing the data, I conducted a paired sample t-test on the complete test results in each class separately to see if there was any evidence of improvement in statistical reasoning. The *Minitab* statistical analysis and output from this paired t-test for the entire CAOS test results fro each class separately are shown in Figure 6. A histogram of the score difference (posttest- pretest) is shown for each class as well the statistical results of a hypothesis test, testing whether these differences equaled zero.



Lecture Class	**N**	**Mean**	**StDev**	**SE Mean**
Posttest	27	18.667	4.169	0.802
Pretest	27	18.593	4.060	0.781
Difference	27	0.074	3.407	0.656

95% CI for mean difference: (-1.274, 1.422)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.11 P-Value = 0.911



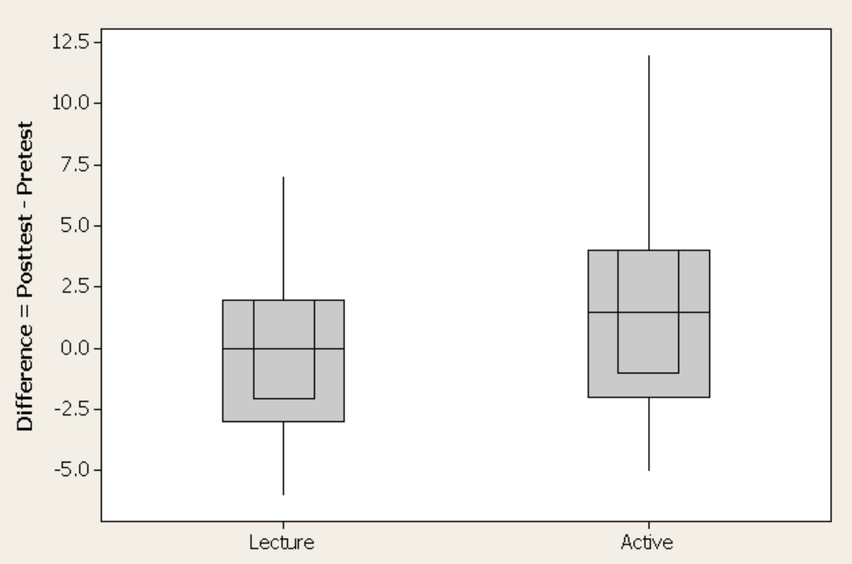
Active Class	**N**	**Mean**	**StDev**	**SE Mean**
Posttest	26	18.615	3.889	0.763
Pretest	26	17.115	2.833	0.556
Difference	26	1.500	3.839	0.753

95% CI for mean difference: (-0.051, 3.051)
T-Test of mean difference = 0 (vs not = 0): T-Value = 1.99 P-Value = 0.057

Figure 6. Statistical analysis of CAOS test results (paired t-test) from both class sections

The results in Figure 6 show that there was a somewhat significant increase in the CAOS scores for the active section with a p-value of 0.057. This seems to indicate that the active class made some reasoning gains, while there is no evidence of a change in the lecture class. Because of the focus of the course and its associated textbook being more calculation-based and not necessarily focused on statistical reasoning explicitly, I was not expecting to see a significant improvement on the CAOS test as a whole. It was encouraging to see even a slight improvement on the CAOS scores for the active section. This may be an indication that the active learning labs had an effect on students' reasoning skills.

In order to compare level of improvement between the sections, I performed a Mann-Whitney test, a nonparametric test that compares the improvement ranks for each class. Since I had attempted to keep the classes content and style roughly the same, with the exception of the sessions that dealt with standard deviation and standard error, I did not expect to have a significant p-value. Figure 7 shows the *Minitab* statistical output from this Mann-Whitney test.



```
Mann-Whitney Test and CI: Lecture, Active

          N  Median
Lecture  27   0.000
Active   26   1.500

Point estimate for ETA1-ETA2 is -1.500
95.1 Percent CI for ETA1-ETA2 is (-3.001,1.000)
W = 652.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1735
The test is significant at 0.1717 (adjusted for ties)
```

Figure 7. Statistical output of Mann-Whitney test comparing improvement levels between each section.

As expected, with a p-value of .1735, Figure 7 shows there was not a significant difference between the level of reasoning improvement between the lecture class and the active class on the overall CAOS test. It was interesting in light of the fact even though the active class had shown a somewhat significant improvement from pretest to posttest, the improvement level between the two classes was not significant. However, after analyzing the histograms of the differences in each section, it became clear that one student did much better on the posttest (12-point improvement on a 40 point test) than on the pretest in the active class. Otherwise, the distribution of differences was roughly the same for each section. In such small sample sizes, one student can have much more of an effect in the paired t-test. When the differences were compared between classes, that student's score had much less of an effect on the results of the Mann-Whitney test.

The results from the paired t-test on each class and the comparison between the two sections demonstrate there may be some evidence that the active class improved their overall reasoning skills over the course of the quarter, while the lecture class showed no significant improvement. However, it is impossible to draw any meaningful conclusions in the context of this study from these statistics based on the entire CAOS test.

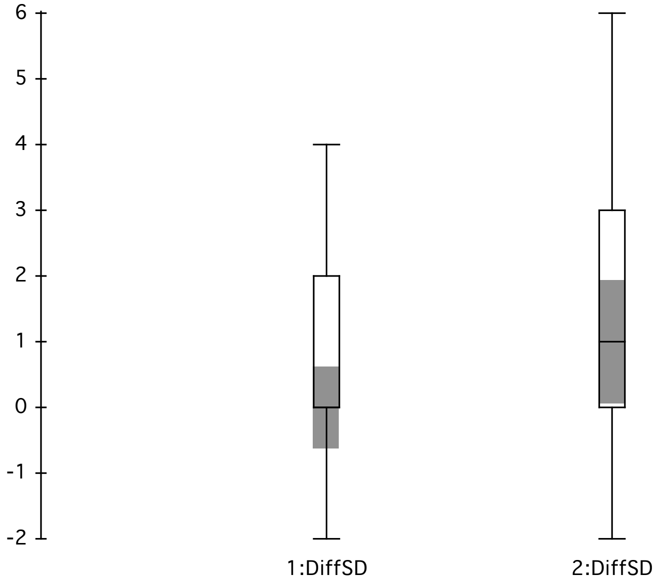
Analysis of research CAOS questions

In order to directly address the research questions, at the conclusion of the quarter, I selected items from the CAOS test that directly assessed student's reasoning about standard deviation, standard error and the difference between these two measures. Only these questions were statistically analyzed for the purposes of answering the research questions. The online ARTIST system provided a question-by-question percentage correct for each of the classes as well as a breakdown of the percentage of the class that chose each of the remaining wrong answers.   I was also able to obtain the raw results for each student from the ARTIST administrators. This allowed me to more extensively and accurately analyze the results. There were three students in each class who dropped the course after they had taken the pretest and therefore needed to be cut from the pretest/posttest analysis. Also the student, who had choosen not to participate in the study, was eliminated from the analysis.

According to my interpretation, there were seven questions on the CAOS test that directly measured reasoning about standard deviation, four that directly measured reasoning about standard error and one question that asked students to reason about the difference between standard deviation and standard error. On these 12 questions, the lecture class did not show significant improvement (p-value = .18) while the active class did show significant improvement (p-value = .004)[8]. This does seem to indicate that the active learning labs improved student's reasoning skills on variability as it applies to standard deviation and standard error. However, once the results were grouped by concept measured, the data showed some interesting trends. Figure 8 shows side-by-side boxplots of the improvement of each class for the questions concerning standard deviation and the statistical results from a Mann – Whitney test comparing the improvement levels. Figure 9 shows side-by-side boxplots of the improvement of each class for the questions concerning standard error and the statistical results from a Mann – Whitney test comparing the improvement levels. And Table 6 shows a contingency table

[8] These p-values are from a paired t-test that was performed on each class separately.

comparing student responses on CAOS Question 32, which addressed understanding the difference between standard deviation and standard error.

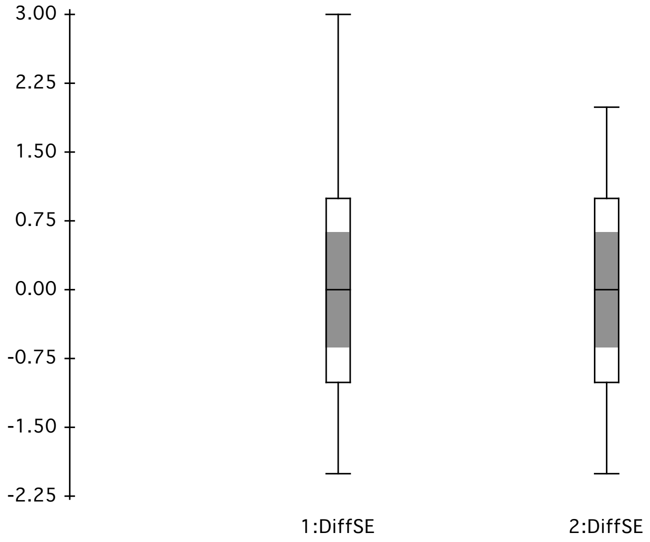


```
Mann-Whitney U
No Selector
Individual  Alpha Level  0.05
Ho: Median1 = Median2  Ha:   Median1 ≠ Median2
Ties Included


1:DiffSD  -  2:DiffSD:
Test Ho: Median(1:DiffSD) = Median(2:DiffSD) vs Ha: Median(1:DiffSD) ≠ Median(2:DiffSD)
                          Rank  Totals   Cases   Mean  Rank
1:DiffSD                           613      27       22.70
2:DiffSD                           818      26       31.46
Total                             2757      53       52.02
Ties  Between  Groups             1326      51          26
U-Statistic:  235
U-prime:  467
Sets of ties between all included observations:  7
Variance:  3159
Adjustment To Variance For Ties:  -170.53
Expected Value:  351
z-Statistic:  -2.1219
p = 0.0338
Reject Ho at Alpha = 0.05
```

(1= Lecture Class, 2 = Active Class, SD = Standard Deviation)

Figure 8. Statistical output comparing improvement between lecture and active sections on standard deviation CAOS test questions.



```
Mann-Whitney U
No Selector
Individual Alpha Level 0.05
Ho: Median1 = Median2  Ha:  Median1 ≠ Median2
Ties Included


1:DiffSE - 2:DiffSE:
Test Ho: Median(1:DiffSE) = Median(2:DiffSE) vs Ha: Median(1:DiffSE) ≠ Median(2:DiffSE)
                        Rank  Totals   Cases   Mean  Rank
1:DiffSE                        702      27        26.02
2:DiffSE                        728      26        28.02
Total                          2757      53        52.02
Ties Between Groups            1326      51           26
U-Statistic:  324
U-prime:  378
Sets of ties between all included observations:  4
Variance:  3159
Adjustment To Variance For Ties:  -252.04
Expected Value:  351
z-Statistic:  -0.49150
p = 0.6231
Fail to reject Ho at Alpha = 0.05
```

(1= Lecture Class, 2 = Active Class, SD = Standard Error)

Figure 9. Statistical output comparing improvement between lecture and active sections on standard error CAOS test questions.

		CAOS POSTTEST (Question 32)	
		Incorrect	Correct
CAOS PRETEST (Question 32)	Incorrect	38	4
	Correct	7	4

Table 6. Contingency table of student responses to CAOS question 32

It is clear from Figure 8 that the active class made significant reasoning gains on the concept of standard deviation over the lecture class. However, the active class performed statistically the same as the lecture class on the concept of standard error (Figure 9) and performed statistically the same on CAOS question 32, the question that addressed understanding the difference between standard deviation and standard error. Table 6 shows that there were 38 students that got CAOS question 32 incorrect on both the pretest and the posttest, four students who were correct on both tests, and four students that went from an incorrect response to a correct response. These four students were distributed evenly between the active and the lecture class (two each). The result that was unexpected was the 7 students who went from a correct response to an incorrect response on the posttest for this question. This needed further analysis.

Analysis of CAOS question 32

CAOS question 32 specifically assessed whether students understood when to apply standard deviation versus standard error (referred to as sampling error below) and is reprinted here.

32) It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches. Which of the following is the most appropriate statistical conclusion?

A. The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.

B. The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.

C. The researchers can conclude that the fish are smaller than what is normal because of the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.

The correct response to question 32 is answer C. Knowing to use the standard error (or sampling error) requires students to understand that the mean of a sample will vary much less than an individual fish. Standard deviation measures how much variability should be expected for one fish at a time, whereas the sampling error or standard error measures how much variability can be expected for the average of a sample. Table 7 shows the distribution of answers to this question in each class. In particular, the table shows how many students changed their answers to CAOS question 32 and how they changed them.

		Lecture Class Answer Posttest	
Answer Pretest	**A**	**B**	**C** (Correct)
A	14	2	1
B	4	3	1
C (Correct)	0	0	2

		Active Class Answer Posttest	
Answer Pretest	**A**	**B**	**C** (Correct)
A	12	1	2
B	4	0	0
C (Correct)	3	4	0

Table 7. Distribution of answers for CAOS question 32

Response A to CAOS question 32 indicated that students understood that some variability is to be expected with a sample but did not understand that standard deviation was an inappropriate measure to use. Response B indicated that students did not understand that there should be some variability in the sample from the population. Since response C was the correct response, Table 7 shows a disturbing trend for the active class. The number of answers to the incorrect responses (A & B) in the active class increased from the pretest to the posttest. Table 7 shows that of the seven students who correctly answered question 32 on the pretest, all answered it incorrectly on the posttest. In each class, two students who answered question 32 incorrectly on the pretest then answered the question correctly on the posttest.

The result in Table 7 that stands out, however, is that the overwhelming number of students in both classes that choose incorrect response A. One of the challenges of teaching students when to use standard error is that the assessment questions usually provide the population standard deviation. Students are supposed to recognize that although the standard deviation is provided and also is a measure of variability, it does not measure the variability among sample means. This requires students to understand the difference between standard deviation and standard error and to recognize that although essentially standard deviation and standard error measure conceptually the same idea, they do so on two different distributions. It is not that surprising then that students overwhelmingly choose A as their response, the response that used standard deviation incorrectly as the measure of variability. This indicates that students generally understood that some variability was to be expected, and that a sample of one hundred from a population may not have the same mean as the population mean. However, students still did not understand how the two measures, standard deviation and standard error, differ or when each is appropriate to apply. A question similar to this had been placed at the end of the sampling distribution lab and in the interview protocol. There were indications from the responses to this lab question, which will be analyzed more extensively later in this chapter, and one of the interview questions, that knowing when standard error versus when standard deviation is the appropriate measure of variability was not well understood by students.

Unfortunately, with a multiple-choice test, it is impossible to understand what a student's thought process was when they answered a particular question. The results from the CAOS test provided an excellent overview of how the two classes performed overall on the topics addressed by this research. I further investigated students' thought processes in this research as they reasoned about variability as they were learning the concepts (through the follow-up questions on the labs) and after they were theoretically comfortable with the concept of variability (through the interview transcripts during the eighth week of the quarter). What follows is an in-depth analysis of the written responses of students in the active class from these research labs and an in-depth analysis of the interview transcripts conducted with students in both classes.

Data Analysis Codes

In the process of analyzing the data, I assigned codes to the research labs and the interview transcriptions. Table 8 shows the key to these codes that will be used throughout the rest of this chapter.

Research Labs	
Standard Deviation Lab	SDL
Sampling Distribution Lab	CLT
Confidence Interval Lab	CI

Interview Transcriptions	
Interview Subject - Lecture Class	LECTURE
Interview Subject – Active Class	ACTIVE

Table 8. Data analysis code descriptions

In this document, each of the codes from Table 8 is followed by a number, i.e. SDL17 or LECTURE08. The numbering for the lab codes was completely arbitrary. Each lab was numbered in the order that they were in the stack of labs. These numbers are crossed referenced across labs and interviews, i.e. SDL14 refers to the same student as CI14.

Research Lab (Standard Deviation)

Overview of instruction

During the second week of the quarter, I introduced the topic of standard deviation in both classes. In both sections, I initially presented the concept of standard deviation by introducing the formula and showing an example calculation on a small dataset. Although many statistical formulas are abstract, the standard deviation formula is unique. In the process of calculating the standard deviation, a student must find the distance of each data point from the mean (and therefore introduce the idea of a deviation) and then average the deviations (and therefore introduce the idea of average distance away from the mean). I discussed, using the structure of the standard deviation formula as a guide, how this measure reflected not only the range of a dataset but also measured how far, on average, data points fell from the mean. I then showed the class how to calculate this number in the graphing calculator and did a couple of example problems from the textbook.

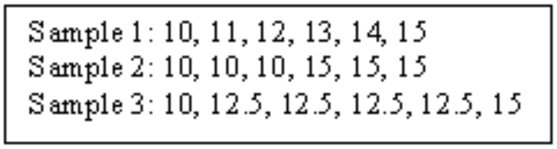
For the next meeting of the lecture class, I lectured on the connection between standard deviation and the distribution of the data, demonstrating that not only does the range effects the value of the standard deviation, but also the location of the data within that range. I had some overhead examples to demonstrate my point[9]. I then asked

[9] These examples were three of the side-by-side distributions that were used in the part b of the standard deviation research lab for the active class. See Appendix D.

students to look at some distributions (constructed from data collected in class) and guess which would have the larger standard deviation. This was similar to the exercise posed to the active class, but done as a class discussion rather than individually and on their own.

