Developing Students' Conceptions of Variation: An Untapped Well in Statistical Reasoning

by

Maria Menelaou Meletiou, B.A., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2000

Developing Students' Conceptions of Variation: An Untapped Well in Statistical Reasoning

Approved by
Dissertation Committee:

Jere Confrey, Supervisor

Dedication

To the memory of my beloved mother:

A mother's love is a blessing

No matter where you roam,

Keep her while you have her,

You'll miss her when she's gone.

[Frank McCourt]

Acknowledgements

May I first acknowledge the debt I owe to my Supervisor Dr. Jere Confrey. Getting to know her and her work has exerted strong influence on my intellectual growth as a mathematics educator. She has been an indispensable source of knowledge and creative ideas. She has more than anyone else taught me the power and relevance of instruction and research that give voice to student perspective. The thesis constructed out of the material of this study was the result of her encouragement and direction.

Dr. Maggie Myers has been a great support for six years now. She has always been available and eager to help with words of encouragement and insightful comments. If I were to name some of my most rewarding experiences during this long journey, getting to know her would definitely be one of them. Many of the ideas in this dissertation have sprung out of the countless conversations I have had with her over the years.

I also owe a lot to the other members of my committee: Dr. Martha Smith, Dr. Uri Treisman, Dr. Rachel Fouladi, and Dr. Walter Stroup. The sound advice and feedback they have all generously provided me with, have helped me grow both as a professional and as a person. I am also very thankful to my Graduate

Advisor Dr. Ralph Cain for his help, especially at the early stages of my doctoral work.

I am especially grateful to Dr. Joan Garfield for her guidance. She was the one introducing me to a big portion of the research literature on statistics education that has influenced the construction of this piece of work.

I am grateful to Dr. Carl Lee for allowing me access to his classroom in such a gracious way and for giving me so much of his time despite his many commitments. Our research collaboration has been a wonderful experience and I am sure will continue for many years to come. His hospitality made my stay in Michigan a very pleasant experience. I only hope to get the opportunity to return some of the many favors that he has done for me.

My special thanks go to the students who participated in this study. I learned a lot from all of them, and I enjoyed being around them. I have developed friendships with a couple of them that I am certain will last for a long time.

I also want to express my gratitude to my fellow “RUMECers”. The experiences and the sense of community I have been gaining from belonging to this unique group of mathematicians and mathematics educators are invaluable.

Thanks to all of my friends both in Austin and at home whose love and thoughtfulness have kept me going: Elena Hadjikakou and her sisters Kika and Yioula, Anna Razatos, Stalo Christodoulou, Melpo Nicolaou, Yiota Avraam, Sylvia Celedon and her husband Marios Pattihis, Robert Jankee, Leanna Loehr, Maria Antoniou and her sister Christiana, Gregorios Gregoriou, Tasos Protopapas, and many others.

I am mostly indebted to my parents Athanasia and Menelaos. The devastating experience of war, of losing everything and having to start life over as refugees, although costing them a great deal both emotionally and financially, did not stop them from providing the most loving and caring environment for their children. Whatever I might have accomplished would have never been possible without the many sacrifices my parents had to make. Dedicated teachers themselves, they instilled in me the love for knowledge from a very early age, and they made sure I got the best education possible. Their life-long commitment to their students has taught me the moral dimension of being an educator. Above all, their sacrificial love for others, has shown me in the most vivid way the truest meaning of life: *"If I speak in the tongues of men and of angels, but have not love, I am a noisy gong or a clanging cymbal. And if I have prophetic powers, and understand all mysteries and all knowledge, so as to remove mountains, but have not love, I am nothing…Love never ends; as for prophecy it will pass away; as for knowledge it will pass way. For our knowledge is imperfect and our prophecy is imperfect; but when the perfect comes, the imperfect will pass away." [1 Corinthians, 13]*

My beloved sisters Elena and Effie and their husbands Christos and Ionas have always given me their unconditional love and support. The cute voices of my two little nephews Charalambos and Nicholas have cheered me up many a times that I was feeling despaired and have reminded me of all the beautiful things waiting for me back home. The members of my extended family – aunts, uncles, and cousins – have also all supported me in many ways. I am very lucky to belong to such a loving family.

Finally, may I say thanks to Stathis Mavrotheris. He has been a constant source of unselfish love and support from the moment I met him four years ago. He has believed in me and has patiently waited for me to complete my studies, always encouraging me with gentle and kindhearted words, despite the many times that my frustrations took their toll on our long-distance relationship. I am thrilled that the end of this stage of my life brings with it the beginning of a new one, with him by my side. A life shared with him is the best graduation gift I could have ever received.

Developing Students' Conceptions of Variation: An Untapped Well in Statistical Reasoning

Publication No._____

Maria Menelaou Meletiou, Ph.D.

The University of Texas at Austin, 2000

Supervisor: Jere Confrey

The conjecture driving this study is that if statistics curricula were to put more emphasis on helping students improve their intuitions about variation and its relevance to statistics, we would be able to witness improved comprehension of statistical concepts (Ballman, 1997). Both the research literature and previously conducted research by the author indicate that variation is often neglected, and its critical role in statistical reasoning is under-recognized.

A nontraditional approach to statistics instruction that has variation as its central tenet, and perceives learning as a dynamic process subject to development for a long period of time and through a variety of contexts and tools, is laid out in this thesis. The experiences and insights gained from adopting such an approach in a college level, introductory statistics classroom are reported.

The prevailing methodology employed by researchers examining conceptions of data and chance of taking snapshots of students' thought processes by posing cognitive tasks to them in order to catalogue their misconceptions, provides little guidance as to how one might systematically research conceptual change. The conjecture-driven research design (Confrey and Lachance, 1999) employed in this study, which sees research and practice as interwoven, and advocates curriculum construction based on an ongoing process of development and feedback, offered an alternative path. It allowed finding similarities and differences between students' informal intuitions and formal statistical reasoning, and working with students' intuitive notions to help them develop ways to map new and richer concepts onto the ones that they already possessed.

The results of the study point to a number of critical junctures and obstacles to the conceptual evolution of the role of variation, including the following: (1) Understanding of histograms and other graphs; (2) Familiarity with abstract notation and with statistics language; (3) Appreciation of the need to be critical of data and always examine the method it was collected; (4) Distinguishing between population distribution, distribution of a single sample, and sampling distribution; and (5) Understanding of the reason behind finding confidence intervals when producing an estimate of some parameter based on a sample.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Pupils in the future will bring away from their schooling a structure of thought that whispers 'variation' matters.
(Moore, 1992, p.426)

In the last decade, there has been a significant move towards modernizing statistics education and a general acknowledgment that learning occurs most effectively when students engage in authentic activities. Although many statistics students from higher institutions are still being taught in traditional classrooms, a large number of statistics instructors have already adopted alternative approaches to teaching statistics, and many statistics classrooms are experiencing wide-scale incorporation of technology. Nonetheless, research on statistical thinking indicates that students' difficulties in reasoning about the stochastic persist despite the reform efforts. The conjecture driving this study is that the reform movement would be more successful in achieving its objectives if it were to put more emphasis on helping students build sound intuitions about variation and its relevance to statistics (Ballman, 1997). The study describes a nontraditional path to statistics instruction that has variation as its central tenet. The experiences and insights gained from adopting such an alternative path in a college level, introductory statistics classroom are reported.

The current study is part of an ongoing research effort to understand the obstacles to the learning of statistics and use this understanding to find ways to create learning environments that facilitate deeper understanding. It builds on a previously conducted study which compared the learning experience of a group of students from a technology-based introductory statistics course following the

PACE (**P**rojects-**A**ctivities-**C**ooperative Learning-**E**xercises) approach developed by Lee (1997a), with that of a group of students with non-technology based instruction. Findings from that previous study (Lee, 1997b; Meletiou, Lee & Fouladi, 1998; Meletiou, Lee & Myers, 1999) indicated that the use of technology had a positive impact on PACE students' motivation and appreciation of statistics, and gave them increased familiarity with the practical aspects of the subject. At the same time however, confusion about the nature and purpose of statistics was observed in both groups of students. Those findings agreed with the main findings of the considerable research literature done in the last thirty years in the area of probability and statistics education. According to this literature, people (i) tend to believe that any difference in means is significant, (ii) have unwarranted confidence in small samples, (iii) have insufficient respect for small differences in large samples, and (iv) underestimate the effect of variation in the real world (Landwher, 1989; in Shaughnessy, 1992).

This led me to conjecture that students' difficulties in comprehending statistical concepts might be due to instructional neglect of variation. As Wild and Pfannkuch (1999) point out, understanding of variation in data includes comprehension of the following ideas: (1) variation is an observable reality; (2) some variation can be explained; (3) other variation cannot be explained based on current knowledge; (4) *random* variation is the way in which statisticians model unexplained variation; (5) this unexplained variation may in part or in whole be produced by the process of observation through *random sampling*; (6) *randomness* is a convenient human construct which is used to deal with variation

in which patterns cannot be detected. Traditional approaches to statistics fail to help students develop understanding of these ideas. Although variation is a critical issue throughout the statistical inquiry process, from posing a question to drawing conclusions (Pfannkuch, 1997), a tendency to ignore variability is hidden in standard approaches to statistical inference (Biehler, 1994). As a result, students do not develop the skills necessary to recognize uncertainty and variation and to distinguish among the different types of variation.

The research literature also neglects variation. There is an almost complete absence of research on variation. Truran (1994) points out that, whereas there has been much investigation of people's understanding of randomness, there has been very little investigation of their understanding of the variability arising when groups of outcomes are observed. Loosen, Lioen, and Lacante (1985) and Batanero, Estepa, and Godino (1994) have also noted the absence of research on variability and the overemphasis that statistics textbooks seem to put on looking at centers in data rather than on variability.

In addition to seeing as problematic the lack of research on variation, I also became dissatisfied with the prevailing methodology employed by most researchers examining people's conceptions of data and chance. This methodology, situated within the misconceptions movement, has been very successful in documenting people's erroneous beliefs and conceptions about probability but has done a very poor job of documenting success (Shaughnessy, 1997a). The common practice is to take snapshots of students' thought processes by posing cognitive tasks to them in order to catalogue their misconceptions.

Little guidance as to how one might systematically research conceptual change is provided. There is hardly any information about the sources of students' difficulties. Rarely does one do any follow up of the students' initial thinking to watch for future transitions (Shaughnessy, 1997a).

The conclusion I drew after reviewing the literature was the same as that drawn by Shaughnessy (1997a). Reflecting on the recent literature on statistics education, Shaughnessy concluded that three areas of opportunity for research on the teaching of probability and statistics that have gone largely unnoticed are:

(i) investigating students' thinking on variability

(ii) posing research questions that begin with what students can do rather than pointing out what they cannot do, and

(iii) following up on students' initial thinking to watch for future transitions

I decided to conduct a study that would provide insights into all three of these areas. The ultimate goal of the study would not be to add to the plethora of existing research documenting people's difficulties with probabilistic phenomena, but to document their successes, to show that "research on variability is an untapped well in research on data and chance" (Shaughnessy, 1997a, p. 137). By investigating introductory statistics students' intuitive understanding of variation and using the knowledge acquired to design, implement, evaluate, and refine some meaningful interventions, this study would aim at helping students develop and expand upon their understanding of variation. I conjectured that if I provided learners with an environment where they experienced the omnipresence of

variation and came to value statistical tools as a means to describe and quantify that variation, I would help them develop statistical thinking that goes beyond the superficial knowledge of terminology, rules and procedures.

In designing and implementing the study, I took the stance that students' conceptions are *transitional* conceptions rather than *mis*-conceptions, their thinking is always under construction (Shaughnessy, 1997a). Since for me intuitions are dynamic rather than static, the nature of evidence in research on learning is also dynamic. The researcher's goal, for me, should be to do "research on the *process* of learning" (Shaughnessy, 1997a, p. 131) and use the data obtained to develop and successively refine curricula. The *transformative and conjecture-driven research design* developed by Confrey and Lachance (1999) as a response to the need for establishing a better connection between educational research and practice was much more suitable to my research purposes than more traditional research models. Just as traditional probability and statistics instruction with its emphasis on formalism fails to establish links with students' intuitive thinking, traditional positivist research methodology has been unsuccessful in helping understand the real reasons for people's impoverished probabilistic and statistical reasoning. The conjecture-driven research model, which sees research and practice as interwoven, and advocates curriculum construction based on an ongoing process of development and feedback, was much better suited for expanding my understanding of the components that promote development and growth of students' understanding.

The study took place in the summer of 1999 in an introductory statistics course at a mid-Midwestern university. I worked jointly with the instructor Dr. Lee – who was also a major collaborator in the research I had previously conducted – towards designing and implementing learning experiences that aimed at improving students' intuitions by raising their awareness of variation. The format of the course was such as to encourage the kind of instruction that extends students' responsibility for their own learning in order to make learning meaningful and promote active knowledge construction. In such an environment, just as content could not be fully captured, learning goals could not be fully pre-specified. The conjecture-driven model permitted both curriculum and conjecture to be revised in light of student responses. Through close listening of the study participants, it offered more than numbers and flat descriptions, it was "able to capture the voices of many" (Nau, 1995), to provide "thick description" (Geertz, 1973) of the classroom setting and the interactions within this setting. The insights it provided led to an ever growing "understanding of themes, patterns, and meanings within context." (Beard, Schmitz, and Domahidy, 1997)

Embracing a growth-and-change view of intuitions allowed us to use the results of the existing literature and our own research, not as proof of innate limitations in students' ability to reason about the stochastic, but as a signal of the areas for which intuitions needed to be strengthened. By identifying similarities and differences between students' informal intuitions and formal statistical reasoning, we were able to work with students' intuitive notions and help them

develop ways to map new and richer concepts onto the ones that they already possessed (Mokros, Russell, Weinberg, and Goldsmith, 1990, 15).

The nature of the study helped me identify the kinds of intuitions students use to make sense of stochastic phenomena and the ways in which their intuitions are shaped by the learning environment. It allowed me to find out the structures that facilitate, as well as those that inhibit, the articulation of intuitions about the stochastic. In contrast to the heuristics research that took snapshots of how people might make sense of stochastic phenomena at a specific point in time, I was able to gain more insights into students' thinking by examining how their intuitions evolved during the course.

At the time the study was conducted, little was known about people's understanding of variation. Employing the conjecture-driven research design has helped bring the learners' voice to the front. By giving validation to the students' personal voice (Confrey, 1991), not only students' notions about variation changed, but my notions were also enriched with varied and ingenious insights offered by the students. By reporting in this thesis the findings of the study, I aim to contribute towards the development of alternative approaches to the notion of variation than the sterile ones dominating both the curriculum and the research literature (Shaughnessy, 1997a).

OUTLINE OF DISSERTATION

The dissertation proposal is organized into six chapters. In the first chapter, an overview of the study was provided. Below is a summary of each subsequent chapter:

Chapter II: Literature Review

This chapter gives a review of the research literature that has provided the framework for the study. I focus on work on the role of variation in statistics, research on statistical reasoning, and work on the role of technology. I also outline some specific aims of the study that emerged out of the review of the literature.

Chapter III: Theory and Methodology

In this chapter, I describe how the conjecture was developed and how it was linked to classroom practice. I also provide an overview of the philosophical foundations underlying the conjecture-driven design model and outline how this approach was employed in the study in terms of research design, data collection, data analysis and rigor.

Chapter IV: Assessment Prior to Instruction

In order to ensure that instruction is adapted to students' existing experience and their pre-knowledge, and also to be able to follow students' conceptual development process, good understanding of their thinking prior to instruction is required. In Chapter IV, I outline the findings from the pre-assessment on variability given on the first day of class and the follow-up

interviews of the primary group. I then discuss how the insights gained led to elaboration of the conjecture and, consequently, the instructional program.

Chapter V: The teaching experiment

In Chapter V, I give a brief description of some teaching episodes and class activities, which are characteristic examples of how the course was organized and how the meaning of main statistical concepts was constituted in social interaction. I also give some examples of how the continuous monitoring, both formal and informal, of student thinking shaped instruction. I also outline and discuss the findings from the assessment given at the end of the course and the follow-up interviews of the primary group.

Chapter VI: Conclusions

This final chapter summarizes the findings of the study and discusses how experience with the setting led to a much better understanding and further refining of the conjecture. It also discusses the implications of this research for statistics learning and pedagogy and examines future research directions.

Chapter II: Literature Review

In this chapter, I consider some existing literature that has provided a framework for my study. This literature includes work on the role of variation in statistics, research on statistical reasoning in general and variability in particular, and work on the role of technology in statistics education. I also outline some specific aims of the study that emerged out of the review of the literature.

RESEARCH ON STATISTICAL REASONING

I give in this section some analysis of why research on variability has been neglected. I then provide an overview of the research literature on students' understandings of samples and centers, and of a very influential body of research on people's reasoning when making judgments under uncertainty that has come to be known as the heuristics literature. Although not having students' notions of variability as their main object of study, these two bodies of literature have offered me some useful insights into people's thinking about variability.

Neglect of Variation

Shaughnessy (1997a) ponders about the almost complete absence of research on variability. He notes that, in contrast to the varied and extremely rich models of central tendency found in the literature, sterile approaches to the notion of variability dominate not only the curriculum, but the research literature also. One possible reason put forth by Shaughnessy may be that research often mirrors the emphases in curricular materials. There is, in the US curriculum in data and chance, a lack of focus on variation and an overemphasis on center. Another

reason may be the over-reliance of many statisticians on standard deviation as *the* measure of spread, a statistic that is computationally messy and difficult for both teachers and curriculum developers to motivate to students as a good choice for measuring spread. A third reason for the neglect of variability might be that, since centers are often used to predict what *will* happen in the future, or to compare two different groups, the incorporation of variation into the prediction confounds people's ability to make clean predictions or comparisons. Finally, Shaughnessy (1997a) concludes, "this whole concept of variability is outside of many people's comfort zone, and may even outside their zone of belief" (134).

Shaughnessy et al. (1999) have come to believe that there may be "an overemphasis in the teaching, assessing, and researching of students' conceptions of center to the detriment and neglect of their development of conceptions of spread or variability" (p. 2). They note the complete lack of items on dispersion, spread, and variation among the Statistical and Probability tasks of the 1996 National Assessment of Educational Progress (NAEP) as an example of this neglect. While there were several items on this assessment involving measures of center, there was only one, low-level computational item on spread. Not a single conceptual question involving statistical variation was included.

I have found only one study in the research literature focusing directly upon students' conceptions of variability. Shaughnessy, Watson, Moritz, and Reading (1999) describe an exploratory study they conducted to investigate elementary and high school students' understanding of variability. They gave a sampling task that was a variation of an item on the 1996 National Assessment of

Educational Progress, to 324 students in Grades 4-6, 9, and 12 in three different countries. Responses to the task, which was also used in this study and will be discussed in more detail in Chapter V, were categorized according to both their centers and spreads. The results from the study indicate a steady growth across grades on center criteria, but no clear corresponding improvement on spread criteria. Shaughnessy et al. (1999) speculate that the observed growth on center criteria might be the result of the emphasis placed on centers in the mathematics curriculum. They attribute the lack of clear growth on conceptions of spread and the inability to integrate the two concepts (centers and spreads) to the instructional neglect of variability.

Insights from Research on Sampling and Centers

Sampling is one of the main determinants of the validity of statistical inference. Because "statistical inference is almost by definition imperfect - all sampling introduces some error" (Jacobs, 1997, p. 4), students need to be aware of the potential threats to valid statistical inferences. A paper by Jacobs (1997) describes two studies that have investigated Grades 4 and 5 children's informal understanding of sampling issues in the context of interpreting and evaluating survey results. The findings show that while many of the children were aware of potential bias issues such as self-selection and restricted sampling and acknowledged the advantages of random sampling procedures, most children seemed to prefer stratified to simple random sampling. They were pre-occupied with issues of fairness and wanted to make sure all types of individuals were included in the sample.

Rubin, Rosebery, and Bruce (1990; in Hawkins, 1997b) observed that many students find it difficult to understand what it means for something to represent something else. They do not understand the "distinction between how a histogram is meant to represent a sample *accurately,* and how a sample is meant to represent a population *probabilistically"* (p. 10) and expect the distribution of the sample to look the same as that of the population. If this does not happen, they conclude that there must have been an error in the sampling process. Bar-Hillel (1982) found in her experiments that students described as "accurate" those samples with a statistic exactly matching the population parameter. Hawkins (1997b) reports encountering among students widespread confusion about the meaning of the statistical term *precision*, for many students often being "semantically indistinguishable from *accuracy*[1]" (p. 12). It is also often hard to convince students that when designing a survey or an experiment, one can manipulate its precision and should do so "if statistical and meaningful or practical significance are to be equated" (Hawkins, 1997b, p. 10), and that such manipulation is not cheating.

Rubin, Bruce, and Teney (1990), after having investigated the teaching and learning of statistical reasoning for several years, have come to the conclusion that grasping the basic concepts of sampling and statistical inference is extremely hard for students. Understanding, according to the researchers, seems to break

[1] Precision of a statistic describes its variation from the population parameter and has a magnitude that is a function of the population variation and the sample size. Students tend to confuse this statistical meaning of precision with the everyday notion of the word. Since the everyday notion of precision is synonymous to accuracy, they think that the two terms have the same meaning. Consequently, they consider as "precise" or "accurate" those samples whose statistic exactly matches the population parameter.

down “as soon as non-determinism enters the classroom” (Rubin et al., 1990, p. 2). In a study where they wanted to see how high school students conceive the relationship between sample and population, they found a tension existing between the ideas of sample variability and sample representativeness. On some instances, students’ comments suggested they believed that a random sample *has* to be representative and that not randomness but some other mechanism must have caused sampling variability. On different situations, however, which actually had the same underlying ideas, students “acted as if sample variability were the most relevant fact about sampling” (Rubin et al., 1990, p. 11).

The purpose of a study by Pfannkuch and Brown (1996) was to investigate the understanding of issues related to sampling of a small group of college students who had just completed an introductory statistics course. Students in Pfannkuch and Brown’s study seemed “oblivious” to probabilistic thinking for problems posed in real-world contexts and to have a minimal understanding of variation in small samples. However, when context was removed, students were comfortable thinking probabilistically. All the participants responded correctly, and “without equivocation”, to a typical coin toss problem and appeared to be, in this context, comfortable with the notion of long-run relative frequency. The authors interpreted the results of the study as a reflection of students’ lack of awareness of the role that variation plays in the social domain.

Garfield and delMas (1990) have also reported that students’ ability to solve problems involving random devices does not transfer very effectively to more applied problems. The two researchers created a computer program named

Coin Toss, which was designed "to repeatedly confront students' intuitions and assumptions" (Garfield and delMas, 1990, p. 6). They found that, when in the context of simulated coin tossing, many of the college students in their study seemed comfortable with the basic concepts of randomness, runs, and sample variability. When, however, the same students were given tasks that required them to apply their knowledge in solving problems that were not based on coin tossing, they were often unable to do so. Other researchers have also shown that students' understanding of probability is more limited in real-world contexts than in the contrived context of standard probability tasks. Some have speculated that the reason is the fact that use of real-world context increases the likelihood of prior beliefs and knowledge about the issues under investigation. Lord, Ross, and Lepper (1979) have suggested that individuals tend not to be equally critical of all sampling methods; they might not be as critical of sampling methods that result in conclusions consistent with their prior knowledge and beliefs as of studies that contradict them.

While the results of the heuristics literature that will be discussed in the next section suggest that students do not have a good understanding of the law of large numbers because they tend to ignore sample size, a study by Nisbett, Krantz, Jepson, and Kunda (1983) made contradictory claims. According to their findings, even people with no formal statistical training are able to use the law of large numbers in real-world situations. In addition, the ability to apply the law of large numbers can be enhanced by training. In order to resolve the contradictory evidence, Well, Pollatsek, and Boyce (1990) did a series of experiments of

statistically naïve college students, which showed that students' intuitive understanding of the law of large numbers is not a simple task and that task variables have a big effect on performance. Students were presented with different versions of the problems. In the "accuracy" version, they were just asked whether the average from a large sample or that from small sample would be closer to the population average. In the "tail" version, however, they had to estimate how likely it is that the sample average was a certain distance from the population mean. Students tended to do well on the "accuracy" version and very poorly on the "tail" version. In a follow up experiment, although students generated and observed computer graphic displays which showed the distribution of 100 samples of size 10 and 100 samples of size 100, many students still believed that the variability will be the same in both sampling distributions. Shaughnessy (1992) interprets the finding of the study by commenting that students' poor understanding of the law of large numbers is the result of the little attention paid to variability issues when teaching about sampling distributions:

> Without explicit teaching on the concepts of variability in sampling distributions, and how sample size affects this variability, students will not improve in their understanding of the law of large numbers. Mere exposure to the graphic displays of sampling distributions is probably not enough. Someone must explicitly point out the patterns in variability in sampling distributions, and the quantitative relationships that are involved. The fact that there are numbers involved in the tail version in the Well study makes it significantly more difficult than the accuracy version. (p. 478)

Biehler (1997) has found that when students make interpretations of summary statistics, they generally lack the flexible knowledge and critical

awareness required to take into account their robustness, or reliability and to realize the importance of sample size even when data do not come from a random sample. When for example, examining a box plot, they fail to realize that the interquartile range tends to be quite robust to outliers that can be the result of either individual inaccuracies or errors in the data. They do not appreciate that "a difference of one hour in the interquartile range has to be taken more seriously than the difference of one or two hours in the range or whisker differences, except for very small sample sizes" (Biehler, 1997, 181). He argues that this lack of understanding is observed because instruction does not encourage students to explore variability issues. Students who have been exposed to properties of distributions such as skewness and symmetry through the demonstration of ideal mathematical distribution curves might not be able to appreciate the boxplot's capability to show these properties by displaying several measures of spread.

The essence of the statistical perspective for Konold, Pollatsek, Well, and Gagnon (1997) is "attending to features of aggregates as opposed to features of individuals" (p. 151). Statistical estimates such as mean percentage of marriages leading to divorce or life expectancy, refer to features of the aggregate and not of individual elements. Although they might be used to make individual forecasts, since there is natural variability in every statistical endeavor, what they really give information about are group tendencies or propensities. The authors suggest that some of the difficulty people have in formulating and interpreting statistical arguments and making statistical comparisons are due to their tendency to think about individual cases or about homogeneous groupings rather than about group

propensities. They report the results of a study with high school students who had just completed a year -long course in probability and statistics to make the point that this tendency might be the result of an avoidance strategy when having to deal with variability issues:

> Our findings suggest that one reason it is easier to think about attributes of objects as opposed to attribute spaces, or dimensions, is that in focusing on attributes one can circumvent the issue of variability. Once there is no variability in collections of values, one can use nonstatistical methods of comparison. (Konold et al., 1997, p. 160)

Konold (1998) argues that to make the "conceptual leap required in moving from seeing data as an amalgam of individuals, each with its own characteristics, to seeing data as a group with emerging properties, properties that are often not evident in any individual members" (in NCTM, 1998, p. 70), students need to experience variation personally.

Insights from Heuristics Literature

The 1970s and 1980s saw the development of a very huge and influential body of research that has examined the informal strategies or heuristics people use when making judgments about the stochastic. This literature has been greatly influenced by the pioneering work of Daniel Kahneman and Amos Tversky (Kahneman and Tversky, 1973; 1982; Tversky and Kahneman, 1973; 1974; 1983). Kahneman and Tversky were the first to discover a number of systematic and persistent errors people – even ones with substantive formal training in probability – commit when attempting to make decisions concerning stochastic events.

Kahneman and Tversky's work suggests that even people who are able to correctly compute probabilities tend to rely on incorrect intuitions when asked to make inferences about uncertain events (Garfield, 1998). Several other researchers have subsequently studied the heuristics that people employ when making judgments of chance. The findings of most of them have confirmed Kahneman's and Trersky's findings and have uncovered additional errors of reasoning when analyzing statistical information (e.g., Cohen, 1979; Nisbett and Ross, 1980; Nisbett et al., 1983).

Garfield (1998) summarized the findings of the heuristics literature. She put the identified errors in statistical reasoning into the following categories: (i) misconceptions involving averages, (ii) the outcome orientation, (iii) good samples have to represent a high percentage of the population, (iv) the law of small numbers, (v) the representativeness misconception, and (vi) the equiprobability bias. Next, I briefly describe each of these of these categories of errors in statistical reasoning as they relate to my topic of interest.

Misconceptions involving Averages

People tend to ignore the possibility of outliers and to think that to find the center of a dataset, one should always add up all the numbers and divide by the number of data values. When asked to compare groups, they focus exclusively on the difference in averages, not considering variability.

The Outcome Orientation

This approach was identified by Konold (1989), who has noticed that people tend to interpret in deterministic terms phenomena that are actually stochastic. Lacking awareness of the stochastic dimension of these phenomena, they make predictions based solely on causal factors (Pratt, 1998). They assign probabilities by focusing on single events rather than looking at a series of events, and disregard frequency information, treating outcomes as either happening or not happening. They deal with uncertainty by predicting what the next outcome will be and then by evaluating the prediction as either right or wrong. A probability of 50% is often assigned when no sensible prediction is possible. Thus, for people adopting the outcome approach, the information that there is a 50% chance of rain tomorrow sounds totally useless, a probability of 30% implies that there is no possibility of rain, whereas a probability of 70% means that it will definitely rain. Konold (1989) does not consider the outcome approach to be a belief system. His extensive research of mainly undergraduate students learning probability and statistics has led him to conclude that individuals tend to switch between different approaches depending on the context of the situation. However, it does seem to him that some people are more likely to adopt the outcome approach than others are.

Good Samples Have to Represent a High Percentage of the Population

No matter how large a sample is and regardless of how well it was chosen, it cannot for many people be a good sample unless it represents a large percentage of the population.

The Law of Small Numbers

People often tend to think that small samples should resemble the populations from which they are sampled and use them as a basis for inference and generalizations (Kahneman, Slovic, and Tversky, 1982). They mistakenly apply the "law of large numbers" to small samples and show unwarranted confidence in the validity of conclusions drawn from small samples. As Shaughnessy (1992) points out, for statistically naïve people the effect of sample size on variation is not a factor to be taken into account. To them, it is not apparent that extreme outcomes are more likely within small sizes.

The Representativeness Heuristic

People who use the representativeness heuristic when making likelihood judgments, estimate the likelihood of an event based on how closely it resembles the population. They "look for an ideal type that represents their answer and then judge probability by closeness to this type" (Wilensky, 1993, p. 22). Thus, they end up predicting the outcome that appears as the most representative under the circumstances (Kahneman and Tversky, 1973; Tversky and Kahneman, 1983). The representative heuristic is a strategy that often is very helpful and leads to valid conclusions. At other times, however, it leads to serious biases even by people with sophisticated knowledge of probability. Some variations of the representativeness heuristic that Kahneman and Tversky have identified include (i) insensitivity to prior knowledge of outcomes, (ii) insensitivity to sample size and (iii) local representativeness.

Insensitivity to prior knowledge of outcomes refers to the phenomenon where people's tendency to look for similarities between outcome and evidence might be so strong that they ignore base-rate frequencies or prior distributions when making their judgment. Kahneman and Tversky (1973) showed through different tasks they gave to their study participants that prior probabilities were often ignored, and saw this as evidence of the use of the representativeness heuristic which dominates any sensitivity to prior probabilities.

When making judgments about the probability of an outcome involving a sample, people often tend to consider the population as a whole and show *insensitivity to sample size*. A classic example of this tendency is people's response to the following question, first given by Tversky and Kahneman (1974) to a group of undergraduates:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the number of days on which more than 60% of babies born were boys. Which hospital do you think recorded more such days? (p. 1125)

Ignoring the difference in the average number of children born everyday in each of these two hospitals, most of the people participating in the study judged that both hospitals had the same probability of obtaining more than 60% boys. Several other researchers have given this or similar questions and have obtained similar results. Kahneman and Tversky consider this as evidence of the

representativeness heuristic; the two events are equally likely since the data provided is equally representative of the general population.

Local representativeness is the phenomenon where "people believe that a sequence of events generated stochastically will represent the essential characteristics of that process, even when the sequence is quite short" (Pratt, 1998, p. 37). For example, when tossing coins, people consider it less likely to obtain HHHTTT or HHHHTH than to obtain HTHTTH, because HTHTTH seems to better represent the two possible outcomes. Similarly, the fallacy of the gambler who, after a long sequence of red outcomes, expects the next outcome to be a black is, for Kahneman and Tversky, the consequence of employing the local representativeness heuristic and perceiving a pattern in random data. The gambler's fallacy is also called the "law of averages" because it describes people's tendency to believe that things should balance out to better represent the population distribution. This is the same idea as that which Shaughnessy (1992) calls an active balancing strategy. For him, an active balancer is the person who, when given the problem "The average SAT score for all high school students in a district is known to be 400. You pick a random sample of 10 students. The first student you pick had an SAT of 250. What would you expect the average to be?" (p. 477), would predict the average of the remaining 9 scores to be higher than 400, in order to make up for the "strangely" low score.

The Equiprobability Bias

People tend to assume that all the different possible outcomes of a possibility space are equally likely (Lecoutre, 1992; in Pratt 1998, p. 44). They,

for example, assume that when shaking two dice, it is equally likely to get a 9 and a 7. Such incorrect notions are resistant to change and are often exhibited by people well acquainted with probability theory.

Although not directly studying people's understanding of variability, the heuristics literature does point to people's tendency towards to the deterministic and their limited understanding of variability. Landwehr (1989; in Shaughnessy, 1992) summarized people's over-reliance on averages and their lack of understanding of variability when he noted in his list of common stochastic misunderstandings that people:

(a) have the misconception that any difference in the means between groups is significant

(b) inappropriately believe there is no variability in the "real world"

(c) have unwarranted confidence in small samples

(d) have insufficient respect for small differences in large samples

Criticisms of the Heuristics Research

Two inferences according to Wilensky (1993) can be drawn from the heuristics literature: either that people make errors when judging under uncertainty because (i) the human brain is "hard-wired" not to be able to think intuitively about the stochastic, or because (ii) intuitions are insufficiently developed to generate probabilistic judgments in a more sophisticated way than that observed. Both views have important ramifications for probability and statistics instruction. If one adopts the first view, then the implication is that probability concepts should be taught in a formal way with no connection to

everyday intuitions. Adopting the second view does pose the question of what learning environments should be used to help sufficiently develop students' intuitions.

Pratt (1998) raises several criticisms of the heuristics literature. One such criticism is the lack of a theoretical framework characterizing this work, which leads to contradictions and confusions. To make his point, he notes the contradictions in the rationale that Tversky and Kahneman claim drives their focus on people's systematic errors:

> They claim that their approach:
>
> - exposes our intellectual limitations and suggests ways of improving the quality of our thinking;
>
> - errors and biases reveal the psychological processes that govern judgment and inference;
>
> - mistakes and fallacies help the mapping of human intuitions indicating which principles are non-intuitive or counter-intuitive. (Pratt, 1998, p. 51)

As Pratt (1998) indicates, whereas the first reason seems to be consistent with the second inference which gives room for improving students' intuitions through the development of better learning environments, the third reason presents a view of intuitions that is very static. Indeed, the literature on heuristics has been interpreted by many people as indication that the human mind is not able to generate accurate judgments concerning stochastic phenomena (Wilensky, 1993).

Pratt (1998) goes on to criticize the methodology that Tversky and Kahneman employed. He argues that if their aim had not been to expose the

limitations of the human brain but to develop a theoretical framework of change in intuitions then, instead of taking "snapshots" of students' intuitions, they should have looked closely at the psychological processes involved. In addition, Pratt (1998) considers as problematic also the fact that the heuristics literature ignores the influence of the setting on the shaping of intuitions:

> If our ultimate aim is to apply research findings to teaching and learning contexts, then it makes no sense to remove setting from the experiment. Indeed, if it were true that we could infer from the heuristics research that setting was not influential (and I do not see how we can possibly make this inference when setting was intentionally ignored) then we are in a position of no hope from a pedagogical perspective. (p. 35)

Lave's (1988) theory of situated cognition, which draws heavily from anthropology and sociology, also sets as problematic the practice of researchers on heuristics to take snapshots of how people might make sense of stochastic phenomena at a specific point in time, without considering the role of context. Her research findings, as well as those of other researchers that have investigated the relationship between situation and cognition (Hiebert and Carpenter, 1992; Brown, Collins, and Duguid, 1989; Nunes, Schliemann, and Carraher, 1993), have indicated that learning is highly specific. They have cast doubt on the validity of research that considers culture as a factor exerting a constant influence on the learner's cognition and have caused a shift of cognitive analysis "from focusing on the abstract, disembedded character of internal representations to describing cognitive actions within the situation." (Hiebert and Carpenter, 1992, p. 76)

Another weakness of the heuristics research according to Pratt (1998) is its emphasis on fallibility, which leaves the reader with the impression that human actions essentially lack rationality. Lopes (1991) also criticizes Kahneman and Tversky's research, arguing that it caused, during the 1970s, a shift in view of humans from one of effective decision makers to one of ineffectiveness. He does not refute the claim that people use heuristics when making decisions concerning stochastic phenomena, but criticizes the emphasis of this literature on the fallibility of judgments and the unfair image it provides about the frequency in which these poor judgments occur. The focus of this research, Lopes argues, is not to identify heuristics but to show the bias in these heuristics, to emphasize errors and irrationality. The language employed is evaluative rather than neutral. The impression it leaves about humans' judgment is a very bleak one and the assistance it offers as far as pedagogy is concerned is minimal.

The heuristics approach is part of the misconceptions movement of the 1960s and 1970s which aimed at discovering fallibilities in students' reasoning. Confrey (1991) raises serious criticisms of this movement. She points out that too often researchers fall into the misconceptions trap and fail to acknowledge the potential legitimacy of students' novel constructions. This leads to the suppression and destruction of student "voice". Smith, diSessa and Rochelle (1993) consider research on misconceptions to be an ineffective way of examining human learning that is applied in a very narrow range of contexts, has little explanatory power, and provides no account of productive ideas that might facilitate learning. Although acknowledging that the misconceptions movement

has significantly advanced our understanding of student learning by producing detailed characterizations of the students' ideas, they see fundamental problems with its tendency to characterize most of students' ideas as misconceptions. Also, they reject the idea inferred from this literature, of replacing misconceptions by appropriate expert learning.

Instead of casting misconceptions as flawed ways of viewing the world that instruction ought to confront and replace, Smith et al. (1993) consider them as useful resources for the acquisition of new knowledge. They disagree with the misconceptions movement's overemphasis on discontinuities between novice students and experts because "it conflicts with the basic premise of constructivism: that students build more advanced knowledge from prior understandings" (Smith et al., 1993, p. 125). They argue that the road from being a novice is a continuous one and use case analyses to dispute some often cited dimensions of discontinuity and to identify important continuities previously ignored. Studies on misconceptions have been unfair to novices since they have tested them in areas unfamiliar to them and in situations where there were no appropriate tools available for them to explore the question. Such conditions make people feel uncomfortable and do not allow them to show their competencies. The authors' analyses indicate that novices do possess intuitive knowledge with abstract features and, given the right environment, they can employ abstract thinking. Hence - the authors suggest - if we find ways to put students in learning environments closer to their own experiential world instead of situations outside their area of competence, we might be able to observe not

failure, but effective search for underlying principles to make sense of the stochastic.

Smith et al.'s (1993) model of cognitive development is one of "refinement and reorganization, rather than replacement" (p. 116). Novice conceptions are not only flawed but productive also, and they can support cognitive growth. The move from novice to expert is a continuous process of transforming and refining of prior knowledge into more sophisticated forms. It is therefore much more useful to study the productive role that novice conceptions continue to play in expert knowledge than to cast them as flawed and try to replace them, something which is impossible anyway, since learning is both constrained but also made possible by prior knowledge. Emphasis should be given not on misconceptions, but on intuitive knowledge that characterizes not only novice but expert knowledge also.

A group at the 5th International Conference on Teaching Statistics (ICOTS V, 1997) was devoted to discussing the state of empirical research on the teaching and learning of probability and statistics. The group noted that although quite a lot of research has been done on students' conceptions and beliefs about chance and data, this research seems to have put too much emphasis on the uncovering of misunderstandings. They raised the question as to whether what we are studying are "misconceptions" or "missed—conceptions" (Shaughnessy, 1997b, p. 219). The conclusion was that if "missed-conceptions" are what we are studying, then it is probably better for us to start thinking of our students' conceptions and beliefs about chance and data as being in transition. Consequently, we should

concentrate our research efforts on longer-range studies of students' growth in thinking over time rather than on snapshots of students' thinking.

A number of more recent studies suggest that, under certain conditions, people are in fact able to use rational statistical reasoning. Nisbett et al. (1983) identified three factors whose presence seems to help increase the possibility that people will employ statistical reasoning appropriately. The first factor is *clarity of the sample space and the sampling process*. People tend to use the representativeness heuristic in cases where the random generating process is not clear and the sample space not well understood, which is often the case in the social domain. They tend to employ statistical reasoning in tasks concerning randomizing devices, since these devices have an obvious sample space and the repeatability of trials can be easily imagined. The second factor is *recognition of the operation of chance factors*. The random nature of social events is often not as explicit as that of randomizing devices, although experience does help identify the role of chance. The third factor is *cultural prescriptions for statistical reasoning*. Cultural knowledge does have an effect on people's ability to reason statistically. In our days, European children are more capable of reasoning statistically than mediaeval children, because of the wide use of statistics in social domains such as sports. The continuous increase of the importance of statistics implies that people in the future might be able to reason more effectively about the stochastic than people today.

Jacobs and Potenza (1991) found that the kind of setting influences the use of heuristics. They witnessed a tendency not to make reference to frequencies

when making social judgments and this led them to distinguish between *social* and *object* judgments. In addition, a group of psychologists at the University of Chicago has been challenging some of the claims of the earlier psychological research, and has found that students tend to be more successful when questions are posed in terms of frequencies (Gigerenzer, 1994, 1996; Gigerenzer and Hoffrage, 1995; in Shaughnessy, 1997a).

The implication of these research findings on the influence of setting is that the nature of the learning environment in which students experience stochastic phenomena can have an important effect on the use of statistical reasoning. This implication leaves open the room for improvement of statistical intuitions, suggesting that by providing the appropriate environment we can increase students' awareness of the underlying random effects.

Moving Away from Misconceptions: Intuitions as Dynamic

The effect of the heuristics literature on probability and statistics education has been profound. However, as already noted, a growing number of researchers have lately become critical of this view. These researchers do not disagree that people's intuitions about probability and statistics often run counter to stochastic reasoning and that this disparity may be partly responsible for the difficulties they encounter in learning probability and statistics. However, they do not take this as indication of cognitive constraints regarding the stochastic, but rather as an indication that more needs to be done in order to deal effectively with intuitions. They believe that students are capable of statistical and probabilistic reasoning

and that their difficulties are primarily due to limitations in the learning methods, tools and cognitive technologies employed (Wilensky, 1997).

Shaughnessy (1992), reminds us that the heuristics people use are not necessarily bad and they do not always result in biases but often give good information. We should not forget that, for example, representativeness is fundamental to the epistemology of statistical events. Being "the very reason we try to draw a random sample from a population" (Shaughnessy, 1992, p. 479), representativeness is the statistical idea which allows us to draw conclusions about the underlying population based on the sample. We should understand that "it is not that there is something wrong with the way our students think, just that they –and we – can carry the usefulness of heuristics too far" (Shaughnessy, 1992, p. 479). We should try to create a curriculum that builds on the strengths of students' conceptions while helping improve those intuitions that are counterproductive.

More research according to Shaughnessy (1997a) needs to be done on what students can do rather on what they cannot do. The problem he sees with most of the research investigating students' statistical reasoning is that the questions researchers have been asking are often posed in the wrong way which exposes what students cannot do rather than what they can do. For example, the problem on the number of boys born in a large vs. a small hospital mentioned earlier, causes lots of confusion because it actually focuses the attention of the responders on centers, while presumably dealing with the concept of spread. Shaughnessy (1997a) believes that if we want to see what our students *can* do

with variability, we should pose questions in a sampling context as a range of likely values, rather than in a context forcing students to compare the point value probabilities of two particular events. He describes one such question he gave to his students that had very encouraging results: “Imagine you have a huge jar of M&M’s with many different colors in it. We know that the manufacturer of the M&M’s puts in 40% browns. If you reached in and pulled samples of 20 M&M’s at a time, what do you think would be the likely range for the number of browns you found in your samples?” (p. 135).

Such types of questions can, for Shaughnessy (1997a), be a point of departure for the kind of instruction builds on students' intuitions in order to help increase their understanding of what the likely spread of outcomes is for a sample from a certain population. By focusing on tasks that elicit conceptions of variability and difference rather than centers and sameness, students can begin to appreciate more the fact that there is lots of variability in the real world but also that we are often still able to detect patterns in the data. They will gradually begin to get some idea of what is likely and what is unlikely to occur by considering the entire distribution of outcomes and to develop some sense of the effect of sample size on the spread of likely outcomes.

Borovcnik and Peard (1996) also believe that instruction should start taking students’ intuitions more seriously into account in order to “describe when and why students encounter problems and how teaching should be designed to intervene with inadequate conceptions” (p. 249). They assert that intuitions are

not necessarily bad but, on the contrary, they are a prerequisite to real understanding that transcends mere recollection of formulas and procedures:

> There is some sort of imagination which seems to act as a driving force when concepts are under construction, when mathematics is in statu nascendi. These intuitive ideas are too vague to be communicated. On the other hand, these abstract concepts seem not to be understandable without sharing similar images. There seems to be an indissoluble interrelation between intuitions (intuitive ideas, intuitive conceptions) and theory (abstract models, concepts). It is not possible to separate these two aspects of the "object". (Borovcnik and Peard, 1996, p. 248)

Borovcnik (1990) perceives intuitions not as something static but as something dynamic that undergoes continuous change. He distinguishes between "primary intuitions" and intuitions that emerge due to instruction. Primary intuitions "have a longlasting effect on our cognitive behavior" which can facilitate or hinder the reconstruction of concepts. Consequently, it is vital for instruction to establish a direct link between primary intuitions and abstract concepts. This will be very beneficial for students both for getting them motivated but also for improving their understanding of abstract concepts. Probability and statistics education will not be successful in helping students "get their world of vague intuitions ordered" and adequately understand abstract concepts, unless it develops a dynamic interplay between the intuitive and theoretical level. Instruction ought to start with the students' primary intuitions and try to develop secondary intuitions that will help raise their awareness of "probabilistic interpretation", and will allow them to understand how stochastic thinking is related to causal and logical thinking.

Borovcnik and Peard (1996) have, along with others, argued that one reason for the underdevelopment of probabilistic intuitions might be the lack of consistent feedback from stochastic phenomena. Since concrete operations are not available when making judgments about uncertainty, the learning process is handicapped and leads to the domination of unreliable intuitions and to causal thinking. Conventional instruction often fails to establish enough links between the learners' primary intuitions and "the clear cut codified theory of the mathematics" (Borovcnik and Bentz, 1991; in Pfannkuch and Brown, 1996). Even if people understand probabilistic theory, we often see them falling back into the trap of causal thinking because, Borovcnik and Peard (1996) remark, "conceptual thinking can be reduced neither to mathematics nor to its applications. To justify a concept necessitates modes of thought different from those required to understand the concept" (Borovcnik and Peard, 1996, p. 243). To make their case, they give the example of the concept of independence. Although independence is mathematically reduced to the multiplication formula and "gains an important role within theory", becoming the basic ingredient of theorems such as the law of large number or the Central Limit Theorem, "its mathematical definition cannot affect the causal ideas individuals relate to it in general." (Borovcnik and Peard, 1996, p. 243)

Borovcnik and Peard (1996) stress that "learning should be more than simply remembering what one has been told" (p. 247) and that the individual formation of concepts cannot be provoked by a hierarchical sequence of actions and reflections. They also warn us that traditional instruction has underrated the

complexity and dangers of using pseudo-real examples: “If the required mapping onto an artificial context conflicts with emotions or with common sense, then a breakdown of teaching is likely.” (p. 274)

Intuitions for Fischbein (1987) play an essential role in the acquisition of new knowledge, they are the key to the understanding and acceptance of theory. He believes that, given the right environment, intuitions can be developed in ways that can make them a very powerful tool in guiding thinking and plans of action. Fischbein (1975) indicates that intuitions about relative frequencies exist from a very early age but they are suppressed by schooling in favor of deterministic thinking. In order to make his point, he reports on a study where participants were asked to predict the outcomes of a repetitive series of stochastic trials, and where even young children were able to make predictions based on the relative frequencies of the different outcomes. The reason that the intuition of chance remains outside of intellectual development is the emphasis of schooling on causality and determinism and its sole focus on deductive reasoning. Due to the lack of nourishment of probabilistic intuitions, learners develop a series of heuristics often subject to bias, in an effort to rationalize stochastic events.

Fischbein (1975) argues that powerful probabilistic intuitions will not develop simply by practicing probabilistic formulae. Neither is it enough to engage students in activities such as throwing dice, playing computer games, or watching sports for probabilistic reasoning to develop. Although unpredictable events happen constantly in these activities, the structure of the activities is such that the underlying probabilistic principles lie hidden. Stochastic experiences

should be structured in ways that make the underlying probabilistic ideas explicit by actively engaging the learner throughout the meaning-making process.

Wilensky (1993) maintains that probabilistic intuitions are not innately made but are constructed. The lack of sound probabilistic intuitions is not the result of inherent limitations but of lack of concrete experiences from which such intuitions can develop. Although many of the results of Tversky and Kahneman hold for many people today, it is unwise to conclude that we are inherently incapable of thinking intuitively about probability. The process of building intuitions takes time and "is not the kind of change that can be measured by a pre test and then a post-test after being instructed in the correct approach" (Wilensky, 1993, p. 187). Also, because of the relative stability of intuitions over time, people tend to forget that these intuitions have developed and often mistakenly assume that "what is intuitive is built-in, is always there, and does not change" (Wilensky, 1993, p. 187). To make his point, Wilensky gives an analogy with conservation experiments of Piaget. In such experiments, young children are typically shown two full water glasses, one taller and one wider glass and are then asked which glass has more water. Children before a certain stage (usually around age 7) say that the tall glass has more. If the tall glass is then poured into another wider glass and fills it, then the children say the two glasses have the same amount of water, but if the water is poured back into the tall glass, they again assert that the tall glass contains more water. A year later, these same children when shown videotapes of their earlier interviews, cannot believe how they could have possibly made such a "stupid" mistake. Students' difficulties

with the stochastic might parallel those of the young children who think that the taller glass has more water. Given the right resources, these same students might end up constructing probabilistic intuitions so sound that their previous difficulties will seem as distant to them and as hard to acknowledge as they seemed to the older children in Piaget's experiments.

Pratt (1998) perceives intuitions as complex and dynamic, "sense-making devices", which "are not static but constantly change in the light of experience, cued by aspects of the setting and shaped by the experience of using those intuitions" (p. 126). He argues that "the notion of misconception ignores the potential for those same intuitions to act as a springboard for successful sense making" (Pratt, 1998, p. 342). His dissertation study suggests that to make sense of long-term behavior of random phenomena, children initially use local meanings such as unpredictability and unsteerability. Although these local meanings appear to be misconceived, Pratt considers them as very important since they are the only resources available to the child. He does not follow the example of the heuristics approach that would "add these misconceptions to the ever growing list of ways in which people, in this case children, behave irrationally when making judgments in the stochastic domain" (Pratt, 1998, p. 341). Rather, he shows how the children in his study used the tools he designed based on the close investigation of their thought processes to make sense of the long-term behavior of stochastic phenomena utilizing local meanings. Thus, he concludes, local meanings "may have the characteristics of misconceptions when the circumstances lie beyond the child's area of competence, but, set in a carefully

designed domain of abstraction, they become the raw materials of new meanings" (Pratt, 1998, p. 342).

Konold (1995a) insists that the "forget-everything" approach probability and statistics teachers often embrace in the hope that what they convey to students will be accurately encoded does not work simply because "learning is both limited and, at the same time, made possible by prior knowledge". The only way people can cope with new information is to relate it to things they already know, since "there is no blank space in our minds within which new information can be stored so as not to "contaminate" it with existing information" (Konold, 1995a). We cannot overwrite students' beliefs and intuitions with more appropriate ones. The only solution to explain and bridge "the frequent gap between what students report and what we, as teachers, thought we clearly communicated" (Konold, 1995a), is to use assessment methods that help reveal the intuitions that hinder learning and then find ways to improve them.

Beliefs about the Nature of Mathematics: Impact on Statistics Instruction

Wilensky (1993) claims that the failure in developing sound probabilistic intuitions is similar to other failures in mathematical understanding and is the result of deficient learning environments and reliance on "brittle formal methods". It is, in my opinion, these same reasons which also cause the neglect of variability observed both in the curriculum and in the research literature. Deep-rooted beliefs about the nature of mathematics are imported into statistics, affecting instructional approaches and curricula and acting as a barrier to the kind of

instruction that would provide students with the skills necessary to recognize uncertainty and variability and be able deal intelligently with them.

