UNIVERSITY OF CALIFORNIA
Santa Barbara

How Do College Students Reason About Hypothesis Testing in Introductory Statistics Courses?

A Dissertation submitted in partial satisfaction of the requirement for the degree of Doctor of Philosophy
in Education

by

Birgit Christina Aquilonius

Committee in charge:
Professor Mary E. Brenner, Chair
Professor Richard Durán
Professor Yukari Okamoto

March 2005

The dissertation of Birgit Christina Aquilonius is approved

_____

Richard Durán

_____

Yukari Okamoto

_____

Mary E. Brenner, Committee Chair

January 2005

How Do College Students Reason About Hypothesis Testing in Introductory Statistics Courses?

by

Birgit Christina Aquilonius

ACKNOWLEDGEMENTS

First I want to thank my advisor, Professor Betsy Brenner, for all her support and expert advice. She gave me great direction without restraining me from what I wanted to do. She also helped me with language at some critical junctions. I also want to thank Professor Richard Durán for encouraging me to apply to the UCSB School of Education and for letting me have a head start on my dissertation in his cognition class. I am grateful to Professor Yukari Okamoto for her detailed suggestions at the dissertation proposal meeting, particularly regarding probability education literature.

The professional discussions about statistics education with colleagues at West Valley College, particularly Jim Wilczak, were invaluable for this dissertation. Thank you, Alyson Clark, Cathy van Hook, Jim Wilczak and Rebecca Wong for letting me recruit participants for my study from your classes. Thank you Wade Ellis for your long-time support and for introducing me to graphing calculators when they first appeared in the educational landscape.

My year with the Functions Group at UC Berkeley's School of Education opened my eyes to the possibilities in analyzing student conversations. Thank you Professor Rogers Hall, Dr. Cathy Kessel, and Professor Alan Schoenfeld for your full inclusion of me in the EMST program at UC Berkeley during 1993-1994.

Thank you Lee Kanner, for all your support, good advice and also for reading and commenting on my work. Your belief in my work was very important to me. Thank you Lee Knutsen for your thorough proofreading of my manuscripts and for your friendship.

My most heartfelt thanks go to my husband Lasse Bergman. His constant support and our frequent discussions about my work made all the difference. Thanks also to our sons Thomas and Ted for their interest, support and inspiration.

VITA OF BIRGIT CHRISTINA AQUILONIUS
January 2005

EDUCATION

Physical Therapy Diploma, Karolinska Institutet, Stockholm, Sweden, June 1970
Bachelor of Arts in Mathematics, University of California, Santa Cruz, June 1979
Master of Arts in Mathematics, University of California, Santa Cruz, June 1981
Master of Arts in Mathematics, University of California, Santa Barbara, June 2001
Master of Arts in Education, University of California, Santa Barbara, March 2002
Doctor of Philosophy in Education, University of California, Santa Barbara, March 2005 (expected)

PROFESSIONAL EMPLOYMENT

1970-72	Physical Therapist, Jönköping County Hospital, Sweden
1980-82	Teaching Assistant, Mathematics Department, UC Santa Cruz
1982	Lecturer in Statistics, Summer Session, UC Santa Cruz
1982-83	Mathematics Instructor, Cabrillo College, Aptos and Monterey Peninsula College (part-time)
1984-present	Mathematics Instructor (full-time) West Valley College, Saratoga
1995-97	Teaching Assistant/ Associate in Mathematics, UC Santa Barbara

PUBLICATION

"Students' Peer Discussions of Statistics. How Do Students Learn From Them?" Unpublished Master's project submitted in partial fulfillment of the requirements for the Master of Arts degree in Education, University of California, Santa Barbara, 2002

PRESENTATION TO PROFESSIONAL ORGANIZATION

"Students Talk About Hypothesis Testing." Presentation at the California Mathematics Council, Community Colleges Annual Meeting, Monterey, December 2004.

AWARDS

Alpha Gamma Sigma Teacher Award, West Valley College, Saratoga, 1990, 1991 and 1992

Office of the President Community College Research Assistantship, UCSB, 2000-01

ABSTRACT

How Do College Students Reason About Hypothesis Testing in Introductory Statistics Courses?

by

Birgit Christina Aquilonius

Many college students are required to take statistics courses for their majors. Hypothesis testing is often taught as the last part of such a course and in a sense becomes the goal of the course. Statistics instructors receive mixed messages about their students' understanding of hypothesis testing. The students in their classes sometimes say or do things that make instructors believe that students have good understanding of the topic. At other times, the same students make mistakes on tests and homework that make the instructor doubt their understanding. In this study, present technology allowed me to go one layer below what can be observed in the classroom. By videotaping students' statistical conversations and viewing them on DVDs, time after time, I was able to analyze students' reasoning at more depth and observe more closely what students understand and do not understand. Two statistics instructors and eight pairs of community college students were asked to solve hypothesis test problems and answer questions about their work.

Regarding sample and population students were able to reason competently in general terms. They knew why one takes samples and about the importance of unbiased samples. They did not realize the mathematical character of random

sampling, but thought about randomness as equivalent to representativeness. In their reasoning they did not exhibit understanding of the qualitative difference between sample mean and population mean which is inherent in the theory of hypothesis testing.

Students' approach to p-values in hypothesis testing was procedural. They considered p-values as something that one compares to alpha-values in order to arrive at an answer. Students did not attach much meaning to p-values as an independent concept. Therefore it is not surprising that though their p-values gave them valid statistical conclusions, they often were puzzled over how to translate the statistical answer to an answer of the question asked in the problem. Their textbooks and instructors gave students scripts to help them formulate their answers. Those scripts were helpful to some students but did not always lead them to the right answer.

TABLE OF CONTENTS

Chapter 1 Introduction …….. 1
1.1 Overview of Problem ……. 2
1.2 Research Questions …… 5
1.2.1 How do students reason about the concepts of sample and population in the context of hypothesis testing? …….. 5
1.2.2 How do students reason about hypothesis testing p-values? ….. 8
1.2.3 How do students reason about answers to hypothesis test questions? ……. 10

Chapter 2 Literature Review …….. 14
2.1 A Model for Statistical Reasoning …….. 14
2.2 Probability Education Research Relevant to Research Questions …….. 15
2.2.1.Misconceptions in hypothesis testing …….. 16
2.2.2 Belief in the law of small numbers …….. 19
2.2.3 Understanding the variability of sample means ……... 21
2.2.4 Making sense of randomness ……... 24
2.2.5 Representativeness ……... 27
2.2.6 The outcome approach to probability ……... 31
2.3 Understanding versus procedures in statistics ……. 33
2.4 What can be learned from mathematical problem solving research? ….. 37
2.5 The statistical register and students' difficulties with semantics ……. 41
2.6 Sense making and construction of meaning in statistical learning …….. 43
2.7 Intuition in mathematics and statistics ……. 50
2.8 Some method considerations grounded in education research literature . 54
2.9 Summary ……...……58

Chapter 3 Methods ……... 62
3.1 Research design ……... 62
3.2 Setting and participants ……... 64
3.3 Procedures ..……...…68
3.4 Data collection ……..… 68
3.4.1 Student interviews ……. 69
3.4.2 Instructor interviews ……. 70
3.4.3 Textbooks ……...71
3.5 Interview questions …….. 72
3.5.1 First research session …….....72
3.5.2 Second research session …… 74
3.6 Analysis …… 76

Chapter 4 Results …… 79
4.0 Some comments on the students' overall performance in the project …. 79
4.1 Research question #1. How do students reason about the concepts sample and population? …….. 82
4.1.1 Students sometimes confused population and sample means ... 84
4.1.2 The abstract character of sample and population in the Tranquilizer problem caused students difficulties ……. 90
4.1.3 One pair placed different values in the null and alternative hypothesis …….. 97
4.1.4 Most of the students were not able to detect that a one-sample problem presented to them had incorrectly been set up as a two-sample problem …….. 98
4.1.5 Two pairs solved the Jail problem correctly ……... 103
4.1.6 All students recognized bias in a given sampling procedure, but their suggestions for doing sampling varied in sophistication ……. 107
4.1.7 Most students gave rationales similar to the instructors' for using samples ……. 110
4.1.8 Summary ……. 113
4.2 Research question #2. How do students reason about the p-value in statistical hypothesis testing? ……... 114
4.2.1 Curriculum leading up to hypothesis testing …….. 116
4.2.2 The *Understandable Statistics* textbook's treatment of the p-value concept ……. 118
4.2.3 The *Workshop Statistics* textbook's treatment of the p-value concept ……. 120
4.2.4 The instructors' treatment of the p-value concept ……... 121
4.2.5 Students' graphical representations in hypothesis testing …... 123
4.2.6 What is the meaning of p-values in hypothesis testing? ……. 129
4.2.7 Difficulties with considering p-values solely from a procedural view …….. 134
4.2.8 Abstract character of p-values causes students difficulties …. 136
4.2.9 Summary …… 144
4.3 Research Question #3.How do students reason about answers to hypothesis test questions? …… 146
4.3.1 Textbook treatments of how to arrive at final answers ……... 150
4.3.2 Instructor treatments of how to arrive at final answers …….. 153
4.3.3 Arriving at the final answer is seldom trivial for the students, even when solving problems correctly …… 155
4.3.4 Sometimes students failed to properly consider their hypotheses when answering …….. 160

4.3.5 Students reacted when they perceived a conflict between the data and their answer ……... 162
4.3.6 Students used expressions reflecting the probabilistic character of their answers …… 167
4.3.7 A few times students preferred drawing conclusions directly from the sample …… 171
4.3.8 Students knew that you do not draw conclusions directly from the sample data, but reasons given were procedural rather than conceptual …… 176
4.3.9 Some concluding remarks regarding how students arrived at their final answers …… 183
4.4 How did the students reason about statistical hypothesis testing? ……. 185

Chapter 5 Discussion ……...191
5.1 Students did not use mathematical models in their reasoning ……..…192
5.2 Most Students had a procedural approach to p-values …….. 194
5.3 Scripts helped students answer hypothesis test questions ……. 197
5.4 Students did not include probability theory in their reasoning about hypothesis testing ……. 199
5.5 The TI-83 Calculator was used more frequently by the students than by their instructors …… 200
5.6 Comparison between the findings in this study and other studies' ….. 203
5.7 Students' reasoning about simulations needs to be studied …… 209
5.8 Limitations and significance of study …… 212

References ……... 214

Appendices …… 221
Appendix I. Hypothesis test problems for students …… 221
Appendix II. Diagnostic problems …….. 222

CHAPTER 1: INTRODUCTION

The introductory statistics course, usually called Elementary Statistics, has become a pivotal course in many community college students' lives. Most students enter the community college with the goal of transferring to a four-year college or university, and many departments at those educational institutions require that transfer students take a statistics course prior to upper division entry. Even when a student's transfer department does not require statistics, the transfer student often uses statistics to satisfy the four-year institution's quantitative reasoning requirement. Because of the increasing student demand for statistics courses, those courses are now one of the most ubiquitous mathematics offerings at the community college level.

Hypothesis testing is usually taught as the last third of the Elementary Statistics course. Hypothesis testing therefore in some sense becomes the goal for the course. The central place that hypothesis testing has in the statistics courses makes sense if you consider how students might use their knowledge later in their academic career or as informed citizens.

(1) They might read research reports in their upper division classes, for example in sociology or psychology, and need to interpret the results or read the results with a critical stance.

(2) They will read reports in newspapers or magazines, or hear reports on radio and TV, and need to interpret the results or to read those results with a critical stance.

(3) They will take a more advanced statistics course for which the Elementary Statistics course will serve as a foundation.

(4) They will carry out some small study as part of their upper division or graduate work.

For each of those purposes the students need to have a rudimentary understanding of inferential statistics. Competence in carrying out computational procedures is not going to be sufficient. Recognizing the student need to understand inferential statistics, most introductory statistics books end with a treatment of hypothesis tests. Hypothesis tests naturally become the goal of the Elementary Statistics course, and research on student understanding of hypothesis tests becomes an important research subject in statistics education.

1.1 Overview of Problem

Two statements, when juxtaposed, coming from the two major review articles on statistics learning research, put the spotlight on the situation of the Elementary Statistics student: Garfield and Ahlgren (1988) write,

> Over the past 20 years, most of the literature on teaching stochastics has been at the college level. This literature has been filled with comments by instructors about students not attaining an adequate understanding of basic statistical concepts and not being able to solve applied statistical problems (Duchastel, 1974; Joliff, 1976; Kalton, 1973; Urcuhart, 1971). The experience of most members in education and the social sciences is that a large proportion of university students in introductory statistics courses do not understand many of the concepts they are studying (p. 46).

and Shaughnessy (1992) writes,

> Most of the courses in probability and statistics that are offered at the university level continue to be either rule–bound recipe-type courses for calculating statistics, or overly mathematized introductions to statistical

> probability that were the norm a decade ago. Thus, college level students with all their prior beliefs and conceptual misunderstanding about stochastics, rarely get the opportunity to improve their statistical intuition or to see the applicability of the subject as undergraduates. University courses may, therefore only make a bad situation worse, by masking conceptual and psychological complexities of the subject (p. 466).

Thus, on the one hand, statistics instructors have for a long time complained that students do not understand the deeper meaning of statistics. On the other hand, available curriculum materials have encouraged teaching introductory statistics courses in a way that prevents such deeper understanding to develop.

Some recent developments might open the way for better teaching and learning of introductory statistics courses such as Elementary Statistics. The use of handheld calculators such as the TI-83, with its built-in statistical functions, is dramatically reducing the time students need to spend on routine computations. The TI-83 calculator will, for example, compute the p-value for all the hypothesis tests normally taught in an introductory statistics course.

The use of hands-on simulations and more frequent use of real life data to build statistical and probabilistic intuition are other important new directions in introductory statistics curriculum development. *Workshop Statistics* (Rossman, Chance & von Oehsen, 2002) by Key Curriculum Press provides an example of using hands-on simulations and real data to build students' statistical understanding. Some students participating in my Master's project were taught from the Minitab version of *Workshop Statistics* (Rossman & Chance, 2000). I later taught from the TI-83 version of *Workshop Statistics* (Rossman et al., 2002) during two subsequent semesters in my

Elementary Statistics courses at a community college. Then I found that students in *Workshop Statistics* classes were more at ease when talking about statistical concepts than were students, whom I had taught from traditional textbooks. The *Workshop Statistics* curriculum encourages writing directly in the textbook, which might help student verbalize their statistical work.

Observing students writing and talking about their statistical work made it more transparent to me that the majority of introductory statistics students are in the process of gaining statistical understanding. In each class, there might be a few students who do not at all understand, for example, hypothesis testing. There also might be one or more students, who have a solid understanding of the concept. However, most students can be found on a continuum between those two extremes. Students who have listened to the same lectures, read the same textbook and been assigned the same homework, have widely varying understanding of the material covered in those lectures and in that textbook. Moreover, the same student can, even within the confines of one test, show a varying degree of understanding. However, there is very little research that details *what* introductory statistics students understand and do not understand. This study will contribute to building a knowledge base of introductory statistics students' understanding. If statistics teaching is to improve, more knowledge is needed about how students think and reason about concepts such as hypothesis testing at different stages of their learning process.

1.2 Research Questions

For my Master's project (Aquilonius, 2002), which also served as a pilot study for my dissertation, I videotaped pairs of community college students solving statistical problems. The focus of that study was on peer interactions in statistics problem solving. The mechanisms through which stronger and weaker students supported each other were analyzed. Though the purpose of the Master's project was to study peer interactions, its corpus of data also raised issues about student reasoning about hypothesis testing, a topic that I already was considering as a dissertation topic.

When searching data bases such as ERIC and Psychinfo, I discovered that very little research had been done concerning students' reasoning about hypothesis testing. The only article, which directly treated the subject, was Falk's (1986) article about misconceptions of statistical significance. So, there was a lack of knowledge regarding students' reasoning about hypothesis testing. Thus there was a need for a study like mine that more comprehensively looked at students' reasoning about hypothesis testing.

1.2.1 How do students reason about the concepts of sample and population in the context of hypothesis testing?

Although research on students' reasoning about statistical hypothesis testing was lacking, there existed probability education studies informing me about issues that would be important to consider in my study. For example, Kahneman and Tversky (1982) described a heuristics that they called *representativeness.* The participants in Kahneman and Tversky's studies used *representativeness* to estimate

probabilities in ways that led to results conflicting with normative probability computations. Sampling is a central concept in hypothesis testing. Students in introductory statistics classes might look at representativeness the same way as Kahneman and Tversky's subjects did. If so, students' misunderstanding of representativeness in samples could negatively affect their reasoning about hypothesis testing.

Similarly, Konold's (1989) study about students' *outcome approach* to probability uncovered non-normative student ways of looking at probability that might interfere with students' forming of appropriate concepts concerning sampling. Such results from probability education research suggested a research question of students' reasoning about the concepts of sample and population in the context of statistical hypothesis testing.

When the concepts of population and sample are first introduced in introductory statistics courses, students do not seem to have much of a problem with the concepts. Students are good at reciting definitions as “In statistics, we use the term **population** to refer to the *entire* group of people or objects about which information is desired. … A **sample** is a (typically small) *part* of the population ….The essential idea of sampling is to learn about the whole by studying a part.” (Rossman, Chance & von Oehsen, 2002, p. 249). Students also seem proficient in distinguishing what constitutes the population and what constitutes the sample in examples given to them. Still, in the context of hypothesis testing, the concepts often seem blurred.

In my statistics classes, I usually give out three mini-projects during a semester. In those projects, the students are asked to design their own small studies, collect data and analyze the data. For the last project I ask the students to think about a question for which they can collect data and for which a hypothesis test would be appropriate. The students are asked if they think the samples that they collected were random, and I hint to them that the samples probably were not random. Then the students are asked how they would have collected a random sample, if they had had the time and money. Many of the students will answer the latter question in a way that indicates that they would attempt to collect information from the whole population, or at least a large part of it. By interviewing students in my dissertation project, I wanted to find out to what degree students understand the power of the statistical theory that allows one to use a rather small random sample to draw conclusions about a rather large population.

Another phenomenon that indicates confusion between sample and population appears when some students set up their hypotheses. An extreme case of this kind of mistake was committed by one of the students participating in my Master's project. Initially the student substituted the sample means for the population mean symbols in the null hypothesis. When he saw his two unequal numbers written as equal to each other on the board, he realized that what he had written did not make sense. Still he spent a substantial amount of time trying to find a way out of his dilemma.

To make a one-sample-test problem into a two-sample-test problem is a much more common mistake. The conventional statistical notations require regular English

letters to be used for sample quantities and Greek letters for population symbols. If a problem contains a sample mean to be tested against a population mean, students will often incorrectly state their null hypothesis as $\mu_1 = \mu_2$. This mistake in setting up their hypothesis suggested to me a confusion about sample versus population and so I decided to investigate this as part of the dissertation project.

1.2.2 How do students reason about hypothesis testing p-values?

The Central Limit Theorem is the theoretical basis for statistical hypothesis testing. Several researchers have written about students' poor understanding of the Central Limit Theorem (Kahneman & Tversky, 1982; Mendez, 1991; Well, Pollatsek & Boyce, 1990). However, there does not to seem to exist any studies concerning how students' lack of Central Limit Theorem understanding affects their reasoning about hypothesis testing. The Central Limit Theorem provides the p-values in hypothesis testing. Because of this central role that p-values play present applications of hypothesis tests, I decided to explore the meaning of p-values with the students in my study.

Present researchers often report statistically significant results by giving a range of p-values rather than basing their claims on test statistics values. For example, Webb (1991) used ranges of p-values to report results in her review article on task-related verbal interaction and mathematical learning in small groups. In her five summary tables on pages 370–376 she used different notations for results, for which $p <. 05$, $p < .01$ and $p < .001$.

Statistics textbooks are gradually following suit. For each new edition, the standard textbooks in introductory statistics give more room to treatments of the p-value approach to hypothesis testing. Of the two textbooks used by the students in the study, *Workshop Statistics* (Rossman et al., 2002) consistently instructs to use the p-value to draw conclusions, while *Understanding Statistics* (Brase & Brase, 2003) shows students the p-value approach *after* it introduces hypothesis testing with the test statistic being the basis for the conclusions.

It was already mentioned that students are much more likely to encounter p-values than test statistics in their future academic career, and as informed citizens. Teaching hypothesis testing using p-values also has pedagogical advantages. The p-value approach offers more of a unified approach than the test statistics approach. With the p-value approach the statistical decision always consists of comparing a p-value and an α–value. Calculators such as the TI-83, used by the students in this study, give the p-value to the user if the appropriate information is entered into the calculator, and no statistical tables are necessary.

With the p-value approach having a central place in introductory statistics courses there is a need to look closer at how students view the p-value in hypothesis testing. How good is their understanding of the role p-values play in answering hypothesis test problems? In instructions to the students in my Master's study I never told students to use any particular approach to make their statistical decisions. Therefore it is worth noting that, even though the four pairs of students came from three different classes, which used two different books, they all used p-values to make

their decisions. The students in my Master's project most often made the correct statistical decision from their computed p-values. Sometimes they would correctly answer a problem directly from the p-value without going through the intermediate process of rejecting or accepting the null hypothesis. Still, even the two strongest students in the Master's project each failed once at this process of answering the posed question using p-values. The weaker students seemed to show real conceptual difficulties in interpreting the p-value. Therefore this dissertation project included examining students' reasoning about hypothesis testing p-values.

1.2.3 How do students reason about answers to hypothesis test questions?

Statistics education literature (Garfield & Ahlgren, 1988; Mendez, 1991; Mevarech, 1983) reports that students often view statistics as a set of computation procedures without those procedures having much meaning for them. Such research results suggest that students might also work hypothesis test problems without much meaning. Hypothesis testing is usually taught as a procedure in introductory statistics classes. However, students are expected to answer the questions in their exercises in plain English to show that they understand what the hypothesis test results mean. Therefore I wanted to study how students discuss their answers to hypothesis test questions. It seemed that those student discussions would reveal student reasoning about hypothesis testing

From my teaching experience I knew that textbook problems involving hypothesis testing could lead to a variety of different student answers. For example, on a recent final exam I gave the following problem:

> During 1995, the average loan for purchasing a home in Greentown, California, was $235,000. The price of homes has increased since then. Using a level of significance of 0.01, test the hypothesis to determine if the average loan for purchasing a home has increased significantly. A random sample of 81 recent home loans produced an average loan of $265,000 with a standard deviation of $25,500.

Nineteen students took the final. Of those, six students wrote the correct answer: "Reject the null hypothesis" and "The average home loan has increased significantly".

Three students were "close" with answers like: "It could be that the average loan for purchasing a home has increased significantly" and "The price of homes has increased since 1995."

One student wrote: "The price level for a house has not raised significantly in California."

Three other students wrote that the loans had not increased even though the students had earlier rejected their null hypothesis. For example, one of them wrote, "I reject the fact that the average loan for purchasing a home has increased significantly". Two students seemed to have made computational mistakes and consequently failed to reject the null hypothesis, but still claimed that the home loans had increased. To have quite varied answers, such as those quoted, is not uncommon on statistics tests.

Also, in my Master's project, the student discussions about hypothesis problem answers seemed particularly informative regarding student thinking and understanding. At the end of solving a statistical hypothesis problem, students often failed to pay appropriate attention to how they had set up their hypotheses in the beginning of the problem. This failure to go back and fully comprehend what they were rejecting or not rejecting made the process of stating the final answer very hard. At other times it seemed that the students might have stated the hypotheses in a mechanical way using rules of thumb as: "If the problem has the word greater in it, H_a should have a $\geq$ sign in it." Again, by not having a clear understanding of the symbolic expressions in their hypotheses it became a difficult task to answer the questions asked in the problem.

Whatever the reason was for the lengthy discussions that students often had at the end of solving their statistics problems, those discussions offered me some insight into student thinking. Therefore, I expected the research question regarding students' reasoning about answering hypothesis test questions to be interesting and useful.

To summarize, my interest in students' reasoning about hypothesis testing originated with my teaching experiences and was enforced by my Master's project (Aquilonius, 2002). However research regarding introductory students' hypothesis testing was lacking. There seemed to be a real need of research in this area that I wanted to help fill. Through my teaching experience and from reading the research studies mentioned above, three foci for my research emerged. The three research questions were: (a) How do students reason about the concepts of sample and

population in the context of hypothesis testing? (b) How do students reason about p-values in hypothesis testing? (c) How do students reason about answers to hypothesis testing questions?

CHAPTER 2: LITERATURE REVIEW

2.1 A Model for Statistical Reasoning

In 2002 Garfield presented a model for statistical reasoning. Her goal for students seems more far-reaching than what one can expect to achieve in a one-semester introductory statistics course. Still, the stages of her model are quite relevant to this study. Her definition of statistical reasoning is also usable for this study, "the way people reason with statistical ideas and make sense of statistical information …. Underlying this reasoning is a conceptual understanding of important ideas, such as distribution, center, spread, association, uncertainty, randomness, and sampling" (Garfield, 2002, ¶1).

Garfield's (2002) model described a process of step-wise cognitive integration of statistical concepts, consisting of five levels. In the first stage a student might scramble statistical words and symbols with unrelated information. In the second stage a student can select or provide a correct definition but does not fully understand the concepts, and so on.

> **Level 1. Idiosyncratic reasoning.** The student knows some statistical words and symbols, uses them without fully understanding them, often incorrectly, and may scramble them with unrelated information.
>
> **Level 2. Verbal reasoning.** The student has a verbal understanding of some concepts, but cannot apply this to actual behavior. For example, the student can select or provide a correct definition but doesn't fully understand the concepts.
>
> **Level 3. Transitional reasoning.** The student is able to correctly identify one or two dimensions of a statistical process without fully integrating these

dimensions, such as that a larger sample size leads to a narrower confidence interval, or that a smaller standard error leads to a narrower confidence interval.

Level 4. Procedural reasoning The student is able to correctly identify the dimensions of a statistical concept or process but does not fully integrate them or understand the process. For example, the student knows that correlation does not imply causation but cannot fully explain why.

Level 5. Integrated process reasoning The student has a complete understanding of a statistical process, coordinates the rules and behavior. The student can explain the process in his or her own words with confidence. For example, a student can explain what a 95% confidence interval means in terms of the process of repeatedly sampling from a population. (¶ 4, Table 2)

In her article, Garfield (2002) claimed that most statistics students lack the integrated understanding needed to make [statistically based] correct judgments and interpretations (¶ 6). She used her model to describe how students reason about sampling distributions. This study examined how students reason about statistical hypothesis testing and will show some of the same lack of integrated understanding that Garfield found. However, her model might be too simple to catch the complexity of students' reasoning about statistical hypothesis testing. In particular, her model seemed to imply that students' statistical reasoning develops linearly, which might not be the case (Cf. Shaughnessy, 1992).

2.2 Probability Education Research Relevant to Research Questions

The theoretical basis for hypothesis testing is probability theory. However, probability education researchers have found that most students have a poor understanding of probability concepts, such as randomness and variation, key

concepts in hypothesis testing. As mentioned earlier, each year thousands of community college students take an introductory statistics course, which includes statistical hypothesis testing. Based on probability education research literature, those students are likely to have problems with some of the fundamental probability concepts underlying hypothesis testing. Since the research questions relate to probability concepts, the following sections will review relevant probability education research.

2.2.1 Misconceptions in hypothesis testing

Falk (1986) provided an illustration of how understanding hypothesis testing builds on understanding probability. She wrote about a common misconception in hypothesis testing. Students often confuse the conditional probabilities $P(H_o| R)$ and $P(R|H_o)$, where H_o stands for the event that the null hypothesis is true and R for the event of rejecting the null hypothesis. Even researchers sometimes fall prey to this misconception. $P(H_o| R)$ is the quantity that would be very helpful to know, because $P(H_o|R)$ would give the probability that the null hypothesis was true though it had been rejected. However, to actually find $P(H_o|R)$, one would need to do some Bayesian computations requiring quantities that are likely to be unknown.

The statistical decision rule is instead based on $P(R|H_o)$, denoted α □□□□□□□□□ the significance level. Usually the researcher selects $\alpha = .05$ or $\alpha = 0.01$. For example, in a one-sample test of a mean, one hypothesizes a value for the population mean μ. The Central Limit Theorem supplies a probability distribution

of sample means based on the null hypothesis being true, and on sample size and standard deviation. Only if the probability is less than α, of getting a sample mean $\bar{x}$ at least as extreme as the one actually observed, does one reject the null hypothesis. "Extreme" here means "in the direction of the alternative hypothesis".

In other words, one assumes that the null hypothesis is true. Then one decides on a level of significance α, which determines the size (and other statistical considerations, the location) of the region of potential sample results that will lead to the rejection of H_0. If R denotes the event "a sample result in the rejection region", then the statistical decision rule implies $P(R|H_o) = \alpha$.

Falk (1986) demonstrated with a numerical example how different $P(H_o| R)$ and $P(R|H_o)$ can be. Her example below can be used in an introductory statistics class.

Prepare 10 opaque urns, each containing 7 beads. There should be two types of urns:

- Nine urns of type A, each consisting of 5 white beads and 2 black beads.
- One urn of type B, consisting of 5 black beads and 2 white beads.

Randomly choose one urn of the ten.
Two complementary hypotheses are formulated with respect to that urn.
H_o: The urn is of type A. H_1: The urn is of type B.
The following decision rule is then applied: Blindly draw two beads (without replacement from the urn in question). If the two beads are black, reject H_o and accept H_1; otherwise H_0 cannot be rejected. (p. 88)

In this example $P(R|H_o) = 0.048$, which would be smaller than the common 0.05 level of significance. On the other hand $P(H_o|R) = 0.47$ - almost 10 times as large.

A more common sense example about speeding on the freeway demonstrates how large a discrepancy there can be between a conditional probability and its

inverse. Let S = the event that a driver speeds on a certain freeway and H = the event of getting stopped by the Highway Patrol for speeding on the same freeway. The P(H|S) = getting stopped by the Highway Patrol while speeding might be below 10%. However, P(S|H) = the probability that the driver was speeding when the Highway Patrol stopped him or her might be over 90%.

Of course, there is a great difference between Falk's (1986) example and introductory statistics problems. In her example, all relevant probabilities can easily be computed. In statistics problems, some of those probabilities will be unknown. Still, working through the urn problem would give introductory statistics students a sense of what the p-value and α stand for, and what those values are not. In this study the meaning that introductory students attach to p-values under current instructional practices was explored. However, the students in my study had not been required to take any course in probability before their statistics course. The two textbooks used by the students in my study had very short treatments of conditional probability, with one of the textbooks, *Understandable Statistics*, (Brase & Brase, 2003) only using one page for its treatment. Thus the study's students had received very little instruction in conditional probabilities and a lack of understanding of those probabilities could impair their reasoning regarding hypothesis testing.

Falk's (1986) main point was that researchers often misinterpret hypothesis testing and that the method does not answer the questions that the researchers really want to have answered. In addition, she maintained from her own teaching experience that hypothesis testing is confusing for students. She suggests some alternative

methods and discusses pros and cons for them. Her discussion of alternative methods is outside the scope of this study.

However, Falk (1986) realized that researchers will continue to use hypothesis testing, which means that it will continue to be taught in introductory statistics classes. Her discussion of the conditional probabilities involved in hypothesis testing theory sheds some light on why this theory is so hard for students. The ideas behind hypothesis testing might not be as straightforward as statistics instructors might like to believe. There are good reasons to look more closely at how students reason about some central concepts in hypothesis testing, and that is what this study did.

2.2.2 Belief in the law of small numbers

The first research question in this study concerns how the students reason about the concepts of sample and population. The relationship between sample and population was the topic for the article *Belief in the Law of Small Numbers* by Tversky and Kahneman (1971). The authors found that people regard a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics.

Tversky and Kahneman (1971) distributed questionnaires regarding random sampling to attendees at meetings of the Mathematical Psychology Group and meetings of the American Psychological Association. The respondents were called psychologists in the article. The psychologists were asked questions about hypothetical replication studies. The given answers indicated that the psychologists

believed that a significant result would easily be replicated, even when both the original sample and the replication sample were small.

Tversky and Kahneman's (1971) main point is that psychologists have "exaggerated confidence in the conclusions based on small samples" (p. 106). In particular, the authors pointed out how using such small samples lead to what they called ridiculously low statistical power. The authors also cited other studies arriving at the same conclusions.

Tversky and Kahneman (1971) found that psychologists had a belief in small samples that is not supported by the mathematical theory underlying hypothesis testing. The "law of large numbers" states that "in the long run, as the sample size increases and increases, the relative frequency of outcomes get closer and closer to the actual probability value" (Brase & Brase, 2003, p. 151). A statistics version of the law of large numbers was stated by Well, Pollatsek and Boyce (1990) as, "The 'law of large numbers' states that as sample size increases, statistics of a sample become less variable and more closely estimate the corresponding parameters of the population from which the sample was selected" (p.289). However, small samples will vary substantially in their characteristics from their parent populations just due to chance. For example, if one tosses a fair coin a large number of times, the proportion of heads will be close to 0.5. However, if one only tosses the coin a few times, the proportion of heads could be quite different from 0.5.

Tversky and Kahneman (1971) invented the term "The law of small numbers" as a parody on the real "law of large numbers". The purpose of this parody was to

draw attention to a tendency of many to inappropriately apply concepts that are true for large samples to small samples. The authors claimed that "naïve subjects" also believe in the “law of small numbers” the same way as the psychologists did in the authors' study.

Tversky and Kahneman (1971) concluded from their studies that many psychologists considered small random samples much more representative of the population than probability theory predicts. The authors' conclusion led to a related, and even more intriguing question: How representative of the populations do the introductory statistics students consider the random samples? It appears that if they have the same beliefs as the many of the psychologists, that the desired information about the population could be found directly from the random sample, without carrying out any hypothesis testing.