For the next meeting of the active class, I had sent students home with a pre-lab questionnaire that asked students to predict which of two variables (based on data that was collected from their classmates) would have a larger standard deviation and why. The next class session was then devoted to completing the standard deviation lab (Appendix C). Students had to complete part b of the lab (stating which histogram had the larger standard deviation) and check their answers with me before I gave them the collected class data for the rest of the lab. They finished the analysis questions at the end of the lab for homework. Included in these analysis questions was a statistical reasoning question designed so that students had to apply what they had just learned on a new and slightly different type of problem. This was the question I used to determine whether students had fully grasped the concept of standard deviation and how it connected to data. This assessment question from the standard deviation lab is reprinted in Figure 10.

3) Consider the following samples of quiz scores and answer the questions without doing any calculations.



Sample 1: 10, 11, 12, 13, 14, 15
Sample 2: 10, 10, 10, 15, 15, 15
Sample 3: 10, 12.5, 12.5, 12.5, 12.5, 15

a. Which sample has greatest standard deviation? Why?

b. Which sample has the smallest standard deviation? Why?

Figure 10. Statistical reasoning assessment question from standard deviation research lab

Analysis

The expectation was that student responses to the two questions in Figure 10 would be the same in terms of logic, i.e. they would either be both correct or both incorrect. However, that was not necessarily the case. Nine students got at one of the two questions incorrect. Table 8 shows the distribution of correct and incorrect responses to the reasoning assessment question in Figure 10.

Number of students	Question a	Question b
18	Correct	Correct
3	Incorrect	Incorrect
1	Incorrect	Correct
5	Correct	Incorrect

Table 9. Distribution of responses to standard deviation lab reasoning assessment in the active class

Twenty-one of the students were consistent in their answers to this question and six were inconsistent.

I included this question not only to measure how successful the lab had been at teaching the desired concepts, but also to analyze how students were articulating their understanding of standard deviation. Makar and Conley (2005) argue that students may articulate a rich and complex understanding of variability, but may do so using non-standard language. In analyzing the written responses, use of non-standard language was one of several commonalities in students' explanations of their reasoning. Several students incorrectly used statistical language, even though they answered the question correctly. For example, several students used the term *range* in the description of their reasoning. Range is technically defined as the distance between the highest and lowest data point in a sample. The statistical definition of range had been discussed briefly in the class meeting prior to the lab. In this assessment question (Figure 10), all three samples have exactly the same range. In the data analysis, responses are prefaced with a or b to indicate which part of the assessment question a student was answering.

> a. Sample 2. The range is the largest.
> b. Maybe sample 3. The range is smaller than sample 1. (SDL2)

> a. Sample 2 + sample 1. Greatest range.
> b. Sample 3. It has the smallest range. (SDL13)

> a. Without doing calculations, I think that sample 2 would have the greatest standard deviation because it has a range from 10 to 15.
> b. I think sample 3 would have the smallest because it doesn't have a very big range in the data set. (SDL14)

All three students correctly identified sample 2 as the answer to question a and sample 3 as the answer to question b. However their usage of the term range was statistically incorrect. While range is a statistical measure, most students appear to use the term *range* in their common vernacular. Perhaps their use of the term range reflects not a lack of conceptual understanding, but a lack of vocabulary to accurately describe the idea of how the data were distributed within the range. Other students were more successful in explaining their reasoning in correct statistical language. For example:

> a. Sample 2 has the greatest standard deviation. All the samples have the same range, so now the distribution of the data is analyzed. Sample 2 only has data at the ends of the range, where the other two samples have data throughout the range. This data shows that sample 2 will deviate more from its average than the other two samples.
> b. I would say that sample 3 would have the smallest standard deviation. Sample 3 would have a bell-shaped graph because the data is evenly distributed on both sides of its peek. Since all the samples have the same range, this symmetric distribution of its data makes it a likely candidate for having the smallest standard deviation. (SDL3)

The students who did use correct statistical language tended to have longer responses. This may be an underlying factor to why so many responses had the correct answer but incorrect reasoning. In my experience, students tend to be hesitant to write more than they absolutely must to answer a question, especially if they were not confident about their answers.

Discussions of distribution were also common in student responses, although again the vocabulary used to describe the shape of the distribution was at times statistically incorrect. Many researchers argue that it is extremely important for students to understand the connection between statistical measures and graphical displays of data to successfully understand statistical concepts (Liu & delMas, 2005; Reading & Reid, 2006). The purpose of the portion of the lab that asked students to guess which graph had a larger standard deviation was to "help students discover that the standard deviation is a measure of the density of values about the mean of a distribution and to become more aware of how clusters, gaps, and extreme values affect the standard deviation" (delMas, 2001, para. 1).

A couple of students referred to the distribution of data explicitly in their explanations.

> b. Sample 3 b/c it has an even distribution. (SDL24)

Although this student explicitly mentioned the distribution, Sample 1 is an even distribution, not Sample 3. It is unclear from this response what this student understood an even distribution to be. Others were clearer in their answers.

> b. Sample 3 because it is a bell-shaped curve and 2/3 of the calculations
> are the same. (SDL12)

Most students, however, described the distribution of the data more indirectly. For example:

> a. 2, because the only numbers are the maximum and minimum, which makes the mean different from all the numbers.
> b. 3, because it has 4 numbers that are the same. (SDL10)
>
> a. Sample 2 has greatest deviation because data more spread out from center. Also the data is adding less middle #'s to provide for standard deviation.
> b. Sample 3 because most the data is already located in the center. (SDL 27)
>
> a. Sample 2 because top loaded and end loaded. (SDL24)
>
> a. #2 because of the jump from 10 to 15 without any numbers in between, resulting in the greatest standard deviation.
> b. Sample #3 because of the repeating 12.5's, results in a smaller standard deviation. (SDL8)

All of these students demonstrated an understanding of the connection between how the data were distributed and the size of the standard deviation, but again did so using mostly non-technical language. It is important to recognize that students may understand the concepts, but may not have the vocabulary to explain their thinking in a statistically correct way. This is consistent with other research findings that introductory students often talk about variability correctly but without a correct technical vocabulary (Makar & Confrey, 2005; Hammerman & Rubin, 2004). It is important for statistics instructors to encourage conceptual understanding first. The technical vocabulary will develop eventually (Makar & Confrey, 2005).

The students who missed either of these questions tended to have either incorrect reasoning as it related to distribution or incorrect reasoning as it related to standard deviation, but not necessarily both. One student understood that Sample 1 had a uniform distribution, but failed to understand that since Sample 3 was more condensed around the mean. A sample, which is more condensed around the mean, would have a lower standard deviation.

> b. Sample 1 because it is level all the way across, representing each number in the range. (SDL22)

Another student did not seem to understand how to analyze the distribution, but did seem to have an idea of what she should be looking for in terms of the standard deviation.

> a. Sample 1, the data is the least consistent, and most different.
> b. Sample 2, there are no other options than 10 and 15. Not much change, as both have their data pieces as well. (SDL 01)

She understood that low standard deviation meant data were similar, but did not understand that it had to be similar to the mean.

So for a few students, the lab was not successful in fully fleshing out their conceptions of standard deviation and how it related to the data distribution, but for the most part students appeared to have discovered the concepts for which the lab was designed. An interesting side-note: in analyzing the lab, the phrase "average distance from the mean" came up in many of the students' reasoning explanations. This is significant because it is the phrase I had used in the previous class meeting to sum up approximately what standard deviation measured. This phrase was used in correct as well as incorrect answers. For example:

> a. Sample 2: Because the average distance away from the mean would be greater without the numbers in between 10 and 15 to bring the average down.
> b. Sample 1: The average of the distance away from the mean would be brought down by the data points between 10 and 15. (SDL06)

It should be noted, however, that while this student technically answered the question incorrectly, they still demonstrated a conceptual understanding of standard deviation, i.e. data points that are closer to the mean will reduce the value of the standard deviation. This highlights the importance of checking to see that students actually do fully understand the concepts being taught. Many researchers have found that students can repeat back what an instructor says and have the appearance of comprehension, but do so without actually understanding the underlying concepts (Ben-Zvi & Garfield, 2004). Unfortunately in some cases, student use of my catchphrase for a definition of standard deviation masked how students were thinking conceptually about standard deviation and limited my ability as the researcher to decipher whether students actually understood what the phrase meant.

Research Lab (Sampling Distributions)

Overview of instruction

It is difficult to separate the idea of standard error of the mean from the ideas of a sampling distribution. For this reason, in both classes, standard error was presented in the context of learning what a sampling distribution of the sample mean was. For the lecture class, I constructed a probability distribution using the years of pennies from a large jar of pennies I keep in my office. I passed out pennies to each student and entered the year of one of their pennies into a dataset of penny years I had already constructed in Minitab prior to class[10]. I then asked each student to calculate the average of five of the pennies in

[10] This was a distribution of about half of the pennies (roughly 250 pennies) in the jar. I was not able to enter every single penny in the jar because of time restraints. This was my

front of them and entered those averages alongside averages of size five I had already drawn from using Minitab's simulation capabilities. I also had averages from samples of size 30 preloaded into Minitab. I then used all of this data to create a probability distribution of the population of pennies, an estimated sampling distribution for the samples of size five and an estimated sampling distribution of the samples of size 30. I used these three distributions as the backdrop to a discussion about what the sampling distribution of the sample means looks like and how it relates to the probability distribution. This led to a definition of all the aspects of the Central Limit Theorem and the formula of standard error as the standard deviation divided by the square root of the sample size.

In the active class, I designed the sampling distribution lab so that students experienced building a sampling distribution from scratch and discovered how standard error measures the variability in a sampling distribution (or at least discovered the idea that the sampling distribution is narrower than a probability distribution and that it varies depending on how big a sample was taken). The lab was adapted from an activity in the textbook *Activity Based Statistics* (Gnanadesikan, Scheaffer, Watkins & Witmer, 1997). As with each of the research labs that were used in the active class, there was a list of questions that students were asked to complete prior to the class in which the lab was actually completed. The extension questions at the end of the lab were designed to measure if students understood the concepts presented in the lab. In this lab, there were

"population of pennies". Students drew their sample of pennies from the same jar. I entered one of each student's pennies so they would feel a part of the population, even though there is a possibility a penny could be counted twice.

three different assessment questions (each downloaded from the ARTIST website) and each had been picked for a particular reason.

Description of reasoning assessment questions

The first assessment question (Post Q1) was designed to assess whether students understood the difference between a population distribution and a sampling distribution and how the variability changes when one looks at the average of a sample versus one data point from one individual. It is reprinted here:

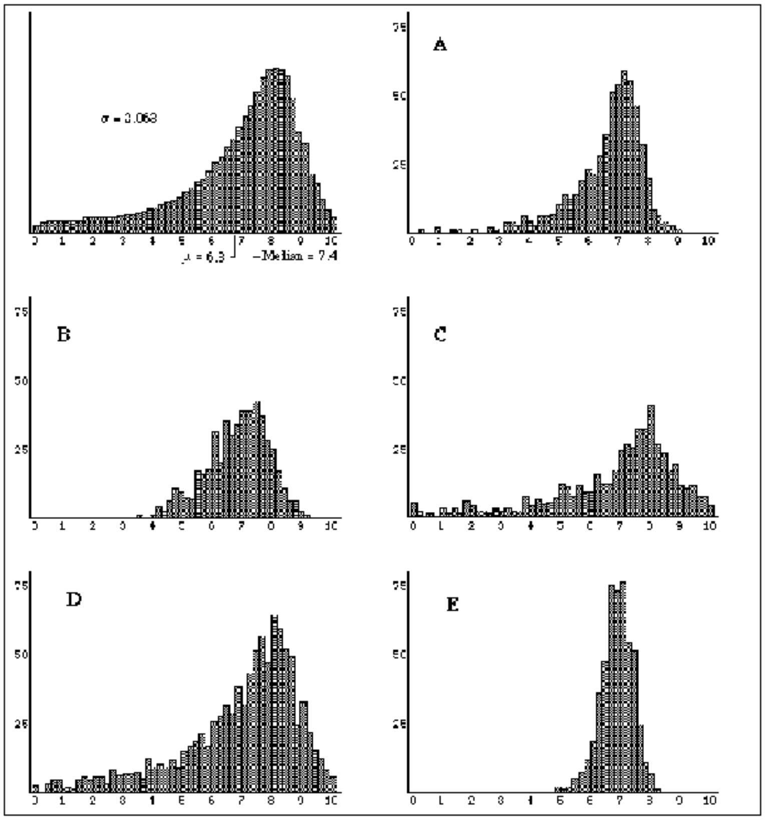
> Post Q1: The distribution of Math SAT scores is Normal with mean 455 and standard deviation 102. Would you expect to find a difference between (a) the probability that a single person scores above 470 and (b) the probability that in a random sample of 35 people, the sample mean (μ) is above 470. Why or why not? Explain.

The second assessment question (Post Q2) was included to assess whether students understood the connection sample size and changes in variability and is as follows:

> Post Q2: Alice, Ben, Connie and Dwayne have each taken a random sample of students from their school to estimate the variability in amount spent on movie tickets this summer. Alice asked 10 people, Ben 30, Connie 50, and Dwayne 70. Whose sample standard deviation probably differs most from true population standard deviation? (Correct answer is Alice.)

The third question (Post Q3) assessed whether students understood the graphical displays of a population distribution and a sampling distribution and whether they fully understood the difference between the two. Since the question asked for a sampling distribution of sample means where only one test score was selected at a time, students had to recognize that this was not a central limit theorem question at all. The sampling distribution should look the same as to the original population distribution. The question is reprinted here:

> Post Q3: The upper left graph is the distribution for a population of test scores. Each of the other five graphs, labeled A to E represent possible sampling distributions of sample means for 500 random samples drawn from the population.



> Which graph represents a sampling distribution of sample means where only 1 test score was selected at a time? Justify your choice.
> (graph C is the correct answer).

It should be noted that later analysis of this question revealed some problems with its wording. It should say the that A through E represent *estimated* sampling distributions. An actual sampling distribution, where the sample size is equal to one, should look identical to the population distribution. This would lead students to correctly answering D. However, an estimated sampling distribution, where only 500 samples are represented, would correctly correspond to Answer C. Also the graphic is difficult to read and it is unclear what the vertical axis scale represents. This question was not extensively analyzed for the purposes of this research, so these problems to do affect a great deal of the analysis. They should be noted nonetheless.

Analysis

Student responses to these assessment questions were disconcerting to me as the instructor. Many students answered the pre-lab questions correctly, but answered the post-lab questions incorrectly. This seems to indicate that the labs hurt conceptual understanding more than they helped. The pre-lab (Pre Q4 and Pre Q5) questions assessed a more general understanding of the nature of sampling distributions without actually defining a sampling distribution or the Central Limit Theorem.

> Pre Q4: If everyone in the class reaches into the same jar of pennies and finds the average age of their handful (about 25 pennies), do you expect that everyone will come in with the same average or different averages? If not, how much variability do you expect between everyone's averages? (A lot, some, or almost no variability?)
>
> Pre Q5: Would changing how many pennies everyone drew from 25 to 100 change how much variability there is between everyone's averages? Explain.

Other researchers, especially in the area of sampling distributions, have found the same result with their own research (Garfield, 2007): that students actually have a basic conception about how a distribution of averages will perform before they learn about sampling distributions, but seem to lose those reasoning skills after they have been taught sampling distributions and the Central Limit Theorem. Table 10 tracks students' reasoning skills as they relate to sampling distributions in the active class from the pre-lab questions to the post-lab assessment questions. Post Q1 and Post Q3 matched best conceptually with Pre Q4 and Pre Q5 and were therefore included in Table 10.

	Post Q1 & Post Q3 Both Correct	Post Q1 & Post Q3 One Correct	Post Q1 & Post Q3 Both Incorrect
Pre Q4 & Pre Q5 Both Correct	6	6	5
Pre Q4 & Pre Q5 One Correct	0	4	0
Pre Q4 & Pre Q5 Both Incorrect	0	0	0

Table 10. Distribution of responses to pre-lab and post-lab reasoning assessment questions in the sampling distribution lab

Perhaps the most striking piece of information gained from the Table 10 is that none of the students went into this lab with completely incorrect reasoning about sampling distributions and only four of the students missed one of the pre-lab questions. The rest (17 students) correctly reasoned generally about the idea of a distribution of averages before completion of the lab. Because sampling distributions and the central limit theorem are traditionally difficult to teach successfully, this was a remarkable finding. However, somewhere in the lab or the discussion afterward of the more technical aspects of the Central Limit Theorem, these students in the active class became very confused and lost their correct reasoning or were simply unable to then apply these ideas into an application question.

In analyzing the incorrect responses to the first post-lab assessment question (Post Q1), it became clear that there were two types of incorrect answers. The first type of answer revolved around reasoning that focused solely on the mean and gave no consideration to the idea of variability. These students grasped the idea the center of a sampling distribution tends to be the same as the population mean, but this seemed to be the only concept of the Central Limit Theorem that they took away from the lab. For example:

> They would be the same because the mean stays relatively the same. (CLT08)
>
> No, because it seems that the average stays the same throughout. (CLT09)
>
> They should be about the same because the mean always stays the same. (CLT16)
>
> No because the mean does not change very much between a population & random sample. The sample is just as likely to have people about 470 as the population (CLT07)
>
> No because the averages stay about the same as the sample size increases, so I think the probabilities of these 2 events will be about the same. (CLT18)

Often in this kind of assessment question, where two means are mentioned (the population mean and the particular sample mean), students will confuse the two and substitute the sample mean as the population mean. In this case, that would mean that students used 470 as both the sample mean and the population mean and may account for why students felt the probabilities were the same. Unfortunately, with such short responses to analyze, it is difficult to tell in the responses quoted above whether students confused these numbers.