Formalist vs. Relativistic View of Mathematics

In recent years, the "formalist" tradition in mathematics and science has come under attack and a second agenda, which aims at "the preservation of meaning in mathematical statements" (Wilensky, 1993, p. 27), has begun to emerge. This interpretive view sees mathematics as a meaning-making activity of a society of practitioners. The emergence of the new paradigm has been the result of developments in the history and philosophy of science which have caused a general shift, in the last thirty years, of virtually every social science and field of humanities away from rationalistic, linear ways of thinking. In the social sciences several critics have attacked formalist tradition in mathematics and science. Hermeneutic critics (Packer and Addison , 1989; in Wilensky, 1993, p. 26) have criticized it, among other things, for its detachment from context, its foundation on axioms and principles rather than practical understanding, and its formal, syntactic reconstruction of competence. Feminists have criticized it for alienating a large number of people and especially women. Sociologists such as Latour (1987) have maintained that science can only be understood through its practice.

Kuhn (1962) has argued that the progress of science rather than being linear and hypothetico-deductive as claimed by logical positivists, is made possible through revolution. New theories are not incremental modifications of existing ones but theories that posit basic entities of the world, which are fundamentally incompatible with old theories. An anomaly occurs, which is an

event that cannot be explained by the existing theory. In effect, the theory/paradigm is disproved. A whole new paradigm takes over, which explains everything the old paradigm explained and also explains the anomaly (Wiesman and Wotring, 1997). Polya (1962) and Lakatos (1976) have argued that, "by placing such a strong emphasis on mathematical verification and the justification of mathematical theorems after their referent terms have been fixed, mathematics literature has robbed our mathematics of its basic life" (Wilensky, 1993, p. 34). Mathematics for Lakatos is a human enterprise and advances happen through the negotiation of meaning among a community of practitioners. It is not given in advance but is constructed through the practices, needs, and applications of this community of practitioners. Proofs are not developed in a linear way, but follow "the "zig-zag" path of example, conjecture, counter-example, revised conjecture or revised definition of the terms referred to in the conjecture" (Wilensky, 1993, p. 34). In the new paradigm, the history of mathematics takes an important role. Its examination reveals that "mathematics is messy and not the clean picture we see in textbooks and proofs", that "the path to our current mathematical conceptions was filled with argument, negotiations, multiple and competing representations." (Wilensky, 1993, p. 17)

In response to criticisms following research findings and reports of the 70s and early 80s exposing students' impoverished understanding of mathematics and science, leaders and professional organizations in mathematics education are now finally promoting a relativistic view of mathematics (Confrey, 1980; Nickson 1981). They have come to believe that current teaching approaches are deficient

in that they do not give students the chance to encounter a number of different perspectives on the nature and uses of mathematics:

> It is difficult for us who have been socialized into the peculiar culture of university faculty to recognize how esoteric we have allowed university mathematics in particular to become. We imagine (incorrectly) that dominance of the abstract over the concrete, absence of ties to applications, and an emphasis on rigor over fluency of use are inherent in the discipline. We value, in Richard Feynman's words, precise language over clear language. Reformers urge a change of culture toward the concrete, toward applications, toward ability to *use* mathematical concepts and tools over rigor of detail. They offer pedagogical reasons, but they are also responding to the pressures of democratization. (Moore, 1997, p. 124)

Reformers argue that the culture of the mathematics classroom should change. Mathematics should be presented as open to discussion and investigation, as a socially constructed discipline which, even at the classroom level, "is not held to be exempt from interpretations that require "reconsideration, revision and refinement" " (Nickson, 1992, p. 104). The emphasis should not be "with mirroring some unknowable reality, but in solving problems in ways that are increasingly useful to one's experience" (Confrey, 1991, p. 136). The teacher should encourage discussion, and allow students to generate and test their own theories.

These ideas embraced by many members of the mathematics education society are influenced by the acceptance on constructivism as a learning theory. Although a wide spectrum of beliefs are covered under the label of constructivism, with traditional constructivists emphasizing individual thinking and creation of meaning (Piaget, 1970) and neo-constructivists incorporating more ideas about culture and social learning (Vygotsky, 1986), there are some

overarching ideas defining constructivism as it relates to instruction. In general, constructivism tends to be more holistic and less mechanistic than traditional information-processing theories. According to constructivism, people make sense out of their world by taking in information from the environment and assimilating it into their pre-existing schemas. This meaning-making process depends on prior knowledge as well as on interaction with others. Learning then is a personal interpretation of the world as well as a collaborative process with meaning negotiated from multiple perspectives.

Wilson, Teslow, and Osman-Jouchoux (1995) warn us that although recent models of cognition are clearly challenging our traditional notions of learning and teaching, changing long-held beliefs and attitudes towards mathematics is not easy. There is an enormous gap between the mechanistic-instrumental portrayal of the nature of mathematics and the more realistic-fundamental view that reform efforts try to advance. The formalist tradition has been around for too long and it runs deeply into people's veins. For people raised in this objectivist tradition, it is very difficult to accept the fallabilist nature of mathematics.

Impact of Formalist View on Statistics Education

In the statistics domain, there has already been a significant move towards modernizing statistics education and a general acknowledgment that learning occurs most effectively when students engage in authentic activities. Although many statistics students from higher institutions are still being taught in traditional classrooms, there is already a large number of statistics instructors who have

adopted alternative approaches to teaching statistics and many statistics classrooms are experiencing wide incorporation of technology. But, as Hawkins (1997a) points out, for reform efforts to be successful, it is "necessary not only to provide the infrastructure and finance to support technological innovations, but also to change attitudes and expectations about statistical education" (viii). The deep-seated beliefs of many people about the nature of statistics as "as a branch of the older discipline of mathematics that takes its place alongside analysis, calculus, number theory, topology, and so on" (Glencross and Binyavanga, 1997, p. 303), hamper the reform efforts.

The linear and hierarchical and approach adopted by statistical courses and syllabuses is testimony to the profound and continuing effect of the formalist mathematics culture on statistics. The structure of almost every introductory statistics course is to first start with descriptive and exploratory data analysis, then move into probability, and finally go to statistical inference. Biehler (1994) warns us that the danger of a curriculum with such a structured progression of ideas is that students get the impression that "EDA (Exploratory Data Analysis), probability and inference statistics seem to be concerned with very different kinds of application with no overlap" (Biehler, 1994, p. 16). This leads to compartmentalization of knowledge: "The degree of networking in some students' cognitive tool system seems to be rather low, otherwise the trial and error choice of methods that we observed quite frequently would be difficult to explain." (Biehler, 1997, p. 176)

In statistics courses, effort is often put on simplifying the process of learning by organizing it step by step, assuming that this helps to remove difficulties from students' path by gradually leading them from more basic to more complex connections (Steinbring, 1990). However, the linear and consecutive structure of the course comes in sharp contrast with "the complex nature of stochastical knowledge which can only be understood as a "self-organizing process" (Steinbring, 1990, p. 8). The static image projected through the formalization of the chance concept to probability is misleading and hides the dynamic and complex nature of chance events. It is inadequate in helping students make the conceptual shift that is needed to understand the difference between long-run stability and variation in finite samples (Biehler, 1994).

The assumptions posed are often too simplistic. Although not necessarily denying underlying causal explanations in case of chance events, a probabilistic approach views them as impractical and, "accepting a current state of limited knowledge, adopts a 'blackbox' model according to which underlying causal explanations, if not denied, are ignored" (Biehler, 1994, p. 10). As Biehler (1994) indicates, the assumption of independence is not plausible in many real-world contexts. Even "coin flipping can also be done in a way that independence has to be rejected in favor of serial correlation, and physical theories can be developed to explain some aspects of coin flipping" (Biehler, 1994, p.10). Similarly, Von Mises principle of an impossible game does not rule out the possibility of improving chance by observing variables from which the roulette result is not independent. As a matter of fact, people have actually constructed computer

programs that gather and process physical data such as the velocity and position of the ball and the wheel in order to make better predictions than those offered by just the uniform distribution.

In standard approaches to statistical inference, "distributions are reduced to the mean values and the question 'are the mean values different' is posed - assuming that distributions are equal in all other respects (normally distributed with known variance). The problem of whether a difference is statistically significant steals into the foreground, masking the basic conceptual question of the difference of distributions" (Biehler, 1994, p. 14). This is problematic and projects a static view of reality that does not take into account the fact that the world is continuously changing and distributions change too.

The over-emphasis of the traditional mathematics curriculum on determinism and its "orientation towards exact numbers" (Biehler, 1997, p. 187) affects statistics instruction, becoming an obstacle for the adequate judgment of stochastic settings. The law of large numbers is often presented as a canon in the statistics classroom, giving students the false impression that the stabilization of the relative frequency of repeated sampling to the ideal value is *guaranteed*. Similarly, instruction leaves students with the impression that a larger random sample *guarantees* a more representative sample. There is a deterministic mindset and an over-reliance on rules and theorems, forgetting that we are dealing with uncertainty, and the variability accompanying all finite statistical processes implies that a sample is almost never totally representative of the population from which it was selected. People have a hard time distinguishing between the real-

world problem and the statistical model. At one extreme are many people who use statistical methods for solving real-world problems in the same way that they would use an artificial mathematics problem coming out of a textbook. On the other extreme, we find people who distrust statistics completely due to the fact that, unlike mathematics, it deals with uncertainty. Both of these two extreme attitudes suggest inadequate understanding of statistics as a decision-support system (Biehler, 1997).

Technology, however, does have the potential to transform both the content and the pedagogy of statistics and to change people's ideas about the nature of statistics. There is currently an ongoing debate among statistics educators on the role of technology in statistics teaching and learning. In the next section, I will give an overview of the increasing literature examining the role of technology in statistics instruction.

ROLE OF TECHNOLOGY

Where is the knowledge that is lost in information?

(T. S. Eliot)

Technology can help statistics turn information into knowledge, by allowing us not only to do old things in new ways but, more importantly, by allowing things that were not possible before (Burrill, 1997a). The developments in computers during the last decades have been so profound that it is not surprising they have had an immense impact on the practice of statistics. The availability of computing technology has freed statistics from many of the constraints of the past and has radically transformed the culture of practicing

statisticians. The calculating power of the computer has relieved the burden of computations, and its ability to generate data from which to make conjectures and simulate the behavior of complex systems has opened new and exciting avenues.

Statistics is not a fixed subject, but one that is ever growing and changing as demands for its application become stronger and as technology enables us to think of new and more revealing ways to process information and make decisions. The different forms of representation and data analysis and the wider range of problems that computers allow, are turning statistics into a "data science" with close ties not only to mathematics but also to computer science and its related fields of application (Batanero et al., 1997). Computers have revolutionized the view of what statistical knowledge means and statistics instruction should adjust accordingly to accommodate the changing nature of statistics:

> Educational technology does afford us with a greater variety of strategies for teaching statistics. Moreover, it offers us new ways of doing statistics. Our education processes often reflect somewhat conservative (if not actually reactionary) ideas of what statistics is and how it should be taught. The changing nature of statistics is an ongoing challenge, often demanding quite radical reforms in statistical education. (Hawkins, 1997a, vii)

Although in the traditional curriculum, it is inconceivable to teach statistics independently of mathematics (Steinbring, 1990), there are now many other requirements in addition to mathematical knowledge for teaching statistics effectively. These include "organizing and implementing projects, encouraging work and cooperation between students, and understanding graphical representations, computation, and so forth, not as didactic tools, but as essential statistical means of knowing" (Batanero et al., 1997, p. 195). While some

statistics instructors still choose to ignore the need for change in their teaching objectives, many others have already responded to the calls for reform and have taken advantage of technology to create new learning environments that adopt to a more data-driven approach to statistics.

Technology has provided the opportunity to create an entirely new learning environment, it has significantly increased the range and sophistication of possible classroom activities (Hawkins, 1997b). Situations in real life are not as simplistic or as black and white as they are presented in many texts. With advances in technology, statistics education can be enhanced to teach, in a flexible manner, skills that are not learned during lectures. Rossman (1997) sees three main uses of technology in the statistics classroom: (i) performing calculations and presenting graphical displays of real datasets, (ii) conducting simulations in order to experience the long term behavior of sample statistics under repeated random sampling, and (iii) exploring statistical phenomena by making predictions and testing and revising these predictions using technology in an iterative manner. Thus, computers and calculators can not only take the burden of calculations away, they are, above all, powerful tools for illustrating concepts and ideas in ways that would not be possible without technology, for "bring[ing] specificity to the abstract language of statistics" (Behrens, 1997). By easily moving among tabular, graphical, and symbolic representations, students can analyze real data, they can make comparisons of expected to observed results, they can create and revise models to describe relationships, and they can perform simulations to help them understand probabilistic phenomena (Burrill, 1997a).

Recent technological advancements have changed both the computer and the image of the computer. Computers are no longer simply number crunchers. They are now multifaceted technologies, which facilitate unlimited opportunities in application, use and vision. They can provide a risk-taking, open-ended climate where there is a shared responsibility, a climate where students' voices can be clearly heard as they affirm their feelings, opinions, and ways of knowing their worlds (Christie, 1997).

Research Findings: Limitations of Technology

The agreement on the potential benefits of technology on student learning of statistics is unanimous. However, having a vision of what technology can do is not the same as knowing how to take advantage of these possibilities in a teaching context (Hawkins, 1997b). While some statistics educators believe that the progress in computer technology has had a significant effect on statistics instruction (Starkings, 1997), others are not as confident that the use of computers has had the expected impact in the classroom (Moore, 1993). Researchers such as Behrens (1997), warn us that "coupling the student with technology alone is generally insufficient to reach the desired effect" (p. 120) and that technological interventions might not work quite as well as we would like to think. They have brought attention to the fact that, despite the wide use of technology in many statistics classrooms, relatively little published research exists describing its actual impact on student learning and curricula are often developed and implemented without the benefit of research on their effects in terms of student learning.

Lipson (1997) notes that, although the trend in statistics education is to replace probability-based courses with courses where computer-based simulation exercises develop the idea of probability distribution, there has been little formal research done to study the kinds of understanding that develop as a result of these exercises. For example, the development of the idea of the sampling distribution is an area that has often been supported by computer simulation exercises. Many instructional programs have taken advantage of the ease of programming computers to draw repeated samples from a population and then summarize the results and draw the emerging patterns to help students empirically develop the idea of the sampling distribution. Lipson (1997) warns us that these approaches, although widely promoted and now commonplace activities in introductory statistics courses, may not have been as successful in developing in students the notion of sampling distribution as statistics educators have hoped:

> ICOTS 2 delegates were treated to "101 ways of prettying up the Central Limit Theorem on screen", but if the students are not helped to see the purpose of the CLT, and if the software does not take them beyond what is still, for them, an abstract representation, then the software fails. (p. 138)

Lipson (1997) goes on to report on a study she conducted to examine the effect of computer-based strategies that were designed to introduce the idea of the sampling distribution to a group of students in an introductory university level statistics course. Students in the study were graduates from a variety of courses, and although some had taken statistics courses in the past, for many others this course was their first experience studying any quantitative discipline. Lipson (1997) found that "many of the propositions that seem paramount in an *a priori* analysis

to an understanding of sampling distribution do not seem to have been evoked by the computer sessions, even though the sessions had been specifically designed with these in mind and students were led to these propositions by the focus questions" (p. 146). The ideas students had a particularly hard time with were that a sample mean has a distribution, parameters are constant, and the spread of the sampling distribution is related to sample size.

At the 5th International Conference on Teaching Statistics (ICOTS V) which had the role of technology in statistics instruction as its central theme, several discussants raised questions about the effectiveness of educational software. Behrens (1997), discussing the implementation of a graphical simulation program developed by Yu and Behrens to help students learn about statistical power, said that their experience showed them that without clear tasks, students simply move the sliders without any real purpose or understanding. Hawkins (1997b) pointed out that computer-based technology has brought with it not only new possibilities, but "many new challenges for the teacher who seeks to determine what it has to offer and how that should be delivered to students" (p. 1). She warned that the belief that introduction to technology automatically enhances the teaching and learning of statistics is simply not true, and that there is still a lot that we have to learn about the use of technology. Hawkins also argued that although there is a large selection of software available which allows students to quickly, efficiently, and under different conditions explore important statistical ideas, one should not take it for granted that students will actually grasp these

ideas without having had some experience with the concrete versions of the experiments symbolized by the computer software.

Other statistics educators have also questioned the effectiveness of computer simulations in helping students develop the important ideas of statistics. An article by Wilder (1994), expresses reservations about the use of computer-based simulations of random behavior in statistics instruction. Wilder indicates that the implicit assumption made by many curriculum developers that students accept the computer-based simulations as exhibiting random behavior is questionable. He also stresses the need for investigating how students' mental models of random behavior compare to their understanding of the computer representation of randomness: "The student needs to relate her own mental model of the problem to the computer representation: how she does this may depend on how she understands the computer generated model of random behavior." (Wilder, 1994, p. 2)

DelMas, Garfield, and Chance (1998) have, for several years, been engaged in extensive research that tries to investigate the ways in which undergraduate introductory statistics students' conceptual understandings are affected as a result of the interaction with an educational software called Sampling Distributions developed by delMas. Their research has been "a continuous cycle in which increases in [their] understanding of how and what students learn about sampling distributions lead to further revisions of the software and research methods, as well as to a better understanding of how to teach this complex and rich topic in statistics" (delMas, 1997, p. 87). They have

come to “recognize that good software and clear directions do not ensure understanding” (delMas et al., 1998, p. 24) and that learning is improved when students are engaged in carefully guided activities structured to help them become aware of their own beliefs and intuitions about chance events.

Similarly to delMas et al. (1998), Jones (1997) also stresses that intelligent partnerships with technology do not seem to be self-generating and that teachers have to develop instructional strategies that encourage their formation. Biehler (1997) conducted a study of twelfth graders at an American high school who had completed a statistics course that used the software DataScope (Konold & Miller, 1994), which was aiming at the reconstruction of different patterns of software use in the context of a data analysis problem. He found that most often students seemed to jump directly to particular methods or graphs offered by the software without much reflection on the things they had learned during the statistics course. His conclusion was that superficial experimentation with the statistical methods offered by the software is a first step, but we should also find ways to “improve the degree of networking in the cognitive repertoire of statistical methods” (Biehler, 1997, p. 175). This is necessary if students are to overcome the belief that it is adequate to use just one method or display.

Wood (1997), suggests that in cases where the underlying ideas are too complex for the users, the computer package has to be treated as a black-box because trying to follow the algorithms employed by the program would be too difficult for the user. He warns, however, that this creates “a potential problem here if the black-box is used incorrectly, if the output is misinterpreted, or if key

assumptions are not recognized" (Wood, 1997, p. 271). He reports on "a large unsuccessful" earlier attempt to build and use an expert system that would enable college students with little background in statistics to use standard statistical distributions. The reason for the failure was that students did not have an adequate understanding of the notions of "sampling size" and "probability distribution". Also, because they "had no image of what a statistical distribution was nor of the types of situations that the standard distributions will model adequately" (Wood, 1997, p. 268), they failed to appreciate the importance of the assumptions underlying the software's answer. The computer output "was correct in their minds, because it was produced by the computer, but also mysterious to them, because they had no idea of the rationale behind it" (Wood, 1997, p. 274). Wood (1997) believes that "the process of experimenting to see how the model works is likely to require encouragement or education", otherwise we run the risk of students "simply keying in the data, looking at the result, and leaving it at that" (p. 271). Other researchers also warn us that use of technology might lead to a lack of intimacy and "feeling for what is being done in the analysis, and a blind assumption that if the computer or calculator has done it then it must be right." (Nicholson, 1997, p. 31)

The black-box approach is the conventional technological approach to probability and statistics. The learner is expected to execute various commands or push certain buttons in order to perform simulations and obtain graphical images. Pratt (1998) gives the following limitations of the black-box approach:

(i) Not convincing

The black-box approach is often not convincing to the user who does not participate in the construction of the computer software. This can lead to a lack of intimacy that does not allow the user to recognize the key assumptions underlying the models generated by the computer.

(ii) *Data is not forceful*

Because, Pratt (1998) argues, our beliefs about stochastic phenomena are resistant to change, "we are more likely to find stories and explanations for the vicissitudes of the data than to believe an alternative interpretation of the data"(p. 113). In order to make the point that the evidence suggested by the data might not be forceful enough in changing people's views about the stochastic, Pratt describes an experiment by Konold, in which Konold placed bets against a student about the outcomes of a series of coin tosses (Konold, 1995b). In this experiment, Konold was the one who was using an incorrect mental model of the situation and although he kept on losing money, he was reluctant to accept the force of the data.

(iii) *Attention is a limited resource*

Technology is not necessarily interesting by itself and how involved the student becomes with a computer activity depends both on the design of the computer-based tools and the nature of the tasks accompanying these tools.

(iv) *Collection of Enough data*

Although it is very easy to collect data in a computer simulation, students may continue to underestimate how much data is needed to draw conclusions that are reasonably sound (Konold, 1995b).

(v) *Variability is typically ignored*

As Pratt (1998) posits, although the computer's ability to repeat experiments is a potential advantage that could be exploited to study variation, this is rarely encouraged. Simulations tend to focus on relative probabilities and to ignore variation.

(vi) *The focus might not be on sense making*

Simulations offer us a way of testing our theories, not replacing them, and the simulation approach runs the risk of leaving the informal intuitions that students bring into the classroom untouched. If people reason from a variety of perspectives and make incompatible predictions, computer simulations might have very little impact on their beliefs (Konold, 1995b).

Need for More Systematic Research

Several statistics educators have stressed the need for more systematic research of the effectiveness of programs incorporating information technologies on student learning of both the theoretical and practical aspects of statistics.

Schuyten and Dekeyser (1997) maintain that very little is currently known about the educational effect of instructional technology and that "case studies, effect studies of software, and evaluation of new curricular materials in natural settings of teaching are needed" (p. 210). Starkings (1997) calls for continued research that tries to determine the purposes for which calculators and computers are best suited and the ways in which technological developments impact statistical courses and curricula. She also points to the need for monitoring and evaluating statistical software that is developed for inclusion into statistical lessons.

Hawkins (1997a) laments the "paucity or lack of synthesis of good quality research to guide *any* developments in statistical education" (p. vii). She regards the belief that research is guiding our progress as simply a myth. Effective introduction of technology "requires empirical evidence about the optimal materials to be used, the methods for presenting them, and how to integrate them into the overall teaching process" (Hawkins, 1997b, p. 6). Carefully carried out cognitive research evaluating in depth the effect of particular technological approaches on students' understanding of some of the fundamental statistical concepts is urgently needed:

> It is one thing to claim that more dynamic and interactive software can allow students to gain insights by exploring and experimenting with statistical concepts. It is quite another to find empirical evidence of how, why, and when these enhanced insights are gained. (Hawkins, 1997b, p. 12)

Delegates of ICOTS V expressed concern that there are still examples of large amounts of money being hastily thrown at development projects of dubious educational or statistical merit, in response to what they perceived as the

"technology expansion crisis in education" (Hawkins, 1997a, viii). They identified the need for research to provide a deeper understanding of statistics learning in technological settings (Garfield, 1997). A working group formed at the conference to discuss the ways in which technology is changing the teaching and learning of statistics in secondary schools stressed that efforts to find ways of using technology to enhance statistical understanding should be research informed. They noted the need for promoting research that intensively studies the relationship among student understanding, statistical reasoning, and the role of technology (Burrill, 1997b). Similarly, among the research recommendations on teaching and learning statistics at the post-secondary level were the following: (i) investigate how to develop intelligent partnerships with technology; (ii) determine how the use of technology could enhance intuitions and understanding of specific probability and statistics topics; (iii) develop more and better methods to measure cognitive and affective effects of technology on all aspects of instruction; (iv) determine what forms of technology are optimal for what topics; and (v) determine what students learn when doing simulations (Blumberg, 1997).

Watson and Baxter (1997) point out that whereas R&D (Research and Development) is a central aspect of successful industrial practice, in education the link between educational practice and theoretical constructs is, at best, tenuous. They stress that the abundance and continuous development of technological innovations makes the employment of R&D nowhere as pressing as in the field of statistics education, since "a good feeling about an innovation is not enough to indicate its validity in terms of producing change" (Watson and Baxter, 1997, p.

285). McCloskey (1997) argues that the failure of many designers of educational software to show any evidence as to whether their stated aims have been achieved is the result of the "lack of a culture of assessing teaching quality in universities" (p. 98). In higher institutions, only student performance is assessed, and no standards against which we could measure the success of a new teaching method exist. Also, most of the assessment is being currently carried out by the software developers themselves. In addition to the partiality of such an approach, McCloskey also points out that software developers can only be held accountable for the quality of the content and the performance of the software. The effectiveness and efficiency of software are, to a big degree, determined by the way and the context in which they are employed and, for this reason, the only way of making a fair and meaningful assessment is to make our assessments "*in situ*".

REDEFINING STATISTICAL EDUCATION

Need for Synergy of Content-Pedagogy-Technology

There are many educators now pushing for introductory statistics courses that put more emphasis "on data collection, understanding and modeling variation, graphical display of data, design of experiments and surveys, problem solving, and process improvements, and less emphasis on mathematical and probabilistic concepts" (Ballman, 1997). This has caused a movement away from statistical content emphasizing the abstract and the memorization of a list of formulas and procedures, toward content emphasizing exploratory data analysis.

Although Scheaffer (1997) finds the emphasis on data now permeating many introductory statistics courses to be "generally positive" compared to more traditional approaches, he stresses that, for effective development of statistical reasoning, there needs to be more emphasis at both ends of the continuum from data to inference. He warns for the tendency in both teachers and students to "grab data wherever they can find it and rush to plot it on whatever plot they may have learned most recently" (Scheaffer, 1997, p. 157). Lack of attention is paid on "how the data originated, what the numbers might mean, if anything, and what plot or numerical summary might be appropriate" (Scheaffer, 1997, p. 157). As a result, one often sees "categorical data get put on stem plots and averaged; age get subtracted from heart rate" (Scheaffer, 1997, p. 157). Also, because so much time is placed on data exploration, not enough attention is given to helping students develop the concepts of statistical inference, which is necessary for them to understand how statistics allow us to make decision in the face of uncertainty:

> Some introductory statistics courses either have become pure data exploration or have remained exercises in formula manipulation with the formulas now residing on a piece of technological equipment. Some instructors view a modern course in the subject as a mixture of the two. All three outcomes are undesirable. The key is to find ways of teaching inference that are in keeping with the notion of construction (naïve, of course) but still allow closure on a few critical ideas. (Scheaffer, 1997, p. 157)

Scheaffer (1997) sees the technological developments as a major cause of the shift in emphasis towards "too much data exploration and too little inference" (p. 157). Technological advances have made data exploration "fun and quick", whereas "inference is still a black box, whether done by hand or on the computer"

(Scheaffer, 1997, p. 157). In fact, Scheaffer (1997) argues, "black-box inference can be even more of a problem now that a computer or graphing calculator can automatically fit a variety of models to data in almost no time" (p. 157). He stresses the need for more thoughtful and constructive use of technology in the classroom, and especially for developing new ways of helping students discover the principles underlying inference. Hoerl, Hahn, and Doganaksoy (1997) also advise for more caution and careful thinking about the purpose for which technology is used in the statistics classroom. Similarly to Moore (1997), they also point that technology should be used not for its own sake but with the purpose of serving content and pedagogy:

> We agree with Moore's fundamental point that technology should serve content and pedagogy. Unfortunately, we are sometimes infatuated with technology to the point where technology becomes the "what" to teach, rather than the "how". Our concern is that computer science will dominate statistics in the next century, just as mathematics has dominated in the past. (Hoerl et al., 1997, p. 151)

Hawkins (1997c) also agrees with Moore (1997) that discussions of one of the triad content-pedagogy-technology are often partial and argues that the nature of any reforms should not be uni-dimensional, but there should be synergy among the three domains:

> Many of the developments in statistical education in the past quarter of the century have been proposed from one particular perspective. Some have indeed evoked controversy. This is not to say that all such developments have been counter-productive. In general, the picture that emerges of present-day statistical education gives cause for optimism. However, there are certainly those, and I would count myself among them, who regret that such developments have largely been made in the absence of

> evidence-based understanding about the teaching/learning process. (Hawkins, 1997c, p. 144)

Current practices in statistics education have evolved from a background quite different from today's needs and possibilities. For this reason, Hawkins (1997c) argues that nothing should be taken for granted. Reform efforts "must have the momentum and energy to challenge even the most fundamental and widely-held ideas about statistical education, and the ways in which these are currently manifested" (p. 142). Both existing and proposed practices should be "open to empirical scrutiny that can sort the 'better' or 'best' from the 'good' or the 'bad', in order to find out when and why content or pedagogy or technology do and do not work." (Hawkins, 1997c, p. 142)

Before achieving the synergy among content, pedagogy and technology, Hawkins (1997c) believes that we should first change the emphasis in our teaching objectives. She is skeptical as to whether we have yet determined the right framework for reform:

> Reform is certainly required, but I am reminded of a traveler who, upon asking a local resident for directions to another town, receives the reply, "Well, to be sure now – I wouldn't be starting out from here at all"! (Hawkins, 1997c, p. 142)

New values and new competencies are necessary for survival and prosperity in the rapidly changing world where technological innovations have made redundant many skills of the past (Ghosh, 1997). The pressure for democratization of mathematics education has created new opportunities for statistics education. The shift of mathematical studies towards a more "utilitarian approach" (Moore, 1997, p.124), has opened up a larger place for statistics which

now, at the post-secondary level, is probably studied by more students than any other topic (Philips, 1999). At the same time, the forces of democratization demand fundamental pedagogical as well as curricular changes that would make statistics instruction more accessible to all students.

Changing of Emphasis in Teaching Objectives

In a world where "recognition of uncertainty as a characteristic of reality and how to behave within, forms a fundamental part of the intellectual development of the individual"(Azcarate and Cardnoso, 1994, p. 1), Hawkins (1997c) believes that we should not remain satisfied with "Statistics for All" policies. She argues in favor of "'Statistical Literacy for all' that emphasizes understanding over facts and tools, with specialists acquiring progressively more 'Statistical Literacy Plus', where the 'plus' possibly relates to more sophisticated/specialized techniques" (Hawkins, 1997c, p. 142). Although, she points out, there is unfortunately no universal agreement even within the statistical community on what statistical literacy means, some of its defining characteristics are the following:

> In its simplest terms, statistical literacy can be interpreted as meaning an ability to interact effectively in an uncertain (non-deterministic) environment. It is not merely the possession of an ever-increasing collection of analytic tools and techniques, although this is the outcome that often results from present approaches to teaching statistics. A statistically literate person must understand the strategies for data collection and analysis, as well as the nature of chance processes and their relevance to data collection, and the assumptions that underlie statistical reasoning. (Hawkins, 1997c, p. 144)

Kettenring (1997) informs us that his many years of experience in industry have taught him that "gross inefficiencies, major tactical and strategic errors, and expensive mismanagement of the enterprise result from a collective inability to learn from relevant data" (p. 153). He argues that if introductory statistics courses are to help improve the situation, they should ensure that sound intuitions about the stochastic "become part of the permanent intellectual bloodstream of the student" (Kettenring, 1997, p. 153). Developing statistical reasoning means having an appreciation of data management. Although this does not require a lot of theory, "it does require ample exposure to real problems in order to gain experience and to develop the instincts that will be needed on the job." (Kettenring, 1997, p. 153)

Joiner and Gaudard (1990) consider awareness of variation and how it affects processes as one of the main determinants of success of business management. Hoerl et al. (1997) point out the gross inefficiencies that occur in industry because most managers and technical personnel tend to think deterministically even though many of them have had formal statistics education:

> They expect mass balances to match exactly, or actual financial figures to exactly equal budget. Any "variance" from budget must be explained. The costs incurred by US businesses searching for "explanations" for random processes should be the cornerstone of statistical education, but most students are not coming out of the current intro course with this understanding. Why not? (Hoerl et al., 1997, p. 149)

A partial explanation Hoerl et al. (1997), see for the shortcomings in statistics education goes back to the math vs. statistics issue:

> Perhaps because of the over-emphasis on mathematics, statisticians seem uncomfortable with statistical concepts which cannot be derived or proven mathematically. The omnipresence of variation is admitted, but often not clearly explained (Simulating a histogram on the computer does not teach business people how to interpret a financial report). Process thinking is generally not taught, hence reducing the motivation to learn and apply statistics. Data quality is ignored while data quantity is emphasized. The concept that in the real world we most frequently sample not from static populations, but rather from dynamic processes is not well documented or taught. Deming (1953) wrote on this last issue, using the terminology "analytic" versus "enumerative" studies over forty years ago, but as a whole the profession still doesn't get it. (p. 150)

Hoerl et al. (1997) argue that we must completely rethink the sequence of topics in order to achieve the objectives for introductory education, which for them should be to "help students unlearn their deterministic view of the world, and view outputs as the result of a process" (p. 152). Instead of treating "data as a commodity", emphasis should be put on teaching students issues of quality of existing data and the importance of proper planning of investigations in order to collect measurements appropriate for the problem at hand. Although students might be taught how to answer the question "What is the required sample size?", the more fundamental questions that need to be addressed are "What information is really required to solve the problem? To what degree do the data at hand meet this need? What additional information needs to be obtained in the future and how?"(Hoerl et al., 1997, p. 149). Such a change in instructional emphases is necessary because otherwise companies will continue to spend significant resources on obtaining large amounts of the wrong type of data.

Changes in Pedagogy

In addition to deep curricular changes, democratization of education demands fundamental pedagogical changes. Democratization "needs new levels of self-discipline and tolerance of different points of view" (Ghosh, 1997, p. 154), it requires making room for student perspectives (Confrey, 1995). Instruction should revolve around student ways of thinking and understandings, and not around a pre-determined curriculum. If we do not want to mute the polyphony of student voice, classroom activities and assessments should allow for a variety of perspectives and approaches, "rather than molding and shaping students to do [statistics] a certain way, and rewarding those with the 'best fit'." (Scarano and Confrey, 1996, p. 32)

As it has been already discussed, the practice of almost every introductory course to present statistics content as a sequenced list of curricular topics fails to communicate to students the interconnectedness of the different statistical ideas they encounter in the course. Therefore, statistics instruction should adopt a more dynamic view of learning. The assumption that by building concepts "separately but directly", students would eventually have an array of statistical ideas at their disposal (Lachance and Confrey, 1996, p. 5), is a very simplistic view of the development of understanding:

> A structural model of the mind that envisions a network with information nodes and connections between them tends to connote a process of adding connections in a cumulative way. Full understanding is achieved by piecing partial understandings together building ever larger networks. Following students' partial understandings over time reveals that the actual process is much more chaotic than this. The connected network may still be a useful analogy, but there seems to be a continuing process of

> reorganization. Disconnecting, connecting, and reorganizing appear to be the rule rather than gradual addition to a stable structure. The appropriate model for the development of understanding may be one of change and flux and reorganization rather than steady monotonic growth. (Hiebert, Wearne, and Taber, 1991, p. 339; in Lachance and Confrey, 1996, p. 3).

Learning about a statistical concept without exploring its connection to the other main statistical constructs can only lead to weak and narrow understandings. As Lachance and Confrey (1996) assert, "the best route between two points is not always a straight line" (p. 23). Instruction should instead provide an interconnected path which, along with an emphasis on building on students' experience, would "better encourage students to follow their own, unique "nonlinear" developmental paths from "smaller" to "larger" ideas" (Lachance and Confrey, 1996, p. 7). Following such a path, rather than a compartmentalized statistics curriculum, should lead students to stronger and deeper understandings.

VARIATION AT THE CORE OF STATISTICS EDUCATION

Ballman (1997) believes that current reform efforts are still unsuccessful in providing the intuitions and understandings necessary to develop statistical reasoning because they do not succeed in helping students develop their understanding of variation and its role in statistics. She argues that the objectives of an introductory statistics course might be better met through topics and activities which help build a sound intuition about the characteristics of random variation and its role on statistics rather on topics and activities that emphasize traditional probability concepts (Moore, 1992; Ballman, 1997). She thinks that, in the introductory classroom, probability should not be viewed a series of topics

with no connection to the rest of the course but rather as a means of quantifying and explaining the variation that is present in almost all processes.

Rubin et al. (1990), emphasize the central role that variation plays in statistical thinking. As they explain, statistical reasoning follows from two notions which, when seen from a deterministic framework, seem antithetical - the notion of sample representativeness and the notion of sample variability. Whereas sample representativeness is "the idea that a sample taken from a population will often have characteristics identical to those of its parent population", sample variability is "the contrasting idea that samples from a single population are not all the same and thus do not all match the population" (Rubin et al., 1990, p. 3). In order to comprehend the purpose behind statistical inference, one has to balance these two ideas and realize that "a sample gives us some information about a population - not nothing, not everything, but something"(Rubin et al., 1990, p. 2). Due to sample representativeness, we can put bounds on the value of a characteristic of the population; due to sampling variability however, we never know exactly what that characteristic is. Balancing the idea of sample variability with that of sample representativeness lies at the heart of statistical inference.

Moore (1990) argues that essential components of statistical reasoning are recognition of "the omnipresence of variation", of the fact that "chance variation rather than deterministic causation explains many aspects of the world" (p. 99), and familiarity with the ways in which variation is quantified and explained. He considers the following elements to be the core of statistical thinking:

(i) the presence of variation in all processes,

(ii) the need for data,

(iii) the design of data production with variation in mind,

(iv) statistical analysis seeks to quantify and explain variation.

Pfannkuch (1997) considers the two essential and inter-linked components of statistical thinking to be:

(a) Recognition of variation, critical evaluation and ability to distinguish between the different types of variation:

  (i) "Special cause variation" – variation that can be assigned to an identifiable source "

  (ii) "Common cause" variation – variation that is hard to link to any particular source

(b) Realization that a sound judgment of a situation can only be made by collecting and analyzing data.

Pfannkuch (1997) discusses the characteristics of statistical reasoning laid out by a practicing as well as teaching statistician during an in-depth interview he conducted to investigate his perspective on the nature of statistical reasoning. Based on the insights obtained from the interview, Pfannkuch (1997) offers the following epistemological triangle, which has the development of understanding of variation at its core, as a model for introductory statistics instruction:

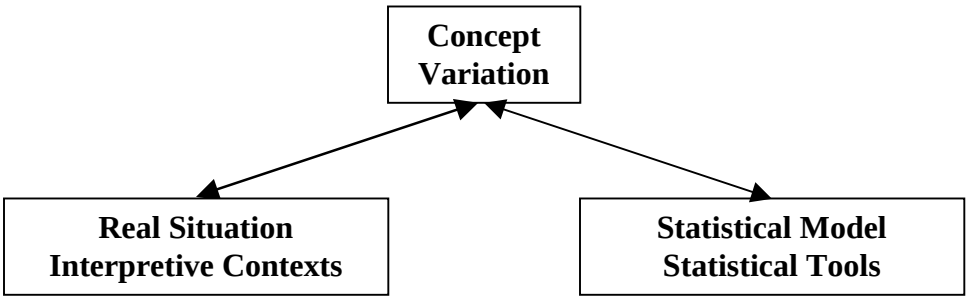


Figure 2.1: Pfannkuch's epistemological triangle

What the epistemological triangle suggests is that a combination of statistical knowledge and context knowledge is essential for conceptualizing variation (Pfannkuch, 1997). The inter-linked arrows indicate the strong linkage that has to be created between the statistical tools and the context of the problem. Emphasizing "the interplay of data and theory" is vital since the main purpose of statistical tools such as graphs and statistical summaries is to help understand or make predictions about stochastic real-world phenomena using a statistical model of them. The assumption underlying the epistemological triangle is that "the concept of variation would be subject to development over a long period of time with different tools and different contexts" (Pfannkuch, 1997, 177). In encouraging students to develop their understanding of the broader construct of variation, instruction should aim to "develop statistical thinking which could be regarded as the interaction between the real situation and its statistical model and

between these and the resulting conceptual development." (Pfannkuch, 1997, p. 177)

CONCLUSIONS AND EMERGING FOCUS FOR THIS STUDY

As the review of the literature indicates, most people, even ones with substantive formal training, tend to think deterministically and to have weak intuitions about the stochastic. Students' difficulties persist despite the significant reform efforts that have led to wide-scale incorporation of technology and interesting activities in the statistics classroom. A reason for students' difficulties in comprehending statistical concepts might be the neglect of variability and the statistical determinism hidden in standard approaches to statistics instruction. Instruction, assessment, and research tend to overemphasize the development of students' conceptions of center, while neglecting their development of conceptions of variability (Shaughnessy, 1997a). Very little is currently known about student understanding of variability; research investigating students' thinking on variation is urgently needed.

Another serious gap identified in the research literature is the lack of studies which perceive learning as dynamic, and intuitions not as static, but as "sense-making devices", constantly changing in light of experience (Pratt, 1998). It often appears that the main objective of much of the research literature on peoples' understanding of probability and statistics seems to be to discover (or confirm) intuitions running counter to stochastic reasoning and catalogue them as misconceptions that ought to be replaced. Snapshots of how students make sense of a stochastic situation at a particular point in time are taken, and there is almost

never any follow-up on student thinking. It seems that it would be more constructive if research findings were rather used as an indication of areas where more needs to be done in order to help students improve their intuitions. By identifying the similarities and differences between the students' informal intuitions and formal statistical reasoning, the researcher could then work with students' intuitive notions and help them develop ways to map new and richer concepts onto the ones that they already possess (Mokros et al., 1990).

> Current practices in statistics education have evolved from a background quite different from today's needs and possibilities. Technological advances and the forces of democratization demand fundamental pedagogical as well as curricular changes that would make statistics instruction more accessible to all students. These changes should come about after careful re-consideration of what the objectives of the introductory course ought to be. Also, unlike the uni-dimensional nature of many current reform efforts, there should be a synergy among content, pedagogy and technology.

Reflecting on the research literature led me, similarly to Ballman (1997), to conclude that reform efforts would have been more successful in achieving their objectives if they had put more emphasis on helping students build sound intuitions about variation and its relevance to statistics. Despite the movement away from statistical content emphasizing the abstract and the memorization of a list of formulas and procedures, traditional probability topics are still being taught, albeit in a later part of the course, as a separate chapter with little connection to what preceded or what will follow. I conjectured that if we provided learners with an environment where they experienced the omnipresence

of variation and came to value statistical tools as a means to describe and quantify that variation, we would help them develop statistical reasoning that goes beyond the superficial knowledge of terminology, rules and procedures.

Pfannkuch's (1997) model, which offers a nontraditional path to statistics instruction that has variation as its central tenet, seemed like a promising alternative to standard approaches. Unlike more conventional approaches, which ignore the influence of the setting, Pfannkuch's (1997) model acknowledges that the nature of the learning environment in which students experience stochastic phenomena can have an important effect on the use of statistical reasoning. Consequently, it views the construction of meanings about variability in particular, and the stochastic in general, as demanding the building of connections between informal and formal views of stochastic knowledge (Pratt, 1998). It perceives learning as a dynamic process subject to development for a long period of time, and through a variety of contexts and tools.

In the next chapter, I illustrate in more detail how I conjectured this radical restructuring of the curriculum as leading to stronger and deeper understandings by helping students see the "big picture" of statistics (Moore, 1997). I also describe the methodology and theoretical framework guiding this study, and the ways in which the model was linked to classroom practice.

Chapter III: Theory and Methodology

INTRODUCTION

The prevailing methodology employed by researchers examining conceptions of data and chance of taking snapshots of students' thought processes provides little guidance as to how one might systematically research conceptual change. There is hardly any information about the source of students' difficulties. Rarely does one do any follow up of the students' initial thinking to watch for future transitions (Shaughnessy, 1997a). Researchers such as Pratt (1998), who perceive cognition as an activity that is socially situated, are casting doubt on the validity of this research tradition, which ignores the influence of the setting on the shaping of intuitions. They stress the need for investigation of students' conceptions and beliefs in natural school contexts, and for a prolonged period of time.

A new trend now witnessed in educational research is an increase in the study of exemplary instructional practices based on the argument that new classroom practices need to evolve from these "best practices". However, this type of research might not be ideal for wide-scale implementation because it requires "a significant period of study and theory building" and "might result in a reform movement that is too constrained, too incremental, or too delayed" (Confrey and Lachance, 1999, p. 231) to meet students' needs. Rather, there is a pressing need for more speculative classroom research where some of the

constraints of typical classrooms are relaxed but others do remain in force (Confrey and Lachance, 1999).

This study has employed a research design model suggested by Confrey and Lachance (1999) called *transformative and conjecture-driven research design*. It is a model developed as a response to the need for establishing a better connection between educational research and practice. This model "utilizes both theory and common, core, classroom conditions in order to create and investigate new instructional strategies...meant to change and even to reform current teaching practices drastically." (Confrey and Lachance, 1999, p. 231)

The conjecture driving the study is a claim that if statistics curricula were to put more emphasis on variation, then we would be able to witness improved comprehension of statistical concepts. In this chapter, I describe how the conjecture was developed and how it was linked to classroom practice. I also provide an overview of the philosophical foundations underlying the conjecture-driven design model and outline how this approach was employed in the study in terms of research design, data collection, data analysis and rigor.

DEVELOPING THE CONJECTURE

Ideological Stance

> *I learned a lot from my teachers, and even more from my colleagues, but from my students - I learned the most.*
> (Talmud; in Ben-Zvi and Friedlander, 1997, p.45)

As Confrey and Lachance (1999) posit, "one's ideological stance informs and saturates one's research design model" (p. 234). I chose the conjecture-driven research model because it is a methodological perspective that fits my

ideological stance both as a mathematics educator and as a researcher. It is consistent with my view of mathematics as a human enterprise, and with my belief that there is a lot that both researchers and instructors can learn by giving voice to students. It is a design well suited for helping me achieve the aims of my study, which are to establish a direct link between intuitive and formal understandings of variation that will allow more students to succeed in statistics.

The ideological stance on which the transformative and conjecture-driven teaching experiment[2] is based is commitment to equity. Confrey and Lachance (1999) believe that the premature emphasis of mathematics education on "the abstract and the formal" leads to unnecessary failure and prevents many students from accessing the mathematical concepts. I am also convinced that an approach to learning that does not allow students to link abstract concepts to their everyday experiences and intuitions impedes the majority of students from doing well in statistics. Epistemological anxiety (Wilensky, 1993) is one of the main factors why students do so poorly in statistics, and research and instruction which "validate [student] opinions, encourage them to think, to wonder, to question, to ask, to agree or disagree" (Goldstein, 1997, p. 74), are needed. This is especially critical given that statistics is now considered to be a crucial part of the education of students at all levels, and from many different disciplines.

The study adopts the view of statistics as an informal, creative, human enterprise and aims at developing alternative ways of teaching statistics, which

[2] Confrey and Lachance (1999) explain that they have chosen to retain the label *teaching experiment* despite the connotations of the term experiment because teaching experiment has established itself with a varied set of meanings in mathematics education.

will raise students' awareness of the stochastic. I chose a research model that recognizes the novelty and power of students' ideas and gives validation to their personal voice (Confrey, 1991) in the anticipation that not only students' notion of variation would change, but my notion also would be enriched with varied and ingenious approaches by students. In contrast to traditional instruction and research models, which "make one unlikely to hear students' constructions", the conjecture-driven research is based on the belief that "a novice can often envision possibilities, arrangements, and logical relations that experts have been trained to overlook" (Confrey and Lachance, 1999, p.234). It is willing, despite curricular constraints and limitations set by the learning environment, to give validation to student voice and to continuously examine and revise the "expert" perspective in light of student voice (Confrey and Lachance, 1999).

The Conjecture

Definition of Conjecture

The conjecture is a very important aspect of the kind of research described in this study. Confrey and Lachance (1999) explain:

> The conjecture is a means to reconceptualize the ways in which to approach both the content and the pedagogy of a set of mathematical topics. Most often, it comes from a dissatisfaction in the researchers' mind with the outcomes of typical practices. It transforms how one views teaching and learning activities. Over the course of the teaching experiment, a strong conjecture should shift one's perspective and bring new events, previously insignificant or perplexing, into relief. At points in its evolution, the conjecture should feel like a grand scheme beginning to emerge from many, previously disparate pieces, making them more cohesive. (p. 235)

The conjecture has two dimensions, a content dimension, and a pedagogical dimension. The content dimension answers the question "What should be taught?" The pedagogical dimension guides instructional decisions; it answers the question "How should this content be taught?" (Confrey and Lachance, 1999, p. 235). The conjecture also has a theoretical aspect. It is situated within a theoretical framework, which helps to structure the activities and methodologies used in the teaching experiment and link together the content and pedagogical dimension of the conjecture. A thorough review of the existing research literature is a critical element of the theoretical framework. A robust conjecture does not "fall full-blown from the sky" (Confrey and Lachance, 1999, p. 236). It comes by carefully and critically reading and reflecting on the existing literature to relate one's ideas about the phenomenon under study to those of other researchers (Romberg, 1992), and "to discern an anomaly that has been overlooked, unsolved, or addressed inadequately by one's colleagues." (Confrey and Lachance, 1999, p. 236)

A conjecture is "*not* an assertion waiting to be proved or disproved", but "an inference based on inconclusive or incomplete evidence"(Confrey and Lachance, 1999, p. 235). In research following the positivistic paradigm, hypotheses or theses are set before the study begins and the whole purpose of the study is to confirm or disprove the truth of these hypotheses. In contrast, a conjecture-driven research design perceives theory development as an inductive process. The purpose of the conjecture is to serve as a guide and not to constrict the collection of data. During the course of data collection and analysis, as

experience with the setting increases, the conjecture is subjected to several alterations and modifications (Confrey and Lachance, 1999).

Variation as the Central Tenet of Statistics Instruction Conjecture

My beliefs and considerations about the epistemology and pedagogy of the stochastic guided the development of the conjecture. I embraced a dynamic view of learning that allowed me to use the results of the existing literature and personal research not as proof of innate limitations in students' ability to reason about the stochastic, but as a signal of the areas for which current intuitions needed to be strengthened. I conjectured that the reason students' intuitions about the stochastic are weak might be the instructional neglect of variation as well as the neglect of students' intuitions. I decided to conduct a study where statistics instruction would have variation as its central tenet, and which would allow me to investigate in depth how students' intuitive notions of variation developed over the course, through the use of different tools and different contexts. Pfannkuch's (1997) epistemological triangle, which calls for the re-structuring of statistics instruction by offering a nontraditional path with variation at its core, seemed well suited for meeting my research aspirations.

This epistemological triangle views variation as the broader construct underlying statistical reasoning and, in encouraging students to develop their understanding of the concept of variation, it at the same time aims to promote richer understanding of all the other main statistical ideas. The epistemological triangle indicates that for conceptualization of variation, a combination of subject and context knowledge is essential (Pfannkuch, 1997). The inter-linked arrows

show the strong connections that have to be created between the statistical tools and the context of the problem. Emphasizing "the interplay of data and theory" is essential since the main purpose of statistical tools is to help understand or make predictions about stochastic real-world phenomena using a statistical model of them.

I conjectured that instruction adopting this model should lead to improved statistical thinking, since statistical reasoning "may not occur unless there is recognition of the underlying variation and an understanding of how to critically evaluate that variation from a real situation and a statistical model perspective" (Pfannkuch, 1997, p. 177). Also, since understanding the context of the situation might be required to operationalize statistical reasoning, teaching variation through a variety of different contexts students have prior knowledge about has to be central to statistics instruction and reasoning:

> Historically probability has roots in two different lines of thought: the solution of gambling problems and the handling of data (Lightner, 1991). Today the gambling root of probability dominates teaching and textbooks. The other root in data needs to flourish alongside with the emphasis being on exploring variation rather than exploring probability as it is now in some curricula (Ministry of Education, 1992). Variation in all its contexts needs to be central to statistics teaching and thinking. And since context knowledge may be needed to operationalize statistical thinking then this implies that students should be taught from contexts that they have knowledge about. (Pfannkuch, 1997, p. 177)

This model, which bases instruction on contexts directly connected to students' experience, seems like a promising alternative to typical approaches to statistics, which attempt to develop probabilistic reasoning through standard probability tasks. Expecting students to transfer the understanding obtained

through coins, dice, and games of chance to everyday contexts is (as the literature review has already indicated) a naïve assumption, since understanding of variation in random devices is very different from understanding variation in data (Pfannkuch and Brown, 1996). Statistics instruction deals not only with difficult concepts, but also "with psychological issues involving chance and data that can be deeply rooted in students' experiences or in their beliefs about chance phenomena" (Shaughnessy, 1997a, p. 130). Students bring to each situation a great variety of prior beliefs, conceptions, and interpretations that instruction has to take into account in order to communicate statistical ideas in a clear and intellectually accessible language (Shaughnessy, 1997a). If the goal is to help students gain sound understanding of statistical concepts, instruction should start by building on their intuitions, even when those are weak. Instruction that neglects or attempts to replace students' intuitions suppresses their intellectual development and is bound to failure.

The underlying assumption of the model I adopted is that "the concept of variation would be subject to development over a long period of time with different tools and different contexts" (Pfannkuch, 1997, p. 176). Rather than viewing learning as a linear and monotonic process that comes once through insights that make everything clear, it proposes a model of understanding composed of "successive cycles of development, modification, clarification and evaluation of conceptual tools" (Confrey, 1996, p. 4). Its dynamic view of learning contrasts the practice of almost every introductory statistics course,

including the ones that made significant moves towards modernizing statistics education, to follow a structured progression of statistical ideas with little overlap.

