The Role of Counterintuitive Examples
in Statistics Education

by

Lawrence Mark Lesser, B.A., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 1994

The Role of Counterintuitive Examples
in Statistics Education

Approved by
Dissertation Committee:

_____

_____

_____

_____

_____

ACKNOWLEDGMENTS

I wish to give recognition and profound thanks to the many people who have seen me through this work and stage of my lengthy academic career:

To my family for their many contributions, especially: my parents, Dr. and Mrs. Herbert Lesser, whose support made it easier for me to complete my education coursework and dissertation in a timely manner; my sister Lori, for taking the train to visit with me during the weekend of my first out-of-state job interview; my mother's parents, Dr. and Mrs. Bernard Farfel; my father's mother, the late Julia Lesser, who taught mathematics with distinction for over 25 years in the Fort Worth schools, and whose memory I hope to honor with my service in the field; and her sister, the late Sadie Streusand, a teacher and guidance counselor in the Houston schools, who encouraged my intial involvement in extracurricular mathematics organizations and contests.

To the distinguished members of my dissertation committee: my supervisor (and expert solver of puzzles), Dr. Ralph W. Cain, for the consistent generosity of his time, experience and humor from the day he welcomed me to the program., Dr. L. Ray Carry, for his sense of integrity and humility that helped me deepen my respect for the power and pitfalls of education research, as well as for many enjoyable and inspiring mathematical software demonstrations and philosophical discussions on "Truth versus truth"; Dr. Charles E. Lamb, for telling it like it is, showing me how "academia is more like the real world than anything else is," and encouraging me to "just start writing" at a time when I was stalled by the inertia of too many options; Dr. Margaret E. Myers and Dr. Mary R. Parker, for their collegial sharing of advice and curricula materials both as fellow statisticians and as fellow educators. I appreciate their timely ongoing feedback and support of study that in some ways was atypical for the department and for someone with my quantitative background.

To all of the instructors I have had in the College of Education for entertaining my many questions inside and outside the classroom. I am also grateful that virtually all of my courses encouraged significant interaction with my peers, a healthy balance between theory and practice, and immediate immersion in the literature. In fact, I recall quite vividly the moment in Dr.

working at the University of Northern Colorado, I know I would have also enjoyed being at any of those other schools.

To the 500-plus students I have taught so far in my career—your questions and answers have been wonderful motivators in my growth as an educator, a researcher, and a human being. And to David Wilcox, for approaching his work with creativity, honesty, love and purpose that inspires me to strive towards more of these qualities in my own work.

The Role of Counterintuitive Examples

in Statistics Education

The purpose of this study was to develop a theoretical model for the use of counterintuitive examples in the introductory non–calculus-based statistics course at the college level. While intuition and misconceptions continue to be of great interest to mathematics and science educators, there has been little research, much less consensus or even internal consistency, in statistics curriculum development concerning the role of examples with counterintuitive results. Because the study intended to provide educators with useful connections to content, instructional methods (e.g., cooperative learning) and learning theory constructs that have been successfully used in mathematics or science education, the model that emerged was organized around a typical syllabus of topics.

The study critiqued and then reconciled "Traditional" and "Alternative" perspectives. The Traditional Position attempts to minimize possible confusion and frustration by avoiding such examples, while the Alternative Position uses them to motivate and engage students in critical thinking, active learning, metacognition, communication of their ideas, real-world problem solving and exploration, reflection on the nature and process of statistics, and other types of activities encouraged by current reform movements.

The study delineated specific criteria and conditions for selecting and using counterintuitive examples to achieve numerous cognitive and affective objectives. Examples explored include the Monty Hall problem, Simpson's Paradox, the birthday problem, de Méré's Paradox, the Classification Paradox, the Inspection Paradox, and required sample size. The study connected many of

these examples (especially Simpson's Paradox) with other counterintuitive examples, with known probability or statistics misconceptions many students have, with representations from other branches of mathematics, and with the constructivist paradigm.

Problematic issues addressed include difficulty in constructing assessment instruments and a multiplicity of terminologies and typologies. Additional directions for research were suggested, including several empirical investigations of various facets of the model. The connections, examples, and representations presented should be extremely useful for teachers of statistics, but should also enrich the pedagogy of teachers of other courses.

CONTENTS

CHAPTER 1

IMPORTANCE OF THE INVESTIGATION

1.1 Introduction

The purpose of this study was to develop a substantial theoretical syllabus-driven model for the use of counterintuitive examples in the introductory statistics course, as teachers and textbooks have expressed neither consensus nor even internal consistency with respect to their use. Even worse, Brewer (1985) lists several examples of best-selling introductory texts which contain any of five types of "myths and misconceptions." This study should be valuable to and readily used by both mathematics education researchers as well as classroom instructors. This is because connections are made both to instructional methods and to learning theory constructs, and because the model of the study builds on the familiar structure of a syllabus typical for the introductory statistics course.

While some connections with content from other mathematics and science courses will be mentioned, the focus of this study is the introductory non–calculus-based statistics course. Also, while many aspects of this study will apply to the introductory statistics courses being taught in a small, but increasing number of high schools, this study focuses primarily on the college level.

Mathematics educators continue to express their interest in the role of intuition at many grade levels. For example, Resnick (1986, p. 162) states: "I will propose that early intuitions about number, although providing a foundation for varied performances, are actually based on a restricted range of mathematical principles. These principles, if activated in school contexts, would provide an intuitive basis for much, but not all, of the elementary school mathematics curriculum, and would need to be enlarged in important ways to support secondary school mathematics. Thus, an expanded body of intuitive mathematical knowledge must be developed if intuitively based concepts are to continue to support learning beyond the first few years of school." Also, the masthead of the debut issue of *Mathematics Teaching in the Middle School* (1994, p. 4) declares:

"The focus of the journal is on intuitive, exploratory investigations that use informal reasoning to help students develop a strong conceptual basis that leads to greater mathematical abstraction."

This interest in intuition has implications for both instructional techniques and curriculum development. As an example of the latter, Fischbein (1987, pp. 212–214) lays out what he perhaps overdramatically calls "a profound, dialectic contradiction" that has played itself out in whether textbooks are dominated by pictorial or axiomatic development: "By exaggerating the role of intuitive prompts, one runs the risk of hiding the genuine mathematical content instead of revealing it. By resorting too early to a 'purified', strictly deductive version of a certain mathematical domain, one runs the risk of stifling the student's personal mathematical reasoning instead of developing it." This conflict is related to the conception of the role of proof in mathematics, which has changed over time (Barbin 1994).

Lee (1989) noticed a similar tension while teaching the introductory statistics course. Lee found that despite the fact that statistics content has been traditionally presented in a hierarchical "rational" sequence, students use a more intuitive style he called the "pattern-forming" mode of learning. For example, Lee's students were able to work hypothesis-testing problems without understanding the antecedent notions of sampling theory or the Central Limit Theorem.

Watts (1991, p. 290, emphasis in original) adds:

the major difficulty that confounds beginning students and inhibits the learning of statistics, and that distinguishes statistics from other disciplines such as mathematics, physics, chemistry, and biology, is that *the important fundamental concepts of statistics are quintessentially abstract. ... .* Can anyone *draw* a random variable? a mean? a variance? probability? ... . Even the most elementary statistics course, however, is concerned with drawing inferences about phenomena in the real world on the basis of data obtained from experiments. Consequently, students in elementary statistics courses must not only grapple with truly abstract concepts, but they must immediately relate and apply these concepts to reality.

While this study indicates ways in which statistics concepts such as means can be made more concrete than Watts suggests, Watts' comments nevertheless are an

additional reason why the role of intuition may be especially crucial in statistics education.

This chapter illustrates some of the ways in which intuition itself has received much attention, but there has been relatively little focus on the role of counterintuitive examples. As Chu and Chu (1992, p. 191) state: "The subject is subtle and probably more difficult than it appears. A seemingly trivial problem has been known to provoke very heated arguments among students, teachers, professional engineers, and scientists, all of whom can come up with apparently flawless arguments to support divergent conclusions."

Indeed, life itself is filled with important situations in which the "true situation" or the "correct action" seems contrary to one's initial intuition: Weightlifters' workouts build muscle by first tearing it down, a patient is inoculated for a disease with an injection of the associated virus, an airline passenger is told to secure her own emergency oxygen mask before attending to her child, medicines (or household cleaning chemicals) may be individually helpful but hazardous in combination, the longer of two long-distance calls (or flights) may actually be less expensive, a skidding driver is told to turn his wheels in the direction he is skidding, a batter is told he can hit a fastball farther than a ball pitched towards him at a slower speed, etc. Perhaps the state of affairs is best expressed by G. K. Chesterton (1959, p. 81): "The real trouble with this world of ours is not that it is an unreasonable world, nor even that it is a reasonable one. The commonest kind of trouble is that it is nearly reasonable, but not quite." Those that insist upon limiting curriculum to intuitive examples are denying Davis and Hersh's (1981, pp. 174–175) examples of how mathematics can and does involve going from order to order, order to chaos, chaos to chaos, and chaos to order. Furthermore, the current curriculum already contains many ideas that are not intuitive, as will be discussed in section 4.7.

Bringing this back to the specific field of concern is Fischbein (1987, p. 96, emphasis in original):

The student has to learn that in science and in mathematics not everything is intuitively understandable, visually or behaviorally representable, that many

statements express logical implications of generalizations going beyond the limited possibilities offered by the empirical, common conditions of our terrestrial life. If there is an intuition to be created here it is the intuition of the *non-intuitive*, the intuitive understanding of the fact that many concepts are by their very nature beyond our intuitive capabilities, although rationally valid. Such an intuitive understanding is also attainable by experience—the experience of the non-representable although intellectually manipulable notion. One lives the conflict and the displeasure, one lives the effort to overcome the conflict, one lives, finally, the acceptance as clear and intellectually consistent of the particular statement or notion. *Such an intuition expressed in accepting the non-intuitive as meaningful on logical grounds represents a fundamental acquisition of science and mathematics education .*

Like Watts' quotation, Fischbein's statement does not perfectly apply to our study in statistics education. For one thing, most examples which are initially counterintuitive can be eventually given an intuitive basis. Also, there is a distinction to be made (which is done in section 3.1) between non-intuitive and counterintuitive. Nevertheless, the spirit of this quotation helps create a context for discussion.

Because of differences in usage, it is necessary to clarify the usage of the term "statistics" that will be used. Shaughnessy (1992, p. 465) follows the European convention of using the word "stochastics" to include probability and statistics. Derry et al. (in press, p. 23) choose "instead to employ the more familiar term 'statistics', but to use it in the broadest sense to refer to both probability and statistics." Indeed, introductory college courses involving both probability and statistics are more likely to use the word "statistics" in the course title than to use the phrase "probability and statistics," and certainly never seem to use the word *probability* alone.

This study adopted Derry's convention because many if not most American researchers and classroom teachers seem unfamiliar with and confused by the use of the term "stochastics." In addition to familiarity and brevity, there is one additional reason for preferring the term "statistics" to "probability and statistics." There is a trend among statistics educators to emphasize data analysis (even perhaps renaming the course "introduction to data analysis") and keep probability theory to a minimal, as-needed basis. One final point is that although

some statistics educators (e.g., Moore 1988, 1993) feel that statistics should not be considered a branch of mathematics, this is a debate that will not be addressed in this study.

The role of intuition in university statistics education has not always received the attention it is now getting. As Shaughnessy (1992, p. 466) relates: "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 (Shaughnessy 1977). Thus, college level students, with all their prior beliefs and conceptual misunderstandings about stochastics, rarely get the opportunity to improve their statistical intuition … . University courses may, therefore, only make a bad situation worse, by masking conceptual and psychological complexities in the subject."

Educators now seem to be stressing the importance of statistics intuition. According to the National Council of Teachers of Mathematics (1989, p. 169): "Students must acquire intuitive notions of randomness, representativeness and bias in sampling to enhance their ability to evaluate statistical claims. These understandings would give students the appropriate tools for rejecting such television advertising claims as one that portrays a series of people choosing the same commercial toothpaste." The NCTM (1991, p. 137) also makes this recommendation for the preparation of middle school and high school teachers: "Potential misuses of statistics and common misconceptions of probability should be discussed." The need for new instructional approaches is in part called for by studies which "show that some misconceptions are quite widespread and can persist in spite of relevant information" (Garfield and Ahlgren, 1988, p. 51).

The importance of statistics intuition is reflected not only in the reform movement, but also in research, assessment, and curriculum development. According to Shaughnessy (1992, p. 465), "Intuitions, preconceptions, misconceptions, misunderstandings, non-normative explanations—whatever one might call them—abound in the research on learning probability and statistics." Joan Garfield and Cliff Konold have been developing (NSF Grant No. MDR-

8954626) an instrument called "Statistical Reasoning Assessment," which has an entire subtest called "Intuitive Thinking." There is even a body of research emerging on "intuitive statistical inference in infrahumans" (Shimp and Hightower, 1990)!

Exercises in some widely used current introductory statistics textbooks explicitly expose students to their misconceptions. For example, exercise 4.26 in Moore and McCabe (1993, p. 304) asks students which of the following sequences is the most likely outcome of rolling a die with four green and two red faces: RGRRR, RGRRRG, GRRRRR. The authors then tell the student that "[i]n a psychological experiment, 63% of 260 students who had not studied probability chose the second sequence. This is evidence that our intuitive understanding of probability is not very accurate." Other exercises (e.g., exercise 4.53, p. 337) force students not only to determine a correct answer, but also to "[e]xplain to the gambler what is wrong with his [incorrect] reasoning."

Operational definitions of intuition as well as a distinction between non-intuitive and counterintuitive are provided in sections 2.2 and 3.1, respectively. Shaughnessy (1992, p. 480) says that "[i]ntuitions can mislead and promote misconceptions of scientific reality, as well as provide simplifying cognitions of that reality." Therefore, an intuition of the counterintuitive "… is particularly important in some branches of mathematics such as probability and statistics in which many phenomena conflict with our initial cognitive beliefs."