The second type of wrong answer revolved around a misunderstanding of the connection between a narrower distribution and a change in probabilities. (In the following responses, answer A was the correct response to the assessment question (Post Q1)).

> No, I don't think there is a difference between A & B because a single person will probably score above a 470 and out of a random 35 peoples sample mean will be above a 470. (CLT14)
>
> (a) P (1 person scores > 470)
> (b) P (random 35 sample mean > 470)
> I think the probability could possibly be the same. (CLT12)
>
> Yes because most of the scores lay around 455, a random sample of 35 people will have scores around 455. Probability a single person has a score of 470 is not as high. (CLT02)
>
> B more likely than A because of mean being 455 more luck w/ one person getting it then 35 being above the average. (CLT27)

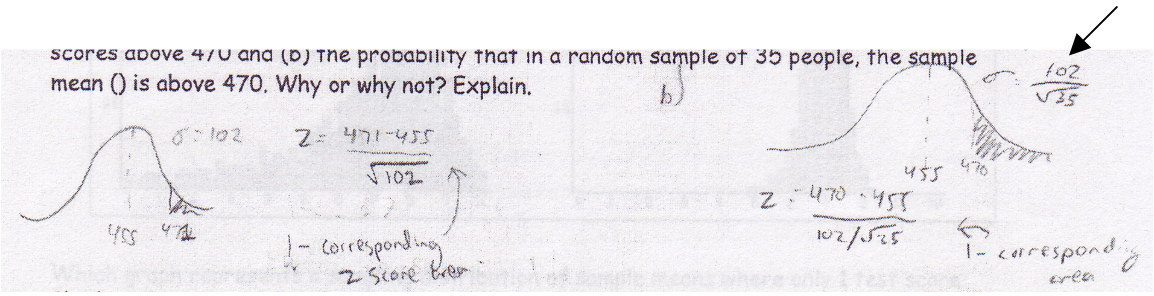
The last student response above (CLT27) appears to be reasoning correctly about the problem, however the student technically gave the wrong answer.

It could be argued many of these students' incorrect statistical reasoning above stemmed from a lack of understanding of what the mean is, i.e. that it is the balancing point of a sample of numbers. However, this was impossible to discern from the fairly short responses to this assessment question (Post Q1). On the other hand, students who answered this question correctly showed a clear conceptual understanding of how a mean differs from an individual data point. For example:

> The probability that a single person scores 470 is greater than the probability that the mean of 35 scores is above 470. In order for the mean

> to be 470, about half of people must score above 470 which is unlikely if the average is 455. (CLT13)
>
> There could be a difference. The single person could be significantly above or below the mean where as the average of 35 people should be pretty close to the mean. (CLT11)

Although calculations were not required to answer the question, by the time students had a chance to answer these assessment questions, there had been a discussion over how to calculate the standard error as a way to numerically measure the variability of a sampling distribution. It is interesting to note that only one student attempted to calculate the standard error in order to answer this question (See Figure 11).



scores above 470 and (b) the probability that in a random sample of 35 people, the sample mean () is above 470. Why or why not? Explain.

σ = 102
Z = 471−455 / √102
455
471
1 − corresponding z score area
b)
σ = 102/√35
455
470
Z = 470 − 455 / 102/√35
1 − corresponding area

Figure 11. Demonstration of the use of the standard error formula (CLT17)

As part of the classroom discussion after the lab, I introduced the formula for standard error ($\frac{\sigma}{\sqrt{n}}$= standard deviation divided by the square root of the sample size) and demonstrated how the real life results showed that this formula worked. The lack of use of the standard error formula in answering Post Q1 seems to indicate is that while students may understand theoretically that a mean of a sample has a higher likelihood to be closer to the population mean rather than one individual data point, students do not necessarily see a mathematical connection between the probability and the standard error. This will be explored further in the student interviews later in this chapter.

Research Lab (Confidence Intervals)

Overview of instruction

The last research lab on confidence intervals was different from the first two as it was not designed to address either standard deviation and standard error directly, but addressed the idea of standard error and margin of error indirectly through the construction and interpretation of confidence intervals. Margin of error is calculated by multiplying the estimated standard error times the number of standard errors away from the mean required to achieve a particular confidence level. Standard error is the “unit” of variability measurement on a distribution of means, margin of error measures the actual distance from the mean one must travel in standard error units to cover a particular percentage of the distribution of means.

In a slight departure from the first two research labs, the lecture class also completed a “lab” in class for the topic of confidence intervals (Appendix F). This lab did not use a conceptual change framework and was merely a guide for how to create a confidence interval for the proportion of green M&M’s. This lab also did not include any statistical reasoning assessment questions. I used the intervals the students constructed in the lab to lead a discussion of how a confidence interval is constructed and how to correctly interpret one. As stated previously, both classes participated in activities and labs that were not part of this research. Because this lab was merely an activity to do in class, I felt that it was appropriate to include in my lecture class.

For the active class, I asked the following pre-lab question pre-lab (CI Pre 4):

> 4) Suppose the following are the results from 10 people with 10 different samples (assume everyone has approximately 70 M&Ms each). Based on these results, would you argue that M&Ms is giving its customers "fewer" green M&Ms. Why or why not?

Sample Percentage ($\hat{p}$)	**95% margin of error**
9.5%	+/- 8.1%
11.5%	+/- 8.8%
12%	+/- 9%
8.9%	+/- 7.8%
8.9%	+/- 7.8%
9.6%	+/- 6.9%
10.6%	+/- 7.2%
14.7%	+/- 8.2%
9.9%	+/- 6.9%
10.4%	+/- 7.2%

In a previous pre-lab question, students had been asked to estimate what percent of M&Ms should be green if M&Ms were equally distributed. All but one student correctly answered that should be about 16.7%, so that did not create any problems with analyzing the responses to question above (CI Pre 4). Then as a follow-up question (CI Post 11) after the lab, I asked the following:

> Look at your answer to number 4 on the preliminary question sheet. Based on what you learned during this lab, would you change your answer from what you said originally? Explain.

Analysis

Based on student responses to this question, this lab appeared to be more successful in accomplishing its learning goals than the previous lab covering the sampling distributions and standard error. Of the 24 students who completed the lab, 15 students answered the follow-up (CI Post 11) completely correct. Seven of the other students unfortunately misunderstood the follow-up question and used the intervals constructed from the lab, not the intervals in the original preliminary question (CI Pre 4), to answer it. Unfortunately, the confidence intervals created by the students using real M&M's during the lab led students to a different conclusion than the table of intervals that I provided the students in the preliminary question. However, even though those students, who had technically answered the question incorrectly, used statistically correct reasoning. Nonetheless, this created some difficulty in analyzing students' reasoning for this lab.

This research lab was also different from the previous two research labs in that I talked a bit more extensively about the concept of point estimates and the margin of error, the two pieces of a confidence interval, before I asked students to complete the preliminary lab questions. So I expected more students would correctly incorporate the margin of error into their reasoning. However, 16 of the 24 students used incorrect reasoning on this preliminary question (CI Pre 4). Several students did not seem to pay any attention to the margin of error, only the sample percentage. For example:

> Yes because all of the 10 samples are less than the predicted 16.7%. 14.7% is the highest and the other 9 are all consistently lower than that. (CI17)
>
> Based on these results I might argue that M&Ms are giving the customers less green candies. I argue this because each sample contains less than 16.67% of green candies, which would be the percentage of green candies if all 6 colors were evenly distributed. (CI03)

However, in retrospect, this was a flawed question. By providing students with ten intervals, all of which had a sample percentage well below 16.7%, it is reasonable for students to argue that the ten intervals taken together do provide evidence that there are less than 16.7% intervals. The question really should have only presented one interval for students to inspect and that would have provided more direct evidence as to whether students were considering margin of error in their responses.

Other students did seem to recognize that the margin of error might need to be considered, but were unwilling to accept that the large margin of error actually made judging whether the greens were less than one-sixth impossible. Again, the flaw in the design of the question makes any analysis of students' reasoning impossible. For example:

> It seems as though M+Ms are giving people less green because $\hat{p}$ is generally less than 16.67, but when you add the margin of error, it still could include 16.67 but it is less likely. (CI08)
>
> Yes, if the average is supposed to be 17%. These numbers, even including the margin of error are low. (CI04)

Other students did not seem to understand how to put the sample percentage together with the margin of error correctly, i.e. the plus and minus part of the interval.

> Yes, because with the margin of error, many results still don't hit the 17%. (CI18)

So although students had been introduced to the concept of a point estimate plus or minus the margin of error through a short lecture, many students may have still been unclear as to the role of margin of error in statistical estimation. And since students were able to go back and re-answer that question after the lab activities were completed and for the most part did so correctly, there is evidence that the concepts and activities of the lab helped students understand the role of margin of error in making a decision. For example:

> Now I might because the margin of error could put all of those numbers to 16.7% (its possible). (CI17)
>
> I would change my answer because I did not take into account the 95% margin of error. 90% results contain the 16.67% mark, so I would no longer argue that M&Ms is dispensing less green candies. (CI03)

However, as I stated previously, there was an issue analyzing some of the responses to the post-lab question (CI Post 11) as students misunderstood what intervals I wanted them to use to answer the question.

> No, the evidence from our samples supports that there are less green M&Ms included than the average (CI04).

This student was still using slightly incorrect reasoning in that while only 60% of the 95% intervals created by the class contained 16.67%, most of the rest of the intervals were actually above the 16.67% mark. Unfortunately, because I did not anticipate students using the data from class to answer the question, it made interpreting the results difficult. This issue, on top of the fact that the initial reasoning question (CI Pre 4) was fundamentally flawed, meant that there was little useful data for analysis from this research lab.

It was not quite as clear whether students understood the concept of margin of error as it related to confidence intervals, since there had been a class discussion about the in-class results and we had discussed specifically the fact that only 60% of the intervals contained 16.67% was unusual if the confidence level was 95%. The students who answered using the in-class data may have just copied down what was discussed in class without any real understanding and it is therefore impossible for me to analyze if these students were truly demonstrating conceptual understanding of the variability concepts.

Interview Overview

A colleague conducted interviews with students from both the active and lecture classes during the eighth week of the quarter. The purpose of the interviews was to have better understanding of students' statistical reasoning skills a few weeks after the topics of standard deviation and standard error had been covered in class. Statistics educators argue that the concepts of variability carry through the entire course (Ben-Zvi & Garfield, 2004; Cobb, 1992; Moore, 1997) and so it is important that students have a good understanding that stays with them through the course and beyond. The student interviews also provided me a chance to compare students' reasoning skills from each of the two classes. Since the research labs were only done in the active class, it was impossible to make any comparisons between students' reasoning using the lab results. The CAOS test only provided me with very superficial information about student's reasoning skills.

The interviews provided some of the most interesting results. The questions were designed to target three specific concepts: standard deviation, standard error, and the difference between these two measurements of variability. The questions were drawn from a database that specifically designs questions to test statistical reasoning skills. Students had seen these types of questions on their in-class labs and on tests. The statistical reasoning interview questions were as follows:

> Q1: Suppose two distributions have exactly the same mean and standard deviation. Then the two distributions have to look exactly alike. Explain whether this is true or false.
>
> Q2: Shelly is going to flip a coin 50 times and record the percentage of heads she gets. Her friend Diane is going to flip a coin 10 times and record the percentage of heads she gets. Which person is more likely to get 20% or fewer heads?
>
> Q3: The distribution of Verbal ACT scores is normal with mean 21 and standard deviation of 5. Which would be more likely:
> A) a single person scores less than a 16
> B) the average score of 25 students is greater than a 22
> Explain your choice.

Question one (Q1) was designed to measure students' conceptual understanding of standard deviation (SD), question two (Q2) was designed to measure students' conceptual understanding of standard error (SE), and question three (Q3) was designed to measure students' understanding of the difference between standard deviation and standard error (SD vs. SE).

Table 10 provides a summary of the responses by each interview subject to each of these interview questions. Again, ACTIVE or LECTURE refers to which class section for each interview subject and the number refers to the order of the interviews. Some numbers are missing because the student scheduled in that interview slot failed to show up for the interview.

	Interview Questions		
Interviewee #	Q1 (SD)	Q2 (SE)	Q3 (SD vs. SE)
ACTIVE02	INCORRECT	INCORRECT	INCORRECT
ACTIVE06	CORRECT	CORRECT	INCORRECT
ACTIVE07	CORRECT	INCORRECT	INCORRECT
ACTIVE10	CORRECT	CORRECT	INCORRECT
LECTURE01	CORRECT	INCORRECT	INCORRECT
LECTURE04	CORRECT	INCORRECT	INCORRECT
LECTURE08	INCORRECT	CORRECT	INCORRECT
LECTURE09	INCORRECT	CORRECT	INCORRECT

Table 11. Overview of the interview results

An initial glance at the Table 11 seems to indicate that students did not understand these measures of variability. Only half answered questions one and two correctly and no one answered question three correctly. Table 11, however, does not tell the whole story. Even students who technically answered an interview question correctly, in many cases, used shaky, incorrect or incomplete reasoning. This was a bit surprising initially, given the fact that some of these students had answered similar questions correctly at the end of each research lab. Because of this, I, as the teacher, was fairly confident that the students (in the active section at least) were grasping the concepts. However, the results on the third interview question were confirmed by the results from the CAOS posttest (see Table 7). It was very clear from both the interviews and the results of the CAOS test that students did not understand the differences between standard deviation and standard error, and only had incomplete understanding of each measure on its own.

The third interview question directly addressed the difference between standard deviation and standard error. After initial analysis of the transcriptions, the responses to the third interview question were the ones that were most interesting for my research questions. The first two questions then gave me some insight into perhaps why students had so much trouble with the last question.

Lack of Confidence in Responses

Students were very unsure of their answers during the interviews. Several students changed their answer several times as they reasoned through the problem, so much so that the interviewer had to ask several different students "Is that your final answer?"

> (Q1) False…Um…if the mean and the standard deviation are the same, then the graphs will appear to be the same, but you could have very different data and a very different problem (trails off)…um true they would look alike, but that doesn't mean they would be the same. (LECTURE04)

> (Q2) You know I am going to take a guess and say Diane….I am just looking at it and it is a smaller number and I am thinking that since there are less flips….actually…..20 percent…you know, I don't know if this is a trick question but I think I want to say that because the flipping of coins is so random that it could be I think that the possibility is the same for both. (ACTIVE02)

Other students simply prefaced or ended their answers with "I know this is wrong".

> (Q3) I really think that they are both equally likely to happen? One person getting…it is kind of like flipping a coin…one person getting above…well…but its below 16, it is not quite above 21 or below 21…. the way I viewed it originally is how one person below 16 and above 16 and the average…that doesn't make sense. Ok honestly I am going to say it is more likely for B (ACTIVE07).

And there was no correlation between correctness of their reasoning and their confidence in their answer. A few students very confidently gave an incorrect answer to a question.

> (Q2) I'd say it is about the same because it is a 50/50 chance and the trials are independent of one another. (LECTURE01)

This was surprising since these were generally students who should be fairly confident in a mathematics class. Almost all of them had a very extensive mathematics background that included calculus, multivariate calculus and linear algebra. Only one student (ACTIVE02) labeled herself as someone who "struggled with math". This illustrates rather clearly the differences between the study of mathematics and the study of statistics and that success in mathematics is not necessarily a predictor of success in statistics.

What is good about this demonstration of lack of confidence is that most students seemed to be aware of how shaky their understanding was of the concepts. That is an indication that this new statistical knowledge was causing cognitive conflict and they were still trying to construct their understanding. However, a confident, incorrect answer was more of a concern, as those students are indicating they were not struggling with the concepts and they had developed incorrect reasoning.

Question Three (Standard Deviation vs. Standard Error)

To correctly answer the last interview question, I expected that students would need to calculate the standard error for the average of 25 students. However, I intentionally made the calculations very easy so students could complete the calculations mentally. (The standard error was $\frac{5}{\sqrt{25}}$ = 1). I expected that students would hopefully then recognize that both scenarios were exactly one standard deviation away from the distribution's center and therefore both scenarios were equally likely. Either they would recognize this after calculating the standard error or some students might go as far as to calculate the probabilities using the normal table. I worded this interview question this way specifically so that students wouldn't be able to guess the correct answer and that would hopefully provide me more insight into their reasoning process.

Only two of the eight interviewees, however, used any sort of a calculation approach to this question and only one of those two used the standard error calculation.