This model, sees the inconsistencies in students' approaches and levels of understanding as indication that developing conceptions are overlapping and concurrent, rather than disjoint and sequential (Confrey, 1988). It recognizes that student knowledge of a certain concept is a complex system:

> For a concept or given property we cannot simplify the possible manifestations that the students make about it by stating that they "know it" or they do not "know it"; it is worthwhile differentiating between the different types of mistakes and strategies, which, generally speaking, cannot be put in order on a numerical scale. (Batanero and Godino, 1994, p.1)

Pfannkuch's model shares the assertion of Lachance and Confrey (1996) that "the best route between two points is not always a straight line" (p. 23). It calls for development of the concept of variation through a variety of experiences and contexts that are related to a variety of interrelated statistical ideas. The interconnected path it provides, along with its emphasis on building on students' experience, "better encourage students to follow their own, unique "nonlinear" developmental paths from "smaller" to "larger" ideas" (Lachance and Confrey, 1996, p. 7). The appreciation of the interconnectedness of variation to all of the different ideas encountered in the course should lead to stronger and deeper understandings than a compartmentalized curriculum.

This study adopts an instructional approach that offers a re-visioning of the introductory statistics instruction. Content is no longer a sequenced list of curricular topics but "an interrelated repertoire of conceptual tools that can assist

one in making sense of, and gaining insight and prediction over interesting phenomena" (Confrey, 1996). It recognizes that adequate statistical reasoning requires more than understanding of the different ideas in isolation. It demands "*integration* between students' skills, knowledge and dispositions and ability to manage meaningful, realistic questions, problems, or situations, both as generators as well as interpreters of data, findings, or statistical messages." (Gal and Garfield, 1997, p. 7)

The theoretical framework within which the pedagogy of the conjecture is embedded is influenced by the acceptance of constructivism as a learning theory. Classroom activities encourage students to explore statistics in familiar and meaningful contexts and to elaborate on and refine their partial understandings and the subtleties in their thinking rather than to suppress them (Confrey, 1996). Use of both physical and technology tools is encouraged. There is a synergy of content, pedagogy, and technology.

DEVELOPING THE TEACHING EXPERIMENT

A transformative and conjecture-driven experiment is a planned intervention, which takes place in a regular classroom over a significant period of time and involves a dialectical relationship between the conjecture and the different components of instruction (Confrey and Lachance, 1999). What makes this research model unique and leads to a re-definition of the research-practice relationship, is that its research questions and methods of data collection are informed both by the conjecture and the components of instruction (Confrey and

Lachance, 1999). Theory and ideology also influence all of the components of instruction (Confrey and Lachance, 1999).

Due to the need to continuously discuss and refine plans and interpretations, a transformative, conjecture-driven teaching experiment requires a team of researchers working together (Confrey and Lachance, 1999). In this study, I worked jointly with the instructor towards designing the different aspects of the curriculum, towards refining and elaborating the conjecture and the components of instruction. The short duration of the summer course meant the instructor would have to cover a huge amount of material in five weeks. Consequently, the different activities used in the course had to be carefully planned so that, while being flexible and open-ended, they also took the time constraints and the confines of the curriculum into account.

One of the main research findings of our previously conducted research (Meletiou, Lee, and Myers, 1999; Meletiou, Confrey, Lee, and Fouladi, 1999; Meletiou, Lee, and Fouladi, 1999) was the need for improving the way in which technology was used in the PACE course. We had found out that although use of technology provided PACE students much more familiarity with the practical aspect of statistics, it did not adequately contribute into improving their statistical reasoning. One software I had become familiar with by participating in a workshop during the 1998 NCTM (National Council of Teachers of Mathematics) meeting, was the object-oriented computer learning environment Fathom (Key Curriculum press; developmental release). Unlike the more conventional black-box use of technology that PACE students had experienced in the past, Fathom

offers a learning environment where students can build and modify their own statistical models. Such an environment provides an explicit interactive feedback-structure for checking, improving and modifying one's comprehension of statistical concepts (Steinbring, 1990). It allows students through interaction with the technology to build on, refine, and reorganize their prior understandings and intuitions about the stochastic.

During the course-planning period, I laid down all the possible benefits in terms of learning outcomes that I could see coming out of students' use of Fathom. The instructor was interested in exploring Fathom's potentials, but his lack of familiarity with the software, and the fact that it was still at the developmental stage with no research having yet been conducted in terms of learning outcomes, made him uncomfortable using it as the primary technological tool. What we ended up deciding was that, other than a couple of activities using Fathom that the instructor would introduce in the classroom, I would work independently with a group of students outside class to assess the effectiveness of the software as an aid to conceptual understanding. These open-ended investigations of individual students interacting with technology, would further help redefine and strengthen the conjecture.

Context

The site for the study was an introductory statistics course in a mid-size Midwestern university. The study lasted over the span of five weeks. The observations took place during the scheduled times for the PACE course. The

course began on the last week of June and ended on the first week of August. The class met four days a week, for two hours each day.

Participants

Recruitment of Students

The number of students in this class was thirty-three. Given the number of students involved, it was impossible to observe closely every single student in the course. Therefore, I chose to study two groups of students. The primary group consisted of a subset of about eight students and the secondary group encompassed the whole class. Although data from both groups were used in the analysis, my focus was on investigating and describing the learning experience of students in the primary group. The selection criterion for the primary group was willingness to participate in the study.

Characteristics of Students

The class was made up of nineteen males and fourteen females. Most of the students in the class (twenty-two students) were majoring in a business related field of study. Very few students had a strong mathematics background. Only thirteen students had taken a pre-calculus, and just seven a calculus course.

The primary group was representative of the students in the class. It was made up of five males and three females. Their degree specializations were similar to those of the rest of the students: five students were specializing in Marketing, one in Economics, one in Manufacturing Systems Engineering, and

one in Sports Medicine. Out of these eight students, only two had taken a pre-calculus course, and none had taken any calculus.

Risk Protection for Students

Any time people are being interviewed, observed, or otherwise studied, their well being must be protected. To that end, I took several procedures to protect the rights of my informants. The choice to take part in the study was voluntary. Only those students who gave their permission to do so participated. I followed all guidelines established by the university's Human Subjects Committee. Participants were given a written statement asserting them that the research was conducted with the consent of their instructor and their university and that their participation was entirely voluntary. They were also assured that participation or non-participation would not affect their grades in any way, and that they would be able to withdraw from the study any time they wished. Their written consent was obtained before they were audio-taped or video-taped.

Finally, the privacy of participants' responses was guaranteed. Students were assured that the notes, audio and video-tapes taken during data collection would be kept locked in a file cabinet and the data would not be made available to any other person than the principal investigator unless specific consent was received from them. They were also assured that, when reporting the results of the study, I would change their names to pseudonyms to protect their identity.

Four Design Components of Instruction

The four components of instruction are (a) the curriculum, (b) the method of instruction, (c) the role of the teacher, and (d) the methods of assessment. I will next offer a general description of how each of the components was conceived and progressively carried out in this study.

Curriculum

In a transformative and conjecture-driven experiment, it is the conjecture that exerts the biggest influence on the content of the intervention, the choice of activities, as well as their sequence and duration (Confrey and Lachance, 1999). However, because the experiment takes place in a regular classroom, there might be need for a modification of the curriculum that is a compromise between the researcher's intentions and the practical demands of the setting (Confrey and Lachance, 1999).

In planning for the intervention with the instructor, we had to make sure our intervention covered the set curriculum that is typically supposed to be covered in an introductory statistics course. However, we expanded the curriculum by including throughout the course activities that aimed at raising students' awareness of variation. We approached the different topics through the lens of the conjecture. Following Pfannkuch (1997), the characteristics of statistical reasoning that instruction should aim at developing were categorized from a modeling perspective of statistics. The three categories used were: (1) understanding the dynamics of the real-world problem, (2) moving towards a statistical model, and (3) using statistical tools.

1. Understanding the context of the problem. When dealing with a real-world problem, the first thing one ought to do is to understand its dynamics. This sets an inquiry process in motion whose success is dependent upon the ability to adopt a view that notices variation, wonders why that variation is present and tries to find ways to collect data or use existing data, that will help answer that question (Pfannkuch, 1997). Acquiring broad background knowledge about the situation requires a mixture of deterministic and non-deterministic thinking. It demands being able to conceptualize variation and distinguish between the different types of variation. "Special cause variation" is variation that can be assigned to an identifiable source, whereas "common cause" variation is hard to link to any particular source. In-depth investigation is necessary for acquiring knowledge on how to improve a system. (Pfannkuch, 1997)

2. Moving towards a statistical model. Once a broad background knowledge about the situation under study has been developed, one can then narrow down the problem and refine it to measures and stratifications "that will capture the essence of the problem, that will reflect a partial truth or model of the actual situation" (Pfannkuch, 1997, p. 173). This stage is very important, and it is because not enough attention is paid to it that so many organizations often end up collecting data using the wrong data production process or collecting and storing data for no reason (Pfannkuch, 1997). The choice of appropriate measurements depends on the ability to recognize the presence of variation and decide how to most effectively deal with it. When using data that has already been collected, one should not take it at face value but should examine it careful and try to

understand what this data means in context and how valid conclusions drawn from that data are likely to be. One should notice, understand, and critically evaluate the variation in the data, looking for biases that might distort the data or measurement errors that are of such a magnitude that they completely obscure any signals (Pfannkuch, 1997).

The model constructed to describe the data will have both a deterministic and a non-deterministic part. The so-called deterministic part is formed by the systematic influences that the person identifies, whereas the causes underlying the remaining observed variation that cannot be directly analyzed, "are conveniently described as random" (Pfannkuch and Brown, 1996). Although random variation is often viewed as unexplained variation, at an intuitive level one could explain variation in individual data, especially if the data was not obtained from random devices: "All variation is caused. Unexplained variation in a process is a measure of the level of ignorance about the process" (Pfannkuch, 1997, p. 175). Probability is employed to model this unexplained variation. One should then view random variation as a model superimposed to deal with variation in which one cannot discern any reliable patterns (Pfannkuch, 1997).

Whether planning for the collection of data or having been presented with existing data, it is very important that one be aware of how one's personal beliefs and biases could influence the interpretation of the situation and the available statistical information, and actively seek alternative explanations (Pfannkuch, 1997). One should always examine differing points of view, employing a combination of deterministic and non-deterministic thinking:

> Much effort must be spent in dialogue with the data, looking for and extracting multiple explanations. One explanation that does not seem to naturally occur to some is the possibility that the difference is due to chance and that some statistical tools will evaluate that possibility or if there is a difference, that statistical tools will take into account the size of the sample. (Pfannkuch, 1997, p. 173)

3. Using statistical tools. When planning for a statistical investigation, one needs not only context, but also subject matter knowledge. To be able to recognize the similarities and differences between the real-world problem and its mathematical model, and the connection between the summary statistics and the situation at hand that generates those patterns, one needs sound understanding of statistical tools. This encompasses understanding that behind those analytical tools lies the concept of variation. Graphs, for example, "should be inspected and fundamental questions asked, such as, what is going on here, is this common cause variation or is this special cause variation?" (Pfannkuch, 1997, p. 174). Adequate understanding of sampling includes appreciation of how statistical tools take into account the variation from sample to sample so that "samples will say something about the population from which they are drawn and will help gauge whether the system is stable (common cause variation) or not " (Pfannkuch, 1997, 174). Understanding variation in relation to significance testing implies "recognition that special cause can be revealed in what is thought to be common cause variation in the summary statistics." (Pfannkuch, 1997, p. 175)

The initial design of the intervention was based on the literature review and findings from previously conducted research. Findings from a questionnaire given on the first day of class to investigate students' intuitive understanding of variation prior to instruction, and which will be discussed in the next chapter, led

to further elaboration of both the conjecture and the instructional design. Also, since in the conjecture-driven teaching experiment instruction changes over the course of the intervention in response to students' needs and inputs, we could not design a complete curriculum before the experiment began. Curricular activities were structured in a way that made them flexible and open-ended in order for us to be able to adapt them in response to feedback from students. Curriculum development was "responsive and emergent" (Confrey and Lachance). Changes in the conjecture also occurred as a result of the insights gained from the teaching experiment. However, they were of a smaller magnitude than curricular changes. Confrey and Lachance (1999) describe changes in the conjecture as "evolutionary refinements" and "elaborations".

Classroom Interactions

The format of the course was based on the PACE model. It was a combination of lecture, group work, and whole class discussions. Each class meeting included laboratory time where students would work collaboratively on activities carefully designed to help them explore different statistical concepts. Classes were sometimes held in a regular classroom and sometimes in a microcomputer-equipped lab. Some of the activities involved the use of technology. Minitab was the main software employed, although Fathom was also used a couple of times.

The structure of the course was such as to encourage students to express their partial and incomplete understandings. Throughout the course, there was validation of personal voice and negotiation of meaning (Confrey, 1991). Both

group and whole-class activities stressed the importance of communication skills. As students worked through the different activities, they would constantly read, write, and talk with one another and with the instructor. Because we viewed learning as a socially constructed meaning, everybody's personal view was taken seriously into account in informing instruction.

The Role of the Instructor

Dr. Lee, the course instructor, is a statistics professor of Taiwanese descent with a strong interest in education. He is an instructor who is continuously striving to improve his course and is always open to new ideas about teaching. At the time of the study, he had already taught statistics for over sixteen years. The instructor's pedagogy, which stresses the importance of students constructing their own knowledge, served to support the research purposes of this study. In addition, the distinction between research and teaching within this methodology was blurred. Dr. Lee is a statistics education researcher with whom I have been collaborating for three years. He is the developer of the PACE model and a major partner in research I had previously conducted. He was therefore very familiar with the conjecture and acted as a research collaborator. He was actively participating in the intervention's development and assisted in both the preliminary and final analysis of the data (Confrey and Lachance, 1999). The fact that he was the one teaching the course gave me more time to spend observing student interactions and evaluating the effects of the intervention. I was continuously presenting him with data that I had collected and analyzed, and we

would jointly draw our interpretations and decide how to use the feedback provided by students to adapt the curricular activities accordingly.

The instructor was continuously trying to actively engage each student with learning the material through reading, thinking, discussing, computing, interpreting, writing and reflecting (Rossman, 1996). His role in the classroom varied. During group activities, he acted as a facilitator. In the whole-class discussion which followed each group activity to summarize what had been learned, the instructor would act as a discussion leader guiding students and helping them see how the specific topic is related to the big ideas underlying statistical reasoning. When necessary, he would also act as an expert source offering "mini-lectures" (Rossman, 1996), he would however always make sure he used examples students would relate to. His instruction would consistently try to make connections to students' everyday experiences and he would adjust it depending on feedback received by the students. The following conversation we had during one of our meetings is indicative of this instructor's teaching philosophy:

> Int.: So, you said that in the past you were trying many different activities, but you saw that they don't transfer from one context to the other.
>
> Inst.: No, no they don't. So I decided to focus on one thing – everything related to their experience in the world and always start from there. And that's what I think introductory statistics is for anyway. You want to do real world, you have to come from their experience. From experience, going to the real world, that's the approach I use. When I see students coming here and pretending listening, I can see from their eyes if they are into it or not. When I see they are not, I quickly think about something different and then immediately change the thing I'm going to do next. That's why my class lectures and activities always change. Today, when I

asked them about that situation to compare the two (a certain activity they had engaged in), it didn't come to my mind till I got there. That's why my notes are so brief. Almost any class you see, I don't have notes. I use different examples each time.

Int.: Not many people can do that, adjust in a second.

Inst.: You have to observe and then analyze right away.

Int.: Not many people have that gift.

Inst.: Not many people are interested in doing that.

Assessment

In order to determine what students know, we need a better understanding of how knowledge of the related statistical concepts develops (Friel, Bright, Frierson, and Kader, 1997). The current literature provides very little knowledge about the development of key concepts related to variation. A major objective of this study was to shed light on this much neglected but urgently needed area of statistics education research. By assessing students' understanding prior to instruction, and then monitoring changes in their thinking throughout the course, it attempted to develop a detailed description of the processes students go through in order to become able to intelligently deal with variability and uncertainty.

A major factor affecting how successful a study is in gaining true insights into students' thinking, is the careful choice of assessment techniques. Assessment tasks that uncover in detail students' understanding of the statistical concepts under investigation and allow the research to distinguish between relational and instrumental understanding (Kelly, Sloane, and Whittaker, 1997) are needed. Multiple assessment practices, "informed by and consistent with the

content, pedagogy, and theoretical framework of the conjecture along with the other components of the intervention" (Confrey and Lachance, 1999, p. 248), were employed in this study and they are discussed in the next section titled "Data Generation".

The assessment strategies that we used to support and evaluate this conceptual development (Friel et al., 1997), which encouraged students to explain in detail their thinking, helped both us and students further clarify their reasoning strategies. The continuous monitoring of the effect of instruction on student learning was constantly supplying valuable information on their levels of concept attainment. This informed instruction, which was adjusted to promote deeper understandings, and it guided the evolution of the conjecture.

The teaching and learning of statistics is remarkably complex because it involves not only new and difficult concepts but also belief systems resistant to change (Metz, 1997). For this reason, I was examining students' emergent understandings of variation not only along the cognitive, but also along the epistemological and cultural dimension (Metz, 1997). The epistemological dimension examined how beliefs came into play in whether or not students thought to apply statistical ideas in their attempts to make sense of patterns. The cultural dimension investigated how the classroom culture supported or subverted students' grasp and utilization of the big ideas related to variation. Although the culture of the society at large also influences the kinds of situations in which individuals are likely to consider a stochastic interpretation and although some

attention was paid to its effect on student interpretations, this was not a focus of the study.

1. Epistemological dimension. Students' understanding of randomness and chance variation involves not only conceptual construction, but also beliefs about the place of chance in the world. The students' *epistemological set* was an important dimension of my assessment of students' understanding and application of ideas related to variation, and especially chance variation. An epistemological set is an individual's inclination to interpret the world in relative deterministic or stochastic terms, based on their beliefs about the place of chance and variation in the world (Metz, 1997). Individual beliefs about the place of chance and uncertainty in the world can affect students' ability to grasp the main ideas behind inferential statistics and their propensity to apply chance interpretative schemas. Understanding, for example, that a population proportion of ¾ is a ratio of the expected relative distribution over an infinite number of repetitions of the event, but that this ratio is only approximated across many repetitions, requires more than having constructed the concepts of randomness and chance. It also requires an inclination to interpret the situation in terms of chance and uncertainty (Metz, 1997).

Although all individuals have both chance and deterministic interpretations within their cognitive repertoire and which one they utilize depends on many factors including the context of the situation, different individuals have varying tendencies of interpreting phenomena toward one or the other end of the stochastic/deterministic continuum (Metz, 1997). There are

people with epistemological sets that are relatively deterministic and others with epistemological sets emphasizing chance and uncertainty. I judged students' epistemological set based on the assumptions they made about causality and chance, on their propensity to assume deterministic explanations vs. their willingness to seriously consider the possibility that the outcomes might be the result of chance. Differences in epistemological stance were manifested in the difference in ways in which students interpreted the same data set – whether they saw "determinism in variability" or whether they saw "probabilistic patterns" or "uninterpretable uncertainty" (Metz, 1997, 234).

2. Cultural Dimension: Assessment of Instruction. To adequately understand how students construct meanings about variation, we need to consider the culture in which students participate. The extent to which individuals assume that deterministic causality underlies variation, as opposed to the possibility of random variation, depends in part upon the orientation of their classroom culture (Metz, 1997). I investigated how the culture of the classroom influenced students' beliefs and ideas. I tried to assess the extent to which the culture of the classroom, in the activities it structured and the interpretations it valued, embodied a deterministic vs. nondeterministic view of the world (Metz, 1997). I considered assessment from the cultural perspective of messages about the place of variation, chance and determination, implicit in the values and habits of the learning environment. Indicators of classroom epistemological set included the choice of subject matter, the structuring of problems, the instructor's reaction to students' claims about causality, the aesthetics of what constitutes a good solution

or explanation, and the instructor's willingness to accept multiple strategies and viewpoints (Metz, 1997).

Understanding of the connections between the different statistical concepts and techniques is essential for statistical reasoning and should be an explicit goal of instruction (Schau and Mattern, 1997). I investigated the degree to which instruction assisted students in gaining understanding of the interrelationship among the different statistical concepts. I assessed the degree of integration of the different topics to the general construct of variation.

DATA GENERATION

In order to enhance the understanding of the research setting and be able to provide answers to the research questions, a transformative and conjecture-driven experiment needs to use multiple forms of data generation. The data gathering techniques I employed included (1) direct and participant observations, (2) interviews with the students and the instructor, (3) samples of student work and (4) other relevant documents. Drawing data from several different sources permitted cross-checking of data and interpretations. This practice, called triangulation of data, increases the strength of the design and, consequently, the credibility of a study.

In addition to data triangulation, I also used methodological triangulation. The data generation process followed a mixed-methods approach. I employed both qualitative and quantitative techniques to gather data from correspondents. The purpose of including quantitative methods was to indicate directly observable relationships and corroborate the findings from qualitative data. Linking the

depth of qualitative data with quantitative breadth provided me with complementary information and a more holistic picture of students' understanding of variation.

There was also investigator triangulation since Dr. Lee was not merely a statistics instructor, but also a research collaborator. Undoubtedly, the comments and suggestions offered by a fellow researcher and a much more experienced instructor provided me with some invaluable insights that have led to a much better understanding and elaboration of the conjecture.

I will now explain the specific data that I generated. I have found it useful to describe the data generation process separately for each of three phases of the course: (a) beginning of course, (b) duration of course, and (c) end of course.

A. Beginning of Course

Questionnaire on Variability

A diagnostic questionnaire (Appendix A) was given the first day of class to assess students' conceptual understanding of variability prior to instruction.

Follow-up Interviews of Primary Group

Soon after the primary group was identified, I conducted individual interviews with each of the members of the group. The interviews, which were audio-taped, were semi-structured. I began by asking some general questions that helped me get information about the participants' background and interests. Then I asked them the set of questions included in Appendix A. The final part of the

interview was a follow-up of the questionnaire on variability taken by the whole class. I went over the questionnaire with the students in order to clarify the reasons for the different responses they had given. Despite the open-ended nature of the diagnostic questionnaire, one-to-one communication with students allowed a more thorough investigation of their reasoning.

B. Duration of Course

Class Observations

In qualitative research, the primary instrument for data collection is the inquirer him/herself. For Lincoln and Guba (1985), the human instrument is the instrument of choice regardless of any imperfections because its adaptability best meets the research requirements tied to the interpretive paradigm. My prolonged and persistent observation of the setting was a major source of information.

Prolonged engagement is necessary if one is to study a setting holistically. Although the study lasted only five weeks, being a constant presence in the class gave me a total of more than forty hours in this classroom setting. This was sufficient time for me to become familiar with the setting and develop what Erlandson, Harris, Skipper, and Allen (1993) call “shared constructions” with the study participants.

Persistent observation is related to prolonged engagement and it aims at providing depth to our investigations. It involves the researcher either as nonparticipant or participant observer of the setting. I assumed the role of nonparticipant observer when the professor was lecturing and leading discussions,

a role that gave me the opportunity to write fieldnotes. Being a nonparticipant observer helped me to get some general idea of the classroom setting and the interactions of the people in the setting. Gaining, however, a deep understanding of a phenomenon under study is hard to achieve from the detached perspective of nonparticipant observer. It is for this reason that my own role during the lab and other group activities was that of participant observer.

Participant observation means immersing oneself in the setting under study in order to closely examine the meanings people give to events and experiences in their social environment. It requires much more than mere observations. It necessitates direct involvement in the daily life of the people under study, getting a direct first-hand experience with events as they occur, listening to what people in the setting have to say, questioning people, "walking in the shoes" of people (Cantrell, 1990). I was interacting with the students while they worked individually or as groups on hands-on and computer-based activities. I was probing in order to get a better idea of their thinking processes and, whenever necessary, to help them reach solutions to their problems.

I exploited different forms of expression (discussed in Pratt, 1998) to help me get a rich picture not only of students' performance, but also of the thinking that stimulates or arises out of their actions:

(i) Discussions among the students;

(ii) Discussions between myself and the students which were often used to validate and probe more deeply into the thinking behind their actions and discussions;

(iii) The button clicks, menu choices and various ways of pointing on screen when using the computer learning environments.

Observing behaviors in their natural context and following these behaviors in detail over time served as check against selective perceptions, prejudice and bias (Cantrell, 1990). Participating in the daily routines of the setting allowed me to collect data on a larger range of behaviors, and develop ongoing relations and more open discussions with the students (Tsourvakas, 1997). These observations complemented formal assessments as a basis for instructional adjustments. The triangulation of sources gave me more confidence in my interpretations.

Fieldnotes

Fieldnotes are "the mainstay of qualitative research...a written account of what the researcher hears, sees, experiences, and thinks in the course of collecting and reflecting on the data in a qualitative study" (Bogdan and Biklen, 1982, p. 74). They begin with jottings (brief written records of impressions, key words and phrases) while in the field and end as expanded notes fleshed out after the field, and have both a descriptive and a reflective aspect. The descriptive aspect captures the details of the observations in a way that tries to give "as full and objective a rendering as possible of the subjects, dialogue, non-verbal communication, behavior, physical setting, events and activities" (Cantrell, 1990). Detailed, thick descriptions of the setting try to counter the fact that there is always subjectivity present when reporting on anything. The reflective aspect, which can be indicated either within the notes or as separate notes, focuses on the researcher as both a person and a researcher. The researcher tries to be self-

reflective and sincere about his or her personal relationship to the setting: his or her feelings, biases, reactions, prejudices, personal meanings. He or she also tries to keep an accurate record of the methods, procedures and evolving analysis, and to reflect critically on initial methodology and analysis.

Since the researcher is the main instrument of data collection and analysis, the role of fieldnotes is critical. The quality of the data depends on the depth and thoroughness of the fieldnotes. It is vital that the researcher "writes down in regular, systematic ways what she observes and learns while participating in the daily rounds of the life of others" (Emerson, Fretz, and Shaw, 1995, p. 1). For this reason, I took during the study every opportunity to write fieldnotes that preserved initial impressions and unique qualities of the setting before those became commonplace (Emerson et al., 1995). This was easier to do while assuming the role of nonparticipant observer during whole-class discussions, when I was able to take extensive notes in my journal. During lab activities I did not have much time for taking notes, I thus put down some jottings to remind me later of things I observed which I considered important to include in the subsequent analysis of the data. Also, since I was not able to take extensive notes during lab activities, all the lab sessions were audio-taped and transcribed. Some of the lab activities, as well as some of the whole-class activities, were also video-taped.

The fieldnotes I took while in the field were of a mostly descriptive nature. After finishing a day's entry in my journal, I would reread the fieldnotes in the evening and fill them in with additional phrases and comments. In order to

facilitate the process of data analysis I was, during the course of data collection, engaged in writing analytic asides, commentary, and in-process memos, which allowed me to capture on the spot valuable insights that could otherwise go unnoticed or fade away (Emerson et al., 1995). Asides were reflective comments inserted in the midst of descriptive paragraphs and commentaries were more elaborate reflections on some specific event or issue, contained in a separate paragraph (Emerson et al., 1995). In- process memo writing involved elaborating, while still actively in the field, the reflective comments included in asides and commentaries, in order "to flesh out ideas and tie them together, specifically as they pertain to emerging theories and patterns." (Cantrell, 1990)

As Emerson et al. (1995) argue, in-process writing should offer "probing reflections, tentative musings, and open questions" (p. 105). When writing commentaries and in-process memos, I tried to remain open-minded and avoid conclusive analytic statements in favor of possibilities and alternatives. I also used the member-checking technique to confirm, modify or correct my initial premises. Member-checking is a very crucial process that is essential to the qualitative research process. It entails asking the informants to verify their own realities and assure that the data obtained are accurate and the interpretations are plausible. The reactions of people close to the setting helped me check for "correctness and completeness" and recognize where the interpretation seemed "overblown or underdeveloped" (Wolcott, 1990, p. 132).

Documents

Any document germane to the investigation, such as instructor's records, files, journal articles, textbooks, and other materials used during the course, was also examined in order to provide additional information as well as to clarify or verify other data. The use of documentation provided a wealth of information, some of which was not accessible through observation or interviewing. Documents served to triangulate the evidence obtained from other sources.

Video-taping of Group Activities

Observing students working in small groups can be a very valuable source of information, since "in small groups, the articulation of students' voices is rich and revealing of their conceptions" (Confrey and Lachance, 1999, 250). Four group activities were carefully chosen to elicit the kind of responses and actions that would help bring to the front the students' mental mechanisms regarding variation. For each of those activities, I chose a group of students to observe closely. Although my initial intention was to follow the same group of students throughout the course, absenteeism and mobility among groups did not permit this. During the group activity, students were video-taped. I also closely observed the strategies students were using to pursue the tasks in the activity and followed their metacognitive processes. My role during the activity was that of a participant observer. My main aim was to allow students to be in control of their explorations, making decisions and moving in directions of their own choice. When students would be turning to me for an explanation of some phenomenon, I always tried to turn back their questions. In a few occasions however it seemed

appropriate for me to intervene. When actions of the students were less than transparent, I would intervene by asking questions that helped them express more clearly the reasons or intuitions lying behind their actions. These interventions were carefully recorded and have become part of the data analysis since they somehow modified and influenced the research context.

In addition to following the conceptual development of one group, by accompanying each of these activities with a worksheet which students had to complete individually, I was able to also collect data from the rest of the class. This allowed me to partially make up for the fact that following the conceptual development of a single group means giving up the chance to broaden the sampling of students' methods by observing multiple groups (Confrey and Lachance, 1999).

Pre- and Post-Activity Assessment

Low-stakes assessments involving short but nonetheless open-ended items were frequently given to students at several points during the course in order to monitor their evolving understandings. In addition, there were some longer, end-of-unit assessments.

Samples of Student Work

The course instructor used multiple methods to assess student learning, such as daily homework, open-ended and essay type questions in exams and quizzes, worksheets completed when engaged in hands-on and computer tasks, and reports on project assignments. At several points during the course, analyzing

samples of student work provided me with valuable information regarding their level of conceptual development.

Intermittent Interviews of Primary Group

These were short, unplanned, and unstructured interviews of students in the primary group that would take place during class whenever I felt I needed some clarification about an emerging issue.

Instructor

Open-ended interviews with the instructor were conducted several times during the course. The purpose of the open-ended interviews, which were audio-taped and transcribed, was to give the instructor the opportunity to express his opinion about the effectiveness of instruction: what aspects he considered to be successful, as well as what concerns he had. In addition to the formal interviews, there were also informal daily conversations with the instructor before and after class, during which collaborative decisions as to how we should proceed with the course were taken. I was writing fieldnotes describing our casual conversations and how my study design and his instruction were affected by the discussions.

Outside-of-Class Data Generation

Four of the students in the primary group and one other student agreed to participate in an outside of class intervention. I met with these students several times during the course. The meetings took place in the computer lab. Students worked either individually or in groups on technology-based activities that made use of the object-oriented computer learning environment Fathom. The activities

were carefully designed and structured to explore and at the same time support students' evolving meanings about the stochastic. During the activities, I assumed the role of a participant observant.

Using the computer as an expressive medium, I studied and analyzed students' actions articulated through button presses, choices from menus and changes to programming code (Pratt, 1998; Noss and Hoyles, 1996). This helped me identify the kinds of intuitions students use to make sense of stochastic phenomena and the ways in which their intuitions are shaped by technology. It allowed me to find out the structures that facilitate the articulation of intuitions about the stochastic and the forging of new connections between intuitions and formalisms (Pratt, 1998). In contrast to the heuristics research that examined how people make sense of stochastic phenomena at a specific point in time, I was able to gain more insight into students' thinking by examining how their intuitions evolved as they came in contact with technology.

My assessment of the effectiveness of technology as an aid to conceptual understanding tried to provide answers to the following questions:

(i) How do students' initial understandings evolve as they interact with the technology?

(ii) What aspects of a technological tool such as Fathom optimize the articulation of intuitions and the building of connections between the students' informal and formal understandings of variation?

(iii) When might technology be confusing?

C. End of Course

In order to assess students' understanding at the completion of the course, multiple forms of assessment were employed:

End-of-Course Questionnaire

The questionnaire included in Appendix D was administered to students in order to investigate their understanding of variation. Several of the items in the questionnaire were taken from a pilot investigation (Meletiou, Lee, and Fouladi, 1999) which had taken place a few months before the experiment began in an attempt to gain insights on students' informal experiences and understanding of variability. Three groups of students across two campuses had participated in that investigation: 44 students enrolled in a statistics course following the PACE model, 106 students enrolled in two non-PACE introductory statistics courses, and 102 students enrolled in a Human Sexuality course where the majority of students (69 students) had not yet taken any statistics courses. In reporting the results of the current study and comparing them to those of the pilot investigation, I use four categories: Non-Statistics (n=69), Non-PACE (n=139), PACE-Previous (n=44), PACE-Current (n=33). The Non-Statistics category includes those students in the pilot investigation out of the Human Sexuality class who had never taken any statistics course. Non-PACE includes both students in the pilot investigation who were taking a non-PACE course, as well as students in the pilot investigation enrolled in the Human Sexuality course who had taken statistics in

the past. PACE-Previous refers to PACE students in the pilot investigation, whereas PACE-Current to PACE students in the current study. In addition to including items from the pilot investigation, items from research conducted by other researchers were also included in the end-of-course assessment to allow additional comparisons with other populations of students.

Follow-up Interview of Primary Group

In this individual interview with each of the students in the primary group, we went over both the questionnaire they took in the beginning and the one they took at the end of the course, and also over some assessment tasks they had completed during the course. Students were probed to explain the reasons behind their answers to different questions. Also, they were reminded of some of the responses that they had given in order to see whether their reasoning changed in any way since then.

Interview of Instructor

An open-ended interview with the instructor was carried out, to get his opinion about the overall effectiveness of the course and how it compared with the kinds of experience his students had in previous semesters.

DATA ANALYSIS

Preliminary Data Analysis and Curricular Revision

In a transformative and conjecture-driven experiment, there are two types of data analysis (Confrey and Lachance, 1999). The first type is the ongoing preliminary analysis, taking place throughout the course, guiding instruction and

pointing towards necessary curricular revisions. This preliminary analysis, which begins simultaneously with the data generation process, "is necessitated by this design's anticipation of emerging issues" (Confrey and Lachance, 1999, p. 251). Throughout the course, I would meet with the instructor on an almost daily basis. Each time we met, I would try to make sure I presented him with some preliminary analysis of the data I had collected since our previous meeting, although it was not easy to keep up with all the new data coming in every day. The implications of the feedback gained from students guided our decisions as to how instruction should proceed and what modifications of our plans were necessary. In addition to substantial revisions of the curricular interventions, this initial analytical work of cycling back and forth the existing data also led to a revision and elaboration of the conjecture. Fledging hypotheses continuously got tested and evidence began to build (Cantrell, 1990). This analysis generated ideas for collecting new and often better quality data (in Cantrell, 1990).

Final Data Analysis

After the data collection stage was completed and all data had been generated and transcribed, the process of analysis continued in a more formal and explicit way. This final stage is the most time-consuming one, since at this stage the researchers "return to the data to attempt to construct a coherent story of the development of the students' ideas and their connection to the conjecture" (Confrey and Lachance, 1999, p. 255). In order to answer the research questions, I used a variety of both qualitative and quantitative data analysis techniques.

Qualitative Data Analysis Techniques

I subjected the raw data to an inductive data processing method called constant comparison analysis (Glaser and Strauss, 1967). This method is concerned with "generating and plausibly suggesting (but not provisionally testing) many categories, properties, and hypotheses about general problems" (Glaser and Strauss, 1967, p. 104). It is designed to aid the analyst generate a plausible and consistent theory, which stays close to the data. The constant comparison method, which involves unitizing, categorizing, chunking, and coding by choosing words, phrases, or sentences that specifically address the research questions, assisted me in the search for patterns and themes that were used to develop the study's interpretation. Once recurring patterns and themes in the data had been identified, they were compared across classifications, and categories collapsed, merged, or were redefined. I developed working hypotheses accordingly by noticing similar patterns across data. The hypotheses were being modified and refined continuously. I reached closure only after many sweeps through the data, which helped me eventually achieve some degree of theoretical saturation. Throughout the coding process, I continued writing memos which helped me keep track of all the categories, properties, hypotheses and generative questions evolving from the analytical process (Pandit, 1996).

During the final data analysis, I was taking measures to avoid the researcher's tendency to quickly reduce the data by focusing only on what is familiar and central to the study. Such a tendency might lead the investigators to miss the opportunity "to know what might not be known to them prior to the

study", to overlook the fact that "the margins of a project often provide some of the most interesting and informative patterns for investigators" (Chenail, 1992, p. 44). In addition to looking for cases that illustrated recurring patterns of behavior and typical situations in the research setting, I was also looking for departures from those patterns. I employed the Chenail Qualitative Matrix (Chenail, 1992), which gave me the opportunity to "discover the serendipitous or unexpected instead of staying focused only on what was known through literature searches and previous observations" (Cole, 1994). The Chenail Matrix has two dimensions, the first dimension covering the Central Tendencies-Ranges spectrum, and second dimension the Expected-Unexpected spectrum. The Central Tendencies-Range spectrum served in reminding me that in addition to describing how the data chunk together into common themes and categories, I also had to describe differences within those themes and categories. The Expected-Unexpected spectrum was set to organize the data presentation, with Expected referring to data which confirmed my assumptions and the findings described in the literature review, while Unexpected referring to data that defied pre-set assumptions and previous research finding.

Quantitative Data Analysis Techniques

In addition to performing qualitative data analysis of the written assessments, I sometimes also looked at them in purely quantitative terms (Confrey and Lachance, 1999). For example, I would draw conclusions about the performance of the class as a whole, or I would make comparisons of the class

performance to the performance of other groups of students who had taken the same assessment task.

As has been already noted, several of the items at the end-of-course assessment were taken from a pilot investigation (Meletiou, Lee, and Fouladi, 1999) that was conducted to gain insights on students' informal understanding of variability, as well as from studies conducted by other researchers. Quantitative analysis was used to compare the performance of students in the current study, to that of students in other studies. Linear model methods were employed using the CATMOD procedure (SAS Institute 1988) which, by default, treats all variables as categorical. These methods are a natural extension of the usual *Analysis of Variance* approach to continuous data. The permitted us to investigate the effect of the course students belonged to (the explanatory variable) on the probability of success in a question (the response probability).

CRITERIA FOR QUALITY OF RESEARCH FINDINGS

While any researcher should strive for results that others would consider rigorous and trustworthy, criteria for assessing those qualities differ depending on the nature of a study. Positivists typically speak of validity, reliability and objectivity when assessing the worth of a study. Based upon the underlying assumptions, the novelty and the emergent nature of a transformative and conjecture-driven teaching experiment, these concepts do not seem to transfer directly. Adhering to the rigid principles of traditional research would not fit this kind of inquiry with its evolving conjectures and shifts in curricula (Confrey and

Lachance, 1999). Validity for example, is not relevant for this type of research where "there is no exact set of circumstances, no single and "correct" interpretation." (Wolcott, 1990, p. 144)

Despite the unconventionality of the conjecture-driven research model, one might still do powerful and relevant research. As Confrey and Lachance (1999) point out, the fact that this kind of research is guided by an explicitly stated and well-developed conjecture makes it worthwhile and significant. However, in order to ensure that the development of the conjecture was indeed guided by the data and not by some independent agenda, the researcher must do a careful demonstration of the quality of the study (Confrey and Lachance, 1999). I will next briefly discuss some of the strategies I have used to ensure the rigor, worth, and trustworthiness of my research and its findings. Following Confrey and Lachance (1999), I employed standards that enhance the quality of both the internal processes of the research and its potential impact on educational practice. These standards address issues related to the components of the research model: the ideological stance, the theoretical framework, and the dialectical interaction between conjecture and intervention.

Ensuring the Quality of the Internal Processes

Perhaps the most important aspect of the conjecture-driven research process that needs to be evaluated in terms of its internal consistency is the explanatory power of its conjecture (Confrey and Lachance, 1999). It is the conjecture, which originated from dissatisfaction with the way typical practices treat a certain set of mathematical topics, that drives the experiment. Therefore,

one of the main questions this type of research must answer is: "Can the conjecture point to a better way to reconceptualize this set of topics that will allow *all students* to construct an understanding of these concepts?" (Confrey and Lachance, 1999, p. 259). This question leads Confrey and Lachance (1999) to identify three targets for evaluating the quality of the conjecture-driven research.

The first target is evaluating the quality of the conjecture in terms of its face validity in relation to peer review (Confrey and Lachance, 1999). An audience of researchers or practitioners can assess the face validity of the conjecture by analyzing both its content and its relationship to the research literature. Hopefully, through the careful elaboration of the conjecture, its content and pedagogical dimensions, and the theoretical framework in which they are situated, I have provided the reader with enough information to be able to make such a judgment in relation to the findings of the study (Confrey and Lachance, 1999).

The second target is judging whether the research process results in a rational reconstruction of the dialectical relationship between the conjecture and the events taking place in the classroom. Audience needs to be provided with evidence of the research process that will allow them to answer the questions: "How closely are the two forces in the dialectic interwoven? Does it result in a coherent story?" (Confrey and Lachance, 1999, p. 260). I have provided ample of evidence in terms of both preliminary and final data analysis to ensure this.

Finally, the study should be judged in terms of its fidelity to its ideological stance. A study which claims to put student voices to the front, should make sure

"student expressions are extensive and authentic enough to convince a reader of the depth of the students' commitment to and ownership of the ideas" (Confrey and Lachance, 1999, p. 260). I have taken all the steps recommended by Confrey and Lachance to allow the emergence of students' voice. I have provided "ample data in the form of quotations from the students and examples of their work in discussions of the research" (Confrey and Lachance, 1999, p. 260). I have included sufficient information "about the characteristics and contexts of the student-speakers, along with comments about how representative of the various student groups a given set of interactions is" (Confrey and Lachance, 1999, p. 260). I have made sure the data I presented came from a wide spectrum of the class population "to demonstrate that the educational benefits have been experienced widely" (Confrey and Lachance, 1999, p. 260). I have also presented the results of many assessment tasks that support this claim.

Since claims for the quality of the internal processes of a conjecture-driven research design are based on interpreted data, one might question the trustworthiness of the data presented. The research report is an interpretation of reality, and the readers might legitimately ask: "How do we know that the researcher saw what she wanted to see or only paid attention to the data that supported her conjecture?" (Confrey and Lachance, 1999, p. 259). The quality of this kind of research is very dependent on how critically the researchers reflect upon the data and challenge themselves about the soundness of their evolving understandings. Therefore, the researcher should provide evidence that re-assures the audience they can trust the methods employed and the interpretations made.

The *credibility*, *dependability*, and *confirmability* of the data obtained from a conjecture-driven research needs to be assured (Confrey and Lachance, 1999; adapted from Guba and Lincoln, 1989). Also, since it is through the interaction of students' voices and the researcher's perspective that the conjecture evolves, the *role of the researcher* should be described.

Credibility

In judging the value of qualitative studies, the correspondence version of truth is replaced with the idea of credible or trustworthy accounts of multiple constructed realities (Lincoln and Cuba, 1985). The researcher must try to show that his or her reconstructions are credible to the constructors of those multiple realities. Erlandson et al. (1993), and Lincoln and Guba (1985) have suggested several different ways of establishing trustworthiness, or credibility. I have used (1) prolonged engagement, (2) persistent observation, (3) triangulation, (4) member-checks, and (5) peer debriefing to enhance the credibility of my study. The first four strategies have already been discussed in the section on data generation. The fifth, peer debriefing, involves sharing preliminary findings with colleagues. Besides Dr. Lee, I also shared developing manuscripts with colleagues who are experts in the fields of mathematics education, statistics education, statistics, educational psychology, and qualitative research methods, who have provided valuable feedback, especially about the interpretations I was constructing from my informants.

Dependability

Dependability is the qualitative parallel to reliability (Guba and Lincoln, 1989). It is the use of sufficient methods and techniques to assure that the study's results can be trusted. In the postpositivist paradigm, reliability means stability over time. However, in a conjecture-driven research experiment changes in the design resulting by increasingly refined understanding of the setting (Marshall and Rossman, 1995) are expected and acceptable. Nonetheless, a publicly documentable record of the change process is required so that "outside reviewers…can explore the process, judge the decisions that were made, and understand what salient factors in the context led the [researcher] to the decisions and interpretations made" (Guba and Lincoln, 1989, p. 242; in Confrey and Lachance, 1999, p. 262). In order to account for the ever-changing context within which this research occurred, I have kept a detailed record of the changes that occurred in the setting and how those changes affected my methodological and analytical decisions.

Confirmability

Confirmability refers to the degree to which the results of the study could be confirmed or corroborated by others. It is the ability for others to examine all data sources and processes to assure that the findings are grounded in data and are not figments of the researcher's imagination (Lincoln and Guba, 1985). Thus, the criterion for judging the confirmability of a study is the degree to which the data confirm the general findings and implications of the study (Marshall and Rossman, 1995). Following Marshall and Rossman's (1995) advice, I have kept a

journal that explicates all the important design decisions taken during the study and the rationale behind them, so that others can judge if they were adequate and made sense. In addition, by keeping all the data collected in a well-organized and retrievable form, I can easily make them available to any researchers challenging the findings and wanting to reanalyze the data.

Role of the Researcher

There were multiple voices in the research setting. Each individual student, their instructor, and I, had our individual perspectives on each situation. Keeping this in mind, I tried to take different observational positions and address both my views, as well as the views of the students and their instructor.

Wolcott (1990) recommends that the researcher should "talk little, listen a lot". He points out that, by talking too much and hearing too little, many fieldworkers "become their own worst enemy by becoming their own best informant" (Wolcott, 1990, p. 128). He warns educational researchers that this "is especially serious problem in school research, where we often presume to "know" what is supposed to be happening and consequently may never ask the kinds of questions we would ordinarily ask in any other research setting" (Wolcott, 1990, p. 128). One method I employed to help make members' views more clearly heard is a method of analysis developed by Confrey (1994) called *Voice and Perspective*. This method has two stages. At the first stage, the researcher articulates the participants' voice through what Confrey calls "Close Listening". At the second stage, the researcher affirms his or her own perspective and how it has been influenced as a result of his or her interactions with the people in the

setting. The structure of the course was ideal for employing Confrey's method. Through providing an environment which gave validation to the students' personal voice (Confrey, 1991), it enabled me to get valuable insights into students' thinking processes that would not have been possible under more conventional instruction.

I acknowledge that both the participants' views and my views were affected by my interactions with the people in the setting. As people interact with each other, there is an "ongoing resocialization", a continuous adjustment of prior views (Eisner, 1985). Whenever I describe members' meanings, these descriptions are interpretive constructions that represent my knowledge and understanding of the participants' experience. However, my prolonged involvement has influenced the perspective from which I am reporting and has acted as warrant against the imposition of "exogenous categories and meanings" and of "a priori theoretical categories" (Emerson et al., 1995, p. 111). Active engagement in the setting has given me the chance to look closely at what students said and did and to record the words, phrases and categories that they used in their everyday interactions. It has enabled me to "attend consistently to members' meanings and concerns" and develop descriptions and analyses that are "sensitive to local concerns, meanings and categories" (Emerson et al., p. 111, 1995). Because no event has a single or invariant meaning, I was constantly trying to put aside my inclination to assume that I knew "what significance members attribute to the events and objects that make up their world" (Emerson et al., 1995, p. 114). I was observing closely and I am documenting multiple stories

in order to examine the different ways in which different members constructed and made meaning of the same event.

Since attitudes and orientation toward the topic or the people studied affect what each researcher writes down, there exists a necessity of including the researcher in one's research. Because "every research report tells a story, and every story has a storyteller" (Christie, 1997), I acknowledge that what I am reporting is not an objective story but an interpretation of reality reflecting my personal worldviews. For this reason I have tried to explicate my beliefs and theoretical stance. Also, at each point where I feel that my personal feelings and reactions are relevant, I try to be open about them (Wolcott, 1990).

Recognizing my personal biases has, hopefully, made me more sensitive to the ways in which my views shaped my interactions with my study participants. Such recognition, according to Emerson et al. (1995), "can better guard against any overriding, unconscious framing of events" (p. 43). I have tried, to the degree that this is possible, to avoid assuming that other people think the way I do and to impose my judgments. I was constantly striving for some balance between my personal impulses and the need for students' voice to be heard. Hopefully, I have achieved what Wolcott (1990) calls rigorous subjectivity which encompasses "elusive criteria like balance, fairness, completeness, creativity" (p. 133). This rigorous subjectivity is much more preferable for me than a detached objectivity. As Wolcott (1990) points out, it is not knowing, but understanding that captures the essence of things and what we should be after: "'To understand', he posits, 'it may not be enough to know.'" (p. 147)

A researcher cannot understand the people she or he is researching unless immersing into their world and trying to make sense of it. As Greene (1994) explains, "it is precisely the individual qualities of the human inquirer that are valued as indispensable to meaning construction" (p. 539). My personal investment in the research setting was a self-conscious choice, and I consider the bias that comes with my being the main measurement instrument as a great asset rather than a weakness of my inquiry.

Assessing the Potential Impact

Bridging the gap between research and practice has never been efficient or easy (Confrey and Lachance, 1999). However, since one of the main objectives of this type of research is to better connect research and practice, assessing the potential impact that this research has on bringing about “achievable” change in the statistics classroom is an imperative (Confrey and Lachance, 1999).

The researchers should find multiple ways of disseminating research findings to educators, students, and other people who initiate changes in the system. This will ensure that a diversity of audiences gets involved in further elaborating the conjecture. It will also ensure that “the dialectical relation between conjecture and instruction carried out in the experiment is reproduced between the practitioner and the researcher as research results are prepared.” (Confrey and Lachance, 1999, 258)

I perceive several potential products originating from this research. Through publications and conference presentations, I will inform other researchers about the study and its findings. I also plan to use the insights

obtained in this study to develop curricular and professional development materials. In my future plans I include a plan to develop and implement a course or a series of workshops for elementary/high school teachers that would aim at improving both their content and pedagogical knowledge of statistics. Since students first introduced to statistics learn similar concepts and procedures regardless of age, findings of a study regarding one age group have implications for instruction with other types of students (Gal and Garfield, 1997).

Confrey and Lachance (1999) identify five criteria for assessing the potential impact of the research products of this type of research. The first criterion is *feasibility*. The implementation of the research products should not make excess demand on financial and human resources, so that they can be implemented and useful for all classrooms. The second criterion is s*ustainability*. The impact of the research products should be endurable and sustainable for a considerable amount of time. In addition, research findings should be *compelling*. The evidence of the research findings should not only attract the interest of practitioners, but its magnitude should be such as to convince them of the urgent need for change. Research products should also be *adaptable*. They should be flexible enough to be applicable to diverse populations and variety of settings. Finally, they should be *generative*, becoming "models for innovation" for practitioners, providing them with "a powerful means of reconceptualizing a variety of classroom events, relationships, and practices." (Confrey and Lachance, 1999, 264)

As Confrey and Lachance (1999) note, "because research has been relatively isolated from practice in the past, these criteria appear too demanding of any one team of researchers" (p. 264). I am convinced of the value of the findings of the study; at this point in time however I cannot fully assess these criteria. The fact that the conjecture guiding this study was tested and refined in a real classroom is a big advantage compared to studies that draw their conclusions by taking snapshots of students' thinking. Nonetheless, my experiences at the setting were unique and could not be replicated. Bowen (1997) stresses the need for reconceptualizing generalizability when dealing with qualitative studies such as mine. Transferability - the degree to which two contexts are congruent - proposed by Lincoln and Cuba (1985) is one such reconceptualization. Thick description of the setting enhances the transferability of my study. Although I acknowledge that this study focuses on a single classroom with unique characteristics, I still believe that the experiences and insights gained can be powerful and relevant for other statistics educators also. By giving a detailed description of the research setting and participants, I allow the readers of this piece of work to determine intuitively the "fit" of my study with their own settings or contexts.

I am well aware of the fact that raising students' awareness of "probabilistic interpretation" and helping them see and feel how stochastic thinking is related to causal and logical thinking (Pfannkuch and Brown, 1996) is not an easy task and takes time to achieve. It is not easy for students to shift from a deterministic view of reality to one that balances deterministic and probabilistic reasoning. Probabilistic thinking is an inherently new way of looking at the world

and, to really learn to thinking in probabilistic terms, students need to undergo a revolution in their thinking (Falk and Konold, 1992). However, I am convinced that providing a learning environment that emphasizes the omnipresence of variation allows students to make some steps beyond their deterministic thinking and start developing their statistical reasoning. I hope that the evidence I provide in this study convinces the reader also that this is the case.

In the next chapter, I outline the findings from the assessment given to students prior to instruction and the follow-up interviews of the primary group. I then discuss how the insights gained led to elaboration of the conjecture and the instructional design.

Chapter IV: Assessment Prior to Instruction

INTRODUCTION

In order to be able to follow the students' conceptual development process, good understanding of their thinking prior to instruction is required. The pre-assessment (Appendix A) on variability given on the first day of class and the follow-up interviews of the primary group (interview protocol in Appendix A), allowed a thorough investigation of student reasoning. A small discussion of the results and the implications for instruction follows.