2.2.3 Understanding the variability of sample means

Tversky and Kahneman's (1971) findings were described in the preceding section, indicating that people have an unfounded belief in information coming from small samples. Well, Pollatsek and Boyce (1990) researched this issue in more detail and stated, "people understand that the means of larger samples are more likely to resemble the population mean but not the implications of this fact for the variability of the mean" (p. 289). Their claim was based on four experiments they did with undergraduate psychology students who had not previously taken a college statistics

course. All the experiments consisted of questions that probed students' understanding of how the sample size affect the variability of the sample mean.

Throughout the first three experiments, Well et al. (1990) successively refined their questions to probe deeper into the students' thinking. From student work, and interviews with the students, the researchers established that a majority of their subjects did not know the general principles guiding the sample size effects on the variability of the mean. The fourth experiment was an instructional intervention, in which the researchers tried to teach students about sampling distributions using computer simulations.

In the training session, students were first given step-by-step instructions to build a probability distribution on a computer screen of all the scores from a population with mean 69. Then, above the graph of the population scores on the computer screen, the students were instructed, again in great detail, how to build a sampling distribution from samples of size ten. By being instructed to compare the two graphs, students were made aware that the variability of the sample means was substantially less than the variability of the population scores. "The interviewer again made sure that subjects understood the difference between the upper and lower distributions and that all subjects noted that the sampling distribution of means was less variable than the population distribution that was displayed below it" (Well et al., 1990, p. 306).

Subjects were then asked questions involving predictions of what the probability distributions of the sample means would look like, if samples of size 100

were taken instead of size 10. To the researchers' surprise, 16 out of the 21 subjects thought that the variability of the mean would be about the same for the samples of size 100 as for samples of size 10 (Well et al., 1990). Even after having recent instruction in sampling distribution ideas, the students did not apply those ideas to the interview questions.

Although few subjects anticipated that the variability of the second sampling distribution would be much smaller than the first, when the computer generated the second distribution, they accepted the result. When asked why the variability of the second distribution was smaller, 18 of the 21 subjects provided explanations. The students' explanations for the decreased variability as sample size increased are summarized below, because they are important background information for this study.

> Nine subjects gave appropriate explanations such as swamping (an extreme score will affect the mean of a larger sample less than the mean of a small sample) or balancing (larger samples provide more opportunity for large and small scores to balance out). Six subjects indicated that the variability was less because “bigger samples are better” and three additional subjects indicated that the larger sample would be more like the population because it represented a greater proportion of the population. (pp. 308-309)

The research results by Well et al. (1990) described above are relevant to this study's first research question about the concepts of sample and population, as well as the second research question about how students reason about p-values in the context of hypothesis testing.

When performing their experiments, Well et al. (1990) divided their questions to students in two categories. The *accuracy* questions asked about probabilities

concerning sample means' proximity to the population mean. The *tail* questions asked how likely it was that the sample average exceeded a value that was at some distance from the population average. In all four experiments, subjects did significantly better on the accuracy questions than the tail questions. The authors thought that the accuracy versions of problems might be answered correctly more often because they map onto the heuristic that larger samples are more similar to the population. The tail questions are closely related to this research question how students reason about p-values in the context of hypothesis testing.

2.2.4 Making sense of randomness

As mentioned in the preceding section, most people do not expect the kind of variability that mathematical theory predicts for samples just due to chance. The word *random* is often used about events whose causes are due to chance. To apply statistical hypothesis testing, one needs *random* samples. Falk and Konold (1997) published an article regarding how people make sense of randomness. They state that, "although people feel that they know what they mean when speaking of *randomness* (Kac, 1983) and they communicate in everyday and professional affairs using their shared intuitive understanding of the term, it is one of the most elusive concepts in mathematics" (p. 301).

Falk and Konold (1997) did experiments to study people's subjective perceptions of randomness. They gave their subjects sequences consisting of 21 symbols, which were either Xs or Os. The subjects were to rate the relative

randomness of the sequences. In the next experiment other subjects were asked to copy the same sequences. The researchers found a high correlation between the first set of subjects' randomness ratings and the second set of subjects' difficulty in copying the sequences from memory. The easier a sequence was to encode, the less random the sequence was considered. In particular, sequences that contained long runs were considered less random than were sequences that contained many alterations. In this respect, the subjective perception of randomness was in conflict with mathematical theory.

Falk and Konold (1997) also gave two real life examples of this conflict between subjective perception of randomness and mathematical theory. Their first example concerned the so-called "hot hand" or streak shooting in basketball. Coaches and players often believe that once a player makes a basket, his chances of making the next shot increase. However, when massive records of individual players in real games were analyzed, the analysis showed that actual hits and misses were largely compatible with the output of the Bernoulli process.

Their second example concerned "luck" in gambling. Again, events that can be described as chance events according to probability theory are attributed to luck. Good (bad) luck is believed to produce longer streaks of wins (losses) than gamblers perceive as random.

There might not be a satisfactory way to define a general concept of randomness. However, in introductory statistics courses a random sample is defined as a sample in which each member of the population has the same probability of

being selected. Also, the sample members are to be selected independent of each other (e.g. Brase & Brase, 2003). Based on this definition there are distinct mathematical models such as the Central Limit Theorem, on which statistical hypothesis testing rests.

Pollatsek, Konold, Well, and Lima (1984) wrote:

> Presumably, an expert's fundamental conception of random variables and random sampling is a process model. Perhaps the most widely used model is the "urn-drawing", or "box" model, in which random sampling is viewed as isomorphic to the process of drawing labeled balls or slips of paper from an urn or box, replacing them, shaking well, and then drawing again. From this model, the idealization of which can be summarized by algebraic expressions, certain conclusions follow. (p. 396)

Introductory statistics courses will usually include activities illustrating this process model. Sometimes actual slips of papers will be used. At other times students will be instructed how to use random number tables or computer random number generators. At the college where this study was taking place, students are required to have TI-83 calculators. The TI-83 calculators have built-in random number generators, which are usually used for simulating random processes during class activities.

In spite of instructional practices to let students simulate random processes, some of the misconceptions reported by Falk and Konold (1997) could interfere with introductory statistics students' understanding of random sampling. Typically, such misconceptions will exhibit a lack of understanding how much samples could vary just by chance.

2.2.5 Representativeness

The word *representative* has appeared frequently in the literature review for this study. When discussing sampling, introductory statistics books emphasize that samples need to be representative of the population. For example, Rossman et al (2002) write, "If the sample is selected carefully, so it is *representative* [italics added] of the population, we will still gain very useful information about the population" (p. 249). Rossman et al. also gave students activities which include deciding if certain samples are representative or not (e.g. p.264).

The preceding quote by Rossman et al. seems to imply that there is a general agreement on what constitutes a representative sample. Unfortunately, such an agreement between statistical experts and more statistically naïve people does not always exist. In Kahneman and Tversky (1982) the term *representativeness* is used to denote a heuristic that often is at odds with normative ways to compute probabilities.

Kahneman and Tversky (1982) wrote, "An extensive experimental literature has been devoted to the question of how people perceive, process, and evaluate the probability of uncertain events…. Perhaps the most general conclusion, obtained from the numerous investigations, is that people do not follow the principles of probability theory in judging the likelihood of uncertain events…. [Instead] people replace the laws of chance by *heuristics* [italics added], which sometimes yield reasonable estimates and quite often do not" (p. 32).

Representativeness was one of those common probability assessment heuristics Kahneman and Tversky (1982) described in their writings. "A person who

follows this heuristic evaluates the probability of an uncertain event, or sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated" (p. 33). The authors described several experiments in which subjects appear to be using representativeness and arriving at non-normative results. Two examples follow below:

> Experiment 1. All families of six children in a city were surveyed. In 72 families the *exact order* of the boys and girls was G B G B B G. What is your estimate of the number of families in which the *exact order* of births was B G B B B B?
>
> Experiment 2. There are two programs in a high school. Boys are a majority (65%) in program A, and a minority (45%) in program B. There are an equal number of classes in each of the two programs. You enter a class at random, and observe that 55% of the students are boys. What is your best guess - does the class belong to the program A or program B? (p. 34)

In experiment 1 the two sequences were about equally likely. However, most subjects in the study considered the second sequence to be less likely, with a median of 30. The researchers interpreted this result as the first sequence being more representative, because of its proportions of boys and girls being closer to the population's 50-50 ratio.

Sixty-seven out of the 89 students in Kahneman and Tversky's (1982) study selected program A as being the most likely program, even though the entered class is slightly more likely to belong to program B. The researchers explained their results by saying that the class is more representative of program A, because both the entered class and program A have a majority of boys.

A third experiment produced results that illustrated the second characteristic of representativeness suggested by Kahneman and Tversky (1982); a person who follows the representativeness heuristic evaluates the probability of an uncertain event, or sample, by the degree to which it reflects the salient features of the process by which it is generated. In this third experiment, the subjects were given the following problem:

> On each round of a game, 20 marbles are distributed at random among five children: Alan, Ben, Carl, Dan, and Ed. Consider the following distributions:

	Allan	Ben	Carl	Dan	Ed
Type I distribution	4	4	5	4	3
Type II distribution	4	4	4	4	4

> In many rounds of this game, will there be more of the Type I or of Type II distribution? (pp. 35-36)

Thirty-six of 52 subjects, a significant majority, answered that there would be more of the Type I distribution results. The normative answer is that there would be more of the Type II distribution results. Kahneman and Tversky (1982) believed that the Type I distribution, in the eyes of the subjects, was more representative of the random process that were stated in the problem. The Type II distribution was too regular.

Kahneman and Tversky (1982) also described some experiments illustrating how their concept of representativeness plays a role in people's perception of sampling distributions. A typical such experiment is stated below:

> Distribution of the sexes. (Binomial, p = .50) Ss were told that approximately N babies were born every day in a certain region. For N = 1000, for instance, the question reads as follows:
> On what percentage of days will the number of boys among 1000 babies be as follows:
> Up to 50 boys
> 50 to 150 boys
> 150 to 250 boys
> …….
> 850 to 950 boys

> Note that the categories include all possibilities, so your answer should add up to about 100%. (pp. 38-39)

For N = 100, the categories were: up to 5, 5-25, etc. For N = 10, each category contained a single outcome, e.g. 6 boys.

Independent groups of probability-naïve subjects were assigned to predict the probability distributions for the different values of N. The researchers then plotted histograms with the percentages of boys on the horizontal axes and the median probabilities on the vertical axes. The histograms for the different sample sizes were close to identical. For Kahneman and Tversky (1982) this insensitivity to sample size showed the subjects' reliance on representativeness. If the sample was truly reflective of the population it would inherit the essential properties of the population, including variance. The subjects' heuristics in this case were in sharp contrast to the normative view of sampling distributions described by the Central Limit Theorem.

Examples such as the ones above indicate that people are not good judges of true representativeness. Of course, people's failure to intuitively decide what a good, representative sample is, is what prompts statisticians and researchers to use random samples. Random sampling has the purpose of sidestepping human biases by using mathematical techniques in selecting members for the sample and drawing conclusions about the data. In this study, students' reasoning about representativeness and randomness was examined, because those concepts are important in considering the relationship between sample and population.

2.2.6 The outcome approach to probability

Konold (1989) found that the representativeness heuristics described by Kahneman and Tversky (1982) could not account for a probability misconception that Konold found among undergraduate psychology students. He called the misconception the *outcome approach* to probability, and contrasted the students' approach with normative ways of looking at probability, such as the frequentist view. "To the frequentist, a probability is meaningful only with respect to some repeatable event and is defined as the relative frequency of occurrence of an event in an infinite (or very large) number of trials" (Konold, 1989, p. 62).

Konold's motivation for the 1989 study came from an earlier study of his, in which he had found that several participants had responded to probabilistic statements as if those statements were true with certainty. The students who participated in Konold's study were given three problems: the *weather problem,* the *misfortune problem* and the *bone problem.* In the weather problem, students were told that a forecaster had said that there was a 70% chance of rain. The researcher then explored the students' interpretations of the 70% chance of rain in the weather forecast with them. Some students expressed the normative frequentist view of probability. However, other students reasoned as if the goal in dealing with uncertainty was to predict the outcome of a single next trial. Konold termed their responses the *outcome approach.* The outcome approach might seem reasonable in an everyday context, such as if one wants to make plans for tomorrow based on the weather forecast. However, Konold noticed that students also used the outcome approach in the other two problems, which were more hypothetical in nature.

Konold (1989) reported that students having an outcome approach to probability often relied on causal explanations, an addition to considering the prediction of a single outcome the goal of probability statements. In the misfortune problem a person had several misfortunes happen the same day. In the interviews with sixteen students, Konold found, that

> Eight students gave other-than-chance explanations of the several low-probability events in the misfortune problem. Six students tried to embed all the events in a causal sequence so that each could be seen resulting directly from a preceding event. Five students relied on explanations that involved causal agents such as God or the stars. (p. 69)

Although no information was given in the problem that linked the events together, students linked the events together through a cause-effect relationship or some third underlying variable to make sense of the problem. A person with a frequentist view of probability would have given a different kind of answer. The frequentist would have considered all humans and a fairly long time period. Then the probability of all the misfortunes happening to somebody by chance might not be that small.

The topic for my study is how students reason about hypothesis testing. A researcher or statistician who uses hypothesis testing might very well want to build a causal argument. However, this causal argument cannot come from the hypothesis test or any other probabilistic reasoning. The causal argument has to come from a research model based on other considerations, such as theoretical constructs. The hypothesis test can only strengthen or weaken the research model.

If students in this study subscribe to the outcome approach to probability, this approach is likely to confuse their view of hypothesis testing. "As long as students believe there is some way they can 'know for sure' whether a particular hypothesis is correct, the better part of statistical logic and all of probability theory will evade them" (Konold, 1989 p. 92).

2.3 Understanding Versus Procedures in Statistics

In section 2.1, the definition of Garfield's (2002) level 4: Procedural reasoning was quoted as, " The student is able to correctly identify the dimensions of a statistical concept or process but does not fully integrate them or understand the process." Such procedural reasoning has been documented in statistical education studies. Of course, procedural reasoning in itself is not wrong. To the contrary, such reasoning is often necessary for intellectual expediency. However, when one lacks understanding of the underlying ideas for the procedures, those procedures might not be applied correctly. To exemplify such procedural reasoning without understanding, two studies that are relevant to this dissertation project will be reviewed. The two studies were carried out by Mevarech (1983) and Mendez (1991).

Mevarech (1983) did a study with 57 freshmen majoring in education and discovered that students would (incorrectly) assume that a closure property, similar to the one for addition and multiplication of real numbers, would apply to means. The students in his study had completed an introductory course, which emphasized descriptive statistics. The participants were given sets of solved statistics exercises.

They were told that some of those problems were solved correctly, while others had conceptual errors. The students were asked to mark solutions as correct or incorrect. The task was much like a True/ False test, except that the students taking it were asked to explain why they thought certain solutions were incorrect. Sixty-five percent of the students thought that a closure law held for calculating the overall mean. They thought that you could find the population mean by averaging the mean of sub-groups, comprising the population, even when the subgroups were not the same size.

In addition to applying the closure property incorrectly, about 80 percent mistakenly applied an associative law to calculate grade point average. And, approximately 30 percent thought that zero (0) was the identity element, i.e. when added to a set of scores, it would not change the mean. A similar set of misconceptions was identified for the computation of variances. Mevarech (1983) said, "The results supported the hypothesis that non-mathematically oriented students mistakenly conceptualized the operations of averaging numbers and calculating the variance as two binary operations satisfying the four laws of an additive group. This model seems quite plausible when one considers that a student might use the mean score and variance as ordinary numbers and forget that they are measures of central tendency and dispersion" (p. 419). To summarize, the students interpreted the words mean and variance as computational instructions as in "to compute the mean you add up all the numbers and divide with how many numbers you have" rather than concepts representing the average or spread of a set. Pollatsek, Lima and Well (1981)

established results very similar to Mevarech's findings in their experiments with undergraduate students.

Mendez (1991) studied students' understanding of the Central Limit Theorem (CLT). He extracted production rules as summarizing CLT from ten introductory statistics books and compared those rules with how students handled CLT problems. "The study provided an enactive statistical experience for the participants in such a way that talk aloud protocols probed by questioning could be elicited. The verbal data were the raw material used by the researcher to write condition-action rules for every informant and produce a CLT mental model characterization" (p. 13–14). Mendez' main finding was that most beginning statistics students failed to consider sample size in applying CLT to compute probabilities. Most students did not show understanding that the probability of a sample mean to be within a certain distance from the population mean will depend on sample size. To ignore sample size in this context is a grave conceptual error. Mendez' beginning students did, as did Mevarech's, treat a conceptual issue as if it only was procedural. Ironically, the conceptual mistakes in both cases led to computational mistakes that allowed the researcher to diagnose the misconceptions.

Mevarech (1983) and Mendez (1991) were able, through their research designs, to diagnose student misconceptions. However, in the classroom, students' lack of statistical understanding is often hidden from their instructors until the course reaches the topic of hypothesis testing. To consistently answer hypothesis test question correctly, students need to understand a range of statistical concepts.

Instructors will usually give a stepwise procedure for hypothesis test problems that students are asked to follow. Such a procedure is presented at the beginning of problem set 9.2 p. 474 in the textbook *Understanding Statistics* (Brase & Brase, 2003). One of the instructions to the students in the hypothesis testing problem template reads, "Explain your conclusion in the context of the problem." To be able to write such an explanation, and answer the question phrased in the problem, is of course the goal of the exercise. The way the students draw the conclusion in context of the problem will show their understanding, or lack thereof, of the hypothesis test concept.

However, when observing students in my classroom I have found that students often can work through all the steps in the hypothesis test template, except they fail to draw the final conclusion. When I was able to study student papers shared by my colleagues, it was noticeable how frequently students worked hypothesis test questions correctly to the last step, but misstated their conclusions. The students in my Master's project often held long discussions before they decided on their conclusions and also erred in a few. As with Mevarech's (1983) and Mendez's (1991) students, the hypothesis test solving students liked to turn hypothesis testing into a computational procedure. The critical issue here is that students somehow fail to attach appropriate meaning to their intellectual activity. This study was designed to explore in which ways students' reasoning reflected understanding of the hypothesis testing procedure.

2.4 What Can Be Learned From Mathematical Problem Solving Research?

Shaughnessy (1992) wrote, in his "Research in Probability and Statistics: Reflections and Directions," that "the current state of research in this area is far too eclectic to admit a complete synthesis" (p.466). It seems today, more than ten years later, that Shaughnessy's statement is still true. There is not enough knowledge about statistics teaching and learning to create a coherent theory. However, Elementary Statistics at the community college level is taught in the mathematics department as a course in mathematical problem solving. Mathematical problem solving has been the subject of substantial research, which has led to some theory building. Schoenfeld's (1985b, 1992) framework might be the most well known (e.g. Pressley & McCormick, 1995). Schoenfeld's framework was helpful for summarizing some of the literature relevant to this study. Schoenfeld's (1985b) original model had four main components: Resources, Heuristics, Control, and Belief Systems. In his 1992 article Schoenfeld (1992) added Practices.

Among the resources mentioned by Schoenfeld *intuition* seemed particularly important to consider in analyzing students statistical reasoning. Therefore intuition is treated separately in section 2.7. *Heuristics*, Schoenfeld's second component, played an important role in Polya's (1988) classical book on problem solving. There are a few studies showing how instruction in particular heuristic strategies improves students' statistics achievement. Hong and O'Neil (1992) reported about the benefit of instruction in using the heuristic "draw a figure" when solving hypothesis testing

problems. The two researchers wanted to help learners build relevant mental models for statistical hypothesis testing. Quilici (1997) showed that giving students schema training, using examples, helped the students to select the right procedure. An important feature of such schema training is to encourage students to use the "exploiting related problems" heuristic.

The literature on mathematical problem solving seems to agree that *control* or *metacognition* is an extremely important aspect of mathematical problem solving (e.g. Schoenfeld, 1987; Vye, Goldman, Voss, Hmelo, & Williams. 1997). Also, one of Polya's (1988) four main steps in problem solving is "Looking Back". Consequently metacognition is important in statistical problem solving too, and finding instances of metacognition in the students' statistical reasoning was part of the analysis in this study.

The studies reviewed in section 2.3 showed how students' *beliefs* about statistics influenced their reasoning about the subject. Reid and Petocz (2002) also reported that students' beliefs about statistics had important implications for the students' learning of statistics. The authors interviewed twenty students from a first-year statistics class and a third-year class in regression analysis. Based on transcripts from open-ended interviews with the students, the authors identified a hierarchy of six "conceptions":

1. Statistics is individual numerical activities.
2. Statistics is using individual statistical techniques.
3. Statistics is a collection of statistical techniques.

4. Statistics is the analysis and interpretation of data.

5. Statistics is a way of understanding real-life [situations] using different statistical models.

6. Statistics is an inclusive tool used to make sense of the world and develop personal meanings. (¶ 3)

Reid and Petocz (2002) grouped the six kinds of conceptions based on their foci. Conceptions 1, 2 and 3 were labeled as having their foci on techniques, conceptions 4 and 5 their foci on using data and conception 6 its focus on meaning. Introductory statistics courses usually place great emphasis on statistical techniques. For example, selecting the appropriate statistical technique for a given problem is a recurring theme in such courses. Therefore it could be expected that students in my study would mainly exhibit behaviors consistent with the first three conceptions. At the same time even introductory statistics instructors also encourage students to reflect over data. Most instructors will also aim for students to share the instructors' meaning behind the statistical techniques. Thus this study also tried to probe student reasoning beyond the techniques, searching for evidence of conceptions 4, 5 and 6.

Reid and Petocz's (2002) conceptions concern students' beliefs about statistics, while Garfield's (2002) model concerns students' statistical reasoning. Still the three researchers seem to share the same perspective on students' learning of statistics. For example, Reid and Petocz (2002), contrast "[the] different ways of understanding statistics [that] range from limiting to expansive views. We use the term 'limiting' to indicate that students who describe such views seem unable to describe any

characteristics of more *integrated* [italics added] and expansive views" (¶ 3). Similarly, Garfield (2002) repeatedly used the word "integrated" in her definitions of the different levels of statistical reasoning. Like Reid and Petocz, she seemed to measure the students' understanding of statistics in terms of how much the students had integrated different statistical concepts into coherent models. This view of what it means to learn statistics showed itself to be fruitful in the discussion of this study's data.

Schoenfeld (1985b) includes *beliefs about self* in his belief category. Low confidence in a student's ability to do statistics tends to create anxiety for that student. Gal and Ginsburg (1994) wrote that they found three Likert-type scales in the statistics education literature describing students' beliefs about statistics and measuring their levels of statistics anxiety. Oathout (1985) wrote that students' affective states in their statistics courses were dependent on students' experiences in earlier courses. Since those experiences often were negative, students often had a statistics anxiety that affected their class performances negatively. Students reported making careless mistakes on exams due to being anxious. Sutarso (1992b) also reported a significant negative correlation between students' course grades in statistics courses and the scores on her STATS (Students Attitude Toward Statistics) instrument (1992a).

The *instructional practices* are changing in introductory statistics courses. Garfield, Hogg, Schau and Whittinghill (2002) reported that graphing calculators were commonly used at 2-year colleges and so were small group activities and

student presentations. There are studies showing that the small group activities improve students' performance in introductory statistics classes compared with classes that only are conducted using a lecture format (Borresen, 1990; Bonsangue 1994; Giraud, 1997; Potthast, 1999). Some of the advantages of such small group activities are demonstrated in this study since the students worked in pairs. In particular the analysis contains statements regarding how working with a peer affects student reasoning regarding Schoenfeld's fourth category, *control.*

2.5 The Statistical Register and Students' Difficulties With Semantics

The concept of the so-called statistics *register* is helpful when analyzing students' statistical reasoning. Halliday (1978) defined a register as "a set of meanings that is appropriate to a particular function of language, together with the words and structures which express these meanings", and concluded that "we can refer to a 'mathematics register', in the sense of the meanings that belong to the language of mathematics (the mathematical use of natural language, that is: not mathematics itself)" (p. 195). Equivalently, this study refers to the statistical use of natural language with the term *statistics register*.

Pimm (1987) described when students encounter the mathematics register, the problem is not "just the use of technical terms, which can sound like jargon to the non-speaker, but also certain phrases and even characteristic modes of arguing" (p.76). He later continued, "Part of learning mathematics is learning to speak as a mathematician, that is, acquiring control over the mathematics register" (p.76). What

Pimm says about the mathematics register is equally true about the statistics register. In this context, statistics can be viewed as a special case of mathematics.

As mentioned by Pimm, part of learning a register is to learn the meaning of technical words. Examples of such technical words in statistics are terms like *histogram* and *quartile*. Even more commonly in statistics, words are borrowed from our natural language and given new meaning. A good example is the word *random*. If one uses natural language in an everyday way and says that one picks something randomly, then one means that one picks it in a haphazard way. In statistics, a random sample is a sample that is selected according to very strict mathematical rules. According to the definition of a random sample, every member of the population has to have the same probability of being selected.

The statistics register often uses two words from our natural language, and juxtaposes them in a way that gives the two-word combination a unique meaning for statistics. Examples of this kind of two-word juxtaposition are standard deviation, normal distribution, and hypothesis testing. All three juxtapositions have quantitative definitions in the statistics register that do not follow from the common sense meanings of the individual words contained in them. Part of the challenge for introductory statistics students is to become competent in the statistics register, a competence that is inseparable from mastering the new statistical concepts.

In the study mentioned in section 2.3, Mevarech (1983) found that, "many college students translate directly from the problem to an equation without due regard for the semantics of the problem" (p. 420). Quilici (1997) also found that statistics

novices were lacking in attention to the semantics of statistics problems. Compared to experts, novices were much more likely to sort statistics problems based on surface characteristics. Her experiments involved sorting statistical word problems (t-test, chi-square, and correlation). Her study provides insight into introductory statistics students' reasoning about how to select the right test for their hypothesis testing problems.

2.6 Sense Making and Construction of Meaning in Statistical Learning

Statistics students are not alone in their attempts to work their problems strictly computationally without regard to meaning. Students' tendency to favor calculation procedures at the cost of understanding has been widely documented in mathematical education research literature (e.g. Schoenfeld 1985b, Selden, Selden, & Mason, 1994). Calculation competency might suffice in many elementary school contexts. However in our increasingly quantitative society citizens benefit at least as much, and probably more, from the ability to interpret numbers than from the ability to calculate them. In addition, the common sense notion that it is easier to learn things if you understand them, has gained ground with mathematics teachers. Consequently a new emphasis on *understanding* mathematics has emerged from policy setting organizations such as the National Council of Teachers of Mathematics (NCTM, 1989, 1991, 2000; Pressley & McCormick, 1995).

The NCTM recommendations were also inspired by research results originated by Piaget and his collaborators (Inhelder & Piaget, 1958) showing that

knowledge could not simply be transferred from one individual to another, that it had to be constructed by the individual. Resnick (1993) expresses a more recent version of how this knowledge construction takes place: "The empiricist assumption that dominated many branches of psychology for decades, the assumption that what we know is a direct reflection of what we perceive in the physical world has largely disappeared. In its place is a view that most knowledge is an interpretation of experience, an interpretation based on schemas, often idiosyncratic at least in detail, that both enable and constrain individuals' sense making" (p. 1).

Schoenfeld (1992) refers to "an emerging body of literature … that conceives of mathematical learning as inherently social (as well as cognitive) activity, and an essentially *constructive* [italics added] activity instead of an absorptive one" (p. 340). The reform statistics movement is based on this assumption that mathematical (and statistical) knowledge has to be constructed by the individual student. Simulation has often been seen as the vehicle for this construction of statistical knowledge, particularly for knowledge related to the Central Limit Theorem, such as hypothesis testing (Gnanadesikan, Scheaffer, Watkins, & Witmer, 1997; Gourgey, 2000; Sterling & Grey, 1991; Webster & Ogden, 1998).

Three of the four books in the Quantitative Literacy Series (Gnanadesikan, Scheaffer & Swift, 1987; Landwehr, Swift & Watkins, 1987; Newman, Obremski & Scheaffer, 1987) are built around simulation activities. The Quantitative Literacy Series was created for grades 6 –10. Around 1990, curriculum writers realized that college students also would benefit from such activities. Gnanadesikan et al. (1997)

described a variety of hands-on activities, mainly simulations, that they and their colleagues used with success in their statistics classrooms. *Workshop Statistics* (Rossman et al., 2001), one of the textbooks used by the students in this study, is based on these kinds of activities to a large degree.

Gourgey (2000) presented a good example of what usually is called the reform approach to teaching inferential statistics. Her article stressed the need to use instructional activities that make statistical concepts meaningful to students, with simulations playing a central role. She said, "Even when students are able to solve conventional textbook problems, they might not understand the underlying concepts; … Therefore it is essential that students be introduced to important statistical concepts in a form that is intuitively meaningful to them" (¶ 1). Below is a description of the activities that she and her students did. Her study is of particular interest in the context of this study for at least two reasons. First, she used her simulation to discuss the issue of using sample proportions to estimate a population proportion. Second, she successfully used her curriculum with community college students in addition to undergraduate university business majors.

Gourgey (2000) chose to use information from the United States Senate Race in New York State, which was at the time between First Lady Hillary Rodham Clinton and New York City Major Rudolph Guiliani. The race was current, highly publicized, historically noteworthy and of interest to students as residents of New York City with knowledge and strong feelings about both candidates (¶ 8). Before the students started the simulation activity Gourgey related the fact that some polls at the

time showed Clinton having a slight lead, while another poll showed Guiliani leading. She phrased the question: “How can polls taken around the same time, all using random sampling, suggest different conclusions about who is leading?” (¶ 9). Her question led to the introduction of the “Margin of Error" concept.

Gourgey (2000) had the students assume, based on a real poll, that the *population* percentage of voters favoring Hillary Clinton was 48%. The students then carried out a simulation activity by pulling tags from boxes containing 48 tags marked “yes’ and 52 tags marked “no”. The students were told that they were going to become pollsters. In four-member teams, the students computed the sample proportion of ten tag samples (¶ 10). At the next class session Gourgey distributed a frequency distribution chart and the class plotted a percentage polygon for the data. The simulation allowed for a discussion what happens when repeated samples are drawn from the same population; individual sample statistics do not always match the true population value, but vary and converge around it (¶ 11).

Gourgey (2000) stated that “previously [students] often had difficulty just stating a conclusion about whether to reject a null hypothesis”(¶ 14). To provide some insight into the improvements she observed in students' reasoning, Gourgey presented some exam questions that her students now were able to answer successfully, along with some of their responses. She pointed out that much less frequently than before she saw students who were able to compute a statistical formula, but unable to draw a conclusion from their results (¶ 18). In the context of this study it is worth noting that she seemed particularly happy with the community college students’ improvement,

both in terms of their performance and attitude (¶s 8 and 23). Her results suggest that activities designed to help students create statistical meanings for themselves do seem to pay off. However, she also reported that questions remain about student understanding of the Central limit theorem, and that she was "continuing to work on developing test questions that tap more deeply into students' understanding of the sampling distribution" (¶ 22). This study used a different approach in exploring the same questions as Gourgey. While she was looking at the final results in terms of answers to test questions, this study looked at the process through which the students arrive at those results.

Verkoeijen, Imbos, Van de Viel, Berger and Schmidt (2002) also tried to use reform statistics principles in their work with health sciences students. However, the authors did not feel successful with their educational experiment that they called the *Constructive Statistical Learning Environment*. Their study covered three instructional units, each teaching an inferential statistics topic, and each lasting for four weeks. Each instructional unit started with an introductory lecture on the topic to be covered and students were provided with an outline of the important aspects of the concept. After the lecture, a week was reserved for individual study of relevant chapters from the course book. In the second week, the students met in a two-hour tutorial group to discuss the studied literature under the guidance of a tutor. The tutor initiated the session by asking the group to collaboratively generate a summary of the topic. Then the tutor fulfilled a monitoring role and did not intervene unless it was strictly necessary. At the end of this meeting practical assignments were handed out

to the students. Students were given one week to use SPSS for solving *individually* [italics added] a set of problems, usually based on real-life data sets. In the third week, students met again to discuss the solutions to the practical problems with their tutor. Finally, in the fourth week the cycle ended with a lecture (¶ 3).

Before the final lecture was given, the students were given the opportunity to get additional explanations on poorly understood concepts. Also, the students were handed blank sheets and asked to write down everything they learned during the instructional cycle. The students had not anticipated this participation in a free recall study. The time constraint for turning in the free call protocols was 30 minutes, but the students took at most 10 minutes to write down everything they could remember of the subject matter (¶ 4).

> Verkoeijen et al. (2002) wrote as part of their discussion,
>
> > The qualitative analyses of the recall protocols showed disappointingly low levels of conceptual understanding. … Statistical terms and formulas took a disproportionately large share of the total recall score while interpretations and background knowledge were hardly mentioned. In addition, the examples of incorrectly recalled elements contained some serious misconceptions. For instance, the idea that the null hypothesis should be rejected in case of a p-value = , does not reflect a particularly good understanding of the subject matter (¶ 6).

The authors' lamentations that interpretations were hardly mentioned by their students are consistent with one of the themes of this study: students find statistical interpretations hard and avoid such if possible.

Three possible reasons for the disappointing results of the Verkoeijen et al. (2002) evaluation study seemed worth considering in the context of this study. First,

inferential statistics is hard for beginning statistics students as has been reported earlier in this proposal. Second, the authors' claim that the students' poor recall protocols appeared, "despite the integration of a well-designed collaborative learning task" (¶ 6), can be disputed. The students were participating in organized collaborative activities *only twice* during each four-week period. In particular, the stipulation that students should work *individually* on their assignments seems like a missed opportunity for collaborative learning. Along the same line, the authors claim, "the learning environment was largely compatible with other small group statistical learning environments such as those described by Giraud (1997) and Magel (1998)" (¶ 3). A reading of Giraud's and Magel's articles revealed that their students regularly worked twice a week in small groups for some time in their classes, but that a substantial time was devoted to lecturing every week too. Thus the format seems quite different from that of Giraud's and Magel's. This difference in instructional format could be part of the reason that Giraud and Magel achieved positive results with their small groups, while Verkoeijen et al. did not.