> Student: I used the normalpdf function, actually it's the normalcdf function and I put in the lower bound as negative infinity and the upper bound as 15 with a mean of 21 and a standard deviation of 5.
> Interviewer: Ok what does it say?
> Student: Hold on one sec, pushed the wrong button (pause) says .11506 so we are going to call about an 11.5% chance and then the average of 25 people scoring greater than 22?
> Interviewer: right.
> Student: Hmm, so I'm going to need use average of the averages, which is actually what we are learning in class right now. So (pause) the mean is still going to be about 21 and the standard deviation would be 5 divided by the square root of 25 people (pause) so the standard deviation of that would be one, so then we could use again the normalcdf function with a minimum boundary of 22, an upper boundary of infinity (pause) the mean is 21 and the standard deviation is one. And you get about about a 15.9% chance.
> Interviewer: Ok so a 15.9% chance.
> Student: So I am going to say it is more likely that an average of 25 people score above 22 than one person scores less than 16. (LECTURE01)

This student (LECTURE01) was the only interviewee that came close to a correct answer on interview question three. One can see in the transcript that she actually did correctly calculate the standard error. Her only mistake was in using 15 as her lower bound in the first part instead of 16. Students often confuse "less than 16" with "less than or equal to 16" on a continuous probability distribution. On a continuous normal distribution, they are essentially equivalent. This is why she most likely chose to use 15 in her calculations. It is interesting that on the second portion of the problem she did not make the same mistake. Even later when she recapped her answer, she still did not see the inconsistency. However, her reliance on the calculator's numbers may demonstrate that although she did seem to understand the process of the problem, there isn't much evidence of statistical reasoning in her explanation. Other students attempted to use a calculation approach but were less successful. For example:

> Well, empirical wise, for one standard deviation you are going to have about 68% of the population, falling one standard deviation outside the 21, which was the mean score. Sixteen is exactly one standard deviation away, which means that below 16 percent, er below 16, we are looking at about 16% of the population left. So, I mean, it is a lower proportion, but it is certainly pretty likely. Otherwise, if you had 25 people's average greater than 22, then it is within that 68% mark. So it's not goin' to be quite 34; it is going a little under 34 so it is going to be about (pause) maybe 46 and 50% likely for people and you have to have 25 of those people. So (pause), the percentage is in the favor of having a greater score but it has to have more people for it, it has to have 25 people whereas only a single person has to score below the 16, so I am actually going t say it is more likely for the single person to score below 16. It is the more likely scenario than the twenty-five people scoring as an average over 22. (LECTURE08)

This student (LECTURE08) demonstrated a very solid grasp of calculations on the normal distribution and recognized that one cannot compare directly a single score to the average of twenty five scores. However, the student did not then make the connection to the need for standard error and the Central Limit Theorem.

Most interviewees simply tried to come up with reasoning based on either the sample size effect: one person versus 25 people, or based on the distance of the numbers from the mean: 16 is further away from the mean than 22 is, without doing much in the way of actual calculations. And these arguments were made to support both scenarios A and B as more likely.

> I think option B…because I think if the, if we are saying the mean is 21, which is the same as saying it's the average, um I think the likelihood that

> picking a single person and will get a score of 15 is going to be less likely (pause) than if we were take twenty five students and average their scores out …(interviewer repeats first part of question)… less than a 16 than a 15? ok I am still going to stay with my answer. (ACTIVE02)

> I going to go with B, the average of 25 is greater, because the standard deviation is five and the mean is 21, so therefore 22 isn't really far from the mean, but to get less than 16 that's more than a whole standard deviation away. (LECTURE04)

> I would say A, that you have a single person might get less than 16, because if you were to go one standard deviation away from the mean, well if you go two standard deviations away from the mean, 16 falls in that range. So therefore, and two standard deviations away is not an unlikely probability, so its likely at least one person will have a score of 16, whereas for the average for 25 people to be greater than 22 to get average that's much higher would be less expected than just getting a single score. (ACTIVE06)

Only one student considered that the two probabilities would be equal, which was the correct answer. However, her initial reasoning as to why the probabilities were equal did not make any statistical sense. Which is perhaps why this student talked herself out of her original answer.

> I really think that they are both equally likely to happen. This is my reason; it is so awkward. One person getting, it is kind of like flipping a coin, one person getting above, well (pause) if it is below 16, it is not quite below 21 or above 21 (pause) I keep getting the same number. Um, the way I viewed it originally is how one person below 16 and above 16 and then the average (pause) that don't make sense. (pause) Ok, honestly I am going to say it is more likely for an average of 25, B, greater than 22. Only because 22 is closer to the mean than 16 is, so in order to get 16 or below, that would probably be like a 5% chance, I don't know. Where 25 people is an average so they could have one below 16, one above 28 and they would average around the mean and the mean would 21, so being above 22 would be, just that it would be higher. It is more likely for you to have an average above 22 of 25 people. (ACTIVE07)

It was difficult to parse out exactly why she (ACTIVE07) talked herself out of her original answer. She mentioned she kept getting the "same number" and that seemed to perplex her. Perhaps she felt that she couldn't possibly get the same probability for both scenarios, and perhaps that was a flaw in the interview question. Her explanation for then picking scenario B was much more difficult to follow. It appeared that she resorted to the argument that many interviewees used: 22 was closer to the mean and therefore more likely.

Making Connections to Graphical Displays of Data Distributions

One of the main concerns of statistics education researchers is looking at how students connect the concepts of variability to the distributions of data. Garfield et al. (1999) did extensive work looking at sampling distributions and how students made the connections between graphical displays and sampling distributions. The first interview question (Q1) was designed to discern how well students conceptually understood the standard deviation and how well they understood its connection to graphical displays of data distributions. The students who answered this question correctly actually had a very rich understanding of how the standard deviation connected to the graphical displays, even though their explanation of their reasoning was a bit uneven at times. The student below (LECTURE04) understood that the data could be different but have the same mean and the same standard deviation. However, she seemed convinced that the graphs would have to look similar.

> False…I guess like the graphs will look alike but that doesn't mean that the data was alike.…with the mean and the standard deviation are the same then the graphs will appear to be the same, but you could have very different data and a very different problem. (LECTURE04)

The following student (ACTIVE06) had a better grasp on the idea of average distance away from the mean and pointed out that the extremes could be different if they were balanced out by data at the other end of the scale. The only part she seemed to lose track of was the fact that the mean might change if you shift the extremes from low end to the upper end of the scale.

> False…I would say false, because your distribution includes all of your sample, all of your data and you could have data that is reasonable similar, enough to give you the same mean and the deviation, but you might have different extremes. You may go a little bit more, farther out on one end than the other, but still have enough high numbers or low numbers to even it out. (ACTIVE06)

The following student (ACTIVE07) had the best conceptual understanding of the entire question and all the pieces, the mean, the standard deviation and the relationship to the distributions of data.

> False…I mean the mean is just the average of where it is…you can have like the same mean, where it has the same number, say this is 2 and this is also 2, and that the average standard deviation away could be two points away….no I am thinking….the standard deviation for a right or left skewed one might actually be (pause) well I don’t know because if this one has an average of two standard deviations away this one could have less here and but more here, which could balance it out….if the bell shape could be wider, where the right skewed could be shorter, but It could still have the same standard deviation, same average length away (ACTIVE07).

The students that missed this question (Q1) generally just accepted that if the mean and the standard deviation were the same then based on the definitions of the two measures, the distributions must be the same.

> If the mean and the standard deviation is the same, I think the two would look the same because the mean for both of them will fall in the same spot on the graph and the standard deviation is basically what tells us or regulates the width of the graph. So if it is the same for both, then the graphs would look the same. (ACTIVE 02)

> It’s true because they have the same mean so the…like in the graph, the middle bar would be in the same place, at the same height, and the same standard deviation means that the next bars are going to be same distance from the middle bar. (LECTURE 09)

Both of these students demonstrated a basic understanding of the mean as the center and the standard deviation as a measure of the range of the data, but actually seemed to lack a complete understanding of a distribution. Reading and Reid (2006) noted in their research that understanding of distribution and variability are inextricably linked. One student, who answered this interview question (Q1) incorrectly, did seem to have a better understanding of distribution but ultimately still chose the wrong answer.

> The mean obviously is going to have to take place in the same spot…although I guess the distributions don’t necessarily have to look the same. It could be in proportion, I imagine (long pause)…well, I believe that is in fact a true statement. (LECTURE 08)

Comparison of interview responses of active class versus lecture class

There were two main purposes of the interviews. The first was to analyze how students verbalized their understanding of the measures of variability and just how well students integrated all the dimensions of variability. The previous analyses focused on this purpose. The second was to compare and contrast responses from the interview subjects in the active class with the interview subjects from lecture class. Table 9, earlier in this chapter, is sorted by section and illustrates how each group of interviewees responded to each interview question.

The only minor difference between the two groups of interview subjects was in the responses to the first interview question (Q1). In general, the students from the active section had a much richer description of the connection between the standard deviation and a data distribution. The three active interview subjects who answered this question correctly discussed how the shape of the distribution affects the size of the standard deviation.

> You may go a little bit more, farther out on one end than the other, but still have enough high numbers or low numbers to even it out. (ACTIVE06)
>
> If the bell shape could be wider, where the right skewed could be shorter, but it could still have the same standard deviation, same average length away. (ACTIVE07)
>
> Theoretically, you could have like a graph that was normal and a graph that was like bimodal, and there would still be, all of the data could still be about the same distance away from the mean and the mean could still be the same. (ACTIVE10)

The responses from the lecture interview subjects were much less rich in their description of the connection to the data display.

> I guess like the graphs will look alike but that doesn't mean that the data was alike (LECTURE 04)
>
> The distributions don't necessarily have to look the same. It could be in proportion, I imagine. (LECTURE08)
>
> The middle bar would be in the same place at the same height and the same standard deviation means that next bars are going to be the same distance from the middle bar. (LECTURE09)

However this difference in rich description was not entirely split along class sections. One active interview subject (ACTIVE02) gave a very short, non-detailed answer to this question, while a lecture interview subject (LECTURE01) had very detailed description of the connection of the standard deviation to the distribution.

> The mean would fall in the same spot on the graph, and the standard deviation is what, is basically what tells us or regulates the width of the

> graph. So if it is the same for both, I think that both graphs would look the same. (ACTIVE02)
>
> The same amount of data will be within the same width on both problems….so I wouldn't say that they would have to be, they wouldn't have to look exactly the same because some groups of data have more data points than others and so there might be some fluctuation in between, you basically a bell shape curve but there might be one bar that is lower one bar that is higher or whatever, but basically they are going to have very similar looks between the two of them. (LECTURE01)

Otherwise, the types of statistical reasoning errors and the different approaches utilized seemed to be the same for both the active and the lecture classes. There were no other outstanding differences between the interview subjects from the active class versus interview subjects from the lecture class that I could discern from the interview transcripts.

Understanding Standard Error

Finally, the analysis of the second interview question has been left to the end because I did not feel that responses provided much data that was useful for answering my research questions. The purpose of the second question of the interview was to determine how well students understood the concept of a sampling distribution and that the standard error varies as the sample size is changed.

> Q2: Shelly is going to flip a coin 50 times and record the percentage of heads she gets. Her friend Diane is going to flip a coin 10 times and record the percentage of heads she gets. Which person is more likely to get 20% or fewer heads?

I originally chose to use proportions instead of means for this interview question because I wanted to vary the questions a little and not focus exclusively on standard errora as it applies to the mean. I also felt that a question about the population mean might be too easy and the responses would not yield me enough data. Both classes had also discussed a sampling distribution for the population proportion and a population sum in class. In hindsight, this added a dimension to the question that I never intended. My intended purpose of the question was to see if a student would realize that as the sample size was increased that the reasonable possibilities of the proportion of expected heads would narrow around 50% and therefore it was more likely that the smaller number of tosses would result in 20% or fewer heads. Although the students never needed to calculate the standard error to answer this question, I felt this was a good question to assess if they understood the concept of standard error and that standard error varies with the size of the sample taken. Unfortunately, students did not respond as I had expected. Instead, most of the interview subjects focused on the fact that it was a series of coin flips and therefore a question of independent trials.

The interview subjects who responded incorrectly to this question all said that the probability was the same for Shelley and Diane because coin tosses are independent and therefore the probability is 50/50 on each toss.

> They have the same chance getting the amount of heads…cause it is always 50-50 chance, regardless of how many times they do it. (LECTURE04)
>
> They have the same percentage…they are both equally likely. I mean the chance is 50-50 so there is no real person more likely than the other. It is 50-50 both ways, they will have the same percentage. (ACTIVE07)
>
> Because they both have the same chance of getting 20% or fewer heads just because it is an independent trials situation (LECTURE01)

Even though these students used an independence argument to argue that both scenarios were equally likely, these responses also demonstrate a lack of conceptual understanding of actual probabilities of an individual versus a sample. In the other questions from the interviews and in the labs that dealt with probabilities with a sample (see interview question 3 (Q3) and first extension question of sampling distribution lab (Post Q1) in Appendix D), students often had trouble distinguishing the probabilities of individuals versus the probabilities of a samples, an issue that goes to the core of understanding sampling distributions.

The interview subjects who responded correctly to this question also all generally used the same reasoning: the law of large numbers argument that as sample size increases, the closer one should get to the expected 50% mark.

> I would say Diane because she is flipping fewer coins and the more times you repeat a experiment, the more likely that you are to get to the expected average. For like flipping a coin you expect to get heads 50 percent of the time, so if you flip a lot more coins, you are more likely to have an average, to get a probability of one-half. Diane is only flipping ten. There is going to be less data figured in which could change her probability if she just happens to flip a lot of tails. (ACTIVE06)

> Theoretically, they both should get about 50 percent, but the more trials you do, the closer you get to the true mean. So I would say probably the person who had 10 because if you do it fifty times, you are more likely to get closer to 50 percent, which is what you should get in a fair coin toss. (ACTIVE10)

These two interviewees (ACTIVE06 and ACTIVE10) also demonstrated more of an understanding of the expected probability versus actual probability. That, yes, in theory, both Shelley and Diane would flip half head and half tails, but that in reality is sample results never quite model the population. One could argue that the law of large numbers argument indirectly demonstrates that those students understand the idea of a narrowing sampling distribution, whether that connection is clear to students is impossible to discern from the data. Unfortunately, the incorrect responses did not necessarily reveal whether students misunderstood standard error itself or just didn't understand that one had to look at the flips as a collection and not individually. Students heard the example of flipping a coin and went straight to the idea of independence and that coins should flip 50 percent heads.

Summary of the Interview Analysis

The interview responses were the most interesting data I collected. Each interview question measured a separate component of student's understanding of variability. From each separate analysis there were some major themes that emerged.

Interview Question 1 (Q1)

The major theme that emerged from the responses to this question was students' understanding of the connection between notions of distribution and the concept of standard deviation. The three interviewees that missed this question demonstrated poor understanding of distribution. The five interviewees who got this question correct showed a much more complex understanding of distribution, i.e. they understood that the height of the bars of a distribution gave a measurement of how much data was in that location. This trend was also reflected in the responses to the third assessment question (Figure 9) in the standard deviation research lab.

Interview Question 2 (Q2)

Although it was not the information I was expecting to gather from this interview question, the major theme that emerged from the responses to this question was students' inability to distinguish the probability of an individual from the probability of a group or a sample as a whole. The four interviewees who answered this question correctly appeared able to make that distinction, although I am concluding that using indirect evidence. This theme recurred in the sampling distribution research lab and the third interview question (Q3) as well.

Interview Question 3 (Q3)

Two major themes appeared in the analysis of the third interview question, which all interviewees got incorrect (although one student reasoned correctly but made a minor mistake in the calculations). One was the inability or the refusal to distinguish the probability of an individual from the probability of an average of a sample, as I stated above. The second was that there was no evidence of students recognizing the necessity of the standard error. However, in most cases, students did recognize that one number was significantly further away from the mean than the other. This at least indicates that students are considering distance from the mean in reasoning about the probabilities. They recognize the importance of considering variability, even if they are not using the proper unit of measurement: standard error.

Connections between the interviews, research labs and CAOS

To summarize the data for the eight interviewees, the Table 11 (active class) and Table 12 (lecture class) records their responses on the interview questions, the post-lab assessment questions in the research labs (active class only) and their percent change on the CAOS test from pretest to posttest.

Interviewee	Number of Interview Questions Correct (out of 3)	Research Lab Standard Deviation Questions Correct (out of 2)	Research Lab Sampling Distributions Questions Correct (out of 2)	Research Lab Confidence Intervals Questions Correct (out of 1)	CAOS % Improvement
ACTIVE02	0	1	1	1	-12.5
ACTIVE06	2	2	0	0	0
ACTIVE07	1	1	0	1	-7.5
ACTIVE10	2	2	1	1	12.5

Table 12: Summary of the active data on all statistical reasoning assessment items

Interviewee	Interview Questions Correct (out of 3)	CAOS % Improvement
LECTURE01	1	+7.5
LECTURE04	1	+2.5
LECTURE08	1	-10
LECTURE09	1	0

Table 13: Summary of the lecture data on all statistical reasoning assessment items

After considering all of the data I collected, there were two important conjectures that I hypothesized from across all my data. First, my data showed that a complete understanding of distribution was very important and intertwined with a complete understanding of standard deviation and variability in general. The five interviewees that got the first interview question correct demonstrated a very complete understanding of the different dimensions of distribution, while the three interviewees who got the first interview question incorrect did not. This same theme emerged from the responses to the assessment questions in the standard deviation research lab.

The second conjecture that emerged was students not recognizing the difference between the probability associated with an individual data point versus the probability associated with the entire sample or the average of the sample. This persisted throughout the responses to both the second and third interview questions and throughout the responses to the assessment questions in the sampling distribution lab. This may indicate that the problem with students being able to recognize when to use standard error in a Central Limit Theorem problem is not an issue with recognizing a sampling distribution and that it narrows with a larger sample size, but an issue with understanding how probabilities change on a sampling distribution. This is supported by the fact that so many students answered the pre-lab questions sampling distribution research lab correctly. Both these conjectures will be discussed in greater detail in chapter 5.