DISCUSSION OF RESULTS

The first question in the pre-assessment asked students to describe based on their experience, what variability means. Several of them defined variability in ways that suggest they viewed it as variety, or as something that takes multiple values: *"Having more than one choice of something or many choices"; "Different things, numbers, amounts."* Others described variability not simply as variety, but also as a measure of how things differ: *"How much or by what something is probable to change/vary from something else."* Some other students described variability as *"range of something… from minimum to max."* Still, others seemed to equate variability with the mathematical notion of variable: *"A variable is something that is not constant. Ex. 2(x) +3, x is the variable as it could be a variety of numbers."*

Students gave very reasonable responses to all five parts of Question 2 of the pre-assessment, where they had to decide whether it was more desirable for

variability to be high or low in each of five different cases. Their responses indicate they recognized that low variability could be a good or bad quality depending on the context of the situation. Twenty-seven students (90%), for example, understood that variability in the diameter of new tires coming off one production line needs to be as low as possible *"so that there are no problems with different sized wheels unbalancing a car."* In contrast, 80% of the students thought it would be preferable for scores on an aptitude test given to a large number of job applicants to have a high variability because this would aid the screening process by making it easier to select the most qualified applicants. The reasoning of the four students (13%) who thought it is better to have low variability of scores, was not faulty either: *"One would hope that the scores would all be in the high range."* Finally, one student who argued that high variability could be either good or bad also gave a good justification: *"A larger range would indicate clearly who will be the best applicant and a small range would mean applicants are all bad or all good."*

In the follow-up interview, several students made remarks indicating some knowledge of quality control processes employed by companies to minimize variability and ensure products stay within specification limits. Tim for example, realized that *"it's very hard to produce anything that is exactly the same every time...even if it's very precise, it can't be perfect."* However, he stressed, companies *"need to have someone who gauges what is permitted and what is not, because certain things have to be close together, otherwise it won't work."*

In Question 3 of the pre-assessment, students were given the sets of scores of two statistics students and were asked to choose a study partner. The aim was to see whether they would notice the difference in the amount of variation between two sets of scores having the same mean. From their responses, it seems they did. Fifteen students (50%) noticed the smaller variation in Student A's scores and chose this student as a study partner "*because although both have the same average, Student A is more consistent with the scores*." On the other hand, ten students (33%) picked Student B as a study partner, viewing his "perfect" scores as evidence of his/her potentials: "*His range is from 40 to perfect…straight As, so he'd do better given the right motivation.*" Finally, five students (15%), although realizing that the two sets of scores differ in variation, thought that in the end "*it wouldn't really matter who you study with since they both are essentially the same grade standing of 74%. They compensate each other.*"

An important first step in data description is assessing shape. Graphs are, along with numerical means, the main statistical tools used to assess the shape of a data distribution. Histograms are among the most important graphical tools used in the statistics classroom. Question 7 in the pre-assessment (taken from Garfield, delMas, and Chance, 1999), was given in order to see whether by looking at the histogram of two distributions of scores students could figure out which of the two distributions has more variability (Figure 4.1):

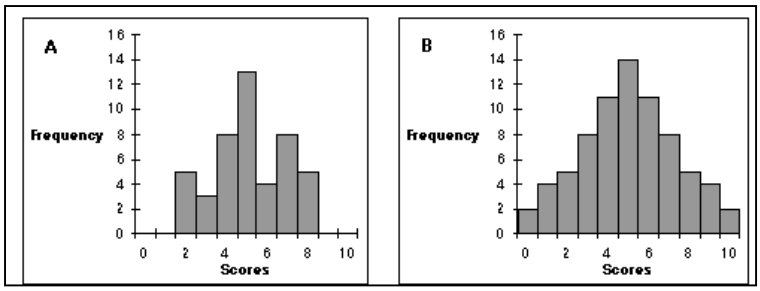


Figure 4.1 – Histograms of Distributions A and B

Twenty-two students (71%) recognized that distribution B has more variability than distribution A. Eight students (26%), however, thought A has more variability. Finally, one student checked both distributions, arguing that *"A has more variability because it's bumpier, whereas B has more variability because it's more spread out and has a larger number of different scores."*

The purpose of the Question 10 (taken from Scheaffer, Gnanadesikan, Watkins, and Witmer, 1996) was, in addition to investigating students' familiarity with histograms and bar graphs, to see how well they could relate features of a distribution to the shape of a graph. Students had to match the following list of variables and set of histograms using their knowledge of the variables:

(i) age at death of a sample of 34 persons
(ii) the last digit in the social security number of each of the 40 students
(iii) scores on a fairly easy test in statistics
(iv) height of a group of adults
(v) number of medals won by medal-winning countries in the 1992 Winter Olympics

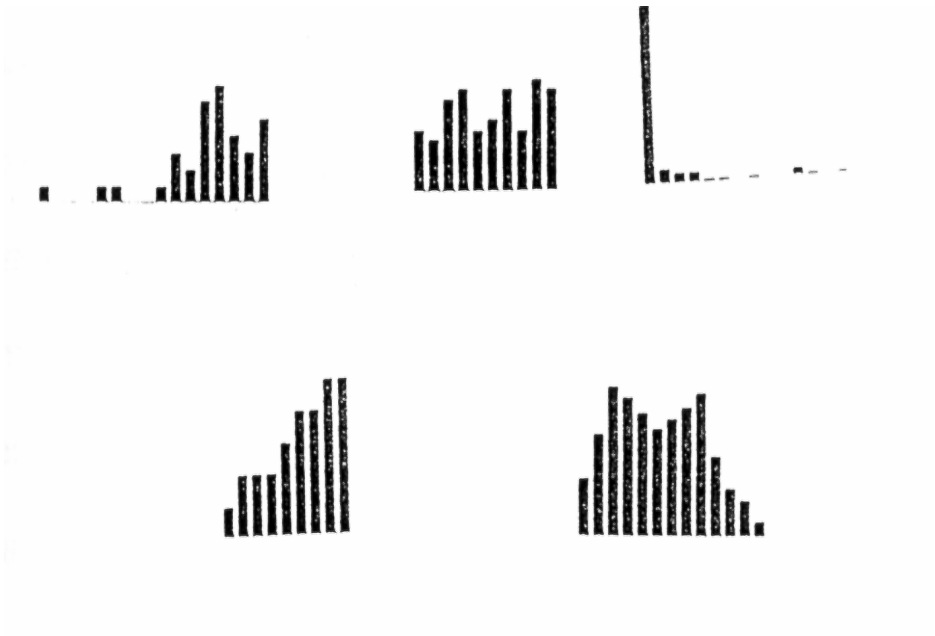


Figure 4.2 – “Matching Histograms to Variables” Task

Table 4.1 – Pre-assessment Results on “Matching Histograms to Variables” Question

Answer	% of Correct Responses
Graph 1	33
Graph 2	33
Graph 3	20
Graph 4	36
Graph 5	30

Only three students (10%) correctly matched every variable to its corresponding graph. For Graph 1, ten students (33%) chose Variables A (age at death) or C (scores on a fairly easy test), which are both reasonable matches. Ten students (33%) correctly matched Graph 2 with Variable B. The 10 bars of the graph seems to be the only reason most of them did since just a couple of students made reference to the relative uniformity of the different outcomes. Only six

students (20%) matched Graph 3 with the number of medals own by medal-winning countries in the 1992 Olympics (Variable E). Even they were not totally correct, since they perceived each bar of the graph as representing the number of medals won an individual country, arguing that in the Olympics, "*not many countries win a high number of medals",* and "*usually one country does much better than the rest."* Eleven students (36%) correctly matched Graph 4 with either Variable A or Variable C. Finally nine students (30%), realizing that the height of a group of adults would have a distribution that is "*higher in middle range, low on extremes",* matched Graph 5 with Variable D.

Especially noticeable was many students' tendency to perceive the graphs as displays of raw data (i.e. as representations similar to dotplots, with each bar standing for an individual observation) rather than as presenting grouped sets of data. For example, several students (23%) matched Graph 1 with variable E, because they perceived each bar of the graph as representing the number of medals own an individual country: "*Some countries win a lot – some a few – and others win a zero."* Similarly, 7 students (23%) matched Graph 5 with Variable C, thinking of each bar of the graph as representing the score of an individual student: "*Almost all test scores are on a high range, with a few remaining low"*. They did not seem aware of the data reduction involved.

One of the only three students who correctly matched all graphs with their variables was Andrew. When I asked him in the follow-up interview whether he knew anything about histograms before taking the class, he responded:

> No, I first just looked at them and said when you look at a bar graph it usually goes …I started out by looking at the 5 examples and making my opinions/thoughts on what the graphs would look like. Then one by one I placed the graphs that looked the most like what I thought they would look like with the 5 labels. I then matched the labels with the variations/ranges and frequencies that I thought would match with the graph.

In contrast to Andrew, most of the other students, whereas realizing that there is variation and thus each variable will assume different values, seemed not to seriously concern themselves with the patterns that emerge within the variation. Of course, the fact that the question was given before most students had ever been formally introduced to histograms is a main reason they did so poorly. Nonetheless if they had thought carefully about how to relate specific features of each variable with the shape of the graphs, they would have made more correct matches like Andrew, who was also unfamiliar with histograms and thought of them as bar graphs. In trying, for instance, to decide which variable matches Graph 2, 5 people (17%) thought this graph represents the distribution of the height of a group of adults (choice D), because *"heights fluctuate a lot."* It is obvious that these students did not seriously think about how people's heights fluctuate, otherwise they would not have matched height with an approximately uniform distribution.

Question 6 (adapted from Jacobs, 1997), investigated students' informal understanding of sampling issues. They had to comment on each of the following six different approaches to conducting a survey by students of a middle school who were trying to estimate how many kids of the whole school would be interested in buying a raffle ticket to win a SEGA video-game system:

Survey 1: Tom asked 60 friends (75%, 25% no).

Survey 2: Shannon got the names of all 600 students in the school, put them in a hat, and pulled out 60 of them. (35% yes, 65% no)

Survey 3: John asked 60 students at an after school meeting at the Games Club. The Games Club met once a week and played different games-especially computerized ones. Anyone who was interested in games could join (90% yes, 10% no).

Survey 4: Ann sent out a questionnaire to every kid in the school and then used the first 60 that were returned to her. (50% yes, 50% no)

Survey 5: Claire set up a booth outside the lunchroom and anyone who wanted to could stop by and fill out the survey. To advertise her survey she had a sign that said: "WIN A SEGA." She stopped collecting surveys when she got 60 completed. (100% yes)

Survey 6: Kyle wanted the same number of boys and girls and some students from each grade. So he asked 5 boys and 5 girls from each grade to get his total 60 students. (30% yes, 70% no)

Surveys 1 and Surveys 3 involve restricted sampling procedures. No student liked Survey 1, because *"it would only show the likes of Tom's friend"* and all but one student found Survey 3 to be very poor, since *"the Games Club is a directly related interest group to Sega."* Several students described the method of sample selection employed by both surveys as biased. A few others pointed out that the sample was not randomly selected. One student, possibly perceiving variability as sample representativeness, did not like either of the two surveys because they led to decreases in variability: "*Friends like the same things, which implies decrease in variability. It does not give a good picture since he only did his friends that may only represent one grade and sex."*

Surveys 4 and 5 utilized self-selected sampling procedures. Sixty-percent of the students argued that Survey 4 does not represent the whole school population because *"the first 60 respondents could have been eager to say yes"*, 30% however considered it to be *"a good survey"* that gives a *"fairly good picture."* Four of these students characterized the sample as random and non-biased since *"everyone in the school was given the same opportunity."* More students were able to detect the dangers of self-selection in Survey 5. Everyone but two students considered it to be a very poor survey that attracts only those who are interested in winning a SEGA: *"Win a Sega" slants participants immediately. Their interest in the Sega brought them to the booth. Not random."*

Surveys 2 and 6 involved random sampling procedures. Almost all of the students viewed positively Survey 2, which used a simple random sampling scheme: *"It was a good way to get an unbiased account…they just pulled them out of the hat…this had nothing to do with them."* Nonetheless, a few of the students that approved Survey 2, still seemed concerned that simple random selection might not lead to a representative sample. There were also four students who did not approve the way Survey 2 was conducted and argued that, due to its randomness, extreme outcomes are possible: *"She could pull out 50 girls and 10 boys and usually girls don't like video games as much as boys do."*

All but two students approved the stratified random sampling scheme employed in Survey 6, emphasizing the *"good diversity in age and gender"*, the *"big variety of people"* that is guaranteed by this method. Some of the students praised the study because it is *"random and a good representation of population."*

A couple used the word variability having again the connotation of sample representativeness to express this idea: "*Good student variability – both sexes and all grades, therefore got good overall picture."* Three students did point out that Survey 6 "*assumes that each grade has equal number of students and boys and girls are also the same in number",* and therefore its quality depends on whether these assumptions hold. In the follow-up interview, I found that students had a good idea of what one has to do to collect a stratified sample when the different strata represent unequal proportions in the population.

The second part of Question 6 was asking students to choose among the six surveys their preferred one. Almost all of the students showed preference for random sampling procedures. Forty-six percent of them chose the simple random sampling method (Survey 2) and 38% the stratified random sampling method (Survey 6), arguing that giving everybody the same chance to be selected should result in a sample more representative of the school. Four students (15%) choose a self-selected method (Survey 5 or Survey 4), with the reasoning that giving everybody the chance to participate would "*show how many people were truly interested."* No student expressed preference for a restricted sampling method.

In the last part, which was asking students to give the best estimate of the proportion of children that will be buying a raffle ticket, several used their personal judgment and ignored the results of the surveys altogether. For example, three of the students that had chosen Survey 6 did not seem to have taken the results of that survey into any account. One wrote "*50-50*", the second one "*40% because less than half usually care for a particular cause"*, and the third one

"62% will buy, 38% won't". Andrew, who was this third student, said in the follow-up interview: *"There is no way to compute it. I just looked and estimated it."*

One of the main things I investigated in the beginning of the course, was students' informal understanding of issues related to sampling variation. Several questions in the pre-assessment of the whole class and the follow-up interview of the primary group called students to make likelihood judgments involving stochastic events in order to get insights into their intuitive notions of randomness and chance variation. In Question 8 (taken from Pfannkuch and Brown, 1996), students were told that a gambler has observed the ball landing on red six consecutive times in a roulette wheel that has 18 black and 18 red numbers, and were asked to predict the next outcome. Only 20% of the students thought black and red are equally likely. Most of the students (67%), excepted black to be the next outcome for things to balance out to better represent the population distribution: *"If red or black have same probability then black is overdue."* Four students (13%), though acknowledging the independence of random events, still found it hard to accept that red is as likely to come up as black: *"Red or black a 50/50 chance either way, I would bet on black though. Seems less of a chance to have 7 red in a row."*

Question 4 in the pre-assessment (adapted from Rubin et al., 1990) examined how students balanced the ideas of sampling variability and sampling representativeness. Students were told that the Easter Bunny was distributing many packets of 6 Gummy Bears at the Easter Parade which he had made up by

grabbing handfuls of Gummy Bears out of a large vat containing two million green and one million red Gummy Bears. They were first asked to estimate the number of green Gummy Bears in a packet. Everyone gave *"4 green, 2 red"* as the estimate. In the next part of the question, where students were asked whether they thought all kids got the expected number of greens, all of them realized that *"not every student got exactly 4 green every time because there's variability"*. They intuitively understood that probability is the limiting relative frequency, which only approximately holds for real data:

> That is just the mathematical way of figuring it, that number will fluctuate.
>
> That is just the probability, the most likely not an exact answer.
>
> Expected ratios are a general rule, not a formula for each individual occurrence.
>
> It is nearly impossible for the ratio to hold perfectly, unless the Easter Bunny uses his Easter magic.

Students recognized that random selection leads to variation: *"There will be a variation on the pattern of green bears in each bag, because of the random grabbing of the beans when they were placed in the bags."* However, when asked to estimate the proportion of packets with 4 greens, almost all of them underestimated the effect of sampling variability and greatly overestimated this proportion. Only two students gave estimates that came close to 33%, the actual probability of 4 greens (found by modeling the situation as a Binomial distribution). The estimates that the rest of the students gave, ranged from 50%-92%. Several students wrote that they expected 66% of the packets to have 4 greens in them. Rubin et al. (1990) who gave this question to high school seniors,

also noticed that “the population percentage itself (66%) seemed to influence their estimate, as they appeared to translate the preponderance of greens in the vat to a preponderance of people getting the representative sample.” (p. 8)

The last part of the question reminded students that the Easter Bunny had started with 2 million greens and 1 million reds and asked whether they thought he ran out of one color long before the other one or whether both lasted until near the end. Three students responded that he should have run out of red first, since they were fewer to start with. All the others understood that, provided that the sample was *“properly mixed and random”*, *“both should last near the end because when filling the bags, on average, you will have twice as many reds.”*

The conclusions drawn from this question are similar to those drawn by Rubin et al. (1990), who found that students answered this question by focusing on samples that mirrored the population proportion of 2G:1R. They over-relied on sample representativeness, underestimating the frequency of samples near the tails of the distribution and overestimating the frequency of the modal sample.

The tendency to underestimate the effect of sampling variability and expect small samples to match population properties was also witnessed in Question 9 of the pre-assessment, taken from Garfield and delMas (1990). The question described how a worker of a student organization went about conducting a survey at a certain college where half the students were women and half were men and the several measures he took to ensure good representation of all students. Students were told that out of the last 20 students interviewed, 13 were women and 7 were men and were asked whether they thought there would be

more women or more men in the next 20 students interviewed. Only 35% chose response C which stated that one should expect about an equal number of men and women, since who has been selected so far does not affect who will be next selected. Thirty-two percent of the students argued that since more women than men were selected so far, they expected the opposite trend to start happening (Response B). Another 16% tried to find causes behind a difference that, given the small number of people interviewed this far, could be easily explained by chance variation (Response A). Another 10%, employing the "law of small numbers", thought that since the trend till now has been more females than males, this trend should continue (Response D). Two students chose E and gave explanations suggesting reasoning similar to that of students choosing D.

The tendency to underestimate the role of chance variation was also partially observed in student responses to Question 5 of the pre-assessment (adapted from Pfannkuch and Brown, 1996):

On average there are 600 deaths due to traffic accidents each year in a city. A person in the city observed the following:

February	Number of deaths
Week 1:	3
Week 2:	12
Week 3:	21
Week 4:	14
March	
Week 5:	2

Assume that none of these weeks contain a holiday weekend. Suppose the headlines in the newspaper claimed that week three was a "disastrous" week and police reported that speed was a factor. The next week was described in the papers as more evidence that the city driving was deteriorating. At the end of week five the police congratulated themselves for the low death rate - their extra patrols had succeeded. What would you say to this person?

Some of the students in the class saw chance variation as the mere cause of the low rate of deaths in Week 5: *"Real life occurrence has a high variability. Deaths don't occur on a quota or required rate."* A couple of students noted that, before congratulating themselves, the police should *"take a look and see that there was a similarly lower death rate in the first week of February as well."* Several students pointed out that the number of weeks was too small and one should *"wait to see what the results are for the next few months before jumping to a conclusion"*. About half, however, of the students in the class gave quite deterministic responses and tried to find reasons to explain the drop in the death rate: *"Patrols may have slowed drivers down by giving them tickets. Many drivers may have been reading the papers and decided to slow their driving; not drive so fast."* Anna was one of the students who thought there must be some reason for the drop in the number of accidents. In the follow-up interview, she repeated this conviction:

> 3 to 21 is also is a big difference. All of a sudden it just sky-rocketed and then all of a sudden it just dropped, so you want to hope that more patrols helped but there is a bunch of other reasons. They cut down on speed, but may be they didn't cut down on that…drunk driving and other things.

Unlike the deterministic mindset with which many students approached the two previous questions, in the follow-up interview I found that students were much more willing to acknowledge the role of chance variation in the following question: *"A fair coin is tossed 50 times resulting in 27 heads. Two days later it is tossed again 50 times resulting in 30 heads. What do you think of the results?"* Similarly to the Pfannkuch and Brown (1996) study, everyone found the results to

be non-surprising. Anna, for example, said this time that neither of the two outcomes sounded suspicious and that *"even 20 heads are fine because you know, you've got to give yourself some variation in there."* That students' understanding of probability is more limited in real-world contexts than in the contrived context of standard probability tasks, was also observed in their responses to three other questions posed during the follow-up interview to investigate intuitive understanding of the effect of sample size on variation. One of these questions (from Pfannkuch and Brown, 1996) was the following:

Every year in New Zealand approximately seven children are born with a limb missing. Last year the children born with this abnormality were located in New Zealand as shown on the map. What do you think, given that one-third of the population lives in the top region and one-sixth of the population in each of the other regions?

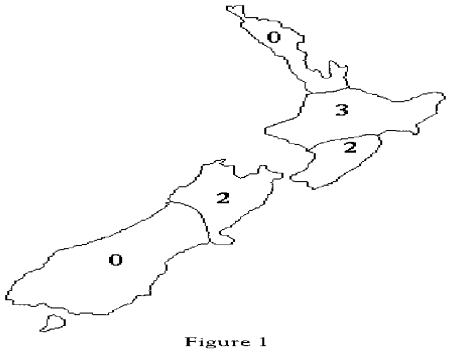


Figure 1

Figure 4.3 – New Zealand Task

Pfannkuch and Brown (1996) found that students' understanding of variation in small samples was very poor in this context. Whereas an analysis combining both probabilistic and deterministic thinking would have been more appropriate, all of the students interviewed gave deterministic explanations, and it

was only after repeated probing that some suggested the need for more data. My findings are very similar. I observed very strong deterministic reasoning in all of the students. George, for example, *"wouldn't want to live in the middle of New Zealand",* and Julie was convinced that there must be an outside factor causing the difference: *"There is always a chance that anything can happen but, 3 and 0 in the other…there must be a reason for that."*

Pfannkuch and Brown (1996) conjectured that students' neglect of probabilistic thinking might be the rich experience they have with similar controversial data's often appearing in the media and seldom being explored from a statistical perspective. When asked what they think of the possibility of obtaining the outcome {3,3,3,4,4,5,5} (order unimportant) when rolling a fair die 7 times, no student found such an outcome surprising. They approached this problem very differently from the New Zealand one although it is analogous – obtaining 1 or 2 on the die corresponds to the top region of the map where one-third of the population lives, and obtaining a 3,4,5, or 6 corresponds to each of the other regions. Similarly, in the follow-up interviews I conducted, students found such a result pretty likely due to the small sample size that allows extreme outcomes: *"I think nothing of the results. After a thousand throws each number will be picked around 1/6 of the total throws."*

The different way in which students approach the two problems indicates how much more prone we are to look from a stochastical perspective at standard probability tasks than problems situated in real-life contexts (Pfannkuch and Brown, 1996). Students are not completely wrong since a lot of other factors

besides chance might influence the occurrence of birth defects, but they should still realize that 7 children is too small a sample for drawing conclusions. They should have shown the same sensitivity to the effects of sample size they showed in the "Child Psychologist Question" (from Garfield and delMas, 1994). In that question, students were asked to judge the validity of conclusions drawn by a child psychologist who, after studying 5 infants and finding that 4 showed preference for the one toy, concluded that most infants would show a preference for this toy. Every single student interviewed challenged the psychologist's conclusions. Tim for example said: *"4 out of 5, I know it's good for like 4 out of 5 dentists prefer this kind of toothpaste, whatever on the commercials, but I would say you need at least a 100 kids...I could get my 5 sons and persuade 4 of them."* His response comes in sharp contrast to how he responded to the "Map of New Zealand" question:

> Int.: Just by looking at the map, do you see any connection between where one lives and how many kids are born with a missing limb?
>
> Tim: Oh, yeah. They correlate because the 1/3 that lives there has 0 because probably there are more doctors and more hospitals and only 1/6 lives there, so there must be something going wrong there. So yes, there has to be a reason.
>
> Int.: Do you see that the numbers are small? Do you think this is something you should take into account?
>
> Tim: Why?

CONCLUSIONS

The general conclusion that can be drawn by looking at students' responses to the pre-assessment, is that they had some sense of variation and of the statistical tools used to deal with it, which however was murky:

- They recognized that low variability could be a good or bad quality depending on the context of the situation.
- They made remarks indicating some knowledge of quality control processes.
- They could notice the difference in the amount of variability between two sets of scores having the same mean.
- They had poor understanding of histograms and bar graphs:
  - (i) Many did not seem aware of the data reduction involved;
  - (ii) Most were unable to relate features of the distribution with the shape of the graph;
  - (iii) A sizable proportion gave the wrong response to a question asking them to look at the histogram of two distributions and figure out which of the two distributions has more variability.

All students understood that different samples from the same population can (and usually do) vary. They realized that the sample mean is not the same thing as the population mean and nobody expected the two to be exactly equal. All students also intuitively understood that larger random samples tend to produce better estimates and stressed that the larger the sample size, the more representative it will be of the population.

Although students did recognize the existence of variation among samples, they tended to underestimate its effect. This tendency, which was more prevalent in real world-contexts, indicates limited understanding of randomness and makes it difficult for students to differentiate between chance variation in the data and variation due to some form of underlying causality. Despite the fact that students seemed to intuitively understand the dangers of drawing conclusions from small samples, when asked to make judgments about real-life situations, they ignored these dangers and did not hesitate to use small samples as a basis for inference and generalizations. For problems posed in real-world contexts, students seemed to expect small samples to resemble the population from which they are sampled, erring thus towards the deterministic side. However, when real-world context was removed, students were comfortable thinking probabilistically. They responded correctly to typical coin toss problems and appeared to be, in this context, comfortable with the notion of long-run relative frequency.

Question 6 was an adapted version of one of the tasks Jacobs (1997) gave to 110 fifth graders. As it has already been pointed out in the literature review, Jacobs found that although children did not like restricted sampling methods, they evaluated positively self-selected methods. In addition, whereas children liked stratified random sampling because it allowed them to specify the mixture of the sample, they mistrusted the "unknown nature" of simple random samples. When asked to indicate their preferred sampling method, 3.6% preferred restricted sampling, 39.1% self-selected sampling, 37.3% stratified random sampling, and only 5.5% simple random sampling. Finally, even when able to identify potential

for bias for individual surveys, children often ignored survey quality when drawing conclusions from multiple surveys. Comparing Jacob's findings with those of this study, we see that older students are much more likely to recognize the potential for bias in self-selection, but that still there were several students either ignoring or not identifying this potential. The mistrust of simple random sampling's ability to produce a representative sample was to some degree also observed in several students. In general, however, students evaluated simple random sampling much more positively than children in Jacobs' study, since they chose Survey 2 more frequently than any other survey as their preferred method. However, we noticed similarly to Jacobs' study that several students ignored survey quality when drawing conclusions from multiple surveys, although they had made correct judgments about the relative quality of data drawing from which of those surveys.

IMPLICATIONS FOR INSTRUCTION: FURTHER ELABORATION OF THE CONJECTURE

In Chapter III, I gave a description of the "Variation as the central tenet of statistics instruction" conjecture, which was based on the literature review and on previously conducted personal research. Here, I briefly describe how insights gained from the pre-assessment and the follow-up interviews, led to further elaboration of the conjecture and consequently the instructional program. Because we viewed "learning [as] a process in which students reorganize their thinking to resolve situations that are problematic for them" (Jones, Thornton, and

Langrall, 1997, p. 43), we utilized the information gained to ensure instruction was adopted to the students' existing experience and their pre-knowledge.

Statistical Thinking is Contextual

Students' assessment prior to instruction as well as the research literature, indicate that students' thinking about the stochastic is "linked to a complex web of personal, social and contextual factors" (Gordon, 1997, p. 146). It is not only probabilistic reasoning that drives students' thinking about the stochastic, but also impressions, prior beliefs, and expectations. The pre-assessment has revealed the "manifold nature of probabilistic thinking." (Jones et al., 1997, p. 42)

In the increasingly many sectors of society relying on data, the purpose of statistical investigation is to help inform decisions and actions by expanding the existing body of context knowledge about the situation under study. Therefore, "the ultimate goal of statistical investigation is learning in the context sphere" (Wild and Pfannkuch, 1999, p. 225). This means much more than collecting new information, it also involves synthesizing new information and new ideas with existing ones in order to gain improved understanding that can then inform decisions and actions (Wild and Pfannkuch, 1999). Thus, we need a re-evaluation of the position that teaching strategies applicable to any particular area of application results through "immersion in the subject matter area, through careful study of statistical applications in that area" (Breslow, 1999, p. 253), since statistical thinking ought to take place within a context. Neither is the usual panacea for "teaching" statistical thinking to students by "let[ting] them do projects" (Wild and Pfannkuch, 1999, p. 224) adequate, although it does give

students the opportunity to experience more of the breadth of statistical investigations. As Wild and Pfannkuch (1999) stress, "the cornerstone of teaching in any area is the development of a theoretical structure with which to make sense of experience, to learn from it, and to transfer insights to others" (Wild and Pfannkuch, 1999, p. 224). In the next section, I lay out the theoretical structure that the instructional program described in this study aimed at helping students develop. At the same time, I give a general overview of the path that, at the beginning of the course, we conjectured instruction should follow in order to optimize the possibility of development of this theoretical structure.

Variation as the Central Tenet of Statistical Thinking

Statistical thinking is concerned with learning and decision-making under uncertainty. Variation is a critical source of uncertainty. It is the fact that all processes vary which creates the need for statistics. It is the need to deal with variation through measurements that provides a (numerical) basis for comparison that produces data (Snee, 1999). We use statistical tools to analyze this data and observe the pattern that exists despite (or because of) the variation. Thus, according to Snee (1999), the elements of statistical methods are variation, data, and statistical tools. Understanding of variation and using this understanding to improve the performance of processes is the core competency and it should be the focus of statistical education, research, and practice (Snee, 1999). Understanding what data is relevant and how to construct proper methods of data collection and analysis enhances successful application of this core competency (Snee, 1999).

Defining Statistics Instruction in Terms of Variation

The central thus element of statistical thinking is variation, and instruction should aim to provide students with the skills necessary to be able to notice and acknowledge it, to explain and deal with it. But, if variation is indeed to be "the standard about which the statistical troops are to rally" (Wild and Pfannkuch, 1999, p. 235), we have to arrive at a common conceptualization of statistics instruction in terms of variation. Wild and Pfannkuch (1999) offer the following three "variation" messages as a starting point: (1) variation is omnipresent; (2) variation can have serious practical consequences; and (3) statistics give us a means of understanding in "a variation-beset world". The subsequent messages of the statistics classroom provide information about tools and methods statistics offers to help us make sense of the omnipresent variation.

1. Omnipresence: Variation is an omnipresent reality that affects all aspects of life and everything around us. In addition to variation inherent to almost any process, whenever we collect data we supplement process variation with variation produced by the data collection and measurement systems (see Figure 4.6).

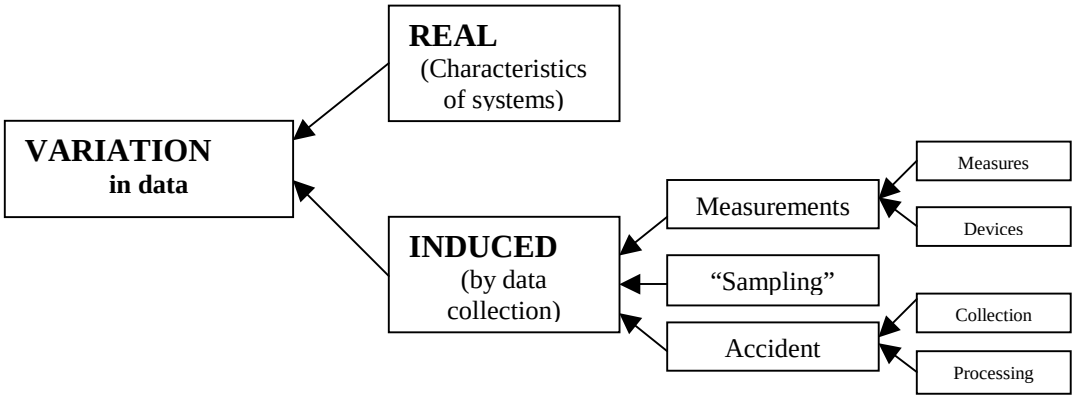


Figure 4.6 – Sources of Variation (from Wild and Pfannkuch, 1999)

2. Practical impact: Once we have established the omnipresence of variation, we have to demonstrate its practical impact on peoples' lives. Students have to understand that "it is variation that makes the results of actions unpredictable, that makes questions of cause and effect difficult to resolve, that makes it hard to uncover mechanisms" (Wild and Pfannkuch, 1999, p. 235). There are three rational responses to variation in a system and Figure 4.4, taken from Wild and Pfannkuch (1999) depicts them.

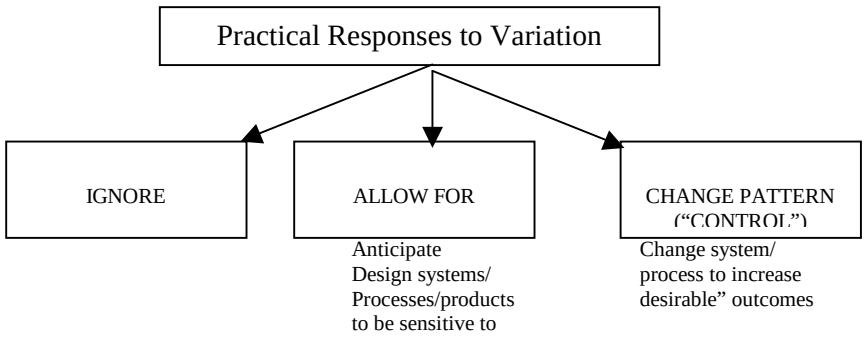


Figure 4.4 – Practical Responses to Variation (from Wild and Pfannkuch, 1999)

In some cases, we can *ignore* variation, pretending that it does not exist (e.g. pretend that every object is the same or differs in a deterministically known way). This, in some circumstances, works wonders. For example, applied mathematics and all its fields of application have made dramatic advances by modeling deterministic variation (Smith, 1999). In other cases, we might decide to investigate the existing pattern of variation and come up with ways to deal with it, to *allow for* it. For example, at the design stage of quality management

approaches to manufacturing they do this when they try to design a product "'robust' to the uses and conditions it will be subjected to (Wild and Pfannkuch, 1999). Finally, we might try to *control* the system and change to something more desirable its pattern of variation by identifying manipulable causes of the variation and controlling them through external treatments. For example manufacturing quality control processes, and much of the medical research have this purpose (Wild and Pfannkuch, 1999).

3. Use of statistical tools to model and understand variation: Whenever we cannot ignore variation, but we have to allow for it, or even control it, statistics comes to our rescue. It provides us with tools to measure and model variation for the purposes of *Prediction*, *Explanation*, or *Control* (Wild and Pfannkuch, 1999). *Control* is when the pattern of variation is changed to something more desirable. *Prediction* is what provides the main source of information to allow for variation. *Explanation*, i.e. gaining some understanding of why different units respond differently, improves the ability to make predictions and at the same time is necessary for control.

Probability

Probability is the relative limiting frequency of an event. However, the theoretical statement that *P(Head on next toss)=1/2*, "seems to be in sharp contrast to the intuitively felt inability to make specific predictions on this outcome" (Borovcnik, 1990, p. 7). Students coming to the statistics class have already experienced the highly fluctuating and irregular pattern of "Heads" and "Tails" in sequential coin tosses. They might have already unsuccessfully tried

some idiosyncratic strategies and this might have led them to conclude that there is lack of any substantial knowledge that could help improve one's situation (Borovcnik, 1990). Thus, they might not be able to clarify at the intuitive level how the statement *P(Head on <u>next</u> toss)=1/2* can be the answer to the "real" need of making a prediction for the next outcome (Borovcnik, 1990). At the same time, their urge to order this chaos, to overcome the uncertainty, might activate "different mathematical "theories", causal links, logical patterns, or even astrological links (or a combination of several of these)" (Borovcnik, 1990, p. 8).

The assumption underlying stochastic experiments of being repeatable under the same conditions causes a lot of confusion on students, who might try to find causal explanations: "Now, what differentiates a stochastic experiment from a physics experiment? Why do probabilists not study the physics of coin tossing and end up with such statements such as 'If you toss a coin at that angle, with that speed, with…, then it will turn up "heads'?" (Borovcnik, 1990, p. 8). Conversely, students might attempt to search for a logical pattern. Such an approach "is highly interwoven with magic belief and astrology (the law of series, a change is overdue etc.), and the search for the signs to detect this early enough" (Borovcnik, 1990, p. 8). Searching for patterns is a very intuitive strategy, similar to that employed by children when learning a language. It is often encouraged in mathematics teaching. We saw manifestations of both logical and causal reasoning in several questions of the pre-assessment.

What we observed in the pre-assessment was students' tendency to over-rely on sample representativeness. We had students reasoning in terms of

patterns, but failing to conceptualize the chance involved in those patterns and hence exaggerating the information given. In the "College Interviewer" question, for example, we saw many students viewing the distributions of males and females in the small sample drawn as an accurate reflection of the college student population. The missing aspect of uncertainty took the form of the student's failure to recognize how chance enters into the sample. In the "Roulette Wheel" question, we witnessed the *Gambler's Fallacy*, which denotes the expectation of local correction to random fluctuation in a sequence. In questions such as the "Map of New Zealand" one, many students assumed fluctuations in the data must be causal and proceeded to develop causal explanations.

Statistics instructors try, either through mathematical derivations or through simulations, to help students understand the law of large numbers. They want students to understand that, in the long-run, relative frequencies vary around the value of ½ and that, since the next trial is representative of the system under study, one could utilize the knowledge provided by the law of large numbers to predict the next outcome. Nonetheless, "this is not an easy secondary intuition: it is a whole bulk of images which is put together…shortcuts to it, or less carefully prepared ways to these intuitions, might lead to wrong associations" (Borovcnik, 1990, p. 8). Unless instruction establishes direct links between the intuitive and theoretical level, students' understanding of probabilistic concepts will be impoverished. Pfannkuch and Brown (1996) also argue that an effect of the clash between students' intuitions and probabilistic reasoning might be that students learn to distrust their intuitions, but because they do not actually understand why

they are wrong, return back to them. We might, for example, have students make claims such as that a "Tail" should follow the series HHHHH "in order to get the relative frequencies nearer to the value of ½ of the probability statement, as 'the law of large numbers suggests'." (Borovcnik, 1990, p. 8).

The literature review as well as the pre-assessment findings suggest a clear tendency of erring toward the side of attributing too much to deterministic causality, and failure to appreciate the extent to which chance operates in what one experiences in the world. It is then crucial for students to understand the idea that "chance variation, rather than deterministic causation, explains many aspects of the world" (Moore, 1990, p. 99). This is a fundamental idea for students' effective handling of data-based curricula, as well as their adequate interpretation and prediction of patterns outside of school. Instruction should promote the development of secondary intuitions that clarify how stochastic thinking is related to logical thinking and causal thinking, both of which often seem to be intuitively more convincing for students than stochastic thinking (Borovcnik, 1990).

Variation, Causation, and Probability

Viewing as a negative quality students' intuitive tendency to come up with causal explanations for any situation they have contextual knowledge about, is an attitude that will not take us far in our efforts to help students improve their intuitions of the stochastic. Since most real-world problems are embedded in a desire to improve something by identifying and controlling causes, we should rather view this impulse of students to find causes for phenomena as a positive resource (Wild and Pfannkuch, 1999). Probabilistic thinking should not be seen

as an alternative to deterministic thinking, but as "something to be grafted on top of the natural thinking modes that directly address the primary goal" (Wild and Pfannkuch, 1999, p. 238).

Statistics instruction should by no means underestimate the fact that conducting any study to uncover possible causal factors always "proceeds from ideas about profitable places to look, ideas which draw almost exclusively on context-matter knowledge and intuition" (Wild and Pfannkuch, 1999, p. 238). The important point that instruction should make is that, "while on the one hand variation may obscure, it is the uncontrolled variation in a system that typically enables us to uncover causes" (Wild and Pfannkuch, 1999, p. 236), by looking for patterns in the variation. Students should understand that "the randomized experiment is the most convincing way of establishing that a mooted relationship is causal" (Wild and Pfannkuch, 1999, p. 238). Statistics helps us in our search for relationships and causes by allowing us to translate ideas into variables to measure, and providing us with methods for appropriate data collection and analysis:

> Solving most practical problems involves finding and calibrating change agents. Statistics education should really be telling students something every scientist knows: "The quest for causes is the most important game in town". It should be saying, "Here is how statistics helps you in that quest. Here are some general strategies and some pitfalls to beware of along the way…" It should not just be preventing people from jumping to false conclusions but also be guiding them towards valid, useable conclusions. (Wild and Pfannkuch, 1999, p. 238)

Instruction that ignores students' prior context-knowledge and intuitions will not take us far from formal knowledge (even that quite shaky as the research

literature suggests) of statistical methods and procedures with no connection whatsoever to reality. One does not measure and model variation in a vacuum but for a purpose which influences how this is done, and thus statistical tools should be taught in context (Wild and Pfannkuch, 1999). Statistical thinking is not a separable entity but a synthesis of statistical knowledge, context knowledge, and the information in the data in order to produce implications, insights and conjectures (Wild and Pfannkuch, 1999). It always necessitates a complementarity of theory and experience:

> One cannot indulge in statistical thinking without some context knowledge. The arid, context-free landscape on which so many examples used in statistics teaching are built ensures that large numbers of students never ever see, let alone engage in, statistical thinking. One has to bring to bear all relevant knowledge, regardless of source, on the task in hand, and then to make connections between existing context-knowledge and the results of analyses to arrive at meaning. (Wild and Pfannkuch, 1999, p. 228)

From Association to Causation

In order to provide students with the tools necessary to make valid judgments when moving from association to causation, statistics instruction needs to supply them with more than the advice "correlation does not imply causation". This advice is nothing but a "'Hey, not so fast' warning" (Wild and Pfannkuch, 1999, p. 240). The search for causes does not need to be explicitly taught since it comes naturally to people. On the other hand, the tendency to "challenge the causal assumption, whether our own or somebody else's", to "rack our brains for other possible explanations and for strategies for testing these explanations" (Wild and Pfannkuch, 1999, p. 240), are dispositions which come naturally to only very

few people. Thus, instruction should put much more effort towards helping students realize that the best way to search for real, non-ephemeral causes, is through an in-depth study and not just by looking among recent changes for a cause (Wild and Pfannkuch, 1999).

Statistics instruction should be aiming at developing in students "a healthy skepticism and the imagination needed for alternative explanations" (Breslow, 1999, p. 253). These dispositions can be taught by gaining experience and seeing ways in which certain types of information turn out to be false and unsoundly based. Also, learning possible threats to the reliability of a study, is something that can be (and is usually) taught. Instruction should encourage students to adopt a critical attitude whenever receiving new ideas and information. This critical attitude means being constantly on the outlook for logical and factual flaws, and it includes learning to counteract the human tendency to be less judgmental of results that agree with their predispositions, expectations, and worldviews. A balance of stochastic and deterministic reasoning, also needs to be supplemented by logical reasoning in order to arrive at valid conclusions. Skepticism needs to be accompanied by the ability to reason from assumptions to implications, which should then be checked against the data (Wild and Pfannkuch, 1999).

Students know from everyday experience that even when studies are conducted under very similar conditions, they will give results which are different in detail, that the patterns observed in one study will never appear identically the same in another study. What instruction should stress is that statistical strategies, based on probabilistic modeling are ways we can use to solve the problem of

distinguishing between genuine patterns and ephemeral patterns that are part of our imagination. Students should come to view probability and statistical inference as ways we can use to counteract our natural tendency to view patterns even when none exists:

> As statistician (and zoologist) Brian McArdle put it so vividly in a personal interview, *"The human being is hard-wired to see a pattern even if it isn't there. It's a survivor trait. It lets us see the tiger in the reeds. And downside of that is that our children see tigers on the wall."* It is not entirely true that no patterns appear in purely random phenomena. These patterns are real to the brain and in the sense that we can recognize features that would help us reproduce them. However, such patterns are (i) ephemeral, and (ii) tell us nothing useful about the problem under study. In other words, they are meaningless. Part of our reasoning from random models is to say that we will not classify any data-behavior as "enduring" if it closely resembles something that would happen reasonably frequently under a purely random model. (Wild and Pfannkuch, 1999, p. 240)

Statistical methods were developed by people in order to help filter out any "signals" in data from surrounding "noise"." The "signal" is the messages, the meanings we find in *explained variation*, the patterns that we have not discounted as being transient. The *unexplained variation* is the variation, the "noise" that remains after we have "removed" all patterns. Although there might be multiple causes for unexplained variation, we are not able to detect them since we do not see any structure, thus we use probability to model it, assuming it had been randomly generated. Even if random sampling was used, although there will be an element of randomness in the noise, unexplained variation will also typically include measurement error and components of the variation in the original process which we do not know whether they behave randomly or not.

Thus, randomness is a "human construct" developed to deal with variation for which patterns cannot be detected and make accuracy predictions. It is "all part of an attempt to deal with complexity that is otherwise overwhelming." (Wild and Pfannkuch, 1999, p. 241)

Assuming that the data was randomly generated according to the model, we can then use probability as the link between the data and the population/process it originated from. Every model is essentially an oversimplification of reality and it involves loss of information, but the hope is that we have caught the essential dynamics of the problem. How successful probability models are, depends on their practically, and their potential to give useful answers to our questions. In order to be able to use probability models, students need to be able to recognize situations in which it would be appropriate and useful to use them. They also need to now how to build and fit an appropriate model, and draw conclusions from it. Deducing implications from model involves some understanding of the behavior of random models. (Wild and Pfannkuch, 1999)

The Behavior of Random Phenomena

Since probability theory was developed as a means to model and describe phenomena for which no patterns can be discerned, "what probability is can only be explained by randomness, and what randomness is can only be modeled by means of probability" (Steinbring, 1990, p. 4). Stochastical knowledge is created as "a *relational form* or *linkage mechanism* between formal, calculatory aspects on the one hand, and interpretative contexts on the other"(Steinbring, 1990, p.5).

However, the classroom culture often comes in sharp contrast with this conception of stochastical knowledge as being developed through a "self-organized" process that balances the objective aspects of a situation and the formal means employed to model and describe it. The linear, completely elaborated and hierarchical structure of knowledge presentation characterizing many statistics curricula and instructional approaches, encourages the development of the chance concept as a concrete, totally clear and unambiguous generalization defined by methodological conventions:

> There is an effort to give a clear-cut definition of the "basic concept of chance" as early as possible. Thus, for instance, textbooks define: "If results cannot be predicted with certainty, but will happen by chance, we speak of chance experiments." In the curriculum, the concepts of "probability" and of "chance" are not organized under a dynamical perspective, but under a static one as ready-made elements. (Steinbring, 1990, p. 8)

Steinbring (1990) analyzed teaching episodes from several different classrooms in order to see how the concept of chance was introduced. The basic pattern he observed in all of those episodes was that "chance" was first introduced through performing and discussing a chance experiment. An attempt was then made to describe the experimental outcomes using a rule or a simple stochastical model. Of course, there was always variation observed between the theoretical predictions and the empirical data. The pattern of justification for the variation, regardless of its size, always was that the observed difference between the empirical result and the theoretical prediction was produced by "chance" (Steinbring, 1990). The difference between theory and experiment was thus neutralized: "Chance degenerate[d] into a substitute for justification, which

serve[d] to deny the importance of the difference between theory and empirical facts in probability." (Steinbring, 1990, p. 14)

As Biehler (1994) points out, shifting from individual cases to systems of events, is a fundamentally new idea for many students. Moving from an individual event to a system of events, "can reveal new types of knowledge, new causes, explanations and types of factors that can not be detected at the individual level also in many cases where a causal analysis of individual events would be informative" (Biehler, 1994, p. 13). However this "picture of a deterministic dependence of long run distributions from conditions in contrast to the problematic individual level" (Biehler, 1994) has a basic limitation when applied to everyday statistical analysis. When drawing inferences from samples, we have knowledge about an intermediate level where, due to the variation present in any finite sample size, conditions do not determine the sample completely (Biehler, 1994).

Instead of emphasizing individual irregularity, it is more constructive to promote in students a way of thinking that perceives a (probability) distribution is based on some conditions, which when changed might lead to changes in the distribution (Biehler, 1994). Such an approach, which Biehler (1994) calls "statistical determinism", is especially useful for dealing with more realistic situations: "Playing roulette with its well-defined chance structure is much different from individual risk assessment, where no unique reference set (system) exists" (Biehler, 1994, p. 13). Students should understand that we can analyze causes of why an individual event took place, but at the same time realize that we

can get something from the transition to a system of events that we cannot get if we just focus on the individual level. This complementarity of individual and system level, which is usually suppressed by instructors, who call attention only to the systems level, is more intuitively convincing and more productive:

> Let us take traffic accidents as another example: We can provide some causal explanation at the level of individual accidents in every instance. We can aggregate data for a longer period, however, and analyze how the number of accidents changed over the years, how different it is on weekdays and weekends, whether there is some seasonal variation. The aggregation makes changes in boundary conditions detectable, which may not be detectable at the individual level. Aggregating individual data or dissecting aggregated data are basic concepts that gain their importance from the above perspective. (Biehler, 1994, p. 14)

As Steinbring (1990) maintains, instruction can indeed begin with the preliminary notion of chance as that which is "*opposite of causal laws*", since this description is a direct continuation of intuitive ideas of chance. This preliminary interpretation does depend on the relation between theory and experiment, and is in accord with the definition that real chance events do not occur with absolute certainty but only with a certain probability (Steinbring, 1990). However, how instruction moves from here needs to be carefully thought out. These intuitively convincing ideas open a direct connection between randomness and probability, but at the same time open relations to everyday notions of chance such as that of having good or bad luck. The students in the pre-assessment who argued that with random sampling one could not make predictions about the likely outcome of a study since everything is possible, seemed to hold this view.

Instruction often fails to help students move beyond the intuitively based idea of chance towards the right direction. Doing stochastic experiments and evaluating experimental outcomes provides a socially constituted teaching context that opens up two main possibilities for further development of the chance concept. The first possibility is what is typically observed in classrooms: "a narrowing reduction of the chance concept to a formalized, conventional label" (Steinbring, 1990, p. 17). Often stark contradictions between theoretical prediction and empirical observation are justified as being the result of "chance" in the naïve sense. The second possibility is for instruction to broaden the chance concept "to become a means of analyzing the relation between experimental situation and stochastical model" (Steinbring, 1990, p. 17). Rather than maintaining the prevailing concept of chance as that of irregularity, instruction could help develop it further by bringing to students' attention the fact that the occurrence of a very rare and improbable event might indicate that there is something wrong with either the experimental conditions or the model.

It is this second possibility that the statistics course should exploit in order to help students develop the *theoretical* nature of this concept in an appropriate way. Development of the chance concept can be first examined through the notion of statistical independence which, according to Steinbring (1990) is "a theoretical generalization of the intuitive chance concept and it introduces a first differentiation between object [experimental situation] and sign [stochastical model] in elementary stochastics" (p. 17). However, unlike conventional instruction which reduces the chance concept to "a universal object for explaining

the connections between outcomes of an experiment and the theoretical prediction", it should take advantage of the contributions offered by students "to unfold and differentiate the development of this concept" (Borovcnik, 1999, p. 4). Probability should not be presented as a body of theory free of any concrete interpretations. Particular attention should be paid to the relation between stochastical theory and empirical outcomes. The role of chance should be "lifted out of naïve magical thinking to become a theoretical concept in scientific stochastical thinking", it should change to that of a device for controlling the underlying connection between the stochastical model and the experimental situation (Steinbring, 1990, p. 18). Contradictions between theoretical arguments and empirical results should not overlooked:

> An empirical outcome is not only a specific, concrete result of a quasi deductive experimental process, but it can be seen as a generalized outcome in the range of many possible outcomes and in this way may give rise to an inversion of the question of justification: Is it necessary to modify some basic assumptions or experimental conditions of the whole process? In principle, all elements of the whole process have to be questioned when a very rare event is observed. (Steinbring, 1990, p. 18)

In order to move stochastical knowledge beyond mere methodological conventions "completely pre-constructed by the teacher's methodical intentions" (Steinbring, 1990, p. 21), classroom processes and interactions should adopt "a proper knowledge-epistemology" that takes the metaphor of "self-reference" seriously. Experiments performed in the classroom and computer simulations should be perceived as fundamental sources for the students and not simply as motivations for step-by-step teaching of the teacher's intended goals (Steinbring, 1990). The self-referent epistemological structure of stochastical concepts should

also be reflected in the social process of the classroom. Stochastical knowledge necessitates direct subjective decisions and interpretations, and it is only through increased involvement by the learner that learning will become powerful:

> The learning subject has to decide how to take the statement: "There is something wrong in the relation between theoretical model and empirical observation!" It is the self-referent character which makes knowledge alive and offers the learning subject participation in this developmental process. Such an understanding of *theoretical knowledge* will permit the re-establishment of an appropriate balance between objective and subjective aspects of knowledge in processes of teaching, learning and understanding. (Steinbring, 1990, p. 21)

Such an approach, of lifting chance from "a naïve intuitive concept, which only is defined negatively as non-existing regularity" (Steinbring, 1990, p. 18), to a theoretical concept that calls for careful analysis of experimental conditions and theoretical assumptions, lays solid foundations for the development of the most important stochastical concepts:

> Future advanced stochastical techniques and concepts can be used in a way of self-application or of feedback to re-analyze the experimental situation of the actual classroom teaching: According to the Bernoullian model, the outcome of the game played by the students…probability less than 0.1%=> plausible to assume discrepancy between assumptions underlying model and actual performing of experiment. (Steinbring, 1990, p. 18)

There is a need for intuitive representations to help students see "the fundamental relationship between chance and regularity, between irregular, unpatterned phenomena on the one hand, and the mathematical intentions to model and describe them in a regular and formal way on the other" (Steinbring, 1990, p. 3). Students have to come to view theory of probability as an attempt to

attain a certain degree of certainty in contexts where “it is no longer possible to advance certain predictions about future events on the basis of strictly causal linkages” (Steinbring, 1990, p. 2). We began the teaching experiment in the hope that the path we had decided to follow, which was based on both the research literature and the assessment of student knowledge prior to instruction, would help build connections between formal mathematical expressions of the stochastic and everyday informal intuitions. In the next chapter, I describe the teaching experiment and how it led to further modifications of the conjecture. I give a brief description of some teaching episodes and class activities which are characteristic examples of how the course was organized, as well as some examples of how the continuous monitoring, both formal and informal, of student thinking shaped instruction and re-defined the conjecture.