This spirit seems to echo the broader statement of Westcott (1968, pp. 197–198):

If one leads an examined cognitive life, one finds that many of these shared logical constraints are probabilistic in nature, that is, they do not always hold. One finds that the information which can be gained in a situation does not always lend itself to the conventional treatment of a classical syllogism or progressive uncertainty reduction. Perhaps the most important constraints one can acquire concern the conditions under which other constraints should be followed and when they should not. An individual should eventually arrive at a point in education where he has a great many useful implicit and explicit constraints, but among them should be some programs which lead to the breaking of other constraints. This may often be the most important step in reaching the solution to a problem: knowing when to ignore the explicit information of the conventional

sources and when to ignore the conventional operations—knowing when to begin flying by the seat of one's pants, while others stare mutely at the obviously broken compass.

One application of this quotation to statistics involves helping students know when, for example, certain heuristics (e.g., availability, representativeness) they may have are appropriate as is and when they need to be modified, as is discussed in chapter 4.

The extent that statistics may have the plurality of counterintuitive results in mathematics may explain why little has been specifically written about counterintuitive results, as statistics education is a much newer field of research than mathematics education in general. Only recently (e.g., Cleary 1992, Lock and Lock 1993) are connections starting to be made between the two. Because researchers have noted some very real differences between mathematics and statistics (e.g., Moore 1993), it is still not clear whether all the work that has been done on mathematics intuition will readily transfer.

As discussed in sections 4.1 and 4.2, the Traditional Position in statistics education has been largely to avoid counterintuitive examples. Many educators and researchers in mathematics and science education, however, have argued such examples can make many significant positive contributions. This so-called Alternative Position is developed in sections 4.3 and 4.4. These two positions will be further analyzed and then reconciled in a new syllabus-driven paradigm in the remainder of chapter 4.

Perhaps the greatest testimony to the importance of having a clear perspective concerning the role of intuitive and counterintuitive examples in statistics education comes from Shaughnessy (1992, pp. 488–490). It is telling that most of the seven items identified in his suggestions for future research in statistics education explicitly refer to intuitions, conceptions or misconceptions. For example, one of Shaughnessy's recommendations is: "At the pre-service level, we will need to develop courses which meet [statistics] misconceptions and beliefs head on, and sensitize our prospective teachers to the prevalent misconceptions they can expect to encounter in their own students. The

instructional experiences we design in [statistics] for teachers should be informed by our research." Also, "Clinical teaching experiments that carefully document changes in students' [statistics] conceptions, beliefs, and attitudes over long periods of time are needed to obtain a clearer picture of the cognitive and affective development in [statistics]." Ulep (1990, pp. 59–60) catalogues some of these teacher misconceptions: "Many education majors who already had a course in statistics thought that the average of numbers that include zero is the same as the average with zero excluded (Mevarech, 1983). A majority of them compute the ordinary average in problems requiring the weighted average. This finding agrees with that of Pollatsek, Lima and Well (1981) … And some of them [preservice mathematics teachers] have difficulties with or misconceptions of permutations (Ball, 1988) … ." It turns out that grappling with a particular counterintuitive situation in statistics, namely Simpson's Paradox (which will be discussed in detail in section 3.3), may have a positive effect on such misconceptions involving weighted averages.

Clearly, there are implications of this study not only for pedagogical strategies (which are discussed in chapters 5 and 6), but also for curriculum design. Currently, textbooks handle this in a multitude of ways: including counterintuitive situations in the main body of the text, briefly describing them in optional enrichment sidebar sections, including one among the end-of-chapter exercises but omitting it from the expository part of the chapter, or omitting them entirely. For example, some introductory textbooks (e.g., Devore and Peck 1990) do not mention Simpson's Paradox at all, some discuss it in a section marked "optional" (e.g., Cryer and Miller 1991), and Moore and McCabe (1993, section 2.5) involve Simpson's Paradox in the only three-way table example in the text as well as in every three-way table exercise following that section! Also, there is not consistency between or even within textbooks with respect to whether or not to "telegraph" that an example will yield a counterintuitive result. (The "telegraphing" issue is discussed in section 4.6.)

Finally, no discussion of the power of paradox would be complete without this quotation from Rapoport (1967, p. 50):

Paradoxes have played a dramatic part in intellectual history, often foreshadowing revolutionary developments in science, mathematics, and logic. Whenever, in any discipline, we discover a problem that cannot be solved within the conceptual framework that supposedly should apply, we experience shock. The shock may compel us to discard the old framework and adopt a new one. It is to this process of intellectual molting that we owe the birth of many of the major ideas in mathematics and science. The paradox of incommensurables (exemplified by the diagonal of a square, which cannot be related to the sides of the square in terms of rational numbers) led to the concept of the continuum. Zeno's paradox of Achilles and the tortoise gave birth to the idea of convergent infinite series. Antimonies (internal contradictions in mathematical logic) eventually blossomed into Gödel's theorem.

Rapoport continues his list by citing paradoxes in science that helped lead to the theory of relativity, quantum mechanics and the link between information and entropy.

Falleta (1983, pp. xvii–xviii) explains the meaning of the term:

The word itself comes from the Greek (*para* and *doxos*), meaning 'beyond belief'. As used today, the term "paradox" covers a range of meanings, with its most general reference being to any statement or belief that is contrary to expectation or received opinion. The definitions of paradox … [involve] basically three meanings: (1) a statement that appears contradictory but which is, in fact, true; (2) a statement that appears true but which, in fact, involves a contradiction; and (3) a valid or good argument that leads to contradictory conclusions.

While a logician might only recognize the third type, this study uses a broader, more everyday definition, corresponding mostly to the spirit of Falleta's first definition.

1.2 Applications Outside the Introductory College Statistics Course

While the focus of this study has been declared to be the introductory college statistics course, Chu and Chu (1992, p. 191) discuss the use of counterintuitive examples before college: "Probability has been suggested for inclusion in the high school or even junior high school curriculum. The suggestion appeals to many because probability is viewed as a natural and intuitive subject manageable with very simple mathematics. It is also a good foundation for understanding statistics, which is in prevalent use in today's

society. … Unfortunately, the apparent simplicity of probability is quite deceiving."

One should, however, not expect the use of counterintuitive examples to be as effective below the secondary-school level. For example, Piaget (1975, p. 214) states (and somewhat overstates): "During the first of these three periods (before seven or eight years), the child does not distinguish the possible from the necessary. … Thus we could not consider his anticipations as judgments of probability, deriving from a greater or lesser degree of subjective certitude, because this certitude is only the product of a failure to differentiate between practical notions of intuitive probability and caprice." Piaget also notes (p. 193) that concepts such as permutations are operations on operations and are thus acquired only at the level of formal thought, usually no sooner than age 14.

Applications can also be made to courses taken after the introductory college course. Counterintuitive examples that can be encountered in an introductory course are often special cases of more general counterintuitive phenomena that can be further encountered and analyzed in later statistics courses. For example, Samuels (1993, p. 87) states that "Simpson's Paradox is actually no more paradoxical than the reversal or distortion of association in other settings, no more, for instance, than the familiar fact that a partial regression coefficient can have a different sign from a simple regression coefficient." Romano and Siegel (1986) discuss a large catalog of examples (e.g., Stein's Paradox) for students in advanced or mathematical statistics courses.

In addition to extension to later statistics courses, there are often connections that can be made between counterintuitive situations in statistics and objects from other branches of mathematics. For example, Lord (1990) shows that Simpson's Paradox can be represented using arguments of complex numbers, linear transformations of the plane, and determinants of matrices. The NCTM (1989, p. 146) strongly supports utilizing such opportunities for multiple representations.

It should be noted that some of the authors who are cited in this study (e.g., Gordon, Fischbein) address all branches of mathematics, not just statistics.

Indeed, most teachers of the introductory course do not have an advanced degree in statistics and typically teach other courses as well. The paradoxes listed in sections 1.1 and 3.1 are just some of the many in other branches of mathematics. Others are presented in forums such as the "Fallacies, Flaws, and Flimflam" column of the *College Mathematics Journal* or in sources such as Gordon (1991), Eves (1990), or Falletta (1990). In summary, the results of this study are by no means intended to be applicable only to statistics courses, although those courses were the study's focus.

1.3 Looking Ahead

Chapter 2 discusses many classifications of intuition and identifies which ones are most relevant to the present study. Chapter 3 presents an operational definition of counterintuitive and a set of four criteria for counterintuitive examples, and thorough discussion of representative examples of these. Chapter 4 analyzes two opposing points of view (mentioned in section 1.1) concerning the use of counterintuitive examples, and offers a syllabus-driven model of synthesis that clearly defines the roles of intuitive examples and counterintuitive examples as well as addresses specific concerns. Chapter 5 explores the model's connections to cooperative learning, structured controversies and constructivism. Chapter 6 summarizes the entire study and then points out problematic issues and directions for future research.

CHAPTER 2

CLASSIFICATIONS OF INTUITION

2.1 The Relationship Between Statistics and Cognition

There are many special connections between learning theory, intuition and statistics. Moses (1986, p. 6) notes that the development of psychology has been "interwoven with the development of statistical theory. Karl Pearson and R.A. Fisher, but also C.E. Spearman, and much later Harold Hotelling and S.S. Wilks, found the source of much of their work in psychological inquiries." Westcott (1968, p. 191, emphasis in original) proposes that intuition is most valuable in situations "in which information, explicitness, and redundancy are *just not available*," situations which often characterize real-world statistics situations. After all, statistics is often thought of as decision-making under uncertainty, or, to quote the title of a popular book (Tanur et al., 1989), "a guide to the unknown." The philosopher A. Ewing (1941, p. 102) concludes that "inference and intuition are linked together. Inference always presupposes intuition to provide the links in inference, but on the other hand inference is needed to support, prepare for, and develop intuition."

Gigerenzer and Murray (1987) explore these and other links as they examine theory construction in psychology through "the metaphor of the mind as an intuitive statistician (p. ix)." Their examples include Neyman-Pearson statistical hypothesis testing (p. 42) in a theory of signal detection and discrimination, random walks (p. 120) in R. Ratcliff's model of memory retrieval and storage, two-way ANOVA (p. 177) in H. H. Kelley's model of causal reasoning and Bayes' theorem (p. 147) in a model of rationality.

Also, Scholz (1991, p. 237) gives a comprehensive table of eight research paradigms on probability learning that includes normative models such as Bernoulli series, contingency measures, and the likelihood principle. Goertzel (1993) cites Monte Carlo and simulated annealing methods in his exploration of

thought as optimization. Finally, Girosi (1994) adds "… the problem of learning to perform some task from a set of examples. In mathematical terms this is equivalent to reconstructing a function from a set of sparse data points (the examples). Therefore, approximation theory and statistics are the appropriate mathematical framework for neural networks."

2.2 Operational Definition of Intuition

Even statistical concepts with the objective reputation of hypothesis testing can involve intuition. As Egon Pearson (1962, pp. 395–396) states: "Of necessity, as it seemed to us, we [Neyman and Pearson] left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters … as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities."

The term *intuition* has gone through many forms in the fields of philosophy and psychology. In philosophy, Westcott (1968, p. 22) explains how the scope of the definition has become progressively reduced, from Classical (which considers intuition as "an experience of ultimate truth, precluded by reason, and is antithetical to reason") to Contemporary ("the immediate apprehension of limited basic truths [e.g., deductive logic, mathematical axioms, causality, etc.] which are applicable to the problems of the intellect"), to Inferential ("rejects both the notion of immediate evidence and the notion of truth. … Truth is to be understood as either a set of conventions or a set of probability statements, both subject to change"). Westcott (1968, pp. 48–53) also offers a history of intuition among mathematicians, such as Pólya and Poincaré, and offers a distinction between mathematical intuitionism and philosophical intuitionism. While of intrinsic interest to scholars such as historians and philosophers, this is of limited direct applicability to the main focus of this study.

In attempting to define this elusive term for cognitive psychologists, Fischbein (1987, p. x) states a working definition which is perhaps closest to the aforementioned Contemporary school: "An intuition is, then, such a

crystallized—very often prematurely closed—conception in which incompleteness or vagueness of information is masked by special mechanisms for producing the feelings of immediacy, coherence and confidence. Such mechanisms have been described in the research literature, but very often without any apparent connection with a theory of intuition. … Studies in overconfidence, in subjective probabilities, findings referring to mental models, to typical errors in naïve physics, to misconceptions in mathematics, to the evolution of logical concepts in children, etc., represent, in fact, rich potential sources for a theory of intuition."

Indeed, Fischbein goes on (pp. 5–6, emphasis in original) to discuss how this feature of immediacy is perhaps the only common property of the many terms (e.g., insight, revelation, inspiration, common sense, naïve reasoning, empirical interpretation, self-evidence) and related areas of cognitive investigation: "*Problem solving* (illumination, heuristics, anticipatory schemas, etc.); *Images and models* (intuitive representations, intuitive models, intuitive didactical means, thinking in images, etc.); *belief* and *levels of confidence*; *developmental stages of intelligence* (Piaget has described intuitive thinking as a preoperational stage)."

Fischbein's working definition of intuition is consistent with usage by decision researchers such as Hogarth (1987, p. 1): "… for the most part judgments are made intuitively—that is, without apparent reasoning and almost instinctively." We will see later in this section that this sense of intuition will correspond to what Fischbein calls a primary intuition. Fischbein (p. 57) reviews previous descriptions and classifications of intuition by Henri Poincaré (1920), M. R. Westcott (1968) and Beth and Piaget (1966).