Chapter Summary

The interview responses provided the most interesting data to analyze. Students' explanations of their statistical reasoning revealed a lack of conceptual understanding sampling distributions and standard error that was surprising. Except for the first question about standard deviation, where the active class showed a slightly better understanding of the connection between standard deviation and the notion of distribution, there were no differences between the lecture class interview subjects and the active class interview subjects. This seems to correlate with the findings on the CAOS test and the lab responses. Essentially, students in the active class came away with a much richer understanding of the how standard deviation works and the idea of variability in the data than the lecture class. However, students in both classes remain very confused about standard error, how it relates to a sampling distribution, how that is different from a data distribution and when it is appropriate to use each measure. What follows in the next is a discussion of my conclusions, the limitations of the study and suggestions for further research.

CHAPTER 5

CONCLUSIONS, LIMITATIONS AND IMPLICATIONS

Overview of Conclusions

Over the course of this research, I collected a significant amount of information and data. So much so, that at times it was hard to make any sense of it all. However, if I were to sum up what I have concluded from my research in an extremely condensed version, this is what I would say:

> 1) The predict/test/evaluate lab model, as I interpreted it, worked very well for teaching the concepts of standard deviation and margin of error in a confidence interval.
>
> 2) The predict/test/evaluate lab model, as I interpreted it, did NOT work well for teaching the concept and importance of standard error and how it differs from standard deviation.

Given the statements above, the next question is invariably why did this work so well on the standard deviation and the margin of error topics, but not on the standard error topics? The next part of this chapter will be devoted to trying to answer that question.

Limitations of the study

General Design Concerns

There were many issues and flaws that have to be discussed about this particular research. The most obvious issue, that is true of so many educational studies that employ a “treatment” and a “control” class, is that my two classes were not exactly the same in structure and format even though I tried to keep them as similar as possible. Each class had its own culture. The lecture class was much quieter, even during activities, while the active class was a friendlier, more talkative group that seemed more willing to ask questions. Having acknowledged these differences, I would argue that the two classrooms were as similar as I could hope to make them. I worked very hard to keep the methods of presentation the same for everything except the topics related to this research. Certainly assignments, homework, and assessments were the same, except for, again, the research labs related to the research.

Both classes were relatively small. Both classes started with approximately thirty students. By the end of the quarter, the lecture class had twenty-seven students and the active class had twenty-six students left. The small sample sizes in each class bring into question the strength of the conclusions of this research, especially since the differences between the two classes on the CAOS test were so small. Ideally, I would like to try these teaching techniques with larger classes or with a larger number of sections. In addition, because two students did not show up for the interviews, I was only able to conduct interviews with four students in each section. Again ideally, I would have liked to interview as many of the students as I could in each section (as many students as were willing). Because the interview protocol was relatively short, this would have simply given me more data. As I analyzed the interview transcripts, I found myself wanting more data from more students. But unfortunately, it simply came down to logistics. Only about fifteen students volunteered for interviews. Since I had asked a colleague to conduct the interviews for me and he was not an instructor at this institution, time availability was a significant factor in how many interviews could be scheduled. In the end, only ten interviews total were scheduled and only eight of those showed up.

The College Effect

Certainly, a major concern of this research is the transferability of the results to other institutions that offer similar introductory courses. The college where this study took place is a small, unique institution that attracts students mainly from Ohio. Because they are such a small college, they only offer one introductory class in statistics. By contrast, The Ohio State University, where I am graduate student, offers at least four different introductory courses that are specifically targeted to groups of majors (for example one introductory course for business students, another for engineering students, etc). As can be seen in the demographics of the two classes (Figure 5 in chapter 4), it is not uncommon to have sports management, business, nursing, psychology, science and mathematics majors all in the same class. With that variety of majors comes a variety of mathematics backgrounds and skill levels. It is a challenge for a statistics instructor at this college to find a balance in the focus of the class, so that all students are challenged but do not feel as though the mathematics requirements are too far beyond their capabilities. The course also has to have a certain level of mathematical rigor as it is currently a course that counts towards a major or a minor in mathematics.

There are several benchmark institutions in the Midwest and beyond that this college measures itself against. It was instructive to search these schools' programs to see what kind of introductory statistics courses they offer. The offerings vary quite a bit. Institutions, such as Capital University and Drury University, offer a course similar to this course at this college. Capital University states in its bulletin that its course has applications that emphasize the behavioral, biological, and management sciences. This covers a wide variety of majors in the same way that the course at this institution does. Other benchmark institutions, such as North Central College offers only an upper level introductory statistics course and then offer a quantitative reasoning course that covers several mathematical topics, including topics in introductory statistics and probability.

By looking at the programs at benchmark institutions, it becomes clear that these issues of teaching an introductory statistics course to an audience that has diverse mathematical backgrounds and skills is not limited to this particular college. Therefore it is important to consider these special issues in statistics instruction at a small, liberal arts types of colleges. How does one as the instructor meet the needs of all of its students in one introductory statistics course? I chose to approach topics from different perspectives and with different pedagogical methods. For example, I lectured about the mathematical theory behind a p-value, but I also conducted an activity where students carried out a hypothesis test and saw a p-value in action. I also challenged them, within this activity, to think about what happens to a p-value when certain characteristics of an experiment are changed, thus bringing some of the mathematical theory into this example application. When so many different perspectives of one topic are discussed, time management and assessment becomes a challenge for instructor. What should I expect all students to know? Is it appropriate for a sports management major to know how to calculate a p-value from the data by hand or is it more important that for this student to focus on the applications and implications of a p-value and leave the calculation part to a computer program or a graphing calculator? I would argue that it is more important for this kind of student to understand p-value conceptually and not have to worry about the calculation part. However, others in this particular Mathematics department and beyond, feel that it is important that students know how to calculate these statistics by hand and understand how these statistical formulas are derived mathematically. In my opinion, this leads students to concentrate on issues related to computation and not statistical reasoning. Students will be more likely to focus on learning the calculations, rather than the reasoning component, because learning to do the calculations is more straightforward and has a "right" answer at the end. Conceptual questions require much more intensive thinking and there may not be a right answer to find in the end.

Instructor Effect

Certainly another major effect to be discussed within the context of this study is the effect of me as the instructor. Being both the researcher and the instructor in this research has its pros and cons. I had an insider's perspective to my classroom culture that only an instructor who is there for every class can have. I was very familiar with all of these students and had even taught a few in previous classes. However, being the instructor may have colored my perspective as a researcher. I was certainly very much aware of interpreting too much from a student's response to a lab question during the data analysis phase; for example, thinking to myself, “oh this is what that student really meant.”

I had originally intended to analyze my data with a rubric based on Chance, delMas, and Garfield (2004) levels of statistical reasoning (see Table 1). However, using this rubric meant I had to make decisions about which category was appropriate for a particular student’s response. I found it nearly impossible to be objective and I found that I simply did not have enough student data to classify students’ thinking neatly into one level or another. Had I actually used this rubric, I felt any conclusions I would have drawn based on these categorizations would have been completely untrustworthy and unscientific.

So instead, I chose to let the data reveal categories and themes. There is still the issue of my instructor biases affecting what I interpreted from the data and what I felt was important. But I think by letting the categories and patterns come directly from the data, the transferability of my conclusions was strengthened. Any other researcher could look at my interview transcripts and I believe, see similar trends. I think if I forced the responses into preset categories, there would a lot of room for argument as to why one response was categorized in a particular way. In the end, it would come down to my opinion and I was not comfortable with that. Ideally, if I were to try to categorize my data using a rubric, I would want an expert outsider’s perspective (such as a fellow statistics education graduate student) in addition to my own, to provide a check to my instructor biases.

I believe it was an excellent choice to have a colleague conduct the student interviews for me. Hopefully, students felt comfortable enough to be honest in their answers. I think if I had conducted the interviews, I would have been too tempted to lead students to the correct response.

Design of the Sampling Distribution Lab

After looking at how all the data related to how students conceptually understood standard error and seeing how little my students understood, my first consideration was to go back and evaluate how the topic of sampling distributions and the central limit theorem were presented to the active section. This particular research lab that involved the sampling distributions (see Appendix D) was one that I had adapted from an activity that was published in *Activity Based Statistics.* The original activity in this book starts with students creating a distribution of the age of a population of pennies. Students are asked to sample five pennies on their own and place the age of each of their pennies on a class histogram. This provides the class with the distribution of ages of the population of pennies. Students are then asked to find the average age of their five pennies and again combine their results with others from the class on a community histogram, but this time only plotting the average. They are then asked to do the same thing with samples of size ten and then size twenty-five. I decided it would better for students to construct a sampling distribution entirely on their own (instead of combining class data) from their own original handful of pennies. Therefore, the samples would be coming from their own individual population of pennies and the results would more likely reflect the concepts of the central limit theorem. The drawback to doing this was that it was a bit time consuming. Finding the average age of a sample of five pennies for fifteen samples (they needed enough averages to create a reasonable distribution) took nearly half an hour and I only had an hour and 20 minutes for the entire lab. However, I felt that it was important that students spend the time constructing a distribution of means as part of the conceptual change framework of the lab. Seeing their own data either confirm or disprove their original conceptions was important.

However, it would have been simply too time consuming to ask students to repeat this process for a sample of thirty pennies at a time, so I made the decision to provide data for students to use. The major drawback was that the mean of sampling distribution data that I provided may not have matched exactly to the mean that students had calculated from each of their own populations. One of the essential components of central limit theorem is that the average of means calculated should be very close to the original population mean, and part of the design of the lab was to have students discover that on their own. I felt this was accomplished by having the students create the sampling distribution for the sample mean for the samples of size five and it was unnecessary to have students repeat this process again for samples of size 30.

It was a conscious choice on my part to switch from their population to my theoretical population for the sampling distribution of size 30, but it may have confused students. I have used different versions of this lab in previous years and have always struggled with how to have each student create a sampling distribution for a large sample size that would directly be drawn from their population of pennies. The obvious answer would be to use simulation software. Minitab, the statistical software package provided in the college computer laboratories, has a feature that allows students to enter data and then select samples of any size from that data. However, there is not a computer lab at this college large enough to hold an entire statistics class. So, at best, students would have to team up on computers. In my experience, this usually means someone is working the computer, while the other is writing down answers. In terms of conceptual change theory, having each student confront their conceptions with their own data is important. Sharing a computer takes away from this process. A secondary issue is that students would have to take a much larger initial selection of pennies to create their population, so that selecting a sample of size thirty would not essentially entail selecting nearly all of the population. If the sample is a high percentage of the population, then the central limit theorem does not work as well. So there would still be a fairly intensive time commitment to create a large enough population of pennies in Minitab so the statistics would work correctly.

The next possible solution would be to use the technology that students have at hand to perform the simulations: graphing calculators. I have tried this in previous quarters and the results have not worked well. The TI-84 does not have a feature (at least that I am aware of) that allows a student to input population data and then sample from that population. It only allows randomly sampling for a uniform, normal or a binomial population. This is nowhere near the usual population distribution of penny ages (generally skewed). In previous quarters, I have attempted to use the uniform distribution sampling feature of the calculator to approximate sampling from a skewed distribution, but as one might expect, the average of the sampled means were always significantly off from the population mean. This defeats the purpose of the lab in the first place: having students discover the Central Limit Theorem through their data. In addition, I have found using the random data generators in the calculators is not intuitive (it requires a little programming knowledge) and it takes time for students to learn how to use it. So while the use of technology seems to be the obvious solution, my previous experience has shown that there is still a time investment required no matter what method of instruction is used to teach sampling distributions in this particular educational setting.

Even though I wrapped up this lab with a discussion and visual demonstration using an online java applet (http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html), it is clear from the CAOS results and the interview results: students were confused. This certainly brings into question the effectiveness of the predict/test/evaluate model, as I interpreted it in this lab. I realized after the quarter was complete, that both the standard deviation research lab and the confidence interval research lab included a checkpoint in the middle of the lab that I think ended up being very important to the success of the conceptual change framework. The sampling distribution research lab did not include a checkpoint. In the standard deviation lab, students had to complete the “pick which standard deviation is larger” section (part b, see Appendix C) and check their answers with me before they could move on to the data analysis section of the lab. If they got them wrong, they had to go back and fix their answers before I would give them the data. In the confidence interval lab, students had to stop and wait to collect the confidence interval data from their classmates before they could continue. I think that checkpoint was essential in forcing students to stop and consider their understanding of the concepts during the lab. So much time was spent on constructing the sampling distributions during the sampling distribution labs and not enough time on the evaluation and feedback. And the idea of a checkpoint, where a student must confront their misconceptions, fits the intended conceptual change framework of the labs better.

The Second Interview Question (Standard Error)

As I stated in the data analysis of the responses to the second interview question (Q2) in chapter 4, I was not expecting that using a question that had population proportion versus a population mean would add another dimension to student responses when I included it. My initial thinking was that I wanted to make the standard error question a little more difficult by talking about proportions rather than means. I had spent a good deal of time in class discussing how standard error varied with sample size when it was applied to a distribution of means and less time discussing how it applied to a sampling distribution of proportions. So I believed this question would require students to transfer what they had learned about the standard error varying with sample size to a less familiar parameter. However, I did not anticipate that discussing coin flips would lead students back to the ideas of independence and probability. I had used the example of coin flips to demonstrate the idea of independence in class several weeks before.

In hindsight, perhaps it would have been better for answering my research questions to have chosen an interview question that related standard error to means. For example,

> The distribution of Math SAT scores is Normal with mean 455. Would you expect to find a difference between (a) the probability that in a random sample of 10 people from this population, the sample mean is above 470 or (b) the probability that in a random sample of 40 people from this population, the sample mean is below 440. Why or why not? Explain.

This question would require students to realize both scores are equidistant from the mean, but that a sample of 10 would have a larger standard error. Therefore, the smaller group would more likely to be in the extremes. I would purposely leave off the standard deviation, since in this case its value does not matter for finding the answer, and that would eliminate any confusion between whether to use the standard deviation and the standard error. I would be able to focus solely on students' understanding of standard error.

Even though this second interview question did not measure what I hoped it would measure, the responses gave me insight into another potential reason for why teaching and learning sampling distributions is so difficult: that students do not distinguish between the probability of an individual and the probability of a group.

Discussion

Teaching sampling distributions and the associated concepts such as standard error is challenging. Every statistics educator I have come in contact with agrees that is one of, if not the most, difficult topics for students to grasp. Understanding sampling distributions and the idea of standard error requires students to synthesize several different concepts from earlier in the class (variability, data distributions, etc.) and it asks them to predict what should happen with hypothetical samples (Chance, delMas, & Garfield, 2004; Reading & Reid, 2006). One suggestion made by previous researchers was if students better understood the prior concepts of variability and distribution, then students' learning of sampling distributions would improve (Chance et al., 2004). Therefore, instructors and researchers in statistics education should focus on improving student learning in these areas. As I stated at the end of chapter 4, one of the two major conjectures I took from my data was that understanding the connection between distributions of data and measures of variability is essential for students in an introductory statistics course.

It is impossible to say, based on my particular results, whether improving student's understanding of variability as it relates to standard deviation early in the course leads to a better understanding of standard error and sampling distributions later in the quarter. But given the small scale of my particular study and the limitations of this particular course at this particular college, I am not implying that this suggestion is invalid. However, I do believe that there is still much more we as researchers do not understand about how students come to understand (or not understand) the concepts sampling distribution and standard error.

What was clear from my interviews and from the lab results is there is a disconnect between understanding what a sampling distribution is and understanding how and when a sampling distribution is applied, and by extension, when standard error is required in a calculation question. Other researchers have theorized this is a direct result of not understanding distribution, variability, the normal distribution, and the idea of sampling (Chance et al., 2004). However prior knowledge of probability is only usually mentioned in the context of the normal distribution; i.e. "students should also be familiar with the idea of area under a density curve and how the area represents the likelihood of outcomes" (Chance et al., 2004, p. 300). And certainly knowing the probabilities change as standard error changes, which changes as the sample size changes in a sampling distribution requires students to make several conceptual connections. This was essential to understanding the assessment questions I set forth in the research labs and the interviews. But a good deal of the responses indicated not an issue with understanding that the probabilities change in a sampling distribution, but rather an issue with understanding that probabilities can be applied to an entire group of data points or on a statistic calculated on a group of data points, such as the mean. Students did not want to consider calculating probabilities for a measure from each sample as a whole. What this may indicate is that in addition to the other topics that researchers propose as prior knowledge required to fully understand the concept of sampling distributions, a more complete understanding of joint probabilities and probability assigned to anything other than an individual item is also required.

Conceptual change theory places a heavy emphasis on being aware of what knowledge students possess prior to a particular learning unit. My research results suggest that students possess a much deeper understanding of the concepts behind sampling distributions prior to learning sampling distributions in an introductory statistics course than was previously thought by researchers. Other researchers are noting this as well in their own studies (Garfield, 2007).

Other researchers have also pointed out that students will attempt to use memorized formulas and definitions in place of reasoning through a problem (Chance et al., 2004), perhaps because they are not confident in their own reasoning abilities. Students think they must use formulas and definitions presented in textbooks and in class to answer a question correctly. If we as educators could find a way to capture and retain this prior intuitive knowledge about the behavior of sampling distributions, perhaps students would more easily understand the concepts of sampling distribution and standard error. If this hypothesis were true, that would imply that a course that focused less on formulas and definitions and more on application and reasoning, would be successful in promoting student's conceptual understanding throughout the entire introductory statistics course. This aligns with the recommendations of the GAISE report, put forth by the American Statistical Association (2005): "stress conceptual understanding rather than mere knowledge of procedures" (p. 10).