Verkoeijen et al. (2002) brought up the free recall assessment as a third possible reason for their disappointing results. The fact that students chose to spend 10 minutes of the 30 allowable minutes indicated to the authors that the students were not motivated to put much effort into the given assessment. There is also the issue of the students having no practice in using the free call assessment earlier in the course, a fact that likely also contributed to the sparse output. Considering the results from

the Verkoeijen et al. study, I made sure that tasks in my interviews were not too different from what students do in their regular statistics classrooms.

2.7 Intuition in Mathematics and Statistics

Fischbein (1987) defined "intuitive knowledge [as being] immediate knowledge; that is a form of cognition, which seems to present itself to a person as self-evident" (p.6). The author pointed out that one cannot doubt every little fact all the time. Therefore one decides unconsciously to take certain facts and ideas for granted. "Those ideas appear to be very robust as an effect of their being deeply rooted in the person's basic mental organization" (p. x). "The survival of intuitive components in scientific reasoning – historically and individually may then be explained by that profound necessity of human beings to rely in their reasoning upon certain, evident, trustworthy conceptions" (p.201).

Fischbein (1987) introduced two classification systems for intuitions, of which the second system, based on the *origins* [italics added] of intuitions is most relevant to this study. "According to this criterion one may distinguish *primary* and *secondary* intuitions…. Primary intuitions refer to those cognitive beliefs which develop in individuals independently of any systematic instruction as an effect of their personal experience" (p. 64).

"The category of *secondary intuitions* implies the assumption that new intuitions, with *no natural roots*, may be developed. Such intuitions are not produced by the natural, normal experience of an individual. Moreover, very often they

contradict the natural attitude towards the same question" (Fischbein, 1987, p. 68). Fischbein took as an example the Newtonian inertia principle that a body will continue its rectilinear constant motion if no force intervenes. "*Intuitively* [italics added], it is difficult to accept such an interpretation. If that interpretation can be transformed from a learned conception into a belief then we refer to it as a secondary intuition. Such a belief will never be acquired in the normal conditions of our terrestrial life" (p. 68).

It seems that a parallel can be drawn to my research question about the meaning of p-values. To statisticians and statistics instructors, it seems very intuitive that you should reject the null hypothesis if the computed p-value is very small. After all, the p-value measures the probability to obtain a particular sample mean value or a more extreme one assuming that the population has a certain mean in one-sample-test of the mean. However, this way of thinking comes from studying probability and statistics. Without those studies we would not have the intuitive meaning associated with the p-value. The students, who are just introduced to probabilistic thinking, on the other hand, are likely to have acquired rather varied levels of intuition associated with concepts such as p-values.

For students to build statistical intuition is closely related to the task of meaning construction that was mentioned as an important part of mathematics education in the preceding section. If rules and procedures are learned as students construct statistical meaning, then those rules and procedures will have a greater chance of staying with the students and be part of a conceptual framework that can be

called intuition. Some typical examples of this kind of article, stressing statistical intuition are: “Flipping Frisbees and Finding Flowers – Developing Statistical *Intuition*” (Wolfe, 1992) and “Statistics and *Intuition* for the Classroom” (Chatterjee & Hawkes, 1996). Other articles claim to use students’ intuition as a tool in teaching statistical techniques. An example of such an article is “An *Intuitive* Approach to Teaching Analysis of Variance” (Johnson, 1989).

So far in this literature review, statistical intuition has been treated as a positive factor in statistics learning. Unfortunately, leaning on one’s intuition might not always lead to a correct solution to a statistics problem. Cobb (1989) calls mathematical intuition "A Double-Edged Sword” in his review of Fishbein’s (1987) book. Often intuition is helpful in problem solving. Schoenfeld (1985b) lists intuition first, when he makes a list of resources in his framework for studying mathematical problem solving. Burton (1999) found, in his interviews with practicing mathematicians, that intuition was very important to them.

However, students’ primary intuitions become a problem when students extrapolate from primary intuitions into a domain where they do not apply. Such an example of incorrect extrapolation was found by Mevarech (1983), in the study described in section 2.3 about Calculation versus Understanding in Statistics. Mevarech’s students incorrectly applied a closure law to means. There is a primary closure intuition: “If Ann has 2 apples and Peter has 3 apples, then together they have 5 apples, not pears.” This primary closure intuition gets strengthened in formal instruction. For example when you add, subtract or multiply any two real numbers,

the result is another real number. Many of the students in Mevarech's study believed that if you computed the mean of two means you would also get a mean, which only works if the two original means come from sets of the same cardinality.

The primary intuition about closure works well in many contexts, even in school mathematics. However, the mean of two means is not necessarily a mean. In some situations it is necessary to suspend the primary intuition in order to start building a secondary intuition that is based on formal thinking.

Thus in building any theory regarding students' statistics problem solving one needs to keep in mind that students' intuitions might differ from the teacher's. The preceding example shows a dilemma for the statistics instructor. The teacher wants statistical problem solving to be a sense-making enterprise. Therefore the teacher appeals to students' intuitions. For example, in hypothesis problems it is often a good exercise to speculate with students which way the statistical decision is likely to go before starting any computations. However, sometimes the students extrapolate to intuitive concepts that are in conflict with the formal development of statistics. Then the resulting conceptions are called misconceptions. The students' use of intuition was considered in the context of my third research question regarding how students reason about answers to hypothesis questions.

In spite of the misconceptions such as the ones found by Falk (1986), the results of this study suggest that intuition is mostly a positive force for students learning statistical hypothesis testing. Well, Pollatsek and Boyce (1990) made an

astute observation that is consistent with this study about how naïve statistical intuitions differ from physics misconceptions.

> The situation seems to be different for statistical intuitions [than for physics intuitions]. The kind of heuristics that naïve subjects apply to statistical problems are not so much wrong as they are crude, inadequate versions of appropriate statistical ideas and in this domain it may be possible, as Bar-Hill (1984) puts it, "to educate intuition to make it more valid and useful". The educational implication of this distinction is that many statistical intuitions may not represent impediments that must be excised before effective learning can take place. Rather, they may represent opportunities for instruction: crude ideas that may be developed and refined. (p. 311)

The research implications of the quoted statement is that more research results are needed about students' statistical intuitions to inform educators about which "crude ideas" might be developed and refined. In this study, students were invited to solve statistics problems in a relaxed environment, in which they were encouraged to talk about their problem solving. Then they were interviewed about statistical concepts related to their work. Due to the ease the students felt in the research session, they made many spontaneous comments giving insight into their probabilistic and statistical thinking. The analysis of those comments provided research results that contribute to the literature on students' statistics understanding. In the long run, such research results might improve statistics instruction.

2.8 Some Method Considerations Grounded in Education Research Literature

In educational research, methodology is more than methods. Methodology includes the theory that drives the way the researcher collects the data as well as the

theory influencing interpretation of data (e.g. Moschkovich & Brenner, 2000). My theory about mathematics learning is social-constructivist. I believe that learning happens in the intersection between the social and the cognitive as presented by Resnick (1991). Vygotsky's (1962) experiments with children led to "the discovery that word meanings *evolve*" (p. 124, italics added). He found that "the relation of thought to word is not a thing but a process of continual movement back and forth from thought to word and from word to thought … Thought is not merely expressed in words; it comes into existence through them"(p. 125). My belief in the intimate two-way connection between words and thoughts described by Vygotsky lied beneath my methodology for collecting data and analyzing data.

Chance and Garfield (2002) wrote, "that little is known about, or has been published on, the methodology of statistics education research" (p. 39). The authors then called for methods that would help to "develop models of how students come to understand statistics" (p. 39). They pointed out that,

> variables interacting with the instructional environment such as instructor attitude, time of the day, resources available, and classroom culture may have dramatic effects on student achievement and attitude and cannot be controlled or even measured as in a laboratory setting… [Therefore] many traditional measurement techniques, such as standardized exams, final course grades, and student ratings, are not sufficient, especially when measuring student reasoning (p. 40).

The quoted statements parallel what mathematics education researchers such as Lester (1985) found earlier, when he made the astute observation that the foremost goal of having reliable and valid data in the quantitative sense has prevented "measuring any but most routine aspects of performance" (p. 53).

Chance and Garfield (2002) further related, “More and more investigators are replacing purely statistical procedures with the collection of rich, diverse data from multiple sources that document the situation being investigated and provide a scholarly account of the situation and/or the intervention. … [Among those methods] are videotaped clinical interviews [that] build on techniques by cognitive psychologists” (p.41). The authors then related an example of how they are using videotape analysis in their research to study students’ statistical understanding in much the same way as was done in this study.

Understanding is often stated as a goal for Mathematics teaching (NCTM, 1989, 1991, 2000; Pressley & McCormick, 1995). Similarly, statistics teachers often say that they want their students to *understand* the statistics. Characteristically Mendez (1991) called his dissertation “*Understanding* the Central Limit Theorem.” In his study, Mendez compared introductory statistics students’ conceptions with more advanced students. This method of comparing novices with experts resembles my comparisons of introductory statistics students’ conceptions with their teachers'. However, Mendez’ focus was on what the introductory statistics students did incorrectly. He made the customary assumption that understanding is to share the meaning of the expert. To help students share the meaning of the expert is likely the teacher's instructional goal. However, this study also valued other points of view. An attempt was done to “understand learners in their own terms and for highlighting the potential in what they [knew]” (Moschkovich & Brenner, 2000, p. 461*).* For example,

how did the intuition students brought to the research session help them in their problem solving?

The "Think-aloud Method" (e.g. Ericsson & Simon, 1984) has often been used in cognitive science when researchers have attempted to access study participants' thinking. Though this method might yield good results with participants that are already competent, it might not be as applicable to students who are learning new material. Students who are still learning need all their cognitive resources to deal with the new material. To talk about their thinking simultaneously with solving their problems is likely to interfere with their work.

Allwood's (1990) study provides an example from literature supporting the argument that individual talk aloud protocols might not be the best choice for studying students' statistical thinking. Allwood conducted a study in which students were given explicit instructions to justify their steps in solving statistics problems. She hypothesized that more problems would be solved correctly if the students were forced to justify their solutions. Forty students studying first year statistics were paid for their participation in the study. All subjects were asked to solve the two statistics problems, which had five and seven sub-steps, respectively. The twenty students in the experimental group were told explicitly, and reminded throughout their solution process, that they had to justify their steps. The students in the control group did not get such justification instructions. There was no significant difference in the number of correct choices made between the experimental group and the control group. One interpretation of her result is that any benefit the students might have gained from

thinking about *why* they were doing the steps, were cancelled out by having to talk about them continuously.

Schoenfeld (1985a) advocated use of student pairs in cognition research rather than individual talk-aloud protocols. He wrote, “In single-person ‘speak-aloud’ protocols, what appears is often the trace of a solution: One sees the results of decisions but gets little insight into how the decisions were made, what options were considered and rejected, etc. When students work together, discussions between them regarding what they should do next often bring those decisions and the reasons for them out in the open” (p.178).

The students in this study were encouraged to talk about their work and results, but were also allowed to work silently when they so chose. When students work together, they naturally switch back and forth between working silently, and talking about their work with the partner(s). There is much for a researcher to learn from those spontaneously occurring conversations, especially when the information from student conversations can be complemented by students’ written work and follow-up interviews with the students. The wealth of such information contained in student conversations became apparent already when I recorded my Master’s project data.

2.9 Summary

When I searched standard databases such as ERIC and Psychinfo, the search only produced three articles related to introductory statistics students learning

hypothesis testing. Those three studies, which were described earlier in this dissertation, are summarized below. From these summaries it can be seen that none of the studies were based on empirical research concerning students' reasoning about hypothesis testing in introductory statistics classes.

Hong and O'Neil (1992) showed that two instructional strategies were beneficial in teaching statistical hypothesis testing. The first strategy tested by the researchers consisted of presenting the ideas behind hypothesis testing before teaching the procedure. The second strategy was to teach students to graph the sampling distribution as part of their solution process. Both those strategies produced better student performance on statistical hypothesis testing exercises than if the strategies were not used. However, those instructional strategies seem to be in common use in introductory statistics classes, including the classes taught at my college. Therefore, although they are beneficial, they did not remove the difficulties so many students have in understanding hypothesis testing.

Quilici and Mayer (1996) studied the role that examples play for students learning to categorize statistics word problems. Some of the word problems in the Quilici and Mayer study were hypothesis test problems. Their study gave some insight into how students decide which statistical test to use when solving such problems. Since this research had been done, I decided not to include a question regarding how students select tests for hypothesis test questions in this study.

Falk's (1986) study was the most relevant one for this study on students' reasoning about hypothesis testing. She focused on a rather subtle misconception that

she had noticed among both researchers and students about the conditional probability involved in statistical hypothesis testing. This misconception, though theoretically important, concerned a very narrow area of the knowledge needed to master hypothesis testing. As a comparison, this dissertation with its three research questions treated a much wider set of knowledge issues related to students' reasoning about statistical hypothesis questions.

Due to the limited existing research on student understanding of hypothesis testing, the literature search had to be widened. This widening took two directions for two different purposes. First, mathematical and statistical problem solving literature, including Garfield's (2002) article on statistical reasoning, provided useful vocabulary and organizing tools as well as models for methods.

Second, probability education literature provided information about students' reasoning regarding concepts crucial for understanding hypothesis testing. Randomness and variability are key concepts in statistical hypothesis testing. Psychological researchers, in their experiments, found that most people, including many of their colleagues, had poor intuition regarding those key concepts. In addition, even when such persons had some statistical training, they often did not use this training when faced with having to make judgments under uncertainty (Shaughnessy, 1992). Other probability education studies described earlier suggested that concepts such as random sampling might not be well understood by students. Also, Falk's (1986) study and studies about students understanding of the Central Limit Theorem (Kahneman & Tversky, 1982; Mendez, 1991; Well et al., 1990)

suggested that p-values might not be well understood. If students have difficulties with the ideas of sampling and of p-values, then they could be expected to struggle with answering hypothesis questions. Studying how students discuss answers to hypothesis questions would likely shed some light over how students reason about statistical hypothesis testing. Thus, the research literature and my teaching experiences converged to the research questions: (a) How do students reason about the concepts of sample and population in the context of hypothesis testing? (b) How do students reason about p-values in hypothesis testing? (c) How do students reason about answers to hypothesis testing questions?

CHAPTER 3: METHODS

3.1 Research Design

The design of this study had the purpose of uncovering students' statistics reasoning a layer below what a statistics teacher sees in the classroom. The teacher will see the students' statistical problem solving, but will rarely know *why* the students do what they do. Are students mimicking the teachers' procedures? How much meaning do the students attach to their work? To explore such issues a qualitative research approach was used. "[This approach] views inquiry as an interactive process between the researcher and the participants, is both descriptive and analytical, and relies on people's words and observable behavior as primary data" (Marshall & Rossman 1999, p. 7). "Printed material and other artifacts are combined with observation and interview records in a process that is widely known as *triangulation*" (Lancy, 1993, p. 20). My study was designed to allow triangulation between what student wrote on their answer sheets, what they said in conversation with each other and in their interviews with me.

Quantitative research methods usually attempt to isolate a limited number of variables and their interactions with each other. In contrast, qualitative research is often is designed to maximize amount information about the topic under study. In this study, students from five different classes were participating. The students showed varying degree of statistical competence as measured by their final grades. They also exhibited different demographic characteristics. Their textbooks differed. However, the students were asked to solve the *same* problems and answer the *same* interview

questions. The purpose of this design was to look for patterns common among students' reasoning although the students had different characteristics.

The student pairs were videotaped while solving hypothesis test problems and answering questions related to my three research questions. By studying videotapes of those student conversations and interviews, I attempted to find out some of the students' underlying thought processes. The purpose of the instructor interviews and textbook analysis was to give background information about factors that shape student thinking. Instructional practices, including textbooks, have persistently been blamed for students' failure to grasp statistical ideas. Both textbooks used by the students in their classrooms study tried, in different ways, to stress meaning above formalism. The instructors in the study were well aware of recent reform efforts in statistics education and incorporated reform ideas in their teaching. How were those attempts to build students' understanding in their classes reflected in this study's student conversations and work? Again, by studying videotapes of student conversations and interviews in great detail (microanalysis), and comparing student data to textbook and instructor information, this study was designed to give some answers to how attempts to build students' understanding of hypothesis testing were reflected in students' reasoning.

3.2 Setting and Participants

The research project was conducted at a community college in Silicon Valley, where I am a mathematics instructor. Approximately 10,000 students were enrolled at the college. Most of the videotaping sessions and interviews took place in empty classrooms. The remaining sessions were conducted in a mathematics study center, which was not used by other students at the time.

At the community college where the study was conducted, the introductory statistics course was called *Elementary Statistics*. The Mathematics department at the college offered fifteen sections of Elementary Statistics the semester when this study took place. Almost all students who were taking Elementary Statistics did so to satisfy requirements set by the four-year colleges to which the students wanted to transfer.

All research sessions involved pairs of students for the purpose of soliciting desired verbal and written data. Six students volunteered with partners to form three pairs. The other ten students volunteered as individuals and I formed pairs from those students based on their scheduling preferences. When possible, I also considered who I thought would be most likely to be comfortable talking with whom. Several of my colleagues helped me recruit students from their statistics classes for the research project. Table 3.1 is an overview of the students in the study with some information about them.

The first column lists the students as pairs the way they participated in the study. When students are analyzed as a pair in the result section their names will be

combined with an ampersand sign. For example the pair Alex and Ben will be denoted Alex & Ben.

The table shows that eight participating students were in classes using the *Understandable Statistics* (Brase & Brase, 2003) textbook and eight students in classes using the *Workshop Statistics* (Rossman et al., 2002) textbook. The students were diverse in terms of age, ethnicity and final grades. Among the ethnicities represented were African-American, Asian, Latino and White. The students were also diverse in their pursuit of educational goals. Professional goals included journalist, lawyer, paramedic, forensic psychiatrist, anthropologist, and special education teacher.

The grades listed in the last column were given to students at the end of their Elementary Statistics classes. The first five pairs listed above received their grades approximately a month after this study was conducted. The three pairs listed last had already received their grades and were interviewed after they had completed their statistics course.

Instructors A and B, whose students comprised the majority of the study's participants, were interviewed. More details about the instructor interviews are provided in section 3.4.2. Four students came from my classes, coded as I in table 3.1. I wrote form letters to my past students after they had received their final statistics grades. The letters described this project and asked the students to volunteer by contacting me by email or phone. I recruited the other instructors' students by going into their statistics classes and asking for volunteers.

Table 3.1
Student participants

Name (pseudonym)	Textbook Used by Pair	Instructor	Approximate Age	Final Grade in Course
Alex	Understandable	A	20s	B
Ben	Statistics	A	20s	A
Cindy	Understandable	C	20s	A
Dana	Statistics	D	20s	A
Elena	Understandable	A	40s	C
Fran	Statistics	D	30s	A
Gus	Understandable	A	20s	B
Hal	Statistics	A	20s	C
Maria	Workshop Statistics	B	20s	B
Nancy		B	20s	B
Rose	Workshop Statistics	I	30s	A
Sylvia		I	40s	B
Tracy	Workshop Statistics	I	40s	A
Ursula		I	40s	B
Vera	Workshop Statistics	B	30s	B
Zoe		B	20s	C

Note. The letter A (or B) in the instructor column means that the student's instructor was coded as Instructor A (or B) in the result and discussion chapters. The letter I in the column means that I myself was the student's instructor. The letters C and D represent instructors who were not interviewed for the study.

The instructors who were interviewed for this study were experienced teachers with more than twenty years teaching experience. Therefore in the analysis of the students' reasoning they were considered experts in the sense of the expert-novice tradition of educational research literature. At the same time, instructors were expected to influence student reasoning. For example, how instructors reasoned about the concepts of sample and population was expected to influence student reasoning about those concepts. Based on instructor teachings, textbook book information and

other factors students formed their own ideas about the relationship the concepts of sample and population. Students similarly constructed their meanings of statistical concepts related to the other two research questions. As mentioned earlier, the analysis of student data did not focus exclusively on what the students did correctly or incorrectly, but also explored the processes that the students used to make sense of textbook and instructor information. The interviews with the instructors were therefore important as a complement to the textbook information also included in this dissertation.

However, most of the mistakes that the students made were difficult to trace back to their textbooks and classroom instruction. My study was designed to discover patterns of student reasoning across textbooks and instructors. To promote this goal of finding general patterns, students using one of two different textbooks and having one of five different instructors were included in the study. The drawback of such a design was that only a few students (in one case only one) had a particular textbook instructor combination. Therefore any connections between a student mistake and his or her textbook or instructor would be quite tentative. Also, each instructor only taught from one of the textbooks, adding to the difficulty of deciding if the instructor or the textbook might have influenced a particular way of student reasoning.

3.3 Procedures

During our first meeting students were asked to work problems that had been chosen to stimulate discussion in order to produce desired verbal data. The order of the problems was also a deliberate part of the research design. The first two problems were chosen as standard problems to put the students at ease so they would start talking about their work. Those problems also gave information about which procedures the students used for hypothesis testing. The four subsequent problems were not as straightforward and were chosen to bring out some of the issues that confuse introductory statistics students.

During our second meeting, I gave the students some statistics problems with incorrect solutions, so called "diagnostic problems". The solutions exhibited the kind of mistakes that I and my colleagues often see in our classes. The students were asked to find the mistakes and explain how the problems should have been set up or solved. The purpose was to elicit student reasoning regarding concepts that I as an instructor found being particularly difficult for students.

3. 4 Data Collection

The three main sources of data for this dissertation project were student interviews, instructor interviews and the students' textbooks. The student interviews were videotaped and the instructor interviews were audio-taped. Eight pairs of students and two of their instructors were included in the study. The data collection resulted in 26 videotapes and two audiotapes.

3.4.1 Student interviews

All students came to two research sessions, which were held one week apart. They were encouraged to bring their textbook, notes and calculator. During the first session the students were asked to solve typical hypothesis test problems (Appendix I). Students were given one problem at a time on a sheet of paper, which contained ample space to work out the problem. Those problem sheets are called answer sheets in the result chapter. Each answer sheet was collected before the students were given another problem. One pair (Rose & Sylvia) chose to do their work on separate notebook paper that they attached to the problem sheets. The students were encouraged to talk to each other about their work during the problem-solving session, and I made notes of topics from their conversations that I wanted to discuss further with the students.

After the students finished the problem-solving part of the session, I collected their written work and gave them a short break. During the break, I looked over their written work for material that might be fruitful to explore further with the students. After the break I first asked the students some theoretical questions. Then a conversation followed about the work that students had done during the first hour of the research session. I emphasized that the goal of our activities was for me to learn more about how students learn statistics, and not to test their knowledge or abilities.

Students were told that because their problem solving was part of a research project they would not be given any feedback during the first hour of the research

session. During the second hour they were allowed to ask questions about what they had done. Also, students were told that if during the first hour they arrived at a point where they could not proceed with the problem solving they would simply be given another problem.

During the second research session the students were given the diagnostic problems stated in Appendix II and asked to grade those problems. If they found mistakes, they were asked to explain to the fictitious problem solver why the problem was solved incorrectly and how it should be corrected. The last diagnostic problem concerned sampling procedures. When the students had finished discussing the last diagnostic question, I asked them some questions regarding sampling. After those questions their work on the diagnostic questions were discussed. For most pairs the second session lasted approximately one hour, while the first session lasted for two hours.

3.4.2 Instructor interviews

After the semester was over, I interviewed two of the students' instructors, one at a time. By waiting to talk with the instructors about my research questions until the semester's teaching was over, an attempt was made not to interfere with their normal way of teaching. The main purpose for collecting the instructor data was to help in the analysis of the student data. As mentioned earlier, the instructors were considered experts in the novice-expert tradition of educational research, and students' problem-solving and interview answers were compared with the instructors'.

The instructor interviews followed a similar format as the student interviews. The main difference was that instructors were only asked to work one complete problem, while the students were asked to do six complete problems. The instructors were first asked to work the Coin problem, then discuss the four diagnostic problems, and lastly to answer the same theoretical questions as the students. The instructors were asked to present the statistical concepts, as they would do in a class lecture. For the diagnostic questions, I played the role of a student complaining to the instructor about points taken off on a test and wanting an explanation why.

3.4.3 Textbooks

Half of the students in this study were in classes using *Understandable Statistics* (Brase & Brase, 2003) as their textbook. The other half of the students were in classes using *Workshop Statistics, Discovery with Data and the Graphing Calculator* (Rossman, Chance & von Oehsen, 2001).

The result chapter in this dissertation contains textbook analysis, which serves as a background for analyzing the students' reasoning. A description of how the textbooks led up to the topic of hypothesis testing is provided in section 4.2.1. For each research question, I discussed the textbooks' treatment before describing the students' reasoning. Chapter 9 (mainly sections 9.1-9.6 on pages 452-519) in *Understandable Statistics* provided the basis for my analysis concerning its treatment of statistical hypothesis testing. *Workshop Statistics*' Topic 21 and Topic 22 on pages

443-479 were the main sources for the textbook analysis concerning its treatment of statistical hypothesis testing.

3.5 Interview Questions

In this section there are a number of references to the research questions. To simplify those references the questions have been numbered as follows:

Research question #1: How do students reason about the concepts of sample and population in the context of hypothesis testing?

Research question #2: How do students reason about hypothesis testing p-values?

Research question #3: How do students reason about answers to hypothesis test problems?

3.5.1 First research session

The first hour of the first research session the students were asked to solve some typical hypothesis test problems. Some of the problems were taken from introductory statistics books and others from my old statistics tests. The problems are stated with their labels in Appendix I. In table 3.2 below is a table summary of the research purposes for the questions.

Table 3.2
Problems that students were asked to solve

Name	**Research Purpose**
Checkbook Problem	Can the students solve a straightforward hypothesis test problem regarding a *mean*? Related to research questions #1, #2, and #3.
Home Value Problem	Can the students solve a straightforward hypothesis test problem regarding a *proportion*? Related to research questions #1, #2, and #3.
Coin Problem	Can students identify the *population proportion*? (This has been difficult for students in the past). Related to Research Question #1.
Home Loan Problem	Can students identify the *population mean*? (This has been a difficult for students in the past, e.g. students have selected to treat this as a two-sample problem). Related to research Question #1.
Sugar Machine Problem	How do students deal differently with a problem that contains raw data comparing with the problems giving summary data? Related to research questions #1, #2, and #3.
Tranquilizer Problem	Can students correctly interpret the p-value? (The word "reduced' in this problem tends to confuse students and I have stopped using this problem on tests. Students will mistakenly treat this problem as a left-tailed test. However, as an in-class-problem it has generated some good class discussions.) Related to research Questions #2 and #3.

(The complete wording of the problems is in Appendix I.)

During the second hour of the first research session students were given the theoretical questions in table 3.3 below. Depending on students' initial answers, the students were also given some follow-up questions. In particular, most students were asked if they could define the alpha- and p-values in hypothesis testing and explain in their own word what those values mean to them.

Table 3.3
Questions to students and instructors regarding key issues

Theoretical Question	**Research Purpose**
#1: Why do you reject the null hypothesis, when $p < \alpha$?	Relates to research question #2
#2 Why is it not good to say, "Accept the null hypothesis" when $p \geq \alpha$?	Relates to research question #2
#3: How does knowing that you reject the null hypothesis help you to answer the question asked in the problem?	Relates to research question #3
#4: How does knowing that you failed to reject the null hypothesis help you to answer the question asked in the problem?	Relates to research question #3

3.5.2 Second research session

During the second research session the students were asked to correct incorrect problem solutions and explain their corrections. Those problems were called diagnostic problems following educational research naming conventions. The instructors were also asked to do those problems.

Table 3.4
Diagnostic questions to students and instructors

(The wording and proposed solutions to those problems are in Appendix II.)

Name	**Research Purpose**
Exercise problem	How well versed are students in notations related to hypothesis testing? Do they clearly distinguish symbols denoting sample and population characteristics? Related to Research Question #1
Jail Problem	Do students distinguish between *population mean* and *sample mean*? Related to Research Question #1
Poll Problem	Do students recognize the importance of hypothesis testing and not to jump to conclusions from a sample value? Related to Research Questions #1 and #3.
Gas Price Problem	Do students recognize the importance of hypothesis testing and that a good sample can yield almost as good information as the population? Related to Research Question #1

After the students (and instructors) completed the diagnostic questions, they were given the questions below. As with the theoretical questions during the first research sessions, the research participants' initial answers led to follow up questions from me.

Table 3.5
Sampling questions

Theoretical Question	**Research Purpose**
#1: Why do researchers and statisticians collect samples?	Relates to research question #1
#2: What kind of samples do researchers and statisticians need if they want to apply methods like hypothesis testing that you learned in class? Why?	Relates to research question #1

3.6 Analysis

This study is an exploratory study of students' statistical thinking in the qualitative research tradition. As such, the analysis was mainly data driven (e.g. Marshall & Rossman, 1999).

The method of analysis was microanalysis. I made a detailed analysis of what the students said in their conversations with each other as well as with me. Moschkovich's (1992) study of Algebra students was used as a model. She analyzed videotapes of student pairs working with linear equations and graphs using a computer-graphing program, and used microanalysis for her analysis of the students' conceptions and language. Similarly I used microanalysis in looking for answers to questions about students' statistical conceptions. I also analyzed their written work along with their taped conversations.

When analyzing the data, at least three resources were used extensively. First, the literature on mathematics and statistics learning provided structure to the analysis. Second, my teaching experiences and, third, my Master's project results provided hints regarding what to look for in the data.

The analysis was done in three phases.

Phase I:

The first phase consisted of organizing the data in *tables*. One set of such tables was the collection of extensive videotape logs. I made twenty-six such logs – one for each videotape. Although those logs were the most important basis for my

analysis, they were too voluminous to be included in the dissertation. However, Table 3.6 shows the organization and the type of content included in those logs. The headings in the tape logs were listed horizontally rather than the vertical way that they are in Table 3.6.

Table 3.6
Tape log headings and contents

Heading	Content
Time	The starting time or starting *and* ending time of a particular tape section
Student activities	Initially this column was to contain a summary of the content on a particular tape section. The column also came to include transcripts of student conversations pertinent to the research questions.
Problem/Question	Hypothesis test problem or diagnostic question students were working on, or question asked by me
RQ	To which of the three research questions, if any, the tape section was relevant
Cat	Coded A if the tape section was expected to be exceptionally helpful in the study, B if section probably was going to be helpful and C if the section might be helpful in the study

Another set of tables is included in Chapter 4, the result chapter. The information in those tables is concise and detailed, mostly originating from the students' answer sheets. Certain behaviors in terms of the students' problem solving procedures and their answers were categorized and counted. For example, Table 4.2 summarizes which students used graphical representations and geometrical arguments in their problem solving. Table 4.3, summarizing the main reasons why students arrived at incorrect answers, provides another example of such a table.

Phase II

Using the purpose of the study as a selection criteria, longer sections of the tapes than would fit in the comment column of Table 3.6 were selected and transcribed. Parts of the instructor interviews that were relevant to the research questions and student conversations were also transcribed.

Phase III:

The third phase consisted of describing themes and patterns that I had noticed to emerge during the viewing and transcription of the tapes. When I combined the table information from Phase I with transcript excerpts of student conversations and interviews, I was able to create a more comprehensive organization of the data. The themes and patterns emerging from this more comprehensive organization became headings for the result chapter's subsections. For example, the heading for section 4.1.1, *Students sometimes confused population and sample means* represented such a theme. I then analyzed the student reasoning in the context of such themes.

In this last phase III, I often went back and watched tape sections several times over again. Some of the student confusions that I had seen in my classes, I was now able to watch on videotapes many times. This opportunity of revisiting the data helped me better analyze the reasoning behind those confusions.

CHAPTER 4: RESULTS

4.0 Some Comments on the Students' Overall Performance in the Project.

The table below lists the six problems in the order that the students were asked to solve them. The check mark in the row of a problem and the column of a pair's initial denotes that the pair solved the problem correctly. The √* symbol means that one of the students in the pair had a completely correct solution, while the other student did not. More details will be provided in the section that treats the results regarding the third research question.

Table 4.1
Which students solved which problems correctly

	Understandable Statistics Textbook				Workshop Statistics Textbook			
Problem	Alex Ben	Cindy Dana	Elena Fran	Gus Hal	Maria Nancy	Rose Sylvia	Tracy Ursula	Vera Zoe
Checkbook		√				√		
Home Value	√	√	√			√		
Coin	√	√						
Home loan	√	√	√*	√ *			√	
Sugar Machine	√	√	√	√	√	√	√	√
Tranquilizer						√*	√	√
	Before Break Interview				After Break Interview			

Note. √ in the intersection between a column and a row means that the pair corresponding to that column answered the problem in that row correctly
√* in the intersection between a column and a row means that one of the students in the pair corresponding to that column answered the problem in that row correctly

The above table highlights at least three important facts. First, all eight pairs solved the Sugar Machine problem correctly. Thus all eight pairs knew enough about statistical hypothesis testing to solve such a basic problem.

The second fact concerns that the Winter break was in the middle of the time period that the student interviews were conducted. The pairs are listed in the order that they were interviewed. The five first pairs listed in the table were interviewed before the Winter break, and the three other pairs were interviewed after the Winter break. Thus the first five pairs were still enrolled in their statistics class, while the three later ones had completed their classes. The table shows that some students interviewed after the break performed as well as the ones before the break, and others as poorly. An interesting pattern emerges among the three pairs interviewed after the break. The longer it had been since they had their instruction, the more problems they had to work on, before they were able to do them correctly.