According to Fischbein, Piaget makes a distinction between empirical and operational intuitions, and the latter category can be further dichotomized either into intuitions expressed by images versus intuitions referring to logico-mathematical concepts, or into geometrical intuitions versus operations with discrete objects. Fischbein (pp. 58, 66) faults Piaget for too much generality in his use of the term *intuition*: "In Piaget's terminology an intellectual activity is

either intuitive or formal. Consequently, almost every intellectual activity the child is

able to perform before the formal operational period may be considered as being achieved on an intuitive basis … Piaget does not explicitly distinguish, at the concrete-operational level, cognitions which are intuitive and cognitions which are operational without being intuitive."

Fischbein then describes his own classification of intuitions based on what he calls roles (actually, the stage in the problem solving process) and also a classification based on origins (i.e., whether it originated before or after formal instruction). This latter classification, which concerns mainly intuitions at the initial stage of the problem solving process, will be the focus of this study for many reasons. Perhaps the strongest reason is compatibility with virtually any study involving interventions to influence conceptions. For example, this distinction between primary and secondary intuitions seems shared in spirit, if not in terminology, by Lembke and Reys (1994), who explore the development and interaction of "intuitive and school-taught ideas."

Fischbein (1987, pp. 64–68, emphasis in original) states:

Primary intuitions refer to those cognitive beliefs which develop in individuals independently of any systematic instruction as an effect of their personal experience. … Primary intuitions may be either pre-operational or operational. … The category of *secondary intuitions* implies the assumption that new intutions, *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. According to our primary intuitions, we tend to consider that in order to keep the velocity of a moving body constant, a force is necessary. … If for a mathematician the equivalence between an infinite set and a proper subset of it becomes a belief—a self-explanatory conception—then a new, secondary intuition has appeared.

As Shaughnessy (1992, p. 480) adds:

Primary intuitions are the ideas and beliefs that we have before instructional intervention; secondary intuitions are restructured cognitive beliefs that we accept and use as a result of instruction or experience within a particular cultural community. … For Fischbein, the process of replacing a primary intuition by a secondary one is not a gradual process [as Piaget might argue]; it takes place as a

whole, all at once. This is very much like the "Aha" experience in gestalt psychology—the moment of discovery or insight in the problem-solving process.

Feffer (1988, p. 40) details how this Gestalt assumption of organization, like the constructivist perspective which will be discussed in section 5.3, has "been advanced in opposition to a major aspect of the Cartesian world view, namely, the assumption that the essential properties of being are those of a clockwork or machine. More particularly, the Cartesian view would have it that nature is comprised of separate self-contained units of 'such and such' properties that can be combined in terms of laws governing the functioning of machines." Later Feffer suggests (p. 61) that "our constructionist and Gestalt assumptions have led to a view of consolidative integration in which the individual is able to anticipate the consequences of his activity in terms of a higher, more inclusive level of organization, namely, in terms of the scheme as a transformation law."

2.3 Pedagogical Applications of Intuition

There are textbooks with titles such as *Statistics: An Intuitive Approach* (Weinberg et al. 1981) and *The Probability Tutoring Book: An Intuitive Course for Engineers and Scientists (and Everyone Else!)* (Ash 1993), and the pedagogical literature is rich with ways in which intuition can be utilized in statistics courses. Most of these fall into the following non-exhaustive set of three categories: conceptual, geometric and numerical.

2.3.1 Conceptual Intuition

In addition to the occasional introductory books (e.g., Haack 1979) which focus on conceptual intuition almost to the exclusion of formulas or symbolic language, there are a number of individual articles which suggest metaphors to enhance communication of statistics concepts. Evans (1986) illustrates the concepts of null hypothesis, Type I error, Type II error, and power with the fairly common "courtroom" metaphor. He then extends it to include a "detective searching for clues" in that chances of discovering significant evidence against null hypothesis increase if the search can be narrowed by a more specific hunch

(one-tailed test), but decrease if the search must extend to both possible locations (two-tailed test). Evans also uses the metaphor of the relative positions of neighboring merry-go-round horses to illustrate patterns of positive, negative and no correlation between two variables. Weaver (1992, p. 178) uses falling leaves to illustrate confidence intervals: "As the trees shed their leaves, piles form around the trunks. … Imagine standing next to a tree's trunk [estimated population mean] and picking up a leaf [sample mean] from the [normal-shaped] pile … . How sure are you that this leaf came from the same tree and not a neighboring one?" Additional examples of such bridging analogies and anchors are included in section 4.7. These examples are particularly helpful to students who are less threatened by conceptual language than by symbolic formulas.

2.3.2 Geometric Intuition

The use of geometric intuition in teaching statistics has seen some increasing popularity as a bridge between the "cookbook" and overly mathematized approaches mentioned in section 1.1, especially in medium-level statistics courses. As Saville and Wood (1986, p. 205) state: "The bulk of commonly used contemporary statistical methods is based on a relatively simple application of the mathematics of Euclidean N-dimensional space." The authors demonstrate how to introduce students to "the theory and methods of analysis of variance and regression in a rigorous but elementary geometric setting, at the same time highlighting the unity of the area," needing only a minimal set of vector geometric tools. Thomas and O'Quigley (1993) use geometry to illustrate correlation and partial correlation, while Schey (1993) uses it to illustrate the relative magnitudes of different regression sums of squares. For example, the correlation coefficient r of the n pairs $(x_1, y_1), …, (x_n, y_n)$ is the cosine of $\angle XOY$, where $X = (X_1, … , X_n)$, $Y = (Y_1, … , Y_n)$, $X_i = x_i - x'$, and $Y_i = y_i - y'$. Finally, section 3.3 discusses geometric representations of Simpson's Paradox. Students in an introductory class can be expected to respond best to those examples in three or fewer dimensions. Johnson and Herr (1993) geometrically illuminate two

initially counterintuitive situations in multiple regression, namely a large R^2 with small regression parameter t-statistics, and vice versa.

2.3.3 Numerical Intuition

One reason for the importance of numerical intuition can be illustrated by a common instructor's lament (Smith 1987, p. 161):

At the computational level, what many of my students lacked was a good intuition about what was a reasonable answer. We meet extreme examples of this lack of intuition all too frequently: negative sums-of-squares in analysis of variance and correlation coefficients greater than unity are examples of impossible results that commonly appear in undergraduates' or even postgraduates' work. … [T]he very people who are most likely to make mistakes in statistical calculations have the most lax criteria for accepting a solution as plausible.

This intuition about what is a reasonable answer seems related to what Greeno (1991) calls "number sense," which includes numerical estimation and quantitative judgment.   Greeno's presentation of number sense as not mere skills, but rather a general condition of knowing in the domain of numbers and quantities, seems quite extendable to basic reasoning in probability and statistics. Just as Perkins and Simmons (1988, p. 307) maintain that "a student with a strong sense that numbers provide the semantic foundation for algebra is considerably more likely to see checking with numbers as a reasonable and rewarding course of action," it seems that students who know their way around the conceptual domain of statistics would be more likely to check their answers against simulations, diagrams, and all the other resources available to them in the domain.

Being able to sense when probabilities are significant is an important example of numerical intuition. Nisbett, Krantz, Jepson and Kunda (1983, p. 342) illustrate this with this thought experiment:

If someone says, "I can't understand it; I have nine grandchildren and all of them are boys," the statement sounds quite sensible. The hearer is likely to agree that a causal explanation seems to be called for. On the other hand, imagine that the speaker says, "I can't understand it; I have three grandchildren and all of them are boys." Such a statement sounds peculiar, to say the least, because it seems transparent that such a result could be due just to chance—that is, there is nothing to understand. Such an intuition is properly regarded as statistical in our view.

Garfield and Ahlgren (1988, p. 52) give another example: "[A] preference for Brand X over Brand Y in four out of five people would typically be believed to be clearly indicative of a general preference—although the probability of getting such an extreme sample [either brand picked by four or five people] of 5 just by chance is 3/8." With more numerical intuition, students would also have a better sense of, for example, when a difference between two groups is statistically significant, practically significant, both, or neither.

Having a ready selection of what Greeno calls "landmarks" with common numbers can be useful in locating oneself in the conceptual domain. For example, it is easily verified that a right-tailed test of H_0: $p = .5$ for sample size $n = 10$ yields p-values of .05, .01, and .001, for 8, 9, or 10 successes, respectively. Also, the number of 4-combinations chosen without replacement from 14 different items is 1001, a nearly-round number that is used in assigning lottery probabilities for the annual draft of the National Basketball Association. Section 3.4 explains why it takes $(\ln 2)N \approx 0.7N$ trials to have at least a 50-50 chance of at least one occurrence of an event with probability 1/N. Also, it takes $(\ln 20)N \approx 3N$ trials for at least a 95% chance of at least one occurrence of an event with probability 1/N. With the normal curve, students should know not only the so-called empirical rule that about 2/3 of the data is within one standard deviation of the mean, but also that these limits occur where direction of curvature changes. Furthermore, Morris (1988) offers a "1/3 rule" that states that the two points on the normal curve at 1/3 the height of the maximum height bound a range corresponding to nearly 3 standard deviations.

It is also useful to have quick and crude, easily applied methods for checking answers to calculations. Moses (1986, p. 137) gives an quick way to estimate a standard error for a small ($n \leq 15$) sample by dividing the range by the sample size n. Schuster (1993) demonstrates that $p \pm 1/\sqrt{n}$ is always at least a 91.0% confidence interval for the proportion p of a finite population of size N having a given attribute, based on a random sample (with or without replacement) of size n from the population, for all n, N, and p.

Another use of numerical intuition is to illuminate formulas, which in turn illuminate connections with parameters and concepts. D. E. Johnson (1989) uses an excellent concrete guided progression of very simple data sets to illustrate ANOVA and his method of illustrating the concepts and relationships between between-groups and within-groups variance has been adapted to other hypothesis testing situations as well. Johnson's classroom results especially suggest his technique may be useful for students who operate at a preformal level of thought, although there are some weaknesses in his study, such as a small, non-random sample with no control group. Several articles (e.g., Read and Riley, 1983) give instructors methods for constructing statistics problems with simple numbers. Another example of numerical intuition is given by Weinberg (1981, p. 280) concerning the formula governing the F distribution, a ratio of independent estimates of the same positive quantity. Moses (1986) presents a formula that illustrates the regression-to-the-mean phenomenon discussed in section 4.7.

It is worth further breaking down the category of numerical intuition into explicit numerical intuition (as we have seen examples of) and implicit numerical intuition. The latter type might be described with a phrase from Piaget (1975, p. 173): "… the equivalent of the discovery of the formative operations themselves, as distinct from the formulation." In discussing children's learning about permutations (without replacement), he says (pp. 173–174): "… they will discover the law $n! = n(n–1)(n–2) … 3 \cdot 2 \cdot 1$; even if they do not arrive at the explicit expression in symbols, they will at least succeed in seeing its operative mechanism, which is all that matters to us from the analytical point of view of probabilistic intuitions."

2.4 Problematic Issues

2.4.1 Terminology

In addition to these aforementioned categories of intuition, there is a plethora of related categories and terms in the literature of learning theory and cognitive science, as applied to mathematics and science education. According to

Roth (1990, p. 149): "Labels for this incompatible prior knowledge have included 'misconceptions', 'preconceptions' 'alternative frameworks', 'alternative conceptual systems', 'alternative conceptions', 'children's science', 'theories-in-action', 'intuitive theories', 'qualitatively different conceptions.'" Roth also describes (pp. 147–148) the variety of terminology applied to the correcting of these: "… conceptual change learning has been called accommodation by Posner et al., following Piaget, and reconciliation or exchange by Hewson and Hewson. Champagne et al. described it as the restructuring of conceptual systems."

In chapters 3 and 4, it will be evident that there are also a number of terms used to refer to specific counterintuitive examples. For example, the Classification Paradox is also known as the false positives paradox (Gonick and Smith 1993, p. 49), the prosecutor's fallacy, and the Taxi Problem. On the other hand, some names that sound similar (e.g., Gambler's Fallacy and Gambler's Ruin) actually refer to different results.

Another point to be made is that "[p]aradoxes generally possess a good measure of ambiguity, and their solutions frequently involve sorting out various meanings or interpretations embedded in the ordinary language or images that form them" (Falleta 1983, xix). Because of this, teachers must take care in defining terms, conditioning events, sampling units, sample space, etc., and certainly never ridicule students whose answers are quite reasonable based on how the student completed the assumptions that were not fully provided (see discussion of the Monty Hall problem in section 3.2, for example).

2.4.2 Measurement

In addition to the profuse number of definitions, there are also challenges in measuring intuitive thinking. By defining intuition as (Westcott 1968, p. 100) "the event which occurs when an individual reaches a conclusion on the basis of less explicit information than is ordinarily required to reach that conclusion," Westcott (p. 101) states the problem as trying "to provide a situation in which individuals may attempt to reach conclusions or solve problems in the presence of varying amounts of information. Furthermore, there must be a way of appraising how much information a given individual requires, and how much is ordinarily

required. Finally, there must be some conclusion or solution which is consensually valid." This may be further complicated by the existence of alternative pathways to a solution.

While Ohlsson, Ernst, and Rees (1992) have claimed success in quantifying the somewhat related concept of "difficulty," their focus on subtraction methods does not seem generalizable to the much richer complexity of the domain of statistics. In claiming that some primary intuitions are more "deep-seated" than others, however, Clement (1993, p. 1242) offers a number of sources for measuring deep-seatedness, all of which seem applicable to measuring intuitiveness. These sources include pre-postcourse tests, student-reported measures of confidence in their answers, "spontaneous expressions of conviction in interviews, resistance observed during tutoring, and historical parallels to students' alternative conceptions." Also, Clement, Brown and Zietsman (1989) operationally distinguish between an individual anchor (a pretest problem for which an individual student both gave the correct answer and expressed at least a minimum score on a confidence scale) and a group anchor (an example found to be an individual anchor for, say, 70% of the students).