Chapter V: The Teaching Experiment

INTRODUCTION

The chapter begins with a brief description of the classroom culture, and how it supported students' grasp and utilization of big ideas related to variation. I then describe some teaching episodes and class activities characteristic of how the course was organized and how the meaning of main statistical concepts was constituted in social interaction. I also try to give some examples of how the continuous monitoring, both formal and informal, of student thinking shaped instruction. I also outline and discuss the findings from the assessment given at the end of the course and the follow-up interviews of the primary group.

CLASSROOM SETTING

The classroom setting was such that it encouraged "statistical enculturation". The instructor's knowledge and behavior contributed towards the creation of an authentic model of the "statistical culture" (Biehler, 1999). It was a setting that modeled realistic statistical investigations, and in which statistical dispositions such as appreciation of data were valued and nurtured. Instead of following the now common approach of progressing from data analysis, to data production, to probability and then to inference, students experienced statistical investigations as a dynamic process. The instructor never taught any method or procedure in isolation. In contrast to more typical approaches, where reference to problems is made to demonstrate statistical content, reference to statistical content in this class was made (in students' mind at least) to help understand a situation,

to assist a statistical investigation. The emphasis was on statistical process and, along the way, students got to learn different statistical methods and procedures. The hope was that by putting students in situations where they needed tools such as the standard deviation, they would realize their usefulness and not wonder why anyone would ever bother to invent them (Erickson, 2000).

The instructor was trying to increase students' awareness of variation, to help them realize that it is the existence of variation which creates the need for statistical investigations. He would keep on emphasizing that the reason we use statistical tools is to describe trends and patterns and deviations from those patterns existing in the data because of the variation inherent in every process. The idea of making conjectures ran throughout the course. Students would state what they believed may or may not be true, and then looked critically at the data to evaluate their statements. While the instructor encouraged students to make conjectures he, at the same time, also tried to help them understand that conjecturing is not enough – one has to evaluate one's predictions by looking closely at the data and making comparisons (Erickson, 2000).

Evaluation of conjectures would typically begin informally by using one or more graphical displays from which the students would get a general idea as to whether their conjectures seemed reasonable. The instructor would encourage students to describe the main features of the distribution displayed by the graph(s), always emphasizing the need to take into account not only the center, but also the spread of the distribution. Students would look at the displays and try to give explanations for the patterns observed, which either confirmed or

challenged their original conjectures. Sometimes these explanations would be proposals for a possible model to summarize the dataset (e.g. a straight line, or a probability distribution). Other times the explanations would be as to why those patterns existed.

The evaluation of conjectures would then become more quantitative. An analysis using appropriate numerical summaries would be made in order to support or refute the conjectures suggested from looking at the displays. In the beginning of the course, the analysis was made using simple numerical summaries. Eventually, more tools were added to the students' repertoire. The mathematization of the data gradually became more and more formal.

Even when the data agreed with their initial conjecture, the instructor would encourage students to also come up with alternative explanations. He tried to help them see that there can be multiple explanations for a phenomenon, in the hope that this would make them "less likely to assume that their data 'proves' the obvious cause" (Erickson, 2000, p.2). He was also trying to raise their caution for conjectures that went beyond the information provided by the data.

A special emphasis of the course was on data production issues. Unlike many other statistics courses where study design issues are discussed as a separate topic and almost never appear again, they were continuously brought up in this course. Throughout the course, the instructor was stressing that data are numbers collected in a particular context that are studied for a purpose (Rossman, 1996), and the quality of the conclusions we draw depends on how the data were obtained. When, for example, students were examining graphs, the teacher would

point to them that patterns in the data depend to a great extent on how the data was obtained and that if data collection had not been properly done the observed patterns might be misleading. When they were discussing inferential methods, he would stress that those methods are based on the assumption of probability-based data production, and if this assumption does not hold, then the inferences drawn might not be sound. With regards to the inferential advantages of a larger sample size, he did repeatedly stress to students that if there is bias in the sample selection process and/or the measurement system then increasing the sample size would probably not lead to more valid conclusions.

The instructor would always situate instruction within contexts familiar to the learners. He would use analogies from students' everyday experience, and would try to simplify mathematical relations in order to help build links to students' intuitions. Borovcnik and Peard (1996) outline the potential benefits of such an approach:

> Starting with a context familiar to the learner and in which there are relations that are quite directly understandable, one has a possible basis to introduce students to the related mathematical concept, which can now easily be understood by referring to this analogous situation. Or, starting with mathematical concepts which are known to the learner, one can thereby structure a vague situation. Then one extends the mapping onto the connection between formal relations on the mathematical side and subject matter relations on the context side. (p. 269)

SAMPLE OF CLASS ACTIVITIES

Distance from Home Class Activity

The instructor began the first day of class by showing students a distorted picture of an elephant and a transparency with the following story:

A Story

Once upon a time, there was a King. One day, he brought several blind men to his kingdom. He asked them to touch an animal and then describe how they thought the animal would look like.

- Those who touched the nose claimed that the animal would look like a huge rope.
- Those who touched the leg said it would look like a huge pole.
- Those who touched the tail said it would look like a snake.
- Those who touched the stomach said it would look like a piece of wood.

He used this story as a way to begin a discussion on the purpose of data collection. He pointed its analogy to the fact that in statistics, when collecting data, we often aim to find information that would help us understand some unknown population. He stressed: "*The key point is that if I want to understand the population using a sample, I better make sure the sample will be representative of the population.*"

The first activity the class engaged in was the "*How far away are you from your home town?*" activity, where students calculated the class average distance from home. Through this activity, which lasted three days, many important statistical concepts and ideas were introduced. When, for example, students were debating what measurement they should use, the instructor stressed

the importance of choosing the right measurement system: *"We should use something useful and objective. Measurement system is very important."* He wrote on the board: *"Garbage In, Garbage Out. Do the right thing. Do the things right."*

After having collected everybody's distance from home, a discussion on how to present data began. At a student's suggestion, they first drew a *"what it's called…the one with dots"* (a dot plot). The instructor asked students to describe what the graph tells them about the distribution of distances from home. Students made some general observations indicating that they were already familiar with this type of plot, and then one student said that they could more easily describe the shape of the distribution if they looked at a *"bar chart"*. The instructor used this as an opportunity to explain the difference between histograms and bar graphs and to discuss how to construct them. They constructed a histogram of the "distances from home" dataset and, through the instructor's prompting, described the main features of the distribution displayed by the graph: its shape, center and spread. Next, the instructor asked students what else other than graphs one could use to describe the data and one student suggested finding *"things like the range and the mean."* The instructor agreed, pointing out that although graphs help us get a general idea about the shape of a distribution, we also need numbers to *"quantify the variation"*.

Through an extended discussion, different numerical summaries such as the mean, the median, the range, the standard deviation and the five-number summary were introduced and were used to describe the center and spread of the

dataset. In discussing these numerical summaries, the instructor's emphasis was on helping students understand their meaning and purpose and not on showing them how to calculate them. The statistical summary students had the biggest difficulty with was, of course, the standard deviation. The instructor told students they do not have to memorize the formula for standard deviation (they could bring a sheet with formulas during exams), but put a lot of effort into helping them understand why the way the formula is set up provides us with a measure of average distance from the mean.

After the class were done with the calculations, the instructor said:

> From the distance data, I find out that the average distance is about 115 miles and the standard deviation 75 miles. Of course, if I take a different sample, would I still get $\bar{x} = 115, s = 75$ miles? No. Different samples have different means and standard deviations. Sample information, for example $\bar{x}$ and s, varies based on different samples. But remember the elephant story? I do not only want to know about the part of the elephant that I study, but about the whole elephant. Inference means to use the sample to make decisions, to predict, to find a pattern in the population. This is what we will be dealing with in this class.

He then asked students whether the sample was representative of the distance from home of all the students in their university. Students argued that it was not and gave reasons such as: *"summer course, students might be closer to home", "sample size of 33 is too small to represent the 18,000", "sample is very subjective (a required class)."* Then the instructor asked them to discuss with their group what they needed to do in order to *"do a better job and make sure they obtain a representative sample of all students."* The group work was followed by a class discussion about the characteristics of the whole university

student body. Students noted that there are different groups of students *("campus-off campus", "freshman-sophomore-junior-senior"* etc.). This was the instructor's opportunity to introduce different sampling schemes. He wrote "REPRESENTATIVENESS" with big letters on the board and asked students to identify the important characteristics of samples that have this property. The different characteristics that were brought up and discussed included "*random selection", "by stratum",* and *"a large sample."*

Matching Statistics to Graphs Activity

Our past research findings were indicating poor understanding of the connection between numerical summaries and graphical representations of the dataset, despite the fact that there had always been plenty of activities in the PACE classroom giving students experience in exploring different features of distributions. For example, an activity the instructor assigned during the first week of the course, where students had to find on the Internet and analyze three datasets, one of which had to be skewed-to-the-left, one skewed-to-the-right, and one symmetric, had also been assigned in previous semesters. Nonetheless, problems persisted. At a departmental seminar of graduate students held at the beginning of the summer session, in which the instructor was leading a discussion on statistics education, he said that for him, understanding of histograms and their relation to variation is one of the stumbling stones in statistics instruction:

> Inst.: So, when I did the interviews, I found out that, you probably don't believe it, but it's a really very simple thing that students miss and continue to miss – histograms. If you think of histograms, they are a transformation from raw data into an entirely different form. And you

think because they see it everyday…Yeah, they do, but if you ask them to describe: "Tell me verbally…"

In order to make his point that understanding of histograms is not as trivial as some instructors might think, he gave as an example students' performance on the following task which he had included in the previous semester's final:

> When constructing a histogram for describing the distribution of salary for individuals who are 40 years or older, but are not yet retired:
> (i) Explain
> What is on the Y-axis:
> What is on the X-axis:
> What would be the proper shape of the salary distribution? Explain why.

What he had found was that only 3-4 students gave the correct response. Everybody else argued that the distribution would be right-skewed, not because they understood how the histogram would look like but because they confused it with scatterplots*: "They say it's going to be right-skewed because people who are near retirement, their salary gets less and less, and therefore the salary is smaller and smaller and therefore it's skewed to the right*." Because "*histograms are related to everything",* he has decided to put even more emphasis on histograms *"and have students do a lot more for homework",* although he was already doing *"a lot more than all the instructors here do in this area."*

Having a good understanding of spread when visually interpreting a distribution displayed in a histogram is necessary to be able to fully grasp the concept of sampling distribution. Students, especially those that in the pre-assessment could not figure which of two distributions had higher variation

(Question 7; Appendix A), had to improve their understanding of histograms. This was necessary since otherwise, when introduced to sampling distributions, students would not be able to recognize the reduction in variance resulting in moving from the population distribution to the distribution of a statistic. In order to assess the effectiveness of instruction, after students had some exposure to histograms, we gave the question again. The results were not very encouraging. There were still six people believing that distribution A had more variation than distribution B.

We tried with the instructor to think of ways to help improve students' ability to relate features of a distribution to the shape of a graph. We decided to include activities that require students to use information they know about the variable to decide how its distribution looks like without actually collecting or analyzing data, as well as activities that require them to look at a distribution and try to estimate its parameters. Such kinds of tasks, which are quite challenging, are typically not included in an introductory course – other than, of course, looking at some trivial cases such as drawing a curve to describe a normal distribution with a certain mean and standard deviation.

One activity we used was the *"Matching Statistics to Graphs"* activity taken from Scheaffer et al. (1996). The purpose of the activity, as the authors state, is to help students estimate the mean, median and standard deviation of different datasets by looking at their histograms, and also to see how boxplots are related to histograms. Students worked on the activity in groups. The group of three students I was observing and video-taping (Anna, Jim, and Tim) was able to

do all the right matches to the first part of the activity, which was asking them to find the variable corresponding to each of the following histograms:

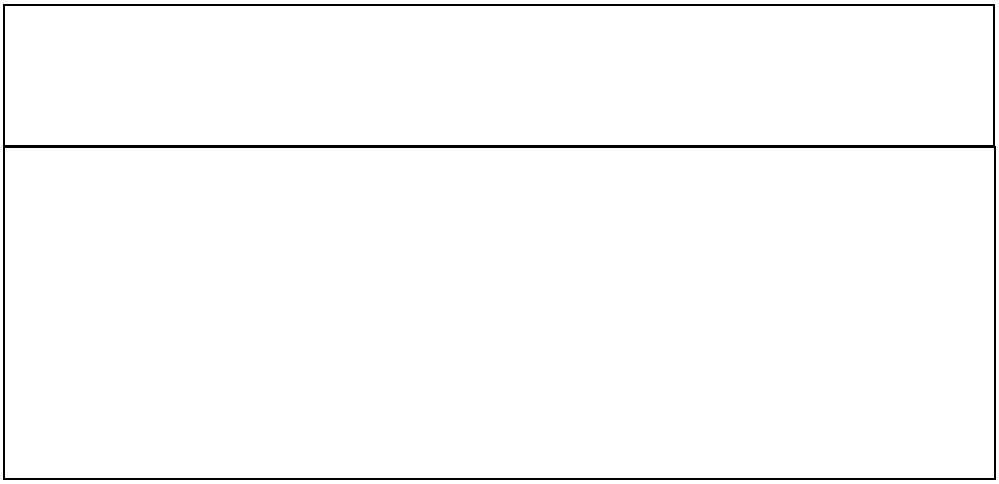


Figure 5.1 – Part A of "Matching Statistics to Graphs" Activity

They first matched Histogram B with Variable 4, because *"it has the most above 60 and the highest mean 53…and the standard deviation is the largest."* Noticing the symmetry and small variation of Histogram F, they matched it to Variable 6: *"This one has to be like a mean of 50 because it has like 38 and it's 12 here and then 12 on the other side."* For Histogram C, they decided that it had to correspond to Variable 5, which has the lowest mean because of the number of data points *"that are way low"*, the fact that "*almost everything is below 60"*, and also because *"if that* (the standard deviation of Histogram F) *was 5, the deviation would be 10 something."* Next, they conferred that Histogram D had to be Variable 2, since *"the mean is in the middle"* and it has the second largest variation due to the high frequency at the tails of the histogram. They matched

Variable 1 with Histogram A because the numbers are spread out *"even"* around 50. Finally, they matched Variable 3 with Histogram E since there are extremes at the high end which make its mean higher than that of Variable 1 – there is *"a whole bunch here that are close to 78."*

Similarly, in the second part of the activity, where they were given the following sets of boxplots and histograms, they correctly matched all the histograms to their corresponding boxplots.

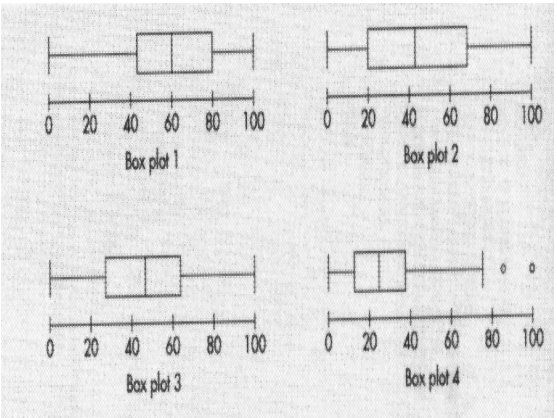


Figure 5.2 – Part B of "Matching Statistics to Graphs" Activity

They decided that Histogram A corresponds to Boxplot 2, Histogram B to Boxplot 3, Histogram C to Boxlpot 4, and Histogram D to Boxplot 1. For example, they decided that Histogram C *"which is low… where you have a couple of high one like this"* corresponds to Boxplot 4, whereas Histogram D corresponds to Boxplot 1 *"because all the high bars are here like that* [at the right

tail].” What is interesting is that the students were able to make the right matches while, as some comments they made indicate, they were under the impression that what the middle line of the boxplot shows is the mean. Originally, for example, Jim disagreed that Histogram D corresponds to Boxplot 1 because, he argued, Boxplot 1 has a mean of 60 but Histogram D does not. Anna however, argued that Histogram D has to be the one corresponding to Boxplot 1 because it has the highest mean. Jim was now convinced and the students moved to the wrap-up.

On the first part, they had to describe the features of a distribution that determine whether the mean and the median will be similar, and the features determining whether the mean exceeds the median. They decided that what makes the mean and the median similar is *“not having like high scores, not having extreme scores”* and that *“mean exceeds median when there is just a few extreme cases…a few high scores above the mean.”* In the second part, where they had to identify the features of the distribution that influence how large the standard deviation is, they decided that what determines the magnitude of standard deviation are the *“wide number of scores”* and also whether there are *“higher bars at the ends.”*

A whole class discussion followed this group activity. The instructor remarked that the purpose of the activity was *“Trying to put the puzzle together. Which puzzle? Mean, median, standard deviation, histogram, boxplots and how they are related to variability.”* He then asked students to describe the strategy their group used to go about doing the matching for the first task and a student responded that what they did was *“to figure out if it’s left or right-skewed and*

then if it's right-skewed the median is less than the mean… If it is symmetric the mean and the median are the same." The instructor agreed and said to the class that their first strategy should then be, looking at the shape of the graph and determining if it is symmetric or skewed. He drew a symmetric, a right-skewed, and a left-skewed curve on the board, and went through the six histograms with students, who decided what type each was. Then, a student noted that the next thing they should do is to compare the mean with the median because "*if it is right-skewed, the median is smaller"*, whereas "*if it is left-skewed, the mean is smaller than the median."* The instructor wrote on the board $m \approx \mu$ under the symmetric curve, $m > \mu$ under to the skewed-to-left and $m < \mu$ under the skewed-to-right distribution, and commented:

> Inst.: Once you understand this, then basically you solved this problem. So, the key is understanding this. Another point that you need to use is this. If you have two distributions that look like this (draws two normal curves with the same mean but with $s_1 < s_2$)…they have the same mean. Is s_1 smaller or is s_2 smaller? (A student says that s_1 is smaller because it is "closer together"). Those are closer together, that means they have a smaller variation… So, that's what this is about. This is what we are going to use. The picture is related to the information, OK? Once you understand these pictures here, basically you will be OK. (He also drew below each of the curves a boxplot corresponding to it).

Impact on Student Learning

A couple of days after this class activity, students took their first test. Overall, they did pretty well, although the test was quite challenging. However, the test results also indicated that student understanding of standard deviation and variation in general was still limited. The question students had the hardest time with was the following:

Cholesterol level Question

A health company is interested in the cholesterol levels for individuals with ages 40 or older in Mt. Pleasant. A random sample of 100 individuals was chosen from the target population and the following information was obtained: Sample Size=100, Sample Average = 158 mg, Sample Median = 159 mg, Sample s.d. = 20 mg. Based on this information, the shape of the cholesterol distribution is more likely to be approximately

(i) skewed-to-the-left

(ii) mound-shaped

(iii) skewed-to-the-right

Only 42% of the students answered this question correctly. Fifty-eight percent of them, not considering the large variation, decided that since *158<159* the distribution has to be skewed-to-the-left. As my following conversation with a student indicates, even among those who did pick the right response, not everyone considered standard deviation:

Keith: I said it's mound-shaped because 1 is a small difference.

Int.: 1 is a small difference here because the standard deviation is 20.

Keith: I don't understand standard deviation.

We met with the instructor to discuss the results of the exam and to decide our next course of action. Regarding students' poor performance on the *Cholesterol Level Question,* he noted: *"Part of this is my mistake because we first talked about skewed distributions before I stressed variation enough."* He remarked that next time he teaches statistics, he will make sure students become well aware of the role variation plays in a distribution before introducing the notion of skewness, so that students will realize they should not follow rules as

recipes without considering the variation involved. He added that understanding variation is very difficult for students: *"Most important aspect of math vs. statistics is that of variation. Unlike math, 21 is not always smaller than 20, and this is hard for students to understand."*

In order to help increase students' awareness of variation, we decided to use the activity described in the next section.

SATs and GPAs: Classroom Activity

In this class-activity, taken from Erickson (1999), students looked at sex differences in SAT scores and grade-point averages (GPAs) for 1000 first-year college students and tried to figure out how meaningful those differences were. The purpose of engaging students in this activity was to help them understand two "basic lessons of statistics" which are the foundation of sophisticated statistical methods: (1) one cannot compare measures of central tendency without taking variation into account; and (2) we can assess the salience of a between-groups difference by comparing it to the within-group variations (Erickson, 1999). Differences in means can be compared to standard deviation, while differences in medians can be compared to the interquartile range.

After passing out the handout found in Appendix B, the instructor first asked students to look at the following Fathom summary table, and describe what they noticed based on the table alone.

SATGPA		Summary Table	
⇩	⇨	**math**	**FYGPA**
sex	**F**	521.94215 484	2.5445868 484
	M	564.59302 516	2.3960659 516
Column Summary		543.95 1000	2.46795 1000
S1 = mean ( )			
S2 = count ( )			

Figure 5.3 – Summary Table of Mean Math and FYGPA scores for Males and Females

A student remarked: *"Because we have more guys than girls, this is not a fair comparison. Guys have a higher average because they are more."* The instructor had to give an example to help this student realize that a larger sample size does not necessarily imply a higher mean. Some other students argued that men do better on the math part of the SAT, while females have a higher FYGPA (first-year GPA). One student claimed: *"Males are better in math"*, and another one: *"Males are better at math, but females on a whole are smarter."*

Unlike the students who were ready to give causal explanations for the differences in the means, several other students argued that knowing mean scores is not adequate information for making comparisons between males with females. We need to look at the actual dataset, one student noted, because we need to find *"things like the standard deviation"*. The instructor agreed with the need for looking more closely at the data and taking spread into account. He pointed out that the explanations some students gave for the reasons behind the difference in means, although echoing different opinions and concerns of society, make claims

which go beyond the information provided simply by comparing mean scores. He added that what they would now do to check their conjectures was to retrieve the actual data and look more closely at it, analyzing it using Fathom.

Graphical Comparisons

After opening the file containing the dataset, students first drew a histogram of math SAT scores:

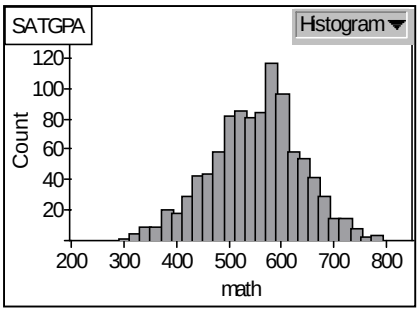


Figure 5.4 – Math SAT scores

The instructor asked:

Inst.: Do you consider the distribution to be normal or not?

Chris: Close to normal.

Inst.: In statistics we use a lot the words close, fair enough. Even though this bar here is very high, this is minor compared to the entire distribution shape. In real situations, that's how we should think. So, this is actually very much a nice picture for demonstrating the normal distribution.

He thus grasped the opportunity to remind students once again about the variability of real world data that makes perfect normal distributions idealizations.

Next, students drew a histogram of the math scores separated by sex:

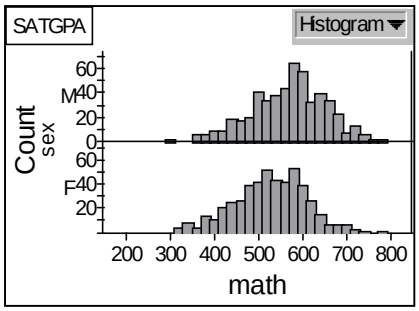


Figure 5.5 – Math SAT Scores separated by Sex

The instructor asked students to write a statement comparing males and females using the new information provided by the graphs. When I later looked at students' work, I saw that most of them wrote that the two graphs are very similar and "*although males are slightly higher than the females they are close.*"

Then, students drew histograms to compare FYGPAs:

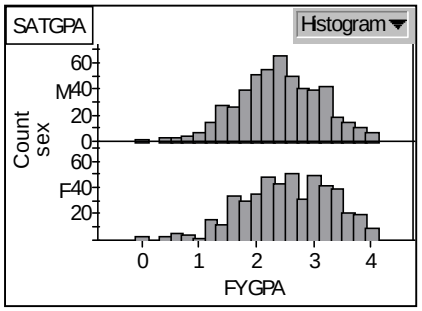


Figure 5.6 – FYGPA scores separated by Sex

They noted that this time it is females who seem to outperform males, with more females than males having a FYGPA of 3 or higher, but that again there is a lot of overlap. In order to get more insight, students compared results of males and females on the verbal part of the SAT and saw that they were similar. The

instructor then asked students to go back and change their graphs to boxplots. It is important, he noted, to look at data from different perspectives:

> Inst.: Distributions will never be perfect in the real world. When you make any comparisons, look at more than one graph and don't look at just two numbers. Look at more than just two numbers. The least you can do is to look at both the mean and the standard deviation.

This is an advice he would often give to students. He was trying to make them realize that different plots and numerical summaries each have different advantages and disadvantages and that using multiple data representations to look at data in different ways and numerical means to summarize it, often provides a much better understanding of the situation explored.

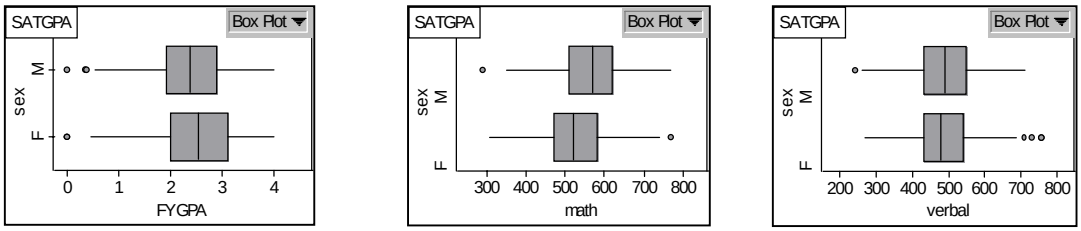


Figure 5.7 – Boxplots of FYGPAs, Math SAT Scores, and Verbal SAT Scores, for Males and Females

Comparing the boxplots of the three sets of scores, students saw that sex-related differences are more noticeable on math SATs and on FYGPAs than on verbal SATs. They then went to the next question asking them to describe how they knew, by looking at the graphs, that the difference in math scores is "more meaningful (not just bigger)" than the difference in verbal scores. One student noted that, looking at the boxplot of math SATs, one sees that the difference in scores between males and females is *"a sizable proportion of the interquartile*

range." Another one thought the difference in math scores is more meaningful "*because the box is not exactly lined up for the math scores like it is for the verbal scores.*" One student remarked: "*Q2* (second quartile) *for males is almost equal to Q3* (third quartile) *of females, showing that the differences are much greater.*" Finally, there was one student who claimed that by comparing the SAT and FYGPA distributions, we see that "*Males are better than Females in overall.*" The instructor challenged this student's statement by saying that based on the displays, it seems that males do in general have a higher SAT scores, but one could not claim that males are better at math, since there is a substantial overlapping of distributions. He added that even if such a claim seemed plausible, looking at a graph is not enough; one also needs numerical summaries to support conjectures drawn by looking at displays.

Making it Quantitative

Students followed the instructions on the handout in order to quantify the differences in mean scores by dividing them with the standard deviation, and used this relative to standard deviation scale to make comparisons.

SATGPA	Summary Table	
⇩ ⇨	**math**	**FYGPA**
	0.50473743	-0.20048576

$$S1 = \frac{\text{mean}(?, \text{sex} = \text{"M"}) - \text{mean}(?, \text{sex} = \text{"F"}}{\text{stdDev}(\ )}$$

Figure 5.8 – Differences in Mean Math SAT Scores and FYGPAs divided by Standard Deviation

The majority of students argued (in both the class-discussion of this question and the responses they put down) that comparing relative scales indicates

that the difference in math SAT scores is higher than that for FYGPAs. Some considered the difference to be large, others however considered it to be insignificant: "*Both are very similar, within 1 standard deviation of each other.*"

Students then compared sex differences in median SAT scores and FYGPAs after they had divided them by the interquartile range.

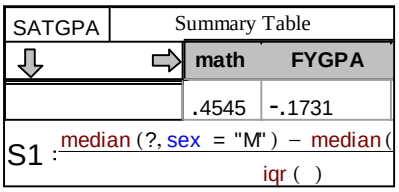


Figure 5.9 - Differences in Median Math SAT Scores and FYGPAs divided by Interquartile Range

Their conclusions again varied. Some students argued that results are very similar for both SAT and FYGPA while others noted that the difference in math SAT scores was higher. One student attempted to give an explanation for women's lower scores on the math part of the SAT: "*The median for males on average is slightly lower, meaning that females were more focused on the task at hand than males were.*" We see here again how students' background knowledge on an issue often discussed in the media affects the way they make sense of the data.

The next question was asking students to include *totalSAT* score (sum of math and verbal scores) in the comparison. Students added *totalSAT* to the summary tables and found the relative difference in mean *totalSAT* scores to be 0.343, and in median *totalSAT* scores to be 0.25. All of the students participating in the class discussion regarding this question argued that *totalSAT* shows males

did better overall than females, but that the difference is not too big: "*The mean and medians are still higher for males, but the verbal helped the females catch up.*" Looking however, at the students' written responses, I found that not everybody in the class shared the same opinion. There were indeed several students who did argue that there was some difference, but it was not large, others however found the difference to be quite significant*: "TotalSAT shows that total SAT of females is definitely lower than that of males."* One student went as far as claiming that "*Male has higher SAT. Male has higher math skills.*" Another one was not as assertive, but still argued: "*Male has higher SAT. Males are better at this type of test format. The test is biased to women. Males may be better at math.*" These are observations that go beyond the data (Erickson, 2000), and if they had been brought up during the class discussion, the instructor would have challenged them. We see here once again both the positive and the negative role that familiarity with the context of the problem can have. Students' world knowledge on this controversial issue allows them to probe more deeply into the meaning of the data, but at the same time leads them to express opinions that are too strong based on the data provided.

The last question on the handout was asking students to explain why one could not argue that, since the difference in SAT scores is over 40 whereas that in GPAs only about 0.15, the difference in SATs is more significant "without using the words, 'interquartile range', 'standard deviation', or 'variance'." Before discussing the question, the instructor asked students to put down an answer to this question. When I analyzed the handouts, I found that many students did not

seem to have understood what this question was asking. For example, some students wrote that we cannot compare high-school level with college-level assessments, and others made the correct but irrelevant here observation that, according to this data, SAT is not a very good predictor of performance in college, or of intelligence. Less than half of the students noted the different scale of SAT and GPA scores:

> SAT scores range from 0 to 1600, GPAs range from 0 to 4. Thus the difference in SAT scores is very large and the GPA difference is very small. Thus a small difference in GPA is greater than a large difference in SAT.

During the class discussion, a student noted that one could not draw conclusions just by looking at absolute differences *"because they have a different scale. The one is out of 800 and the other one is out of 4."* Another student added*: "Different range of scores."* Other students agreed. The instructor stressed that dividing the difference in the means by the standard deviation and the difference in the medians by the interquartile range allows us to transform scores coming from different scales to the same scale.

This activity provided a link to the topic of standardization of scores coming from a normal distribution. A student who was absent the day that the class engaged in the "SAT and GPAs" activity, but who did the assignment on his own and turned it in on a later day, saw the connection:

> Basis of scores on SAT and GPA are different. GPA is on a 4.0 scale while SAT is out of 1600 – that's why the formulas we used did a better job of comparison. Similar to the purpose of Z-scores, comparing apples to oranges.

Impact on student learning

The following question (from Pfannkuch and Brown, 1996) was given to students right after the class had engaged in the SAT activity:

> A small class was given a test on arithmetic and the results were recorded. The same test was given a few weeks later. The box plots for both sets of results are shown.
> Have the results changed significantly?

Figure 5.10 – "Test Results" Task

Five students who had just completed a statistics course and were participating in a study by Pfannkuch and Brown (1996) were also asked this question. The authors report that when first interviewed, all five students relied on their experience and produced deterministic explanations for the difference in test results. In this study, on the contrary, the proportion of students judging that test results changed significantly was only 25%. Most students concluded that the almost identical interquartile ranges meant the difference in scores was small compared to the variation, and therefore might have simply been the result of chance. The few students who saw significant changes, gave arguments such as: "*Changes did occur – the high scores dropped and the median is also lower – it seems the students either forgot what they knew or were out in the sun too*

much."; "Perhaps they did not worry about the second test as much because they did all right on the first one."

At the same time though, this assessment task helped us see that a number of students had misunderstandings about what the boxplots represent. Several students gave responses that suggested they were still confused about boxplots and thought that what their middle line shows is the mean of a dataset: "*They have changed but not much. The average score on the second test fell slightly."*

The "Matching Histograms to Variables" task (Question 10, Appendix A), on which students had done extremely poorly in the pre-assessment, was given again along with the task discussed above. Student improvement from pre- to post-assessment was substantial, as can be seen in Table 5.1, which shows the proportion of students correctly matching each of the five histograms with a corresponding variable:

Table 5.1 – Post-assessment vs. Pre-assessment Results on "Matching Histograms to Variables" Question

Response	**Pre-assessment %**	**Post-assessment %**
Histogram I	33	97
Histogram II	33	59
Histogram III	20	63
Histogram IV	36	72
Histogram V	30	75

Seventeen students (53%) correctly matched every variable to its corresponding graph compared to only three students (10%) in the pre-assessment. Students seemed much more aware of how features of a distribution are related to the shape of a graph. For example, all of the students who matched

Histogram II with Variable B (last digit of Social Security Number) noted the approximate uniformity of the graph, whereas in the pre-assessment it seemed that the 10 bars of the histogram were the only reason behind the choice of most students who had made the same match. Similarly, most of the students matching Histogram III to the number of medals won by medal-winning countries, now recognized what the graph represented: *"Olympics are usually dominated by 1 or 2 countries and many countries get 1 or 2 medals."* Nonetheless, the tendency to think of bar graphs and histograms as representations of raw data, and not appreciate the data reduction involved, was still observable despite the improvement. For instance, there were again some students who thought of Histogram III as representing the number of medals won by individual countries. There also some students who still did not to seem to think carefully enough about how the features of a distribution affect the patterns emerging within the variation. For example, one student thought it is Histogram V (approximately normal) that describes the distribution of the last digit of the Social Security Number because *"it should vary a lot"*.

The assessment showed that students still had some difficulties interpreting histograms and boxplots. Instruction continued to put a lot of emphasis on helping students improve their ability to read and understand graphs and relate features of a distribution to the shape of its graph (Scheaffer et al., 1996).

What's Common Here? "Discovering" the Binomial Distribution.

Practical application of any concept or technique involves three steps: (1) Recognizing applicability, (2) Applying method, and (3) Interpreting results (Wild and Pfannkuch, 1999, p.3). Unlike conventional instruction that tends to focus on Step 2 - mechanical application of methods – the emphasis on this course was on Steps 1 and 3. Recognizing applicability and interpreting results in context are much more challenging than learning techniques which one can teach "simply talking about them, establishing them with a few exercises and then moving on" (Wild and Pfannkuch, 1999, p. 231), but are necessary if Step 2 is to have any utility. For this reason, the instructor's main goal in introducing binomial distributions was for students to be able to recognize a binomial setting. He began by giving students a description of five different situations, all of which could be modeled using the binomial distribution, and asked them to figure out *"the common properties that these different situations have."* Group work was followed by a whole class discussion during which the four main properties of the binomial distribution were laid out: (1) Number of trials is finite; (2) For each trial, there are only two possible outcomes, *Success* and *Failure*; (3) *P(Success)* remains the same from trial to trial; and (4) Each trial is independent of other trials. The instructor remarked:

> Inst.: Many real world problems have these properties. We have now identified some properties that give us a nice distribution that applies to many real world situations including the five cases you have investigated.

He pointed to students that *P(Success)* is nothing but *"an assumption we make typically using experience, prior information."* It was only after students

had brought up several examples of different situations which could be modeled using the binomial distribution, that he introduced the probability formula of the binomial distribution. He did that noting that using the formula it would now be very easy to calculate, for example, the probability that Spurs wins 2 out of 10 games against Nicks (a situation they had talked about when discussing combinatorics): *"See how easy now? We can use this simple formula to solve the problem."* Students used the formula to calculate $P(X=2)$, where X is the number of victories by Spurs in 10 games. Then the instructor asked them to find $P(3 \le X \le 9)$, in order for them to see that this would be time-consuming to calculate using the formula: *"Too complicated for me. So, I'm going to give you a way. Statisticians are smart in this regard. Statisticians are smart in this way."* This was his way of leading to the introduction of the binomial table.

In order for students to recognize that the probability of success is a key feature of a binomial distribution that helps deduce the likely shape of its distribution without actually collecting or analyzing data, the instructor asked students to build the probability distribution table for a certain binomial distribution with $P(Success)=0.2$. After filling up the table, students drew the bar graph and saw that the distribution was skewed-to-the right. The instructor then asked students to look at the binomial table and describe what the shape would be like for different probabilities. This helped them see that when $P=0.5$ the distribution is symmetric, when $P<0.5$ it is skewed-to-the right, and when $P>0.5$ it is skewed-to-the-left.

Is the Student's A Score Rare? What About Student's B?

The way the instructor approached normal distributions was not the typical approach used in introductory statistics courses. The emphasis was not on teaching the formal properties of the normal distribution, but on helping students understand why one could use the normal distribution to model a certain variable and in what ways this is useful. An idea he used to help students appreciate the usefulness of normal distributions was that of the "rare event":

> Inst.: If they are going to learn something about the normal distribution, I'd rather have them think: "Oh, yeah, based on normal distribution, that means that we are way out, so we can consider this to be a rare situation." That's why once I understand that my situation is to find some probability, then I know I have the table to use. But use of the table should not be the thing. Otherwise, you spend hours and hours showing them how to use the table. This is not the goal. It doesn't make any sense.

The instructor introduced normal distributions through the following problem:

> Inst.: How can standard deviation be applied? Other than measuring range and spread, standard deviation can help us do more. Let's give an example. Let's take the SAT scores distribution (drew a normal curve with a mean of 500 and a standard deviation of 80).

He asked students whether it is a fair assumption to make that the SAT scores are normally distributed and students recalled that the distributions of SAT scores they had analyzed in the "SATs and GPAs" activity were. He then asked:

> Inst.: Suppose now student A scored 750, and Student B scored 600. Is A a rare event? Is B a rare event? In many situations, this application is very important. When you go out to work you will use that in manufacturing processes to decide if something is rare or out of specifications. Just by looking at this picture, what can we say? Is 750 an extreme score? What do you think?

S: Yes.

The instructor noted that they could do more than just looking at the picture. He introduced the Empirical Rule as a means *"to quantify"*, *"to help us see whether our performance is extremely high or not."* He wrote down the rule, emphasizing that it is only approximately that it holds. He then drew a normal curve with mean 500 and s=80, highlighted the area where X>750, and asked students whether based on the following rule 750 is rare:

A case is rare if it falls outside 2 *s* of $\bar{x}$

Students responded that it is, since there is less than 5% chance to get a 750. A student then asked *"Do I still have the same results when I don't have bell-shaped?"* and this led to a discussion where the students pointed out that if the distribution is skewed the rule would be wrong. The instructor noted: *"You brought up a very important point. When you have some data, look at its characteristics so that you won't do a trivial mistake."*

The instructor then introduced *Z-scores*. He drew a *N(500, 80)* curve corresponding to the distribution of SAT scores, and labeled the horizontal axis *Scores*, and the vertical axis *Relative Frequency*. He also drew a histogram that had approximately the same distribution. He noted that the normal curve is *"a smooth curve"*, whereas *"this is real data"* and that *"with the smooth curve, I don't do the histogram, I'm just looking at the curve that passes through it."* He explained to students that, unlike real datasets that are only approximately normal, this normal curve is a formal model, it has a perfect normal distribution.

Next, the instructor asked the students to look at the normal curve and try to estimate *P(X>450)*. Students gave different guesses and he remarked that although their guesses "*sound*[ed] *OK*", they "*could do it more precisely.*" He introduced "*the yellow table, which gives you much more than the empirical rule can do.*" He explained that using this table "*we can find the information we need not only for 67% or 95%, but for any point on the curve.*" He described how the *Z-score* is a standardized scale free of measurement by outlining the properties of the *Z-distribution*, gave students the formula for finding Z, and showed them how to use the table. Students calculated *P(X>450)*. The instructor stressed that this probability holds exactly for "*the curve*", but only approximately for "*the histogram*". Use of the empirical rule and the *Z-scores*, he explained, gives us theoretical percentages or chances of different values of the "*ideal*" normal distribution, which only approximately hold for the actual data distribution.

Students then worked in filling up a table where they were given the *X* score, found the corresponding *Z*-score, and decided whether it was rare or not. For 680, which translated to a *Z-score* of 2.25, the instructor asked: "*So, you have 2.25. What would you do if you had to make a decision based on this?*" A student responded: "*I'd say it's rare.*" The instructor remarked that one could do this, but they could also think "*Is the 2 standard deviations rule realistic?*" because "*sometimes it is, sometimes it is not.*" This led to some more discussion where it was pointed out that statistics provides an aid in decision-making, and not rigid rules that hold regardless of the nature of the situation.

Understanding that the tails of the normal distribution correspond to unusual outcomes is analogous to the idea of rejecting the null hypothesis when the distribution it assumes would make the data be at the tails of the sampling distribution (Cobb, Witmer, and Cryer, 1998). Thus, the idea of looking for "rare events" when exploring normal distributions provided a link between exploratory and inferential statistics**.**

Probability, Causation, and Variation

For this instructor, the most important ideas about probability that need to be addressed in order to link what has already been taught to what will follow, are those of *chance* and *independence*. Most statistics books and many statistics instructors, however, teach the chapter on probability as a separate topic, not connected to the rest of the concepts introduced in the course:

> Probability is a very important aspect of statistics, but the important part is to introduce the ideas of chance and independence. For example, rare event…this won't probably happen because it seems very unlikely…very small probability of occurring. Many people teach probability without understanding how it is linked to what follows. The point of independence is that random outcomes are i.i.d.[independent, identically distributed]. The way you take your sample will dictate whether your sample is i.i.d. or not.

Introduction to Probability

Students were introduced to the relative frequency definition of probability through the following experiment. One student tossed a coin and they marked the number and ratio of heads, then two students tossed a coin and they did the same, etc. Before the experiment began, the instructor asked students to

make predictions about the relative frequency of heads. After they had completed the experiment, he drew a graph of the relative frequency of heads against the number of tosses and students confirmed their prediction. They noted that the graph shows that *"the more the experiment the closer the relative frequency of heads to 1/2"*, but *"when sample size is small, the relative frequency fluctuates a lot… there is high variation."* The discussion continued and eventually the idea of probability came in. They discussed the difference between probability and actual relative frequency. The instructor wrote on the board:

$$\boxed{\begin{array}{l}\text{P(H) is the ideal situation that we cannot observe.}\\ \text{Proportion(heads)} \approx \frac{1}{2} \text{ as \# of experiments} \uparrow\end{array}}$$

He stressed that such a claim is based on the assumption that the coin is balanced and that to figure out whether this or other claims are indeed true, one ought to do experiments similar to the one they did, and/or use prior knowledge. It was only after extended discussion that he introduced basic ideas and conventions of probability such as conditional probability and mutual exclusiveness.

Independence

When they were discussing independent events, the instructor made sure he emphasized the complexity of real-life situations rather making simplistic assumptions that would conflict students' common sense. After he had asked students to give examples of events that are independent and they had given typical examples such as coin tossing and die rolling, he asked them whether the

success of a *"free throw"* of a basketball player is independent from the success of his previous *"free throw"*. Students argued that it depends on how the player responds to pressure, on how well he did on the previous throw etc. The instructor remarked:

> Inst.: You see, in real world situations it's hard to tell because for example here, it depends on the person's psychology. So, in real life it's hard to say with a straight yes or no.

Notice how the instructor did not reject students' causal explanations. Nonetheless, "hot hand" is one of the main examples many statistics educators often use in their pleading for probabilistic reasoning. Tversky and Gilovich (1989), using empirical data, showed that a binomial model well explains runs (streaks) in basketball player failures. According to this model, the chance of success in a shot is independent from the previous shot, and Tversky and Gilovich, and subsequently many teachers and researchers, concluded that people's tendency to detect patterns (hot hands) is often unwarranted. One need not look for specific causes like nervousness since there is no other "pattern" than chance pattern explaining the data. Some have even gone as far as concluding that the belief in "hot hands" is an illusion. However, this instructor understands what Biehler (1994) has pointed out – that even when the binomial model well explains the variation in a dataset, it does not mean it is the "correct" and unique model for this phenomenon. One cannot exclude alternative models which give better prediction and which suggest causal dependence of individual throws.

Unlike some instructors who only emphasize the similarities of the streaks that sports fans see in sports data to the "gambler's fallacy", this instructor also

emphasized their difference. This is a real-life situation he pointed out, and *"in real life it's hard to say with a straight yes or no."* Similarly, when talking about slot machines in the casino, he noted that, although in theory *"when you put a coin and you pull it down and then you put another coin and you pull it down, although those 2 events should be independent, mechanically they might not be."*

Next, discussion revolved around random sampling. They decided that randomly selecting a sample out of a small population would mean that there is no independence, but as long as the population size is large, we do have approximate independence. The instructor stressed that statistical methods depend on the assumption of independence characterizing random sampling. He emphasized again the importance of the data production stage:

> Inst.: That's why random sampling works. What I'm trying to point out is that it's important how you pick the sample. Once you have the sample, you've already decided that. It's very important to understand independence before you collect the data…This kind of concept is not easy when coming to applications. But, the way you should think about it in real situations is that independent or not is determined when you do sampling. Independence is determined by sampling. These things here are the consequence, OK?

Then students split in groups and worked on the following problem:

> In the basketball championship games, Spurs won in 5 games (4 to 1). In the regular season, Spurs played with Nicks 5 games and Spurs won 3 games. Assume that Spurs winning probability is 0.6 when against Nicks.
>
> (1) Find out all possible situations (combinations) for Spurs to win in 5 games (e.g. SSNSS)
>
> (2) If Spurs has a winning probability of 0.6, what is the probability Spurs will win in 5 games?

When the class got back together to discuss the problem, the instructor stressed that the different formulas for calculating probabilities are based on the assumption of independence. It is because we assume that *"each game is a fresh start…it's not affected by previous outcomes"* that we can use the formula $P(S \cap N) = P(S)P(N)$, to calculate the probability that *"in the first game Spurs wins, and in the second Nicks wins."* He continued:

> Inst.: What if we have 5 games? Each game is a fresh start and so $P(NSSSS) = P(N)P(S)P(S)P(S)P(S)$. Independence plays an important role here. That's why I keep on repeating that I assume each game is a fresh start...Independence plays a very important role. If you don't have independence then you cannot do it like this.

Independence of random events was not a topic that was introduced on one day, and then never discussed again. The instructor knew that grasping the statistical notion of independence is not that easy for students. Students' responses to a question given to them the day after independence was introduced, which was asking whether in a marketing survey where 400 individuals are randomly chosen from a large city we have approximate independence, were testimony of their difficulties. Seventy-two percent of the students agreed, but only half of them gave an adequate explanation. These were the students who explained that *"because you are choosing such a small sample out of a very large population, you would be approximately starting from a clean slate."* Several students gave explanations that were either too vague or wrong. A couple of students thought independence means not including the same person twice in the sample, and argued that in a large city this would most likely be the case. A few others thought of independence as the opinion of one person in the sample not

exerting an influence on the opinion of another, and concluded that we do have independence here because of the large population size. Four of the eight students (28%) who argued that we do not have independence, gave explanations indicating they had a similar notion of independence: *"There is a chance that the individual selected is not independent of the individual selected previously because the two individuals could have same job and same interests."* The other four correctly put down that, since we are sampling without replacement, the samples are actually not independent of each other, but missed the idea on which random sampling is based – that when the population is large, this is not a serious problem and we can assume approximate independence.

The concept of independence is very important, and for this reason the instructor continued grasping any available opportunity to help students understand statistical independence and not confuse it with connotations that the word independence has in everyday speech. He would be constantly reminding students that many statistical methods are based on the assumption of independence, of *"a fresh start"*.

Sampling Distribution

The logic of inferential statistics is based on the notion of sampling distribution. Sampling distribution is perhaps the hardest concept introduced in the introductory statistics course. Comprehending sampling distributions means understanding the relationship between "three similar-seeming but in fact fundamentally different sets of numbers, each set with a different role and meaning – the population, the sample, and the set of values of the statistic"

(Cohen and Chechile, 1997, p. 208). In the meeting I had with the instructor before formal introduction to sampling distributions, he stressed how difficult this idea is for students, since to grasp it *"they need to understand the concept of distribution, shape, normal distribution, sample, of sampling variability."* In addition, they need to realize that, with the sampling distribution *"the X-axis has changed"*. The idea that, on the horizontal axis, *"the scale has changed from one single observation to an average of observations"* is very difficult for students.

In regard to the role of technology, the instructor said he believes that *"computer simulation helps students understand that different samples give different means, but that does not help transfer into the distribution and the standard error concept."* He personally does not think students have a difficulty understanding that different samples have different means because of variability, and that a larger sample size is preferable. It is *"the degree of the fluctuation and how that should be quantified they have a hard time with"*, because *"quantification means precision and that's very difficult, and so that is the hardest part."* Standard error is a *"very, very difficult"* concept for students. It involves ideas not used in the everyday world. In order for students to understand standard error, they first need to understand histograms very well, to understand how standard deviation is related to a distribution, and also to realize that we have a transformation from single observations to a function of a set of observations. It is much more complicated than estimating likely intervals for individual observations coming from the population:

> Inst.: So, interval estimation, I think students are able to understand that concept much more easily because they use that in the real world. The standard error concept is much harder. Changing from descriptive to inferential statistics is a big jump. They stick their mind to the descriptive. Even though we say confidence interval is an estimation, to them it's nothing but descriptive because it's coming from a sample. Even though we talk about sampling distribution, to them it is just descriptive. They don't consider that as inferential, and then the jump is on the sampling distribution of $\bar{x}$ and that jump is just a heck of a job to do.

He was hoping this time students would do better, but did not expect miracles: *"I hope, I don't know. I used all different ways to do this before…very few students really grab the concept."* He noted that the short duration of the summer course makes it even harder: *"We just covered normal distributions a couple of days ago and we're going to do this tomorrow, you know?"* However, he was hopeful that the emphasis he had put this time on *"investigating distribution and shape and variability"*, had given students *"some understanding they could apply here so that we don't need to build everything from scratch."*

SOS Scores Activity

This activity, which lasted two days, was the main activity the instructor used to introduce the ideas related to sampling distribution to students. Students retrieved a file that included 120 SOS (Student Opinion Survey) scores, belonging to 5 faculty members, who each had taught the same course 24 times. They first drew a histogram of the 120 individual SOS scores, which revealed a skewed-to the left distribution. Then, the instructor guided students as to how to get the average SOS scores of the 5 faculty members for each of the 24 semesters. They then drew a histogram of the 24 averages. The instructor asked students to compare this histogram of mean scores to the histogram of individual scores.

Because Minitab is not a dynamic learning environment, students got confused trying to bring both graphs right next to each other and also, at the instructor's request, to adjust the graphs to make sure they both had the same scale. Several students started complaining that they *"do not have a clue about what is going on."* It was the first time I was seeing students reacting like this. Till then, they all seemed to enjoy the course and to feel quite confident. Eventually, they managed to adjust the two graphs and bring them next to each other for comparison. A student noted that the distribution of average SOS scores had a smaller spread and looked *"more normal"*. They then calculated the mean and standard deviation of the set of 120 individual scores, and the mean and standard deviation of the set of 24 average scores and found that the two means were very similar, but that the standard deviation for the set of averages was smaller.

Next day, class began with the instructor reminding students about the previous day's activity, where they saw that there was variation among sample averages which, however, was smaller than the variation among individual scores. Since there is variation among sample averages, he added, *"there is a distribution about sample average."* In contrast to the population mean, *"sample average varies for different samples"* and *"our goal is to make sure we do not come too much away from the truth – that's the key."* If the distribution of the sample mean has a big spread, we have the chance to be far away from the truth but if the spread is small then we are *"guaranteed"* to be close to the truth:

> Inst.: The key is we don't know the truth, but it is guaranteed, it is guaranteed that if I have a small variation, that sample mean is not very far

away, therefore I don't make a big mistake regardless of what I pick. Sampling scheme will guarantee my sample is good.

(Note how here, even this instructor who put so much emphasis on helping students increase their awareness of variation, uses the word "guarantee", which has no place when dealing with finite statistical processes. This statement of the instructor, might have left students with the wrong impression that a larger random sample *guarantees* a more representative sample.)