Implications for classroom practice

This research has implications for classroom instructors. The first implication is that instructors need to find ways to make students communicate their reasoning throughout the course. Statistical reasoning is very difficult to assess and is often masked in traditional assessments. There needs to be ongoing, authentic assessment of students' statistical reasoning skills. The students that participated in the interviews for this research were mostly doing very well in the course on the traditional tests and quizzes. It was then very surprising to hear their struggles with the statistical reasoning questions during the interviews. It is essential for instructors, who are trying to tech statistical reasoning skills, to receive constant feedback from students, so that instructors can try to make the necessary adjustments in the course.

Writing an assessment item, that authentically measures statistical reasoning are difficult to create. Fortunately there are now several resources for instructors for statistical reasoning assessments. The online database, ARTIST (Assessment Resource Tools for Improving Statistical Thinking), provides many good statistical assessment questions in their assessment builder and provides links to other published assessment tools. Joan Garfield and Ido Gal published a book, *The Assessment Challenge in Statistics Education,* which highlights both recent research on assessment and models for assessing student learning. The Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) also provides a wealth of information about assessment on their website. Some of the suggestions that have already been made in other research have included a more extensive use of statistic projects and a greater amount of writing in statistics courses (Mackisack, 1994; Sharpe, 2007).

While the conceptual change, active learning lab for sampling distributions was not as successful in the classroom, the standard deviation lab was. I believe the key to the success of the standard deviation lab over the sampling distribution lab was the idea of the "checkpoint" within the lab. During the sampling distribution lab, students more or less worked straight through the activities. And while the sampling distribution lab did encourage them to revisit their earlier predictions, this still may not have been enough to create a cognitive conflict in the student's mind and thus sparking any lasting conceptual change. Feedback from the teacher did not occur until after submission of the lab. On the other hand, the worksheet that was built into standard deviation lab forced students to check their answers with the teacher during the lab before they could move on the next part. This immediate feedback was crucial, I believe, to sparking the process of conceptual change and to creating a deeper understanding of the concepts of standard deviation. Teachers need to keep this notion of a "checkpoint" or immediate feedback in mind as they design their own activities to use in the introductory statistics classroom.

Suggestions for further research

Based on what I discovered in this study, I think there are several new areas of research that could be pursued. Research needs to be conducted to study how strengthening students' understanding of probability affects students' understanding of sampling distributions. There is research on students' understanding of probability (Reading & Shaughnessy, 2004) and research on students' understanding sampling distributions (Chance, delMas & Garfield, 2004), but little that focuses on the connection between the two. There also needs to be more research into other prior knowledge that might need to be strengthened.

Given that there are so many different topics that are prerequisite knowledge for understanding sampling distributions and that conceptual change theory appears to have been successful with the topic of standard deviation in this study, a more comprehensive study that employs these types of active learning research labs would provide much more data on their effectiveness. Three labs interspersed in a 10-week course are not enough of a change to draw any meaningful conclusions about active learning techniques.

There also appears to be very little data on the unique challenges that a liberal arts statistics instructor faces. Although it wasn't a focus of my study, my review of the literature found little information on how these courses might differ from introductory courses at larger institutions that have a specific target audience.

What I learned through the dissertation process

When I started this process, I have to admit I was not entirely convinced of the value of conducting research for my improving my skills as a statistics instructor. I was simply conducting research in order to complete the requirements for my doctorate. However, this dissertation experience has completely changed my beliefs on teaching, the value of research on teaching, and where I want to go with my career.

I have been teaching in some capacity for almost ten years and I always believed I was doing a good job. I got excellent student evaluations and great job reviews. Students really seemed to appreciate all of the effort I put into making the course interesting and fun, and I thought they were learning what I was trying to teach. This research, however, showed me just how much I don't know about how students learn and the teaching process. It was very hard not to feel like a terrible teacher the first time I listened to the student interviews. It shocked me, but it also motivated me. This one research project provided me with many more questions than answers. And I want to pursue answers to these questions. If there is one thing I will take from this experience, it is the belief that conducting research on teaching and being an effective teacher are inextricably linked. This will not be my one and only research experience, but rather hopefully the first in a long career in statistics education.

Summary

It is my hope that this research will both encourage statistics instructors to continue trying new teaching techniques and but also caution them against trusting any one teaching method as *the* method that will work. My research had mixed results. The gains the active learning students made in conceptually understanding standard deviation is quite encouraging. However, gains were not as much as I would have hoped in the area of standard error and sampling distributions. However, it has motivated me to continue researching how students learn to reason statistically, how to promote statistical reasoning in my role as the instructor, and to research how reconcile these ideas with the needs and skills of a diverse liberal arts audience.

REFERENCES

American Statistical Association. (2005). *GAISE College Report.* Retrieved October 16, 2007, from http://www.amstat.org/education/gaise/

American Statistical Association (2007). *Using statistics effectively in mathematics education research.* Retrieved December 2, 2007, from www.amstat.org/research_grants/pdfs/SMERReport.pdf

Assessment Resource Tools for Improving Statistical Thinking (ARTIST). (n.d.). Retrieved October 20, 2007, from https://app.gen.umn.edu/artist/

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ball, D. L. (2000). Working on the inside: Using one's own practice as a site for studying teaching and learning. In A. Kelly & R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp 365 – 402). Mahwah, NJ: Lawrence Erlbaum Associates.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning*, *2(1&2),* 127–155.

Ben-Zvi, D. (2004). Reasoning about data analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 121–146). Dordrecht, The Netherlands: Kluwer Academic Publishers.
Ben-Zvi, D. & Friedlander, A. (1997). Statistical thinking in a technological environment. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 International Association for Statistics Education Round Table Conference* (pp. 45-55). International Statistical Institute.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 3–15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Biehler, R. (1997). Students difficulties in practicing computer supported data analysis. In J. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: 1996 Proceedings of the 1996 IASE Round Table Conference* (pp. 169-190). International Statistical Institute.

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press.

Binnie, N. (2002). Using projects to encourage statistical thinking. In B. Phillips (Ed.) *Proceedings of the sixth international conference on teaching statistics: Developing a statistically literate society.* Retrieved October 16, 2007 from http://www.stat.auckland.ac.nz/~iase/publications/1/10_69_bi.pdf

Bradstreet, T. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician 50(1),* 69-78.

Brown, J. S, Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Education Researcher, 18*(1)*,* 32-42.

Boeree, C. G. (2006). Personality Theories: Jean Piaget. Retrieved October 16, 2007, from http://webspace.ship.edu/cgboer/piaget.html

Borresen, C. (1990). Success in introductory statistics with small groups. *College Teaching, 38(1),* 26-28.

Burrill, G. (2002). Simulation as a tool to develop statistical understanding. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics, Developing a statistically literate society*. Retrieved October 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/1/7d1_burr.pdf

Burrill, G., & Romberg, T. (1998). Statistics and probability for the middle grades: Examples from Mathematics in Context. In S. Lajoie (Ed)., *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 33-59). Mahwah, NJ: Lawrence Erlbaum Associates.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). Retrieved October 16, 2007, from http://www.amstat.org/publications/jse/v10n3/chance.html

Chance, B., delMas, R., & Garfield, J. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7.* Retrieved May 15, 02, from http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295 – 323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chi, M., & Roscoe, R. (2002). The processes and challenges of conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3 – 28). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, G. (1991). Teaching statistics: More data, less lecturing. *Amstat News, 182,* 3-4.

Cobb, G. (1992). Teaching statistics. In L.A. Steen (Ed.), *Heeding the Call for Change: Suggestions for Curricular Action* (pp. 3 – 43). Mathematical Association of America.

Cobb, G. & Moore, D. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly 104,* 801-823.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher, 23*(7), 13-20.

Collins, L., & Mittag, K. (2005). Effect of calculator technology on student achievement in introductory statistics course. *Statistics Education Research Journal, 4*(1), 7-15.

Cross, P. K., & Steadman, M. H. (1996). *Classroom Research: Implementing the Scholarship of Teaching.* San Franscisco, CA: Jossey-Bass.

Davis, J. (2001). Conceptual change. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology*. Retrieved October 16, 2007, from http://projects.coe.uga.edu/epltt/

De Jong, E. J., & Gunstone, R. F. (1988). A longitudinal classroom study of mechanics concepts and conceptual change. Paper presented at the 61st Annual Meeting of the National Association for Research in Science Teaching. (ERIC Document Reproduction Service No. ED291580)

delmas, R. C. (2001). What Makes the Standard Deviation Larger or Smaller? Retrieved October 20, 2006 from http://www.causeweb.org/repository/StarLibrary/activities/delmas2001

delMas, R. C. (2002). Statistical literacy, reasoning, and learning. *Journal of Statistics Education, 10*(3), Retrieved October 16, 2007, from http://www.amstat.org/publications/jse/v10n3/delmas_intro.html delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. Reasoning about sampling distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 79 – 96). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Denzin, N K., & Lincoln, Y. S. (2000). Introduction: The discipline and practice of qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research, 2nd Edition* (pp. 1-30). Thousand Oaks, CA: Sage Publications.

Dietz, E. (1993). A cooperative learning activity on methods of selecting a sample. *The American Statistician, 47*(2)*,* 104-108.

Doerr, H.M., & Tinto, P. P. (2000). Paradigms for teacher-centered, classroom-based research. In A. Kelly & R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp 403 – 427). Mahwah, NJ: Lawrence Erlbaum Associates.

Fendel, D. & Doyle, D. (1999). Welcome to our focus issue on statistics. *Mathematics Teacher, 92(8),* 658-659.

Field, C. (1985). Projects in elementary statistics classes. *American Statistical Association: Proceedings of the Section on Statistical Education* (pp. 146-149). Washington, DC: American Statistical Association.

Fong, G, Krantz, D. & Nisbett, R. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253-292.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70,* 1-51.

Gal, I., & Garfield, J.B. (1997). *The Assessment Challenge in Statistics Education.* Amsterdam, Holland: IOS press.

Gal, I., & Short, T. (2006) Editorial. *Statistics Education Research Journal, 5*(2), 2-3.

Garfield, J. B. (1995). How students learn statistics. International Statistical Review, 63, 25-34.

Garfield, J.B. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). Retrieved October 18, 2007, from www.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J.B. (2007, May). *Connecting Research to Practice: Collaboration, Mixed Methods, and Lessons Learned.* Paper presented at the meeting of the United States Conference on Teaching Statistics, Columbus, OH.

Garfield, J. & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education 19(1),* 44-63.

Garfield, J. B. & delMas, R. C. (1991). Students' conceptions of probability. In D. Vere-Jones, S. Carlyle, & B. P. Dawkins (Eds.) *Proceedings of the Third International Conference on Teaching Statistics: Vol. 1* (pp. 340-349).

Garfield, J. B., delMas, B., Chance, B., & Oooms, A. (2006). *Comprehensive Assessment of Outcomes in a first course in Statistics (CAOS).* Retrieved October 20, 2007, from https://app.gen.umn.edu/artist/caos.html

Gnanadesikan, M., Scheaffer, R., Watkins, A., & Witmer, J. (1997). An activity-based statistics course. *Journal of Statistics Education, 5*(2), Retrieved April 25, 2004, from http://www.amstat.org/publications/jse/v5n2/gnanadesikan.html

Groth, R. (2006). An exploration of students' statistical thinking. *Teaching Statistics, 28*(1), 17-21.

Gunstone, R. F. (1994). The importance of specific science content in the enhancement of metacognition. In P. Fensham, R. Gunstone, & R. White (Eds.), *The content of science: A constructivist approach to its teaching and learning,* (pp. 131-146). London: The Falmer Press.

Hammerman, J.K. & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41.

Hawkins, A. (1996). Myth-Conceptions. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: 1996 proceedings of the 1996 IASE round table conference* (pp. 1-14). International Statistical Institute.

Hewson, P., Beeth, M., & Thorley, R. (1998). Teaching for conceptual change. In B. Fraser & K. Tubin (Eds.), *International Handbook of Science Education* (pp. 199-218). Great Britain: Kluwer Academic Publishers.

Hogg, R. V. (1991). Statistical education: Improvements are badly needed. *The American Statistician, 45*(4), 342-343.

Ivarsson, J., Schoultz, J., & Saljo, R. (2002). Map reading versus mind reading: Revisiting childrens' understanding of the shape of the earth. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 77 – 100). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Jones, G.A, Langrall, C.W., Mooney, E.S., & Thorton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 97 – 118). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Jun, L., & Pereira-Mendoza, L. (2002). Misconceptions in probability. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics, Developing a statistically literate society.* Retrieved October 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/1/6g4_jun.pdf

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement under uncertainity: Heuristics and biases.* New York: Cambridge University Press.

Keeler, C.M., & Steinhorst, R.K. (1995). Developing material for introductory statistics courses from a conceptual, active learning viewpoint. *Journal of Statistics Education, 3*(3). Retrieved October 20, 2007, from http://www.amstat.org/publications/jse/v3n3/steinhorst.html

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6(1),* 59-98.

Konold, C., Pollatsek, A., Well, A, Lohmeier, J. & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24(3),* 392-414.

Lajoie, S., Lavigne, N., Munsie, S., & Wilkie, T. (1998). Monitoring student progress in statistics. In S. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 199-231). Mahwah, NJ: Lawrence Erlbaum Associates.

Lajoie, S., & Romberg, T. (1998). Identifying an agenda for statistics instruction and assessment in K-12. In S. Lajoie (Ed.), *Reflections on Statistics* (pp. xi – xxi). Mahwah, NJ: Lawrence Erlbaum Associates.

Lan, W., Bradley, L., & Parr, G. (1993). The effects of a self-monitoring process on college students' learning in an introductory statistics course. *Journal of Experimental Education, 62(1),* 26-40.

Leavy, A. (2006) Using data comparison to support on a focus on distribution: Examining preservice teachers' understanding of distribution when engaged in statistical inquiry. *Statistics Education Research Journal, 5*(2), 89-114.

Lecoutre, M. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23(6),* 557-568.

Lee, C., Meletiou, M., Watchel, H., & Zeleke, A. (2002). The issue of motivation and expectation in the introductory statistics – obstacles and opportunities. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics, Developing a statistically literate society*. Retrieved October 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/1/8a1_lee.pdf

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the sixth international conference on teaching statistics, Developing a statistically literate society*. Retrieved October 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/1/6c1_lips.pdf

Limon, M., & Mason, L. (Eds.). (2002). *Reconsidering conceptual change: issues in theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Liu, Y., & delMas, R.C. (2005) Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55-82.

Lovett, M. (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In D. Klahr & S. Carver (Eds.), *Cognition and Instruction: Twenty-Five Years of Progress* (pp. 347-384). Mahwah, NJ: Lawrence Erlbaum.

Lovett, M. & Greenhouse, J. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54(3),* 196-206.

Mackisack, M. (1994). What is the use of experiments conducted by statistics students? *Journal of Statistics Education, 2*(2), Retrieved April 25, 2004, from http://www.amstat.org/publications/jse/v2n1/mackisack.html

Mackisack, M., & Petocz, P. (2002). Projects for advanced undergraduates. In B. Phillips (Ed.) *Proceedings of the sixth international conference on teaching statistics: Developing a statistically literate society.* Retrieved October 16, 2007 from http://www.stat.auckland.ac.nz/~iase/publications/1/3e4_peto.pdf

Magel, R. (1996). Increasing student participation in large introductory statistics classes. *The American Statistician, 50*(1), 51-56.

Magel, R. (1998). Using cooperative learning in a large introductory statistics class. *Journal of Statistics Education, 6*(3). Retrieved October 20, 2007, from http://www.amstat.org/publications/jse/v6n3/magel.html

Makar, K., & Confrey, J. (2005). "Variation-talks": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27-54.

Marasinghe M.G., Meeker, W.Q., Cook, D. & Shin, T. (1996). Using graphics and simulations to teach statistical concepts. *The American Statistician, 50,* 342-351.

Marcoulides, G. A. (1990). Improving learner performance with computer based programs. *Journal of Educational Computing Research, 6,* 147-155.

Martinez-Dawson, R. (2003). Incorporating Laboratory Experiments in an Introductory Statistics Course. *Journal of Statistics Education, 11(1),* Retrieved November 19, 2004 from http://www.amstat.org/publications/jse/v11n1/martinez%2Ddawson.html

Mayer, R. E. (2002). Understanding conceptual change: A commentary. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 101 – 114). Dordrecht, The Netherlands: Kluwer Academic Publishers.

McClendon, M. A. (1992). The development of a graphics calculator study guide for calculus students. *Dissertation Abstracts International, 52*, 2450A.

Metz, K. (1998). Emergent understanding and attribution of randomness: comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction 16(3),* 285-365.

Mickelson, W.T., & Haseman, A. (1998). Students' conceptual understanding of the sampling distribution of the sample proportion in a constructivist learning environment. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Mills, J. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education, 10*. 1-20. Retrieved June 2, 2002 from http://www.amstat.org/publications/jse/v10n1/mills.html#Giesbrecht.

Moore, D. (1992). What is statistics? In D.C. Hoaglin & D.S. Moore (Ed.), *Perspectives on Contemporary Statistics: MAA Notes Number 21* (pp. 1-17). Mathematical Association of America.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65,* 123-165.

Moore, D. S. (1998). Should Mathematicians teach statistics? *College Mathematics Journal, 19*(1), 3-35.

Moore, D., & Cobb, G. (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly, 107,* 615-630.

Mvududu, N. (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education, 11*(3). Retrieved October 20, 2007, from http://www.amstat.org/publications/jse/v11n3/mvududu.html

National Council of Teachers of Mathematics. (2003). *National Council of Teacher's Principles and Standards for School Mathematics*. Retrieved December 7, 2003 from http://www.nctm.org

Nicholson, J. (1996). Developing probabilistic and statistical reasoning at the secondary level through the use of data and technology. In J. B. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics: 1996 proceedings of the 1996 IASE round table conference* (pp. 29-45). International Statistical Institute.