The pairs are listed above in the order they were interviewed. The first four pairs used *Understandable Statistics* as their textbook, while the last four pairs used *Workshop Statistics.* Thus a third fact shown in the table shows that some students using each of the two textbooks performed well in this project, while other students did not. Performance on the project problems was closely related to the final grades of the students. For example, the only pair (Cindy & Dana) in the project who worked five problems out of six correctly was also the only pair that consisted of two A students. On the other hand, the two pairs (Gus & Hal and Vera & Zoe) who were

only able to solve two of the problems correctly were composed of one B student and one C student.

As can be seen from the table Maria & Nancy was the only pair that had only one correctly worked out solution to the study's problems. This low performance of Maria & Nancy requires an explanation. The two textbooks cover their topics in slightly different order. In particular, *Workshop Statistics* treats hypothesis testing as its very last topic, while in *Understandable Statistics*, hypothesis testing is followed by correlation and regression theory. The poor performance of the pair Maria & Nancy could likely be explained by the fact that the pair's class had not completed the unit on hypothesis testing at the time of this project. Unfortunately, their instructor became ill and was absent at a very strategic time. Results from Maria's & Nancy's research sessions are still included, because they provide some insight into what students struggle with when they are in the middle of learning hypothesis testing.

As was mentioned above, Cindy and Dana both constituted the pair with the highest grades in their statistics classes and the pair exhibiting the strongest performance in the study's interviews. What distinguished the pair Cindy & Dana from the other pairs in the study? Three characteristics stand out. First, Cindy & Dana focused on the question in the problem, while most pairs paid more attention to the numerical information. For example, Cindy & Dana reread the question aloud several times. Second, they were more reflective of their work than most of the other students. They clearly expected statistics to make sense, often using the expression, "That makes sense." They also acted on this belief. Partly, they looked back

individually on their work, and partly they acted as instruments of metacognition for each other. That they were continuously giving each other feedback is worth particular attention, as they came from different classes and did not know each other before the research session. Third, they seemed to be the most fluent among the pairs in using the statistical register as a tool to organize their thought. For example, the use of the word "mu" for population mean and the word "x-bar" for sample mean was helpful to them in solving the study's problems.

4.1 Research question #1. How Do Students Reason About the Concepts of Sample and Population?

The relationship between sample and population is a complex one, but this relationship is also at the heart of statistical hypothesis testing. When statisticians or researchers are faced with one-sample hypothesis test situations they focus on comparing a sample mean with a hypothesized population mean. The statistician or researcher will first determine which value to hypothesize as the population mean and then locate or compute a sample mean for comparison. In contrast, when the students read the problems, they would try to identify the statistical quantities in the order they appeared in the question. For example, when they saw a number expressed as a percentage, they would decide if it was a level of significance or some kind of proportion. Often during this identification process both hypothesized means and sample means were merely spoken about as means.

As a consequence of not specifying which kind of mean was spoken of, it was not unusual for students in the study to confuse sample means and population means. In section 4.1.1 examples of the otherwise competent pair Cindy & Dana exhibiting such confusions are discussed. The same pair also had difficulties with the last problem in the study, the Tranquilizer problem. Most pairs in the study were not able to solve the Tranquilizer problem correctly. The pairs' difficulties to do so, seemed largely to be caused by the abstract character of the population mean and the sample mean in this problem. An analysis of several pairs' treatment of the Tranquilizer problem is presented in 4.1.2. In some problems, a few students placed sample means in the hypotheses, a phenomenon that is discussed in 4.1.3.

In section 4.1.4 the students' treatment of the Jail problem is discussed. The examples in section 4.1.4 demonstrate, in a different way from the earlier examples, that students often do not make clear for themselves the differences between sample and population. While the instructors immediately identified one of the means in the Jail problem to be a sample mean and the other a population mean, six of the eight student pairs did not. Section 4.1.5 describes how two students in different pairs were able to find the mistake in the Jail problem. One of the conversations analyzed in 4.1.5 suggests that some conceptual change from students' standard way of reasoning is needed for students to fully understand statistical hypothesis testing.

Sections 4.1.6 and 4.1.7 discusses students' responses to some general questions about samples and populations. The students were asked questions about *how* they would collect random samples and *why* one takes samples. Section 4.1.8

provides a summary regarding the students' reasoning about the concepts of sample and population.

4.1.1 Students sometimes confused population means and sample means

Every student in the study confused population means and sample means at some time during the problem solving sessions. In this section it will be shown how Cindy, though one of the most competent students in the study, also exhibited such confusions. However, the analysis of Cindy's & Dana's problem solving below also shows how Cindy improved her ability to distinguish between sample mean and population mean as the problem solving session proceeded.

The first problem that students were given was called the Checkbook problem.

> *Checkbook Problem*. In a discussion of the educational level of the American workforce, someone says, “The average young person can’t even balance a checkbook.” The NAEP survey includes a short test of quantitative skills, covering mainly basic arithmetic and the ability to apply it to realistic problems. The NAEP survey says that a score of 275 (out of 500) reflects the skill needed to balance a checkbook. An NAEP random sample of 840 young men (between 21 and 25 years) yielded a mean score of 272 with a standard deviation of 60. Is this sample result good evidence that the mean for all young men is less than 275?

Like the other student pairs, Cindy & Dana, spent a fair amount of time (6 minutes) before deciding which method to use for the problem. However, almost right away, Cindy reread the question: "Is this sample result good evidence that the mean for all young men is less than 275? " and added her preliminary answer, "I would say 'No', though I can't think of why, right off". Thus, Cindy showed some good intuition in determining what a likely answer might be to the question. Most likely she was using the fact that an average score of 272 is not very much less than 275.

Then Cindy & Dana discussed which kind of problem the Checkbook problem might be, much the same way the other pairs did. Dana looked over the programs in her TI-83 calculator to (in her own words): "see if there is something in my calculator that I can type the numbers in." Several of the participating pairs frequently used this way of trying to fit the given problem data into a calculator program in order to solve a problem.

As Dana kept entering numbers on her calculator, Cindy asked Dana if the latter was trying to compute the mean on her calculator. Dana answered that she wanted "to compare the *mean* to this" and pointed to something on the problem sheet. When Dana said mean, she most likely was referring to the sample mean. The students in the present study did, more often than not, talk about means without qualifiers. The students' habit not to specify whether talking about a sample mean or a population mean was in stark contrast to their instructors' way of talking. The instructors, in their interviews, usually would indicate if they were talking about a sample mean, a population mean or a hypothesized mean. All the problems in this study, and many real applications, involve comparing sample means (or sample proportions) with population means (or population proportions). Some of the students' confusion in their hypothesis test solving seemed to be caused by their not specifying what kind of mean they are talking about.

Contrary to most of the other participants, Cindy was able to focus on the importance of the question in the first problem. After looking in her notes and exchanging some ideas with her partner she said, "I think we are over-thinking this,

because the question just asks, "Is this sample result good evidence that the mean for all young men is less than 275?" Cindy & Dana pondered for another minute, after which I asked if it would help them if they were given a level of significance. The students said, "Sure" and they were given $\alpha = 0.05$.

After the students were given a significance level, they knew they were dealing with a hypothesis test. Dana said, "It could be a z-test" and Cindy agreed. The transcript below shows how both students initially places the sample value in the hypotheses, but how Dana catches their mistake almost right away.

Excerpt 4.1
Cindy and Dana discuss the Checkbook problem

Cindy: So, the mean is 272 and the alternative is… So the question is that *the mean is less than 272.* So it is a z-test.

Dana: But the question is 275, less than 275. Do you see?

Cindy: Oh, yeah (laughs embarrassingly).

Dana: That is what it is supposed to be, the skill required to balance the checkbook (reads form the problem).

Cindy: So should μ be equal 275?

Dana: I think so (Both students erase the sample mean of 272 in their hypotheses). Do you think so?

Cindy: Probably. But I am kind of stuck, still.

Dana: With our standard deviation is going to be 60.

Cindy: Yes, our standard deviation is 60.

Dana: And here it is. Our mean is 272. It goes in here. (Dana shows Cindy where she entered x-bar in her calculator).

Cindy: Oh yes! Alright! (sounding like something is coming back to her). Oh gosh, it has been a week since I have been doing these kinds of problems.

Dana: And our number is 840. Now we are on a roll, right?

Cindy: Wow, we are getting there!

Neither the word "sample", nor the word "population" was used by Cindy or Dana in this dialog. Still, Cindy & Dana were aware that there was only one sample in this problem. Other pairs tried to work the problem as a two-sample problem. However, Cindy first used the *sample mean* to set up her hypothesis and Dana went along. Not until Dana had reread the claim about the mean for all young men being less than 275 did she realize that they had confused the sample mean with the population mean. This kind of confusion between sample measurements and population measurements is typical in this study's data. Students were not clear about the distinct quality difference between sample and population. This lack of clarity suggests that the relationship between sample and population in the context of hypothesis testing is often an unformed concept for introductory statistics students.

Cindy again initially placed the sample value in the null hypothesis in the third problem, the Coin problem.

Coin Problem. You suspect that a certain coin, when tossed, favors heads. You toss it 50 times and find 31 heads. At the 0.05 significance level, does it favor heads or is it a fair coin?

Excerpt 4.2
Cindy and Dana discuss the Coin problem

Cindy & Dana read the problem and decided that it was a "proportion" problem

Dana: So we are going to do 31 over 50 to find the proportion.

Cindy: That's good

Dana: So that is 62 percent

Cindy: So the null is *p equal 62*

Dana rereads the problem and reflects over Cindy's suggestion

Cindy: Oh, wait

Dana: point five

Cindy (almost simultaneously): point five. (She erases the .62 on her answer sheet and replaces it with .5)

Dana: Yes, that way it is a fair coin.

The transcript shows that Cindy first wanted to use the sample measurement, this time a proportion, to set up the hypothesis. However, she realized her mistake faster in this problem than in the earlier checkbook problem.

In the fourth problem, the Home loan problem, Cindy put words to the kinds of quantities she had difficulty in distinguishing in the two earlier problems described above. In the excerpt below it can be seen how using the statistical register helps the students to clarify the concepts of sample and population.

Home Loan Problem. During 1995, the average loan for purchasing a home in Greentown, California, was $235,000. The price of homes has increased since then. Using a significance level of 0.01, test the hypothesis to determine if the average loan

for purchasing a home has increased significantly. A random sample of 81 recent home loans produced an average loan of $265,000 with a standard deviation of $25,500.

When setting up the hypothesis for the Home loan problem Cindy seemed to be talking to herself.

Excerpt 4.3
Cindy and Dana discuss the Home loan problem

Cindy: μ□□□□□Oh, wait! No, it's two hundred…no, wait (Dana giggles and rereads the problem silently). *μ is for the population and x-bar is for the*

(Cindy hesitates, and as she does so, Dana brings the conversation back to what kind of test they should use)

Cindy: I know it is a z-test, but I am trying to figure out what μis.

Dana: OK (rereads the problem).

Cindy: So, it would be 235,000

Dana (almost simultaneously): 235,000, yes, because it says that the prices of homes have increased since then, right?

Cindy: Yes.

Cindy: Determine if the average loan for purchasing a home has increased significantly (reads from the answer sheet). So μ is greater than

Dana: 265,000

Cindy: You put 265?

Dana: Yes.

Cindy: That's x-bar. Isn't it?

Dana: Yes. Sorry. It was just the one that was staring me in the face.

Cindy: Yes

Cindy's comment that "μ is for the population" helped her to set up her hypotheses correctly in the Home Loan problem. Her use of the word "x-bar" was also productive in sorting which information in the problem concerned the sample and which concerned a population. Most of the less successful students used the single word "mean", when referring to the sample mean. When Dana made the possibly careless mistake of writing the sample mean 265,000 in the alternative hypothesis instead of the population mean 235,000 Cindy corrected her immediately.

The fifth problem, the sugar machine problem, was worked correctly by Cindy & Dana with great efficiency. The pair completed the problem in four minutes. However, they spent 20 minutes on the tranquilizer problem, which was the last problem during the session. In the next section, there will be a description and analysis of how Cindy & Dana, and other student pairs, handled issues regarding sample versus population in the tranquilizer problem.

4.1.2 The abstract character of sample and population in the Tranquilizer problem caused students difficulties

> *Tranquilizer Problem.* In an experiment with a new tranquilizer, the pulse rates of 25 patients were taken before they were given the tranquilizer, then again five minutes after they were given the tranquilizer. Their pulse rates were found to be reduced on the average by 6.8 heart beats per minute with a standard deviation of 1.9. Using the 0.05 level of significance, what could we conclude about the claim that this tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute?

All the student pairs in the study took substantial time to decide which method to use on the tranquilizer problem. The words "before" and "after" steered students toward what *Workshop Statistics* calls "Matched Pairs Test" and *Understandable Statistics* calls "Tests Involving Paired Differences". The students' conception of the problem as a Matched Pairs situation is essentially correct. However, when the students did not realize that the mean and standard deviation of the paired differences were already computed for them then they became confused.

It is somewhat surprising that after working through five problems more and more fluently, the pair Cindy & Dana spent more than 20 minutes on the tranquilizer problem, without being completely successful. As mentioned above, several of the student pairs approached the tranquilizer problem as a matched pairs problem. Cindy & Dana also took that approach as can be seen in the transcript below.

Excerpt 4.4
Cindy and Dana discuss the Tranquilizer problem

Cindy: The differences problem? Do you think that that is what this is?

(They look through their notes)

Cindy: Differences of two means? There are no two means here. Is that what you mean?

Dana: I don't know. Let me find it real quick (looks through her notes).

It is on page 539.

(Both students look in their books.)

Dana: There it is, page 717. Before and after.

Cindy: Let me see. (She looks in her book.)

Dana: Differences

Cindy: Yes, we have done those. Do you think that this is what it is?

Dana: I don't know.

The "before and after" or matched pair approach to the tranquilizer problem made it difficult for students to set up their hypotheses. In their examples from class all such problems had tested for no difference or $\mu_d = 0$. If the tranquilizer problem was to be viewed as matched pair situation, then the hypothesized difference would be 7.5. Cindy & Dana initially wrote $\mu_d = 0$ and $\mu_d > 0$ in their hypotheses, modeling example problems from their class work. Only after they were unhappy with an extremely low p-value, did they backtrack and change their hypotheses. Unfortunately, their new hypotheses were not correct either.

Why did the tranquilizer problem cause trouble for several of the pairs, including the otherwise competent pair Cindy & Dana? One possible explanation could be that the sample and population in this problem were more abstract than in the preceding problems. As pointed out in the section on the statistical register, there is a difference between the concept of population in everyday language and in the statistical register. In everyday language the word "population" means a group of people or animals. In the statistics register, “a *population* can be thought of as a set of *measurements* [italics added] (or counts), either existing or conceptual. ….A sample is a subset of measurements from the population” (Brase & Brase, p. 336).

In the five first problems, the link between the objects and corresponding sample and population measurement was straightforward, as with the weights of the sugar bags in the sugar machine problem. However, in the tranquilizer problem, the sample and population measurements were *differences* in pulse values. The link between the patients and the differences in their pulse rate before and after administration of the tranquilizer is rather abstract. The students preferred to think more concretely about two sets of values, one set before and one set after the tranquilizer was given. However, the students were not given those set of values. Therefore, they had difficulties solving the problem.

Elena & Fran had similar difficulties with the tranquilizer problem as did Cindy& Dana. In the transcript below Elena & Fran also discuss sample mean versus population mean in a similar way that were used by Cindy & Dana.

Excerpt 4.5
Elena and Fran discuss the Tranquilizer problem

Elena: What kind of test is this? It's two-tests!

Fran: Yes.

Elena: Yes. On one we have a sample of 25 patients. Reduced … So the mean of the first sample is 6.8 with a standard deviation of 1.9 and the level of significance alpha is 0.05. What could we conclude about the claim that this tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute?

Fran: OK, so means …

Elena (interrupting): No, there is only one sample. I was wrong about that. Not two samples. Only 25 patients.

Fran: Yes, only 25 patients, so sample size, N is 25.

As was mentioned above, the students' first impulse was to look at this problem as if it involved two samples. However, they saw only one sample mean, one sample standard deviation and one sample size. Therefore they had to revise their first decision and contemplate a test involving only one mean. However, as can be seen in the transcript below, Elena was not able to change her mind right away.

Excerpt 4.5 Continued
Elena and Fran discuss the Tranquilizer problem

Elena: Oh, you know, we are doing two-tests. It is two-tests because first we did it. We took the pulse 5 minutes after the treatment with the tranquilizer. Ok? And then we want to say, to conclude the tranquilizer will reduce pulse rate. So when we are doing this we are going to test by how it is affecting the beats per minute, which is the mean? Right?

Fran: Hmm. It is a sample

Elena: Yes, it is a small sample, I agree with that.

Fran wanted to stay with the idea that there was only one sample and Elena tried to accommodate her. Fran then reread the question of the problem aloud, after which they summarize the information as follows.

Excerpt 4.5 Continued
Elena and Fran discuss the Tranquilizer problem

Elena: OK, so we are saying greater than 6.8

Fran: Yeah, 7.5

Elena's omission of saying *what* is greater than 6.8 is typical of the student data collected in this study. Often, the word "it" was used as a placeholder. This kind of omission caused the students difficulties in their problem solving. On many occasions, the omission of certain essential words also covered up lack of conceptual understanding. Yes, the statement that 7.5 is greater than 6.8 was true. However, this fact did not have any bearing on the solution process, and consequently did not move the students any closer to a solution. In fact, Elena & Fran stumbled around for quite some time in their conversation including using the sample mean 6.8 in their hypothesis. Then, Fran reread the problem, looked at her calculator screen and seemed to have an aha-experience, actually saying "aha", as can be seen the following transcript.

Excerpt 4.5 Continued
Elena and Fran discuss the Tranquilizer problem

Fran: aha

Elena: and the standard deviation. The first standard deviation was 1.9

Fran: this was the sample.

Fran: *Oh! Got it! Mean is now 7.5*, they are asking this. So μ is 7.5.

Elena: so you are saying?

Fran: the average is supposed to be different. This is the mean, x-bar, (points to her calculator display) the standard deviation is 1.9, and we *compare* to this 7.5

Elena: so for … That is what our hypotenuse [sic!] is

Fran: Yes. This tranquilizer will reduce the pulse rate on the average by 7.5.

Elena: So our null is that μ equals

Fran: Yes, equal to 7.5 (erases her 6.8)

Elena: and for the alternative

Fran: maybe not the 7.5, we don't know

Elena (reads the question aloud again): yes it says different. So we'll do it that way. So μ is different than 7.5.

Fran: The words they use are confusing a little bit. I think that is what it is asking.

In the beginning of this last exchange, Fran took out her calculator and entered the numerical information as the students were reading it off from the answer sheet. As Fran read off her calculator display, she realized the calculator also asks for μ, and she seems to gain clarity in what the problem is all about.

As often was seen in this study the calculator offered more than computational power. The calculator helped the students organize their thinking. Maybe the use of their calculator even helped them build conceptual structures, if they internalized the process they went through in their interaction with their calculator.

Another observation that can be made from the above excerpt concerns the statistical register. After Fran has entered the sample information in her calculator she said, "this was the sample". Her utterance of the word sample seems to prepare her

for realizing that she also needs to consider a population mean in order to solve the problem.

Gus & Hal also started to solve the tranquilizer problem thinking it was a two-sample problem. This student pair depended heavily on their calculators for their problem solving. In the tranquilizer problem their dependence on their calculators seemed useful in that it prevented them from spending as much time on finding a suitable test as other pairs did. Not finding a second standard deviation to enter into the two-sample tests on the calculator, they promptly settled for a t-test.

4.1.3 One pair placed different values in the null and alternative hypothesis

It was demonstrated earlier that students would sometimes place a sample value into a hypothesis. However, Hal made a different kind of mistake on the tranquilizer problem that seems to indicate a lacking understanding of the probability theory behind hypothesis testing. He wrote his null hypothesis as H_o 7.5 and H_1 6.8. Thus, he correctly placed the hypothesized population mean in the null hypothesis, even though the symbol for the population mean was missing. However, he incorrectly placed the sample mean in the alternative hypothesis.

Somehow this practice of having different values in the null hypothesis and the alternative hypothesis seems more of a serious mistake than using sample means in the hypothesis. When students place sample means in the hypotheses it can sometimes be attributed to careless reading of the problem, compounded by the bad habit of skipping the qualifier in front of the mean. On the other hand, when students,

like Hal, place different numbers in the null and alternative hypothesis it more clearly indicates a poor understanding of hypothesis testing. Hypothesis testing is based on a sampling distribution with the null hypothesis mean as its center. The alternative hypothesis is always concerned with that same sampling distribution. Never do you specify a particular value for the alternative hypothesis.

Hal also placed a different value in the alternative hypothesis from the one in his null hypothesis, in the other problems involving means. His repetitive mistakes show that the mistakes were not just careless ones, but indicate a lack in his understanding of hypothesis testing. His partner Gus also committed the same mistake on the first problem, the Checkbook problem. The Checkbook problem asked if a sample score of 272 (with a standard deviation of 60) from 840 young men provided good evidence that the mean for all young men is less than 275. Both Hal and Gus wrote their hypothesis as H_0: 275/500 and H_1: 272/500. Thus they used the sample value in their alternative hypothesis. Curiously, this student pair set up the hypotheses correctly for the Home Value problem. However, the Home Value problem involved a hypothesis about a proportion, which could have made a difference from the "mean" problems.

4.1.4 Most of the students were not able to detect that a one-sample problem presented to them had incorrectly been set up as a two-sample problem

During the second research session the student pairs were given diagnostic problems. Those were problems in which students were presented work done by

fictitious statistics students that contained mistakes. The participants in this study were to find and correct those mistakes. The student pairs in this study were also asked to explain to the fictitious students, why their work was wrong and why it should be corrected the way shown. The instructors, when given the same diagnostic tasks as the student pairs, worked the problems correctly with speed and efficiency. After all, they had a lot of practice in grading student papers. However, the first two of those same tasks presented serious challenges to the student participants. The second diagnostic problem was called the Jail problem.

> *Jail Problem.* A student is given the following problem:
> Pre 1990 records show that the average time in jail spent by a first time convicted burglar was 2.5 years. A random sample was taken to see if the average time increased in the 1990's. From the sample of 25 first time convicted burglars in the 1990's, the average length of time in jail was 3 years with a standard deviation of 9 years. Did the average length of jail time increase in the 1990's?
>
> The student sets up the hypotheses as:
> $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 < \mu_2$
> Do you think that is right? If you don't think it is right, please correct the student's work and verbally explain to the student why the way the student set up the hypothesis was wrong.

Both instructors corrected the hypotheses to $H_0 : \mu = 2.5$ and $H_a : \mu > 2.5$. In her explanation to the fictitious student, instructor A emphasized that the 2.5 years was a "historical claim" that "you want to test with your sample data." At least one of her students used the term "historical claim" in the student interviews, which suggests that the instructor used that expression as part of her instructional strategy. Instructor B emphasized that the problem only supplied information about *one* population and *one* sample.

The students were introduced to the diagnostic tasks, by being asked to "play the teacher", which intrigued and amused the participants in the study. Still, most of them found the Jail problem hard. The next two transcript excerpts show two unsuccessful pairs' conversations about the Jail problem. The first pair, Maria & Nancy, performed poorly in the problem solving session, while Tracy & Ursula performed much better. Still, their reactions to the Jail problem were very similar.

Excerpt 4.6
Maria and Nancy discussing the Jail problem

Nancy: Yes. No, wait. We want to know if it increased, that would be $m_1 > m_2$

Maria: Because that's the first one pre-1990s, that this is bigger than this (points to her answer sheet)

Nancy: Yes, you are right. But does the average time in jail increase? So there is nothing really wrong.

Maria (turns to me): Is there something wrong with all of these?

I: Yes.

Nancy: So then there is supposed to be $m_1 > m_2$, right?

Maria: No

Nancy: Is H_o always equal to?

Maria: Yes

Nancy: So then m_1 should be greater than m_2

Maria: Yes

Nancy: Because we are testing whether it went up or not.

The main point of showing this transcript is to show that the students were not able to correct the fictitious student's mistake of treating the jail problem as a problem involving two population means. However, Nancy's use of *it* in her last sentence is also worth noticing. By not specifying what went up, she avoids confronting the contradiction in her statement.

The next transcript gives an example of another pair, who was not able to find the mistake in the Jail problem set up. The more confident Tracy took the lead in the conversation, which might have prevented her partner Ursula from thinking through the problem enough to find the mistake.

Excerpt 4.7
Tracy and Ursula discussing the Jail problem

Tracy: I believe this is correct

Ursula reads through the problem carefully

I: It all has to do with the hypothesis. So you do not need to work out anything.

Ursula: So basically … OK

Tracy (turns to Ursula): You want me to explain it?

Ursula: OK

Tracy (showing Ursula her notes on her answer sheet) : Having mean1 being pre-1990s and mean2 being 1990s and of course the null hypothesis that would make the two equal (points to her paper) so the jail time in the 1990s

would be equal to the time in the pre-1990s and the alternative hypothesis would be, "did the jail time increase in the 1990s?' So this would be the pre-1990s , you know, mean1 less would be less than mean2, the 1990s would be the alternative hypothesis. *So I say it's correct.*

Ursula (seems not to be completely convinced): So basically you are saying that…(falls silent)

Tracy: That's what the question is, "Did the average jail time increase in the 1990s?' And this is the 1990s and this is the pre-1990s (points to information on the answer sheet.)

Ursula: Yes.

Tracy & Ursula, just like Maria & Nancy, did not find the mistake on the jail problem, though Ursula might have been able to do so, if she had been given time to work out the problem. A curious similarity between the two conversations above is that neither of the pairs used the American pronunciation of the Greek letter μ as mu. The μ_1 and μ_2 were spoken of as m1 and m2 by the first pair, and mean1 and mean2 by the second pair. Maybe, if they had spoken about the means as "mus", using the given Greek or foreign symbol, then the notation could have helped them to discover that one of the means in the problem was a population mean, while the other was a sample mean. Or, from a different perspective, the practice of using the Greek symbols for population means instead of words could be part of the difficulty for the students.

Cindy & Dana could not find the mistake on the jail problem either, though they seemed to be on the right track. Cindy said that μ was for "a population group of all the people who were in jail for that particular crime. That's why it is mu." Later she said, "OK, there is a sample of 25 first time convicted burglars in the 1990s". Still the pair decided to run a two-sample test. When they realized they did not have a second standard deviation, they did not see that as a hint on how to proceed, but just said that they were giving up.

4.1.5 Two pairs solved the Jail problem correctly

Only Alex & Ben and Rose & Sylvia were able to correct the fictitious student's mistake in the jail problem. Ben pointed out almost immediately that, "They were comparing μ_2. It should be greater than 2.5." To check if his conjecture was correct, Ben suggested that they "check if they are doing the right test, t-test, I think. Sample of 25." Alex agreed, "They give a mean and standard deviation." Ben ran the t-test and found the information given in the jail problem matched information asked for by the calculator's t-test program. Alex & Ben concluded as written on Ben's answer sheet, "The problem was a t-test, the hypotheses were set up incorrectly. The t-test only deals with μ_0, $\bar{x}$, σ, n, …There is no comparing to a μ_2. "

Rose & Sylvia were also able to diagnose what the mistake was in the jail problem, but it took them longer as can be seen in the transcript below. They also decided to work the problem out separately, after their introductory remarks, but before most of the conversation was taking place.

Excerpt 4.8
Rose's and Sylvia's discussion of the Jail problem

Sylvia: This is exactly opposite of that one, because we have a sample size and we have the mean. This is on those particular 25, which is exactly what you tried to tell me on the last one, right?

Rose: Yes

Sylvia: So this one is exactly the opposite. On this *one it needs to be x-bar* because it's *pertaining to this particular sample*, right?

When Sylvia spoke about this problem being opposite to the "last one", she referred to the first diagnostic problem that did indeed concern two populations. Thus she was already paying attention to the fact, at the beginning of the pairs' discussion, that one set of information in the jail problem "pertained to this particular sample."

Excerpt 4.8 Continued
Rose's and Sylvia's discussion of the Jail problem

Rose: Wait a minute. I need to read the problem. I find it easier to list all the information first and then set up your hypothesis. Don't look at those hypotheses there. Set up your own and see what is different.

Sylvia: OK.

Thus, the students separately worked out the problem. (The students were not expected to work out the diagnostic problems from beginning to end. The

methodological purpose for those problems was to focus students on particular concepts of interest to me. Still, the students often worked the complete problem, somewhat circumventing the methodological purpose, but being more successful in the process.) After Rose was done with her work she said, "OK" to signal that she was ready to talk to Sylvia again. She leaned over to look at Sylvia's work and Sylvia read off her listed information.

Excerpt 4.8 Continued
Rose's and Sylvia's discussion of the Jail problem

> Sylvia: And because the mean is pertaining to this particular sample, that is why it should be x-bar.
>
> Rose: I put it as a μ, because our μ is 3. μ is 2.5 because we want to find out if it is greater than the pre-1990s, so I put equal to because that's supposed to be your null hypothesis. And for the alternative I put μ is greater than 2.5. Do you see what I mean?

Sylvia kept stressing (correctly) that the 3 was an x-bar, an idea critical to understanding what was wrong with the Jail problem. Rose said (incorrectly) that it is a "mu". Rose contradicted Sylvia somewhat, but still made the appropriate correction to the fictitious student's hypothesis by writing "$H_0 : \mu = 2.5$ and $H_a : \mu > 2.5$". Thus she seemed to "solve" the jail problem correctly without being clear about the underlying ideas. Sylvia, on the other hand, who had initially not made the

appropriate corrections on her answer sheet, seemed to have an insight when she listened to Rose and looked at her work, as can be seen in the excerpt below.

Excerpt 4.8 Continued
Rose's and Sylvia's discussion of the Jail problem

Sylvia: Oh! Because *you can't compare these because they are apples and oranges* (referring to the sample mean of 3 and the hypothesized mean of 2.5). You need to set up the ones that you are testing against. I still think that this should be x-bar

Rose: I put it as mu

Sylvia: But I think it does not matter

Rose: Your mean is still 3 years. The average is your mean.

Sylvia: Yes, but the reason that is correct, because we would be comparing apples and oranges the way I was doing it. We can't do it because we are looking for an increase or decrease. We are not really comparing them equally. That is why I think that will be correct (points to Rose's correct work). You are seeing if there is an increase in mu versus if they are equal or different. You are looking for that particular population, so you are already given the original one.

Rose, in spite of solving the jail problem correctly, was still not distinguishing appropriately between sample and population means. In contrast, Sylvia seemed to

have experienced some conceptual change and grasped that there was a qualitative difference between sample means (apples) and population means (oranges).

4.1.6 All students recognized bias in a given sampling procedure, but their suggestions for doing random sampling varied in sophistication.

As described in the methodology section, the last research session ended with the students being asked some questions about the concepts of sample and population. Those questions were asked after the problem solving tasks, giving students the opportunity of using examples from the problem solving sessions in answering some of the more theoretical questions. As a transition to the more theoretical questions, the students were asked to analyze the two scenarios in the Gas Price problem.

> *Gas Price Problem.* Suppose a statistics instructor asked her students to compare gas prices in Santa Clara County with those in Santa Cruz County, and decide if there was a significant difference in gas prices between the two counties. Two students, working together on the project, decided on the following procedure: Since they were living and going to school in Santa Clara County they decided to get their sample from Santa Clara County by recording gas prices as they went about their business during the week. On the weekend, they would go over to the beach in Santa Cruz, and would record their sample gas prices on their way to and from their destination in Santa Cruz.
> (a) Would you consider the samples that the students collected to be random samples?
> (b) Suppose that a consumer organization would want to decide if there is a significant difference between gas prices in Santa Clara County and Santa Cruz County. Suppose the organization has quite a bit more resources in terms of money and time than the students have. How would you recommend that the consumer organization collect their random samples?

All the students answered "no" to part (a) of the question, stating that the fictitious students did not collect random samples. The most common reason students

gave was that the samples were not random because the data were collected on different days of the week. Students pointed out that prices are likely to go up in a beach town like Santa Cruz during weekends. Most of the students also said that the geographical areas covered would be seriously limited by the chosen method of data collection. The excerpt below shows a typical student conversation about part (a) of the gas price problem.

Excerpt 4.9
Alex and Ben discussing the Gas Price problem

Alex: They are not random samples.

Ben: Yes, they are not random. They are just going where they are living. What is that called? The easy thing. The easy collection.

Alex: They are doing different times

Ben: Yes. It might rise over the weekend or something. Yes, I don't think they are random. What is the technical term for that? That easy sampling. Convenience, right?

Alex: Yes

For the (b) part of the gas price problem, where the students were asked to design a procedure for collecting random samples, the answers were much more varied. Tracy stated, as was done in the students' textbooks, that "everybody in the population should have the same probability of getting chosen for it to be a random sample." Even though Tracy was the only student quoting the definition of random sampling, the other students' conversation reflected the same sentiment. However, the

suggested methods of achieving such a sample were quite varied. The interviewed *instructors*' suggestions reflected what Pollatsek, Konold, Well, and Lima (1984) wrote about as a process model:

> Presumably, an expert's fundamental conception of random variables and random sampling is a process model. Perhaps the most widely used model is the "urn-drawing,' or "box,' model, in which random sampling is viewed as isomorphic to the process of drawing labeled balls or slips of paper from an urn or box, replacing them, shaking well, and then drawing again. From this model, the idealization of which can be summarized by algebraic expressions, certain conclusions follow. (p. 396)

Instructor B's student Zoe also expressed methods reflecting this view. She said, "You could have the people that work for you print out lists of [the gas stations], cut up the lists in pieces and draw by lottery. Or you could stick them up on a wall and throw darts at them and decide which one you want to call this way." Nancy, also instructor B's student, said, " To make it totally random, you would have to have a list of all the gas stations." Her partner added, "and put them in a computer to select them randomly."

Thus, instructor B's students knew the expert's way of drawing random samples that allows the application of the formulas and calculator programs that the students were taught in their statistics classes. However, those four students were in a minority in the study. The other students suggested methods that they thought would minimize bias, but at the same time ignored that the methods taught in class were built on mathematical models requiring simple random sampling. Some students recalled methods such as cluster sampling, whose applications require its own formulas not taught in introductory statistics classes.