The instructor added that how well the sample mean estimates the truth depends on its distribution and that the group activity they were about to engage in, which was a continuation of what they had done the previous day, would help them investigate further the properties of sampling distributions. He asked students to assume that the 120 SOS scores they looked at the previous day, are the entire population of SOS scores. He explained that although we do not actually know the population size, since *"there are so many SOS scores every semester"*, by assuming that those 120 scores are the entire population, we could then *"make comparisons and summarize patterns."* Students split into groups and worked together on the activity, where they had to:

(1) Take 1000 random samples of 2 SOS scores from the population of 120 SOS scores and obtain the sample mean for each of the 1000 samples. Draw a histogram and a boxplot of the 1000 sample means of size 2.

(2) Repeat the same process for samples of size 5, 10, 15, 20, 25, 30.

(3) Create a table with numerical summaries (mean, median, standard deviation, min, max, Q1, Q3) for the original population and each of the 7 sets of sample means.

(4) Use the graphs and numerical summaries to answer different questions about the properties of sampling distributions.

After all the groups had completed the activity sheet, the whole class resumed. The instructor told students that he hoped the activity helped them understand the idea of sampling distribution, the most difficult concept of the class. He then went on to explain that the purpose of the course up until then was to equip them with the tools they could now use to make inferences. If we need to make inferences about some unknown population mean for example, he continued, knowing the sample mean is not adequate. We also need to know its distribution properties to decide how good that sample mean is. He asked students to describe some properties of this distribution, and a student remarked *"When your sample size gets larger, then your curve gets taller and narrower."* The instructor used the following example to get across to students the advantages of taking a large sample:

> Inst.: Does anybody have an orchard field at home? If you sell apples, you will probably be able to have 1-2 rotten apples hidden in a large bag, but not in a small bag. You can sell more easily the large bag. That's what average says. Average of a large sample size takes care of those rotten apples. If you have a large sample, it smoothes out extremes. That's why the distribution is so close.

The other properties of the sampling distribution were also brought up by students and were discussed. He finally introduced the Central Limit Theorem as a summary of these properties:

> Putting these together becomes a notation of what we've been talking about:
>
> When n large: $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Confidence Intervals

Shaughnessy (1997a) has found the idea of an interval of likely values in a sampling situation to be quite accessible for students. In a previously conducted study (Meletiou, Lee, and Fouladi, 1999), we used one of the items Shaughnessy had used in his research. The conclusions we drew were very similar to his. We also found the idea of a range of likely values in a sampling context to be very accessible even for students with no statistics background.

Although students seem to be quite comfortable with the idea of likely intervals, confidence intervals are among the most confusing topics in introductory statistics. Erickson (2000) sees the way "official confidence intervals" are defined being a problem. Understanding that 95% confidence interval means that if we were to draw many samples and calculate the confidence interval for each, 95% of the intervals would contain the true value, and not that there is a 95% chance that the true value is in our interval is hard for students. Also, the way we explain what makes a confidence interval work is even more confusing:

> The definition above is all about taking multiple samples from a single unknown population. Yet we use the interval to define a range of reasonable population values – derived from a single known sample. It's confusing, so we explain it by taking multiple samples from multiple hypothetical populations, which, while correct, can confuse students all the more. (Erickson, 2000, p. 269)

Erickson (2000) believes that "although the tumult about confidence intervals is important for AP or college students" (p. 269), the most important issue for every student to understand is that whenever producing an estimate of some parameter based on a sample, our estimate is probably wrong due to chance error. Therefore, the estimate is misleading unless we figure out how big the chance error is likely to be. This is a belief that the instructor shared too. His emphasis was on helping students understand the logic behind confidence intervals, instead of teaching them how to *"plug in numbers and formulas"* without understanding the real purpose behind what they are doing.

No. of Raisins in a Box

The instructor introduced confidence intervals through a whole-class activity where students tried to guess the number of raisins in a box. He gave one packet of raisins to each student. First each student estimated the number of raisins in their box and then counted them. When students were done counting, they each gave both the estimated and the actual number of raisins. The instructor had overhead projection and they entered the values in Minitab. Students compared the column of *Guess* with that of *Actual*, and noted that *"Guess is too low"* since almost everybody underestimated the number of raisins. They also noticed that *"Guess has more variation."* The instructor then told students that he

would now *"use some of the tools* [they] *learned before, basic statistics."* Using Minitab, he drew graphs and found summary statistics for each of the two datasets. Students copied the summary statistics as well as the histograms of *Guess* and *Actual*. Before drawing the two histograms, the instructor remarked:

> We have to use the same scale. We need to. It's so easy to cheat by manipulating shape. Tell you what: 3-4 years from now your boss asks you to present something. You know he wants to see major improvement. Make your Y much wider. Then you're going to show a huge improvement, even if it actually was only 1%. OK! You've got a trick here. And then you will be fired.

Students looked at the two histograms and noted that it was obvious *Guess* had a more varied distribution. *Actual* was more "*compact"* and also "*closer to normal"*. The instructor agreed*: "I've noticed it for many semesters that it comes very close to a normal distribution."* Students also examined the boxplots of the two distributions and noted that the one for *Actual* had higher values, was tighter, and more symmetric.

The sample mean of *Actual* they found was $\bar{X}$ =33.5. The instructor asked*: "How am I going to get* μ*?"* A student suggested to "*count everything"*, but at the teacher's remark that this would mean "*opening every box"* he changed his mind and said instead that "*you estimate it by* $\bar{X}$*."* The instructor approved the student's idea but added*: "It is an estimate, it is not the truth, right? I say that's a pretty good estimate, but how do you do better than that?"* What he tried to do here is an introduction to confidence intervals. He continued:

> Inst.: Every time people ask you to estimate something, let's say I ask you to estimate your average GPA, I'll probably say it's between 0 and 4.

S**:** Sure.

> Inst.: Can I do better than that? Well, if you are able to stay in this class, chances are you have at least a C. So, the average is probably at least 1.5 to 4. So, I can narrow it down. What am I doing? I'm giving you an interval. In real life, if I'm not sure this is it, I say I think it's between this and that.

He then asked students to discuss with their group ways to come up with *"a good interval"*. A class discussion followed, during which it was decided that intervals that are too wide are not very informative, whereas ones that are too narrow are very risky. The instructor then asked students to come up with a reasonable interval for the mean number of raisins in a box. He wrote on the board *"True mean number of raisins is between ____(a number) and _____(a number)"* and remarked:

> Inst.: That's what an interval is. Everybody write an interval and then we'll discuss whether your estimate is good or not. You can use any information here that you think is useful. This is more than a guess now.

Different students gave different intervals. One student suggested *"30 to 36…one standard deviation on each side"* (They had found s=3.34). Another student suggested *"27 to 40. 2 standard deviations. Events that are not rare."* Then instructor remarked: *"27 to 40 is from min to max…quite a big interval for true mean, right?"* and the student responded laughing: *"It's guaranteed to be there."* The instructor had to remind students that *"we talk about sample mean, we are not talking about individual raisins, individual boxes"*, for them to remember that they need to use the standard error and not the population standard deviation *"to estimate how close* $\bar{X}$ *is to* μ*.."* He told students that their idea of *"going up and down"* a certain number of standard deviations was right, but they

should have used *"some multiple of standard error."* This, he noted, is the most important concept in confidence interval estimation and the rest is computation: *"Be sure you understand that we add and subtract a certain amount of SE from* $\bar{X}$, *because that measures error between* μ *and* $\bar{X}$. *Here, I have 67% and 95%."* He then added: *"Of course, I can do more than that. I can do any confidence interval I want"* and introduced statistics notation to generalize the idea of confidence interval, *"to cover any case"*. Whereas students seemed pretty comfortable till that moment, once notation was introduced, they got very puzzled. Although the instructor told them that they "*should not be confused by that notation"*, since if they *"understood how* [they] *came to this part, this will be nothing but a simple formula"*, students seemed quite frustrated.

After the class ended, I had a meeting with the instructor. He remarked that one of the biggest obstacles is that students have a weak mathematical background and are intimidated by abstract notation. I asked him whether he thinks statistics courses should change so as not to include as much mathematical notation. He replied that although he tries to use mathematical symbols as little as possible, some use of notation is necessary because "*if you don't do that, they will be talking about 2 standard deviations and that's it. You've got to have some kind of notation*."

Hypothesis Testing

The instructor used the idea of rare event (corresponding to a small p-value) to introduce students to hypothesis testing. He tried to help students understand that a rare event could mean either that the initial assumptions are true

and an unlikely event happened by chance, or the initial assumptions are incorrect (Ballman, 1997). He also tried to use as simplified a language as possible. Other than that, the approach used in this class was similar to the one typically used to introduce hypothesis testing, as the short description of the following activity which was the main class activity related to hypothesis testing indicates.

Drug for Reducing Cholesterol Level

Students worked in-groups on this activity, which can be found in Appendix C. They had to decide whether the reduction in the average cholesterol level of a sample of 64 high-risk patients from 285 mg (Sample S.D. = 100 mg) to 250 mg, was significant enough for FDA (Federal Drug Administration) to consider the drug effective in reducing cholesterol level. First they had to make this decision based solely on their "common sense". Then, the handout informally introduced them to hypothesis testing, by helping them see that a way to decide whether the drug reduced average cholesterol to a significantly lower level is to check if a sample average of 250 mg is rare when the true average is 285 mg. Students found the z-score corresponding to a sample mean of 250 mg and standard deviation of 100 mg, and decided that the reduction in average cholesterol level was "a rare event" (it fell outside the two s.d. of 285 range). Based on that, they drew their conclusions regarding the effect of the new drug.

In the second part of the activity, the court case scenario was used to introduce students more formally to hypothesis testing, and the rationale it is based on. The ideas of null and alternative hypothesis, of a decision rule based on sample information, and of the two Types of errors and the relative seriousness of

each were discussed. In the third part of the activity, students returned to the "New Drug" case, and the analogy to the court case scenario was pointed out – that this problem also involves two decisions to choose from (Ho: the drug is effective, Ha: the drug is not effective), and there is always the possibility of drawing the wrong conclusion. Students then had to define what "Not effective" and "Effective" meant in the context of the "New Drug" problem. The handout then provided them with hints as to how a formal hypothesis test would be set up $(H_o : \mu = 285\ mg, Ha : \mu < 285\ mg)$ and how the decision as to whether to reject the null hypothesis or not would be reached. Finally, some more formal terminology such as *critical value*, *level of significance*, and p-value were introduced.

Learning with Fathom: Outside-of-Class Investigation

As already mentioned in the Methodology Chapter, in addition to investigating what happened in the PACE course, I also worked independently with a group of five students outside class to assess the effectiveness of the technological tool Fathom as an aid to conceptual understanding. I met with these students – either individually or in small groups – several times during the course. More than thirty hours of open-ended investigations of students interacting with technology were audio-taped and/or video-taped and transcribed. The information I gained about the kinds of intuitions students use to make sense of the stochastic and the ways in which their intuitions are shaped by technology is so rich that it could be the main theme of a dissertation study. Since, however, technology is not the main theme of this dissertation, I only provide in this section a very brief description of just few of the many activities that students engaged in.

Before doing this, I first give some background information, taken from Erickson (2000), about Fathom and the way it is structured.

Structure of Fathom

A Fathom document lives in a window and contains different components. The main one is the collection. The collection looks like a box with gold balls in it. Each of these gold balls is a *case.*

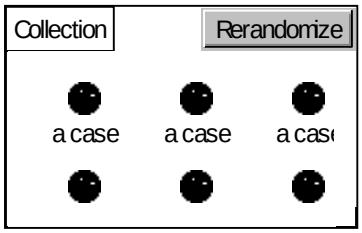


Figure 5.11 – A Fathom Collection

A case has one or more *attributes* (i.e. variables) which can be continuous or categorical. There are two kinds of attributes: regular attributes, which are *Case Attributes* – they have a separate value for each case in the collection of data points, as well as *Collection Attributes* - attributes with one value for the entire collection. Collection attributes are then statistics summarizing a collection. In Fathom, we also have *Derived Collections* – collections that automatically fill with data according to rules the user specifies in the *Formula Editor*. The two most important kinds are *Sample Collections* and *Measures Collections*. A Sample Collection is a sample of another collection, called its *Source* (the population it originates from). A Measures Collection, on the other hand, converts Collection Attributes into Case Attributes in order to enable the user to

record statistics (collection attributes) and thus build distributions of statistics. The user adds Collection Attributes to the collection using the *Inspector,* which he/she can access by double-clicking the box containing the collection.

Graphs, *Case Tables*, and *Statistical Tests* are also components. They do not contain data, but they provide ways of exploring data. A Case Table looks like a regular spreadsheet, with each case appearing as a row in the table and each attribute appearing as a column. Fathom also supports several different kinds of graphs, which the user can draw by dragging attribute names from either the Case Table or the Inspector to the appropriate axis of the graph, and then choosing from the menu in the corner of the graphing tool the plot desired. The user can also add things such as functions to the graph by using the Formula Editor.

Sliders are also components that can be used to control variable parameters.

I now give the "Coin Toss" activity, which was one of the first activities students engaged in, as an example of the kinds of interactions students had with Fathom. I then give a sketch of some of the other activities students worked on.

Coin Toss Activity

In this activity, students tested the predictions they had made in the pre-interviews as to what outcomes are likely when tossing a coin 50 times. I used the activity as an opportunity to informally introduce the notions of sampling distribution. Students first collected a single sample of 50 coin tosses by building a Table with 50 *cases*. Each of these cases had one *attribute.* The attribute, named *CoinToss*, was a binary variable whose value was either "Head" or "Tail".

Students entered the command RandomPick(“Head”, “Tail”) in the Formula Editor to simulate the tossing of the 50 coins. Thus, the column containing the *CoinToss* value of each Case, is a *Sample Collection* whose *Source* (the population distribution it comes from) is a binomial distribution with P=0.5.

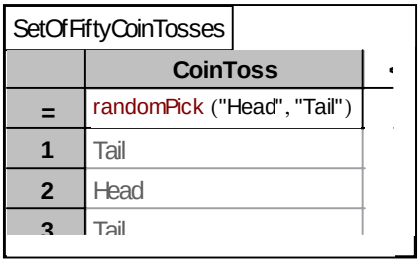


Figure 5.12 – A Sample Collection of 50 Coin Tosses

Students “dragged” a graph on which they “dropped” the *CoinToss* attribute, in order to get a bar graph of the outcomes.

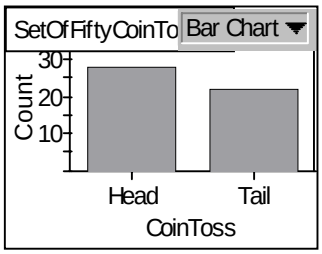


Figure 5.13 – A Bar Graph of the Sample Collection of 50 Coin Tosses

Pushing Control-Y, re-randomizes the values in the table. Students pushed Control-Y several time and saw how the bars of the graph, and the numbers in the table, would change. This helped them get an idea as to what values were likely. Also, I asked students to write down the number of heads that came up each time. After they had repeated this re-randomization several times, I

let them know that Fathom has a feature, the *Measures Collection*, which actually allows us to record statistics – in this case the count of heads. I explained to them that using Measures Collection one could collect many such statistics and then draw a histogram to see how those statistics are distributed. I showed them how to use the *Inspector* to ask Fathom to add the Collection Attribute *CountHeads.* They specified it by using the command *Count(CoinToss= "Head")* and checked that it gave them the same number as the number of heads displayed on the bar graph. They then used the *Collect Measures* command, which automatically collected the counts of 5 samples of coin tosses.

Before continuing, I made sure students understood what those counts represented: *"It's 5 times…each time out of 50"* (Lucas).

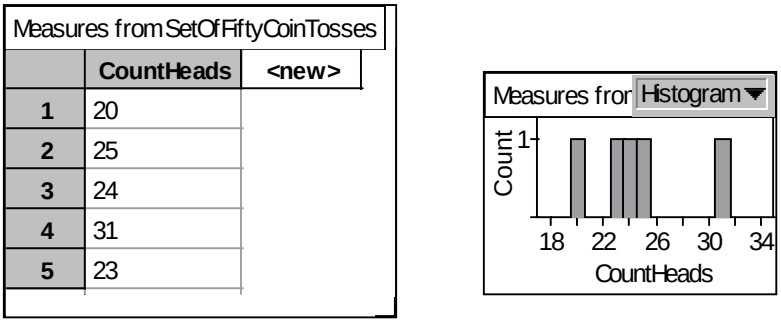


Figure 5.14 – A Measures Collection of 5 Sample Statistics and the Corresponding Histogram

Table 5.2 –Sample Collection vs. Measures Collection in Fathom

Coin Toss Sample Collection	**Measures collection**
Each case is a single coin toss	Each case represents a collection of 50 coin tosses
Whole collection is a collection of 50 single coin tosses	The collection summarizes many samples of 50 coin tosses (5 by default)
The measure is a single number (a statistic) that describes the number of heads in the collection.	Each case contains the count of heads of one collection of 50 coin tosses, so the collection has many counts of heads.
You can't estimate the mean count of heads here because this collection is only one sample	Can calculate the mean count of heads in this collection- by averaging the set of counts.

Students then collected the count of heads for a large number of samples and drew the resulting distribution.

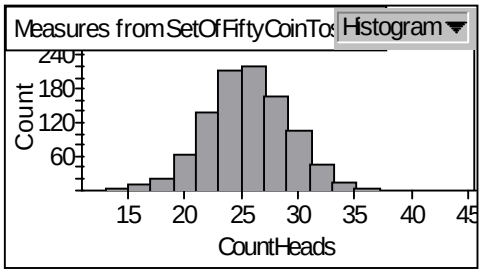


Figure 5.15 – The Distribution of a Measures Collection of Counts of Heads for a Large Number of Samples

Looking at the graph, students would make observations such as, for example, that 30 heads in 50 tosses was pretty likely, but that a number as high as 40 or higher would almost never appear. They also saw that the distribution of the numbers they collected looked *"like bell shaped."*

In the beginning, when students started working with Fathom, they did confuse single samples with the samples of statistics obtained with the Measures collection. Differentiating between the two was a key breakthrough for students,

that helped build connections to inferential statistics. At first, students created informal confidence intervals from distributions and decided what range of values of that parameter they would consider reasonable, and which values seemed implausible and begging for another explanation (Erickson, 2000) or, in students' terminology, which were "rare events". Eventually, processes became more formal, and students began comparing empirical probabilities with the theoretical ones.

Sample of Other Fathom Activities

1. Roll of a die: Students did simulations and verified the conjecture they had made during the pre-interview that it is quite likely to only get 3 out of the 6 possible outcomes in 7 rolls of a die. They then made predictions as to what would happen if, instead of 7 times, they rolled the die a bigger number of times. They would repeat the rolling of an increasingly bigger number of dice and see how the distribution would become increasingly uniform.

2. "Love is Not Blind" (from Scheaffer et al., 1996): This is another activity that I used to informally acquaint students to the idea of sampling distribution. Students first read the article "Love is not blind, and study finds it touching" (Associated Press, 1992; in Scheaffer et al., 1996). The article was describing an experiment where 72 blindfolded people tried to distinguish their partner from two other people of the same age, weight, and height. Students used a Binomial Distribution with $P(Success)=0.33$ to model this situation. They simulated the experiment many times, collected the proportion of correct

recognitions each time, drew a histogram of this set of proportions and checked to see how likely it is that 58% correct recognitions could have happened by chance.

3. Aunt Belinda: This activity is also an inference problem based on a simple binomial situation. Students were told that "Aunt Belinda" claims she has psychic powers that allow her to make heads come more often than tails when a coin is tossed. They were asked to decide how likely her claims are, given that during an experiment in her presence a fair coin was tossed 20 times and heads came up 16 times. Students did simulations similar to what they had done in the "Coin Toss" activity, and built the distribution of the Count of Heads. Looking at the graph of the distribution, they concluded that obtaining 16 heads out of 20 just by chance is "rare".

4. Drug for Reducing Cholesterol level: This is the same hypothesis testing situation as the one that the whole class had engaged in. Students used computer simulations to build the distribution of average cholesterol levels for a large number of samples coming from a population distribution they defined, which had a population mean of 285 and a standard deviation of 100. They put the statistics they collected on a histogram and saw that the distribution of those statistics was approximately normal even when the original population distribution was not. They looked at the histogram and saw that 250 seemed rare. In order to verify that one would reject the null hypothesis at the 95% confidence level, they drew a vertical line where the 5^{th} percentile was located and saw that 250 was below that line. They also compared the computer results with the ones they got in class and saw that the two were very similar.

END-OF-COURSE ASSESSMENT

In this section, I outline the findings from the end-of-course assessment, which can be found in Appendix D, and the follow up interview of the primary group. The analysis includes a brief discussion comparing the understanding of inferential statistics by students who participated in the outside-of-class activities using Fathom with that of the rest of the students.

Exploratory Data Analysis

All the students interviewed at the end of the course had a notion of distribution adequately close to the statistical one: "*Frequency is like the amount of people in each category. The distribution is like how the graph looks like…how many people are here and how many people are there.*" They all noted that graphs help us figure out what the distribution looks like and stressed that, in addition to graphs, numerical measures are also necessary to adequately describe a distribution. Most of them had good understanding of mean and median and of how the relationship between the two affects the shape of the distribution. They also understood that, in addition to knowing about the center of a distribution, one also needs information about "*the overall spread.*" They recognized that, when comparing measures of center, one should always take spread into account: "*You can't say, since the mean is 30 and the median is 40, the mean is less than median. If the range was from 0 to 50, yeah that's a big difference, but if you're going from like 0 to 1,000,000 that's different.*"

When in the follow-up interview I asked students to give examples of measures of variation, standard deviation was not the only measure they

mentioned. They all noted that measures such as the *"range of the box"* also give us information about the spread of a dataset. When asked to describe the meaning and purpose of standard deviation, Tim was the only one whose response indicated poor conceptual understanding: *"Like the…I don't know…I need my notes."* Showing the formula did not help this student: *"Symbols confuse me…I don't know…I don't know the definition of standard deviation."* All the other students had a pretty good grasp of what standard deviation is and how it is used. Peter, for example, said that standard deviation *"gives information about the distribution between the scores, the distance…outside of the center."* George explained that one calculates standard deviation *"to figure out the deviation, the average deviation of the scores from the mean"*, and this *"makes it easier to compare."* Anna understood how standard deviation is connected to normal distribution and how *"if it's not normal then the* [Empirical] *rule is just not right."* To make her point, she drew a skewed distribution and noted: *"It would not be good because here's the mean and you will have two standard deviations here but you will have to keep going over here."*

Students' performance on Question 10 of the end-of-course assessment asking whether two distributions with the same mean and standard deviation ought to look exactly alike, compared to that of students in the pilot investigation, is another indication of the positive effects of the emphasis of the course on variation.

Table 5.3 – Results of Current Study vs. Results of Pilot Investigation on "Same Mean and Median" Question

Response	Non-PACE %	PACE-Previous %	PACE-Current %
Yes	29.8	29.5	9
No	68.9	70.5	91
No response	1.2	0.0	0.0

Only three students (9%) in the current study claimed that having the same mean and standard deviation implies identical distributions. This is in contrast to the pilot study, where 29.5% of the PACE students and 29.8% of the other statistics students, argued that knowing the two parameters "*decides the shape of the distribution.*" A linear model was fitted using CATMOD (SAS Institute 1988), in order to investigate the relationship between course type and probability of giving the right response to this question. Course type was the independent variable and it had two categories. The first category was *Allprev*, where all statistics students in the previous study (i.e. both non-PACE and PACE-Previous) belonged, and the second category was *PaceCurr*, which encompassed students in the current study (PACE-Current). The analysis of variance indicated that course type was a significant factor in determining success (p-value = 0.0004).

In the pilot investigation, although the question mentions nothing about normal distributions, several students gave responses such as: "*Of course, because they have the same mean and standard deviation. If one has a normal distribution, the other one should look the same"*, or *"Yes, because mean and standard deviation are what makes the curve."* Those students seemed to be thinking only in terms of "perfect" normal curves where the mean is the middle

point and the standard deviation does determine how the data is spread around that middle point: *"Yes, because the mean is the middle point of the distribution and the standard deviation is the spread of each and if they are the same they have the same distribution."* In this study however, almost everyone recognized that *"many different shapes can occur from samples that have some similar attributes."* Several students also did drawings to make their point. One student, for example, drew two distributions, one uniform and one with just two bars at a distance from each other, that had the same mean and standard deviation, and wrote: *"Standard deviation is similar to an average distance from the mean and may not explain shape."* There were also several students who argued that two distributions with the same mean and standard deviation could look quite different *"because the variability might differ."* It is encouraging that those students perceived variability as something more than standard deviation.

Despite all the emphasis of the course on helping students improve their ability to construct and interpret graphs, they still had some difficulties. Even at the end-of-the course, when given again the seemingly easy question (Question 1) of having to decide, by looking at the histogram of two distributions of scores which one had more variability, five students (15%) gave the wrong response (distribution A). Even among those who chose distribution B, some might have had misunderstandings, as I found in the follow-up interview of Tim. This is how he explained why distribution B has more variability:

> Tim: I mean, it has more variability cause like the people here…the highest frequency is here…and the highest frequency on this one is 2…12 different variables here. This has more people… Because this is 14 and

> this 2…a difference of 12. And this is 3 and this is 13, and so…B should have more variability.

Instead of looking at the horizontal axes of the histograms to compare their spread, Tim was looking at their vertical axes and was comparing differences in the heights of the bars (i.e. differences in frequencies among the different categories). The explanation Lucas gave for having chosen histogram A in the previous assignments of the task, showed that he had also been looking at the wrong axis: *"I had chosen it because…the height is very… there is a lot of different heights. I was looking at the height of the graph…but now I understand it that it's here that we should be looking at…I was looking at the wrong side."* He, however, now realized that *"variability means spread…and also* [he] *would include the range of the scores",* and for this reason gave the right response at the final assessment. Tiffany also said that she had, in the pre-assessment, claimed that histogram A has more variability because she *"didn't know then the definition of variability."* She explained: *"I was confused between this…like the height and the width. I just got totally confused because…in some ways I was thinking that this has more variability and it doesn't make any sense right now…I just got confused from the height of it…"*

At the completion of the course, we also gave students the same question the instructor had given at the end of the previous semester, where they had to describe how a histogram of the distribution of salaries for individuals 40 years or older but not retired would look like (Question 6). As already discussed, with the exception of 3-4 people, all students had argued that the distribution would be skewed-to-the right, not because they understood how the histogram would look

like, but because they confused it with scatterplots. In this study, 42% of the students realized that, on a histogram describing distribution of salaries, *salary* goes on the *x-axis* and *(relative) frequency of people* on the *y-axis*, and that the distribution should be right-skewed, since *"most would make around the same but a few would make lots more."* Nonetheless, the rest of the students, similarly to the students of the previous semester, saw the graph as a scatterplot of *salary* vs. *age*. One of those students was Peter. In the follow-up interview, he confessed that he still *"struggle*[d] *with histograms"*. Although with my help he did realize that his reasoning was wrong, he was still not very optimistic about his ability to construct or interpret histograms: *"May be you give me another histogram today, I'd probably still mess it up."*

The "Test Results" Question from Pfannkuch and Brown (1996), which was given to students right after they had engaged in the "SATs and GPAs" activity was also included at the end-of-course assessment (Question 11). This was the question asking students to compare the boxplots of two sets of results on a test, given to the same small class twice, and decide whether the results changed significantly or not. The first time students answered this question, the proportion of them judging that test results changed significantly was only 25%. This proportion dropped further to 9% at the end-of-course assessment. All but three students recognized that the almost identical interquartile range of the two boxplots meant the change in test results was probably not very significant. At the same time, we again found several students who thought that the middle line of the boxplot shows the mean. In the follow-up interview of the primary group,

two of the eight students interviewed had serious misunderstandings about boxplots. They made claims such as "*25th percentile means that 25% in the class scored 43"*, *"most of the scores are in the middle box"*, or *"Q1 is the average of the lower part."* The other students gave a good description of the features of a boxplot and explained that the difference in the median scores would have been more important if the interquartile range had been narrower, because then *"the percentiles, whatever, would be a lot closer together. So if you just move it a little bit, you know, a quarter of an inch, that would be like a difference of 20*%."

Data Production

Question 2 at the end-of course assessment, adapted from Jacobs (1997), asked students to compare the quality of two surveys, both conducted to determine how many higher institutions in Texas are recycling. In the first survey, postcards were sent to all the deans of higher level institutions in Texas and about half of them responded, 91% of those responding stating that their school was recycling. The second survey used a medium sample size and a random sampling method, and found the proportion of schools recycling to be 37%. This question had also been included in the pilot study. The same coding used for categorizing students' responses in that study was also employed here.

Table 5.4 – Results of Current Study vs. Results of Pilot Investigation on "Recycling" Question

Response	**Non-Statistics %**	**Non-PACE %**	**PACE-Previous %**	**PACE-Current %**
A: Second method better; other method biased	67	72	68	79
B: First method representative of more schools and thus a better indicator	20	14	27	6
C: First method good because of larger sample; second because more random	3	7	0	6
D: First better even though biased because of bigger sample size	0	1	2	3
E: Other responses	9	5	2	3
F: No responses	1	1	0	3

Almost four-fifths of the students in the current study (79%) expressed preference for the random sampling method and only two students (6%) chose the first method without mentioning the selection bias characterizing it. This is unlike the pre-assessment, where several students did not recognize the dangers of self-selection. Analysis of variance was performed using the CATMOD procedure, with the binary variable *ResponseType* (1 = giving an "A" type response, 2 = not giving an "A" type response) being the response variable, and the categorical variable *CourseType* (1= Non-Statistics, 2=Non-PACE, 3=PACE-Previous, 4=PACE-Current), being the explanatory variable. It did not reveal any significant effect of course type on probability of a correct response (p-value = 0.5729).

The response of one student who, although acknowledging the potential for selection bias of the first method, still preferred it, is interesting: "*Large sample with 50% response – If we assume those not responding do not recycle, you really have a population statistic at 45.5% recycling. The smaller random*

sample only yielded 37% - I would go with the larger method." This student assumed that everybody who did not respond was not recycling and adjusted the results of the study accordingly. With this adjustment, he concluded, it is better to use the results of the first method, since it utilizes a larger sample size.

Concept of Independence

By the end of the course, students' notions of independence were much closer to the statistical one than when the concept was first introduced. In the end-of-course assessment, Question 5 asked students what "independent events" means, and all but two students (who confused independent with mutually exclusive) gave a fairly good explanation. Also, students' performance on the "Roulette Wheel" question at the end of the course (Question 15), compared to their performance on the same question at the beginning of the course, shows that the course did contribute a lot towards improving their notion of independence:

Table 5.5 – End-of-Course vs. Pre-assessment Results on "Roulette Wheel" Question

Response	**Pre-assessment No. of students**	**Pre-assessment %**	**End-of-course No. of students**	**End-of-course %**
Black	20	67%	3	9%
Red	1	3%	0	0%
Either	6	20%	30	91%
Other	3	10%	0	0%

Students' improvement was impressive. The proportion stating that black and red are equally likely to come up on the next landing rose from 20% to 91%. Also, students' comments show they did grasp the idea that previous outcomes do not affect the probability of the next outcome of random events: "*Because it is*

balanced. Each time is a fresh start"; "There is an equal chance for each color. It could land on red 30 times in a row, the probability remains the same."

Despite the improvement, several students confused independence of a single event with long-term frequency of random events, as the class performance on Question 3 of the end-of-course assessment, adapted from delMas and Garfield (1990) indicates. In this question, students had to decide who, among two friends, is more likely to get 80% or more heads, Shelly who is going to flip a coin 50 times, or Diane who is going to flip the coin 10 times. They had four responses to choose from.

Table 5.6 – Results of Current Study vs. Results of Pilot Investigation on "Shelly vs. Diane" Question

Response	**Non-Statistics %**	**Non-PACE %**	**PACE-Previous %**	**PACE-Current %**
A: Diane *because the more you flip the closer you get to 50%.*	21.7	24.5	36.4	39
B: *Shelly because the greater the sample size, the greater the variability in results.*	17.4	21.6	6.8	9
C: *Neither, because each coin flip is a separate event and the probability of heads is not affected by the number of times flipped.*	53.6	47.5	56.8	52
D: Other.	7.2	6.5	0.0	0

Comparing the performance of PACE students that participated in the current study on this item to that of PACE students in the pilot study, we see that they are quite similar. In both studies, there was a higher proportion of PACE students giving the right response A. Analysis of variance using CATMOD was carried out to examine the effect of course type on the probability of choosing the correct response C. The independent variable *CourseType* was modeled to have

two categories. The first category *Allother* was made up of all the students in the previous study who did not come from the PACE course (i.e. non-Statistics, non-PACE), and the second category *PaceAll* was composed of PACE students (PACE-Previous and PACE-Current). The analysis of variance suggests that *CourseType* had a significant effect on the probability of choosing the right response (p-value = 0.0242).

There was also in this study a very low proportion of students choosing response B, which claimed that the bigger the sample size the greater the variability in results. Nonetheless, in both studies, there was a high proportion of PACE students choosing C, which stated that since each coin flip is a separate event, the probability of heads is not affected by the number of times flipped.

In the follow-up interview, I asked students who had chosen response C to look at the question again. They all recognized that Diane actually has a higher chance of getting 40% heads because of the smaller number of tosses. Lucas recalled a related activity we had done together using Fathom: *"Oh, just like we did it on the computer with Fathom, we did it with 10 tosses, but when we did it many times and we kept on pressing Control-Y, Control-Y, it showed us that it would become more symmetric I guess, 50-50 chance."* Peter remarked that he should have chosen A, *"because the larger the n the better, and 50 is pretty large, so the proportion should come close to 50%"*, but he chose C *"because* [he] *saw independent and it sounded good."* Similarly, Zoe said she chose C because *"*[she] *was thinking that it's independent and all that staff."*

Sampling Variation vs. Sampling Representativeness

Several tasks were given to students at the end of the course in order to investigate their understanding of the relationship between sampling variation and sampling representativeness. I will discuss only a few representative ones here**.** One such question (Question 12), was also given to students prior to introduction to probability, and was taken from Shaughnessy et al. (1999). It was one of three versions of a task used in a series of exploratory studies on student understanding of variation with groups of primary and secondary students from the US and Australia. A total of 235 primary students (grades 4 to 6) and 89 secondary students (grades 9 and 12) had participated in that study. This version, which the authors called the CHOICE version, was given to a total of 105 students. Students had to choose, among five possible lists, the one that is most likely to present the number of reds drawn by four students who each drew 10 candies out of a bowl of 100 wrapped candies that had 50 reds.

In analyzing student responses, the same procedure as that of Shaughnessy et al. (1999) was followed. Responses were scaled both on the basis of their use of centers and of their use of spreads. For the "centering" scale, student responses were categorized as LOW, FIVE or HIGH. Responses for which the mean number x of reds was $4 < x < 6$, were classified as FIVE, otherwise they were classified as either LOW or HIGH. For the spread scale, the following categories were used: NARROW, REASONABLE, and WIDE. Responses in which the range was 7 or more are pretty unlikely to occur and were classified as WIDE, and so are those with ranges less than or equal to 1, which were classified as

NARROW. Ranges between 2 and 7 were considered REASONABLE. According to the scale, responses can be classified as follows:

Table 5.7 – Classification of Responses in Shaughnessy's "Candies from Bowl" Question

Response	Center Classification	Spread Classification
A: {8,9,7,10,9}	HIGH	REASONABLE
B: {3,7,5,8,5}	FIVE	REASONABLE
C: {5,5,5,5,5}	FIVE	NARROW
D:{2,4,3,4,3}	LOW	REASONABLE
E: {3,0,9,2,8}	FIVE	WIDE

The best response is therefore B, which is centered on 5 and is also a reasonable response in terms of spread.

The following table compares the performance of students in the Shaughnessy et al. (1999) study, with that of students in this study:

Table 5.8 – Results of Current Study vs. Results of Pilot Investigation on "Candies from Bowl" Question

Classification	Shaughnessy et al. Study %	Pre-assessment %	End-of-Course %
Center			
Low	13	10	0
Five	56	87	100
High	27	3	0
Unclear	4	0	0
Spread			
Narrow	16	19	12
Reasonable	76	81	88
Wide	4	0	0
Unclear	4	0	0
Correct			
Five, Reasonable	35	68	88

Students in the current study did better in estimating both center and spread. Instruction seems to have been particularly effective in helping them take

both spread and center into account. Whereas in the Shaughnessy et al. (1999) study, only 35% of the students belonged to the FIVE, REASONABLE category (i.e. chose response B), in this study the percentage was 68% in the pre-assessment and 88% at the end-of-course assessment. Linear model analysis using the CATMOD procedure was performed to investigate the relationship between probability of choosing the FIVE, REASONABLE category and course type (1= from Shaughnessy et al. study, 2= PACE-Current). The analysis of variance indicates a significant effect ($p<0.0001$) of course type on the probability of success.

Since helping students move away from "uni-dimensional" thinking and be able to integrate center and variability into their analyses and predictions, should be one of the main goals of statistics instruction, the results of this question are encouraging. It is an important accomplishment of instruction given that in the Shaughnessy et al. (1999) study, although most students' measures of spread were reasonable, they predicted values that were either too high or too low on the centering scale. Also, in that study, the use of words explicitly referring to variation was quite rare. In contrast, students in this study gave explanations that indicated they were integrating ideas of spread and center:

> Because 50% of the candies are red, the handfuls should be close to 5 reds each time so B. Not C because it's random, there is a margin of error.
>
> Because they all range around 5 per pick, as would a sample with 50% reds. The others seem too far away or impossible, like C.
>
> Because the average that would be expected should be 5 with some variation above and below the expected value.

> It’s unlikely with .50 probability of reds that anyone got 0 or 10 or straight 5's. There are .50 reds and so we would expect to see more of those but this is a random sample and thus there should be some variability. We expect to have some below 5 and some above. B shows that.
>
> It's all about variance, but "central-tendency" must always be counted.

Of course, students in Shaughnessy’s study are primary and high school students, whereas the present study deals with college students taking a statistics course. However, in that study, while a steady growth across grade levels on the “centering” criterion from 34% at Grade 4 to 83% at Grade 12 was observed, there was “an apparent oscillation on the variability criterion across grade levels.” The researchers noted “a high spike occurring in our Grade 9 students, and then a drop off at Grade 12, for both the REASONABLE and the FIVE, REASONABLE categories.” They speculated that the steady growth in the FIVE category, is an indication of the considerable focus of school curricula on “center”. A possible explanation they saw for the oscillation at Grades 9 and 12 is that Grade 9 students participating in their study were spending more time on data analysis than the higher level mathematics students, whose school work on probabilities might have interfered with their reasoning about this problem. The exposure to probability did not seem to interfere with the reasoning of the students in the current study. There were, of course, four students (12%) who at the end of the course chose response C [“5,5,5,5,5”], but all of them (as well as two additional students) had given the same answer in the pre-assessment, which was taken by students before formal introduction to probability. The reasons that these four students gave to justify their choice of response C involve, similarly to Shaughnessy’s study, misapplication of probability arguments. Students seemed

to be calculating the probability for a particular outcome rather than predicting the likely range for the number of reds.

Students' performance on Question 8 at the end-of course assessment, compared to their performance on the same question prior to instruction, is another example of the positive effect of instruction in helping improve students' probabilistic reasoning. Sixty-four percent of the students at the end of the course, compared to 35% of them in the pre-assessment, realized that due to the independence of random samples, one should still expect that, out of the next 20 students interviewed, about half should be men and half women (choice C). Of course, there was still a considerable proportion of students (24%) employing the balancing strategy and arguing that they expected the opposite trend to start happening (Response B), but in general, students' performance was much improved. Also, comparing results with the pilot investigation in which this task was also included, we see that students in the current study did better.

Table 5.9 – Results of Current Study vs. Results of Pilot Investigation on "College Interviewer" Question

Response	Non- Statistics %	Non-PACE %	PACE – Previous %	PACE-Current %
A	35	30	30	6
B	26	22	9	24
C	33	40	52	64
D	6	9	9	6
E	0	0	0	0

A significantly higher proportion of students in the current study gave the correct response C. Analysis of variance using CATMOD was carried out to examine the effect on course type on the probability of choosing the correct

response C. Course type was modeled to have two categories. The first one was *Allprev*, which included all the students in the previous study (i.e. non-Statistics, non-PACE, and PACE-Previous), and the second one was *PaceCurr*, and included students in the current study (PACE-Current). The analysis of variance indicated that course type had a significant effect on the probability of a correct answer (p-value = 0.0095).

Students' increased awareness of sampling variability is also seen if we compare their performance on the M&M problem (Shaughnessy, 1999) given to them at the end of the course (Question 9), with the performance of students in the pilot study:

Table 5.10 – Results of Current Study vs. Results of Pilot Investigation on "M&M" Question

Response	**Non-Statistics %**	**Non-PACE %**	**PACE-Previous %**	**PACE-Current %**
A: Exactly 8	11.6	15.1	4.5	3
B: 0-8	10.1	15.1	15.9	9
C: 8-20	2.9	2.9	4.5	0
D: 6-10	52.2	52.5	63.6	58
E: 0-20	20.3	14.4	11.4	27
F: Other	2.9	0	0	3

Only one student (3%) in the current study, compared to several Non-PACE and Non-Statistics students in the previous study, stated that they expect exactly 8 browns. Similarly to all groups of students in the pilot study, the majority (58%) of students in the current study chose 6-10 (response D) as the most likely range. Students choosing D explained that, since 40% is the population proportion of browns, what is most likely to happen is to get approximately 40% brown: "*Sometimes there will be a couple more, other times a*

couple less"; *"There will be variation but probably not too severe."* Some students even noted that 6-10 *"would be the likely range"*, but this does not mean all sample outcomes would fall in that range. On the other hand, that a proportion of students as high as 27% chose E ["0-20"], might be the adverse effect of a course putting so much emphasis on sampling variability. Some of the students' comments hint that this might be the case: *"Although 40% of 20 is 8, you wouldn't expect to get 8 every time and the possibility is between 0-20 with a size so small."* The student who chose F wrote something along the same lines of the students who chose E, emphasizing the small number of M&Ms sampled*: "Take and do many more. This is too small."*

In the follow-up interview, I reminded students of the *"New Zealand Question"* in the pre-interview (Appendix A) and asked whether their reasoning had changed in any way. Most of the students did change their mindset about this situation and pointed out that the number of children is so small that one cannot draw any conclusions. Tim, for example, who had argued in the pre-interview that the probability of giving birth to a child with a missing limb correlates with where one lives, now said: *"There is not enough information to…it's not a big enough…it's only 7 people. It's not enough number of subjects to understand what's going on."* Lucas remembered that in the pre-interview he was thinking *"there might be something wrong with the sanitation or the water, something like that."* Now though, he realized that *"this is only one year so, last year or the year before, they could have had 3 down here and 2 up there and 2 over here.*

You have to look at many years to see what's happening." He added: *"That's why I liked this class. I learned to look at the big picture of things."*

Despite the students' overall increased awareness of sampling variation, not everybody was as apt to let go of the deterministic reasoning with which they had approached this question during the pre-interview. Zoe's first reaction was: *"I don't know, most of the population lives on this side, so…I don't know."* However, when I noted the very small number of children, she realized that the differences might have just occurred by chance. When next I reminded her of the "Roll of the Die" Question (Appendix A) and explained its analogy to the "New Zealand" problem, she remarked that it is much easier for her to think in terms of chance about dice than about real life problems. Similarly, Andrew at first argued: *"There is a higher probability to get a limb missing here cause it's not a regular type of civilization where there is doctors' offices"*, but then recognized that the number of births is very small to draw conclusions based on a single year.

Inferential Statistics

Several tasks given at the end of the course assessed student understanding of sampling distributions. Most students' performance on those tasks indicated poor understanding of this so important, but yet so difficult concept. I just give the following task (Question 7 of end-of-course assessment) as an illustration of this.

> The amount of time it takes to take an exam has a skewed-to-left distribution with a mean of 65 and a standard deviation of 8 minutes. A sample of 64 students will be selected at random.

PART I: Which of the following describes the distribution of the amount of time it takes to take an exam? (a) N (65, 8); (b) N (65, 1); (c) a skewed distribution with a mean of 65 minutes, but unknown variance; (d) a skewed distribution with a mean of 65 minutes and a standard deviation of 8 minutes. Explain your reason.
PART B: Which of the following describes the sampling distribution of the sample mean based on n=64? (a) Approximately N (65, 8); (b) approximately N(65, 1); (c) approximately N(1, 65); (d) a skewed distribution with a mean of 65 minutes and a standard deviation of 1 minute. Explain your reason.

Only 64% of the students answered the first part correctly, although the answer was explicitly given in the definition of the problem. The rest confused population distribution with sampling distribution. Thirty-percent chose A (a *N(65,8)* distribution) and most of them gave a justification along the same lines as Matthew who wrote: *"Sample size is large enough to force Normal distributions by Central Limit Theorem."* These students were confusing population distribution with sampling distribution. Also, they were confusing standard error with standard deviation since, given that they had thought the question was asking for the sampling distribution of the mean for samples of size 64, they should have chosen the *N(65,1)* distribution. Barely more than half the students (54%) answered the second part correctly by choosing B. Some of these students, did so as a result of applying the Central Limit Theorem: *"Since n=64, the sample variance is 1 because: N (65, 1) by CLT. We know our sample is large enough for normal distribution."* However, not everyone choosing B did so for the right reasons as I found out in the follow-up interview:

> Int.: Over there it was the distribution of the amount of time it takes. What about this one?

Andrew: This is the distribution of the sample means?

Int.: Yeah. What does that mean?

Andrew: Well, just how the…I don't…I don't …I don't really know what that means…I understand why I picked that answer, but…

Int.: Why did you pick that?

Andrew: I just kind of guessed... Whatever it was I got lucky in the sense that, I just thought it is like 65 and 1 and you get 64, I don't know…I'm sure it probably has more to do with the numbers back here. I don't really know.

Similarly to Andrew, George's choice of the correct response was the result of meaningless manipulation of the information provided: *"I was just guessing on that one, because the mean is 65 and so I said (n-1) is 64. I knew it wasn't that one, because of the question before."* Tim's response to the second part of the question, also seems to be correct as a result of guessing and not of true understanding: *"I don't remember how I did this staff, but I…I guessed it actually. I guessed, that's pretty much what I did."*

The students I interviewed, also seemed not to realize that the reason so much emphasis is put on normal distribution (and consequently on standard deviation) is its connection to the sampling distribution of the mean. Although with some prompting most of them did point out that with increase in sample size, the sampling distribution *"will be closer together"* and, when the sample size is large, it will be *"symmetric...normal"*, they had to be pointed out that this is what makes normal distribution so important. They did realize that the property of a distribution to be approximately normal is a very useful one, because then *"if you know the standard deviation you can find out if you scored this much where it will*

be on the graph." They also knew how to apply the empirical rule and understood the usefulness of the Z-scores*: "You can standardize and then you can compare ranges that are different, distributions with different means and standard deviations you can compare."* Nonetheless, the connection of the properties of normal distributions to the implications of the Central Limit Theorem was missing.

The following question ("Nicotine Level") was given to students just before formal instruction on inferential statistics began:

> FDA has a maximum upper limit for nicotine contents to be 12 mg. A company is manufacturing a new brand of cigarettes. FDA sent an evaluator to test the nicotine content:
>
> (a) The evaluator took a random sample of 10 cigarettes and found the mean nicotine content to be 13 mg with a standard deviation of 2 mg. Based on this sample of 10, do you think the FDA should conclude that the average nicotine level is not acceptable (is significantly higher than the acceptable brand)? Why or why not?
>
> (b) What about if the evaluator takes a sample of 100 and again finds the mean to be 13 mg? Why or why not?
>
> (c) The company filed a complaint that, based on their test, the mean nicotine level is 11.8. Is it possible that the FDA has made a mistake: (i) When basing their decision on a sample of size 10? (ii) When basing their decision on a sample of size 100? Explain why or why not.

In the pre-assessment, only six people (19%) thought FDA could conclude that the average nicotine level is unacceptable based on a sample of size 10, but 66% of them argued that based on a sample of size 100 the evidence is strong enough to draw this conclusion. Seven students (22%) did not think FDA should

conclude that the nicotine level is above acceptable limits even with a sample of size 100, because this is still too small of a sample. In the third part, all but one student thought it is possible that FDA had made a mistake when basing their decision on a sample of size 10, since it is such "*an extremely small sample*" that *"even one cigarette with higher levels will throw results off."* In the last part, however, there were nine students (28%) who argued that, when basing the decision on the sample of size 100, the sample is large enough to eliminate the possibility of reaching a wrong conclusion.

Out of the eight students interviewed at the end of the course, four of them belonged to the group of students who had worked with me in the outside-of-class activities using Fathom. Three of these four students agreed to meet in the computer lab and work together on the "Nicotine Level" task. This meeting, which took place the day before I interviewed students, was video-taped, and will be described later on. The four "Non-Fathom" students, as well Lucas who was a "Fathom Student" but did not make it to the lab meeting, were asked to work on this task during their interview.

Non-Fathom Students

Despite all the work done on sampling distributions and hypothesis testing, these students' responses were almost identical to the ones they had given before their formal introduction to inferential statistics. When, for example, I asked Andrew how he could justify quantitatively his claim that based on a sample size of 100 cigarettes FDA would have enough evidence to conclude that the nicotine level is above acceptable limits, but not based on a sample of 10

cigarettes, he had no answer. I had to point to him that this is a situation where we could use a hypothesis test. Although he did agree that it is "*because it could be either yes or no*", he had a hard time formulating the test, and had to be reminded to use the standard error instead of the population standard deviation**.** He did note that *"they want to see if the level is 12, so they're going to take a sample of cigarettes to see if they fall within the legal limits of 12"*, but claimed that if they reject the null hypothesis they will conclude *"that the mean is 13."* Similarly to Andrew, George also claimed that *"what we are trying to see here is whether the mean nicotine level is 13."* He also had no answer when I asked him whether there is a way, with the things that they had learned in the class, to check how likely it is to get a sample mean of 13 by chance. Once, however, I reminded him about using Z-scores, he went ahead and did the calculations which show that we would reject the null hypothesis for a sample of size 100 but not for a sample of size 10, because "*13 is not rare enough to…put the two standard deviations."*

Fathom Students

Lab meeting

Four students participated in the activity (In addition to Anna, Tiffany, and Zoe who belonged to the primary group, Lucia, who had also been working with Fathom, participated). The four students worked as a single group. I told them that they could answer this whichever way they wanted, but that I would not offer them any help other than technical assistance. Students started discussing the problem:

Lucia: So guys, would you like to name it nicotine drug or what…?

Anna: We're gonna make like a bunch of cases for the 10 cigarettes. Because we have a sample size of 10 cigarettes.

Zoe: But you have to do it more than once to…

Lucia: Right.

Lucia: Go to New Cases/Data and make 10 cases.

Zoe: Now you have to give it a mean of 13 and do it you know, 1000 times.

Anna: The mean for this is 13.

Zoe: I think you just click on the formula box.

Lucia: Mean of 13 and a standard deviation of 2.

It was very obvious they were not giving much thought yet into what they were doing. Then, however, they decided to read the problem again. Zoe exclaimed:

Zoe: Oh, yeah, we don't want 13, we want 12.

Lucia: The maximum acceptable limit is 12. "Based on the sample of size 10, can you conclude…" So, we have to see if…

Zoe: If 13 is out of range.

Anna: Oh, yeah, yeah, we want to see if 13…

Lucia: If it's rare. So…we want to see if…if 13 is over the upper limit.

They started working on the simulation. The following discussion is an interesting one as it relates to students' understanding of the role of variation in hypothesis testing:

Lucia: So we put Normal Random here, and put 12.

Zoe: There is no standard deviation.

Lucia: Oh, you're right. Just 12. We don't put the standard deviation?

Tiffany: Should we use the sample standard deviation of 2 if we don't have the actual?

Lucia: If you don't have the population, then you use the sample.

Anna: OK, let's do it.

Lucia: How are you going to know upper and lower limits if you don't have the standard deviation? We've got to have standard deviation. How can you know? What do you guys think?

Anna: I'm mad Maria is not telling us what to do (laughs).

The following conversation, where students explain that what they are doing here is simulating a sample from the population they have defined and that this sample will change each time they repeat the simulation process, shows how the activity facilitated their understanding of sampling:

Lucia: OK. We are going to make a histogram of this, so that we see the distribution, so…You drop nicotine here and then you go to the dotplot and you choose Histogram. It shows you the distribution of the population.

Zoe: Of the sample.

Lucia: Of the sample.

Anna: Now we are going to show why…

Lucia: Not of the sample distribution, of the sample population.

Anna: Do Control-Y.

Lucia: Control-Y changes values.

Zoe: Because it's random chance. It gets a different sample each time.

They then went to the Measures Collection in order to build the distribution of sample means from samples of size 10, and test whether a sample mean of 13 indicates that the nicotine level is above acceptable limits:

Zoe: Let's do a collection of measures.

Lucia: Is that how you save them?

Zoe: Yeah, go to collect measures.

Lucia: No, you've got to do this first (goes to Measures Collection), because you've got to tell it what measures to collect. We are doing the mean drug of…

They gave a command to the computer to collect "*500 measures of size 10"*, made *"a table of measures"*, and drew a histogram of the collected statistics. They then collected 500 sample means of size 100. While the computer was collecting the measures, Lucia remarked:

Lucia: So, we know that for sure, the standard deviation for 100 is going to be smaller because we have a 100 instead of 10 – divided by a bigger number.