Ng, V., & Wong, K. (1999). Using simulation on the internet to teach statistics. *Mathematics Teacher, 92,* 729-733.

Pfannkuch, M. (2005). Characterizing year 11 students' evaluation of a statistical process. *Statistics Education Research Journal, 4*(2), 5-26.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal, 5*(2), 4-9.

Pollatsek, A., Konold, C.E., Well, A.D., & Lima, S.D. (1984). Beliefs underlying random sampling. *Memory & Cognition, 12,* 395-401.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211- 227.

Raymondo, J. C., & Garrett, J. R. (1998). Assessing the introduction of a computer laboratory experience into a behavioral science statistics. *Teaching Sociology, 26,* 29-37.

Reading, C., & Reid, J. (2006). A emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Reading, C., & Shaughnessy, J. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 169 – 200). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rosen, E., Feeney, B., & Petty, L. (1994). An introductory statistics class and examination using SPSS/PC. *Behavior Research Methods, Instruments, & Computers, 26,* 242-244.

Rossman, A. (1997). Workshop statistics: Using technology to promote learning by self-discovery. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: 1996 Proceedings of the 1996 IASE Round Table Conference* (pp. 226-237). International Statistical Institute.

Scheaffer, R., Watkins, A., Gnanadesikan, M., & Witmer, J. (1996). *Activity-Based Statistics: Instructor Resources.* New York: Springer-Verlag.

Sharpe, N. (2007). *How can writing be used effectively in statistics courses.* Paper presented at the CAUSE webinar conference. Retrieved December 18, 2007, from http://www.causeweb.org/webinar/2007-10/

Shaughnessy, M. (1977) Misconceptions of probability: an experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics 8(3),* 295-316.

Shaughnessy, M., & Zawojewski, J. (1999). Secondary students' performance on data and chance in 1996 NAEP. *Mathematics Teacher, 92,* 713-718.

Shirley, L. (2000, April). *Reviewing a century of mathematics education: Ready for the future.* Paper presented at the meeting of the National Council of Teachers of Mathematics, Chicago, IL.

Smith, G. (1998) Learning statistics by doing statistics. *Journal of Statistics Education, 6(3).* Retrieved November 20, 2004 from http://www.amstat.org/publications/jse/v6n3/smith.html

Snee, R. (1990). Statistical thinking and its contribution to total quality. *The American Statistician 44(2),* 116-121.

Snee, R. (1993). What's missing in statistical education? *The American Statistician 47(2),* 149-154.

Sylvester, D. L., & Mee, R. W. (1992). Student projects: An important element in the beginning statistics course. *American Statistical Association: 1992 Proceedings of the Section on Statistical Education* (pp. 137-141). Washington, DC: American Statistical Association.

United States Department of Education. (2007) *The Nation's Report Card: Mathematics 2007.* Retrieved October 18, 2007, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007494

Utts, J., Sommer, B, Acredolo, C., Maher, M., & Matthews, H. (2003). A study comparing traditional and hybrid internet-based instruction in introductory statistics classes. *Journal of Statistics Education 11(3),* Retrieved November 15, 2004 from http://www.amstat.org/publications/jse/v11n3/utts.html

Varnhagen, C. K., & Zumbo, B. D. (1990). CAI as adjunct to teaching introductory statistics: Affect mediates learning. *Journal of Educational Research, 6,* 29-40.

Von Glasersfeld, E. (1997). Homage to Jean Piaget. *Towards an Ecology of Mind.* Retrieved December 2, 2007, from http://www.oikos.org/Piagethom.htm

Vosniadou, S. (2002). On the nature of naive physics. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 61 – 76). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson, J.M. (1997). Assessing statistical thinking using the media. In I. Gal & J.B. Garfield (Eds.), *The Assessment Challenge in Statistics Education.* Amsterdam, Holland: IOS Press.

Watson, J.M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 257 – 276). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watson J. M. & Kelley, B. A. (2002). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society.* Retrieved October 13, 2007, from http://www.stat.auckland.ac.nz/~iase/publications.php?show=1

Watson, J.M., Kelly, B.A, Callinham, R. A., & Shaughnessy, J.M. (2003). The measurement of statistics students' understanding of variation. *Canadian Journal of Science, Mathematics, and Technology Education, 34,* 1-29.

Watson, J.M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal of research in Mathematics Education, 31*(1), 44-70.

Wild C. & Pfannkuch, M. (1999) Statistical thinking in empirical enquiry. *International Statistical Review, 67,* 223-265.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement, 45*, 401 – 405.

Wolfe, C. (1993). Quantitative reasoning across a college curriculum. *College Teaching 41(1),* 3-9.

APPENDIX A

ORIGINAL INFORMED CONSENT

Original Informed Consent Form
Student's Conceptual Understanding of Variability

The Department of Mathematical and Computer Sciences at Otterbein College supports the practice of protection for human subjects participating in research. The following information is provided for you to decide whether you wish to participate in the present study. You should be aware that even if you agree to participate, you are free to withdraw at any time without penalty.

I am interested in researching how student's conceptual understanding of variability in an introductory statistics course develops. You will be completing three lab activities and an online (secure) statistics assessment as part of the course requirements for this class. By participating in the present study, you are allowing me to further analyze your work (beyond normal grading) for the purposes of my research.

I am also asking volunteers to participate in one twenty-minute interview outside of class time. You may choose to be in the study without signing up for an interview. The interview will involve a few questions about mathematical background and a few statistical analysis questions. You would not need to prepare anything for this interview

Your participation in any part of the study is solicited but strictly voluntary. I assure you that your name will not be associated in any way with the research findings. Only a code number will identify this information. If you choose to not be in the study, that will in no way have any bearing on your grade in this class and no one in the class will know who choose to participate in the study and who did not.

If you would like additional information concerning this study before or after it is complete, please feel free to contact me by phone or mail.

Sincerely,

Leigh Slauson, Co -Investigator
Otterbein Library
(614) 823-1624
lslauson@otterbein.edu

_____Patti Brosnan, Principal Investigator

Signature of subject agreeing to participate (with my signature I affirm that I am at least 18 years of age). Please sign on line that is your choice.

_________________________I am willing to participate in the study and I would be interested in being interviewed.

_________________________I am willing to participate in the study but I do not wish to be interviewed.

_________________________I do not wish to be a part of this study.

APPENDIX B

SUPPLEMENTAL INFORMED CONSENT

Supplemental Informed Consent Form
Student's Conceptual Understanding of Variability

The Department of Mathematical and Computer Sciences at Otterbein College supports the practice of protection for human subjects participating in research. The following information is provided for you to decide whether you wish to grant me access to demographic data records for the purpose of calculating class statistics. Please be assured that your individual information will be kept in the strictest confidence. You should be aware that even if you agree to participate, you are free to withdraw at any time without penalty.

I am asking your permission to access demographic and grade point average data from the Otterbein computer database for the purposes of my research. I am only requesting this information so that I may aggregate it and present a picture of this class overall (e.g. our class' overall grade point average, distribution of major, etc). Your individual information will never used or identified in my research or subsequent publications in any way.

Your participation in this part of the study is solicited but strictly voluntary. I assure you that your name will not be associated in any way with the research findings. Only a code number will identify this information. If you choose to not grant this request for access to your records, that will in no way have any bearing on your grade in this class. No one else in the class will know who chooses to grant access and who did not.

In addition, I am soliciting participants to take the online CAOS test three months after the end of Winter quarter. This is strictly voluntary. I will email the link and the access code and you would have a week to complete the test. If you would be willing to do this, please check the box below.

If you would like additional information concerning this study before or after it is complete, please feel free to contact me by phone or mail.

Sincerely,

Leigh Slauson, Co -Investigator
Otterbein Library
(614) 823-1624
lslauson@otterbein.edu

_____Patti Brosnan, Principal Investigator

Please sign on the appropriate line below. (Your signature indicates that you, as the signee, are at least 18 years old.)

_________________________For the purposes of statistical summaries in the research described previously, I grant the researcher permission to access my Otterbein records.

_________________________I do not grant the researcher access to my Otterbein record.

□ Please check the box if you would be willing to be contacted beyond the end of the quarter to take the online assessment (CAOS) test again.

APPENDIX C

STANDARD DEVAITION LAB

Standard Deviation Lab (Part 1)

Goal: You will discover how the measure of standard deviation relates to a dataset and what it looks like in histogram form.

Part 1: Initial Conjectures (Do this part at home)

For each of the following pairs of datasets, pick the dataset you believe will have the larger standard deviation. Explain your choice in a sentence or two.

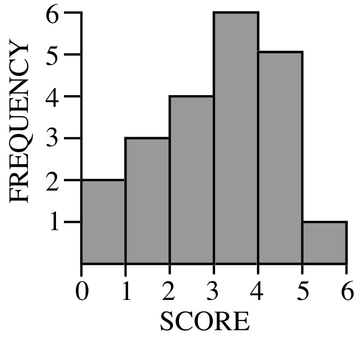
Dataset A	Dataset B
Amount of change (coins only) students currently have with them (our class only)	Amount of money students spent on their last haircut (our class only)
Explain	
Number of CD's students own (our class only)	Number of pairs of shoes students own (our class only)
Explain	
Student's height in inches (our class only)	Student's shoe size (our class only)
Explain	

Part 2: Connecting Standard Deviation to a Histogram.

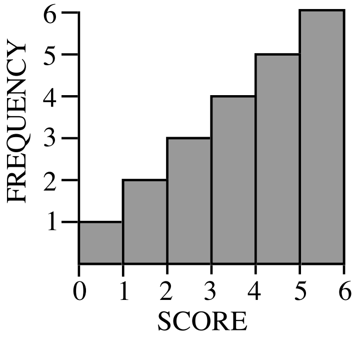
For the following pairs of histograms, choose which one you think will have the larger standard deviation. Check with me that you have the right answers before you move on to Part 3.

1.

A μ **= 2.57**



B μ **= 3.33**



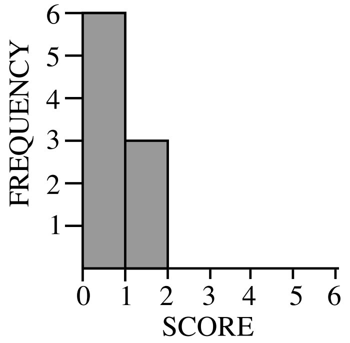
A has a larger standard deviation than B

B has a larger standard deviation than A

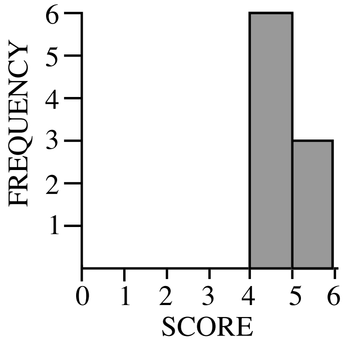
Both graphs have the same standard deviation

2.

A μ **= .33**



B μ **= 4.33**



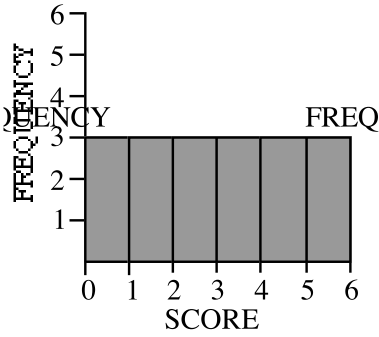
A has a larger standard deviation than B

B has a larger standard deviation than A

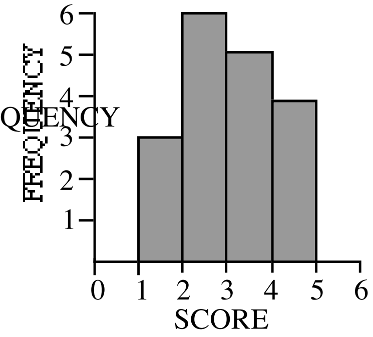
Both graphs have the same standard deviation

3.

A μ **= 2.50**



B μ **= 2.56**

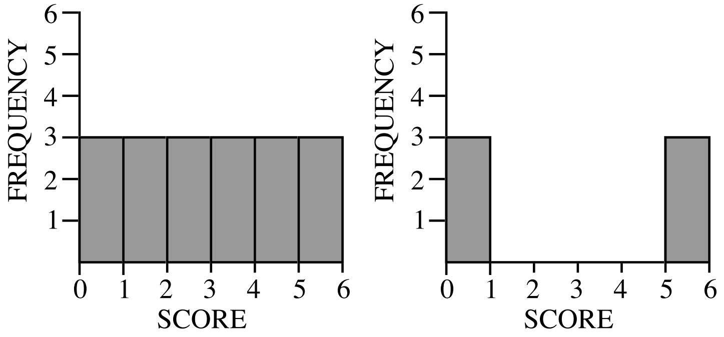


A has a larger standard deviation than B

B has a larger standard deviation than A

Both graphs have the same standard deviation

4. **A** μ **= 2.50** **B** μ **=2.50**

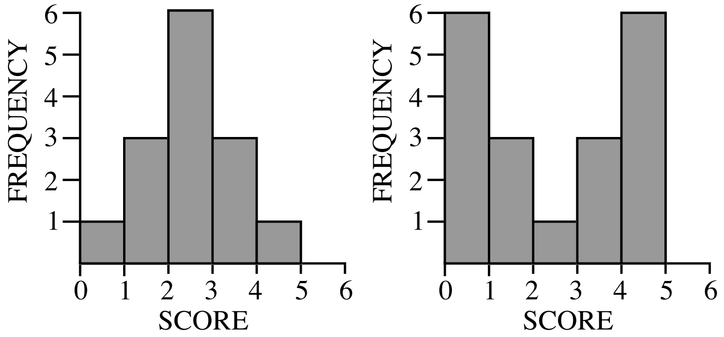


A has a larger standard deviation than B

B has a larger standard deviation than A

Both graphs have the same standard deviation

5. **A** μ **= 2.00** **B** μ **= 2.00**

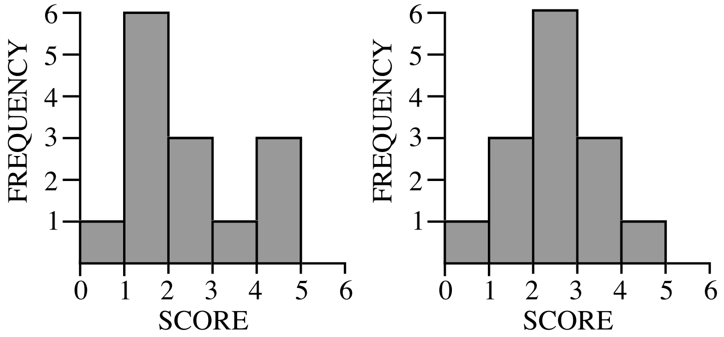


A has a larger standard deviation than B

B has a larger standard deviation than A

Both graphs have the same standard deviation

6. **A** μ **= 1.93** **B** μ **= 2.00**



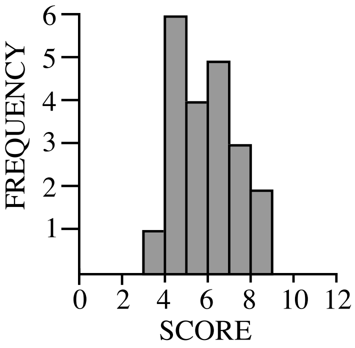
A has a larger standard deviation than B

B has a larger standard deviation than A

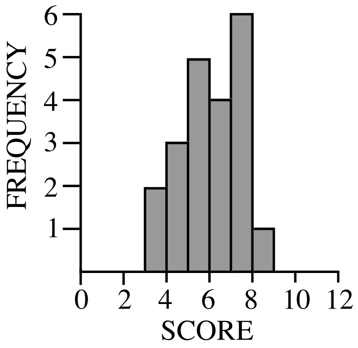
Both graphs have the same standard deviation

7.

A μ **= 5.43**



B μ **= 5.57**



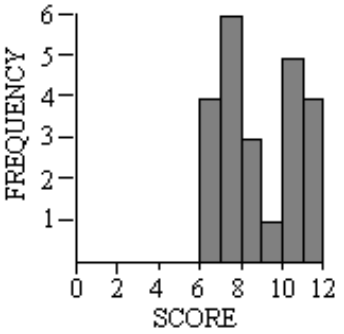
A has a larger standard deviation than B

B has a larger standard deviation than A

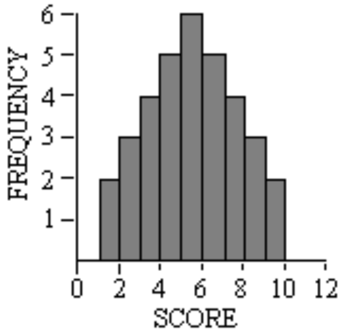
Both graphs have the same standard deviation

8.

A μ **= 8.33**



B μ **= 5.00**



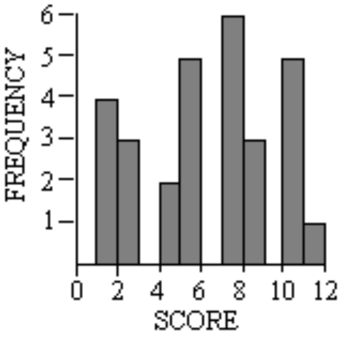
A has a larger standard deviation than B

B has a larger standard deviation than A

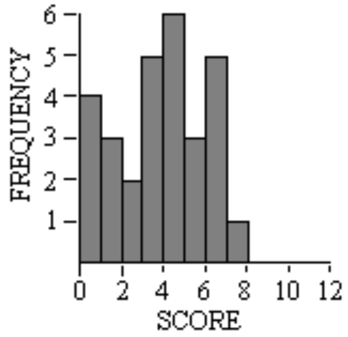
Both graphs have the same standard deviation

9.