Dana suggested that, "they should send a bunch of people out". Her partner Cindy added, "at the same times", and Dana finished with, "at a lot of different areas." Ideas, such as those that Cindy & Dana expressed, were dominant among the student responses. Such responses showed that the students knew the importance of avoiding biased samples. However, probability education research is full of examples how the human mind is not good at avoiding bias when using common sense methods such as those suggested by Cindy & Dana. Statisticians have therefore devised methods such as simple random sampling, which sidesteps human judgment and lets machines pick the samples.

4.1.7 Most students gave rationales similar to the instructors' for using samples

Instructor B, like the students, was asked, "Why do you take samples?" His response was,

> Here is the issue you have when you do inferential statistics versus descriptive statistics. The big difference between the two is that when you do descriptive statistics you collect data on a group and then make a statement about that group, where in inferential statistics you collect data on a smaller group and make a statement about a larger group. So you have this whole idea of sample. So is it better to make a statement on the group that you collected all the data on if you could? Yes, of course the answer is yes. Because then you know that what you say is absolutely true. Whereas for this inferential stuff you don't know for sure. There is always a possibility that there is error. But oftentimes you can't collect the information from the whole population because it costs too much money [or] you might be talking about a future population [such as] all the people who might get a particular disease. So the only option you have is looking at a sample.

Thus, Instructor B's rationale was that the focus of interest, the population, was unavailable for some reason. Therefore a sample was needed to be collected from the population that would still allow for conclusions about it.

Cindy gave an answer that closely resembled Instructor B's, " To find out about the population (She shrugs her shoulders.) They broaden their information to include the population.” Also, Nancy said, "Because there is too much data to find everything out. That's why you need to take random samples." A typical student exchange about why one takes samples took place between Tracy and Ursula.

Excerpt 4.10
Ursula and Tracy speak about why you collect samples

Ursula: To get data

I: Right (with an intonation implying I want more).

Ursula: To get the data they are trying to research

Tracy: And to *get a sense of the whole*, by spreading out (She gestures to describe the spreading out.) and get samples they get a sense of what the overall situation is.

I: Very good. I can hear that you are a journalist. What kind of sample do they want? What is the technical term for the sample that we always want?

Tracy: Random sample

Ursula: Statistical and random sample.

I: Why do we want a random sample?

Tracy (as a delayed answer to the preceding question): A representative sample that represents all the different areas. (She again gestures to show the spread)

Ursula: *Sometimes we can't get everyone*, so, I don't know if I am saying it correctly, so that is why we want to do a random sample.

Ursula started her answer by only saying that the sample would give you data. However, after listening to Tracy she recalled a phrase frequently used by her instructor (me), " Sometimes we can't get everyone".

Tracy also exhibited a good understanding of why one would take samples when she explained to Ursula what the Coin problem was about. Ursula was trying to solve the Coin problem as it were a hypothesis problem concerning two means and the following exchange arose.

Excerpt 4.11
Tracy and Ursula Discuss the Coin problem

Ursula: But what is μ_1 and μ_2 then?

Tracy: I would say it is the number of heads and tails that land.

So the H_0 would be that you get *fairly consistent* number of heads and tails and H_a would be that you would not get consistent number of heads and tails, and that would be that coin favors heads So you get more heads.

Thus Tracy knew that one takes a sample to test if the characteristics of the sample are *fairly consistent* with some hypothesized characteristics for the

population, from which the sample is drawn. Tracy was able to express a key idea in hypothesis testing. Still, just a minute later, as is very typical for the data in this study, she had problem applying this idea in a formal way. She entered the alpha value for $p_{o.}$ instead of the correct .5. Fortunately, Ursula questioned Tracy's choice of value for $p_{o.}$. Then Tracy corrected her mistake and successfully completed the Coin problem.

A majority of this study's students in some way expressed that the purpose of taking samples was to find out information about the population. However, a few students also talked about samples without directly connecting them with populations the way Ursula initially did. For example, Rose suggested you took samples to support a hypothesis, and Sylvia to show tendencies.

4.1.8 Summary

A few students in this study talked about collecting samples in order to gather information without connecting the concept of sample with the concept of population. However, most students in the study did connect the two concepts and competently talked about them in general terms. Students also knew the importance of getting unbiased samples, if one were to draw valid conclusions from a sample to a population. However, in the context of their problem solving, the distinction between the sample and population was not always kept in mind. For example, students confused sample means with population means at times. The confusion between sample means and population means was noticed in at least two contexts. First, several students would use the sample means in their hypotheses. Second, only two of

the student pairs were able to identify the jail problem as a one-sample problem instead of a two-sample problem. In addition, most students had difficulties deciding which test to use for the tranquilizer problem because they did not see differences in pulse values as sample and population values.

The students who were the most successful problem solvers were also the most proficient in using the statistical register. Using words like sample means and population means instead of only referring to means seemed to help students clarify the purpose of hypothesis testing. Also, looking at the calculator screen of different tests, such as the t-test, seemed to remind students about the distinction between sample and population, because the students needed to input sample characteristics for the American letters, while the Greek letters required inputs of population values.

4.2 Research Question #2
How Do Students Reason About the p-value in Statistical Hypothesis Testing?

The second research question concerns how students reason about p-values in statistical hypothesis testing. Nowadays researchers generally use p-values to decide if their results are statistically significant or not. Therefore the p-value concept has become an important topic in introductory statistics classes. To understand the meaning of the p-value one needs to have a solid understanding of sampling distributions. At community colleges the prerequisite course for the Elementary Statistics course is Intermediate Algebra. The curriculum does not assume that students have the mathematical background to follow derivations of formulas

describing sampling distributions. Since the instruction regarding sampling distributions cannot be built on strictly mathematical arguments, other ways of providing meaning to them are sought. Sampling distributions in Elementary Statistics courses are described in graphical terms as generalized histograms, which of course is not far from the way they are mathematically derived. To further support students in building some intuition for sampling distributions and p-values most statistics instructors let their students do simulation exercises in which they build histograms using hands-on activities such as dice throwing.

Section 4.2 will analyze how the students in this study reasoned about p-values. First there will be a short overview of the curriculum that leads up to hypothesis testing in section 4.2.1. Then, the two textbooks' treatments of p-values will be presented in sections 4.2.2 and 4.2.3. After the textbook discussion, a summary of how the instructors handled the p-value concept in the interviews will be given in 4.2.4. The students' reasoning about p-values will begin with a discussion of their graphical representations of p-values in section 4.2.5 and the meaning students attach to p-values in 4.2.6. Some examples of how student make mistakes because they consider p-values from a strictly procedural view follow in sections 4.2.7 and 4.2.8. In section 4.2.9 some students' difficulties with the abstract character of p-values will be demonstrated, and section 4.2.10 provides a summary.

4.2.1 Curriculum leading up to hypothesis testing

Before analyzing the student data on the p-value concept it is prudent to make a few comments about the material in the Elementary Statistics curriculum that leads up to hypothesis testing. In both of this study's textbooks a treatment of probability precedes inferential statistics. First, the textbooks have a general exposure to probability, which is followed by a treatment of the binomial and normal distributions. The normal distribution is always described in geometric terms, where probability is expressed as the "area under the curve". This view of the normal distribution was particularly useful when tables were used to compute normal distribution probabilities. Consistent with this geometric tradition the TI-83 calculator provides an option to graph the normal curve with the desired probability shaded.

After the probability curriculum, the textbooks cover the Central Limit Theorem, which also is described in geometric terms showing how the spread of the sampling distribution shrinks as sample size increases, e.g. figure 7-2, p. 343 in *Understandable Statistics. Workshop Statistics* has several simulation activities, such as 17-1 and 17-2, pp.367–375, showing how as sample size increases, any sampling distribution approaches the normal distribution. The ideas of the Central Limit Theorem are all described in geometric terms, with probability represented as "the area under the curve."

Inferential statistics then starts with a treatment of confidence intervals. In developing the theory of confidence intervals the textbooks continue to depend on graphs for their arguments (e.g. figure 8-3 p.377 in *Understandable Statistics*). In

addition to the normal distribution graphs, *Workshop Statistics* also refers to box plots and dot plots in an attempt to build students' intuitions. Hypothesis testing follows the confidence interval topics. After some introductory material, *Understandable Statistics* summarizes its information by giving a four-step procedure for hypothesis testing on page 463. All the students in the study had been given versions of this procedure by their instructors. When students follow those procedures, the students' reasoning will be called procedural reasoning in the subsequent pages. This definition of procedural reasoning is consistent with the more general definition given by Garfield (2002) as quoted on page 32 in this dissertation. However, it is more specific to the study's purpose. As an example of a procedure given to the study's participants, I will list the steps I give my students.

(1) List the information given in the problem

(2) Set up the hypotheses

(3) Decide on a level of significance

(4) Sketch the graph showing the critical region

(5) Use the calculator to compute the p-value

(6) Decide to reject or fail to reject the null hypothesis

(7) Answer the question asked in the problem

As books and instructors move into the hypothesis testing curriculum they continue to describe probability in terms of area. Since the mathematical derivation of the probability measures is beyond an Elementary Statistics course, an appeal to

graphs seems a good approach. The two next sections will summarize how the two textbooks treat the p-value concept.

4.2.2 The Understandable Statistics textbook's treatment of the p-value concept

When *Understandable Statistics* first introduces hypothesis testing for one mean, the textbook teaches the students to draw their conclusions based on test statistics such as z-values or t-values and tables from the book's appendix (p. 463–67). Then, students are taught the p-value approach in the following section (p. 477–489). As with many of the concepts, *Understandable Statistics* introduces the p-value approach with an example. The textbook chooses a hypothesis test problem example for which the statistical conclusion is different whether the significance level is $\alpha = .05$ or $\alpha = .01$ (p. 478). Then the textbook gives its definition of p-values:

> For the distribution described by the null hypothesis, the P *value* is the smallest level of significance for which the observed sample statistic tells us to reject H_o. consequently, if
> P value $\leq \alpha$, then we reject H_0
> P value $> \alpha$, then we do not reject H_0 (p. 479).

After the formal rule the book explains what the p-value means for a right-tailed test of the mean:

> The P value [for a right-tailed test on μ] is simply the probability that the sample mean from any random sample of the same sample size will be greater than or equal to the observed sample mean, $\bar{x}$. (p. 480).

The p-value is also described in graphical terms as areas in the tail or tails of sampling distributions referring to illustrations on pages 479 and 480 in the textbook. The example that follows includes a sketch of the corresponding sampling

distribution (but without the horizontal axis labeled). However, in the conclusion part of the example, neither the distribution described by the null hypothesis (cf. the definition above) nor its graphical representation is referred to. Only the following statement is made: “Since the P value is the *smallest* level of significance for which the sample data tell us to reject H_o, we reject H_o for any $\alpha \geq 0.0256$. For any $\alpha < 0.0256$, we fail to reject H_o”. (p. 481). However, students are unlikely to have developed much meaning with the level of significance at this point of instruction. Therefore, the quoted statements and similar ones in subsequent examples cannot be expected to further students' understanding of p-values.

After the section dedicated to p-values, *Understandable Statistics* devotes several sections to hypothesis testing of means and proportions. For each test that the textbook introduces, it first explains how to work problems using the traditional method involving a test statistic and tables, then explains how it would be done using the p-value. It might be worth noting that the instructors at the college where this study was conducted, who are using *Understandable Statistics*, mainly teach the p-value approach to hypothesis testing in spite of the textbook putting more emphasis on the traditional method of using tables. The p-values are easily available to students because they are required to have TI-83 calculators, which give the p-value for the hypothesis tests taught in Elementary Statistics classes.

4.2.3 The Workshop Statistics textbook's treatment of the p-value concept

Workshop Statistics defines the p-value as the probability, assuming the null hypothesis to be true, of obtaining a test statistic at least as extreme as the one actually observed. "Extreme" means "in the direction of the alternative hypothesis," so the p-value takes on one of three forms (corresponding to the appropriate form of H_a):

(a) $Pr(Z \leq z)$ (area below z-score) or

(b) $Pr(Z \geq z)$ (area below z-score) or

(c) $2Pr(Z \geq |z|)$ (area more extreme than the z-score in both directions) (p. 449).

On the next page *Workshop Statistics* gives the following template to students for them to use in solving their hypothesis problems:

> One judges the strength of the evidence that the data provide against the null hypothesis by examining the p-value. The *smaller* the p-value, the stronger the evidence against H_o (and thus the stronger the evidence in favor of H_a). For instance, typical evaluations are:
>
> p-value >.1: little or no evidence against H_o
> $.05 <$ p-value $\leq .10$: some evidence against H_o
> $.01 <$ p-value $\leq .05$: moderate evidence against H_o
> $.001 <$ p-value $\leq .01$: strong evidence against H_o
> p-value $\leq .001$: very strong evidence against H_o (p. 450).

Right after this template, the textbook introduces the significance level α as a "cut off" level for the p-level, and gives the more traditional decision rules of "Rejecting H_o" versus "Failing to reject H_o" on page 450. In the subsequent activities involving hypothesis tests the students are asked to compute p-values and use them to draw their conclusions. Occasionally students are also asked to interpret the p-value in the context of a particular problem as for example in part (e) on page 467.

4.2.4 The instructors' treatment of the p-value concept

During their interviews, the instructors talked more informally about p-values than the textbooks did. When asked about decision rules in hypothesis testing Instructor A used an example about an automobile manufacturer. She said, "The automobile manufacturer claims that the average mileage per gallon for a certain model of a car is 45. But you think that is too high. So you would be doing a left-tailed test. So let's say that alpha is 0.01. That means you are willing to reject the manufacture's claim only if the sample mean is in the lowest 1 percent of the sample mean distribution." Thus, Instructor A was talking about p-values as being a percent of the sampling distribution. When she spoke about the sample mean being in the lowest 1 percent of the sampling distribution her presentation was compatible with placing the sample mean on the horizontal axis of a sampling distribution graph and visualizing the p-value as the percentage of the area under the graph left of the sample mean. In other words, she expressed a graphical view of the p-value, as a shaded part of a sampling distribution graph.

Instructor A was also asked about the p-value in the context of solving a problem. Then she said that the p-value is "the probability that if you take a random sample that you will get results that look like this" (pointing to the sample mean).

Instructor B expressed similar views about the p-value when he explained the process of hypothesis testing using the Coin problem.

Excerpt 4.12
Instructor B discussing the Coin problem

> Instructor B: Step number four is always to find the particular probability and it is called the p-value of the test. It is basically the probability that if the null hypothesis is true that you would get the sample result you got, or greater than it. So, what is the probability that you under the null hypothesis would get a sample value p-hat greater than or equal to .62 [the sample proportion]. So that is called getting the p-value of the test, and that is basically what you have to do in step number four.

Then a calculator was brought out and the instructor commented about how he would discuss what to do on the calculator for this particular problem.

Excerpt 4.12 Continued
Instructor B discussing the Coin problem

> Instructor B: It is always very important to understand what is going on in this process, that the test you are doing is a test of the null hypothesis. You are trying to say something about the null hypothesis here, the question here is about the population proportion. Is the population proportion .5 or not? So, when you are looking at the null hypothesis, under p_o, that is always the population proportion

Then he told some calculator instructions ending with an instruction to select "Calculate".

Excerpt 4.12 Continued
Instructor B discussing the Coin problem

> Instructor B: And if you do that, what you find is that the p-value is 0.0448. What you find out, this is a little hard to read, because there is both a p and a p-hat specified. What does that mean? And what I would hope to elicit from them is, and usually there is somebody in the class that responds, that this is the likelihood to get .62 if the null hypothesis was true. So, that is hopefully what they are going to tell me. Then the question is: The chance of getting what we got, is that a *rare* event or not, what we got under the null hypothesis. We would already have discussed this issue that we need a definition of what we mean by rare event. And the definition of what we mean with a rare event is what we set the alpha level at. So the last step is to make a conclusion. And the conclusion is made by comparing the p-value with the alpha -value and it is basically about deciding, is this a rare event or not, if the null hypothesis was true.

Thus Instructor B added a connection between p-values and everyday language by introducing the term "a rare event". *Workshop statistics* also uses "rare" as a term for low probability events in the textbook's introductory activity on page 308. P-values as such are rather abstract. However, by using terms such as "rare" events for low probability events, and "probable" events for high probability events, statistics instructors, such as Instructor B, attempt to help build student intuitions regarding p-values.

4.2.5 Students' graphical representations in hypothesis testing

As was mentioned in the sections about the textbooks' treatment of sampling distributions and p-values, those concepts were often referred to in graphical terms. In fact, even persons with statistical training, such as Instructor B, seemed to think about normal probabilities as area. Consequently, the p-value often was referred to as the "area in the tail." In line with the above-mentioned geometric approach to probability as area, the analysis of the student data will start with an overview of how students used graphs when talking about p-values.

Below is a table summarizing the frequency with which students in the study used graphs of the normal curve and geometric arguments in making statistical decisions. As can be seen in the table, all but one pair drew graphs on their answer sheet. The graphs mostly consisted of a normal distribution sketch with shading of the tail(s). However, not all of those graphs had p-values marked on them. A special row in the table shows which students had p-values marked on their graphs, because only

those students could be expected to use their graphs to make their statistical decisions. Also, the word "explanation" in the table refers to explanations that students gave when interviewed by me.

The first two rows of the table contain information collected from the students' answer sheets. Therefore it was possible to give information for individual students. In each box the top check mark corresponds to the student listed first in the column headings. The bottom check mark belongs to the student listed below. This expansion of the table was necessary, because in some pairs one student used graphs while the student's partner did not. For the remaining three rows the information comes from the videotapes. In most of those instances the decision to use graphs seemed to be more of a joint decision by the pair than by individuals. Therefore one large check mark was used to denote that the pair used graphs or geometric arguments.

Table 4.2.
Students' graphical representations

	Alex Ben	Cindy Dana	Elena Fran	Gus Hal	Maria Nancy	Rose Sylvia	Tracy Ursula	Vera Zoe
Graphs on Answer Sheet but no p-values on Graphs	√ √	√	√ √	√ √		√ √	√ √	
Some Graphs with p-values on Answer sheet	√ √	√	√	√ √	√ √			
Geometric Argument for Several Answers	√			√				
Spontaneous Graphs With Explanations	√							
Geometric Arguments in Explaining Decision Rules	√		√	√	√		√	

Note. √ means that *both* students in the corresponding column had made graphs of the sampling distribution on their answer sheet
√ on the *upper* part of a cell means that the student listed *first* in the column heading had made graphs of the sampling distribution on her or his answer sheet
√ on the *lower* part of a cell means that the student listed *second* in the column heading had made graphs of the sampling distribution on her or his answer sheet
√ means that the *pair* listed in the corresponding column exhibited the behavior listed in the corresponding left heading

As can be seen from the table a great majority (81%) of the students drew graphs when they solved hypothesis test problems. The only pair that did not draw graphs at all was Vera & Zoe. However, most graphs did not have p-values marked on them, nor areas corresponding to p-values shaded in. Only half of the students

used detailed graphs that included some representation of p-values. Both students with good final grades (Ben and Dana) and students (Elena and Hal) with final grades of C used detailed graphs.

Student conversations revealed diverse attitudes towards the usefulness of graphs in solving hypothesis testing. At one end of the spectrum was Alex who said, "The visual thing helps a lot" when he was solving his first problem during the research session. At the other end of the spectrum were Rose & Sylvia who said that the graphs "did nothing for them" as far as understanding hypothesis testing. Even within pairs, opinions about the sampling distribution graphs differed. For example, Zoe suggested to Vera that they should do a diagram during their work on the Home Loan problem. Vera responded: "I never liked the diagrams. I always thought they were tedious. They annoyed me." To which Zoe responded: "They are nice."

Alex & Ben used geometrical reasoning more than any other pair. The pair called their graphs "the visuals" and used expressions such as "landing in the reject region" and "landing in the accept region". Those expressions were used both during the problem solving session and when the pair was interviewed. When asked about the decision rules they also used the expression "passing the point where you are willing to accept". The graphical approach served the pair well in their problem solving. From a strictly procedural view the pair's problem-solving work was flawless. Their two incorrect answers were due to other reasons. One of their incorrect answers was to due to a lack of care in reading the text of the problems and the other to an incorrect alternative hypothesis.

The only other pair that used geometric reasoning for their statistical decisions was Gus & Hal. For example, when the pair was asked why you reject the null hypothesis when p is less than alpha, Gus answered, "Because if it is in the alpha region, from alpha all the way to infinity, that's the reject region." Gus and Hal gave me the impression of being the pair in the study that were least serious about their statistical studies, which was reflected in their final grades of B and C. Still, as for Alex and Ben, the graphical approach served them well. As sometimes happened to other students in the study Gus & Hal were temporarily confused about the statistical decision rules during one of their problem solving attempts. While other students either referred to their notes or checked with their partners, Gus & Hal took advantage of their geometrical approach as can be seen the transcript below.

Excerpt 4.13

Gus and Hal discuss the statistical decision rule in hypothesis testing

Hal: Is it less than when you reject *it*, or is it when *it* is greater than. It is when *it*'s greater than, that you reject *it*, right?

Gus: *It* is right-tailed, right?

Hal: Yes. So you reject *it*, when *it* is greater than or less than.

Gus: You reject *it* if *it* is greater than.

There is a pause, under which Gus keeps writing.

Hal: So reject *it*. Even though *it is on the right side?*

Gus: But *it* will be in the reject region.

Hal: *It* will be way out here (points to the right tail of Gus' graph).

Gus: Yes.

Hal: So reject *it*.

After this exchange Gus and Hal silently wrote up their answers. It seems like Gus' & Hal's geometric approach helped them to make a correct statistical decision, when their memory failed them regarding the algebraic version of the decision rule. Hal says, "it is all the way out here" and points far to the right of the center on Gus' sampling distribution graph and draws the correct statistical conclusion from this idea. It is tempting to interpret his statement to mean that he was talking about the t-statistic being far to the right on his partner's graph. Such an interpretation would explain why the pair settled on that "'greater than' is reject". However, there is no trace of the pair considering the t-statistic in either the pair's conversation or on the answer sheets. Both their answer sheets only have the right tails shaded in with $\alpha = .01$ and 1.7×10^{-26} marked to the right of the .01. Thus it seems that Gus & Hal the week before their statistics final exam still have not fully thought through the ideas of hypothesis testing. Their persistent use of "it" for statistical concepts and concepts in the conversation seems to cover up their lack of understanding. Still, in none of the six problems they were asked to solve did they draw the wrong statistical conclusion from their calculated p-value.

In the student interviews, most students said that they had made more detailed graphs when they were in the beginning of the hypothesis test curriculum than at the time of the study. However, this study's data does not allow for any conclusions if the students actually did rely more on graphical representations earlier in their courses.

When analyzing the students' conversations it became clear that most students depended more on algebraic rules than graphical representations when making their statistical decisions.

Creating graphs as part of problem solutions seemed to help students to make correct statistical conclusions based on p-values. However, there was almost no evidence in the student conversations that those graphs helped build meaning for the whole hypothesis testing process. As with most of the students' work, shading the tail on the graph to represent the p-value was just another step leading to the answer of the problem.

4.2.6 What is the meaning of p-values in hypothesis testing?

In general the students in this study did not attach much meaning to p-values. One indication of this lack of meaning could be seen in the temporary confusion that most students showed regarding the statistical decision rules for hypothesis testing that are based on p-values. One such example was exhibited in Excerpt 4.12. Another such example is shown below in Excerpt 4.14, in which Rose and Sylvia discuss the answer to the Home Value problem.

Excerpt 4.14
Rose and Sylvia discuss the answer to the Home Value problem

Sylvia: So it is .008. So this time it is less [than alpha]

Rose: So we *cannot reject.*

Sylvia: Oh, wait a minute!

Rose: Right?

Sylvia: Doesn't that mean that we have to?

Rose is quiet for a moment reflecting on Sylvia's question and then responds.

Rose: Yes, yes, you are right. I got it backwards.

Rose was a good statistical problem solver. She even said in one of the sessions that she enjoyed solving statistics problems. Still, as was seen in the above excerpt, she made a mistake in drawing a conclusion from her p-value. When a student makes such a mistake after having finished an introductory statistics course, it suggests that p-values do not carry much meaning for that student.

Students were asked in the study what p-values were, usually in context of some problem. None of the students were able to give a definition of p-values without consulting notes. Cindy gave two definitions from her notes. Her first definition read, "the p-value is the probability of committing a type I error based on the data collected". Her second definition read, "the smallest value of significance for which you would reject." The way she read off the definitions verbatim in a rather monotone voice, without any comment, suggests that she did not comprehend those definitions very well.

Before I raised the p-value issue, I had discussed the concept of the significance level or alpha with Cindy & Dana. The pair seemed to have a fairly good understanding of the level of significance, a fact that was stated by me in initiating the following exchange.

Excerpt 4.15
Cindy and Dana speak about p-values

I: So you pretty much knew what alpha was, but you couldn't express it. It seemed that you had a pretty good idea. But you don't have the same intuition for what the p-value is.

[Both students shake their heads in agreement.]

Cindy: Yes, I honestly don't know what it is. I just know that if it's greater than or less than, it's just a rule …

Dana: It's just what you compare to the alpha

I: Yeah, it takes you through the problem fine. One way to think about it is graphically …

Dana: it's the shaded part under the curve

Cindy: Oh, yeah.

When Cindy said, "I honestly don't know what it is", referring to the p- value, she represented the majority response of the participants in this study. I also asked Cindy & Dana, "Why do you reject the null hypothesis when p is less than alpha?" Then Cindy answered, "Because our teacher told us to", and both students laughed. In the pair Vera & Zoe, Zoe said, "That's because what the teacher told me to do." Similarly, when Maria & Nancy were asked the same question, Nancy answered, "Oh my gosh, because that's the way we were taught. …It was a rule [the instructor] told us in the very beginning, but I don't think [the instructor] ever talked about why." Thus, it was common for the students in this study to consider p-values as a tool to

get to the answer without actually understand *why* the comparison with alpha would lead to an answer.

When pressured by me to produce explanations, those explanations often sounded confusing. The excerpt below shows such an example, in which Nancy's partner Maria tried responding to my request for more of an explanation.

Excerpt 4.16
Maria and Nancy talk about p-values

Maria: Maybe because the value that we are testing is 0.05 (she holds up her hands in the air, as to show the two "cut off lines" that alpha creates on the normal curve, when you do a two-tailed test.) I don't know …

Nancy: (encouraging Maria to continue): No, no, I think you are on the right track, keep going with that.

Maria (making a hand gesture suggesting a confidence interval): it's our little comfort zone to see if it fits in there.

Nancy: The sampling variability.

For quite some time Nancy and Maria tried to find some sense in the statistical decision rule without much success. The closest Nancy came to expressing a normative view was when she said, "Yes, because of your alpha value and your sampling variability, which is your alpha value, which is your significance level, if it less than your alpha level then you are probably off from what you are trying to find."

There were exceptions to the mechanical applications of p-values and unclear rationales for those applications. For example, Dana was able to show that for her a

small p-value meant something more than a step towards the answer. When asked by me for an explanation of why you reject the null when p is less than alpha she said, "Because there is such a small chance that it is true". Prompted by me, she was also able to fill in that "it" referred to the null hypothesis. Other students were able to express similar ideas as Dana about why a small p-value implied rejection of the null hypothesis. For example, Ursula answered that you rejected the null hypothesis when p was less than alpha because "when p is very small *it* cannot have happened by chance." Gus started to answer the question with a geometric argument, "It is in the rejection region". Then he added, "it means that the probability is slim to none for *it* to even happening, so you reject *it*." When asked by me what "it" is, he and Hal answered, "The μ that you are testing." Hal also added, "The null hypothesis."

The examples above show that in the context of making statistical decisions, some of the students were able to attach meaning to a small p-value, even though they did not know the p-value definition. However, for a majority of the students in the study, a p-value only made sense as a quantity to compare with alpha when making statistical decisions. Also, the fact that the students required alpha-values from me to do the hypothesis test problems indicates that most of the students did not see p-values as a self-contained concept.

The majority of students reasoned differently about p-values than statisticians do. The statistician view is expressed in *Workshop Statistics*. *Workshop Statistics* describes the p-value as a measurement of the strength of the evidence against the null hypothesis. The textbook also gives guidelines how to interpret p-values

depending on which interval they fall in. For example, a p-value between .05 and .10 is said to present some evidence against the null hypothesis, while a p-value between .001 and .01 presents strong evidence against the null hypothesis. Then the textbook introduces alpha as an "*optional* [italics added] cut off" level for the p-value on page 450. Thus while most students only think about p-values in relation to alpha-values, *Workshop Statistics*, statisticians, and a few students prescribe meaning independently to p-values.

4.2.7 Difficulties with considering p-values solely from a procedural view

If p-values only serve as tools to get to the answer, then the issue of p-values mainly becomes computational and procedural. Maria & Nancy exemplify the kind of confusion that appear when p-values are considered solely from a procedural point of view. The pair routinely divided all p-values by two, because they remembered their instructor dividing p-values by two. When I asked the pair why they divided their p-values by two the only response was that what their teacher did. Recall that Maria & Nancy was the pair with the least instruction in hypothesis testing. Likely they remembered the instructor solving two tailed test problems using graphs to make the statistical decisions. When solving such problems he would have had to divide the p-values and the alpha-values by two before writing the values on the tails of the graphs. When Maria & Nancy did not use graphs for their statistical decisions, but just tried to remember rules, the pair solved all but one of their problems incorrectly. Gus & Hal did an "opposite" mistake in the home value problem. They divided the

alpha-value by two but not the p-value. Then, when they used their graph to make their statistical decision, they arrived at the wrong answer to the Home value problem.

Alex also was confused by the experience of seeing his teacher sometimes divide the p-value by two. His partner Ben tried to sort things out for Alex, as can be seen in the following dialogue.

Excerpt 4.17
Alex and Ben discussing when you divide p-values by two

Alex: When you don't use the visuals, you don't have to divide by two?

Ben: Yeah, you don't. Yes, you never really do. Only for the visual thing.

Alex: Only for the visual?

Ben: Yes. Because if you are going to divide alpha by half then you have to divide p by half. But if p is already bigger than … then you already know it is going to be bigger. I don't think it's going to be bigger, though. So we don't have to divide by anything. (Both start working on their calculators).

Ben correctly pointed out that if you use the algebraic decision rule of rejecting the null hypothesis when the p-value is less than alpha, then you do not "divide by anything". The p-value given by the TI-83 simply needs to be compared with the given alpha. Only when using the graphical representation of the sampling distribution, which the pair calls "the visual", is dividing the p-value by two appropriate. In that context one would already have divided the alpha by two as part of graphical representation.

Later, while finishing the problem, Alex watched Ben draw the sampling distribution. As Ben was making his graph he commented on how the graph helped him making the statistical decision, "I get a p of .03, rounding up, and alpha was 0.05, yes. … So, it lands in the reject region." Alex is still was watching when Ben shaded both tails of his sampling distribution and said, "That's here and that's here." Alex says "yes", agreeing that those shaded areas constitute the reject region. Then he asks, "Would you divide p with two on this side too?" Ben answers in the affirmative and nods. He then adds, "So we are rejecting H_o. The pair then turned to a discussion of how to formulate the final answer. The discussions about the p-values and the graphs that Ben and Alex had in the second problem, the Home value problem, seemed to increase Alex's understanding of how p-values are used for making statistical decisions. His work on the subsequent problems seemed to reflect this increased understanding.

4.2.8 Abstract character of p-values causes students difficulties

Elena and Zoe (not in the same pair), who said they enjoyed doing the graphs, were among the weakest students in the study. Were they weak because they unsuccessfully tried to depend on the graphs? Or were they trying to depend on the graphs because they had great difficulties with the abstract concepts of hypothesis testing? The data suggest more of a yes to the second question than to the first. Both Elena and Zoe were serious, re-entry students who worked hard. It was not for lack of effort that they only received grades of C in their statistics classes. Ironically, their

effort to understand what they were learning might at times have led them into difficulties, which they could have avoided by using more of a mechanical approach that some of their younger counterparts used.

Zoe was well aware of this dilemma when she said, "I find that if I analyze math problems too much, then I get all wound up, because I don't understand. They don't make any sense to me. And then hard to do. So what the teacher says." A couple minutes later she provided a concrete example of her statements, which is related below in Excerpt 4.18.

Zoe's and Vera's instructor used particular notations in his class to stress that the p-values were conditional probabilities. For example, to denote the probability of getting a sample proportion of .62 or more if the hypothesized proportion was .5 he had the students write $P_{H_o}(\hat{p} > .62) = .0448$. Vera & Zoe dutifully used this notation for all their problems. Below is an exchange between Zoe and me about the p-value notation used in her class.

Excerpt 4.1
Zoe is asked about conditional probability notation

I: You had some nice notation in your work, where you had P and then it said a parenthesis and then it had x-bar or p-hat greater than something. So what does that stand for?

> Zoe: Yes, I know what you are talking about. And I have no idea what it stands for. What I know is that when [the teacher] showed us the process that is how we wrote it out and that is how he expected it to be written out on the test or we would get docked points. I don't know what it means. It just gives us the answer and we had to write it out. The p-h-oh-thing, something with the H_0. Oh, there goes bye-bye [She makes a gesture with her hands showing how the abstractions elude her.]

Zoe's and her partner Vera's work on the answer sheets supports Zoe's claim that the conditional probability notation did not carry much meaning for the pair. For example, in the work on the tranquilizer problem both students wrote $P_{H_o}(\bar{x} \neq 6.8) = .0779$. Since the students did not understand the ideas behind the notation, they mimicked what they had seen for one-tailed tests when solving a two-tailed test problem. The resulting statement was faulty and suggests that the pair had a poor understanding of p-values. A correct statement would have been that the p-value was $2\,P_{H_o}(\bar{x} < 6.8) = .0779$ (Cf. Rossman et al., 2002, p. 449). To consider the probability of having a sample mean that is not 6.8 as the students' notation suggested is irrelevant, and probably not even what the students had in mind.