Tiffany: Yeah, by $\sqrt{100}$.

Anna: I think we are done.

Lucia: We'll close this and then we'll do a histogram of this.

Zoe: See how it is more like the normal thing?

Lucia: So, this is Collection 1 from the population, and that's the sampling distribution from the population.

Tiffany: OK.

They then turned to the handout and tried to answer by hand the first part of the question that asked whether a sample mean of 13 for a sample of 10 cigarettes indicates that the nicotine level is above the acceptable limits:

Zoe: "Do you think the nicotine level…?"

Lucia: We don't know yet, until we…

Anna: Oh yeah.

Lucia: Too small of a sample.

Zoe: Too small.

Lucia: We have to justify our answer. Is it acceptable?

Lucia: (She writes on paper) So, it's a random sample of 10 from a population of mean 10 and standard deviation of 2. So, do we guys want to standardize this one (shows histogram) to see…do we have to standardize? If I'm crazy just let me know.

Anna: After yesterday I'm just drained.

Zoe: Yeah, I wish we had this before. The final just drained my brain.

Lucia: You know, 13-12 over the standard deviation…I always screw that up…because it's $\bar{x} - \mu$ …

Tiffany: Divided by $\sigma/\sqrt{n}$.

Lucia: So, I always forget, $\bar{x}$, if it is the 12 or the 13. Which one is it?

Tiffany: $\bar{x}$ is the 13.

Lucia: You're right. So we standardized it. So, we'll use the table.

Zoe: So what is it, 1.58?

Lucia: But do we use *Z*- or *t*- because it's a small sample size?

Zoe: We would use t.

Anna: *Z.*

Lucia: *Z.* She's right. You go to 1.58 and look in there.

They found *P (Z>1.58)* and decided it was too high, so they did not reject the null hypothesis. They compared results they got when they did the calculations to the ones they got by doing the computer simulation and concluded that they were very similar. They did the same for samples of size 100, and concluded that this time the evidence was enough to reject the null hypothesis.

Interviews of individuals

In addition to the lab activity, my conversation during the follow-up interviews with those students who had been using Fathom to explore sampling distributions made it clear that, unlike students like George and Andrew, these students had a much better grasp of the ideas of sampling distribution and hypothesis testing. Anna, for instance, gave a very good explanation of sampling distribution, although she admitted that when the concept was first introduced she was completely lost:

> Anna: I don't think I got it at first. I was confused. I understood that the larger the sample the closer you get to μ. But I didn't understand how from the population graph you get to the sample mean graph…When we started talking about sampling distribution, I would just be "I have no clue."

Anna then commented that although with ideas like "*the probability stuff he* (the instructor) *did a really good job*", it was her experience with Fathom that helped her grasp the concepts introduced during the last part of the course. Zoe too said that she would not have understood what sampling distribution means if

it had not been for the Fathom activities, "*if* [she] *didn't really see it.*" For the "Nicotine level" question, she set up the hypotheses, did the calculations, and then gave a good explanation of the purpose of doing the hypothesis test.

The following excerpt from my interview with Lucas (the only "Fathom Student" who did not participate in the lab activity) is, in my opinion, evidence of how powerful Fathom can be as an instructional tool. Although Lucas had one of the lowest grades in the course, and had a very hard time understanding abstract notation, his conceptual understanding of the notion of sampling distribution was better than that of most of the students in the class:

> Int.: Can you tell me how we would do it on the computer?
>
> Lucas: You want samples of size 10.
>
> Int.: That's right. What would the mean of the distribution be?
>
> Lucas: That should be 12, right?
>
> Int.: OK. And what do you want to see?
>
> Lucas: What are the chances of getting 13, if 13 is extremely high.
>
> Int.: So, you do that. Then what do you do?
>
> Lucas: You keep on going. You look at like 100 or 1000 sample means, how many you want to look at.
>
> Int.: Thousands of what?
>
> Lucas: Thousands of samples of 10.
>
> Int.: And what do you do for each of those samples?
>
> Lucas: You find the mean of each of those and then you put those on the histogram.

Int.: When you look at histogram what do you try to see?

Lucas: If 13 was an extreme high.

DISCUSSION

Findings from the study suggest that the course has been quite successful in helping improve students' statistical reasoning. Students in this study recognized that in addition to knowing about the center of a distribution, one also needs information about its spread. They acknowledged that, when comparing measures of center, one should always take spread into account. Most of them had good understanding of the meaning and purpose of the different numerical summaries they had learned in class. It was for example very impressive that, in contrast to our previous research findings where almost no student really understood what standard deviation means, most of the students in this study had a pretty good grasp of the meaning and use of standard deviation. Also, they all knew that, in addition to standard deviation, measures such as the interquartile range also give us information too about the spread of a dataset. They understood that mean and standard deviation are not the only two measures that define the shape of the distribution. And although they still did have some difficulties with constructing and interpreting graphs, their understanding was much more sophisticated than that of students in the previous studies we had conducted.

Instruction proved quite effective in achieving one of its main goals – helping students move away from "uni-dimensional" thinking and integrate center and variation into their analyses and predictions. Although not totally letting go of their deterministic mindset, students were much more willing to interpret

situations using a combination of stochastic and deterministic reasoning. The course increased significantly their awareness of sampling variation and its effects (in some cases, such as the M&M problem to a degree higher than the desirable one).

Despite the positive effects of instruction on students' skills and dispositions, instruction did not succeed in helping most students adequately develop the important ideas related to inferential statistics. Inference is a very important part of statistical reasoning and for this reason the next section deals exclusively with this topic.

Student Understanding of Inferential Statistics

The experience I gained from the course has led me to agree with the instructor who thinks that the idea of rare event linked to everything, from the beginning to the end of the course is very promising in helping improve student understanding of inferential statistics. I think that the idea of a rare event was quite intuitive for students. It helped them see the purpose and usefulness of standard deviation and of the *z-scores*. It was also quite effective in helping students make connections between exploratory and inferential statistics. For example, when confidence intervals were first introduced, and the instructor asked students to give a likely interval for the number of raisins in the box, students used the empirical rule in order to include in the likely interval those values that are not "rare events". Of course, they used the standard deviation of the population and not the standard error, but still they understood the logic behind confidence intervals. Also, when doing hypothesis testing, the idea of rare event

was much easier for students to grasp than the idea of a p-value. As the instructor pointed out, that students in this type of course always need some reference to experience, they *"have to have a concrete message to tell them about what's going on."* It is much easier for students to relate to their everyday experience an event that occurs rarely than it would have been if instruction had on hypothesis testing would have begun by giving students the definition of a p-value.

Despite the benefits of having approached inferential statistics using the intuitive idea of "rare event", difficulties of students persisted. Next, I make some conjectures for the reasons behind these difficulties.

Conjectures for Students' Difficulties

A. Meaning Attached to Variation

What I have found as being one of the problems is how students perceive variation. What meaning do they attach to population variation, to variation of a single sample, and to variation of the sample mean? Do they realize the relationships/differences between them?

Just before formal instruction of sampling distribution began, students were given the following task (from Garfield et. al, 1999):

> The distribution for a population of measurements is presented below. Suppose that 10 values are going to be sampled from this population and the sample mean calculated. Some possible means for this sample are 1, 6, 8, and 10.

PART A: Which of the four possible means is MOST likely to be calculated? (i) 1 (ii) 6 (iii) 8 (iv) 10. PART B: Which of the four sample means is LEAST likely to be calculated? (i) 1 (ii) 6 (iii) 8 (iv) 10. PART C: Looking at the graph above, what would you guess to be the value of μ, the population mean? PART D: Suppose now that in addition to the sample of 10 values, you take another sample of 100 values and you calculate its mean. (i) In Part B, you stated the sample mean out of 1, 6, 8, and 10 that you believe is LEAST likely to be calculated when you sample 10 values. Will it still be the same when you sample 100 values and you calculate their mean? Why or why not? (ii) When is there more chance to get this LEAST likely mean value, when you sample 10 or when you sample 100 values? Explain why.

Figure 5.16 – Pre-assessment Task on Sampling Distributions

All students did well on the first three parts of the question. They all gave 1 as the mean least likely to be calculated when drawing a sample of 10 values from this population. All their guesses for the value of the population mean were also reasonable, ranging between 7-8, with most of them being around 8. Their responses however to Part D of the question were quite varied, indicating different notions of sampling variation for different students.

Although 78% of the students did agree that when taking a sample of 100, the least likely mean value they chose in Part B would still hold, there were seven students (22%) who disagreed, arguing that variation goes up with increase in sample size: *"With a larger sample there is more chance for extreme scores."* Using the same reasoning, these students also argued that it is more likely to get the least likely mean of 1 with a sample of size 100: *"there is more variation…more values and therefore more chance of people doing badly."*

These students' notion of variation in this context seems to be that of range, which indeed usually goes up with increase in sample size. It is a different notion from that of several students who claimed that a sample of size 100 is less likely to give you an extreme estimate "*because when you have 100,* [there is] *more chance for variation and more spread out distribution."* For these students, variation had the connotation of sample representativeness, which increases with increase in sample size because "*there is more to choose from."*

The different meanings students attach to variation indeed proved to be one source of the difficulties students had with comprehending sampling distributions, as the analysis of their written responses to the group activity on sampling distributions they did in class in order to estimate the average SOS score for CMU faculty indicates. Another source of difficulty that the analysis revealed was confusion between variation of individual observations and variation of sample means.

In the first set of questions students had to choose among three sampling schemes the one that has a higher chance to give a sample mean closer to the true average SOS score: (1) the mean of a sample of 2 faculty scores; (2) the mean of a sample of 5 faculty SOS scores; or (3) the mean of sample of 20 faculty SOS scores. All of the students recognized that the sampling scheme with the largest sample size is more likely to give a sample mean closer to the population mean. Several students justified this by arguing that a larger sample size means there is "*smaller variation and so it's closer to the truth"*, while others argued that a larger sample size implies higher variation. The responses of several of the

students who argued that increase in sample size leads to decrease in variation, suggest they referred to the distribution of an individual sample and not to the distribution of sample means: *"As sample size gets larger sample variation gets smaller."* Among those who believed that increase in sample size leads to increase in variation, there were two groups of students. The first group were students who perceived variation as sample representativeness. The second group were students who perceived it as the range of values in the sample: *"More possibility of wider range but* $\bar{x}$ *should be more accurate."* According to the meaning that both of these groups attached to variation, an increase in sample size would usually indeed lead to increase in variation.

These beliefs regarding variation of individual samples, affects how students perceive the relation between sample size and variation of sampling distribution. When asked to describe this relation, ten students (36%) argued that *"the larger the sample usually the larger the variation."* They were the same students who had previously claimed that increase in sample size leads to increase in the variation of a sample. The arguments they made were almost identical to the ones they had made before, this suggesting that they did not understand the changes involved when moving from the distribution of individual values to the distribution of sample means. This confusion between individual observations and sample means was also hinted in the responses of some of the eighteen (64%) students who argued that variation of sampling distribution decreases with increase in sample size: *"5 points are more likely to cluster in than only two points whose variation is the distance between the two points."*

As part of the activity, students had to look at the graph of the standard deviation of the distribution of sample means as a function of sample size and predict the trend if we continued to increase sample size.

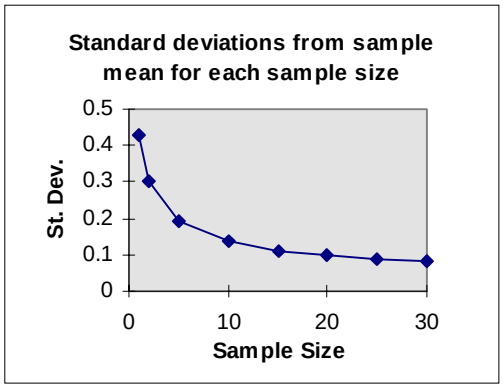


Figure 5.17 – Graph of Standard Deviation away from Sample Mean against Sample Size

The group of students I video-taped made the important observation that eventually, *"sample mean will form a straight line"*, and standard deviation will be zero. All the other students in the class also wrote that standard deviation would decrease with increase in sample size and that *"standard deviation of sample mean would be 0 when the sample size equals the population size."* Nonetheless, in a subsequent question, which was asking students to explain the relation between the standard deviation of sample means and that of the original population, we witness their confusion between population standard deviation, sample standard deviation, and standard error.

Only few students pointed out that *"S.D. of sample mean will become smaller as the average of averages size increases – S.D. of population is constant and set by population distribution."* Many students argued that, as the sample size increases, the standard deviation of the sampling distribution gets closer to the standard deviation of the original population. What several of these students seemed to have in mind was actually the standard deviation of a single sample, which they thought decreases with increase in sample size. Some of them even claimed that, as the sample size increases, the standard deviation of the sample will eventually reach the population standard deviation which is 0: *"As your standard deviation gets closer to zero, you know you are getting closer to the standard deviation of population which is 0."* On the other hand, there were a few students claiming that *"standard deviation of sample mean tends to be larger than that of original population because original population is the truth, and sample is an estimate, subject to error."* These students also seem to be referring not the distribution of sample means but the distribution of a single sample.

Students' confusion between population distribution, distribution of a single sample, and sampling distribution makes understanding of hypothesis testing more difficult. In the first question of the handout students worked on in class (Appendix C), which was asking them to use their knowledge and common sense to decide, given the information provided by the company, whether the drug is effective or not, everybody considered the population standard deviation instead of the standard error. The discussion of the group I video-taped, is indicative of this confusion between standard deviation and standard error:

Um: 3 deviations away is lower than 0.

Kristin: But isn't everything that is beyond 2 standard deviations rare?

Kristel: Right. So, no way it's rare…

Kristin: Don't we have like the 95%?

Kristel: The 95% confidence interval would be between 50 mg and 450 mg. This is not conceivable. A few might benefit, but not the population as a whole.

It was the hint provided in the second question that alerted this group as well as the rest of the students to the fact that *"this is not talking individually, but about the average."* They did the calculations using now the standard error and concluded that a reduction in the mean cholesterol level to 250 by chance when 285 is the true average is rare *"cause 250 to 285 is about 2.8 standard deviations. So we are 49.77% away from the mean. 2 standard deviations is already 47.5% away."* However, they concluded that the drug is not effective because "*despite the significant reduction of the few, many do not benefit… It is rare to have such good benefits.*" Their conclusion suggests they go back into viewing the sampling distribution as representing cholesterol levels of individuals and not average cholesterol levels. Of course, it also suggests failure to realize that if the drug is effective in reducing cholesterol level, the sampling distribution of average cholesterol levels will no longer be centered at 285.

That different students attach different meanings to variability was also observed at the end-of-course interviews. Out of the eight students interviewed, four said that for them now, at the end of the course, variability means distance from the center: *"Variability is like the variance on the left and right from the*

center"; "I think of it as like a set mean, or median standard, and variability would be the amount, the distance away, higher or lower than the median or the mean, that would be variability, like the ranges and staff." Two other students thought of variability as the way in which the data is distributed: "*Variability is the way in which the data is distributed and if we say small variability then the data is close together, if we have large variability it's far apart."* Two others defined variability as the range of values of a distribution: "*High variability means there is a high range of values…like if the lowest score is 5 and the highest is 95, there is high variability…it's spread out."*

All of the students' definitions are reasonable and describe different aspects of variability; they are not however all appropriate in the context of sampling distributions. Instruction needs to put more emphasis on helping students realize that when dealing with sampling distributions, the notion of variation one should use is that of variation as distance from the center of the distribution. In the same way that being literate means being able to differentiate between different connotations of the same word depending on context**,** statistical literacy means the person is able to use the right meaning of the word variability depending on the situation.

B. Abstract Notation

Just before instruction on sampling distributions began, students were given Question 14 (Appendix D), taken from Garfield et al. (1999), which aimed at investigating whether they understood what the basic statistical symbols (σ, s, μ, $\bar{x}$) stand for. The same question was given to them at the end of the course.

Only 48% of the students in the pre-assessment and 64% of the students in the end-of-course assessment circled s and $\bar{x}$ as symbols they would be able to calculate if they took a sample of data from a certain population. Only 59% of the students in the pre-assessment and 70% of the students at the end-of course assessment were able to distinguish symbols representing parameters from those representing statistics. The proportion of students correctly circling the symbols that vary from sample to sample increased significantly from 39% to 73%. Despite the improvement, the fact that at the end of the course so many students were still confused about basic notation helps explain why understanding seemed to fall apart each time the instructor would use some mathematics symbols.

C. Statistics Language

The instructor considers statistics language to be *"a big headache"* and a main reason for students' difficulties in comprehending the logic of hypothesis testing:

> Inst.: Hypothesis testing you have to test if it's Ho or not…and you set up these…and you set up these…and then you need a rule…and then the rule and then… OK, the rule depends on this, depends on that, under this condition we do this, under that condition we do that, that's a very difficult approach.

Looking at the how students responded to Question 10 of the hypothesis testing activity (Appendix C), where they had to define *effective* and *not effective* in the context of the "New Drug" problem, suggests that the statistical notion of effective might be quite different from what students have in mind. Whereas in this context, effective would mean that the drug reduces the mean cholesterol of patients to a level less than 285 mg, most students defined effective quite

differently. For example, students in the group I video-taped had a small debate as to what they should consider effective. One student argued that the drug should drop the cholesterol level to normal levels (≤200 mg), but another one argued that if somebody's level is reduced from 400 to 250, this should be considered effective, although not being down to 200. They ended up deciding to define effective as *"dropping the cholesterol level of at least 90-98% of the population of patients by at least 10-20%, without side effects."* Looking at the responses of the rest of the students, I found that several of them specified levels to which the mean cholesterol level should drop for them to consider the drug effective (e.g. lowering the mean cholesterol level by at least 50 mg). Several others argued that the drug has to go through much more extensive testing before deemed effective, while others wrote that they wanted side-effects to be at a minimum. These students had a different notion of effective than a test of hypothesis would; they nonetheless brought up considerations which, though not addressed by a statistical test, play an important role in making the final decision of whether to allow a new drug to enter the market or not.

When the video-taped group read a subsequent question, they realized that the definition of effective given by their handout is different from theirs: *"According to what they are saying there, it's effective if it lowers it at all and we are saying that it's not effective unless it lowers it by 20%."* They decided that *"according to their definition of effective, then yeah it is effective"* and, although they "*still don't know the side effects*", they should re-evaluate their answers to the previous questions. Consequently, they decided to reject the null hypothesis,

since *"250 is rare compared to 285"*, and conclude that the drug is indeed effective in reducing the cholesterol level. Unlike this group, many other students in the class were not convinced by the hints provided in the handout, and still argued that the drug is not effective. Forty-eight percent concluded that the evidence is *"not strong enough to claim that the drug has significant effects"*, although almost all of them had, in the early part of the activity, found the reduction in the mean cholesterol level to be *"very significant and rare"*.

Students' responses to Question 14 of the handout on hypothesis testing, also point to the difficult time students have in comprehending the language of hypothesis testing. Only half of them identified correctly the two possible types of errors that can occur in decision-making for this study: *"(1) Legalizing a drug that is not effective (Type I); (2) Not legalizing an effective drug (Type II)."* The rest gave two cases that were actually the same type of error: *"(1) Approving a bad drug, (2) Probable use of a bad drug"; "(1) Could have side effects, (2) Could kill someone."* Analysis of students' responses also suggests that the court case scenario some statistics textbooks and instructors use as an example in order to point to students the seriousness of committing a Type I error, might not be convincing to everybody. Although the majority of the students (73%) did think that falsely convicting an innocent person is worse than letting free a guilty one, there were five students who argued that letting a guilty man go free is more serious, and one student who thought that *"both errors are critical"*.

Question 13 of the end-of-course assessment (Appendix D), asking whether it is true that *"A statistical test of hypotheses correctly carried out*

establishes the truth of one of the two hypotheses, either the null or the alternative one", is another example of poor performance because of difficulties in comprehending statistical language. Seventy-nine percent of the students agreed regarding the veracity of the statement. Out of the 7 students (21%) that responded with a "No" to the question, only two gave a satisfying response. The results were very similar to those obtained when this question was given to students in the pilot study. In that study, there was again an extremely high proportion of students (80% PACE, 66% non-PACE) who had agreed that hypothesis testing establishes the truth of one of the two hypotheses:

Table 5.11 – Results of Current Study vs. Results of Pilot Investigation on "Hypothesis Testing" Question

Response	**Non-PACE %**	**PACE-Previous %**	**PACE-Current %**
True	66	80	79
False	31	18	21
Unsure	3	2	0

Analysis of student responses to this question in the previous study had led me to conclude that many students did not understand the difference between mathematical proof and a statistical test. Responses such as "*hypothesis testing tests the validity of what you are testing and the results are evidence enough to support either the null or alternative"* were typical. Jimenes and Holmes (1994), drew similar conclusions when they gave this question to 436 students in seven different departments of a university in Spain. They divided student responses into six categories. Their modal category was also that of students whose

response they interpreted as a claim that the test procedure is a logical proof of one of the two hypotheses.

When analyzing the results of the current study, I found it hard to believe that such a high proportion of students did not recognize that, because of variation, a sample is almost never completely representative of its population and there is always the risk of drawing the wrong conclusions. Having followed these students' thinking so closely, I knew that they were all well aware of the effects of sampling variability. Looking more closely at students' responses, I realized that they might have not understood what the question was asking. This became evident in the follow-up interviews. When, for example, Tiffany re-read the problem she said: *"So, here we could have a Type I or a Type II error. I got it now. I was just thinking in the sense of, if you don't conclude the one you conclude the other."* Zoe said she had interpreted the question as "*asking if the hypothesis is either wrong or right"* and, though she knew that wrong conclusions are always possible *"because of the two Types of errors"*, she responded with *"Yes."* Peter said he *"did not like the way this problem was worded."* After I re-phrased the question, he responded: *"No, no. Always no…because there is always the chance of error."* He used an example the instructor had given in class (*Ho: Carl Lee does not give you A, Ha: Carl Lee gives you A*) to explain Type I error: *"Carl Lee gave you an A when you deserved it…that one there would not be an error. You got an A when you did not deserve it, that would be like a Type I error."*

Students' performance on this question is one of many cases in the study that made me realize that if we base assessment merely on written examinations and do not listen closely to students, we might end up concluding something entirely different from what they actually mean. Although I only report this example, I have observed in several tasks I gave students which I had taken from the research literature, a discrepancy between researchers' interpretations of student responses and the actual reason behind those responses. The fact that I used multiple-sources of assessment, allowed me to actually find out why a student made a certain choice, rather than simply speculate as to what might have led the student to this choice.

D. Use of Technology

Comparing the understanding of sampling distributions and the logic of inferential statistics by the group of students who explored these ideas using Fathom to that of the other students in the class, has led me to conclude that the choice of computer-based tools is a crucial decision. I see several advantages to using an object-oriented learning environment such as Fathom.

Fathom is a *dynamic learning environment.* All of its objects are continuously connected and thus selection of data in one representation means the same data is selected in all representations. Students can interact with the data and see the immediate impact that their actions will have on the different representations of the data on the screen. The heavy reliance on the *"drag and drop" interface*, the fact that students have to drag rather than choose things, is another advantage of Fathom that helps increase immediacy of data analysis

(Pratt, 1998). Using the *Formula Editor* also has many advantages and comprises a core Fathom activity. Telling the computer what to do gives the user the opportunity to "express fuzzy ideas in a formal, conventional and rigorous language"*(*Pratt, 1998, p. 108), and this makes the ideas become more concrete. The Formula Editor is a control mechanism that gives students a sense of ownership and brings them into direct contact with the fundamental notions of the stochastic (Pratt, 1998). At the same time, formulas resemble everyday language enough to be easy to learn, making Fathom a *user-friendly* environment. But what makes Fathom really unique is the idea of *Measures Collection*. Unlike black-box simulations, students get to appreciate "the different layers of simulations" involved when building the distribution of a statistic (Erickson, 2000).

When using Fathom, students first simulate a single sample from some population that they specify. Then, in the Measures Collection, they specify what statistic is relevant for the problem. They get to see how the Measures Collection calculates this statistic and turns it into a case attribute, which they can graph to see how it behaves. They can then build the distribution of the statistic by repeating this process many times, collecting the statistic for each sample. Since Fathom is a dynamic system, students get to see how the graph of the distribution of the statistic changes as the number of samples increases. If they collect a large number of samples, the distribution they get will be very close to the sampling distribution of the statistic.

Fathom's structure is such that it helps understand the relationship between population, sample, and sampling distribution. This is something that most of the students in the class who did not experiment with Fathom, as well as too many university students who take introductory statistics courses do not understand. Although they may be able to calculate a standard deviation and a standard error, they *"do not understand how these concepts are related (and distinguished) and so make application mistakes such as using one concept when they should use the other."* (Schau and Mattern, 1997, p. 91)

Cohen and Chechile (1997) warn us that if students do not have intuitions about the variables that a distribution represents, they might get lost. This is more likely to happen, of course, when dealing with sampling distributions, for which students are much less likely to have intuitions, than with population distributions, where world knowledge about the variable it represents often helps deduce the likely shape and spread of the distribution. Use of technology in the classroom did not allow students to interact with the concept of sampling distribution; it was weak in helping build their intuitions about sampling distributions. In contrast, the group of students I worked with who made use of Fathom developed a stronger conceptual understanding of both sampling distributions and the logic of inferential statistics and became "able to interpret the results of abstract manipulations in terms of concrete reality." (Wild, 1994, p. 209)

Pratt (1998) differentiates between constructive and instructive formal representations, "the former offered as a means for the learner to build new representations whilst the latter offered as a finished article, an expression of

culturally sanctioned mathematics" (p. 108). Fathom, in contrast to more conventional computer learning environments, puts more emphasis on constructive representations. Pre-packaged simulations, no matter how well designed, "can not take into account all of the possible 'what-if' questions that users will want to ask", and thus limit students "in how far they can improvise and explore" (Resnick, 1994). Fathom, on the other hand, is not a tutorial program, but a general-purpose learning tool that students can use to build and modify their own simulations.

Black-box simulations, where students simply observe the computer build the sampling distribution, do not allow them to make direct connections between the formal and the informal. Fathom, on the other hand, allows this connection by providing a medium for the design of activities that integrate experiential and formal pieces of knowledge. Formal mathematics can be expressed in both symbolic and iconic forms, providing the chance for connecting the symbolic and the iconic. Abstract mathematical ideas can be presented on the screen in forms that are concrete and visible and allow the user to directly manipulate and use them (Pratt, 1998). Students can articulate their informal theories, use them to make conjectures, and then use the results to test and modify these conjectures. The interaction between the data and the theoretical model seems to be much more convincing than black-box simulations, and helps students construct more powerful meanings for the stochastic (Wilensky, 1993).

In the next chapter, I will summarize the findings of the study and discuss how the insights gained led to a much better understanding and further refining of the conjecture. I will then present my thoughts about the implications of this research for statistics learning and pedagogy and for future research directions.

Chapter VI: Conclusions

SUMMARY

The results of a previous study of PACE and other statistics students, which agreed with the main findings of research in the area of stochastics education, gave the motivation for the study described in this thesis. Similarly to the research literature, we found in that previous study that the students we interviewed, regardless of whether they came from a lecture-based classroom or from the PACE course which made wide incorporation of technology and engaging activities, had poor intuitions about the stochastic and tended to think deterministically. This led me to conclude that the reason behind students' difficulties might be the instructional neglect of variation. I conjectured that the reform movement would be more successful in achieving its objectives if it were to put more emphasis on helping students build sound intuitions about variation and its relevance to statistics.

The thesis described how the conjecture driving the study was developed and how it was linked to classroom practice. It reported the experiences and insights gained from adopting an alternative path to statistics instruction that had variation as its central tenet in a college-level, introductory statistics classroom. Chapter III described the "variation as the central tenet of statistics instruction" conjecture, which was based on the literature review summarized in Chapter II and on previously conducted research. Chapter IV described how insights gained from student assessment prior to instruction led to further elaboration of the

conjecture and, consequently, the instructional program that was described in Chapter V.

The findings from student assessment at the beginning of the course further supported the conjecture that variation is neglected, and its critical role in statistical reasoning is under-recognized. At the outset of instruction, we witnessed the tendency of the students participating in this study to think deterministically and have a hard time differentiating between chance variation in the data and variation due to some form of underlying causality. Although students did recognize the existence of variation among samples, they tended to underestimate its effect. This tendency was more pronounced in real world-contexts. Although students did seem aware of the dangers involved when drawing conclusions from small samples, when asked to make their own judgments based on data, they often ignored these dangers and, exaggerating the reliability of the information provided, did not hesitate to use small samples as a basis for inferences.

We decided that the problem of people focusing on center and of erring toward the side of attributing too much to deterministic causality when investigating real-life situations needs to be addressed throughout the statistics course, through conceptual evolution of the role of variation. We re-defined statistics instruction in ways that we thought would help increase students' awareness of variation. The main aim of instruction became to communicate to students three "variation" messages: (1) variation is omnipresent; (2) variation can have serious practical consequences; and (3) statistics provides with tools that

help us make sense of the omnipresent variation, allow for it, or even control it (Wild and Pfannkuch, 1999).

Both the research literature and the assessment of student knowledge prior to instruction, led us to conclude that in order to build connections between formal mathematical expressions of the stochastic and everyday informal intuitions, we had to base instruction on the following principles:

1. Complementary of theory and experience: Statistical thinking always necessitates a complementarity of theory and experience. It should not be viewed a separable entity but a synthesis of statistical knowledge, context knowledge, and the information in the data in order to produce implications, insights and conjectures. If the statistics classroom is to be an authentic model of the statistical culture, it should model realistic statistical investigations, rather than teaching methods and procedures in a sequential manner and in isolation. The emphasis should be on the statistical process. The teaching of the different statistical tools should be achieved through putting students in authentic contexts where they need those tools to make sense of the situation. Statistical techniques should come to be viewed as a means to describe trends and patterns and deviations from those patterns existing in the data because of the variation inherent in every process. Probability should not be presented as a body of clear and unambiguous generalizations free of any concrete interpretations. Probability distributions should be presented as models based on some assumptions which, when changed, might lead to changes in the distribution.

2. Balance between stochastic and deterministic reasoning: Instruction should view as an important precursor of statistical reasoning, students' intuitive tendency to come up with causal explanations for any situation they have contextual knowledge about. It should present statistical thinking as a balance between stochastic and deterministic reasoning. It should stress that statistical strategies, based on probabilistic modeling, are the best way to counteract our natural tendency to view patterns even when none exists, to distinguish between real, non-ephemeral causes and ephemeral patterns that are part of our imagination.

The study did not follow the common research practice of taking snapshots of students' thought processes, and almost never doing any follow up of their initial thinking to watch for future transitions. The aim was not to identify errors in student thinking to catalogue as misconceptions that ought to be replaced, but to work with students' intuitive notions and help them develop ways to map new and richer concepts onto the ones that they already possessed. The conjecture-driven research model, which sees research and practice as interwoven and advocates curriculum construction based on an ongoing process of development and feedback, provided a way to systematically research conceptual change. It allowed thorough investigation of introductory statistics students' intuitive understanding of variation and use of the knowledge acquired to design, implement, evaluate, and refine meaningful interventions that helped students develop and expand upon their understanding of variation.

By examining how students' intuitions evolved during the course, I was able to identify structures that facilitated the articulation of intuitions about the stochastic. Findings from the study suggest that the instructional approach employed, with its emphasis on the omnipresence of variation, did help students develop statistical thinking that goes beyond the superficial knowledge of terminology, rules and procedures. Students' understanding of graphical tools and numerical measures of center and spread was much more sophisticated than that of students in the previous study we had conducted. Instruction proved quite effective in achieving one of its main goals – helping students move away from "uni-dimensional" thinking and integrate center and variation into their analyses and predictions. Although not totally letting go of their deterministic mindset, students were much more willing to interpret situations using a combination of stochastic and deterministic reasoning. The course increased significantly their awareness of sampling variation and its effects.

The experience I gained from the course has also led me to agree with the instructor who thinks that the idea of "rare event" linked to everything, from the beginning to the end of the course, is very promising in helping improve student understanding of inferential statistics. It helped students see the purpose and usefulness of standard deviation and of the z-scores. It was also quite effective in helping them make connections between exploratory and inferential statistics. For example, when confidence intervals were first introduced, and the instructor asked students to give a likely interval for the number of raisins in a box, students used the empirical rule in order to include in the likely interval those values that

are not “rare events”. Of course, they used the standard deviation of the population and not the standard error, but still they understood the logic behind confidence intervals. Also, when doing hypothesis testing, the idea of rejecting the null hypothesis when the statistic seems unlikely, was quite understandable to students since it was directly analogous to their familiar idea of a “rare event” falling at the tails of a normal distribution.

The investigation of students’ conceptions and beliefs in a real school setting has also allowed me to gain wealth of information about the source of student difficulties. I found, for example, the different meanings that students attached to sampling variation as being one of the main sources of difficulties they had with comprehending sampling distributions. Several students viewed variation as sample representativeness and thus argued that the variation of a sample increases with increase in sample size. Similarly, others who viewed variation as range also argued that variation goes up with increase in sample size. These beliefs regarding variation of individual samples affected how students perceived the relation between sample size and variation of sampling distribution. Both of these groups of students shared the belief that the bigger the sample size, the higher the variation of a sampling distribution.

Students’ difficulty with abstract notation was another source of difficulty I identified. The fact that at the end of the course so many students were still confused about basic notation helps explain why understanding seemed to fall apart each time the instructor would use some mathematics symbols. Statistics language was also a hurdle to effective learning. Much of the terminology of

statistics borrows words widely used in everyday speech that have different connotations than their statistical meaning. The everyday connotations of words such as independent, effective, bias, and error, was a main source of students' difficulties.

The outside-of-class investigation that assessed that effectiveness of the technological tool Fathom as an aid to conceptual understanding, has enabled me to explore the different ways in which student intuitions are shaped by the computer learning environment. Comparing the understandings of the "PACE group" to those of the rest of the students has led me to the conclusion that the choice of computer-based tools is a crucial decision. Use of technology in the classroom did not allow students to interact with the concept of sampling distribution, it was weak in helping build student intuitions about sampling distributions. Use of black-box simulations did not prove very successful in helping students understand the relationship between population distribution, sample distribution, and sampling distribution and, consequently, between population standard deviation, sample standard deviation, and standard error. Many of the students in the class did not seem aware of the transformation involved when moving from the distribution of individual values to the distribution of sample means. In contrast, the students who made use of Fathom, developed a stronger conceptual understanding of sampling distributions and the logic of inferential statistics in general.

Fathom's structure facilitated learning. Features such as the *"drag and drop" interface* increased immediacy of data analysis (Pratt, 1998). Use of the

Formula Editor helped formal ideas become more concrete by providing students with a control mechanism that gave them a sense of ownership and brought them into direct contact with the fundamental notions of the stochastic. Students built and modified their own simulations, and this allowed them, by working at their own pace, to make direct connections between the formal and the informal. Unlike "black-box" simulations, where students simply observe the computer build the sampling distribution, the structure of Fathom and, especially the *Measures Collection*, helped students understand the relationship between population, sample, and sampling distribution. The dynamic nature of this learning environment allowed students to interact with the data and see the immediate impact that their actions had on the different representations of the data. Abstract mathematical ideas were presented on the screen in symbolic and iconic forms that were concrete and visible and allowed students to directly manipulate and use them. Students articulated their informal theories, used them to make conjectures, and then used the results to test and modify these conjectures.

IMPLICATIONS FOR INSTRUCTION

Since instruction that aims to build direct links with students' intuitive reasoning has to take into account the unique characteristics and background of learners, a model of student understanding is not possible. I acknowledge that this study focuses on a single classroom with characteristics that could not be replicated. Nonetheless, I still believe that the experiences and insights gained can be powerful and relevant for other statistics educators and curriculum

developers also. The fact that the conjecture guiding the study was tested and refined in a real classroom is a big advantage compared to studies that draw their conclusions by taking snapshots of students' thinking.

This study emerged due to dissatisfaction about the neglect of variation by both statistics curricula and the research literature. In light of the results of this study, it seems that instruction built around the central tenet of variation does indeed lead to improved learning. According to Landwher's summary of the main findings of the research literature, people tend to believe that any difference in means is significant, to have unwarranted confidence in small samples and insufficient respect for small differences in large samples, and to underestimate the effect of variation in the real world. In contrast, students in the current study, who came in direct contact with the omnipresence and serious practical consequences of variation, were found to be much less likely to compare differences in measures of center without taking spread into account, or to be confident in conclusions drawn from small samples. They recognized that in addition to knowledge of the center of a distribution, one always needs information about its spread also. Instruction managed to get across to students the idea that "thinking about variability is the main message of statistics" (Smith, 1999, p. 249).

The structure of the course, with its simultaneous focus on variation and on the process of statistical investigation, proved a promising alternative to more conventional instruction, where the linear and consecutive structure of the course comes in sharp contrast with the complex nature of stochastical knowledge. In

our previous study of PACE and other statistics students, we had witnessed superficial and not well-interconnected knowledge of statistical concepts. In this study, most of the students had good understanding of the meaning of the different measures of center and spread introduced in the course and of the relationship between them. In addition, although still having difficulties, students acquired a much better understanding of graphical tools. More importantly, students recognized the connection between numerical summaries and graphical representations of the data and appreciated the usefulness of graphical representations in helping us make inferences and predictions. They came to view graphical and numerical tools as parts of the process of statistical investigation and not as ends in themselves (Friel et al., 1997).

The findings of this study also point to the benefits of an instructional approach that takes students' intuitions more seriously into account. The emphasis of the course on the complementarity of theory and experience, and its efforts to situate instruction within contexts familiar to the learner, proved helpful in building bridges between students' intuitions and statistical reasoning. For example, although in the pre-assessment 67% of the students argued that a gambler who has observed a ball landing on red six consecutive times in a roulette wheel having 18 black and 18 red numbers, should bet on black for the next outcome, only 9% of the students shared the same belief at the end of the course. The proportion of students stating that black and red are equally likely to come up on the next landing rose from 20% to 91%. The so much documented in the literature *gambler's fallacy,* which denotes the expectation of local correction to

random fluctuation in a sequence, was hardly at all observable by the end of the course.

The efforts of instruction to present statistical thinking as a balance between deterministic and stochastical reasoning, did prove useful in helping students better understand the relationship between chance and regularity. Students were, at the end of the course, much less prompt to assume that short-term fluctuations in the data must be causal and to develop causal explanations compared to the beginning of the course (e.g. "Map of New Zealand" question). The structure of the course led to the emergence of a functional view where short-term and long-term behavior are not discrete entities, but the one merges into the other (Pratt, 1998). Nonetheless, similarly, to Pratt (1998), we also saw that in what he calls "the region of large small numbers" problems remained, since it was unclear for students whether such situations should cue local or global meanings. In this study, students' acute awareness of uncertainty resulted, at some instances, in failure to realize that as long as the sample was randomly selected, it does not take an extremely large sample for patterns to begin to emerge. Future instruction should take additional steps to help avoid the danger of some students seeing everything as being the result of "random variation" (Pfannkuch and Brown, 1996).

Although instruction did undoubtedly help improve students' statistical reasoning, we saw that their understanding of methods of inference was still not well developed. The short duration of the course might have contributed to this.

Nonetheless, in our previous study of PACE students who had completed a semester-long course, we observed similar confusions about the nature and purpose of inferential statistics. The format of the course allowed students to use their intuitions to make sense of population distributions and of distributions of single samples. However, students found themselves lost when looking at sampling distributions since they did not have intuitions about them (Cohen and Chechile, 1997). The use of technology in the classroom was not successful in helping build students' intuitions about the distribution of statistics. The implications from this, as well as many other studies, are that the inferential paradigms typically employed in the statistics classroom are subtle and difficult for students to grasp (Wild and Pfannkuch, 1999) and that more intuitively sound paradigms, that make more constructive use of technology, are required.

The experiences I gained from this work have led me to conclude that a major limitation of the instructional approach was its use of technology. I am convinced that instruction would have been much more effective if it had adapted a technological tool such as Fathom rather than the more conventional use it made of technology. Use of a software such as Fathom, which is specifically designed to encourage students to build, refine, and reorganize their prior understandings and intuitions about the stochastic, would have been much more suited with the overall structure of the course. It would have been more successful in achieving the desired synergy among content, pedagogy, and technology than "black-box" simulations.

I advocate use of a more informal approach using a technological tool such as Fathom, because it can help "make the abstractions more concrete" by using experimental rather than theoretical probabilities (Erickson, 2000). At the beginning of the course, students can create informal confidence intervals from distributions and decide, by looking at graphical displays, what range of values of that parameter they would consider reasonable, and which values seem implausible and beg for another explanation (Erickson, 2000). The instructor could eventually ask students to find ways to quantify their thinking. It would be preferable if students are first asked to create their own test statistics instead of using the standard ones. As we have seen, students might have quite different notions of what, for example, "statistically significant" means, than an official test of hypothesis would. It is likely that students will use their intuitively convincing idea of a "rare event" as a basis of whether to reject the null hypothesis or not. The instructor should encourage students to also come up, depending on the context of the situation, with alternative ways to the "two standard deviations from the center" rule of defining whether an outcome is rare or not.

The conceptual difficulties students have to overcome when working with experimental probabilities are similar to those they encounter when dealing with theoretical probabilities. Students still have to understand the real problem in order to model it properly. They have, for example, to decide whether the events are independent, or whether they should sample with or without replacement (Erickson, 2000). Or, as we saw with the students working with Fathom, when first building distributions of statistics students are likely to confuse single

samples with the samples of statistics obtained with the Measures collection. Still, the experimental approach "gives students a more concrete take on the problems", and should "work hand-in-hand" with the more formal approach (Erickson, 2000).

Of course, since simulations take longer and empirical probabilities do not exactly match their theoretical counterparts, it is still useful for students to learn "the streamlined statistics of t tables and professional packages" (Erickson, 2000, p. 227). Nonetheless, because the logic of inference is so much more understandable through the simulation approach used by Fathom, students will get insights that will also help them better understand what is really going on when using traditional inferential tools. By moving more slowly and less abstractly, students can come to appreciate in ways they never did before the meaning and power of tools such as z, t, and chi-square (Erickson, 2000).

Employing such an approach that encourages students to explore the entire distribution of a statistic to see what is likely and not likely to occur, is more effective than standard simulation approaches to statistical inference, where there is a tendency to focus on relative frequencies or on means and to ignore variation. This focus on the entire distribution of outcomes should apply not only to activities involving technology, but also to other kinds of tasks posed in the class or used to assess student learning.

This study something has confirmed that students' performance is dependent on the type of question posed. For example, in both this study and the pilot investigation, students did quite well in the M&M question, taken from

Shaughnessy (1997a). The majority of even students with no statistics background chose the interval more likely to be drawn when taking samples of 20 M&M's from a bag where 40% of the M&M's are brown. In contrast, students in both studies did poorly in the question asking them to decide who, among two friends, is more likely to get 80% or more heads, Shelly who is going to flip a coin 50 times, or Diane who is going to flip the coin 10 times. A very high proportion of students in both studies chose the response stating that since each coin flip is a separate event the probability of heads is not affected by the number of times flipped. As already discussed in the literature review, Shaughnessy (1997a) argues that a reason students do poorly in problems of this type is that the question is posed in the wrong way, which exposes what students cannot do rather than what they can do. The difficulty he sees with problems such as this is that they cause confusion because, although presumably dealing with the concept of spread, they focus attention on center. Implicitly, this question is asking for the likelihood of a particular outcome (80% heads).

The results of this study, concur with Shaughnessy (1997a), who believes that to find what students can do with variability, instructors and researchers should start posing questions that can be answered in a sampling context, and not in a context forcing students to compare point values of particular outcomes. Questions such as the M&M one can become a point of departure for the kind of instruction that likes to build on students' intuitions in order to help increase their understanding of what the likely spread of outcomes is for a sample from a certain population. By focusing on tasks that elicit conceptions of variability and

difference rather than center and sameness, and encourage consideration of the entire distribution of outcomes, students can gradually begin to get some idea of what is likely and what is unlikely to occur.

IMPLICATIONS FOR FUTURE RESEARCH

Although this study has provided some valuable insights into students thinking of variation as part of the overall statistical investigation, we still need to learn a lot more about this neglected area of statistical reasoning. A limitation of the study is that it focused on a single group of students, for a short duration of time. The students in this study were undergraduate statistics students with weak mathematical background. Understandings of students of different age groups and of a different background can be varied in future work.

This study is only part of an ongoing research effort to understand the obstacles to the learning of statistics and use this understanding to find ways to create learning environments that facilitate deeper understandings. The findings from the study have provided answers to some questions, but they have at the same time raised other questions. For example, we have seen that a main source of difficulties for students in comprehending sampling distributions were the different meanings they attached to sampling variation. Research should be carried out to investigate ways that could help students differentiate between the different notions of variation, and use the appropriate ones depending on the context of the situation. We have also seen that despite the emphasis of the course on graphical tools, students still had some difficulties constructing and interpreting them and confused different graphical representations, for example

scatterplots with histograms, or histograms with bar graphs. The current literature tells us very little about how understanding of graphical representations develops (Friel et al., 1997), although research has shown that even medical researchers often confuse histograms and bar graphs (in Kelly et al., 1997). A possible direction of future research is thus to find ways to help students recognize the different functions of the horizontal and vertical axes across different graphical representations (Friel et al., 1997)

In the review of the research literature (Chapter II), the need for more systematic research to guide developments in statistical education was pointed out. The prevailing methodology employed by researchers examining conceptions of data and chance of posing cognitive tasks to students in order to catalogue their misconceptions, provides little guidance as to how one might systematically research conceptual change. There is hardly any information about the source of students' difficulties. My experiences from this study have led me to conclude that a research model such as the tr*ansformative and conjecture-driven research design,* which views learning as dynamic rather than static and the researcher's goal as doing research on the process of learning, is a preferable alternative to the prevailing methodology. Such a research model can be very useful for expanding our understanding of the components that promote development and growth of students' understanding.

The wealth of information that emerged out of this study shows the advantages of using a variety of assessment tasks in order to triangulate student thinking. Students come to a situation with a wide range of skills and knowledge

and offer responses that are difficult to anticipate (Cohen and Checile, 1997). Future research, as well as instruction, should also use assessment items that complement each other in order to provide a more complete profile of what students are and are not learning and why (Cohen and Checile, 1997). A students' response might, for instance, be erroneous due to misunderstanding about the question is asking (e.g. the question at the end-of-course assessment asking whether a test of hypothesis establishes the truth of one of the two hypotheses). Conversely, it might be the result of poor context knowledge of the situation under study and not of statistical content. Open-ended tasks and multiple form of assessments are required to get a reliable picture of students' thought processes.

CONCLUDING REMARKS

My firm belief is that students are capable of statistical and probabilistic reasoning and that their difficulties are primarily due to limitations in the learning methods, tools and cognitive technologies employed (Wilensky, 1997). Tracing people's understanding of abstract concepts such as variation is a very difficult task and empirical research might have failed to discover the real obstacles to students' fuller comprehension and to find ways to effectively link their intuitions (Wilensky, 1993). More research is needed to gain better understanding of the sources of learners' difficulties with the stochastic and use this understanding to develop improved learning environments that will help learners establish adequate and stable intuitions that would allow access to the theoretical level (Borovcnik, 1990). The study described here is a step towards this direction.

Appendix A: Assessment Prior to Instruction

QUESTIONNAIRE

Question 1

Based on your experience, what does variability mean to you? Give a verbal explanation and/or an example.

Question 2

Would it be more desirable for variability to be high or low for each of the following cases? Explain your decision.

(a) Age of trees in a national forest.

(b) Diameter of new tires coming off one production line.

(c) Scores on an aptitude test given to a large number of job applicants.

(d) Daily rainfall

(e) Weight of a box of cereal.

Question 3

Two students who took a statistics class received the following scores (out of 100):

Student A - 60, 90, 80, 60, 80
Student B - 40, 100, 100, 40, 90

If you had an upcoming statistics test, who would you rather had as a study partner, A or B? Support your answer.

Question 4

Suppose you took your little nephew on an Easter parade. At the parade, the "Easter Bunny" handed out packets of Gummy Bears to all of the students. Each packet had 6 Gummy Bears in it. To make up the packets, the Easter Bunny took 2 million green Gummy Bears and 1 million red Gummy Bears, put them in a very big barrel and mixed them up from night until morning. Then he spent the next few hours making up the packets of six Gummy Bears. He did this by grabbing a handful of Gummy Bears and filling as many packets as he could. Then he reached into the barrel and took another handful, and so on, until all the packets were filled with 6 Gummy Bears.

(a) When you get home from the parade, you open up your packet. How many green Gummy Bears do you think might be in your packet? Can you tell me how you got that?

(b) Do you think all the students got *n* greens, where *n* is the number of Gummy Bears you gave in part (a)? Can you explain that to me?

(c) If you could look at the packets of 100 students, how many students do you think got n greens?

(d) Remember that the Easter Bunny was starting with 2 million greens and one million reds. Did he run out of one color long before the other when he was filling the bags or did they both last until near the end? Why?

Question 5

On average, there are 600 deaths due to traffic accidents each year in a city. A person in the city observed the following:

February	Number of deaths
Week 1:	3
Week 2:	12
Week 3:	21
Week 4:	14
March	
Week 5:	2

Assume that none of these weeks contain a holiday weekend. Suppose the headlines in the newspaper claimed that week three was a "disastrous" week and police reported that speed was a factor. The next week was described in the papers as more evidence that the city driving was deteriorating. At the end of

week five the police congratulated themselves for the low death rate - their extra patrols had succeeded. What would you say to this person?

Question 6

Students in a middle school are trying to raise money to go on field trip to Great America (an amusement park). They are considering several options to raise money and decide to do a survey to help them determine the best way to raise the most money. One option is to sell raffle tickets for a SEGA video-game system. Consequently, nine different students each conducted a survey to estimate how many students in the school would buy a raffle ticket to win a SEGA. Each survey asked 60 students but each sampling method and results were different. The six surveys and their results were as follows:

(1) Tom asked 60 friends. (75% yes, 25% no)
(2) Shannon got the names of all 600 students in the school, put them in a hat, and pulled out 60 of them. (35% yes, 65% no)
(3) John asked 60 students at an after school meeting at the Games Club. The Games Club met once a week and played different games - especially computerized ones. Anyone who was interested in games could join (90%, 10% no)
(4) Ann sent out a questionnaire to every kid in the school and then used the first 60 that were returned to her. (50% yes, 50% no)
(5) Claire set up a booth outside the lunchroom and anyone who wanted to could stop by and fill out her survey. To advertise her survey she had a sigh that said "WIN A SEGA". She stopped collecting surveys when she got 60 completed. (100% yes)
(6) Kyle wanted the same number of boys and girls and some students from each grade. So, he asked 5 boys and 5 girls from each grade to get his total of 60 students. (30% yes, 70% no)

(a) What do you think about the way that each survey was conducted? Do you think it was done in a proper way? Do its results give a good picture of how

many students in the school would want to buy a raffle ticket to win a SEGA? Explain why or why not.

Survey 1:

Survey 2:

Survey 3:

Survey 4:

Survey 5:

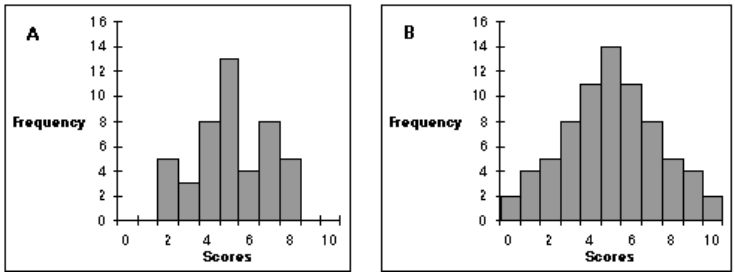
Survey 6:

(b) If you were to pick one of the six ways to do the survey, which one would you choose? Explain why.

(c) What do you think is the best estimate of what percentage of kids will buy a raffle ticket?

Question 7

Which of the following distributions shows MORE variability? Check one of the choices:



A has more variability_____ B has more variability_____

Question 8

A roulette wheel has 18 black (B) and 18 red (R) numbers. The probability of a ball landing on a red is the same as landing on a black. A gambler observes the ball to land on red six times in a row, that is RRRRRR. What do you expect the next color to be? Why?

Question 9

Circle the best answer to the following problem:

At a nearby college, half the students are women and half are men. A worker for a student organization wants to interview students on their views about recent changes in the federal government's funding of financial aid. The worker wants to get a good representation of the students, and goes to as many different areas on campus as possible. Three or four students are interviewed at each place the worker visits. Out of the last 20 students interviewed, 13 were women and 7 were men. Now, you do not know what time of day it is, to which part of campus the worker has already gone, or where the worker is going next. Out of the next 20 students the worker interviews, do you think more will be women or men?