2.5 Looking Ahead

Now that a broad backdrop on intuition has been presented, chapter 3 will distinguish between non-intuitive and counterintuitive and give some concrete representations of counterintuitive examples from the area of statistics. Chapter 3 describes only some of the many examples from the model in chapter 4, so that more in-depth understanding of their power and pitfalls can be obtained.

CHAPTER 3

COUNTERINTUITIVE EXAMPLES IN STATISTICS

3.1 The Meaning of *Counterintuitive*

Complementing Rapoport's examples of paradox listed in section 1.1, Fischbein relates (1987, p. 10): "The Copernican revolution, the non-Euclidean geometries, the special and the general theories of relativity, the findings related to the Cantorian concept of actual infinity, etc.—all these ideas and representations have contributed to the notion that self-evidence (i.e., intuitive evidence) is not synonymous with certainty. More and more non-intuitive or counterintuitive concepts have invaded science and mathematics."

While chapter 2 sets forth the operational definition of intuition (namely Fischbein's origin-based classification), the terms *non-intuitive* and *counterintuitive* (often used interchangeably by some authors) also need to be operationalized. For this study, the term *non-intuitive* refers to a topic or situation in which the student's foundation is so minimal that there is no intuition of any kind for what type of results to expect or how to interpret them. (Indeed, this is consistent with the use of the term in the quotation by Fischbein in section 1.1.) A possible example of such a concept might be a statistic based on higher moments, such as skewness or kurtosis, especially for a multimodal distribution. In practice, before classifying the intuitiveness of a situation, one needs to ensure that causal and chance factors are clearly identified, because as Hogarth (1987, p. 21) states, "even experts can make responses similar to novices when problems are complex." In general, it seems that the term non-intuitive is associated with broad topics or concepts, while counterintuitiveness is associated with surprising results of particular tangible situations.

The concern of this study, however, is with counterintuitive examples. The term *counterintuitive* will assume both that a student does indeed have an initial expectation or primary intuition (a directional hypothesis, so to speak) and that that primary intuition with respect to a result contradicts and is, at least initially, very resistant to the normative view. While not all students, even in a specific target audience, can be expected to find the same things counterintuitive, the fact that experienced teachers of statistics notice patterns in students' difficulties yields reasonable justification for the idea that there are patterns in what students find counterintuitive. This parallels the group anchor of

section 2.4.2. Fischbein's earlier research uncovered some distinctions about what is often found to be intuitive. Fischbein (1987, pp. 67–68) states:

subjects aged 12 more (formal operational period) possess a correct, natural, intuitive understanding of the following probabilistic concepts: the concept of chance and of the quantification of chances as the relationship between the number of favorable and of all possible equally likely outcomes; the fact that increasing the number of conditions imposed on an expected event diminishes its chances (which corresponds to the multiplication of probabilities). By contrast, there is no natural understanding of the compound character of some categories of events nor of the necessity to inventory the different situations which can produce the same event (for instance, when throwing a pair of dice, there is no intuitive understanding of the fact that there is a difference between the probabilities of gettting the pair 5-5 and the pair 5-6) (Fischbein, 1975, pp. 138–155).

Sometimes, distinctions of intuitiveness can be made even within the same statistical topic. For example, Garfield and Ahlgren (1988, p. 52) illustrate differences between the combinational and sampling forms of the representativeness misconception. Also, the experiments of Well et al. (1990) showed that students tended to do well on the so-called "accuracy version" but not on the "tail version" of a question concerning the law of large numbers. As Shaughnessy (1992, p. 478) states, "Thus, the emerging picture of students' intuitive understanding of the law of large numbers is not a simple one; task variables affect student performance a great deal."

What is "counterintuitive" has had very little attention from researchers, but will be operationally defined as a result that seems "surprising" to a high percentage of people in a particular population at a particular time.

As stated in section 1.1, the focus of this study is students in college introductory statistics courses, most of whom will be naïve-statistical. Shaughnessy (1992, p. 485) lists as indicators for this level of stochastics conception: "use of judgmental heuristics, such as representativeness, availability, anchoring, balancing; mostly experientially based and nonnormative responses; some understanding of chance and random events."

It is a suggestion for future research that a survey be undertaken to see if people in this category (or other categories, for that matter) consistently find the same examples or results counterintuitive, and whether or not they would rank

them in the same pattern or order of "counterintuitiveness." This empirical ordering can be compared with a theoretical ranking based on a priori features. For example, the averaging-the-averages misconception [falsely assuming that the average of a set of averages equals the overall average of the individual original numbers] is less counterintuitive than Simpson's Paradox in that a student may come to accept that the average of averages may not be the same as the true overall average and yet still be quite startled to find that the direction of a comparison of overall weighted averages can actually be the reverse of the individual weighted averages.

To illustrate the example in the preceding paragraph, consider the data from section 3.3. The overall male average is .55, while the average of the males-by-department numbers 5/20 and 50/80 is .4375. The overall female average is .45, while the average of the females-by-department numbers 30/80 and 15/20 is .5625. Now if the female ratios exceed the male ratios within each department, then the (unweighted) average of female ratios will always exceed the (unweighted) average of male ratios, thus preserving the direction of the comparison. Thus, if a student has the averaging-the-averages misconception, he or she will always be susceptible to Simpson's Paradox.

A second example is that a student who masters the Inspection Paradox will surely also master the average class size paradox (discussed in chapter 4), as the latter is a special case of the former.

As Fischbein (1987, p. 70) suggests, what is considered counterintuitive may be relative to a particular era. "It is, for instance, easier today to get used to the Newtonian understanding of inertia—which was originally counterintuitive— than to the relativistic interpretation of space and time." Also (p. 63): "… the intuitive acceptance of the fifth Euclidean postulate was so strong that it inspired two thousand years of research in a wrong direction! It took 2000 years of unsuccessful efforts until mathematicians dared to consider some intuitively incredible alternatives!"

Hans Hahn (1956, p. 1976) adds:

If the use of multi-dimensional and non-Euclidean geometries for the ordering of our experience continues to prove itself so that we become more and more

accustomed to dealing with these logical constructs; if they penetrate into the curriculum of the schools; if we, so to speak, learn them at our mother's knee, as we now learn three-dimensional Euclidean geometry, then nobody will think of saying that these geometries are contrary to intuition. They will be considered as deserving of intuitive status as three-dimensional Euclidean geometry is today.

Hahn's statement blurs the distinction between constructs that are initially intuitive and initially counterintuitive. Rice University mathematics professor Reese Harvey liked to tell his students that some results in mathematics start off seeming difficult, and after successful reflection or instruction, some of these results can be "refiled" as "easy," but others will still seem difficult. In any case, the survey suggested earlier in this section should clarify much of this.

3.2   Criteria for Counterintuitive Examples

There are a large number of situations in statistics which can result in counterintuitive results and sometimes widespread interest. For example, the conditional probability problem now referred to as the "Monty Hall problem" or the "car and goats problem" generated an outpouring of popular interest (including front page newspaper stories) when it was published by vos Savant (1990). It appeared in several mathematics magazines, and Barbeau (1993) recently offered a 63-item list of references! It has also been used to demonstrate the fallibility of intuition (e.g., Kohn 1992). Shaughnessy (1992, p. 475) describes the problem as follows:

During a certain game show, contestants are shown three closed doors. One of the doors has a big prize [e.g., a car] behind it, and the other two have gag gifts [e.g., goats] behind them. The contestants are asked to pick a door. Then the game show host, Monty [who always knows where the big prize is], opens one of the remaining closed doors and shows it to the contestant, always revealing a gag gift. The contestants are then given the option to stick with their original choice or to switch to the other unopened door. What should they do?

The typical interpretation of the problem assumes that before opening your door, Monty always first opens a different door (chosen at random from the 1 or 2 remaining doors that hide a gag gift), and then gives you the chance to

switch. In this scenario, there was a 1/3 chance you were right in the first place (in which case you will win if and only if you stick to that choice) and a 2/3 chance you were wrong in the first place (in which case you will win if and only if you switch). Therefore, switching is the best strategy, winning 2/3 of the time. By modifying assumptions about Monty's options and motives, Foster and George (1994) show that alternative answers such as 0, 1/2, or 1 can be reasonable for the probability of winning with the switching strategy.

There are a number of other counterintuitive situations involving conditional probability (Shaughnessy 1992, pp. 473–474), as well as situations involving probabilities of disjunctive events, comparisons, randomness, and averages. This study focuses on a representative (intended to be substantial but not exhaustive) sample of situations chosen to meet four criteria: (1) the situations actually occur in real-life contexts; (2) they can all be (but often are not) discussed in an introductory statistics course; (3) they (initially at least) seem counterintuitive to a large majority of students before instruction (i.e., a large majority would give an nonnormative, incorrect answer; at this point, this is supported largely by didactical writings and anecdotal observations, although it would be straightforward to test this observation empirically on a larger scale; see section 2.4.2); and (4) the situation be readily explained, demonstrated or experienced through tangible heuristic explanation and/or experimentation, rather than depending only on technical theoretical proof.

This first criterion is encouraged by the NCTM (1989, pp. 87, 105): "… learning should be grounded in experience related to aspects of everyday life … The data to be gathered, organized and studied should be interesting and relevant. …" Also, findings by Mevarech (1983, p. 425) suggest that one reason students may "misconceive a set of given means under simple mean computation as a mathematical group satisfying the four properties of closure, associativity, identity and inverse" is that they treat means as "decimals that mean nothing to them," as opposed to Moore's (1993, p. 15) view of data as "numbers with a context." Beins (1985, p. 168) gives an additional motivation: "One of the most obvious ways to overcome the anxiety associated with statistics is to focus

students' attention on various kinds of information they already have," such as "the myriad of claims proffered on television and in magazines." Greeno (1991, p. 177) insists that abstractions and symbols "should not replace experience in conceptual environments as the main learning activity that we provide for students."

Some real-life situations may have important political or historical ramifications, even if they are not likely to be personally encountered by the students in their own lives. Konold (1991, p. 6) relates:

Students often balk when given the standard introductory problems—"What has this got to do with anything?" This is not to say that getting students to seriously consider the standard problems is unimportant. But if we want to demonstrate the broad range of probability applications, then the situations we ask students to consider must become more complex than flipping coins, rolling dice, and blindly selecting socks from drawers.

Konold goes on (also see Konold 1994) to give a very rich real-world situation for the geometric distribution, namely a proposed policy in China to limit families to one son (rather than one child). Instructors (and especially researchers), however, should be wary of real-life situations that "are laced with contextual traps" (Shaughnessy 1992, p. 473). An example he gives of such a trap is (emphasis in original): "The language 'had a heart attack and is over 55' may be interpreted by some people as 'had a heart attack *given that* they are over 55.' "

It also serves many pedagogical purposes if the situations are amenable to a diversity of representations, but that is not the main focus here. The Classification Paradox, for example, can be set up using Bayes' theorem, a contingency table, a Venn diagram, a tree diagram or a reverse flow diagram. These representations are all fairly common, except for the reverse flow diagram, an example of which is given by Chu and Chu (1992).

There are books full of examples (e.g., Romano and Siegel, 1986) that go well beyond an introductory course, thus violating condition #2, but nevertheless suggesting ideas for future investigations. And, as mentioned in chapter 1, examples such as Simpson's Paradox can be examined in an introductory class and examined with more generality in later classes. An example of a situation

that would fail condition #1 is the St. Petersburg Paradox (e.g., Weaver 1963, Falletta 1990), which is essentially that bettors should (if they behave consistently with the expected value criterion) be willing to pay an arbitrarily large amount of money to play a game in which the bettor flips a coin until it finally lands on heads (say this occurs on the nth flip), at which point the bettor is given 2^{n-1} dollars. The expected gross payoff of this bet is $\sum (1/2)^n 2^{n-1} = \infty$. The reasons no real-life casino offers this bet are given by Weaver (1963, p. 165): "Although $1 million is, from the point of view of the formal theory, a very cheap entrance ticket, it is an impossible price for me, partly because I just haven't that kind of money, and partly because it doubtless would ruin me to lose that amount, even if I had it. Second, the so-called 'infinite value' of the St. Petersburg game depends essentially upon the house's being able to pay off, no matter what happens." The field of behavioral decision theory investigates situations in which what is counterintuitive has more to do with human behavior than with the underlying mathematics.

3.3 Simpson's Paradox

As Moore and McCabe (1993) explain, relationships among three categorical variables can be described by a three-way table of counts of percents, which is printed as separate two-way tables for each level of the third variable. A comparison between two variables that holds for each level of the third variables may be changed or even reversed when the data are aggregated (i.e., summed over all levels of the third variable). When this happens, Simpson's Paradox has occurred. According to Falletta (1990, p. 137), this situation is "named after the British statistician E. H. Simpson, who first wrote about it in 1951." There is much literature concerning examples of Simpson's Paradox involving real-life comparison of overall rates, ratios, percentages, proportions, probabilities, averages, or measurements that are weighted averages of subgroup counterparts.

Bickel et al. (1975) showed that when University of California at Berkeley graduate school admissions were analyzed by department, women were accepted at a higher rate then men, but were accepted at a lower rate overall (due to the lower admission rates of departments that had more female applicants).

Freedman et al. (1991, p. 16) discuss this same situation but do not use the term Simpson's Paradox. Cohen (1986, p. 34) lists many other examples that have occurred: "Rural fertility and urban fertility can both be rising while (as a result of population movements) aggregate fertility is falling. The morbidity of both young and old can be improving while (as a result of shifts in the age structure) aggregate morbidity worsens. … The federal income tax rate for taxable income tax returns in each of five categories of adjusted gross income declined from 1974 to 1978, but (because of category creep) the overall tax rate increased."   The introductory textbook by Moore and McCabe (1993, pp. 188–191) gives additional examples. Simpson's Paradox has also been mentioned in publications for more general audiences such as *Discover* (Paulos 1994).