Elena had a much more positive attitude towards her instructor and the curriculum than did Zoe, but no fewer problems with the abstractions and the mathematics. The graphs that Elena made on the answer sheets all showed the p-values in the wrong position, i.e. the p-value was written on the wrong side of alpha.

All the graphs illustrated either two tailed tests or right-tailed tests. Thus in the cases where the p-value was less than alpha, the p-value should have been listed to the right of alpha. However, Elena listed the p-value to the left of alpha. In the interview with Elena it became clear that she had marked the alpha and p-values on the horizontal axis as if those values had been values on a real number line. Elena did not think through how those values being represented by areas ordered them differently than, for example, z-values on graphical representations. When graphing on the number line, smaller numbers are graphed to the left of larger numbers. However, because alpha values and p-values are graphed as *areas* in the tail of the sampling distributions, a p-value that is smaller than an alpha-value should be listed to the right of that alpha value. It is not uncommon for me to see students in my classes reversing the order of p-value and alpha-value notations on sampling distribution graphs when they are just learning hypothesis testing. However, when Elena made such a mistake two weeks after the relevant instruction was completed, it indicated a lack of understanding.

Elena would not have been able to derive correct answers from her incorrect graphs if she had tried. However, her partner Fran preferred making the statistical decisions using algebraic arguments, such as "you reject H_0 if p is less than alpha". Elena followed her partner's lead and the pair solved half of the problems correctly and did some good work on the other half.

As mentioned earlier, Elena was a serious student, and the fact that her graphs did not match the decisions she and her partner Fran made, bothered Elena, especially

on the last problem. However, on the first problem, where Elena made a detailed graph (the home value problem), she forgot to divide her p-value by two for graphing purposes. Thus her p-value of .027 to the right of $\alpha = 0.025$ on her graph brought her no conceptual conflict with the algebraically made decision of rejecting the null hypothesis. If she had correctly divided .027 by 2 to get .0135 she would probably have realized that something was wrong with her graph. Instead her two mistakes cancelled out to make a correct statistical decision, and she was encouraged to retain her incorrect way of graphing p-values.

On the subsequent problems Elena made some comments about large p-values being "way over there", indicating that marking those p-values so far from the center of the graph did not make complete sense to her. Since her partner Fran used the algebraic rule to make her statistical decisions she only politely acknowledged Elena's comments without reflecting much over them. However, on the last problem, the tranquilizer problem, a discussion about graphing p-values follows after they made their statistical decision and answered the problem.

Excerpt 4.18
Elena and Fran discussing p-values

Elena: And we said p was equal .08. That is all the way over here (writing p=.08 and an arrow far to the right, below the shaded tail. The tail has α = .025 written above it).

Elena: Oh, Oh! [realizing something is wrong] Wait a minute! We said p = .08 and alpha = .025.

[Elena shows her graph to Fran.]

Fran: Yes.

Elena: This is *reject*. This is *reject* alpha, right? [Elena points to the shaded right tail of the graph, where her p=.08 is written.]

Fran: Yes. [She reruns the numbers on the calculator.]

Elena: You are saying accept the hypotenuse [sic]?

Fran [looking at her calculator display]: .08

Elena: Now, where is .08 on your picture? [She points to a graph on her partner's sheet with no numbers marked on it - only a normal curve with shaded tails]

Fran: So .08 is here.

Elena: So where are your critical region?

Fran [responds with an algebraic statement] : p-value is bigger than alpha. Do not reject.

Elena: Which is the same as accept.

[Fran looks in her notes]

Elena: Oh yes. I wrote that down wrong. [She replaces the *correct* labeling of the tail as "reject H_0" with the *incorrect* label "Accept H_0" to match the algebraic decision rule.]

Elena made the same mistake here as she had consistently done during the problem solving session. She forgot that because the p-values were graphed as areas, the larger a p-value was, the closer to the center the shading should reach and in this

case closer to the center than the alpha of .05. Elena's mind was still in the "larger-means-to-the right-of" mode.

However, Elena continued being unhappy with the answer to this problem. She looked in her notes for something that might help her out in her confusion. At one point she seemed to have found support for her original (correct) position of having the "reject" region in the tails of her graph. She then told her partner "I don't think that is right" [referring to the way they answered the problem]. However, since she did not get a supporting response from Fran she did not change the incorrect labeling of her graph.

This conversation displays a striking lack of awareness of a key concept in statistical hypothesis testing. Statistical hypothesis testing is based on sampling distributions with the hypothesized mean in the *center* of those distributions. If Elena had been aware that the mean they had hypothesized in the null hypothesis was the center of her graph, then she most likely would not have been willing to move away from her "Accept H_0" label from the center of her graph.

Statistics instructors usually make their graphs with the hypothesized mean or proportion marked below their horizontal axis in the center of their figure. By including the hypothesized mean or proportion this way in the graphs, the instructors stress that the tests are centered around sampling distributions that are created by assuming the null hypothesis is true. The pair Rose & Sylvia made most of their graphs the way the their instructor (I), had instructed them in class. The pair wrote out the hypothesized mean or proportion in the center of their graphs. Some other

students did a few of their graphs this way too, while other students did graphs with no mean or proportions marked on the horizontal axis. It is worth noting that the Rose & Sylvia pair was one of the more successful problem solver pairs in the study. The pair's successful problem solving indicates that marking the hypothesized mean or proportion on the sampling distribution graph might help students in solving one-sample hypothesis test problems.

The scarcity in the students' textbooks of graphs with means or proportions marked in the center of sampling distributions might explain why so few of those graphs were found on the students' answer sheets. *Workshop Statistics* has *no* graphs in the sections on hypothesis testing. *Understandable Statistics* does have graphs in the hypothesis testing chapters. However, those graphs always use z- or t-scales for the horizontal axis, which is not instructive for students whose hypothesis test instruction is built around x-bar and p-hat distributions and p-values.

The calculator allows students to draw the sampling distribution associated with any test the students might do and will shade the areas corresponding to the appropriate p-values. Some of the students used this feature during the problem solving sessions. However, the calculator graphs suffer from the limitation of not having the horizontal axis labeled at all.

Instructors try to make statistical hypothesis testing concrete by giving students procedures to follow, and by using graphical representations. However, as Falk (1986) eloquently pointed out, the ideas behind hypothesis testing are quite complex. The mathematics courses that the average community college student have

taken before his or her Elementary Statistics course do not provide for much practice in abstract mathematical thinking. Therefore it is not surprising that most community college students find statistical hypothesis testing challenging.

4.2.9 Summary

All students in this study used p-values to make statistical decisions for their problems. The only exceptions were two solutions by Maria & Nancy. This pair made decisions on two problems by drawing the sampling distribution and rejecting the null hypothesis based on the sample data being more than three standard deviations away from the hypothesized proportion. In all other cases the students (including Maria & Nancy) based their statistical decisions on a p-value given to them by their calculators.

Students also made the correct statistical decisions based on their calculators' p-values. A few times a student became temporarily confused and did not know if he or she should reject the null hypothesis or not. However, before the problem was completed that student was always set straight either by his or her partner or by looking in his or her notes. This kind of temporary confusion is not surprising if one considers that p-values did not carry much meaning for many of the students. On the other hand, the students knew their rules sufficiently well that those temporary confusions did not prevent the students from correctly reject or not reject their null hypotheses.

In spite of being able to make correct decisions based on p-values, the students could not explain to me what p-values were. The students neither know formal definitions of p-values nor could explain in everyday language what they meant. Several students did know the meaning of a *low* p-value as a small likelihood for the null hypothesis to be true. Several also said that a *high* p-value meant, "There was not enough evidence to reject the null hypothesis."

For most of the students the p-value did not have any meaning if no significance level alpha was given. For those students the p-value was something you compared with alpha to get to the answer. For two pairs of students, drawing graphs with sampling distributions helped in this process and might also have added some meaning of the p-value as a probability represented by area. One student, Elena, drew her p-value incorrectly on her graphs and could possibly have run into difficulties, if her algebraically inclined partner had not dominated the decision process. Considering that most weak students would not volunteer for a study like this one, there are most likely a higher percentage of weak students with the same confusion as Elena in introductory statistics classes than in this study. I have also seen the larger-means-to the-right concept incorrectly being applied to p-values in my statistics classes. This difficulty in moving from one abstraction, the number line, to another, probability as area, is an example of how the abstract thinking required in inferential statistics can be a hindrance even for hardworking students to gain competence.

4.3 Research Question #3.How Did the Students Reason About Answers to the Hypothesis Test Questions?

The third research question concerns how students reasoned about the answers to the problems. Instructors will usually ask students to complete the hypothesis test questions by answering the questions asked in plain English. In the discussion of the third research question those answers in plain English will be called *final answers*.

As an instructor I have often observed students solving a hypothesis test problem correctly, except for the last step of writing the answer using words from the given problem. In my Master's project (Aquilonius, 2002), I noticed that when students were formulating their answers, it led to discussions that might offer insight to the students' reasoning.

Table 4.3 below summarizes at which step students made mistakes that led to incorrect answers. When two mistakes were made on the same problem, I made a judgment call as to which mistake seemed to be the direct cause of the incorrect answer. Thus only one mistake is indicated for each incorrect answer.

As in the introductory section of this chapter, the stars symbolize situations where one student made a mistake, while the other did the work correctly. For example, Rose set up the hypothesis in the tranquilizer problem correctly, while Sylvia did not. The tape indicates that Sylvia agreed with Rose's answer in the end, but the appropriate corrections were not made on Sylvia's answer sheet.

The labels in the right-hand column of the table mean the following: *Incorrect hypothesis* means that the student pair set up their problem with at least one incorrect hypothesis and that this mistake was the main reason why the final answer was incorrect. *Incorrect p-value* means that some incorrect information was entered into

the calculator, which led to an incorrect p-value. *Incorrect final answer from statistical decision* means that the students, for example, correctly rejected the null hypothesis, but did not state the final answer that would logically have followed from this rejection. One of the *incorrect method* solutions consisted of a one-sample-problem treated as a two-sample problem, while the other incorrect solution consisted of a "mean" problem treated as a "proportion" problem.

From Table 4.3 it can be seen that none of the pairs in classes that were using the *Workshop Statistics* book, i.e. students listed after the dividing line, wrote the incorrect final answer after they decided to reject or not reject the null hypothesis. This finding suggests that the hypothesis testing might carry more meaning for the students using *Workshop Statistics* than for the ones using *Understandable Statistics*.

Table 4.3
Reasons for students' incorrect answers

Textbook for / In- student pair / correct answers due to:	Understandable Statistics				Workshop Statistics			
	Alex Ben	Cindy Dana	Elena Fran	Gus Hal	Maria Nancy	Rose Sylvia	Tracy Ursula	Vera Zoe
Incorrect hypothesis	1	1	1	1		2*	2	1
Incorrect choice of method						1		1
Incorrect p-value				1	2			1
Incorrect final answer from statistical decision	1		2*	2*				
No concluding sentence							1	

Note. 1(2) in a column means that the pair listed at the head of the column arrived at one (two) incorrect answer(s) due to the reason listed in the corresponding row
2* in a column means that one of the students listed in the pair listed in the column arrived at *two* incorrect answers due to the reason listed in the corresponding row, while the other student only had *one* incorrect answer for the reason listed to the left

All pairs, except Maria & Nancy, arrived at an incorrect answer, at least once, because they set up an incorrect hypothesis. Rose & Sylvia and Tracy & Ursula wrote their hypothesis for the Coin problem as a two-tailed problem. However, assuming that a two-tailed test would have been appropriate, as some statisticians might have claimed, the pairs' answer would actually have been correct. The choice of method was only the cause of two incorrect answers. These few method mistakes need to be seen against the background of the pairs' often lengthy discussions about which method to use. The pairs were always able to correctly reject or fail to reject the null

hypotheses based on achieved p-values. Therefore the table does not contain an entry for incorrect answers caused by incorrect statistical decisions.

An important finding, not reflected in the table, is that Maria & Nancy wrote correct answers to three of the problems without having correct work supporting their answers. Maria & Nancy was the pair who used *Workshop Statistics*, but had not received instruction on the whole hypothesis testing section at the time of this study. Four other pairs also wrote correct answers to one problem each without correct supporting work.

As a background to the analysis of the students' reasoning about answers I present the textbooks' instructions to students about how to answer the hypothesis test questions is given in section 4.3.1. In section 4.3.2 there is information from the instructors about how they speak to their students regarding arriving at final answers to hypothesis test problems. In section 4.3.3, I give an example of how the students often spent a substantial amount of time to formulate the final answer, even though they had worked the problem correctly up to that last step. Students sometimes failed to look back to their original hypothesis at the end of a problem to clearly ascertain what they are rejecting or not rejecting. An example of such a case is presented in section 4.3.4. On the other hand, when the students perceived a conflict between what they saw in the data and what their computations told them, they often reflected on the contraction as described in section 4.3.5. Section 4.3.6 discusses some of the phrases students used in their answers. In section 4.3.8 it is described how students were aware that they were not supposed to draw conclusions directly from the data

without using the hypothesis testing process, but still preferred to do so a few times as can be seen in section 4.3.7. Section 4.3.9 provides some concluding remarks about how students arrived at their final answers.

4.3.1 Textbook treatments of how to arrive at final answers

It is hard to find any explicit instructions on how to find the final answer to hypothesis test questions in either of the two textbooks (Brase & Brase, 2003; Rossman, Chance, & von Oehsen, 2002). The students often find this part challenging and will not always understand that to do so is the most important part of the exercise. *Understandable Statistics* gives table 4.4 below. Some of my colleagues tell their students to use this table as a template for their answers, substituting appropriate words from the given problems.

Table 4.4
Meaning of the Terms *Fail to Reject H_0* and *Reject H_0* (Brase & Brase, 2003, p. 458)

Term	*Meaning*
Fail to Reject H_0	There is not enough evidence in the data (and the test being used) to justify a rejection of H_o. This means that we retain H_o with the understanding that we have not proved it to be true beyond all doubt.
Reject H_0	There is enough evidence in the data (and the test employed) to justify rejection of H_0. This means we choose the alternative hypothesis H_1 with the understanding that we have not proved H_1 to be true beyond all doubt.

A few of the students in the study made good use of the phrases given to them in this table. However, more students in the study had difficulties formulating their

final answers. The students using *Understandable Statistics* spent more time on formulating their final answers than the *Workshop Statistics* students. Also, as was seen in table 4.3, only students from the *Understandable Statistics* classes wrote incorrect final answers even though their work leading up to those answers was correct.

Understandable Statistics is heavily dependent on examples for its presentation of statistical topics. Thus, in particular, the students are expected to learn how to answer hypothesis questions in the contexts of the presented examples. However, the hypothesis testing examples in *Understandable Statistics,* with very few exceptions, use test statistics such as z-values or t-values for their decisions, while the students were instructed to use p-values in their classes. Therefore the students might not have been able to benefit much from their textbook's examples. In addition, the textbook at the end of their example problems never directs students to go back and clarify the meaning of the hypotheses stated in the beginning of the problem. Without clarifying the meaning of the stated hypothesis, it is hard for the students to draw the final conclusion after they rejected or failed to reject such hypotheses.

Workshop Statistics gradually introduces students to hypothesis testing throughout the book. For example, the book defines the concept of *significance* in *Topic 16,* SAMPLING DISTRIBUTIONS I: PROPORTIONS. In *Topic 16,* there is also the Activity 16-4 ESP Testing in which students carry out a hypothesis test without the standard symbolism and without the activity being called a hypothesis test. In the ESP activity the students are given a probability distribution for correct guesses

in the task of identifying which of four shapes (star, circle, wave, or square) appears on a card unseen by the subject. The students are then given a sequence of questions, culminating in, "Suppose that a particular subject gets 50% (20) correct in a test. How convinced would you be that she actually possesses the ability to get more than 25% correct in the long run? Explain!"

This approach of introducing hypothesis testing ideas without the conventional symbolism is consistent with the authors' goals as expressed in their preface. In the preface the authors write that they aim to give the students "learner-centered activities through which students can discover statistical concepts" (Rossman et al., 2002, p. xi), and for students to "construct their own knowledge of statistical ideas as they work through the activities" (p. xix). The implication seems to be that if the students have a better sense of the reasons why one might want to do hypothesis testing, then they will be better prepared to write the answers to subsequent problems.

Workshop Statistics leads the students through a number of hypothesis test activities step by step. A typical example of how students are prompted to answer a hypothesis test question is found in Activity 21-2 part (g), page 453: “Based on this p-value, would you say that the *sample data provide strong evidence* [italics added] to support Marilyn’s contention that the proportion cited by the grandfather is too high to be the actual value? Explain.” This example and other similar prompts are likely designed to help students connect the original question with the final answer. Using terms such as "sample data providing strong evidence" *Workshop Statistics* supplies

students with appropriate vocabulary to use in their answers similar to what *Understandable Statistics* did with the templates quoted in Table 4.4.

4.3.2 Instructor treatments of how to arrive at final answers

Instructors and students were asked two questions that were directly related to how one answers hypothesis test questions. The excerpt below describes the exchange between instructor A and me about those questions.

Excerpt 4.20
Instructor A instructions to students regarding answering problems

I: How does knowing that you reject the null hypothesis help you to answer the question asked in the problem? Use an example, if it helps you to explain.

Instructor A: I actually had the students, when they decided that they were rejecting the null hypothesis, lightly draw a line through the null hypothesis, which left them with the alternative hypothesis. So when you reject the null hypothesis, then you accept the alternative hypothesis and then they were to write a sentence that stated in words the answer to the question in the problem that they were supposed to answer.

I: So here is the opposite situation again. How does knowing that you failed to reject the null hypothesis help you to answer the question asked in the problem? Use an example, if it helps you to explain.

Instructor A: That's where the word “accept” was nice to use, because if they failed to reject it, then you were accepting it in the sense that this is a keeper. The claim is OK. We don't have enough evidence to throw it out.

I: Now I wished I had a video camera here because now instead of drawing a line through the null hypothesis, you are circling it, right?

Instructor A: OK, we don't have enough evidence to reject the manufacturer's claim that the miles per gallon for this model is 45.

I: I know now why it is good to have a video recorder, because your circles will not show up on the audio recording.

Instructor A: Yes, I have always made sure before they start … They often want to just start plugging in numbers in the calculator…that they write the null and the alternative hypothesis first. And when they finished that they go back up to that. And either they reject it and I would have them lightly draw a line through the null hypothesis. Or fail to reject it. Or circle it. This is what you are keeping. Write this in a sentence.

From the transcript at least three important pieces of information can be derived. First, Instructor A stresses the importance of going back and reading the hypotheses at the end of the problem. She gives her students a hands-on strategy by telling them to physically go back and either line out or circle the null hypothesis depending on whether they reject or do not reject it.

Second, she tells her students to write a sentence in which they are to state in words the answer to the question in the problem. Third, she gives them some vocabulary to use for their answers, such as "we don't have enough evidence to reject the claim". When asked to do one of the diagnostic problems, the poll problem, she used similar vocabulary. In the poll problem, a sample proportion of .52 turned out not to be significantly larger than .5. When discussing this problem, Instructor A said, "though the sample showed slightly more than 50 percent, it was not enough more than 50 percent to contest that figure of 50 percent".

Instructor B talked about how to answer hypothesis questions similarly to Instructor A. In particular, he also used the phrase "There is not enough evidence" in the context of discussing the poll problem. In addition, Instructor B stressed the issue of sampling variability when answering questions about populations based on sample information. When discussing the Coin problem he said, "I always have them say, 'Based on the data', because it is important, as I teach this, to make sure that they all understand this, that our decision is made based on our particular sample. But if we had a different sample, we might have a different conclusion. It is based on a single sample and what we know, and there is always this issue of sample variability."

4.3.3 Arriving at the final answer is seldom trivial for the students, even when solving problems correctly

When the instructors solved a problem in this study they spent very little time formulating the final answer. For the instructors, writing the answer was just another step in the problem solving process requiring no more effort than the other steps.

Sometimes, a pair of students also would effortlessly be able to correctly answer a problem in the study. However, more commonly, students discussed how to formulate their final answer. Some of those discussions were fairly lengthy. These discussions not only provided information about the students' process of arriving at a final answer, but also more generally about the students' statistical thinking. The conversation quoted below exemplifies such a discussion.

Alex & Ben had correctly completed most of the Coin problem, which read: "You suspect that a certain coin, when tossed, favors heads. You toss it 50 times and find 31 heads. At the 0.05 significance level, does it favor heads or is it a fair coin."

The pair had correctly hypothesized, in the null hypothesis, that the proportion of heads is 50% and in the alternative hypothesis that the proportion of heads was greater than 50%. The students also had used their calculators to find a p-value of .04 and correctly decided to reject the null hypothesis. The transcript starts as the students begin to discuss how to formulate the final answer.

Excerpt 4.21
Alex and Ben answering the Coin problem

Ben: Does it favor heads or is it a fair coin?

Ben: What is it asking? It is not saying "different than", is it?

Ben: It says, "Does it favor heads?" It is kind of weird. It could say, "Does it favor heads" or "Is it a fair coin?" But it is in the same sentence. See, it's in the same sentence.

Alex: Yes, here it says, "Does it favor heads?" That's the first one.

Ben: So we are rejecting that it favors heads, because our H_0 is 50%. And we are saying …

Alex: But we are rejecting it, because we are trying to see if it favors heads, right? So aren't we testing that it isn't?

When Ben reread the question in the problem he was puzzled. He was accustomed to textbook problems where there was *one* claim in the problem that should be rejected or not rejected. Ben therefore complained that this question had two competing claims, and that they were "in the same sentence." Alex & Ben earlier made the statistical decision "to reject", and Ben considered (incorrectly) rejecting the claim mentioned by Alex, that the coin favored heads. However, Alex then interrupted Ben and (correctly) pointed out that, as always, they are rejecting the statement in the null hypothesis. However, Alex's utterance with its double negative seems to have been too convoluted for Ben to grasp. Therefore Ben did what many of the students did in this study when they were unsure how to proceed, Ben started to enter information into his calculator.

Excerpt 4.21 Continued
Alex and Ben answering the Coin problem

Ben: Let's see what we get, though (reruns his computations on the calculator) by doing this thing. Well, I don't know what we are testing. What is the problem? Oh, no no, this is what we are testing. We are testing getting 50% of heads - or more. That the proportion of heads is greater than 50%. It's got to be. That's what we are testing. Which makes it favor heads. So what we are

testing here is that it is greater than p_0. So, right-tailed? OK. Let's see what we got (looks at calculator display) I got the p equal to .04, that's to the right. So we are rejecting that it favors heads and it *is* a fair coin?

Alex: Yeah. (nods) Yeah.

Ben: (nods as well)

Alex: We are rejecting it?

Ben: Yeah, because our p is less than alpha. It lands in the rejection region

Alex: What's our H_0? That it favors heads? (does not sound convinced)

Ben: Yeah.

Ben was still (incorrectly) rejecting what the students had as the alternative hypothesis. Alex was not happy with this rejection and suggests that they are rejecting "it is equal to".

Excerpt 4.21 Continued
Alex and Ben Answering the Coin Problem

Ben: But if we are testing that our proportion is greater than fifty. And it gave us reject. That it is not greater than 50. That it is a fair coin.

Alex: Does that mean that we have to change our hypothesis?

Ben: Yeah, there is something screwy here.

Alex finally brought Ben's attention to the fact that there was "something screwy" in his reasoning. Ben paused to reflect. After a couple more minutes of

bouncing ideas back and forth between the students, Ben comes to the correct conclusion.

Excerpt 4.21 Continued
Alex and Ben Answering the Coin Problem

> Ben: And we put .5 that is 50% for the H_0. And we are rejecting H_0, we are rejecting that it is a fair coin. So we are accepting that it does favor heads. And this is wrong (refers to the hypotheses that the students incorrectly changed) Right? Because if we are rejecting H_o, H_0 is .5 in here (refers to calculator) then we are rejecting that it is a fair coin, agreeing that it favors heads. Which is weird because no coin favors heads, it is always fifty-fifty (shakes his head in disbelief). I would love to see the hint for this one later on. I want to stick to what we had originally for this one. I think it was right. This one makes more sense now. That p_0 is equal to .5 and H_1 is that it favors heads. Yeah, it does favor heads. I don't want it to but you know what? It has to.

Ben found it counter-intuitive that a coin would favor heads because "it is always fifty-fifty". He therefore had a hard time moving away from his original belief that the rejection referred to the coin favoring heads. However, his partner Alex's prodding helped him clarify what the data were telling him and he laughingly ended the problem by saying, "I'm writing, the coin significantly favors heads."

4.3.4 Sometimes students failed to properly consider their hypotheses when answering

The symbolic character of the hypotheses seemed to be one source of difficulty for Alex & Ben in answering Coin problem. In fact, in the beginning of the solution process, Ben said, "H_0 is heads? H_0 is?". He wanted his hypothesis in words, not in symbols. He had his focus on the question asked in the problem, "Does it favor heads or is it a fair coin?", and wanted to place this question in the null hypothesis, However, the null hypothesis would only be able to encompass one of the two sub-questions contained in the main question. Ben initially wrote one of the sub-questions his null hypothesis in (almost) correct symbols as $p_0 = . 5$. However, the pair's discussion, quoted in section 4.3.3, shows that Ben had difficulties connecting the meaning of his symbols with the subsequent solution of the problem. Thus the pair needed a lengthy process to complete the problem.

Elena & Fran had similar difficulties with the Coin problem. Fran wrote her null hypothesis both in words and in symbols. In words she wrote, "It favors heads or is a fair coin". In symbols she wrote, "p = .62" (making the additional mistake of using sample information in her hypothesis). Her computations led her to not reject her null hypothesis. In her answer, she chose to use the word version of the null hypothesis, "There is not enough evidence to conclude that it favors heads or it is a fair coin". If she had considered and comprehended the symbolic version of her null hypothesis she might have realized, as did Ben, that "there was something screwy here", and corrected her mistakes. As it happened she did not try to reconcile her

worded null hypothesis with her symbolic one and the pair left the problem with an incorrect answer.

Alex's and Ben's incorrect answer to the Checkbook problem provides another example of how a problem was answered incorrectly because the problem solvers did not pay sufficient attention to what the symbolic null hypothesis represented. The Checkbook problem was worded as follows:

> *Checkbook Problem.* In a discussion of the educational level of the American workforce, someone says, “The average young person can’t even balance a checkbook.” The NAEP survey includes a short test of quantitative skills, covering mainly basic arithmetic and the ability to apply it to realistic problems. The NAEP survey says that a score of 275 (out of 500) reflects the skill needed to balance a checkbook. An NAEP random sample of 840 young men (between 21 and 25 years) yielded a mean score of 272 with a standard deviation of 60. Is this sample result good evidence that the mean for all young men is less than 275?

Ben correctly worked the problem all the way up to the final answer including writing, "H_0: $\mu = 275$. However as a final answer, he wrote, "We accept H_0, that a score of 275 reflects the skill needed to balance the checkbook." He did not seem to consider what his symbolic hypothesis represented in words. His partner Alex wrote his null hypothesis as H_0: 275, leaving out the μ.In this way, Alex avoided dealing with some of the symbolism. However, his shorthand writing might also have prevented him from fulfilling the monitoring function that led the pair to the correct solution in the Coin problem.

4.3.5 Students reacted when they perceived a conflict between the data and their answer

In the preceding section Ben was shown to expect a coin always to be fair, to come up "fifty-fifty". His reasoning might have had its origin in him being introduced to theoretical probability by discussing how tossing fair coins give rise to two equally likely outcomes. Alternatively, he might have thought of how coins are used to make unbiased decisions such as who should kick off in a football game. In contrast, Rose & Sylvia's did not have a problem with the concept of an unfair coin. Because the sample proportion of heads was .62 they thought that the coin favored heads. As was noted in the preceding section, the correct answer was indeed that the coin favored heads. However, because Rose & Sylvia set up their test as a two-tailed test, their problem solving procedure did not give them the correct answer and led them to the following conversation.

Excerpt 4.22
Rose's and Sylvia's answering the Coin problem

Sylvia: Now here, I have problem. Yes, it is greater than .05, which means we can't reject the null hypothesis, right?

Rose: Say it one more time.

Sylvia: OK

Rose: No, we cannot reject it, because .089 is greater than .05

Sylvia: But it is not a fair coin. Because they got it 31 times.

Rose: Yes.

Because Sylvia and Rose set up the problem as a two-tailed test the calculator gave them a p-value double the one received by Alex & Ben. Following the decision rules taught in class, Sylvia and Rose correctly failed to reject the null hypothesis. However, this decision was counter-intuitive to Sylvia in light of the sample proportion of heads being .62. The students' intuition was right. On the other hand, they could also have considered the possibility that the deviation of the sample from the hypothesized proportion of .5 was due to sampling variability. When instructor B discussed the coin problem he mentioned this possibility before he did any computations.

The continuation of excerpt 4.22 shows Rose and Sylvia attempting to find a reason for their cognitive conflict between their intuition and their calculator result.

Excerpt 4.22 Continued
Rose and Sylvia answering the coin problem

Sylvia: Oh darn, remember what that magic number was? It was something about the sample size. If it was too small it was not accurate.

Rose: That was 30.

Sylvia: And we got 50. (laughs)

The students recalled that the Central Limit Theorem requires a sample of 30 for its application. Though not directly the cause for their difficulties at this point, sample size does play a role in how samples vary from the corresponding populations

(Cf. Tversky & Kahneman, 1971). Therefore, Sylvia's comment was pertinent to the solution of the problem.

Excerpt 4.22 Continued
Rose and Sylvia answering the Coin problem

Rose: Well, it's not necessarily a fair coin. It all depends what's showing up when you flip it, how many times it flips. So it's not a fair coin. Do you see what I mean?

Sylvia: Yes.

Rose: So it could be that it is not a fair coin. Right?

Sylvia: But if you cannot reject the null hypothesis? But isn't that what it says?

Rose: That it is. That it is a fair coin. We cannot reject that.

Sylvia: OK. So

Rose: Because of …

Sylvia: It is not sufficient to prove that it is not a fair coin.

Rose: That it isn't a fair coin.

Rose was leaning heavily on the sample result here when she said, "It could be that it is not fair". Of course, with a correct "greater than" in the alternative hypothesis, Rose could have shown her statement to be true. However, with the students' two-tailed alternative hypothesis the logic would lead to a different conclusion. Sylvia led Rose through this logical process and the students agreed that

they could not reject the null hypothesis. They said the answer would have to be, "It is not sufficient to prove that it is not a fair coin". Even when they agreed on this answer Sylvia was not happy with it. After a detour talking about confidence intervals, she voiced her doubts again.

Excerpt 4.22 Continued
Rose's and Sylvia's Answer to Coin Problem

Sylvia: That's what's throwing me off. It is 62.

Rose: That must be 62 here, and not .5

Sylvia: But we already know it's 62. So why would that be a big deal? Either it's 62 or not, but 62 isn't fair.

Rose: Right

Sylvia: It has to be 50.

Sylvia raised again the problem with having to call a coin fair that came up heads 62% of the time. Rose tried to help out by suggesting that they would change the proportion in the null hypothesis to .62. Sylvia correctly responded negatively to Rose's suggestion, indicating that you already knew the sample proportion is .62 and that it would not make sense to use a number you know in a hypothesis. She did not fall into the trap, as some of her fellow students did, of using the sample information in the hypothesis. However, she did not realize what their mistake was and the pair's puzzling over the coin problem continued.

Excerpt 4.22 Continued
Rose and Sylvia Answering the Coin Problem

Rose: So that would make it not a fair coin. If it is equal to .62, then it wouldn't be a fair coin.

Sylvia: How would we write that up?

Rose: You mean down here? (refers to the step of writing the final answer to the problem)

Sylvia: No. Here. I follow everything but when we get to the answer we both agree that this is not a fair coin. But the way we did it, it is supposed to be.

Rose: I guess this is one we have to talk to [me] about.

(Students look at me who follows the research protocol, which calls for waiting to give any feedback until the end of the research session.)

Sylvia: No hints?

I: Write something, though.

Sylvia: We did.

I: Write a summary of what you just said. You don't have to write everything.

Sylvia: OK. So we do not reject the null hypothesis (writes). And a question mark. Because it does not sound right for us. Answer: supports the coin is fair even though theta is not equal to fifty. I can't see where we went wrong.

Sylvia was still not happy with their final answer. She went back and made sure that they used the right test, which they did. She also checked that they entered the correct information in the calculator. They did enter the numerical information

correctly, but incorrectly selected the not-equal-sign for their alternative hypothesis. Sylvia failed to reconsider the alternative hypothesis, in which the students made the mistake to not use theta *greater* than .5.

Rose & Sylvia were not the only students who reacted when they saw a conflict between what they saw in the data and the answer the hypothesis testing procedure gave them. Gus & Hal had a very similar discussion about the Home value problem. In fact, in most of the problems students compared the final answer with what they perceived the sample information was telling them.

4.3.6 *Students used expressions reflecting the probabilistic character of their answers*

Certain words and phrases kept recurring in the students' answers. Cindy & Dana supplied the most extreme example on using scripted answers. For any problems, in which the pair had rejected the null hypothesis, the students wrote their answer as, "There is evidence …" (Cindy), or "Evidence suggests …" (Dana). For example, Cindy answered the coin problem with "There is evidence that the coin favors heads", while Dana wrote, "Evidence suggests that the coin is not fair".

For all the problems in which Cindy & Dana had failed to reject the null hypothesis, they used the phrase "There is not enough evidence". For example, in the Checkbook problem they wrote, "There is not enough evidence that the mean for all young men is less than 275". Cindy and Dana had different instructors. When they were to write the answer to the checkbook problem, which was their first problem, Dana asked Cindy if her class wrote the statement with "There is not enough

evidence" and Cindy said yes. It appears that both of the students' instructors had taught the students to use the phrases from *Understandable Statistics* (Brase & Brase, 2003, p. 458).