(a) The worker seems to interview more women than men. There could be several reasons for this. Perhaps women are more willing to talk about their opinions. Or, maybe the worker goes to areas of campus where there are more women than men. Either way, the worker is likely to interview more women than men out of the next 20 students.
(b) Since half of the students on this campus are men and half are women, you would expect a 50/50 split between the number of men and women the worker interviewed. Since there tended to be more women than men so far, I expect the opposite trend to start happening. Out of the next 20 students the worker interviews, there will probably be more men than women so that things start to balance out.
(c) Half the students on this campus are men and half are women. That means that the worker has a 50/50 chance of interviewing a man or a women. It should not matter how many men or women the worker has interviewed so far. Out of the next 20 students interviewed, about half should be men and half women.
(d) So far, the trend seems to be more women to be interviewed than men. Out of the next 20 students the worker interviews, I would expect the same thing to happen. The worker will probably interview more women than men out of the next 20 students.

Question 10

Consider the following list of variables:

(a) age at death of a sample of 34 persons
(b) the last digit in the social security number of each of the 40 students
(c) scores on a fairly easy test in statistics
(d) height of a group of adults
(e) number of medals won by medal-winning countries in the 1992 Winter Olympics

Use your knowledge of the variable (i.e. ask yourself if the distribution is likely to be symmetric or not) to match the variables with the following histograms. Justify your choice.

INTERVIEW PROTOCOL

Map Question

Every year in New Zealand approximately seven children are born with a limb missing. Last year the children born with this abnormality were located in New Zealand as shown on the map. What do you think? (In New Zealand, it is common knowledge that one-third of the population lives in the top region and one-sixth of the population in each of the other regions.)

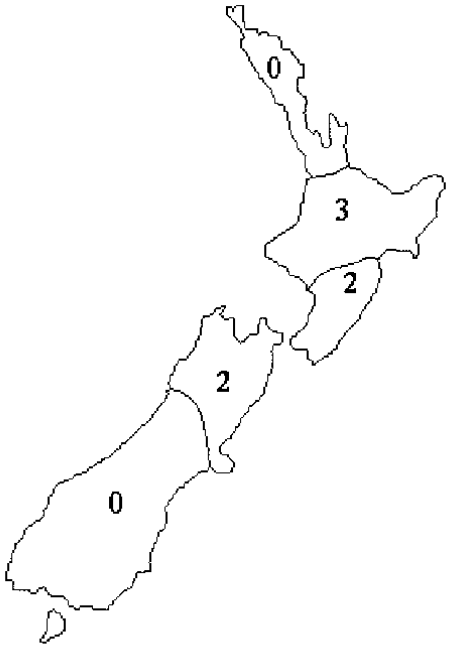


Figure 1

Die Toss Question

A fair die is tossed 7 times resulting in the outcome 3,3,3,4,4,5,5 (order is unimportant). What do you think of these results?

Coin Toss Question

A fair coin is tossed 50 times resulting in 27 heads. Two days later it is tossed again 50 times resulting in 30 heads. What do you think of these results?

Child Psychologist Question

A child psychologist is engaged in studying which of two toys infants will prefer to play with. Of the first five infants studied, four have shown a preference for this toy. The psychologist concludes that most infants will show a preference for this toy. Do you think the psychologist has drawn a valid conclusion?

Appendix B: SATs and GPAs

This activity is about sex differences in SAT scores and grade-point averages (GPAs) among first-year students from an un-named college a few years ago. Here are some important statistics from the data, as displayed in a Fathom summary table:

SATGPA	Summary Table		
		math	verbal
sex	F	521.9 484	486.0 484
	M	564.6 516	492.5 516
Column Summary		543.9 1000	489.3 1000

1. What do you notice based on these statistics alone?

2. Now let's look at the raw data and examine further. Open the file *C/Program Files/Fathom/Sample Documents/Exploring Mathematics with Fathom/SATGPA*

3. Verify the table above by constructing it. To get the empty table, choose *Summary Table* from the *Insert* menu. Drag the attribute (variable) *math* dropping it on the right-pointing arrow. Do the same for *FYGPA*. Drop sex on the down-pointing arrow.

4. Means are never the whole story. Look at a distribution. Make a new graph by dragging the graph tool (fourth icon from left) and then putting *math* on the horizontal axis, over the spot labeled *Drop an attribute here*. You get a dot plot. Change it to a histogram by choosing *Histogram* from the popup menu in the corner of the graph.

5. Drag *sex* onto the vertical axis to get separate graphs for males and females. Write down what you notice.

Make a second graph and do the same for *FYGPA*. The bin widths may be too big to compare the two graphs easily. Drag on the edge of a bin to adjust it. You'll have to adjust the vertical scale as well. What do you notice?

How can we decide who is better overall? Let's look at *verbal* for more insight.

6. Replace *FYGPA* with *verbal* on the second graph. Now the males and females are nearly identical (though the men's is slightly higher). But both the *math* and *FYGPA* sex differences are much bigger than the verbal one.

7. Change this graph to a box plot by choosing *Box Plot* from the popup menu in the graph. Do the same for the *math* graph. Sketch the two graphs below. Label and scale the axes.

8. Describe in words how you know from the graph that the difference in math scores is more meaningful (not just bigger) than the difference in verbal scores.

Making it Quantitative

(i) Difference in medians

Standard deviation is a measure of spread of the distribution that usually goes with the mean (a measure of center). One way to describe the difference between males and females is to compare the difference in their mean *math* and *FYGPA* scores to the standard deviation of each of these distributions. Armed with that functions *mean ( )* and *stdDev( ),* we can make this difference quantitative.

1. Make a new summary table by Selecting *Summary Table* from the *Insert* menu. Drag *math*, *verbal*, and *FYGPA* on top of the right-pointing arrow of the table, but do not drag *sex*. The overall means of the attributes appear.

2. We will now find the ratio of the mean differences over the standard deviation. Click on the summary table and then go to *Summary* menu and choose *Add formula*. Click the divide sign on the calculator. Enter the following formula in the formula editor (do not type the question marks; they appear automatically):

$$\frac{\text{mean}(?, \text{sex} = \text{"M"})\ \ \text{mean}(?, \text{sex} = \text{"F"})}{\text{stdDev}()}$$

3. Close the formula editor.

4. What do you conclude about sex differences in SAT scores and first-year college GPA?

(ii) Difference in medians

5. Another way to make the comparison numerically, is to compare the difference in median (SAT, FYGPA etc.) scores to the interquartile range (the size of the boxes), which is the *inq()* function in Fathom. The interquartile range is a measure of spread that usually goes with median.

6. We need to find ratio of the difference in medians between males and females over IQR. Make a new row in the summary table that will calculate that ratio by click on the summary table and then going to *Summary* menu, choosing *Add formula* and entering the following formula:

$$\frac{\text{median}(?, \text{sex} = \text{"M"})\ \ \text{median}(?, \text{sex} = \text{"F"})}{\text{iqr}()}$$

7. Close the formula editor.

8. What do you conclude about sex differences in SAT scores and first-year college GPA?

9. Make a new attribute, *totalSAT,* which is the sum of math and verbal. To do this, first go to the case table (the table containing all the scores), and replace <*new*> with *totalSAT*. Then choose *Show Formulas* from the *Display* menu. Double-click the shaded rectangle below *totalSAT* and enter the formula *math+verbal* in the formula editor. Perform this activity's analysis on that new attribute-that is drop *totalSAT* onto the right-pointing arrow of the summary table. What does this add to our analysis of sex differences?

10. Someone could argue that the difference in SAT scores is over 40, but the difference in GPAs is only about 0.15. Therefore, the difference in SATs is more significant. Explain why that isn't important without using the words, "interquartile range," "standard deviation," or "variance".

Appendix C: Drug for Reducing Cholesterol Level

FDA is the gatekeeper to make sure any new drug is thoroughly tested and proven effective before it can be made available to patients. Companies have to submit their data to demonstrate their new drug indeed significantly improves patients' condition. This usually involves many years of experiments on animals as well as on patients. Here, I am presenting a very simplified example of a new drug developed to reduce cholesterol level.

Company A submitted the following data and claimed that the new drug is very effective in reducing cholesterol level:

A new drug by company A was given to a random selected sample of 64 patients whose cholesterol levels were at the high risk level of more than 250 mg before receiving any drug. Before receiving the new drug, the average cholesterol level of these 64 patients was 285 mg. The new drug was given to patients for a month. Their cholesterol levels were measured again. 64 cholesterol levels were obtained and summarized into Sample Mean = 250 mg, and Sample S.D. = 100 mg.

Now, you work for the FDA and your job is to find out whether the claim made by company A is appropriate or not.

(Q1) Based on your knowledge and common sense, when you see the information provided by the company, do you think this new drug is effective or not? Why?

(Q2) Based on the average cholesterol level before taking any drug, and assuming that the s.d. of cholesterol levels before taking any drug is also 100 mg, how likely is it that the average cholesterol level of a sample of 64 high risk patients will be lower than 250 mg? (Hint: this is a probability problem based on the distribution of sample mean: $\overline{X} \sim N(285, 100/\sqrt{64})$

(Q3) Based on the results in (Q2), is the average cholesterol level of 250 mg a significant reduction from 285 mg? (Hint: Consider this to be the case if the 64 patients' average cholesterol level of 250 mg is a rare case when 285 mg is the true average, i.e. if 250 mg is in the lower 5% of the population of patients whose average cholesterol level is 285 mg.)

(Q4) Let us consider this situation in terms of standardized *z-scores* for 250 mg based on the sampling distribution of the sample mean from 64 patients: $\overline{X} \sim N(285, 100/\sqrt{64})$. What is the corresponding standardized z-score for 250 mg? Does this *z-score* fall outside the two *s.d.* of 285 range?

The approach we used above to deciding whether the decrease in the average cholesterol level of the patients is significant, is based on the idea of checking whether a sample mean occurs rarely or not. We have learned how to solve this type of problem before, when we were discussing the Empirical Rule and Normal Distributions in general. This result helps us decide if the drug reduced the average cholesterol level to a significant level.

(Q5) Based on your answers to Q2-Q4, what is your conclusion about the effect of this new drug?

More formally, this is a hypothesis testing problem. Hypothesis testing is similar to a court case. It involves a process of decision making based on data information. The rules that are applied to make the decision are based on some probability rules. Before we look into how to conduct a hypothesis test, let us go to the court house to observe how a judge decides if someone is innocent or not.

When a criminal case comes to court, the person is first assumed INNOCENT, and will be eventually judged either innocent or guilty based on the INFORMATION (or EVIDENCE) presented by the prosecutor and the defendant. The rules that are used by the judge are the LAW.

The two choices for the judge are (1) This person is INNOCENT, or (2) This person is GUILTY. At the beginning, the person is ASSUMED INNOCENT.

Therefore, to simplify the discussion, we use **Ho** for the assumed situation (that is: The person is assumed INNOCENT) and we call it NULL HYPOTHESIS.

The alternative that the prosecutor is trying to prove (The person is GUILTY) is called the ALTERNATIVE HYPOTHESIS, and the notation is **Ha**. Complete the following blanks for the court case:

(Q6)

Ho: _____ Ha:_____

Decision Rule: _____

Sample Information: _____

Final Decision: Made by Judge, who applies the Decision Rule to the Sample Information, and decides to conclude Ho or Ha.

NOTE: It is important to understand that no matter what the final decision made by the judge is, there is ALWAYS some chance of making errors.

(Q7) There are two possible errors in this decision making process. What are they?

(Q8) Based on the types of errors described in (Q7), which type of error is considered more critical, that is, if that type of error was made, it would have consequences that are more serious in general?

(Q9) Based on (Q8), the judge would like to reduce the type of the more critical error, so that the judge will not have too high a probability to make this critical error. In order to reduce this type of more critical error, what suggestion (s) do you have?

Now back to the New Drug case. This is indeed a hypothesis testing problem. There are also two possible decisions to choose from: One is the NULL hypothesis: "The new drug is NOT effective", the other is the ALTERNATIVE, which is what the company tries to prove: "The new drug is effective". We therefore set up the NULL and ALTERNATIVE hypotheses as:

Ho: The new drug is NOT effective Ha: The new drug is effective

(NOTE: Ha is what the study tries to prove. In this case, it is "The New drug is effective")

(Q10) The above statements are somewhat vague for making a decision. We will need to be more specific as to what "Not effective" and "Effective" mean. In this study, what do you think we should define as "Not effective", and "Effective"?

Since, before taking the drug, the average cholesterol level was 285 mg, if a drug is to be effective, it must provide strong evidence that the average cholesterol level after taking the drug is much lower than 285 mg. So, one way to lay out Ha is:

Ha: The true average cholesterol level after taking the drug < 285 mg.

Ho would be the opposite of Ha, that is,

Ho: The true average cholesterol level after taking the drug >= 285 mg

However, to prevent confusion, we usually set up Ho as:

Ho: The true average cholesterol level after taking the drug = 285 mg.

NOTE: It is a good idea to determine Ha first, then set the equal sign for Ho, so that we need to worry only about setting Ha.

The statements above are written as:

$H_o : \mu = 285\ mg$ $\qquad$ $Ha : \mu < 285\ mg$

Here we use the notation μ to represent "The true average cholesterol level after taking the new drug".

NOTE: There are three kinds of hypothesis tests: Left-side test, Two-side test, and Right-side test.

The general from of Ho and Ha are respectively:

Left-side Test: $H_o : \mu = \mu_o$ $\qquad$ $Ha : \mu < \mu_o$

Two-side Test: $H_o : \mu = \mu_o$ $\qquad$ $Ha : \mu \neq \mu_o$

Right-side Test: $H_o : \mu = \mu_o$ $\qquad$ $Ha : \mu > \mu_o$

(Q11) What kind of test does the New Drug study involve?

The idea of making this type of decision is similar to the decision-making by a judge in a court.

You have solved the problem by figuring out if the sample average 250 mg is rare when the true average is assumed to be 285 mg.

In fact, what you did there is an informal procedure of solving a hypothesis problem.

If the new drug has no effect, it is similar to saying that it is as if these patients did not take any drug, and hence, the true average cholesterol level can be assumed the same as the average cholesterol level before taking any drug, that is our Ho. So, (Q2) asks:

If the true average cholesterol level after taking the new drug is 285 mg (meaning no effect), how likely is it observe a sample of 64 patients having an average cholesterol level of 250 or lower? Now, if this probability is really small (e.g. smaller than 5%), then, to have an average from 64 patients lower than 250 mg is very rare, and we can therefor claim that there is a strong effect made by the new drug. Otherwise, we will not make such a statement, and conclude that the sample evidence is not strong enough to claim that the drug has a significant effect.

(Q12) Based on the results you got in (Q2) and (Q3), what would be your decision?

In Q4, you also obtained a standardized z-score of the sample mean, 250 mg. Does this z-score fall outside two s.d. of 285 mg among all possible sample means? (Hint: Compare the observed z-score with the z-value –1.645 (the z-value at which there is a 5% chance to be lower). Since the z-value –1.645 is the cut-off point for us to decide if the observed z-score is rare or not, we call this z-value at α **=5% the critical value, and denote it by** $-z_{\alpha}$**.** (NOTE: This is for Left-side test. There will be different for Two-side test for right-side test).

(Q13) Based on the comparison between z_{obs} and the critical z-value, do you draw the same conclusion as you did in Q12?

(Q14) As we discussed in the Court Case, there are always possible errors coming with the decision. For this New Drug study, what are the two possible types of mistakes involved with the decision?

(Q15) What type of error is considered more serious in general?

The 5% that is used to define "a rare case" is called *"the level of significance"*. The notation is α. This is common is real world decision making. Another commonly used "level of significance" is α = **1%.**

The probability value you obtained by computing $P(\bar{X} < 250)$ in (Q2) is the chance of obtaining such a rare event in a sample of 64 patients. Since 250 mg is observed from the sample average, this rare event probability is called **"the observed level of significance"**, and the notation is used to describe it is **p-value**.

Your final decision is made by comparing this p-value with the given α using the following rule.

If p-value < α , then conclude that there is a significant event.

If p-value >= α , then conclude that there is no significant event.

When you compute the corresponding z-score for the observed sample mean, the z-score from the observed sample mean 250 mg is called the **observed z-score**, z_{obs}. The same decision as the one we made above, can also be made by using the observed z-score. (NOTE: This is for the Left-side Test. Two-side and Right-side tests will be different.)

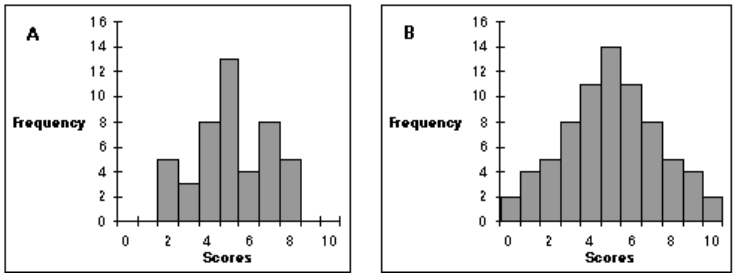
If $z_{obs} < -z_\alpha$, then accept Ha, and decide that the new drug is significantly effective.

If $z_{obs} >= -z_\alpha$, then the sample data do not provide enough evidence that the new drug is effective.

Appendix D: End-of-Course Assessment

Question 1

Which of the following distributions shows MORE variability? Check one of the choices:



A has more variability_____ B has more variability_____

Question 2

Two surveys were conducted to determine how many higher level institutions in Texas are recycling. The first survey used a large sample size and a self-selected sampling method by sending out postcards to all the deans of higher level institutions in Texas. About half of the deans sent them back, and 91% of those that returned the postcards said that they recycled. The second survey used a medium sample size and a random sampling method (the names of all the higher level institutions were put into the computer and a program which gave each institution an equal chance of been chosen, selected the specified sample size). Thirty-seven percent of the schools said that they recycled. Evaluate each survey and decide which of the two you think is better.

Question 3

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 80% or more heads?

a) Diane because the more you flip the closer you get to 50%.
b) Shelly because the greater the sample size, the greater the variability in results.
c) Neither because each coin flip is a separate event and the probability of heads is not affected by the number of times flipped.
d) Other (please specify): _____

Question 4

FDA has a maximum upper limit for nicotine contents to be 12 mg. A company is manufacturing a new brand of cigarettes. FDA sent an evaluator to test the nicotine content.

(a) The evaluator took a random sample of 10 cigarettes and found the mean nicotine content to be 13 mg with a standard deviation of 2 mg. Based on this sample of 10, do you think the FDA should conclude that the average nicotine level is not acceptable (is significantly higher than the acceptable brand)? Why or why not?

(b) What about if the evaluator takes a sample of 100 and again finds the mean to be 13 mg? Why or why not?

(c) The company filed a complaint that, based on their test, the mean nicotine level is 11.8. Is it possible that the FDA has made a mistake:

(i) When basing their decision on a sample of size 10?

(ii) When basing their decision on a sample of size 100? Explain why or why not.

Question 5

If events A and B are independent and P(A) = .6, P(B) = .4, which of the following is correct?

a. $P(A \cap B) = 0$ b. $P(A \cup B) = .76$

c. $P(A \cap B) = 100$ d. $P(A \cup B) = .24$

Explain the statement 'A and B are independent'

Question 6

When constructing a histogram for describing the distribution of salary for individuals who are 40 years or older, but are not yet retired.

Explain:

(a) what is on the Y-axis :

(b) what is on the x-axis:

What would the proper shape of the salary distribution? Explain why.

Question 7
The amount of time it takes to take an exam has a skewed-to-left distribution with a mean of 65 minutes and a standard deviation of 8 minutes. A sample of 64 students will be selected at random.

PART A
Which of the following describes the distribution of the amount of time it takes to take an exam?
a. N(65,8)
b. N(65,1)
c. A skewed distribution with a mean of 65 minutes, but unknown variance.
d. A skewed distribution with a mean of 65 minutes and a standard deviation of 8 minutes.
Explain your reason:

PART B
Which of the following properly describes the sampling distribution of the sample mean based on n=64?
a. Approximately N(65,8)
b. Approximately N(65,1)
c. Approximately N(1,65)
d. Skewed distribution with a mean of 65 and standard deviation of 1.
Explain your reason:

Question 8
At a nearby college, half the students are women and half are men. A worker for a student organization wants to interview students on their views about recent changes in the federal government's funding of financial aid. The worker wants to get a good representation of the students, and goes to as many different areas on campus as possible. Three or four students are interviewed at each place the worker visits. Out of the last 20 students interviewed, 13 were women and 7 were men. Now, you do not know what time of day it is, to which part of campus the worker has already gone, or where the worker is going next. Out of the next 20 students the worker interviews, do you think more will be women or men?

a. The worker seems to interview more women than men. There could be several reasons for this. Perhaps women are more willing to talk about their opinions. Or, maybe the worker goes to areas of campus where there more women than men. Either way, the worker is likely to interview more women than men out of the next 20 students.
b. Since half of the students on this campus are men and half are women, you would expect a 50/50 split between the number of men and women the worker interviewed. Since there tended to be more women than men so far, I expect the opposite trend to start happening. Out of the next 20 students the worker interviews, there will probably be more men than women so that things start to balance out.
c. Half the students on this campus are men and half are women. That means that the worker has a 50/50 chance of interviewing a man or a woman. It should not matter how many men or women the worker has interviewed so far. Out of the next 20 students interviewed, about half should be men and half women.

d. So far, the trend seems to be more women to be interviewed than men. Out of the next 20 students the worker interviews, I would expect the same thing to happen. The worker will probably interview more women than men out of the next 20 students.

Question 9

Imagine you have a huge jar of M&M's with many different colors in it. We know that the manufacturer of M&Ms puts in 40% browns. If you reached in and pulled samples of 20 M&Ms at a time, what do you think would be the likely range for the numbers of browns you found in your samples?

a) 8 because the proportion of browns in the bag is 40% and you would expect the sample to represent the population.
b) 0-8 because 40% of 20 is 8.
c) 8-20 because 40% of 20 is 8.
d) 6-10 each time because 40% of 20 is 8 and it would vary a little bit each time because you are taking a sample.
e) The range would be 0-20 because there is no way to make predictions with such a small sample.
f) Other (please specify):_____

Explain your reason:

Question 10

Suppose two distributions have exactly the same mean and standard deviation. Then the two distributions have to look exactly alike.

(a) True
(b) False

Explain the reason:

Question 11

A small class was given a test on arithmetic and the scores were recorded. The same test was

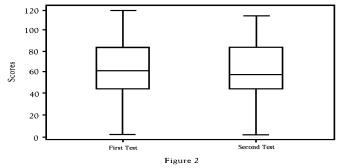


Figure 2

given a few weeks later. The box plots for both sets of scores are shown.

Have the scores changed significantly? : (a) Yes (b) No

Explain the reason:

Question 12

A bowl has 100 wrapped hard candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Jenny pulls out a handful of 10 candies, counts the number of reds, and tells her teacher. The teacher writes the number of red candies on a list. Then, Jenny puts the candies back into the bowl, and mixes them all up again. Four of Jenny's classmates, Jack, Julie, Jason, and Jerry do the same thing. They each pick ten candies, count the reds, and the teacher writes down the number of reds. Then they put the candies back and mix them up again each time.

I think the teacher's list for the number of reds is most likely to be (please circle one):

a. 8,9,7,10,9
b. 3,7,5,8,5
c. 5,5,5,5,5
d. 2,4,3,4,3
e. 3,0,9,2,8

Explain your reason:

Question 13

A statistical test of hypotheses correctly carried out *establishes* the truth of one of the two hypotheses, either the null or the alternative one:

(a) True (b) False

Explain the reason:

Question 14

If you take a sample of data from the population described above, what information will you be able to calculate from these data? Check as many as apply:

(i) μ (ii) s (iii) σ (iv) $\bar{x}$

Circle all of the symbols below which represent parameters:

μ s σ $\bar{x}$

For the symbols listed below, circle the ones, which vary for sample to sample:

μ s σ $\bar{x}$

Question 15

A roulette wheel has 18 black (B) and 18 red (R) numbers. The probability of a ball landing on a red is the same as landing on a black. A gambler observes the ball to land on red six times in a row, that is RRRRRR. What do you expect the next color to be:

a. Red b. Black c. About equal chance

Explain your reason:

References

Azcarate, P., and Cardenoso, J. M. (1994). Why Ask Why? *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Ballman, K. (1997). Greater Emphasis on Variation in an Introductory Statistics Course. *Journal of Statistics Education, 5*(2).

Bar-Hillel, M. (1982). Studies of representativeness. In Kahneman, Slovic and Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 69-83). Cambridge: Cambridge University Press.

Batanero, C., and Godino, J. D. (1994). The Use of Multivariate Methods to Analyze Students' Stochastic Conceptions. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics*. Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

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

Batanero, C., Estepa, A., and Godino, J. D. (1997). Evolution of Students' Understanding of Statistical Association in a Computer-Based Teaching Environment. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 198-212). Voorburg, The Netherlands: International Statistical Institute.

Beard, H. V., Schmitz, C. L., and Domahidy, M. R. (1997). Interdisciplinary Evaluation of Collaborative, School Based Family Support Centers. In L. D. Labbo and S. L. Field(Eds.), *Proceedings of the 1996 Conference on Qualitative Research in Education* [On-line]. Athens: Georgia. Available: http://www.coe.uga.edu/quig/Christie.html

Behrens, J. T. (1997). Toward a Theory and Practice of Using Interactive Graphics in Statistical Education. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics*

(pp. 111-121). Voorburg, The Netherlands: International Statistical Institute.

Bennet, A. (1997, October). Research Design Tasks in Case Study Methods. *Presented at the MacArthur Foundation Workshop on Case Study Methods* [On-line]. Bleeder Center for Science and International Affairs: Harvard University. Available: http://www.georgetown.edu/bennett

Ben-Zvi, D., and Friedlander, A. (1997). Statistical Thinking in a Technological Environment. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 54-64). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes: Do we need a probabilistic revolution after we have taught data analysis? In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics*. Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Biehler, R. (1997). Students' Difficulties in Practicing Computer-Supported Data Analysis: Some Hypothetical Generalizations From Results of Two Exploratory Studies. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 176-197). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (1999). Discussion: Learning to Think Statistically and to Cope with Variation. *International Statistical Review*, *67*(3), 259-262.

Blumberg, C. J. (1997). Discussion: How technology is changing the teaching of statistics at the college level. In J. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 279-283). Voorburg, The Netherlands: International Statistical Institute.

Bogdan, R. C., and Biklen, S. K. (1982). *Qualitative Research for education: an introduction to theory and methods*. Boston: Allyn and Bacon.

Borovnik, M. (1990). A Complementarity Between Intuitions and Mathematics. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics*. University of Otago, Dunedin, New Zealand.

Borovnik, M., and Peard, R. (1996). Probability. In A.J. Bishop (Ed.), *International Handbook of Mathematics Education* (pp. 239-287). Netherlands: Kluwer Academic Publishers.

Bowen, T. J. (1997, September). Understanding qualitative research: A review of Judith Meloy's Writing the Qualitative Dissertation: Understanding by Doing [On-line]. *The Qualitative Report, 3*(3). Available: http://www.nova.edu/ssss/QR/QR3-3/bowen.html

Breslow, N. E. (1999). Discussion: Statistical Thinking in Practice. *International Statistical Review, 67*(3), 252-255.

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

Burrill, G. (1997a). Graphing Calculators and Statistical Reasoning at the Secondary Level Through the Use of Technology. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 15-28). Voorburg, The Netherlands: International Statistical Institute.

Burrill, G. (1997b). Discussion: How Technology is Changing the Teaching and Learning of Statistics in Secondary School. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 71-74). Voorburg, The Netherlands: International Statistical Institute.

Burrill, G. (1997c). Discussion: Technology, Reaching Teachers, and Content. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 71-74). Voorburg, The Netherlands: International Statistical Institute.

Cantrell, D. C. (1990, November). *Alternative Paradigms in Environmental Education: The Interpretive Perspective* [On-line]. Presented as part of a symposium entitled Contesting Paradigms of Environmental Education Research at the Annual Conference of the North American Association for Environmental Education, San Antonio, Texas. Available: http://www.edu.uleth.ca/ciccte/naceer.pgs/pubpro.pgs/alternate/pubfiles/08.Cantrell.fin.htm

Carr, J., and Begg, A. (1994). Introducing Box and Whisker Plots. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference*

on Teaching Statistics. Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Catterall, M., and Maclaran, P. (1997). Focus Group Data and Qualitative Analysis Programs: Coding the Moving Picture as Well as the Snapshots [On-line]. *Sociological Research Online, 2*(1). Available: http://www.socresonline.org.uk/socresonline/2/1/6.html

Celedon, S. (1998). *An Analysis of a Teacher's and Students' Language Use to Negotiate Meaning in an ESL/Mathematics Classroom.* Unpublished doctoral dissertation: The University of Texas at Austin.

Chenail, R. (1990, Summer). Introduction [On-line]. *The Qualitative Report, 1*(1). Available: http://www.nova.edu/ssss/QR/QR1-1/editorial.html

Chenail, R. J. (1992, Spring). Qualitative research: Central tendencies and ranges. *AFTA Newsletter*, 43-44.

Christie, A. A. (1997). Using Telecommunications to Break Down Gender Stereotypes. In L. D. Labbo and S. L. Field(Eds.), *Proceedings of the 1996 Conference on Qualitative Research in Education*. Athens: Georgia. Available: http://www.coe.uga.edu/quig/Christie.html

Cobb, G. W., Witmer, J. A., and Cryer, J. D. (1997). *An Electronic Companion to Statistics.* New York: Cogito Learning Media Inc.

Cohen, L.J. (1979). On the Psychology of Prediction: Whose is the Fallacy? *Cognition, 7*, 385-407.

Cohen, S., and Chechile, R. A. (1997). Overview of ConStatS and the ConStatS Assessment. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 110-119). Voorburg, The Netherlands: International Statistical Institute.

Cole, P. M. (1994, Spring). Finding A Path Through The Research Maze [On-line]. *The Qualitative Report, 2* (1). Available: http://www.nova.edu/ssss/QR/BackIssues/QR2-1/cole.html

Confrey, J. (1980). *Conceptual change analysis: Implications for mathematics and curriculum inquiry.* East Lansing, MI: Institute for Research on Teaching, Science-Mathematics Teaching Center, Michigan State University.

Confrey, J. (1988, February). *The concept of exponential functions: A student's perspective.* Invited address to the conference Epistemological Foundations of Mathematics Experience, University of Georgia.

Confrey, J. (1990). A review of the research in students' conceptions of mathematics, science and programming. *Review of Research in Education, 16,* 3-32.

Confrey, J. (1991). Learning to listen: A student's understanding of powers of ten. In E. von Glaserfeld (Ed.), *Radical Constructivism in Mathematics Education* (pp. 111-136). Netherlands: Kluwer Academic Publishers.

Confrey, J. (1995, July). Student voice in examining "splitting" as an approach to ratio, proportions and fractions. *Proceedings of the 19th Annual Meeting of the International Group for the Psychology of Mathematics Education*. Recife, Brazil: Universidade Federal de Pernambuco.

Confrey, J. (1996, April). *Strengthening Elementary Education through a Splitting Approach as Preparation for Reform Algebra.* Presented at the annual meeting of the American Educational Research Association, New York, NY.

Confrey, J., and Lachance, A. (1999). Transformative Teaching Experiments Through Conjecture-Driven Research Design. In A. E. Kelly and R. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education.* Mahwah, N. J.: Lawrence Erlbaum Assoc.

Confrey, J., and Scarano, H. Constructivism and the Practicing Teacher. *In Support of Excellence: Views From the Field* (CD-ROM). Columbus, OH: Eisenhower National Clearinghouse.

Confrey, J. (1994). Voice and Perspective: hearing epistemological innovation in students' words. *Revue des Sciences de L'education, 20*(1), 115-133.

Confrey, J. and Smith, E. (1995). Splitting, Covariation, and their role in the development of exponential functions. *Journal for Research in Mathematics Education. 26*(1), 66-86.

Conti, G. J. (1997, Summer). Research in the Tribal Community. Two Research Paradigms. *Native Research and Scholarship Symposium Papers* [On-line]. Available: http://www.fdl.cc.mn.us/tcj/summer97/GC.html

delMas, R., and Garfield, J. (1990). The Use of Multiple Items to Identify Misconceptions in Probabilistic Reasoning. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

delMas, R. C. (1997). A framework for the evaluation of software for teaching statistical concepts. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 75-90). Voorburg, The Netherlands: International Statistical Institute.

delMas, R., Garfield, J., and Chance, B. (1998). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Submitted to the Journal of Statistics Educatio*n.

Denzin, N. K., and Lincoln, Y. S. (1994). Introduction: Entering the field of qualitative research. In N. K. Denzin and Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 1-17). Thousand Oaks, CA: Sage.

Eisner, E. (1985). *The Educational Imagination*. New York: Macmillan, 216-252.

Emerson, R. M., Fretz, R. I., and Shaw, L. L. (1995). *Writing Ethnographic Fieldnotes*. Chicago: The University of Chicago Press.

Erickson, T. (1999). Data in Depth. Exploring Mathematics with Fathom. Draft. Emeryville, CA: Key Curriculum Press.

Erickson, T. (2000). *Data in Depth. Exploring Mathematics with Fathom*. Emeryville, CA: Key Curriculum Press.

Erlandson, D. A., Harris, E. L., Skipper, B. L., and Allen, S. D. (1993). *Doing naturalistic inquiry: A guide to methods.* Newsbury Park, CA: Sage.

Estepa, A. and Batanero, M. C. (1994). Judgments of Association in Scatterplots: An Empirical Study of Students' Strategies and Preconceptions. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Falk, R. and Konold, C. (1992). The Psychology of Learning Probability. In F. and S. Gordon (Eds.), Statistics for the Twenty-First Century. *MAA Notes, 29* (pp. 151-164). USA: The Mathematical Association of America.

Fischbein, E. (1975). *The Intuitive Sources of Probabilistic Thinking in Children*. Dordrecht, The Netherlands: Reidel.

Fischbein, E. (1987). *Intuition in Science and Mathematics.* Dordrecht, The Netherlands: Reidel.

Friel, S. N., Bright, G. W., Frierson, D., and Kader, G. D. (1997). A Framework for Assessing Knowledge and Learning in Statistics (K-8). In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education* (pp. 55-63). Burke, VA: IOS Press.

Gal, I., and Garfield, J. (1997). Curricular Goals and Assessment Challenges in Statistics Education. In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education.* Burke, VA: IOS Press.

Garfield J. and delMas, R. (1990). Exploring the Stability of Students' Conceptions of Probability. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

Garfield, J. B. (1994). Beyond Testing and Grading: Using Assessment To Improve Student Learning. *Journal of Statistics Education, 2*(1).

Garfield, J. B. and delMas, R. C. (1994). Students' Informal and Formal understanding of Variation. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Garfield, J. (1997). Preface. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. ix-xi). Voorburg, The Netherlands: International Statistical Institute.

Garfield, J., and Chance, B. L. (1998). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Submitted to the Journal of Statistics Education.*

Garfield, J., delMas, B., and Chance, B. L. (1999). *Tools for Teaching and Assessing Statistical Inference: Simulation Software* [On-line]. Available: htpp://www.gen.umn.edu/faculty_staff/delmas/stat_tools/stat_tools_software.htm

Geertz, C. (1973). *The interpretation of cultures.* New York: Basic Books.

Ghosh, J. K. (1997). Discussion. *International Statistical Review, 65*(2), 154-155.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky. *Psychological Review, 103* (3).

Glaser, R. (1962). Psychology and instructional technology. In R. Glaser (Ed.), *Training research and education.* Pittsburgh, PA: University of Pittsburgh Press.

Glaser, B. G. and Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago, Illinois: Aldine Publishing Company.

Glencross, M. J., and Binyavanga, K. W. (1997). The role of technology in statistics education: A view from a developing region. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 301-308). Voorburg, The Netherlands: International Statistical Institute.

Goldstein, L.S. 1997. More than gentle smiles and warm hugs: Applying the ethic of care to early childhood education. *Journal of Research in Childhood Education. 12* (2): 244-261.

Gordon S. (1997). Students' Orientations to Learning Statistics – Profiles of Experience. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 171-178).

Ghosh, J. K. (1997). Discussion. *International Statistical Review, 65*(2), 154-155.

Greene, J. (1994). Qualitative program evaluation practice and promise. In N. Denzin and Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 530-544). Thousand Oaks, CA: Sage

Guba, E. G., and Lincoln, Y. S. (1989). *Fourth Generation evaluation*. Newsbury Park, CA: Sage.

Hawkins, A. (1997a). Myth-conceptions. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics*

(pp. vii-viii). Voorburg, The Netherlands: International Statistical Institute.

Hawkins, A. (1997b). Children's Understanding of Sampling in Surveys. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 1-14). Voorburg, The Netherlands: International Statistical Institute.

Hawkins, A. (1997c). Discussion. *International Statistical Review, 65*(2), 141-146.

Hiebert, J., and Carpenter, T. (1992). Learning and Teaching with Understanding. In D. Grows (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 65-100). New York: Macmillan

Hoerl, R., Hahn, G. & Doganaksoy, N. (1997). Discussion: Let's Stop Squandering Our Most Strategic Weapon. *International Statistical Review, 65*(2), 147-153.

Jacobs, V. (1997). Missed opportunities on the teaching and learning of data and chance. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 3-37).

Jacobs, J.E., and Potenza, M. (1991). The Use of Judgement Heuristics to Make Social and Object Decisions: A Developmental Perspective. *Child Development, 62*, 166-178.

Jimenes, V. A., and Holmes, P. (1994). Students' Understanding of the Logic of Hypothesis Testing. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Jones, G., Thornton, C., Langrall, C., and Mogill, A. T. (1997). Using Students' Probabilistic Thinking to Inform Instruction. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 171-178).

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. (1973). *On the Psychology of Prediction. Psychological Review, 80*(4), 237-251.

Kahneman, D., and Tversky, A. (1982). *On the Study of Statistical Intuitions. Cognition, 11,* 123-141.

Kelly, A. E., Sloane, F., and Whittaker, A. (1997). Simple Approaches to Assessing Underlying Understanding of Statistical Concepts. In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education*. Burke, VA: IOS Press.

Kettenring, J. (1997). Discussion. *International Statistical Review, 65*(2), 153.

Konold, C. (1989). Informal Conceptions of Probability. *Cognition and Instruction, 6,* 59-98.

Konold, C., Pollatsek, A., Well, A., and Hendrickson, J. (1990). The Origin of Inconsistencies in Probabilistic Reasoning of Novices. In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Konold, C. (1995a). Issues in Assessing Conceptual Understanding in Probability and Statistics. *Journal of Statistics Education, 3*(1).

Konold, C. (1995b). Confessions of a Coin Flipper and Would-Be Instructor. *The American Statistician, 49*(2), 203-209.

Konold, C., Pollatsek, A., Well, A., and Gagnon, A. (1997). Students Analyzing Data: Research of Critical Barriers. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 159-175). Voorburg, The Netherlands: International Statistical Institute.

Kuhn, T. (1962). *The Structure of Scientific Revolution.* Chicago: University of Chicago Press.

Lachance, A., and Confrey, J. (1996, April). *Mapping the Journey of Students' Explorations of Decimal Notation via Ratio and Proportion.* Presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Lakatos, I. (1976). *Proofs and Refutations.* Cambridge: Cambridge University Press.

Latour, B. (1987). *Science in Action.* Cambridge, MA: Harvard University Press.

Lave, J. (1988). *Cognition in Practice.* Cambridge: Cambridge University Press.

Lee, C. (1997a). *Promoting Active Learning in Introductory Statistics Using the PACE Model* [On-line]. Available: http://www.cst.cmich.edu /users/ leec/PACE-1.html

Lee, C. (1997b). *An Assessment of the PACE Strategy for an Introductory Statistics Course* [On-line]. Available: http://www.cst.cmich.edu/users/ leec/PACE-2.html

Lee, C. M., Fouladi, R T., & Meletiou, M. (1998, September). *Teaching and learning in introductory statistics: Impressions, attitude, learning style and motivation.* Presented at the Third Annual Conference on Research in Undergraduate Mathematics Education, Southbend, Indiana.

Lincoln, Y. S., and Cuba, E. G. (1985). *Naturalistic inquiry.* Newbury Park, CA: Sage.

Lipson, K. (1997). What do students gain from simulation exercises? An evaluation of activities designed to develop an understanding of the sampling distribution of a proportion. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 137-150). Voorburg, The Netherlands: International Statistical Institute.

Loosen, F., Lioen, M., and Lacante, M. (1985). The standard deviation: some drawbacks of an intuitive approach. *Teaching Statistics, 7*, 29-39.

Lopes, L. (1991). The Rhetoric of Irrationality. *Theory and Psychology, 1*(1), 65-82.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37* (11), 2098-2109.

Marshall, C., and Rossman, G. B. (1995). *Designing Qualitative Research.* Thousand Oaks, CA: Sage Publications.

McCloskey, M. (1997). QERCUS and STEPS: The Experience of Two CAL Projects From Scottish Universities. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 99-109). Voorburg, The Netherlands: International Statistical Institute.

Meletiou, M., Confrey, J., & Lee, C. M., & Fouladi, R. T. (1999, April). What do Introductory Statistics Students Gain from Technology? In Confrey, J. (Chair), *Making Teachers Smart About Reform: Developing Their Knowledge of Mathematics, Pedagogy, and Assessment.* Interactive Symposium conducted at the annual meeting of the American Educational Research Association.

Meletiou, M., Lee, C. M., & Fouladi, R. T. (1998, September). *Interviews as a pedagogical and research tool.* Presented at the Third Annual Conference on Research in Undergraduate Mathematics Education, Southbend, Indiana.

Meletiou, M., Lee, C. M., & Fouladi, R. T. (1999a, June). *Students' intuitive understanding of variability.* Presented at the Second Biennial Midwest Conference on Teaching Statistics. Oshkosh, Wisconsin.

Meletiou, M., Lee, C. M., & Fouladi, R. T. (1999b, September). *Statistics Instruction Informed by Student Thinking: The Case of Statistical Variation.* Presented at the Third Annual Conference on Research in Undergraduate Mathematics Education, Chicago, Illinois.

Meletiou, M., Lee, C. M., & Myers, M. (1999). The Role of Technology in the Introductory Statistics Classroom: Reality and Potential. *Proceedings of the International Conference on Mathematics/Science Education and Technology.* San Antonio, Texas.

Metz, K. E. (1997). Dimensions in the Assessment of Students' Understanding and Application of Chance. In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education.* Burke, VA: IOS Press.

Mokros, J. R., Russell, S. J., Weinberg, A. S., and Goldsmith, L. T. (1990). What's Typical? Children's Ideas about Average. In J. B. Garfield (Ed.),

Research Papers from the Third International Conference on Teaching Statistics. University of Otago, Dunedin, New Zealand.

Moore, D. (1990). Uncertainty. In L.Steen (Ed.), *On the shoulders of giants: new approaches to numeracy* (pp. 95-137). USA: National Academy Press.

Moore, D. S. (1991). *Statistics Concepts and Controversies.* New York: W. H. Freeman and Company.

Moore, D. (1992). Statistics for All: Why? What and How? In D. Vere-Jones (Ed.) *Proceedings of the Third International Conference on Teaching Statistics: Volume 1* (pp. 423-428). Voorburg: International Statistical Institute.

Moore, D. S. (1993). A generation of statistics education: An interview with Frederick Mosteller. *Journal of Statistics Education, 1*(1).

Moore, D. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review, 65*(2), 123-165.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (1998). *Principles and Standards for School Mathematics: Discussion Draft.* Reston, VA: Author.

Nau, D. S. (1995, December). Mixing Methodologies: Can Bimodal Research be a Viable Post-Positivist Tool? *The Qualitative Report, 2*(3). Available: http://www.nova.edu/ssss/QR/QR2-3/nau.html

Nicholson, J. (1997). Developing Probabilistic and Statistical Reasoning at the secondary level through the use of data and technology. In J. B. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 29-44). Voorburg, The Netherlands: International Statistical Institute.

Nickson, M. (1981). *Social foundations of the mathematics curriculum: A rationale for change.* Unpublished doctoral dissertation, Institute of Education, University of London.

Nickson, M. (1992). The culture of the mathematics classroom: An unknown quantity? In D. A. Grows (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 101-114). New York: Macmillan.

Nisbett, R. E., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R., Krantz, D., Jepson, C., and Kunda, Z. (1983). The Use of Statistical Heuristics in Everyday Inductive Reasoning. *Psychological Review, 90*(4), 339-363.

Noss, R., and Hoyles, C. (1996). *Windows on Mathematical Meanings: Learning Cultures and Computers.* London: Kluwer Academic Publishers.

Nunes, T., Schliemann, A.D., and Carraher, D.W. (1993). *Street Mathematics and School Mathematics.* Cambridge: Cambridge University Press.

Pandit, N. R. (1996, December). The Creation of Theory: A Recent Application of the Grounded Theory Method [On-line]. *The Qualitative Report, 2*(4). Available: http://www.nova.edu/ssss/QR/QR2-4/pandit.html

Pfannkuch, M. and Brown, C. M. (1996). Building on and Challenging Students' Intuitions About Probability: Can We Improve Undergraduate Learning? *Journal of Statistics Education, 4*(1).

Pfannkuch, M. (1997). Statistical Thinking: One Statistician's Perspective. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 171-178).

Philips, B. (1999, November). Report from the IASE President. *IASE Review*, 2-3.

Piaget, J. (1970). *Genetic Epistemology*. New York. Columbia University Press.

Polya, G. (1962). *Mathematical Discovery.* New York: John Wiley and Sons.

Pratt, D. C. (1998). *The Construction of Meanings In and For a Stochastic Domain of Abstraction.* Ph.D. Thesis, University of London

Resnick, M. (1994). Learning About Life [On-line]. *Artificial Life Journal, 1*(1-2). Available:http://el.www.media.mit.edu/groups/el/Papers/mres/ALife/ALife.html

Rossman, A. J. (1996). *Workshop Statistics: Discovery with Data.* New York: Springer-Verlag.

Rossman, A. J. (1997). Using Technology to Promote Learning by Self-Discovery. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 226-237). Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., Bruce, B., and Tenney, Y. (1990). Learning about Sampling: Trouble at the Core of Statistics. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

Scarano, G. H., and Confrey, J. (1996, April). *Results from a Three-Year Longitudinal Teaching Experiment Designed to Investigate Splitting, Ratio and Proportion.* Presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Schau, C., and Mattern, N. (1997). Assessing Students' Understanding of Statistical Relationships. In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education.* Burke, VA: IOS Press.

Scheaffer, R. L. (1997). Discussion. *International Statistical Review, 65*(2), 156-158.

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

Schuyten, G., and Dekeyser, H. (1997). Computer-Based and Computer-Aided Learning of Applied Statistics at the Department of Psychology and Educational Sciences. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 213-222). Voorburg, The Netherlands: International Statistical Institute.

Shaughnessy, J. M. (1992). Research in Probability and Statistics: Reflections and Directions. In D. Grows (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 465-494). New York: Macmillan.

Shaughnessy, J. 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,* 285-316.

Shaughnessy, J. M. (1997a). Missed opportunities on the teaching and learning of data and chance. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 129-145).

Shaughnessy, J. M. (1997b). Discussion: Empirical research on technology and teaching statistics. In J. Garfield and J. Truran (Eds.), *Research Papers on Stochastics Education* (pp. 217-219).

Shaughnessy, J. M., Watson, J., Moritz, J., and Reading, C. (1999, April). School Mathematics Students' Acknowledgment of Statistical Variation. For the NCTM Research Presession Symposium: *There's More to Life than Centers*. Paper presented at the 77th Annual NCTM Conference, San Francisco, California.

Smith, T. M. F. (1999). Discussion. *International Statistical Review, 67*(3), 248-250.

Smith, J. P., diSessa, A.A., and Rochelle, J. (1993). Misconceptions Reconceived - A Constructivist Analysis of Knowledge in Transition. *Journal of Learning Sciences, 3*(2), 115-163.

Snee, R. (1990). Statistical Thinking and its Contribution to Quality. *The American Statistician, 44*(2), 149-154.

Snee, R. D. (1999). Discussion: Development and Use of Statistical Thinking: A New Era. *International Statistical Review, 67*(3), 255-258.

Starkings, S. (1997). How technological introduction changes the teaching of statistics and probability at the college level. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 233-254). Voorburg, The Netherlands: International Statistical Institute.

Steinbirg (1990). The Use of Chance-Concept in Everyday Teaching - Aspects of a Socially Constituted Epistemology of Mathematical Knowledge. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

Tsourvakas, G. (1997, September). Multi-Visual Qualitative Method: Observing Social Groups in Mass Media [On-line]. *The Qualitative Report, 3*(3). Available: http://www.nova.edu/ssss/QR/QR3-3/tsour.html

Truran, J. (1994). Children's Intuitive Understanding of Variance. *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Tversky, A., and Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance, 2(1),* 16-21.

Tversky, A., and Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology, 5,* 207-232.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Tversky, A., and Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement. *Psychological Review, 90*(4), 293-313.

Vygotsky, L. (1986). *Thought and Language*. MIT Press, Cambridge, MA.

Watson, J., and Baxter, J. (1997). Learning the Unlikely at Distance as an Information Technology Enterprise: Development and Research. In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 288, 302). Voorburg, The Netherlands: International Statistical Institute.

Well, A. D., Pollatsek, A., and Boyce, S. (1990). Understanding the effects of sample size on the mean. *Organizational Behavior and Human Decision Processes, 47,* 289-312.

Wiesman, D. and Wotring, E. (1997). *The Scientific, Humanistic, and Critical Paradigms: A Second Look* [On-line]. Available: http://mailer.fsu.edu/~ewotring/com5312/notes.html

Wild, C. (1994). Embracing the "wider view" of statistics. *The American Statistician, 48,* 163-171.

Wild, C. J., and Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*, 3, 223-265.

Wilder, P. (1994). Students' Understanding of Computer-Based Simulations of Random Behavior. *Research Papers from the Fourth International Conference on Teaching Statistics.* Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Wilensky, U. (1993). *Connected Mathematics - Building Concrete Relationships with Mathematical Knowledge.* Ph.D. Thesis, Massachusetts Institute of Technology.

Wilensky, U. (1997). What is Normal Anyway? Therapy for Epistemological Anxiety. In R. Noss (Ed.), *Educational Studies in Mathematics. Special Issue on Computational Environments in Mathematics, 33*(2) (pp. 171-202).

Wilson, B., Teslow, J., and Osman-Jouchoux, R. (1995). The Impact of Constructivism (and Postmodernism) on ID Fundamentals. In B. B. Seels (Ed.), *Instructional Design Fundamentals: A Review and Reconsideration* (pp. 137-157). Englewood Cliffs NJ: Educational Technology Publications.

Wolcott, H. F. (1990). On Seeking - and Rejecting - Validity in Qualitative Research. In E. Eisner and A. Peshking (Eds.), *Qualitative Inquiry in Education*. New York: Teachers College Press. Pages 121-152.

Wood, M. (1997). Computer Packages as a Substitute for Statistical Training? In J. Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 267-278). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes: Do we need a probabilistic revolution after we have taught data analysis? In J. B. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics*. Minneapolis: The International Study Group for Research on Learning Probability and Statistics.

Biehler, R. (1999). Discussion: Learning to Think Statistically and to Cope with Variation. *International Statistical Review, 67*(3), 259-262.

Borovnik, M., and Peard, R. (1996). Probability. In A.J. Bishop (Ed.), *International Handbook of Mathematics Education* (pp. 239-287). Netherlands: Kluwer Academic Publishers.

Erickson, T. (2000). *Data in Depth. Exploring Mathematics with Fathom*. Emeryville, CA: Key Curriculum Press.

Gal, I., and Garfield, J. (1997). Curricular Goals and Assessment Challenges in Statistics Education. In I. Gal and J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education.* Burke, VA: IOS Press.

Hoerl, R., Hahn, G. & Doganaksoy, N. (1997). Discussion: Let's Stop Squandering Our Most Strategic Weapon. *International Statistical Review, 65*(2), 147-153.

Lee, C. (1997). *Promoting Active Learning in Introductory Statistics Using the PACE Model* [On-line]. Available: http://www.cst.cmich.edu /users/ leec/PACE-1.html

Meletiou, M., Lee, C. M., & Myers, M. (1999). The Role of Technology in the Introductory Statistics Classroom: Reality and Potential. *Proceedings of the International Conference on Mathematics/Science Education and Technology.* San Antonio, Texas.

Steinbirg (1990). The Use of Chance-Concept in Everyday Teaching - Aspects of a Socially Constituted Epistemology of Mathematical Knowledge. In J. B. Garfield (Ed.), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

Wild, C. J., and Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*, 3, 223-265.

Vita

Maria Menelaou Meletiou was born in Nicosia, Cyprus, on July 30, 1968, the daughter of Athanasia Pontou Meletiou and Menelaos Meletiou. After completing her work at the American Academy of Larnaca, in 1987, she entered the Elementary Education program of the Pedagogical Academy of Cyprus. She earned her Teaching Certificate in Elementary Education in June 1990, and taught at the 5th Aglantzia Elementary School, in Nicosia, Cyprus, for one academic year. In August 1991, she came to the United States to pursue a Bachelor's of Art degree in Mathematics, at the University of Texas at Austin, which she earned in May 1993. She began graduate work at the University of Texas in August 1993, and earned the degree of Master's of Science in Statistics in August 1994. She entered the doctoral program in Mathematics Education, at the University of Texas, in the Fall of 1994 semester. While pursuing her doctoral degree, she also joined the Department of Mechanical Engineering at the University of Texas, and received a Master's of Science in Engineering degree in August 1998.

Permanent address: 3 Ayios Georgios St., 5380 Dherynia, Cyprus

This dissertation was typed by Maria Menelaou Meletiou.