As a final justification for the real-life importance of this particular counterintuitive example, consider that a major provision of the $28 billion anti-crime bill passed by the House of Representatives (Thomma 1994) is that "[d]efendants facing the death penalty would be allowed to use racial statistics on capital punishment as evidence of discrimination." Moore and McCabe (1993, p. 197) list an example where Simpson's Paradox has in fact occurred concerning that issue.

As stated in section 3.1, this situation can be thought of as a more pathological case of the averaging-the-averages misconception. And as Ulep (1990, pp. 59–60) notes, "Many education majors who already had a course in statistics thought that the average of numbers that include zero is the same as the average with zero excluded (Mevarech, 1983). A majority of them compute the ordinary average in problems requiring the weighted average. This finding agrees with that of Pollatsek, Lima and Well (1981)."

While Simpson's Paradox itself, although not some aspects of related generalizations (see Samuels 1993, p. 87), is well understood by statisticians, the difficulties it poses to students have not been seriously examined or explicitly connected to difficulties when the sample mean must be calculated as a weighted average. Falk and Bar-Hillel (1980) are among the very few researchers who seem to suggest such a connection.

31

The following is a brief numerically streamlined illustration, involving the three categorical variables of gender (male or female), department (social sciences or physical sciences), and employment status (hired or denied). From the 2×2×2 table below, it is routine to verify that within each department, women are hired at a higher rate than men (since

$$30/80 = .375 > .25 = 5/20 \ \text{ and } 15/20 = \ .75 > .625 = 50/80),$$

yet are hired at a lower rate than men for the overall aggregate situation:

$$\{ 80(.375) + 20(.75) \} / 100 \ = \ .45 \ < \ .55 \ = \{ 20(.25) + 80(.625) \} / 100$$

department:	S		P	
gender:	m	f	m	f
hired	5	30	50	15
denied	15	50	30	5
applied	20	80	80	20

As Paik explains (1985, p. 53): "The paradox is more clearly visualized by the circle graph [in which each circle represents a gender-department combination, the y-coordinate of the center of each circle is the subgroup-specific hiring rate, and the area of each circle is proportional to the sample size of its associated subgroup] when we use the … ordinary correlation coefficient r applied to two dichotomous variables." The circle graph in Figure 1 shows that the within-group correlations (represented by the top two circles and the bottom two circles) each have the same sign, a sign which is different from the overall correlation (represented by all four circles), "since the two larger circles on the negative diagonal dominate the positive ones. … By varying the positions and sizes of the circles in Figure 1, one can easily see that all of the 3^3 combinations for the three correlations are actual possibilities." Despite the clear insight provided by this representation (which relies only on informal uses of scatterplots, correlations and regression slopes, which are all standard topics in an introductory course) it has not been incorporated into introductory textbooks.

Figure 1

Simpson's Paradox: Circle Graph

(adapted from Paik 1985)
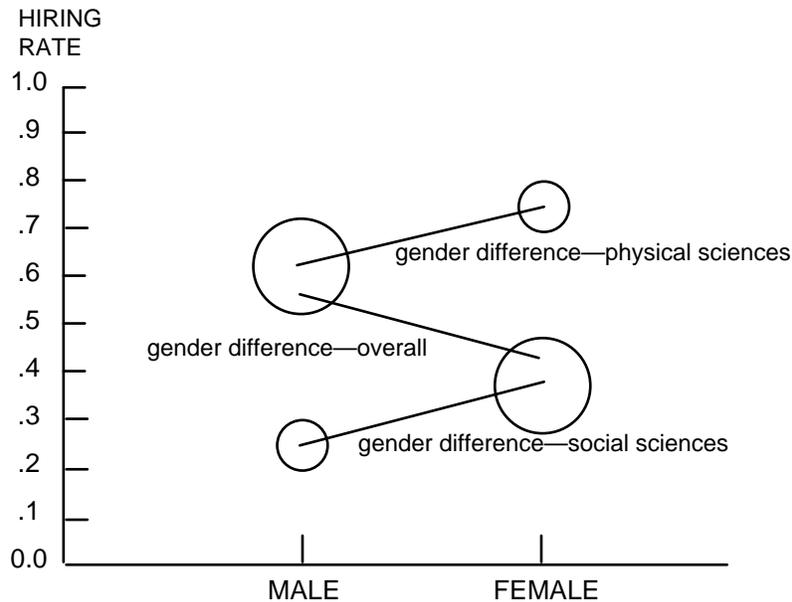
Tan (1986) provides yet another geometric representation of Simpson's Paradox which is built only on the observation that "[t]he length of any line segment which is parallel to the two bases and has its endpoints on the nonparallel sides of a trapezoid is the weighted mean of the lengths of the two bases." This relationship can be quickly derived algebraically by setting the usual formula for the area of the overall trapezoid equal to the sum of the areas of the two smaller trapezoids formed by the new segment. Applying this to our university employment example, each gender would have a trapezoid in which the two bases represent the two departments. The trapezoids have two bases and one leg in common, as shown in Figure 2:

Figure 2

Simpson's Paradox: Geometric Representation

(adapted from Tan 1986)
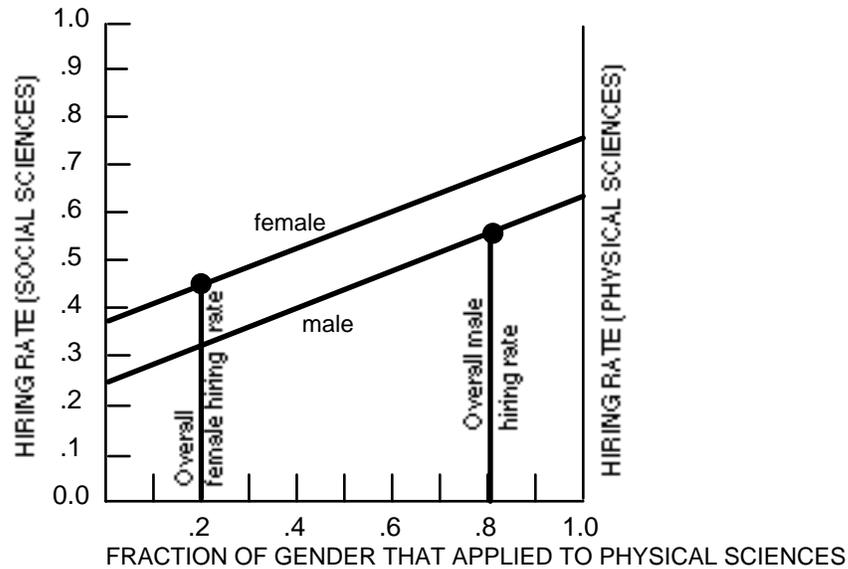
Finally, Falk and Bar-Hillel (1980, p. 107) suggest a concrete representation involving a platform scale:

Suppose a set of uniform blocks arranged in stacks of varying heights is located on a weightless platform, which is balanced on a pivot located at the center of gravity. … One can … shift the entire construction to the *right*, while simultaneously moving individual blocks to other stacks on their *left*. If done appropriately, the net result could then be a new center of gravity which is to the left of the old one.

Figure 3

Simpson's Paradox: Platform Scale Representation

(adapted from Falk and Bar-Hillel 1980)

PROPORTION HIRED
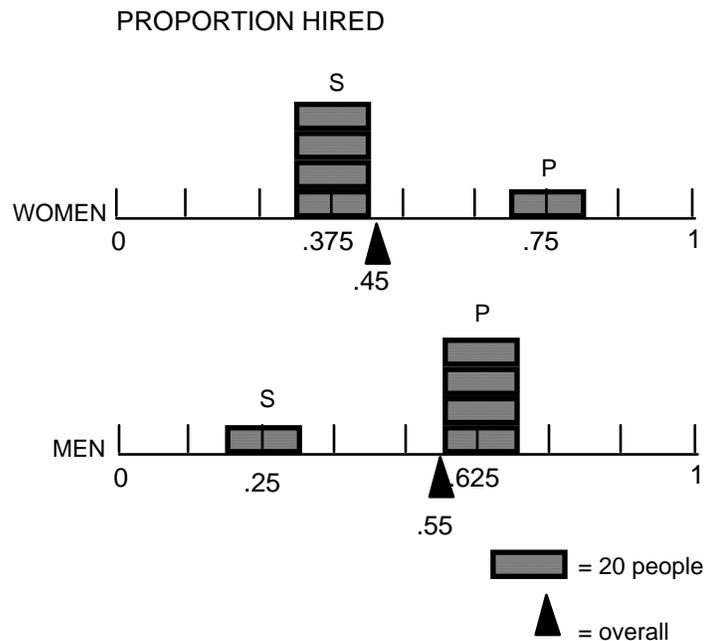
This third representation is limited to numerical examples in which the subgroup weighting numbers are multiples of each other (20 and 80, in this case) and the total number in each overall group is the same (there are 100 men and 100 women). The platform scale representation, however, is readily extended to more than two stacks (departments, in this case).   This example is convenient to construct because the four gender-department hiring proportions are all multiples of one-eighth and the two overall gender hiring proportions are nearly multiples of one-sixteenth. In fact, a single physical model could be built with a horizontal scale that goes in both directions, and be turned 180° to represent the other gender's situation. On the other hand, a side-by-side comparison of two platform scales keeps the subgroup and overall comparison in view.

The platform scale representation is clearly the most concrete of the three discussed in this section, and should therefore be the first one used (following Bruner's suggested progression of concrete before iconic and abstract) in a classroom setting. Furthermore, the fact that the unweighted mean is often described in terms of a platform scale model makes the representation very natural to build on to generalize to the weighted means that Simpson's Paradox involves. Students can certainly see with this representation that, for example, the weighted average and unweighted average of two stacks (i.e., averages) will be the same (i.e., have the same balance point) only if the sizes of the stacks are equal. Algebraically, $(nx' + my') / (n + m) = (x' + y') / 2$ implies $m = n$.

In any case, students can readily verify for themselves that the paradox exists, and they often respond (perhaps as much to this counterintuitive example as any other, before the introduction of a clarifying representation such as the "circle graph") with statements such as: "It's correct, but I still don't believe it." As Confrey (1990, p. 111) states, "Ironically, in most formal knowledge, students distinguish between believing and knowing. To them there is no contradiction in saying, 'I know that such and such is considered to be true, but I do not believe it.' To a constructivist, knowledge without belief is contradictory." Thus, the relationship between constructivism and the use of counterintuitive examples needs further examination in light of the current call in mathematics education reform for more constructivist styles of teaching. This will be examined further in section 5.3, and may also be related to proofs which are convincing yet produce no understanding (Barbin 1994).

The particular reaction to a conflicting comparison may have an affective component as well, such as anxiety or cognitive dissonance in the face of two competing claims, a situation that often occurs in the media with stories on what increases cholesterol or cancer risk. A teacher could have a structured contro-versy with Simpson's Paradox by giving each four-student group, the Bickel data set (as simplified in Mitchem 1989), for example, and assign students within the group roles such as "women's advocate," "university counsel," etc. After a while, students will believe the paradox "can happen" and can then be asked (in the unlikely event that they themselves don't ask), "**When** does it happen?"

At this point the instructor can suggest that the students explore this paradox not only algebraically (e.g., Mitchem 1989, Lord 1990), but also geometrically (e.g., Paik 1985, Tan 1986) and physically (e.g., Falk and Bar-Hillel, 1980). Lord (1990), for example, shows that Simpson's Paradox can be represented using arguments of complex numbers, linear transformations of the plane and determinants of matrices. Falk and Bar-Hillel (1980) not only make a connection between Simpson's Paradox and weighted averages, but also provide a concrete representation with blocks on a platform that can be adapted for classroom use (as previously discussed in this section).

Other questions that can be explored include whether the true mean of a set of data could be higher than the true mean of another set of data if all that is known is that the first set of data has a lower mean based on the "grouped data" formula that is a weighted average of class interval midpoints by class frequencies. Formulas for grouped data are not uncommon in introductory textbooks (e.g., Lapin 1987), yet this connection to weighted averages and Simpson's Paradox (all involve levels of aggregation) is virtually never made. Specifically, if f_i, L_i, U_i represent the frequency count, lower class limit, and upper class limit, respectively, for the ith class interval, then the grouped mean formula $\sum f_i (U_i + L_i)/2$ could produce a value as large as $\sum U_i f_i$ or as small as $\sum L_i f_i$.

(Connections to the midpoint Riemann sum can be made here for calculus students.) Also, introductory statistics courses for business students often include weighted aggregate price indices, which are typically ratios of weighted averages of prices.

It would also be useful to have a long-term follow-up to see whether or not students who grappled with Simpson's Paradox would in fact be more critical users of statistics, and less likely to accept uncritically a number on their calculator or a graph in the newspaper. Would it make a student more likely to wonder when presented with a university's 15:1 student-faculty ratio: "Is 15 the mean or the median? Is it the average per class (and thus the average from a professor's point of view) or per student?" As Hemenway (1982) shows, the expected class size for a student always turns out to be at least as big as the

expected class size for a professor! For a quick concrete example, consider a college with a 90-student class and a 10-student class. The average class size from a professor's view is $(90 + 10) / 2 = 50$, but from a student's view is $[90\,(90) + 10\,(10)] / 100 = 82$! Considering the immediate relevance of this to a college student's situation, it is surprising that it is rarely included in introductory courses or textbooks. It will, however, be included in the model in section 4.7. Finally, it would be useful to investigate whether the process of realizing how different weights make Simpson's Paradox possible in turn is effective in helping students avoid the common errors with weighted averages referred to earlier in this section by Ulep (1990).