In the second hour of the first research session, I asked Cindy & Dana about the phrasing of their answers. Below are their responses.

Excerpt 4.23
Dana's and Cindy's comments about their scripted answers

Dana: They are basically phrased the same way. (referring to her and her partner's answers)

Cindy: Yes. For my teacher the only difference in each problem is that there is evidence or there is not evidence and then you take the question (points to one of the questions on her answer sheet) and form a sentence.

I: It really helps a lot.

Cindy: Yes, it is really easy.

Most Elementary statistics students find writing the answers to the hypothesis questions difficult. Therefore, it is worth noting that Cindy found using the script made writing answers easy. However, the script did not guarantee correct answers for the students in the study, even when all the work up to the final answer had been done correctly. Fran had the same instructor as Dana. However, when she used the script to answer the tranquilizer problem, it led her to an incorrect answer. Recall the wording of the tranquilizer problem.

> *Tranquilizer Problem.* In an experiment with a new tranquilizer, the pulse rates of 25 patients were taken before they were given the tranquilizer, then again five minutes after they were given the tranquilizer. Their pulse rates were found to be reduced on the average by 6.8 heart beats per minute with a standard deviation of 1.9. Using the 0.05 level of significance, what could we conclude about the claim that this tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute?

After Elena & Fran correctly worked most of the problem, they wrote their answers as, "There is not enough evidence that the tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute." Fran's conversation with her partner revealed that she associated the phrase "Do not reject H_0" with the phrase "There is not enough evidence". However, the claim of a 7.5 beats per minute pulse reduction was, as correctly stated by the students, in the null hypothesis. Thus the failure to reject the null hypothesis implied "there was not enough evidence to say the pulse reduction was *differen*t from 7.5 beats per minute" if one were to use the students' scripted kind of answer.

The scripted answers, used by Cindy & Dana and some of the other students, contained expressions that expressed the probabilistic nature of answers to inferential statistics questions. Other students used words such as "significantly" for the same purpose. The table below gives a summary of the most common such phrases and technical statistical vocabulary used in the students' answers. An **M** in any of the three first rows symbolizes that the respective students used the corresponding phrase in a majority of the answers. An **S** in the two last rows means that the respective students used the corresponding words in several of their answers. The phrases and words in rows #1, #2 and #4 all reflect the probability character of inferential

statistics answers. Thus all the students showed awareness that inferential statistics cannot give definitive answers. In fact there were very few answers reflecting certainty on the students' answer sheets.

Table 4.5
Students' probabilistic answers

	Alex Ben	Cindy Dana	Elena Fran	Gus Hal	Maria Nancy	Rose Sylvia	Tracy Ursula	Vera Zoe
Evidence suggests or not enough evidence to conclude		M	M					M
Data supports or data does not support …						M		
Based on the data …					M			M
Significant or significantly	S		S	S	S		S	
Accept	S			S			S	

Note. **M** symbolizes that the respective students used the corresponding phrase in a majority of the answers. **S** means that the respective students used the corresponding words in several of their answers.

In spite of the students showing awareness of the probabilistic character of hypothesis test answers, the table also shows that three student pairs did use the word accept in their answers. The interviewed instructors differed in their opinions about the use of "accept" in the context of hypothesis testing. In fact, that was the only area in which the two instructors expressed different views. Instructor A admitted that it was not statistically proper to use the expression "Accept H_0", because "Accept has a stronger meaning in English than we want". Still, she said, as was shown in the

transcript in section 4.3.2, that "accept" was nice to use as a transitory tool for arriving at the final answer.

On the other hand, instructor B was adamant that his student would not use the phrase "accept H_0" and none of his students did. Instructor B said that when you fail to reject H_0 it means, "you can't really tell. You get this really kind of wimpy statement. The statistics here does not allow us to make a stronger statement."

The students were asked why it is not good to use the phrase "Accept H_0." All but one pair, Gus & Hal, gave statistically meaningful answers to this theoretical question, even the students, who had used "accept" in their answers. For example, Tracy said, "Because it is statistically probable, so you can't say that you accept it, because if you accept it, then you say *it is.*" However, Hal answered that, "you are concentrating more on if you are rejecting or not rejecting than if you are accepting". When asked if there was a mathematical reason for this concentration, Gus answered that the reason for not using "Accept H_0, was "to keep it to a straight yes or no answer." It seems that Gus & Hal were rephrasing my question rather than answering it.

4.3.7 A few times students preferred drawing conclusions directly from the sample

In spite of the students' use of probabilistic terminology as demonstrated in the preceding section, a few times students preferred to draw their conclusions directly from the sample result without using statistical theory. For example, Gus answered the coin problem, "it is not a fair coin because the probability of the coin

landing on heads is greater than landing on tails." His answer sheet has $\hat{p} = .62$ next to his written answer indicating that the sample proportion was used to answer the question.

Gus' work on his answer sheet shows that he used a calculator program that gives a proportion confidence interval if information from a sample is entered. He also has a list on his answer sheet that shows what information he entered into his calculator. He correctly entered the number of successes as 31 and the sample size as 50. However he incorrectly used the significance level .05 for the confidence level. Those numbers would give the very narrow confidence interval of (.6157, .6243), but Gus never wrote this interval on his sheet. He chose just to use the sample proportion for his answer.

When Gus and Hal handed in their coin problem work, they expressed doubts that their solution was correct. They might have sensed that their work was incomplete. Other students in the study expressed similar ambivalence. Below are a couple of examples in which students were not clear why they had to go the extra step of doing a hypothesis test after they had studied the sample information.

Both examples below concern the sugar machine problem. The first transcript starts with Vera reading the problem.

Excerpt 4.24
Vera and Zoe start to discuss the sugar machine problem

Vera (reading the problem): A machine can be adjusted so that when under control, the mean amount of sugar filled in a bag is 5 pounds. To check if the machine is under control, six bags were picked at random and their weight (in pounds) were found to be as follows: 5.3 5.2 4.8 5.2 4.8 5.3 Has the machine slipped out of control?

Zoe: A little bit. Because those two equals 5 each. (She points to sample values on her sheet, and she averages one sample value of 5.2 with one of 4.8 and another 5.2 with another 4.8, which leaves her with two unaccounted for 5.3s). So it is a little off, but not much. But now we have to prove it (starts laughing). Because my theory does not work. Mathematicians do not accept that.

Vera: But it still works.

Zoe did some mental computations and concluded that the machine had slipped "a little bit" out of control. Statistics instructors encourage these kinds of preliminary estimations. Instructors would be supportive of her expressing an expectation of what the decision might be, based on her estimates. However, Zoe also knew that "mathematicians do not accept that". Thus the students continued solving the problem following the hypothesis test procedure taught by their instructor. Still, Zoe and Vera did not show any understanding of *why* "[Zoe's] theory does not work". In fact, Vera claims that Zoe's theory *does* work.

However, when Zoe said that the machine had slipped "a little" out of control she assumed that the mean of all sugar bags would be the same as for the sample. Her theory was consistent with Kahneman's and Tversky's heuristic *representativeness* (Kahneman & Tversky, 1982). Her instructor's (Instructor B) considerations of sampling variability were not brought up by either of the two students. The instructor, on the other hand, would have voiced the possibility that the mean of all sugar bags could still be five pounds, even though the *sample* was "a little off".

In the next example, Tracy made a preliminary decision based on the sample values. In her case, her partner Ursula alerted her that "there is more to the problem", as can be seen in the following transcript.

Excerpt 4.25
Tracy and Ursula work on the Sugar machine problem

Ursula: What did you do?

Tracy: I added up the numbers (refers to the sample values) and divided by six and we got 30.6 divided by six, and got 5.1.

Ursula: Right.

Tracy: So I would say the machine has not slipped out of control. Because we are looking for the mean amount of sugar to be five pounds.

Ursula: Aha.

Tracy: And the mean amount, the average amount, is 5.1 for that.

Ursula: Yes, maybe that is the easier way but I'm thinking there is something more to this problem (both students laughing).

Ursula: Because we have to get the mean, and the standard deviation, and after we got the mean and standard deviation we have to put it into …, you know, and put it into the test and get the answer.

Tracy rereads the problem silently.

Ursula: We want to set it equal to five. And not equal to five. Don't you think?

Tracy: Aha, Oh, I see what you are saying. So the H_o, of course, would have to be (starts writing) that the mean is equal to five.

Ursula: Aha. And H_A not equal to five.

The students finished the problem by completing the hypothesis procedure and answering the problem correctly, writing, "There is not enough evidence to prove that the machine has slipped out of control".

Later, I asked Tracy how she had concluded right away from the sample mean of 5.1 that the machine had not slipped out of control. The following conversation ensued.

Excerpt 4.26
Interview of Tracy and Ursula about the Sugar machine problem

Tracy: Actually I was saying since we only had six bags then the average was probably five pounds, but then [Ursula] took me into what was a little more statistically accurate and acceptable.

I: Yes, so what you did, using intuition, she did in a formal way.

Tracy: Yes, she has been trained better than me (laughs).

Ursula: Yes, somebody trained me that way. (Refers to me who also was the pair's instructor).

Thus Tracy started with computing the sample mean, and drew the conclusion that the machine had not slipped out of control. When Ursula was not satisfied with answering the problem that way, Tracy followed Ursula's lead and helped complete the hypothesis testing process. However, just as in the case with Vera & Zoe, Tracy & Ursula were not motivated by some probability or statistical rationale. Rather, they justified their hypothesis test procedure with "that's how their teacher had trained them".

4.3.8 Students knew that you do not draw conclusions directly from the sample data, but reasons given were procedural rather than conceptual

In an attempt to probe further into students' reasoning about *why* you need to carry out hypothesis testing the students were given the following "diagnostic" exercise.

> *Poll problem.* A student was given the following problem:
> In a survey conducted by Louis Harris of LH Research, 1250 US adults were polled regarding their view on banning handgun sales. The results were that 650 of those sampled favored a ban. At the 5% significance level, do the data provide sufficient evidence to conclude that a majority of US adults (i. e. more than 50%) favor banning handgun sales?
> The student worked the problem correctly all the way to the last step, including setting up $H_0 : \theta = .5$ and $H_a : \theta > .5$. Also the student correctly failed to reject the null hypothesis because the p-value (.0786) is greater than .05.
> At the end the student says, "Yes, a majority favors the ban, because $\hat{p}$ = .52".

> Do think that is right? If you don't think it is right, please correct the student's work and verbally explain as to the student, why the way she or he answered the problem was wrong.

This problem was presented to the two interviewed instructors in the following way: "Pretend that I comes to your office complaining that you had taken points off on my test for the answer given above. Please explain to me why the way I answered the problem was wrong." When explaining to me why my answer was wrong both instructors emphasized that what was of interest in this problem was the proportion of a *population.* Instructor A said, "They took a sample of 1250 and found that 650 favored the ban and they got a point estimate of the proportion to be .52, but you are seeing if a sample of this size 1250 with 650 favoring it, if it is significant enough to believe that the *population* proportion is greater than .5."

Instructor B started his explanation with, "Here is what you have to remember about hypothesis testing. Your hypothesis is about a *population*." Later he continued with

> So you took *one* sample. It was a big sample it was 1250 people and in those 1250 people you found that 650 people favored the ban, that is that 52% was in favor of this. However, as we talked about over and over back to the Central Limit Theorem and how it fits in all of this. One of the things you see all the time is this whole issue of sampling variability even though this was a rather large sample, 1250, how would you know that if you picked another sample of 1250, how do you know that you would also get .52? Is there a possibility that you would get something less than it?

Both instructors emphasized that you need to do hypothesis testing because samples can vary substantially from each other and their characteristics might vary from their parent population. The hypothesis testing using one sample will help to decide the likelihood of the parent population having a certain characteristic such as a mean.

In contrast to their instructors the students mostly focused on which number should be entered for which symbol in the calculator similarly to what Gus & Hal are shown to do in excerpt 4.27. When Gus & Hal were presented the poll problem, Gus first read the problem as instructed by me. Hal meanwhile entered the information from the problem in his calculator. Then Gus almost immediately pointed out the student's mistake and his partner agreed with him.

Excerpt 4.27
Gus's response to the Poll problem

Gus: Because they are using p-hat and we are using p.

Hal: Yes

Gus (directing his comment to me): I think he used the wrong p.

I: Right. So can you pretend you are the tutor and explain. You are on the right track there.

Gus: Right. So if I was telling him what to do I would say, "Look at your screen with the results that you got. Because you did the test. You set up the hypothesis and everything right, but the only thing was that you looked at the wrong set of results. Because you don't want to look at the p-hat, because p-

hat is the mean of the probability, but you want to look at the real probability, which is regular p."

While the instructors phrased their comments in terms of sample and population, Gus referred to numerical quantities in the statement of the problem. He clearly assumed that those quantities came from a calculator display. He neither did the computations himself, nor looked at his partner's display. He had so internalized the calculator functions and the calculator's display that the calculator had become a conceptual tool to arrive at answers for hypothesis testing. He knew which p in the poll problem was the "real probability", the one you used to answer the problem. There is little, though, in his explanation that shows understanding of the underlying ideas behind hypothesis testing. The only exception might be his use of the term "mean probability". His expression "mean probability" for p-hat can be interpreted as an understanding that p-hat plays a similar role as the sample mean in hypothesis testing.

Cindy & Dana also had no problem finding the mistake in the poll problem, as can be seen in the following transcript.

Excerpt 4.28
Cindy's and Dana's discussion of the poll problem

Dana: I think they are wrong, because you cannot conclude that a majority favors the ban, because you cannot make that kind of a statement.

Dana: Because they said they worked the problem correctly all the way until the statement that they made. So what they should have said was …

Cindy: Yes, the p-value was greater than .05 so they should say . Yes, they correctly failed to reject

Dana: So what they should be saying is "there is not enough evidence to conclude"

Cindy: Oh yes, I just wrote he or the person just got confused

(Cindy writes "there is not enough evidence to say a majority favors banning hand gun sales [p-value = .0786]")

Dana: I wrote, "There is not enough evidence to conclude that the majority of US adults don't favor banning hand gun sales".

Cindy: You are making it, like a double negative. Because the sentence says (reading from the text) "does a majority of US adults favor banning hand gun sales?" So you would say it is not enough evidence to say this (points to the sentence of her answer sheet).

Dana: OK

Cindy: You don't have to switch it all around. So I think they just got confused on … first of all they were twoish like their sentence was too much of a Yes answer. The null and the alternative are not supposed to ever having a Yes or No definite answer. And, second, they just confused on Reject and Fail-to-Reject part. That's what I would say was wrong.

Just as Gus & Hal used the calculator as a way of organizing their thoughts and work, Cindy & Dana used their expressions "enough evidence" or "not enough

evidence" to organize theirs. Cindy & Dana also used those same phrases when they were asked to solve complete problems, as was demonstrated earlier in this section. However, this particular transcript shows that Dana initially inserted one negation too much in her answer. This mistake suggests again, as in the case with Fran's answer to the tranquilizer problem, that using the scripts does not guarantee a correct final answer, even though all preceding work has been done correctly.

In addition, Cindy's comment that the fictitious student's answer was "twoish," "too much of a yes-answer" is worth noting. As suggested earlier, the students' repetitions of the mentioned phrases seem to enforce the impression that the answers achieved by statistical hypothesis testing are probabilistic in nature.

Both Gus & Hal and Cindy & Dana largely ignored the sample proportion .52 in their discussion of the poll problem. This omission was in contrast to the instructors' discussions. The instructors both pointed out that .52 was close to the .5 in the null hypothesis. The pair Rose & Sylvia considered the sample result in their discussion. Sylvia's first comment was, "I don't remember to ever put p-hat as an answer". Then Rose & Sylvia decided to work through the problem as if no work had been shown to them. After they finished their work, the following conversation occurred.

Excerpt 4.29
Rose's and Sylvia's discussion of the poll problem

> Sylvia: So it is correct up to here, so that obviously changes the answer. So the way I wrote it, "Data given is insufficient to support that a majority favors a ban."

Rose (after a short period of reflection): You are right.

Sylvia: Because it is pretty close. What was the proportion?

Rose: .52

Sylvia: No, I mean the 650 divided with 1250. Let's see (she initially does not realize that the sample proportion was already computed for her in the problem. She enters the sample data in her calculator and is temporarily thrown by the .52 she gets on the calculator). Wait a minute!

She then realizes the connection between the given .52 and her computed .52 and continues:

Sylvia: I know we are right, but it is one of those things with the wording. That's why we have to include the level of significance.

Rose: Yes

Rose & Sylvia seemed to attach more meaning to the poll problem than Gus & Hal did. However, to do so they had to rework the problem. When they reworked the problem they saw that the sample proportion was indeed greater than 50%. Still it was not significantly greater than 50%, since significance depends on the level of significance. "That's why we have to include the level of significance".

Elena & Fran also decided to work through the poll problem in much the same manner as Rose & Sylvia. In the beginning of their computations, Elena said, "I agree that p-hat is .52, so I am not going to question that, but I am questioning the statement

that a majority favors the ban" In the end Fran (correctly) said, "So there is not enough evidence that a majority favors the ban. It could be 50%."

All pairs except Maria & Nancy were able to find the fictitious student's mistake and correct it. Maria & Nancy had most likely not had enough practice solving hypothesis problems to discover the mistake. The other student pairs' discussions showed varying degree of understanding of the underlying principles for hypothesis testing. For example, when Fran said, "It could be 50%", she is correctly expressing that the population proportion could be 50%, even though the sample proportion was .52. However, in contrast to their instructors, none of the students mentioned the words sample and population in their conversations.

4.3.9 Some concluding remarks regarding how students arrived at their final answers

The student pairs spent considerable time formulating their final answers. As was observed in Aquilonius (2002), students benefited greatly from discussing with a peer how to answer the problems. Metacognition, which often is lacking in students' mathematical problem solving (e.g. Schoenfeld, 1985), was promoted in at least two ways in this study by students working in pairs. First, one student usually took the lead and made a running commentary of the problem solving process. As the student did so, she or he would sometimes realize there was a problem with her or his reasoning. Second, students often asked their partners why they were doing what they were doing. Such questions often caused both the questioner and the questioned to reflect.

The pair Cindy & Dana, which was labeled as the most statistically competent pair early in this result chapter, consistently used phrases such as "there is not enough evidence" in their answer. This pair had the easiest time arriving at answers. It is not clear from the data if the pair had a better understanding of hypothesis testing as a whole and therefore found the afore-mentioned phrases natural to use. Alternatively, it could be that the repetitive use of the phrases helped the students understand the ideas behind hypothesis testing. Consistent with Vygotsky's (1962) theories, both of these explanations for the connection found between statistical competence and use of scripted answers might be simultaneously true.

Students' difficulties in formulating their final answers were often caused by a lack of clarity about what they had rejected or failed to reject. Sometimes the students did not reread their null hypothesis. At other times, the symbolic character of the null hypothesis prevented students from seeing clearly what they had rejected or failed to reject, as was shown in section 4.3.4.

In some cases the hypotheses were stated incorrectly, which led to incorrect answers. For some of those problems with incorrect hypotheses, the students reflected back trying to find where they had gone wrong. For other problems, students wrote out the wrong answers without reflection.

Sometimes students preferred to answer questions by referring to the sample data without using the hypothesis testing procedure. Even when they continued their problem solving process and did the procedure, their conversations indicated that doing so was to please me rather than seeing a statistical need for the process.

4.4 How Did the Student Reason About Statistical Hypothesis Testing?

The most striking difference between the students and their instructors regarding their approach to hypothesis testing concerned the modeling aspect of the topic. The instructors had a theoretical model of hypothesis testing consistent with the textbooks. The instructors' reasoning and problem solving followed from this model. The students had many of the pieces of their instructors' model but those pieces were not integrated to a whole. Students' lack of a coherent mental model showed in the way they were sometimes able to do one part of a problem correctly, and not another. In this section four examples of such partly correct solutions will be discussed.

The first example concerns Elena's & Fran's work on the Tranquilizer problem. In section 4.1.2 there was a detailed discussion regarding the pair's selecting a method and setting up the hypotheses for the problem. The pair was successful in correctly using the t-test and at setting up their hypotheses. However, as was discussed in section 4.3.6, Elena & Fran still answered the Tranquilizer problem incorrectly, demonstrating that they did not fully comprehend what their hypothesis testing result meant.

The second example concerns Maria's and Nancy's work on the Checkbook problem. Their Checkbook problem solution contained two mistakes. Their alternative hypothesis was incorrectly stated as $\mu_1 > 275$ instead of $\mu < 275$, which then was used for their calculator computations. At the end they incorrectly divided their p-value by two before making their decision. Otherwise their work on the

Checkbook problem was correct, including their final answer. Maria's final answer read, "Based on the data do not reject H_o. there is no evidence that the mean for all young men is *less than* 275".

The fact that their alternative hypothesis spoke about the mean being *greater than* 275 and the final answer used the words *less than 275* did not seem to bother Maria & Nancy. Their conversation clearly indicated that their use of the symbol > was used to express "greater than" and was not a careless mistake. Therefore the contradiction between having "greater than" in the null hypothesis and "less than" in the answer suggests that Maria & Nancy did not look at the Checkbook problem as a whole. Instead they were focusing on doing each step correctly.

The third example concerns Vera & Zoe, who also did two mistakes in a problem, but still finished with a correct final answer. In fact, their two mistakes in the solution of the Home Value problem cancelled out to provide them with the correct final answer. Their first mistake was to use the given alpha value as a hypothesized population value in their hypotheses. Then they worked the problem correctly based on their chosen incorrect hypothesized proportion. However, because of the incorrectly hypothesized proportion their calculator gave them the p-value 0.497. Following the decision rules taught in class they wrote, "based on the data do not reject H_o." Then, surprisingly, they finished the problem with writing, "undervalued homes are different than 18%." This last answer is a correct one, but would not logically follow from Vera's and Zoe's work. Again, as with other students in this study, Vera & Zoe had a good general idea what hypothesis testing is supposed

to achieve and showed good intuition for what the answer might be. They were also able to carry out most of the steps correctly. However, they did not demonstrate knowledge how those steps together create a model that will lead to the desired answer.

Cindy's and Dana's work on the Tranquilizer problem provides the fourth and last example of how students can miss the connecting logic of hypothesis testing though most of the steps are done right. As mentioned earlier, of the six problems that the students were asked to work completely, the Tranquilizer problem was the only one that Cindy & Dana did not do correctly. However, as will be shown in excerpt 4.30 they were very close to taking the right route at one time. The pair, as did several other pairs in the study, approached the Tranquilizer problem as a "Matched pairs" problem. If one is conceptualizes the Tranquilizer problem as such a problem, then the key step is to realize that the mean of the paired differences has already been computed, and is given in the problem as 6.8.

Cindy and Dana did take note of the important fact that the mean of the paired differences was already computed in the Tranquilizer problem, as can be seen in the following excerpt. However, somehow they were not able to capitalize on this important insight. Instead they fell back on mimic a class problem where the hypothesized difference was zero.

Excerpt 4.30
Cindy & Dana discuss the Tranquilizer problem (continuation from Excerpt 4.4)

Cindy: If it is a test, it is going to be t . It is under 30, anyway. It is 25.

Dana (picking up her calculator): I'll see what I can enter, you know. *If the data fits.* So, if you go to STAT the mean would be. So you see that's what it is reduced by. The mean is reduced by 6.8.

Cindy: OK, alright, OK we can do this, because the differences which is d-bar. Did you do that?

Dana shakes her head as if saying no

Dana (showing Cindy her calculator display): We go into … what is it?

Cindy: d-bar is just the average of differences.

Dana: OK, so when we do that we calculate it in here, though (refers to her calculator).

Cindy: The average of differences is 6.8.

Dana: Oh, because *it is already given?*

Cindy: *Yes, it's already given.*

Dana: So, we don't have to enter it (refers to that they do not have to compute the mean).

Cindy agrees

Cindy: the standard deviation of the differences is 1.9.

Dana: Cool

Cindy: And the *sample* [sic!] of differences is 25.

Up to this point Cindy & Dana seemed to be on the path to a correct solution. However, as the continued transcript shows, Cindy made a suggestion that surprised

Dana, but not enough for the pair to backtrack. The pair seemed to be aware that they something was missing in their understanding of the problem. This awareness showed, for example, in Cindy's embarrassment for using her notes to answer Dana's question rather than giving a statistical rationale.

Excerpt 4.30 continued
Cindy & Dana discuss the Tranquilizer problem

Cindy: So the null is that the average of the differences equals zero.

Dana (sounding surprised): OK, so the null is what?

Cindy: I'm just going off my notes. That is exactly that kind of problem. (laughs embarrassingly). These are the steps we have (points to her notes)

Dana: Oh, yes.

Cindy: Null is that the mean of the differences is zero.

Dana: Right, and then …

Cindy: As if there was no differences.

Dana: And the alternative is that you have to test it against the claim, right?

As Cindy & Dana tried to set up their alternative hypothesis they realized that something was wrong. However, they were not able to correct their mistake and did not successfully complete the problem. This last example shows that even the strong students did not always keep in mind the model that is to be used when testing the sample mean against the hypothesized mean.

Two other occurrences in the above excerpt are worth noticing, because other students in the study were using similar expressions. When Dana spoke about checking "if the data fits" (referring to a calculator program), she used an approach that all students in the study used at times.

Also, several students in the study spoke about sample size only using the word "sample", the way Cindy did. Though, the use of the word "sample" for sample size did not cause difficulties for the students in this study, one could see that this abbreviation could cause confusion. Samples have other important characteristics in addition to size. Therefore, the use of the word sample for sample size suggests that many students do not have a fully developed the concept of sample when they reason about statistical hypothesis testing.

The examples in this section indicate that introductory students know pieces of the statistical hypothesis test model. However, those pieces do not fit together as a whole for the students. In the discussion chapter, there will be a more detailed treatment of what hypothesis testing knowledge the students seem to lack. There will also be some suggestions how one might improve the instruction in hypothesis testing.

CHAPTER 5: DISCUSSION

The purpose of this study was to gain insight into students' reasoning about statistical hypothesis testing. I had, in my experience as an instructor, received mixed messages about students' understanding of hypothesis testing. The students in my classes would sometimes say or do things that made me believe that they had good understanding of the topic. At other times, the same students would make mistakes on tests and homework that made me doubt their understanding. In this study, present technology allowed me to go one layer below what I have been able to observe in the classroom. By videotaping students' statistical conversations and viewing them on DVDs, time after time, I was able to analyze students' reasoning at more depth and see more what students understand and do not understand.

Section 5.1 reviews some of the study's findings suggesting that many introductory statistics students have not fully grasped the relationship between sample and population in the context of hypothesis testing. Section 5.2 reviews some of the results about students reasoning about p-values suggesting that many students have a procedural view on p-values, which might prevent them from fully understanding what the result of their hypothesis test is telling them. A discussion of the kind of scripted answers that several student pairs used follows in section 5.3. Section 5.4 relates how the students did not seem to include probability theory in their reasoning about hypothesis testing. Section 5.5 concerns the students' use of the TI-83 calculator in this study. The discussion also suggests that statistics teachers might want to make more instructional use of the calculator than is now standard practice.

Section 5.6 is an overview of how some findings in this study relate to some published findings in the education literature. The need for some detailed studies of students' reasoning while they perform simulation exercises is presented in section 5.7. The last section, section 5.8, discusses the limitations and significance of this study.

5.1 Students Did Not Use Mathematical Models in their Reasoning

Most of the students did not seem to realize that hypothesis tests build on mathematical models requiring simple random sampling. The students were well aware of the importance of avoiding bias in sampling. However, they seemed to equate randomness with representativeness. They did not seem to understand how hard it is for the human mind to select an unbiased or truly representative sample. Therefore they did not fully appreciate why the simple random sampling method is needed as part of the hypothesis testing model. Of course, simple random sampling is not always practically feasible. However, the textbook problems all state that the samples are random, and as with all mathematical modeling it is important to understand the assumptions made when applying the model.

When students confuse sample means and population means on a classroom test or a homework assignment, their instructors will usually not know if those errors are just careless mistakes that students make because they were in a hurry and under stress, or if the errors are conceptual. When, in the past, I have pointed out such mistakes to students in my classes, the students have claimed those were due to lack

of care rather than lack of understanding. However, this study casts serious doubts on such claims. Even one of the best students, Cindy, systematically made such errors as incorrectly using sample values in the hypotheses for the first problems. Not until the fourth problem, did the repetitive pattern of the study's problems seemed to make her, at least temporarily, competent in deciding what was $\bar{x}$ and what was μ in a problem. Cindy did not bring a clear picture to the study session of how one uses the sample mean from the data to test it against a hypothesized population mean. Her lack of clarity existed even though she had had recent classroom instruction in the relationship between sample and population. Most of the students in the study made similar mistakes as the ones Cindy did. The diagnostic Jail problem was particularly designed to test if students know how to distinguish between sample means and population means. Six of the eight pairs in the study were not able to discern that one of the means given in the problem was a population mean and the other a sample mean. Those mistakes indicate that the students do not include the mathematical model of sampling distributions in their reasoning. Population means and sample means play such different roles in the theory of sampling distributions.

In addition to the confusion between sample means and population means, there were other indications in this study that the students did not use sampling distribution models in their reasoning. For example, the students seemed to have little awareness of sampling variability as they were doing problem solving. In principle, they knew that their answers to the hypothesis questions were not definitive. However, as they were solving the problems, they did not consider sampling

variability as a cause for the sample mean being quite different from the hypothesized population mean. For example, Ben did not like his answer to the Coin problem, because, in his words, "[The result] is weird because no coin favors heads, it is always fifty-fifty." Regarding the same problem, Rose & Sylvia did not mention sampling variability as a possible reason for the sample proportion .62, though their statistical hypothesis test indicated that the coin might be fair. In contrast, when Instructor B discussed the coin problem, he said that, "for a proportion test, 50 is a relatively small sample size and that you find more variation, more sample variation when we have smaller samples [like this one] … and it could just be happenstance, just a normal variation among samples". Instructor B expressed a key idea of hypothesis testing, which the students did not seem to have fully developed. The finding that students did not usually consider sampling variability in their problem solving is consistent with probability education research (Tversky & Kahneman, 1971, 1982a; Well et al., 1990).

5.2 Most Students Had a Procedural Approach to p-values

How do students reason about hypothesis testing in introductory statistics classes? Mainly they think about this topic as a procedure. That students think about hypothesis testing as a procedure should not be surprising, because that is how we instructors teach the topic to them. All five instructors, whose students participated in this study, teach their students very similar step-wise procedures to use for solving

hypothesis testing problems. An example of such a procedure was provided on page 116 in section 4.2.1.

The students' procedural thinking was most visible when they made their statistical decisions based on p-values. When asked about reasons for those decisions, most of the students referred to rules given by their instructors. None of the students spoke about sampling distributions or the Central limit theorem. Most of the students drew sampling distribution graphs, but those graphs were seldom used for statistical decisions. Neither were those graphs used in answering theoretical questions unless I suggested that the students do so.

As mentioned earlier in this dissertation, using p-values in statistical hypothesis testing has several pedagogical advantages compared with using test statistics such as z-values and t-values. In particular, the p-value method allows for a more unified approach to hypothesis testing. For example, the methods taught for tests about means and proportions can easily be extended to tests about two-way tables and ANOVA. However, the results from this study suggest that using only p-values in an introductory statistics course might also have at least one serious disadvantage.

The graphical representation of the sampling distribution seemed to be much less helpful in providing meaning to the hypothesis testing procedure when p-values were used than when, for example, z-values are used. When using z-values for making statistical decisions, the z-value can be interpreted as the number of standard errors from the hypothesized sampling distribution mean. In my experience as an

instructor I have noticed that students find it helpful to make statistical decisions by marking z-values and critical values on sampling distribution graphs. In such a case, the values involved are marked in the horizontal axis and dealt with as if it were on a real number line. As discussed in the result chapter, the comparing of p-values and alpha-values graphically follows a rather different pattern, which seemed to be much less intuitive to students.

On the one hand, the students competently made statistical decisions, such as "to reject" or "cannot reject", based on p-values using rules given by their instructors. One the other hand, when students had to move beyond the phrases "reject" and "cannot reject," those phrases often seemed more of a barrier than a help in formulating good English answers. Students struggled to give answers that seemed meaningful to them. Sometimes students even failed. Since their answers were to be concluded from p-values and the phrases "to reject" or "cannot reject", the students' difficulties suggest that the p-values did not contain much meaning for the students.

If students are to use their knowledge about hypothesis testing in their future academic career or as informed citizens the instruction about p-values in hypothesis testing need to be improved. Simulation exercises are already being tried and can probably be improved on. Also, better use could be made of the TI-83's capacity of drawing the sampling distributions with shaded in p-values. In particular, this TI-83 capacity could be used in conjecture with instructors speaking about the "graph as representing a sampling distribution based on the null hypothesis being true". The textbooks and the instructors used expressions like the quote above. However, we

instructors take for granted what the graph represents, but the students might need to be reminded in each example that we present to them. In their problem solutions, students could be asked to include phases such as " the probability of observing a sample mean as large as [the given sample mean] or larger, if the null hypothesis is true is [p-value]". Then students could formulate their decisions directly from p-values along the lines suggested by *Workshop Statistics* on page 450. This decision process might work better than using phrases containing the term "reject", which appear to be confusing to students. For example, if the p-value is between .001 and .01, *Workshop Statistics* declares that there is strong evidence against the null hypothesis. In addition, it might be helpful to have students write out what the null hypothesis represents in words at the start of solving the problem.

Another instructional suggestion concerns the z-statistic. The TI-83 calculator will give the z-statistic (or t-statistic for the t-tests) along with the p-values. The connection between the z-statistic and the p-value could be pointed out to students. Z-values measure distances to the hypothesized mean or proportion. The higher the absolute value of the z-statistic, the lower the p-value, because the sample mean is less likely to come from the hypothesized distribution.