A μ **= 5.86**



B μ **= 3.38**



A has a larger standard deviation than B

B has a larger standard deviation than A

Both graphs have the same standard deviation

Part 3: Using Your Own Collected Data

Using the attached data that we collected in class (these match to the variables in part one), construct a histogram and calculate the both the mean and the standard deviation using your graphing calculators. Be sure to provide a scale of each of your histograms.

1) Pocket Change

Histogram:

Mean________ SD_______

2) Haircut

Histogram:

Mean_______ SD_______

3) CD's

Histogram:

Mean___________

SD___________

4) Shoes

Histogram:

Mean___________

SD___________

5) Student's heights

Histogram:

Mean___________

SD___________

6) Shoe size

Histogram:

Mean___________

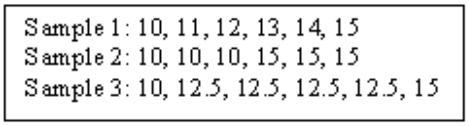
SD___________

Part 4: Analysis

1) Compare your histograms in Part 3 to your conjectures in Part 1. Are the results different than what you predicted or are they what you expected? Explain.

2) Name at least one characteristic of a data set or a distribution that has an effect on the size of the standard deviation. Give an example using your data above.

3) Consider the following samples of quiz scores and answer the questions without doing any calculations.



Sample 1: 10, 11, 12, 13, 14, 15
Sample 2: 10, 10, 10, 15, 15, 15
Sample 3: 10, 12.5, 12.5, 12.5, 12.5, 15

a. Which sample has greatest standard deviation? Why?

b. Which sample has the smallest standard deviation? Why?

APPENDIX D

SAMPLING DISTRIBUTIONS LAB

Preliminary Questions for Lab 5

(answer before coming to class)

1) We will be looking at the distribution of the ages of pennies in the population. What do you think the shape of this distribution will look like? Why?

2) Does it matter where we sample our pennies from (jar at home, the bank, etc)? Will we get different distributions? Why?

3) Suppose the average age of a penny in the general population is 12 years old. When you take a random sample of 5 pennies, do you think the average age will also be 12 in your sample? What if we took a sample of 50 pennies? Explain your answers.

4) If everyone in the class reaches into the same jar of pennies and finds the average age of their handful (about 25 pennies), do you expect that everyone will come in with the same average or different averages? If not, how much variability do you expect between everyone's averages? (A lot, some, or almost no variability?)

5) Would changing how many pennies everyone drew from 25 to 100 change how much variability there is between between everyone's averages? Explain.

Cents and Sampling Distributions Lab

Due Friday February 16, 2007

Objective: In this activity you will discover the central limit theorem by observing the shape, mean, and the standard deviation of the sampling distribution of the mean for samples taken from a distribution that is decidedly not normal.

Materials:
Population of Pennies
Graphing Calculator

Procedure:

1) You should have a list of the dates on a random sample of 25 pennies. Next to each date, write the age of the penny by subtracting the date from 2006. Use your calculator to graph a histogram of the ages of your pennies. This gives you at least an idea of what the population distribution of pennies looks like. Sketch a picture of the histogram on the next page. Please label the bars using the TRACE feature of your calculator. What appears to be the shape of your histogram?

Year	Age

Distribution of penny ages:

2) Use the 1-Var Stats feature on your calculator to find the mean and standard deviation of your distribution. This is your estimate for the population mean and standard deviation.

Mean ________________
Standard Deviation ________________

3) From your collection of pennies, sample 5 pennies at a time. Write down the age of each below and then calculate the average for the sample of 5 pennies Repeat these steps in order to obtain 15 random samples of 5 pennies (without replacement) and compute the mean age of each of your samples. This is your sampling distribution of sample means

n = 5

Sample	Age Penny 1	Age Penny 2	Age Penny 3	Age Penny 4	Age Penny 5	**Average** **Age**
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

4) Look at the last column of averages you just generated. Do you think the average of those numbers will be larger than, smaller than, or the same size as mean age of all pennies you calculated in step 2? Do you think the standard deviation will change? (see your answers to #2) Why?

5) Graph and sketch below a histogram of the average ages (last column of data only). Calculate the mean and standard deviation of those means. What is the shape of your data?

Sketch and Shape: Mean ______________

Standard Deviation ________________

6) I repeated this experiment for you but instead of sampling 5 pennies at a time, I sampled 30. (Trust me this would have taken you a while ☺). Below are the average ages of those samples of 30 pennies each.

```
Variable    Mean
Sample 1   15.30
Sample 2   15.40
Sample 3   16.67
Sample 4   18.83
Sample 5   19.90
Sample 6   18.87
Sample 7   16.73
Sample 8   15.97
Sample 9   15.73
Sample 10  20.70
Sample 11  16.83
Sample 12  16.80
Sample 13  16.77
Sample 14  15.00
Sample 15  17.27
Sample 16  17.33
```

Using the data above, graph and sketch a histogram of the average ages (second column of data only). Calculate the mean and standard deviation of these means. What is the shape of your data?

Sketch and Shape:

Mean ______________

Standard Deviation ______________

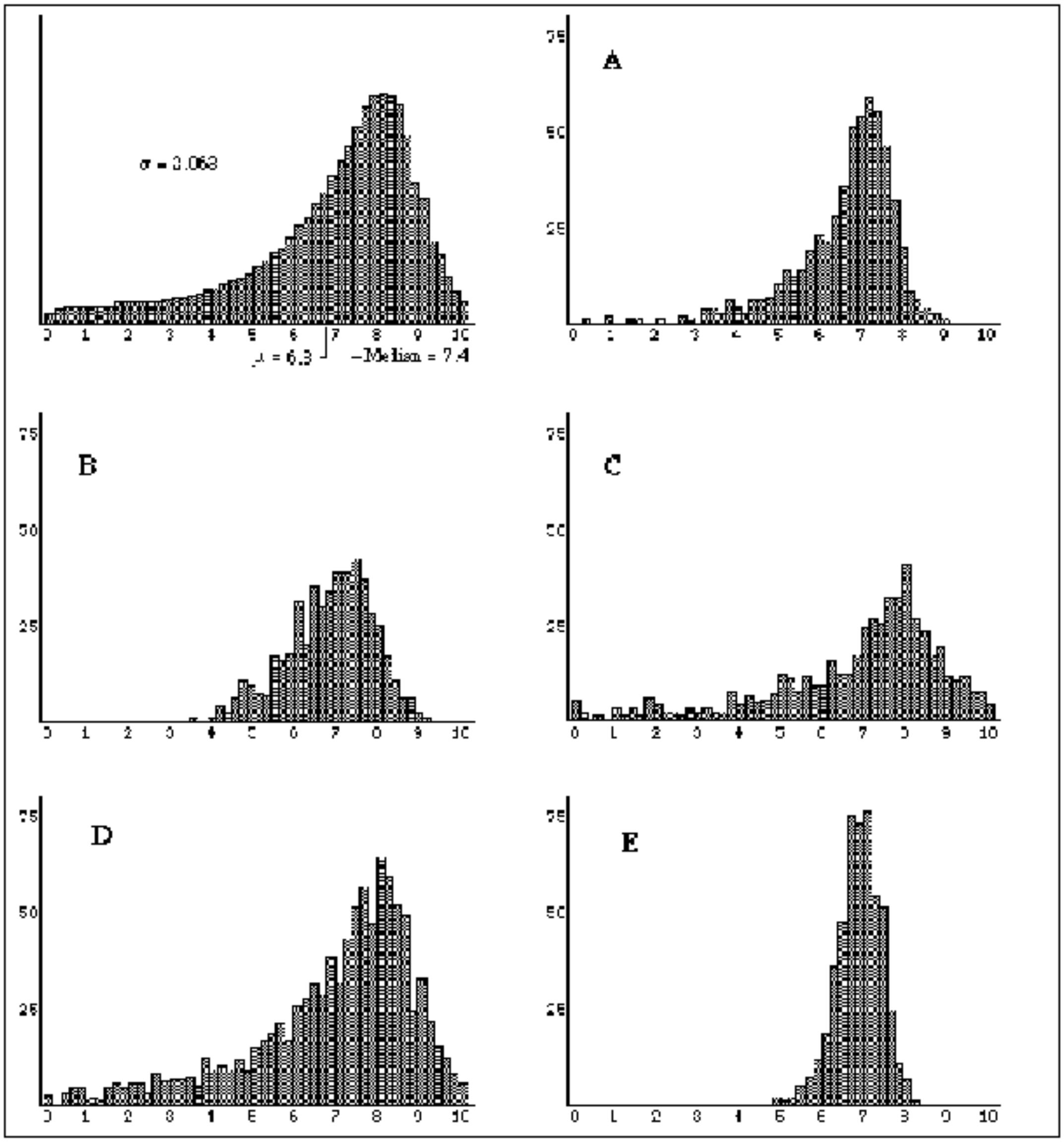
7) Compare the two distributions of means you have just created (n= 5 and n = 30). Look at the shape, the mean and the standard deviation. What is changing and what is staying approximately the same?

Extensions

1) The distribution of Math SAT scores is Normal with mean 455 and standard deviation 102. Would you expect to find a difference between (a) the probability that a single person scores above 470 and (b) the probability that in a random sample of 35 people, the sample mean () is above 470. Why or why not? Explain.

2) Alice, Ben, Connie and Dwayne have each taken a random sample of students from their school to estimate the variability in amount spent on movie tickets this summer. Alice asked 10 people, Ben 30, Connie 50, and Dwayne 70. Whose sample standard deviation probably differs most from true population standard deviation?

3) The upper left graph is the distribution for a population of test scores. Each of the other five graphs, labeled A to E represent possible sampling distributions of sample means for 500 random samples drawn from the population.



Which graph represents a sampling distribution of sample means where only 1 test score was selected at a time? Justify your choice.

APPENDIX E

CONFIDENCE INTERVAL LAB – ACTIVE CLASS

Lab 7 – Confidence Intervals for the Population Proportion

Preliminary Questions:

1) If the M&M's company distributes the colors equally in its packages, what percentage of the M&M's should be green? (HINT: There are six basic colors)

2) If each person in our class randomly samples a package of M&M's, will each person have the same percentage of green M&M's? How close to the percentage you found in number 1, do you think everyone should be if there is indeed an even number of all the colors (You can just guess here….you don't need to calculate anything)?

3) Suppose one person has 10% green M&M's in their sample. Is this enough evidence for you to argue that M&M's is giving its customers "less" green M&M's. Explain.

4) Suppose the following are the results from 10 people with 10 different samples (assume everyone has approximately 70 M&M's each). Based on these results, would you argue that M&M's is giving its customers "less" green M&M's. Why or why not?

Sample Percentage($\hat{p}$)	**95% margin of error**
9.5%	+/- 8.1%
11.5%	+/- 8.8%
12%	+/- 9%
8.9%	+/- 7.8%
8.9%	+/- 7.8%
9.6%	+/- 6.9%
10.6%	+/- 7.2%
14.7%	+/- 8.2%
9.9%	+/- 6.9%
10.4%	+/- 7.2%

Lab 7 – Confidence Intervals for the Population Proportion

Due Monday February 26, 2007

Materials: M&M's (Do NOT eat your data until you complete this lab!!)

Goal: To construct a confidence interval for the population proportion and to understand how to interpret it.

Steps:

1) Open your packages of M&M's and count how many total M&M's you have. This is your n. n = _________

2) We will decide as a class which color we will estimate the proportion of. Once this is decided count the number of that particular color in your sample. This is x. x = _________

3) Calculate the sample proportion ($\hat{p}$) of that color.

4) Calculate the standard error of your sample.

5) Construct a 95% confidence interval for the population proportion. (See your notes for the formula.)

6) Collect the other groups' confidence intervals and fill them into the left column of the table below.

95% Confidence Intervals	Contain p?

7) Check off how many intervals actually contain that value 16.7% in the table above. What percentage of those intervals contained that value?

Discussion Questions

8) If the percentage of intervals that contain that value is not close to 95%, then there is evidence that something is amiss. In other words, there may be evidence that you are not getting 16.7% green M&M's. Comment on what the results from the entire class seem to be saying. Do you think there is evidence that the colors are not equally represented? Are you getting more green or less of it? How do you know?

9) Do you believe that that this a random sample of M&M's? Explain why or why not? (The M&M's were bought at the Walmart in Hilliard on Thursday night). If it is not a good random sample, explain how you might draw a better random sample of M&M's.

10) Constructing confidence intervals in this way requires that we assume that the sampling distribution of $\hat{p}$ be approximately normal. Using your n and the value of p from number 7, check to see if this is a valid assumption. (Hint: there are two checks you have to do).

11) Look at you answer to number 4 on the preliminary question sheet. Based on what you learned during this lab, would you change your answer from what you said originally? Explain.

Extensions:

12) It is thought that 65% of adult Americans drink alcohol. In a random sample of 50 college students, 39 say they drink alcohol of one kind or another. The dean of students wants to study whether a higher percentage of college students drink alcohol than in the adult population at large and calculates a 95% confidence interval to be (.640, .885). Suppose we plan to take another random sample from this same college.

A) How do you think the width of the new confidence interval will differ from the first if they take a random sample of 50 students instead of 25 (narrower, wider, or the same)?

B) How do you think the width of the new confidence interval will differ from the first if they use a 90% confidence level instead of 95% (narrower, wider, or the same)?

13) The USA Today AD Track (3/1/00) examined the effectiveness of the new ads involving the Pets.com Sock Puppet (which is now extinct). In particular, they conducted a nationwide poll of 428 adults who had seen the Pets.com ads and asked for their opinions. They found that 36% of the respondents said they liked the ads. If you wanted to cut the margin of error in half the next time you took a poll like this, what sample size would you need? (**Hint**: First, calculate the 95% margin of error of the current poll)

For more information on the history of M&M's and other fun facts, go to www.mms.com.

APPENDIX F

CONFIDENCE INTERVAL LAB – LECTURE CLASS

Lab 6 – Confidence Interval for the Population Proportion

Materials: M&M's (Do NOT eat your data until you complete this lab!!)

Goal: To construct a confidence interval for the population proportion and to understand how to interpret it.

Steps:

1) Open your packages of M&M's and count how many total M&M's you have. This is your n. n = _________

2) We will decide as a class which color we will estimate the proportion of. Once this is decided count the number of that particular color in your sample. This is x. x = _________

3) Calculate the sample proportion ($\hat{p}$) of that color.

4) Calculate the standard error of your sample.

5) Construct a 95% confidence interval for the population proportion. (See your notes for the formula.)

6) Collect the other groups' confidence intervals and fill them into the left column of the table below.

95% Confidence Intervals	**Contain p?**

7) If you were getting equal numbers of each color of M&M's, what percentage of the color we counted should you be getting? (This would be our claimed population proportion, p).

8) Check off how many intervals actually contain that value in the table above. What percentage of those intervals contained that value?

Discussion Questions

9) If the percentage of intervals that contain that value is not close to 95%, then there is evidence that something is amiss. In other words, there may be evidence that you are not getting equal numbers of each color of M&M's. Comment on what the results from the entire class seem to be saying. Do you think there is evidence that the colors are not equally represented? Are you getting more of this particular color or less of it? How do you know?

10) Do you believe that that this a good random sample? Explain why or why not? (The M&M's were bought at the Walmart in Hilliard on Monday night). If it is not a good random sample, explain how you might draw a better random sample.

11) Constructing confidence intervals in this way requires that we assume that the sampling distribution of $\hat{p}$ be approximately normal. Using your n and the value of p from number 7, check to see if this is a valid assumption. (Hint: there are two checks you have to do).

12) Besides taking a better random sample, what might be the best way to make our calculations more accurate and produce a narrower confidence interval?

For more information on the history of M&M's and other fun facts, go to www.mms.com.

APPENDIX G

INTERVIEW PROTOCOL

Interview Questions

Math Background

1) Why are you taking introductory statistics? Is it required for your major?
2) How many math classes have you taken at Otterbein before enrolling in Math 230?
3) Have you had any exposure to statistics before taking this class? Did your high school offer a statistics class?
4) Generally do you enjoy taking math classes?
5) What were your feelings about taking statistics before you started this quarter?

Statistics Assessment– I am going to ask you a few statistical reasoning questions. The point of the interview is for me to try understand your thought processes as you work through a statistical analysis question. Please try to explain as much of your reasoning out loud as you can. Don't worry about whether you are right or wrong. You are welcome to use whatever tools you need to work these out (paper, calculator, etc), but do you best to verbalize as much of your reasoning out loud. I may ask you follow up questions as you answer the question.

1) Suppose two distributions have exactly the same mean and standard deviation. Then the two distributions have to look exactly alike. Explain whether this is true or false.

2) Shelly is going to flip a coin 50 times and record the percentage of heads she gets. Her friend Diane is going to flip a coin 10 times and record the percentage of heads she gets. Which person is more likely to get 20% or fewer heads?

3) The distribution of Verbal ACT scores is normal with mean 21 and standard deviation 5.
Which would be more likely:
A) a single person scores less than a 16
B) the average score of 25 students is greater than a 22
Explain your choice.