For further work which can increase student interest and confidence, students can be led through the work of Moses (1986) or Friedlander and Wagon (1993). Moses (1986, p. 429) who, after the standard warning against combining 2×2 contingency tables themselves, then goes a step further by demonstrating what he calls "a rather clear, intuitive way" to combine the information (not the original data) in several 2×2 contingency tables. Friedlander and Wagon (1993, p. 268) pose the possibility (it is actually not possible) of a "double Simpson's Paradox":

It is possible for there to be two batters, Veteran and Youngster, and two pitchers, Righty and Lefty, such that Veteran's batting average against Righty is better than Youngster's average against Righty, and Veteran's batting average against Lefty is better than Youngster's average against Lefty, but yet Youngster's combined batting average against the two pitchers is better than Veteran's. … [I]s it possible to have the situation just described and, at the same time, have it be the case that Righty is a better pitcher than Lefty against either batter, but Lefty is a better pitcher than Righty against both batters combined?

Finally, an instructor can ask students to compare the dynamics of Simpson's Paradox with other reversal paradoxes such as the "test-drive paradox" of Falletta (1990, p. 142):

The man's test results clearly show that the attribute of being luxuriously comfortable is found more frequently among American-made cars than among European-made cars, as is the attribute of being economical. However, there is

only one American-made car tested that has both attributes, whereas there are two European-made cars with both attributes. Thus, the conjunction of the two desired characteristics is found more frequently among European-made cars despite the fact that taken separately each individual attribute occurs more frequently among the American-made cars.

Or the paradox of voting (Falleta 1990, pp. 182, 185) in which "the moderate candidate in pairwise races against either the liberal candidate or the conservative candidate wins the election. Yet, in the three-way race, it is the moderate candidate who is excluded from the runoff race. … In other words, although the ranking of the candidates is transitive for an individual, the society's ranking is non-transitive." Falleta explains that a voting system in which the majority's choice always wins has been proven to be impossible without violating certain basic conditions of democracy.

3.4 Probabilities Involving Disjunctive Events

Another major source of counterintuitive situations involves what Shaughnessy (1977, p. 306) refers to as "the deceptive nature of the probability of disjunctive events." Disjunctive refers to the "or" logical operator, or union, and is involved in the probability that at least one of a number of simple events occurs. In simplified language, when there are a large number n of opportunities for an unlikely event (with probability p) to happen, the overall probability that it happens at least once can become high surprisingly quickly. Students often overestimate the probability, by using the incorrect formula np—which can exceed the logical upper bound of 1—instead of the correct formula $1 - (1 - p)^n$. Also, students tend to overestimate the number of trials needed for there to be at least a 50-50 chance of the event happening at least once. An example of an important situation in real life is given in Moore and McCabe (1993, p. 297, example 4.16) involving AIDS testing. Bar-Hillel and Neter (1993) relate the history of research on a fallacy often committed with disjunctions, including its relationship to the more famous conjunction fallacy researched by Tversky and Kahneman (1983).

One famous example in the history of misconceptions involving probabilities of disjunctive events is the paradox of the Chevalier de Méré from

the mid-17th century. De Méré asked Blaise Pascal why it is advantageous to bet on the occurrence of at least one ace in four rolls of a die, but not to bet on the occurrence of at least one double-ace ("snake eyes") in $(4)(6) = 24$ rolls of two dice. The two respective probabilities are in fact $1 - (5/6)^4 \approx .518$ and $1 - (35/36)^4 \approx .491$.

Kunoff and Pines (1986, p. 211) paraphrase the paradox as follows:

In throwing two perfectly balanced dice, how many tosses are needed to have at least an even chance of getting a pair of sixes at least once? … From $1 - (35/36)^n \geq .5$, de Méré found $n \geq 24.6$ and concluded that $n = 25$ was needed. … He also concluded that 24 tosses was the correct answer, based on the widely accepted gambler's rule: If the chances are one in N of success in a single trial of an experiment, and n is the number of trials need to have at least an even chance of success, then n/N is a constant.

Freedman et al. (1991, p. 231) state: "Fermat saw that de Méré had used the addition rule for events that were not mutually exclusive. … pushing de Méré's argument a little further, it shows the chance of getting an ace in 6 rolls of a die to be 6/6 or 100%. Something had to be wrong." This example actually was very important, both at the time (the two probabilities involved are on opposite sides of the 50% mark, and so were of great interest to gamblers), and because it led to work by Pascal and Fermat that (Weaver 1963, p. 52) "can properly be regarded as the real start of the mathematical theory [of probability]."

As students can hypothesize with their calculators, when N is large enough, the ratio n/N is indeed almost constant. Originally shown by Abraham De Moivre, this result can be justified by setting $1 - (1 - 1/N)^n$ equal to .5, which yields $n \ln(1 - 1/N) = - \ln 2$. Using power series to approximate $\ln (1 - 1/N)$ by $- 1/N$, the result follows. In de Méré's case, $N = 6$ was not large enough for $n/N = 4/6$ to sufficiently approximate the limiting constant of $\ln(2) \approx .6931$. This value is a useful landmark, as discussed in the context of numerical intuition in section 2.3.3. In fact, had Frances Bobnar known of this particular benchmark, she might not have sued the Pennsylvania Lottery Commission (see Cohen 1994) after her family and friends bought some $1.5 million in tickets without winning a jackpot.

A more famous example involving disjunctive events (and perhaps the most famous problem mentioned in this study) is the "birthday problem": estimating the number of people (23 is the answer) needed so that there is at least a 50% chance that at least two people have the same birthday. Students often tend to use 50% as a representative multiplier (see Kahneman and Tversky 1972) in such a situation. For example, in a pretest administered by Shaughnessy (1977, p. 306), "62 out of the 80 subjects responded that it would take 183 or more people." In that same experiment, Shaughnessy found that after $6^{1}/_{2}$ weeks of a small-group activity-based course, the students' "tendency to use 50% as a representative multiplier of the total population had practically disappeared in the experimental group."

As Slonim (1960, pp. 10–11, emphasis in original) explains:

Experience, as well as mathematics, in this instance discloses the error in one's intuitive feeling that the occurrence of multiple birthdays in a group of 30 people is rare. … Picture the 30 people lined up in a row. Number One states his birthday. The remaining 29 then compare their birthdays with his. If there are no matches, Number Two then announces his birthday. The remaining 28 now have a second chance to compare their natal dates with that of Number Two. If, again, none match up, one or more of the remaining 27 may still duplicate Number Three's birthday. So on down to Number Twenty-nine, whose birthday may still be the same as Number Thirty's. Each of the 30, therefore, has 29 separate chances of matching his birthday with another's. … [T]he average person sees the problem as 'What are the odds that any one of the other 29 has the same birthday as *mine*?" whereas they more properly should ask, "What are the odds that any one of the 30 has the same birthday as *any other one of the 30? "*

A correction must be made to Slonim's statement that each of the 30 has 29 separate chances of matching his birthday with another's, as this would be double-counting many of the chances. Now, applying the corrected scenario to the 23 people needed to answer the original birthday problem question results in $22 + 21 + 20 + ... + 1 = 22(23)/2 = 252$ separate chances. It is no coincidence that 252/365 corresponds well to the n/N approximation discussed in connection with de Méré's Paradox.

Assuming birthdays are uniformly distributed among 365 days (and Berresford 1980 shows that the birthday problem is robust to real-life deviations

from this assumption) and disregarding February 29 birthdays, the probability of at least one match among n people is given by the formula:

1 – Pr(no matches) = 1 – [(365/365)(364/365) ⋯ ((366 – n) / 365)] .

Students can easily do computer simulations of this situation. They can also collect data from groups of 23 people (easily obtainable from classrooms) recording whether there was at least one match in about half of the groups of 23. As Falletta adds (1990, p. 123): "In addition to getting twenty-four people together, you could check a reference work such as Who's Who or an almanac for the birth dates of twenty-four randomly-selected individuals or groups of presidents, writers, inventors, and so on."

Shaughnessy's (1977) dissertation study included an experiment in which students "guessed the number of cards they would have to turn over to have at least a 50% chance of getting at least one ace from an ordinary well-shuffled deck of cards." Shaughnessy reports: "The guesses made by the subjects [the guesses ranged mostly from 12 to 15] indicated that they were more aware of the deceptive nature of the probability of disjunctive events than they had been at the beginning of the course. Only one student guessed that it would take 26 cards to have at least a 50% chance." While Shaughnessy reports that "the tendency to use 50% as a representative multiplier [as most students did in the classical birthday problem on a pretest] of the total population had practically disappeared in the experimental group … about 6.5 weeks into the experimental course," we cannot conclude that student answers were due solely to the use of counterintuitive examples in instruction.

Because of its fame since being proposed in 1939 by Richard von Mises, the birthday problem usually appears in some form in most textbooks. However, it is often presented more as a historical footnote or enrichment exercise, rather than as an example of a problem worth serious attention. It is not clear whether this is due more to its counterintuitive result or due to the complexity of the calculation and application of both the complement and independence rules. The examples of disjunctive events calculations are usually both simpler computationally but also far less relevant to a real life situation. In any case,

rarely does a textbook follow up with an additional real-life though less-famous situation analogous to the birthday problem, although numerous examples exist.

One real-life illustration of probabilities of disjunctive events arose from a highly publicized course (e.g., Elliot 1993, Ringo 1993) on the Texas Lottery taught by the author at the University of Texas at Austin for University of Texas Informal Classes in the fall of 1993. In the "Pick Six" Texas Lotto game, 6 balls are randomly drawn without replacement from a set of 50 balls numbered 1 to 50. Students noticed that it was not uncommon for at least one number to be drawn that had also been drawn in the previous 6-ball drawing and asked its probability. Using 50-space spinners, students quickly simulated a large number of drawings and were surprised that this event seemed to occur as much as half the time. The exact probability is actually greater than 1/2:

$$1 - [(44/50)(43/49)(42/48)(41/47)(40/46)(39/45)] \approx 5/9$$

3.5 Classification Paradox

The Classification Paradox (so named in articles such as Reinhardt 1981) involves confusing a conditional probability $P(A|B)$ with its inverse conditional $P(B|A)$. This is one interpretation for the confusion students often encounter in "the interpretation of what it means to reject the null hypothesis" (Shaughnessy 1992, p. 474). The Bayesian focus is on the probability of a hypothesis given particular data (the "inverse probability"), while the frequentist focus is on the direct probability of the data given a particular hypothesis (e.g., Gigerenzer and Murray, 1987, p. 8).

In any event, the counterintuitive fact is that, when testing for a condition (HIV, drug use, birth defect, etc.) that is rare in a particular population, even with a fairly accurate test, most positive results will be false positives. After an exercise, Moore and McCabe (1993, p. 364) declare that the example "illustrates a fact that is important when considering proposals for widespread testing for AIDS or illegal drugs." In their exercise with real-life numbers, $P(+ \text{ test} | \text{HIV+})$ = .997, $P(- \text{ test} | \text{HIV}-)$ = .985, and $P(\text{HIV+})$ = .01, yet $P(\text{HIV+} | + \text{ test })$ is only .40. One reason students tend to overestimate $P(\text{HIV+} | + \text{ test })$ is that they may

be committing the fallacy of (see Shaughnessy, 1992, p. 471) ignoring the base rate information that P(HIV+) = .01. The probability P(HIV+ | + test) would be even less for less accurate tests or for populations in which the trait of interest is less common. For a concrete illustration of this example, assume 100,000 tests are given. On the average, 1000 will be HIV+ and 997 of these people will yield a positive test result. Of the 99,000 who are HIV–, (.015)(99,000) = 1485 will test positive anyway. Therefore, there are more false positives (1485) than true positives (997)! The mathematical approach of the probability "chain rule" known as Bayes' theorem—

$$P(A \mid B) = P(A \text{ and } B) / P(B) = P(A) \, P(B \mid A) / P(B)$$
$$= P(A) \, P(B \mid A) / \{ \, P(A) \, P(B \mid A) + P(\text{not } A) \, P(B \mid \text{not } A) \, \}$$

where A = actually having the trait

and B = test says "positive" for the trait

—can be supported by a number of representations, such as a contingency table, a Venn diagram, a tree diagram (e.g., Moore and McCabe 1993, p. 350), or a reverse flow diagram (e.g., Chu and Chu, 1992). These same techniques can certainly be applied just as well to conditional probability problems that do not lead to counterintuitive results (e.g., Moore and McCabe 1993, p. 378, exercise 5.75), yet such problems are not as likely to capture students' attention.

Just as Simpson's Paradox has occurred in publicized real-life situations (as discussed in section 3.3), the Classification Paradox has also received its share of attention (e.g., vos Savant 1993, Paulos 1994, Pringle 1994). Pringle (1994, p. 51) describes how a rape case was actually overturned after judges were convinced that the paradox had cast doubt on the verdict of the original trial:

… forensic evidence answers the question "What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?" But the jury must try to answer the question "What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?" … a very small answer to the first question does not necessarily imply a very small answer to the second. … it is wholly wrong to imply that the DNA match probability is the same as the probability of the defendant's innocence. "A DNA test showed that the

chances of the defendant not being the attacker were 859 million to one" is a typical newspaper error. This type of statement has been dubbed "the prosecutor's fallacy."

A final example is offered by Kernighan (1994): "[I]f you correctly predicted two coin flips in a row—a 1-in-4 stroke of luck—would you conclude that your chances of possessing ESP were 3 in 4?"

3.6 Required Sample Size

Freedman et al. (1991, p. 336) concisely state a counterintuitive result involving required sample size required for reasonable precision: "When estimating percentages, it is the absolute size of the sample which determines accuracy, not the size relative to the population. This is true when the sample is only a small part of the population, which is the usual case." The authors then provide an example:

There are about 1.2 million eligible voters in New Mexico, and about 12.5 million in the state of Texas. Suppose one polling organization takes a simple random sample of 2,500 voters in New Mexico, in order to estimate the percentage of voters in that state who are Democratic. Another polling organization takes a simple random sample of 2,500 voters from Texas, in order to estimate the percentage of Democratic voters there. Both polls use exactly the same techniques. … It does seem that the New Mexico poll should be more accurate than the Texas poll. However, this is one of the places where intuition comes into head-on conflict with statistical theory, and it is intuition which has to give way. In fact, the accuracy expected from the New Mexico poll is just about the same as the accuracy to be expected from the Texas poll.