5.3 Scripts Helped Students Answer Hypothesis Test Questions

Scripts helped the students in this study to answer hypothesis test questions, but were not always sufficient. Vygotsky's (1962) theory about the intimate relationship between learning words and understanding ideas seems applicable to the

data in this study. Students would clarify their thoughts by searching for appropriate statistical terms.

As was pointed out earlier, the students Cindy & Dana, who were the most successful problem solvers in the study, consistently used statistical phrases such as "There is not enough evidence". Sylvia and Tracy were articulate in a general sense, which helped them articulate some important statistical ideas (e.g. Excerpt 4.8, p. 105 and Excerpt 4.11, p. 111). Still, they were not as successful in solving the problems as Cindy and Dana, who used more terms from the statistical register. At the same time, using Cindy & Dana's type of scripted answers was not a guarantee for correct answers. As was described in 4.3.6 on page 162, when Elena & Fran answered the Tranquilizer problem they used the same terms as Cindy & Dana. However, they used them in a way that gave them an incorrect final answer in spite of this pair solving the rest of the problem correctly.

Also, for future use of statistical hypothesis testing, students need to move beyond the kind of scripted answers that Cindy & Dana used. Dana, herself, expressed some of her dissatisfaction with the scripted answers when asked about how you arrive at a conclusion from a statistical decision. She said, "It is that terrible evidence thing again." Also, Sylvia was not satisfied with her correct, scripted answer to the Poll problem, because she found the answer too convoluted. However, as students are learning statistical hypothesis testing, for them to use the appropriate vocabulary seems to help them in the process. Therefore following the suggestions in section 5.2 might be good instructional strategy. If students use the phrases supplied

by *Workshop Statistics* on page 480 regarding different p-values' implications, then students might better understand hypothesis testing results.

5.4 Students Did Not Include Probability Theory in Their Reasoning About Hypothesis Testing

As mentioned earlier, most students articulated well why one needs to draw samples in order to say something about populations. In section 4.1.6, it was also shown that students were well aware that samples need to be unbiased. However, most students seemed to equate randomness with representativeness and did not recognize the mathematical character of random sampling. Therefore the hypothesis testing procedure was somewhat of a black box that when fed the appropriate information would give the correct answer. The students seemed to trust their instructors and the textbook to give them a procedure that would produce the right answer. The connection with the probability theory that the students studied right before the inferential statistics seemed to be missing. In contrast to their instructors and textbooks, students *never* used the expressions "sampling distributions" or "Central limit theorem" during their problem solving.

One might argue that students do not really need to understand the probability theory behind hypothesis tests in order to carry out such tests. However, the difficulties students have to write their final answers suggest the contrary. Such difficulties have been documented in this dissertation and are well known by statistics instructors. If the answers to hypothesis questions do not naturally follow from

students' reasoning about their problem solving, then there is something wrong with the reasoning. When students can work a problem correctly all the way except for writing the final answer (e.g. 4.3.6, p. 168), they lack the understanding, which would allow them to use their statistical knowledge competently in their future academic career or in interpreting statistical results a informed citizens.

5.5 The TI-83 Calculator Was Used More Frequently by the Students Than by Their Instructors

In addition to their textbooks and their teachers' instruction, the students' reasoning was also shaped by their use of calculators. Students used their calculators much more than their instructors did. The instructors only used their calculators at one step of the hypothesis test procedure, to compute the p-value. The students used them at several different occasions and for several different purposes. All student pairs, at some time, were calling in different hypothesis testing programs to see which one requested the kind of information given in the problem. Even when students were not actually calling in the programs, they were speaking in terms of those programs. The student conversations indicated that the calculator menus and programs helped students construct their statistical knowledge.

The TI-83 calculator is required of all students taking statistics at the college where the study was conducted. This calculator has built-in programs that will compute the p-value for all the hypothesis tests taught in Elementary Statistics courses. The first step to finding the p-value is to call in the menu that is called

TESTS. This menu also contains confidence interval selections, but the tests are listed first. Therefore it is much more common that students incorrectly select a hypothesis test instead of a confidence interval, instead of the other way around. For example, in this study a pair only once incorrectly selected to use a confidence interval. Gus & Hal selected to do a one-proportion z-interval on the Coin problem in spite of me telling the students that all the problems were hypothesis test problems.

After a test has been selected, a list of variables appears on the screen. For the One Sample Proportion Test the calculator asks for the sample size n and for the number of "successes", called x. Similarly, the One Sample Z-tests and T-tests ask for sample size n, sample means ($\bar{x}$), and standard deviations (σ, s). Those variables are listed as letters or symbols without any words. The same is true for the hypothesized parameters called p_o and μ_0. The calculator also asks for the alternative hypothesis. Again, due to the small calculator screen, the words "alternative hypothesis" are not on the display. Instead the user is asked to select from the three symbols: $<$, $>$ and $\neq$.

After entering the required information, the user selects "CALCULATE" and the calculator will display a list of information. In the case of the One Sample Proportion Test the calculator will list the alternative hypothesis, the z-value, the p-value, the $\hat{p}$-value and the sample size. For the One Sample Z- and T-tests the calculator returns the alternative hypothesis, the z-value or t-value, the p-value, the sample mean and the sample size.

All the students spent time trying different tests to see which might fit with the given information. They also looked at the calculator display trying to judge if the

resulting lists made sense in the context of the problem. For example, extremely high or extremely low p-values often raised suspicion that some mistake had been made in choosing the test or matching variables with given information. Some student pairs spent a large amount of time trying out different scenarios on the calculator, while others spent less. However, no student pair used their calculators only to find p-values the way the instructors did.

This study was not designed to investigate calculator use in introductory statistics classes. Still, the findings suggest that the calculators, like the TI-83, could be used not only for computational purposes, but also as an instructional tool. The students in this study modeled how the TI-83 could be used for instructional purposes. For example, discussions of population mean versus sample mean could be organized around the calculator's Z-test screen the way students did in this study. Similarly, discussions regarding the difference between p_0, $\hat{p}$ and p-values could be organized around calculator displays in ways also modeled by the students. Since special overhead projectors are available for instructors using the TI-83 in their classrooms, such discussions are easy to arrange.

There are articles suggesting use of graphing calculators in introductory statistics courses (e.g. Iossif, 1999). However, there does not seem to be any systematical, empirical studies about what such use means for student learning. A study of students' interaction with their graphing calculators seems to be prudent if graphing calculators continue to be used in introductory statistics classes at the extent they are now.

5.6 Comparison Between the Findings in this Study and Other Studies'

Garfield's (2002) article about statistical reasoning described many of the characteristics that students showed in this study. In particular, the students in this study showed a lack of integration of statistical concepts into a model – a lack that seems consistent with Garfield's findings. In this study, almost all of the problems that were *not* marked as correctly done had several pieces that were done right. Also, students were often able to answer correctly when asked about concepts, but were not always able to apply those concepts correctly. When incorrect solutions were discussed with me after the problem solving, students often needed a minor hint to see where they had gone wrong.

Though much of the student behavior described by Garfield (2002) was found in this study, none of the students in this study showed enough consistency to be placed at any one of Garfield's statistical reasoning levels. Her model of statistical reasoning consists of five levels: idiosyncratic, verbal, transitional, procedural, and integrated process reasoning. Those five levels were described in this dissertation's literature review. None of the students in this study performed at Garfield's integrated process reasoning level – the highest level. However, most of the behaviors described in Garfield's four lower levels of statistical reasoning were observed in the study. On the other hand, it does not seem possible to place any of the students at only one particular level. Two examples below will exemplify how the same student would

exhibit characteristics of different levels of statistical reasoning within the same research session.

The first example concerns Gus. In the result chapter two instances were described at which Gus behaved at two different levels (see section 4.2.6, p. 132 and section 4.1.3, p. 96). He explained that one rejects the null hypothesis when p is less than alpha because then the μ tested in the null hypothesis is unlikely to be true. Judging from this explanation Gus seemed to be at the fourth of Garfield's levels. She described level 4 as, "The student is able to correctly identify the dimensions of a statistical concept or process but does not fully integrate them or understand the process" (¶ 4). Yet, in the Checkbook problem Gus entered the sample mean in the alternative hypothesis i.e. was not able to apply his knowledge to actual behavior - a characteristics of Garfield's level 2.

The second example of a student showing behaviors from two different levels concerns Tracy. As related in section 4.1.7, Tracy showed good understanding of hypothesis testing in discussing the Coin problem, when she said the purpose of the hypothesis testing was to find out if the coin were showing *fairly consistent* numbers of heads and tails. Those utterances by Tracy seemed to suggest an understanding of hypothesis testing at level 4. Yet, later in the same problem, Tracy entered the alpha value for p_0 in her calculator i.e. "scrambled a symbol with unrelated information" (Garfield, 2002, ¶ 4) - a characteristics of Garfield's level 1.

Those examples show that Garfield's (2002) model does not apply very well to this study's results. As was discussed in the preceding sections, this study's students

had developed some of the hypothesis testing ideas well. Other ideas were almost completely missing. Garfield's emphasis on statistics learning as integrating different concepts is very much in line with the results of this study. However, the way she defined her levels makes it not possible to use her model on this study's findings.

Garfield (2002) did not concern herself about how factors like intuition and metacognition affected students' reasoning, some issues that were considered in this study. One way that students showed good intuition concerned the probabilistic character of their answers. The students expressed, in different ways, that the results they arrived at by hypothesis testing were not definitive. Several students also understood that a low p-value signifies a rare event. The students in this study showed awareness of the uncertainty built into hypothesis testing. In one sense, this awareness contrasts with student reasoning in some probability education studies. For example, Konold (1989) found that an "outcome" approach was common in some students when they were asked to discuss predictions based on probabilities. In his study those students responded to probabilistic statements as if they were true with certainty. At the same time, my study's students' tendency to not much consider sampling variability in their problem solving resembles Konold's outcome-oriented students' way of thinking.

The students in the present study were interviewed by me either towards the end of their statistics courses or after those courses were completed. By that time it seemed that the students had gained some probabilistic reasoning from the instruction in their introductory statistics courses. However, in order to know if this ability to

reason about decisions under uncertainty indeed was a result of the instruction, a future study would have to test the same students before and after they had taken their introductory statistics course.

Also, the students in this study often used intuition in constructive ways. For example, three pairs did some preliminary computations using the list of weights in the Sugar machine problem and used their intuition to conjecture the correct answer to the problem. As reported in the literature review, probability education research often has found students' intuitive ideas to be contrary to mathematically derived probabilities. In this study incorrect solutions were caused by a failure to use intuition, rather than by faulty intuition. At times students' processes became too mechanical, and careless mistakes were not caught.

In the results of this study, intuition and metacognition were intimately related. Fischbein (1987) defined “intuitive knowledge [as being] immediate knowledge; that is a form of cognition, which seems to present itself to a person as self-evident” (p.6). When students were working the problems in this study, there were times when given information or achieved results were counterintuitive to them. In those cases, what students had in front of them conflicted with what they saw as self-evident. Most of the time their intuition was correct and the reason for the counterintuitive result were mistakes earlier in their problem solving. When the students chose to use their intuition, it worked as a control mechanism or metacognition.

In much of the problem solving, one of the students in the pair, not always the stronger one, made a running commentary on the problem solving process, while the other student interjected questions and comments. This method of problem solving, used by the pair, supported the students' metacognition in at least two ways. First, the student leading the problem solving would sometimes question her or his own statements, often stopping with a "Wait!". Second, the other student listening to the commentary would ask her or his partner questions that would make both students stop to think and clarify their reasoning. All the students came from classes in which small group work was part of the instructional practice, a fact that might have made it natural for them to talk about their work in this study. Observing them clarify their reasoning in conversations with each other confirmed the findings found in Aquilonius (2002). As in this earlier study, the observation of students' metacognition and clarifying of thought shed light on why small group work in introductory statistics classes produces more competent statistics students than lecture classes (Bonsangue, 1994; Borresen, 1990; Giraud, 1997; Magel, 1998 Potthast, 1999).

Metacognition and intuition were important categories in Schoenfeld's (1985b, 1992) model of mathematical problem solving as described in the literature review of this dissertation. Another category of importance was heuristics. Among the ones mentioned by Schoenfeld, the students used "drawing a picture" most frequently. Most students drew sampling distribution graphs with their problems, but only two pairs actually used them to make their decisions. The students instead relied on algebraic rules, such as "reject H_o when the p-value is less than alpha" for their

statistical decisions. One student, who expressed a preference for graphical representations (Elena), labeled her graphs incorrectly in a way that would have given incorrect results if she used them for decisions. The reluctance to use the graphs for decisions by many of the students and the incorrect graph by one student supports a major claim of other statistics education researchers: Students have a poor understanding of the central limit theorem, on which hypothesis testing builds. Of course, if students become frequent users of statistical hypothesis testing, using the algebraic rules become routine. However, when students are just learning inferential statistics it seems a pity that the graphical method of making decisions does not have more of a following. If students were willing to make the effort of doing graphs comparing alpha- and p-values as areas, one would expect more meaning to be attached to the hypothesis testing procedure (cf. Hong & O'Neil Jr., 1992).

Schoenfeld (1985b) also had a category called "beliefs" as part of his model. He lamented that school mathematics consists so much of rules and so little of understanding. The students in this study also believed that statistics mainly consisted of rules. They acknowledged that their instructor wanted them to understand the ideas behind the rules. Several students tried hard to understand the statistical ideas but felt that they were not very successful and therefore preferred to rely on rules. My opinion is that most of the students understood more than they were willing to give themselves credit for. In two cases the students' low opinion about their competency resulted in an anxiety that interfered substantially with their performance in the study. The anxiety expressed itself differently for the two students. One student (Elena)

talked incessantly during the research session in a way that not only prevented her, but also her partner from thinking clearly. The other student (Ursula), from another pair, would periodically stop participating in the conversations when her anxiety became overwhelming, saying that her brain had stopped working. She told me that her brain would sometimes cease to operate in the same way during tests.

Anxiety might be the most researched factor that influences introductory students' statistical reasoning. Even if this study was not designed to study statistics anxiety, my observations were consistent with other studies in the field. For example, one of Oathout's (1995) findings were also seen in this study; negative prior experiences in previous mathematics courses negatively affected students' reasoning in this study.

5.7 Students' Reasoning About Simulations Needs to Be Studied

The difficulty students have in grasping sampling distribution theory was documented in the literature review (Mendez, 1991; Tversky & Kahneman, 1971, 1982) and in the results of this study. As also mentioned in the literature review, simulation is the main instructional strategy recommended for building meaning with respect to sampling distributions (Gnanadesikan et al., 1997; Gourgey, 2000; Sterling & Grey, 1991; West & Ogden, 1998). The *Workshop Statistics* curriculum is designed around simulations. Therefore it is worth noting that all the students in this study having used the *Workshop Statistics* as there textbook drew correct final conclusions from their decisions to reject or not to reject the null hypothesis. The only pair in the

Understandable Statistics classes that also drew all correct conclusions from their statistical decisions was Cindy & Dana. An interesting aside is that Maria & Nancy, from a *Workshop Statistics*, class answered both the Checkbook problem and the Coin problem correctly, though their work leading up to the conclusions was faulty. Their correct answers could be taken as additional support for the claim that the simulations in the *Workshop Statistics* curriculum helped build statistical intuition.

However, I could not detect any direct conceptual impact from the simulation exercises on the students' reasoning. The students rarely mentioned their class simulations in their conversations with each other or with me. When simulations were spoken of, the students' associations between class simulation exercises and this study's problems seemed superficial. For example, when Vera & Zoe were asked to do the Sugar Machine problem it reminded Zoe of a simulation her class had done estimating proportions of differently colored Reese's Pieces. Her comments indicated that she connected the Sugar Machine problem with the Reese's Pieces problem because they both concerned random samples. However, the idea of randomness did not enter the pair's further discussion.

On one hand, there exists anecdotal evidence, from this study as well as in statistical education literature, that simulations help students to build probability and statistical intuition. For example, the *Workshop Statistics* students in this study did well on answering hypothesis test questions if they had overcome earlier obstacles in the problems. On the other hand, no systematic study of the effects of different kinds of statistical simulations on student thinking seems to exist. There are quite a few

articles describing simulation exercises that the authors claim help students understand sampling distributions and hypothesis testing, but those articles do not contain evaluation data. The article by Anderson-Cook and Dorai-Raj (2003) exemplifies this kind of article. The authors describe some computer applets and related exercises that they have used in their introductory statistics classes. The authors write, "Having the students work with the applets themselves has a dramatic impact on helping them to reinforce the concepts and to better prepare them to be able to solve problems that they are likely to encounter after completing the introductory statistics course"(¶ 3), and "*Anecdotally* [italics added], students' performance on test questions related to the concepts of power, sample size and hypothesis testing in recent years has improved "(¶ 5). Studies, using the same methods as this one, with detailed analysis of students' reasoning about their simulation activities are needed.

Instructors know what ideas simulations are designed to illustrate. Students most often do not know the point of a simulation that they are asked to perform and might not have the conceptual schema in which to place its result. In addressing instructors, the *Workshop Statistics* authors both encourage them to "Allow students to discover." (p. xxi) and to "Be proactive in approaching students." (p. xxii). I agree that a balance between those two recommendations is good instructional strategy. However, the more that is known about how students reason about the activities now given to them, the more productive instructors can be in striking such a balance.

5.8 Significance and Limitations of Study

This study was based on data from community college students. Thus the study's results apply to many, many students. At my college, 15 sections of Elementary Statistics are taught every semester to approximately 35 students each, a total of about 525 students. During Fall 2003, approximately 31,000 students were enrolled in 882 California community college statistics classes (T. Lu, California Community Colleges Chancellor's Office, personal communication, July 30, 2004.)

As mentioned earlier many of those students need the course in order to transfer to a university. Thus for them, as well for society at large, any research results that might improve statistics learning are significant. In addition, there are more intangible benefits to society from a more statistically literate populace. For example, if people understand statistics, they can see through advertisers' misrepresentations and judge the validity of politicians' claims more easily.

However, caution should be used in drawing conclusions about other student populations than community college students. The average community college student tends to be academically less prepared than, for example, a student, who has been accepted into a competitive university directly from high school. Many community college students have a weak mathematics background with accompanying mathematics anxiety. Therefore community college students can be expected to have a harder time learning statistics than their university counter parts.

As stated in the methods chapter, this study was not designed to analyze how particular instructional methods compare in effectiveness of delivering statistics

instruction. Instead the analysis was centered on student reasoning that seemed consistent across variables such as instructional methods and students' demographic characteristics. Future research could extend my study by systematically analyzing some of those variables' effects on students' statistical reasoning. This study did not allow conclusions whether a particular instructor behavior or textbook presentation might have led to certain student mistakes. For example, each of the instructors taught only out of one textbook. Therefore one could not conclude whether the instructor or the textbook might account for particular student difficulties.

Sections 5.5 and 5.7 discussed two areas that would also benefit from further study. Firstly, the prevalent use of graphing (and programmable) calculators in the statistics classroom has changed the way statistics is taught, but little is known about how this instructional practice is affecting student reasoning. Secondly, simulations are used extensively as part of statistics instruction, but much more needs to be known about what students do cognitively with those simulations.

Methods such as were used in this study would lend themselves well to studies about students' reasoning about simulations. Digital videotaping provides a powerful tool for probing student cognition. With a greater knowledge about what students do with their present instruction, future instruction can be improved.

REFERENCES

Allwood, C. M. (1990). Justification and choice of solution method. *Scandinavian Journal of Psychology, 31*, 182–90.

Anderson-Cook, C. M., & Dorai-Raj, S. (2003). Making the concepts of power and sample size relevant and accessible to students in introductory statistics courses using applets. *Journal of Statistics Education, 11*. Retrieved August 1, 2004 from http://www.amstat.org/publications/jse/v11n3/anderson-cook.html

Aquilonius, B. C. (2002). *Students' peer discussions of statistics. How do students learn from them?* Unpublished master's project, University of California, Santa Barbara.

Brase, C. H., & Brase C. P. (2003*). Understandable statistics. Concepts and methods* (7th ed.). Boston, MA: Houghton Mifflin.

Bonsangue, M. V. (1994). Symbiotic effects of collaboration, verbalization, and cognition in elementary statistics. In J. J. Kaput & E. Dubinsky (Eds.), *Research issues in undergraduate mathematics learning. Preliminary analysis and results. MAA notes Number 32.* (pp. 107–117). Washington, DC: Mathematics Association of America.

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

Burton, L. (1999). Why is intuition so important to mathematicians but missing from mathematics education? *For the Learning of Mathematics, 19*(3), 27– 32.

Chance, B. L., & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal, 1*, 38–41. Retrieved September 3, 2004 from http://www.stat.auckland.ac.nz/~iase/serj/SERJ1(2).pdf

Chatterjee, S., & Hawkes, J. (1996). Statistics and Intuition for the Classroom. *Teaching Statistics, 18*(2), 34–38.

Cobb, P. (1989). A double-edged sword. [Review of the book Intuition in science and mathematics: an educational approach.]. *Journal for Research in Mathematics Education,* 20, 213–218.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: The MIT Press.

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

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301–318.

Fischbein, E. (1987). *Intuition in science and mathematics: an educational approach*. Dordrecht, Holland: D. Dreidel Publishing Company.

Gal, I., & Ginsburg, L. (1994). *Journal of Statistics Education, 10.* Retrieved July 30, 2004 from http://www.amstat.org/publications/jse/v2n2/gal.html

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10.* Retrieved July 18, 2003, from http://www.amstat.org/publications/ jse/v10n3/garfield.html

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education, 19*, 44–62.

Garfield, J., Hogg, B., Schau, S., & Whittinghill, D. (2002). First courses in statistical science: the status of educational reform. *Journal of Statistics Education, 10.* Retrieved July 18, 2003, from http://www.amstat.org/publications/ jse/v10n2/garfield.html

Giraud, G. (1997). Cooperative learning and statistics instruction. *Journal of Statistics Education, 5*. Retrieved January 2, 2001, from http://www.amstat.org/publications/ jse/v5n3/giraud.html

Gnanadesikan, M., Scheaffer, R. L., & Swift, J. (1987). *The art and techniques of simulation.* Palo Alto, CA: Dale Seymour Publications.

Gnanadesikan, M., Scheaffer, R. L., Watkins, A. E., & Witmer, J. A. (1997). An activity-based course. *Journal of Statistics Education, 5.* Retrieved January 2, 2001, from http://www.amstat.org/publications/ jse/v5n2/. gnanadesikan.html

Gourgey, A. F. (2000). A classroom simulation based on political polling to help students understand sampling distributions. *Journal of Statistics Education, 8.* Retrieved June 15, 2003, from http://www.amstat.org/publications/ jse/v8n3/gourgey.html

Halliday, M. A. K. (1978). *Language as a social semiotic. The social interpretation of language and meaning.* London: Edward Arnold.

Hong, E., & O'Neil Jr., H. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology, 84,* 150–159.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic books, Inc.

Iossif, G. (1999). The graphics calculator as a teaching aid in statistics. *Teaching statistics, 21(*2), 45-48

Johnson, D. E. (1989). An intuitive approach to teaching analysis of variance. *Teaching of Psychology, 16*(2*),* 67–68.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 32–47). New York: Cambridge University Press.

Konold, C. (1989). Informal Conceptions of Probability. *Cognition and Instruction,* 6, 59–98.

Lancy, D. F. (1993). *Qualitative research in education: An introduction to the major traditions*. New York: Longman.

Landwehr, J. M., Swift, J., & Watkins, A. (1987). *Exploring surveys and information from samples.* Palo Alto, CA: Dale Seymour Publications.

Lester, J. K. (1985). Methodological considerations in research on mathematical problem-solving instruction. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving* (pp. 41–69). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Magel, R. C. (1998). Using cooperative learning in a large introductory statistics class. *Journal of Statistics Education, 6.* Retrieved January 2, 2001, from http://www.amstat.org/ publications/jse/v6n3/magel.html

Marshall, C., & Rossman, G. B. (1999). *Designing qualitative research*. Thousand Oaks, CA: Sage Publications

Mendez, H. (1991). Understanding the central limit theorem. (Doctoral dissertation, University of California, Santa Barbara, 1991). *Dissertation Abstracts International, 53,* 2722.

Mevarech, X. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics,14,* 415–429.

Moore, D. S. (2000). *The basic practice of statistics.* New York: W. H. Freeman and Company.

Moschkovich, J. M. (1992*). Making sense of linear equations and graphs: An analysis of students' conceptions and language use.* Unpublished doctoral dissertation, University of California, Berkeley.

Moschkovich, J. N., & Brenner, M. E. (2000). Integrating a naturalistic paradigm into research on mathematics and science cognition and learning. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 457– 486). Mahwah, NJ: Lawrence Erlbaum.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (1991*). Professional standards for teaching mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2000*). Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

Newman, C. M., Obremski, T. E., & Scheaffer, R. L. (1987). *Exploring probability*. Palo Alto, CA: Dale Seymour Publications.

Oathout, M. J. (1995, April). *College students' theory of learning introductory statistics: phase one.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Pimm, D. (1987). *Speaking mathematically: Communication in mathematics classrooms.* London: Routledge.

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

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12,* 191–204.

Polya, G. (1988). *How to solve it.* Princeton, NJ: Princeton University Press.

Potthast, M. J. (1999). Outcomes of using small-group cooperative learning experiences in introductory statistics course. *College Student Journal, 33*(1), 33–42.

Pressley, M., & McCormick, C. B. (1995*). Cognition, teaching, & assessment*. New York: Harper Collins.

Quilici, J. L. (1997). The influence of expertise and schema training on how students categorize statistics word problems. (Doctoral dissertation, University of California, Santa Barbara, 1997). *Dissertation Abstracts International, 58,* 3941.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144–161.

Reid, A., & Petocz, P. (2002). *Journal of Statistics Education, 10*. Students' conceptions of statistics: A phenomenographic study. Retrieved December 21, 2004, from http://www.amstat.org/publications/jse/v10n2/reid.html

Resnick, L. B. (1991). Shared cognition: Thinking as social practice. In L. B. Resnick, J. M. Levine, & S. T. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 1–20).
Washington, DC: American Psychological Association.

Rossman, A. J., & Chance, B. L. (2000). *Workshop statistics. Discovery with data and minitab.* Emeryville, CA: Key College Publishing Press.

Rossman, A. J., Chance, B. L., & von Oehsen, J. B. (2002). *Workshop statistics. Discovery with data and the graphing calculator*. Emeryville, CA: Key College Publishing Press.

Schoenfeld, A. H. (1985a). Making sense of "out loud" problem-solving protocols. *Journal of Mathematical Behavior, 4,* 171–191.

Schoenfeld, A. H. (1985b). *Mathematical problem solving*. New York: Academic Press.

Schoenfeld, A. H. (1987). What's all the fuss about metacognition? In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 189–215). Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.*), Handbook of research on mathematics teaching and learning* (pp. 334–370). New York: MacMillan.

Selden, J., Selden, A., & Mason, A. (1994). Even good calculus students can't solve nonroutine problems. In J. J. Kaput & E. Dubinsky (Eds.), *Research issues in undergraduate mathematics: Learning. Preliminary analysis and results. MAA notes number 32.* (pp. 17–26). Washington: Mathematics Association of America.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: MacMillan.

Sterling, J., & Gray, M. W. (1991). The effect of simulation software on students' attitudes and understanding in introductory statistics. *Journal of Computers in Mathematics and Science Teaching*, *10* (4), 51–56.

Sutarso, T. (1992a, November). *Some Variables in Relation to Students' Anxiety in Learning Statistics.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Sutarso, T. (1992b, November). *Students' Attitudes toward Statistics.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Knoxville, TN.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Verkoeijen, P. P. J. L., Imbos, T., van de Viel, M. W. J., Berger, M. P. F., & Schmidt, H. J. (2002). Assessing knowledge structures in a constructive statistical learning environment. *Journal of Statistics Education, 10*. Retrieved July 16, 2003, from http://www.amstat.org/publications/jse/v10n2/ verkoeijen.html

Vygotsky, L. (1962). *Thought and languag*e. Cambridge, MA: The M. I. T Press.

Vye, N. J., Goldman, S. R., Voss, J. F., Hmelo, C., & Williams, S. (1997). Complex mathematical problem solving by individuals and dyads. *Cognition and Instruction, 15,* 435–484.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22*, 366–389.

Well, A. D., Pollatsek, A., & Boyce, S. (1990). Understanding the effects of sample size on the mean. *Organizational Behavior and Human Decision processes, 47,* 289 –312

West, R. W., & Ogden, R. T. (1998). Interactive demonstrations for statistics education on the World Wide Web. http://www.amstat.org/publications/jse/v6n3/west.html

Wolfe, C. R. (1992). Flipping frisbees and finding flowers: Developing statistical intuition. *College Teaching, 40*(2), 60.

Appendix I. Hypothesis Test Problems for Students

1. Checkbook Problem (Moore, 2000, p.299 and p. 331)
In a discussion of the educational level of the American workforce, someone says, "The average young person can't even balance a checkbook."
The NAEP survey includes a short test of quantitative skills, covering mainly basic arithmetic and the ability to apply it to realistic problems. The NAEP survey says that a score of 275 (out of 500) reflects the skill needed to balance a checkbook. An NAEP random sample of 840 young men (between 21 and 25 years) yielded a mean score of 272 with a standard deviation of 60. Is this sample result good evidence that the mean for all young men is less than 275?

2. Home Value Problem
A city council member said that 18% of all homes in the city had been undervalued by the assessor's office. The local newspaper conducted a random sample of 98 homes and found that 26 had been undervalued. At $\alpha = 0.05$, test the claim that the proportion of undervalued homes in the city is different from 18%.

2. Coin Problem
You suspect that a certain coin, when tossed, favors heads. You toss it 50 times and find 31 heads. At the 0.05 significance level, does it favor heads or is it a fair coin?

4. Home Loan problem
During 1995, the average loan for purchasing a home in Greentown, California, was $235,000. The price of homes has increased since then. Using a significance level of 0.01, test the hypothesis to determine if the average loan for purchasing a home has increased significantly. A random sample of 81 recent home loans produced an average loan of $265,000 with a standard deviation of $25,500.

5. Sugar Machine problem (Paraphrased from Khazanie, 1990, p. 414).
A machine can be adjusted so that when under control, the mean amount of sugar filled in a bag is 5 pounds. To check if the machine is under control, six bags were picked at random and their weight (in pounds were found to be as follows:
5.3 5.2 4.8 5.2 4.8 5.3
Has the machine slipped out of control?

6. Tranquilizer Problem
In an experiment with a new tranquilizer, the pulse rates of 25 patients were taken before they were given the tranquilizer, then again five minutes after they were given the tranquilizer. Their pulse rates were found to be reduced on the average by 6.8 heart beats per minute with a standard deviation of 1.9. Using the 0.05 level of significance, what could we conclude about the claim that this tranquilizer will reduce the pulse rate on the average by 7.5 beats per minute?

Appendix II. Diagnostic Problems

1. Exercise Problem.
A student was given the following problem:

Twenty subjects are randomly assigned to an experimental condition and a control condition. The ten experimental subjects do aerobics exercises for 10 weeks. The ten control subjects do not exercise during the ten weeks. After the ten weeks all the subjects are brought into the lab and asked to solve mental arithmetic problems. The measure of stress is the subjects' heart rate during the task. The mean heart rate for the experimental group was 79.40 with a standard deviation of 5.25. The mean heart rate for the control group was 85.40 with a standard deviation of 6.69. Does this data indicate that aerobics exercise will help subjects to tolerate stress if the reaction to stress is measured in terms of heart rate?

The student sets up the hypotheses as:

$H_0 : \bar{x}_1 = \bar{x}_2$ and $H_a : \bar{x}_1 > \bar{x}_2$

Do think that is right? If you don't think it is right, please correct the student's work and verbally explain to the student why the way the student set up the hypothesis was wrong.

2. Jail Problem
A student is given the following problem:

Pre 1990 records show that the average time in jail spent by a first time convicted burglar was 2.5 years. A random sample was taken to see if the average time increased in the 1990's. From the sample of 25 first time convicted burglars in the 1990's, the average length of time in jail was 3 years with a standard deviation of 9 years. Did the average length of jail time increase in the 1990's?

The student sets up the hypotheses as:
$H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 < \mu_2$

Do think that is right? If you don't think it is right, please correct the student's work and verbally explain to the student, why the way the student set up the hypothesis was wrong.

3. Poll Problem

A student was given the following problem:

In a survey conducted by Louis Harris of LH Research 1250 US adults were polled regarding their view on banning handgun sales. The results were that 650 of those sampled favored a ban. At the 5% significance level, do the data provide sufficient evidence to conclude that a majority of US adults (i. e. more than 50%) favor banning handgun sales?

The student worked the problem correctly all the way to the last step, including setting up $H_0 : \theta = .5$ and $H_a : \theta > .5$. Also the student correctly failed to reject the null hypothesis because the p-value (.0786) is greater than .05.

At the end the student says, "Yes, a majority favors the ban, because $\hat{p}$ = .52".
Do think that is right? If you don't think it is right, please correct the student's work and verbally explain as to the student, why the way that she or he answered the problem was wrong.

4. Gas Price Problem

Suppose a statistics instructor asked her students to compare gas prices in Santa Clara County with those in Santa Cruz County, and decide if there was a significant difference in gas prices between the two counties. Two students, working together on the project, decided on the following procedure: Since they were living and going to school in Santa Clara County they decided to get their sample from Santa Clara County by recording gas prices as they went about their business during the week. On the weekend, they would go over to the beach in Santa Cruz, and would record their sample gas prices on their way to and from their destination in Santa Cruz.

(a) Would you consider the samples that the students collected to be random samples?

(b) Suppose that a consumer organization would want to decide if there is a significant difference between gas prices in Santa Clara County and Santa Cruz County. Suppose the organization has quite a bit more resources in terms of money and time than the students have. How would you recommend that the consumer organization collect their random samples?