Freedman et al. (1991, p. 337) explain this fact by first using a "box of tickets" model to show why population size is completely irrelevant if the sampling is done *with* replacement, and then why the multiplicative correction factor that must be used for sampling *without* replacement is nearly one (and thus can be ignored) when the sample is not a substantial fraction of the population. The fpc ("finite population correction" factor), is $(N - n) / (N - 1)$, which is indeed nearly 1 when the sampling fraction n/N is small, where n and N are sample and population sizes, respectively. As Cochran (1977, p. 25) states: "In

practice, the fpc can be ignored whenever the sampling fraction does not exceed 5% and for many purposes even if it is as high as 10%. The effect of ignoring the correction is to overestimate the standard error. …" Cochran (1977, p. 52) derives this formula for estimating the variance of the sample proportion: $[pq / (n-1)][(N-n)/N]$, where p is the sample proportion and $q = 1 - p$. If n/N is small, this formula clearly depends on n, not n/N.

Perhaps a more dramatic illustration is offered by Paulos (p. 35, 1994): "Although it seems counterintuitive, a random sample of 500 people taken from the entire US population of 250 million is generally far more predictive than a random sample of 50 people out of a population of 2,500." The formula in the preceding paragraph will verify that in Paulos' example, the sample of 500 people has an estimated variance that is about one-tenth as large as (and therefore considerably more predictive than) the sample of 50 people.

Freedman et al. (1991, p. 339) later offer this accessible analogy to defuse the dissonance of the overall situation: "Suppose you took a drop of liquid from a bottle, for chemical analysis. If the liquid is well mixed, the chemical composition of the drop [i.e., the sample] would reflect quite faithfully the composition of the whole bottle [i.e., the population], and it really wouldn't matter if the bottle was a test tube or a gallon jug." This conceptual intuition can be reinforced with comparing binomial and hypergeometric distributions (which represent the "with replacement" and "without replacement" sampling schemes, respectively) as in Berenson and Levine (1989, p. 236).

Since the sample proportion can be considered as a sample mean of 0's and 1's, it should not be surprising that these results about the sample proportion can be generalized to the sample mean. Cochran derives the following formula for the estimated variance of the sample mean when sampling is done *with* replacement: $(s^2/n)(N-1)/N$, where s^2 is the sample variance. Cochran derives the following formula for the estimated variance of the sample mean when sampling is done *without* replacement: $(s^2/n)(1-(n/N))$, which is equivalent to the "with replacement" formula times the fpc factor, $(N-n)/(N-1)$. The confidence intervals in Cochran (1977, p. 27) are of the form $x' \pm z(s)\sqrt{[1-(n/N)]}/\sqrt{n}$, which depends on n, not n/N, if n/N is negligible. Slonim (1960, p. 74) offers a specific

data set to illustrate that "as the universe increases in size the sample size (needed to obtain a specific target of precision a with a certain level of confidence) remains remarkably constant." The coefficient of variation was apparently known for this data set, and precision is defined as the quantity after the $\pm$ sign divided by the mean. While this is useful to demonstrate if time permits, the example with proportions has a simpler formula, and perhaps is more common to introductory students' daily experience, given the ubiquitous presence of opinion polls in the media.

3.7 Looking Ahead

This chapter has illustrated several key counterintuitive examples in statistics. The next chapter presents two points of view on their use, followed by a model that incorporates these and several other examples to the framework of a syllabus for the introductory statistics course.

CHAPTER 4

A PAIR OF PARADIGMS ON PARADOX

4.1 The Traditional Position: "Paradox Lost"

One perspective on the role of counterintuitive examples is expressed in the fifth of Gail Burrill's (1990, p. 113) ten suggestions for teaching statistics and probability in the spirit of the *Standards*: "The emphasis in teaching statistics should be on good examples and building intuition, not on probability paradoxes or using statistics to deceive." This study labels this perspective the Traditional Position for two reasons. First, it appears to be used by most instructors, although there has not been a formal survey to confirm this. Second, Burrill's ten suggestions also appear in *Guidelines for the Teaching of Statistics K–12 Mathematics Curriculum*, which was published by the American Statistical Association's Center for Statistical Education in March 1991, and therefore carry the status of official sanction. Because this position is the status quo and is well-known, this study will give less attention to explaining the Traditional Position relative to the Alternative Position.

It is the spirit of Burrill's statement that is the key. Many of the items listed as counterintuitive in the model (SPICE) in section 4.7, for example, are in fact typically covered in a Traditional course (e.g., "correlation does not imply causation") and are marked with an asterisk in the model. The important

distinctions to be made are that Traditional teachers will consistently avoid many of the paradoxes in the model of section 4.7, and will not take full advantage of the learning opportunities possible (see section 4.3 for a discussion of these) with the counterintuitive items they do cover.

Some textbook authors seem to specifically invoke this Traditional Position in the context of the introductory course. Even Moore and McCabe (1993), whose book certainly includes some counterintuitive examples, nevertheless advocate a somewhat streamlined approach (p. xv): "Because judgment is developed by experience, an introductory course should present firm guidelines

and not make unreasonable demands on the judgment of students … as is appropriate in a first course, most exercises are straightforward even at the cost of some oversimplification."

Burrill's thoughts are also echoed in Falk and Konold (1992, p. 161): "It is tempting to bring some of the more devious problems to the classroom to demonstrate to students their erroneous tendencies and perhaps enlighten them. However, if a teacher persists in pointing out to students how prone they are to inferential errors, they may become so convinced of their incapacities that they despair of ever mastering more appropriate techniques." There does not seem to have been but there should be research to investigate whether in fact teachers that follow the Traditional Position do so because of their belief in this statement. Extending the weightlifting analogy introduced in section 1.1, a teacher from the Traditional Position would avoid virtually all weightlifting because some people have been injured by lifting too much or too often.

A teaching model that is consistent with or sympathetic to the Traditional Position might be the "bridging analogies" approach from science education (e.g., Duit 1991, Brown 1992, or Clement 1993), in which teachers use carefully chosen concrete anchoring examples from students' experience to induce the student to make an analogical jump to the correct target intuition. Indeed, some of the metaphors in section 2.3.1 or section 4.7 might well be candidates for such bridging analogies. As Falk and Konold state (1992, p. 161): "[I]t seems reason-

able to begin instruction in probability by building on students' sound intuitions. Valid probabilistic intuitions are not hard to come by. Despite the abundance of studies describing people's inferential biases and shortcomings, many of the rules prescribed by probability theory are compatible with common sense."

This spirit of building on students' sound intuitions needs to be further related to the ample research that has been done on heuristics. For example, Shaughnessy (1992, pp. 478–479) states:

Although there are circumstances where reliance upon heuristics such as representativeness and availability can result in biased, nonnormative probability estimates, in some contexts these heuristics are very useful … Availability … is

often a very useful organizer for decision-making. … The very reason we try to draw a random sample from a population is so that it will be representative of the population. … Representativeness is, therefore, fundamental to the epistemology of statistical events . … [I]t is not that there is something wrong with the way our students think, just that they—and we—can carry the usefulness of heuristics too far."

Piagetian developmental theory also seems consistent with the Traditional Position in that it implies that probability concepts are developed in natural, gradual ways, without the need of confrontative intervention. The three-stage model of Piaget and Inhelder (1975) involves stage transitions near ages 7 and 14.

4.2   Limitations of the Traditional Position

To the extent that the Traditional Position is based on an exaggerated fear of intimidating students, it needs to be confronted. The statement in section 4.1 by Falk and Konold exaggerates the Alternative Position and can be addressed by the constructivist views in Roth (1990, p. 160):

Instruction for meaningful learning cannot be a simple matter of pointing out to students the conflicts between their own ideas and scientists' ideas. Telling students their ideas are 'wrong' and explaining to them why other explanations are better will not engage students in the process of actively constructing meaning. Researchers exploring conceptual change models of instruction emphasize the need to devise ways to engage students in actively thinking and puzzling about the phenomena they are studying.

Duit (1991, p. 665) claims that "[a] main problem with the approach is that there may not be enough good anchoring situations and bridging analogies available" and also students may not be able to understand or apply the analogies as intended by the teacher.   Clement et al. (1989, p. 558) bring up another problem, namely when students "refuse to believe that the prediction [about the anchor situation] applies to the target situation. Apparently, they cannot transfer the key relationship to the target. In such a case we refer to the anchor as *brittle."*

Also, appeals to developmental stages must be made carefully, as these have been found to be problematical by some researchers. Scholz (1991, p. 246) details how information processing models reveal the inadequacy of aspects of cognitive-developmental models by Piaget and Inhelder.   Scholz quotes the results of M. Scardamalia (1977), who "showed that even concrete-operative children are ready to solve combinatorial problems provided that the task demands do not exceed their information processing space." In any event, many researchers (e.g., Hudak and Anderson 1990; Allen et al. 1987) have found that substantial numbers of college undergraduates had not achieved the level of formal operations that Piaget associates with age 14 and up.   Also, even misconceptions studies that suggest "a developmental scheme of stages" (Garfield and Ahlgren 1988, p. 51) nevertheless call for an active role on the part of the teacher to facilitate "the transition to the next stage."

A final disturbing threat to the Traditional Position is a study by Greeno (1983), who found that intuition with respect to randomness actually decreased with age among students aged 11–16 years.   Finally, the position of Burrill (1994) is based mostly on "what happens in my classroom" rather than on formal research.

4.3   The Alternative Position: Counterintuitive Examples

Based on the reasons given in section 4.1 for labeling the avoidance of counterintuitive examples as the Traditional Position, the use of counterintuitive examples is considered the Alternative Position in this study. Borasi makes a related point (1994, p. 170, emphasis in original) when he says: "To fully appreciate the radical nature of approaching errors as springboards for inquiry in

mathematics instruction, it is also important to realize that such an approach is at odds with most teachers' and students' current views of errors and also differs considerably from the uses of errors made by most mathematics education researchers to date."

Nevertheless, mathematics educators in general are finding many motivations for counterintuitive examples. Hansen (1994, p. 202) states: "Whenever we can surprise our students with counterintuitive results—Why did *that* happen?—their increased interest level presents an exceptional learning opportunity." Konold (1994, p. 233) adds: "Finding a surprising result, students are more motivated than they otherwise would be to understand what is going on. They eagerly express opinions in class discussions." The active discussions that much more naturally accompany a counterintuitive example (see section 5.2) very much support the increased emphasis on mathematical communication called for by the NCTM (1989, Curriculum Standard #2).

Gordon (1991, p. 511) lists many additional benefits of the Alternative Position:

For not only do instances that run counter to intuition gain students' attention because of the disequilibrium experience when what had been imagined to be true turns out not to be so, but such examples also help students challenge habits of thought and practices, thus leading to their becoming better thinkers (Marzano et al. 1988, p. 128). By presenting students mathematical moments that challenge common sense and common practice, the teacher gives them the opportunity to gain a greater appreciation of the need for exploration, reflection and reasoning.

In statistics, this further translates to an opportunity to gain a greater appreciation of statistics as an empirical science, and in valuing this, be less likely to make decisions based on intuition alone (a further point on this is made in section 5.2). For example, Kohn (1992, p. 218) presented a version of the Monty Hall problem (discussed in more detail in section 3.2), in which he put a $1 bill in one of three envelopes, had a volunteer pick an envelope, revealed one of the unchosen envelopes to be empty, and asked if the volunteer would have the best chance of getting the dollar if she stayed with her initial choice, switched to the other unopened envelope, or if it made no difference. Most students initially did not have the correct view that switching is best. After conducting an experiment

that tallied the subjects' choices and consequences of the choices, and observing that switching resulted in a significantly greater proportion of wins, students completed a survey in which students "rated whether they would base their actions on intuition only (1) to research only (9)." "Results of the Trust in Research Survey indicate [$p < .1$] that students who participated in the demonstration had higher trust in the empirical technique than students before the demonstration."

Counterintuitive examples may also help students appreciate the relationships and differences between empirical investigation and deductive mathematical logic. According to the NCTM (1989, p. 187): "It is essential that students come to understand the difference between the right-or-wrong quality characteristic of most mathematical thinking and the qualified nature of outcomes in statistical analysis. It is equally important, however, that students do not extrapolate beyond this fact to reject statistical thinking because it allows counterexamples. Instead, they should recognize that statistics plays an important intermediate role between the exactness of other mathematical studies and the equivocal nature of a world dependent largely on individual opinion."

Romano and Siegel (1986, p. vii) make a related point:

Counterexamples can also provide insight into a problem—for example, by showing which hypothesis needs to be strengthened in order to achieve a true result, by helping to establish a result as 'best possible' or by clarifying the need for a particular choice of definition. A less technical but perhaps more important theme is that in statistics we really have no set principles that work in all situations.

A well-chosen trip through the temporary "fog" of paradox can lead to deeper, lasting insight, as Rapoport (1967, p. 52) relates:

... [A]n advance in generalization can exorcise a paradox by restoring a familiar law that may have been lost in a preceding generalization. ... It was found, paradoxically, that in some number domains, algebraic integers could be factored into prime numbers in more than one way. ... Further study showed, however, that algebraic integers could be generalized into objects called "ideals," and it turned out that these can be factored into primes in only one way.

As section 3.3 demonstrated, the pedagogical device of geometric representations restored and strengthened confidence that might have been temporarily jarred by Simpson's Paradox.

Counterintuitive examples share many similarities to what Borasi (1994, p. 168) refers to as anomalies, and thus may share some of their benefits (listed at the end of section 4.3) and theoretical framework:

First of all, important contributions were provided by Dewey's and Peirce's view of knowledge as "a process of inquiry motivated by doubt" (Dewey, 1933; Siegel & Carey, 1989; Skagestaad, 1981). Within this view, *anomalies* (i.e., 'things that do not make sense or perceptual judgments or observations that seem unexpected' [Siegel & Carey, 1989, pp. 23–24]) play a key role because they are considered likely to create the kind of doubt that can set the inquiry process in motion. … Anomalies also play a key role in Kuhn's thesis that scientific knowledge is achieved through an alternation of "normal science" and "scientific revolutions." Kuhn (1970) observes that in the history of science some unacceptable results or unsolvable problems have occasionally challenged the very paradigm within which a certain area of science was developing, thus motivating a search for an alternative paradigm that revolutionized research in that area.

Thus, a student who experiences counterintuitive examples gains appreciation for the typical and historical process by which statistics and mathematics grows. (Certainly a counterintuitive or unsatisfying result from frequentist inference has led more than one statistician to embrace the Bayesian paradigm.) Kunoff and Pines (1986, p. 210) state that, more than for most areas of elementary mathematics, "[i]t is possible to present problems that puzzled the experts of the time in which they originated but which can readily be solved by students once some elementary probability concepts are developed." Indeed, students' motivation is often connected to this very fact (ibid., p. 214): "Many students are anxious to try Fra Paccioli's problem [the problem of "points," which at about 500 years old, is one of the oldest probability problems to draw the attention of mathematicians] when they learn that it was not solved by either Cardano or Tarataglia, both being important mathematicians of the sixteenth century who worked on it extensively."

To the extent that a counterintuitive example might be readily be exploited by advertisers or activists, a teacher can provide a strong sense of empowerment

to students by allowing them to actively confront and demystify the example. In the words of Huff (1954, p. 9): "The crooks already know these tricks; honest men must learn them in self-defense." This is also one of the suggestions of Garfield and Ahlgren (1988, p. 48). If it is worth warning students about, for example, "abuse of statistical tests" (Moore and McCabe, 1993, pp. 475–477), then surely it is also worth warning students about, say, the abuse of contingency tables when Simpson's Paradox is neglected.

A problem that is not obvious, that takes a while to unravel and understand, is very much in the spirit of the NCTM's recommendation (1989, p. 6) that students become mathematical problem solvers: "… students need to work on problems that may take hours, days, and even weeks to solve. Although some may be relatively simple exercises to be accomplished independently, others should involve small groups or an entire class working cooperatively. Some problems also should be open-ended with no right answer, and others need to be formulated." While it may not be completely accurate to describe a particular counterintuitive example as "having no right answer," it certainly tends to have multiple representations (e.g., Simpson's Paradox in section 3.3, the Inspection Paradox in section 4.7) and multiple pathways towards resolution. Fischbein's "origins" paradigm of intuition (discussed in section 2.2) is nicely compatible with this position in that there is a sense of a "before" and "after" created by something powerful and immediate in between. Also, the extent to which structured controversies (discussed in section 5.2) have been found superior to concurrence seeking certainly supports the Alternative Position over the Traditional Position.

Also, the conflict between intuitive and formal approaches mentioned in section 1.1 may be alleviated with the Alternative Position. While invoking the "cognitive sense" of intuition, a counterintuitive example may also activate deductive cognitive processes by forcing a learner to go through the comparing, contrasting, sorting, separating, exploring and testing that a counterintuitive example induces. Steen and Seebach (1978, p. iii) certainly found this true in topology: "Not only are examples more concrete than theorems—and thus more

accessible—but they cut across individual theories and make it both appropriate and necessary for the student to explore. …"

Metacognition is yet another benefit of the Alternative Position, as described by Falk and Konold (1992, p. 157):

Probability is especially rich in counterintuitive examples, which often entail fallacies and paradoxical conclusions. Some of these examples played an important role in the development of probability theory. Students may likewise benefit from comparing their intuitions concerning puzzles and paradoxes with normative solutions. This activity requires increased awareness of one's own thought processes. … Metacognition is no less important than learning the right solution, and reflective thinking is a vital step toward achieving abstract mathematical ability.

The Alternative Position is more consistent than the Traditional Position with the "conflict-discussion" lessons that have been successfully implemented by Bell (1986, p. 28):

Many people feel that class discussion in mathematics is difficult. This may be partly because we try to make it too convergent, aiming at appreciation not only of a single correct result but also the single correct line of reasoning towards it. But, as can be seen from any discussion based on a strongly felt conflict, many factors and many connections contribute to pupil's convictions. These all need to be brought out and aired. Discussions may have an element of repetition; normal discussions on any subject do, because it takes more than one cogent argument to shift an established view. Perhaps the most striking observation from all this work is that back-sliding is the norm. Even after clearly effective lessons with learning visibly taking place, in the next lesson, most of the class could slip again into the original error. True, the second recovery was quicker than the first. The method of conflict-discussion promises to provide a more effective way of dealing with this widely recognized phenomenon than simply reteaching.

Although Bell's original context concerned neither college students nor statistics, his use of group discussion and multiple paths to the answers are certainly qualities for which the NCTM (1989) has called for greater emphasis.

Another use for counterintuitive experiences has been found by Osberg (1993, pp. 110–111), who attempted to help "introductory psychology students overcome the misconception that psychology is just common sense. Early in the course, I recount Festinger and Carlsmith's (1959) classic cognitive dissonance

experiment and ask students to guess the outcome of the study." After the class is given feedback about their collective responses (which are almost unanimously intuitive but wrong!), the class is told the counterintuitive result of the actual experiment.

Evaluative data from the most recent time I used this demonstration suggest it achieved its aim and that students were very engaged by it … . Students' open-ended comments included: … "It helped me to understand that I must study in order to do well in this course because it is not just common sense but can involve surprising results." "You could have just told us psychology does not mirror common sense and I probably would have forgotten it. I won't forget it now."

In a related example, Carkenord and Bullington (1993) tried to convey an appreciation for the phenomenon of cognitive dissonance by having students complete a survey which was explicitly designed to induce cognitive dissonance in the students themselves. After debriefing the students with discussion, they surveyed the students and found that students considered the activity highly effective in understanding cognitive dissonance.   Such an assessment of student evaluations would be valuable in the specific context of statistics courses. Also, affective studies should be undertaken to assess student feelings of satisfaction and empowerment after understanding a counterintuitive example.

A summary of the learning opportunities that are offered by activities involving counterintuitive examples would include all of those listed by Borasi (1994, pp. 185–186):

experience constructive doubt and conflict regarding mathematical issues, engage in challenging mathematical problem solving, pursue mathematical explorations, reflect on the nature of mathematics, experience the need for monitoring and justifying their mathematical work, experience initiative and ownership in their learning of mathematics, recognize the more humanistic aspects of mathematics, verbalize their mathematical ideas and communicate them.

Indeed, the relationship between how students deal with counterintuitive results and how they deal with their errors should be further explored.

4.4 Limitations of the Alternative Position

Further research is needed to better understand and reconcile the positions of these authors, especially in light of recent studies by Konold et al. (1993, p. 412), who report that "[t]he results of the present study suggest one limitation to the cognitive-conflict approach—a situation designed to contrast normative with informal reasoning may produce no conflict. … Incompatibilities and contradictions at this level will probably not be noticed by students unless they are reasoning from a single, coherent framework."

Also, as discussed in section 3.1, what is counterintuitive is relative to the accumulated knowledge and experiences of a group of people. Thus, if a standard set of counterintuitive examples published for use in the classroom becomes widely circulated, it could eventually lose some of its effectiveness, as students become aware of the "right answers" without fully grappling with the under-lying issues.

The deeper kind of understanding that exploring a counterintuitive example can yield is rarely associated with an introductory college course. Beginning with Gelbaum and Olmsted (1964), books of counterexamples have been published in many areas of mathematics (including probability, graph theory, real and complex analysis, and topological vector spaces) and most are intended to serve as a reference or supplement (not as a text) for senior mathematics majors, graduate students, and professors. In his critique of the book of probability counterexamples by Stoyanov (1987), Durrett (1989, p. 405) states: "I personally find it undesirable to dwell exclusively on what can go wrong without talking about how things are proved." This call for a context and balance is echoed by Burrill (1994): "… students need to see probability paradoxes. The critical thing is not to screen them but to make sure they have the right tools to analyze the paradoxes in a meaningful way. Most of the time they are presented as magic with very little foundation for kids to work from."

4.5 Moving Towards Synthesis: Lesser of Two Evils

Garfield and Ahlgren (1988, p. 48) reveal a tension by recommending that teachers "recognize and confront common errors in students' probabilistic thinking" and yet also "create situations requiring probabilistic reasoning that correspond to the students' views of the world." It is, however, encouraging that the use of bridging analogies and anchors consistent with the Traditional Position and the use of conceptual change teaching strategies consistent with the the Alternative Position have been combined by some researchers in science education such as Brown (1992, 1993).

It is not clear a priori that it is any less contrived to provide "nice" data sets (e.g., Read and Riley 1983) or data sets which have surprising aspects. Shaughnessy quotes (1992, p. 479) Clifford Konold as noting that "psychologists actually have to search for situations that will lead people astray," but does not offer justification or specific reference. Furthermore, something that would consistently lead one astray may, by its very nature, be difficult to detect without a careful search.

It is also not clear a priori whether Burrill and Gordon would be so far apart in practice. After all, Burrill's guideline in section 4.1 stated what the "emphasis" should be on, not necessarily that counterintuitive situations should be ruled out completely. And yet, Burrill's phrase "building intuition" suggests that only "nice" situations can build intuition. On the other hand, it is unlikely that Gordon intended to suggest presenting nothing but counterintuitive examples.

There seems to be no hard evidence that a course that completely avoids counterintuitive examples would be most effective in achieving certain affective and cognitive goals. On the other hand, while there is some support for the use of well-placed selected counterintuitive examples, there seems to be no hard evidence that they should be used all or even most of the time. If one accepts Mevarech's premise (1983, p. 423) that "[t]wo of the main characteristics that distinguish deep understanding and rote learning are the ability to deal with familiar and unfamiliar situations and the capacity of thinking in concrete and abstract manner (Klausmeier, 1980)," it seems to be a very reasonable extension to add the phrase "(and deal with) intuitive and counterintuitive examples."

The spirit of moderation and synthesis in statistics education is not limited to counterintuitive examples. For example, introductory classes traditionally may mention Bayes' theorem, but do not allow for discussion of the Bayesian paradigm of inference in general, even though it is often closer to students' intuition than the frequentist interpretations. Nevertheless, Albert (1994, p. 4) states he is "not planning a Bayesian revolution. Generally, statisticians use methods that are developed from a frequentist viewpoint and we should continue to teach methods that are used in practice." Albert (p. 5) simply suggests "[w]hen appropriate, give and encourage Bayesian interpretations to frequentist inferential procedures." It is this same "when appropriate" spirit that guides the building of the model of this study.

Further support to this spirit is provided by Falk and Konold (p. 161, 1992):

In probability, as in physics, "not all preconceptions are misconceptions." That expression is borrowed from the title of a paper by Clement, Brown, and Zietsman. They propose that in teaching physics it is desirable to ground new material in students' intuitions that are in agreement with accepted theory. Likewise, in teaching probability, one expedient strategy is to offer nontrivial probability problems for which students can guess whether the answer is greater or smaller than a given number. This would allow the student to experience the satisfaction of having the prediction borne out by the results of the probabilistic calculation. They may realize that commonsense can still be a good guide though it should be exercised with caution.

This is related to certain heuristics, as discussed in section 4.1, and is also supported by Moses (1986, p. 7):

Most of the statistical methods encountered in a first course have the nice property of making sense, at least after due consideration. It is fair to say that statistics is largely an extension of common sense. You should *expect* the statistical procedures in this book to seem reasonable and not to conflict with your intuition (at least after some reflection). If you *do* find that your eyeball appraisal of a body of data differs sharply from the result you obtain by applying some formula, that is a warning flag. … If careful restudy confirms the original analysis, then your intuition will have received a useful corrective. But only

rarely should your thoughtful judgment and statistics say quite different things about a set of data.

Unfortunately, there may be a double-bind situation here for the Traditional Position in that the above quotation requires "reflection" and "due consideration" for most methods to make sense, an entire process that may be difficult to elicit from students if the instructor is overly concerned with not challenging students too much.

4.6 Issues Considered in Building the Model

The model of synthesis proposed in section 4.7 attempts to address the issues raised in this chapter. The model reflects Albert's "when appropriate" spirit of balance by insisting on criteria for counterintuitive examples (e.g., section 3.2) and by suggesting guidelines for their use in the classroom. For example, the model suggests structured controversies to explicitly address Konold's concern that contradictions could go unnoticed (mentioned in section 4.4).

The concern of overuse of the same counterintuitive examples can be addressed in many ways. First, new examples are constantly being created (indeed, the syllabus portion of the model has more than one example for most topics). Also, more than one context is possible for most counterintuitive examples (see chapter 3) and students are often slow to recognize an old concept in an updated context. Second, examples need not always be telegraphed as counterintuitive. Third, and perhaps most important, is that students can and should be taught in a manner in which the process and reasoning is a primary focus, not merely the correct answer. Some students may have already heard the "punchline" before taking a statistics class, but probably do not already have a deep understanding of the structure and context and why their primary intuition is wrong.

In section 4.3, the relationship between instructional goals and appro-piateness of counterintuitive examples was discussed at length. During the model building process, the following issues were also considered:

Should the quantity or frequency of counterintuitive examples relative to intuitive examples be HIGH, MEDIUM or LOW?

There is certainly precedent (e.g., Tennyson 1990) for linking cognitive learning theory with instructional prescriptions. However, Tennyson's idea of assigning a percentage of time for different instructional prescriptions may not transfer to statistics in that it may be impossible to discuss a counterintuitive example thoroughly without referencing "intuitive" ideas as well, thus rendering the allocation distinction meaningless. Another position explored was that the model should reflect the proportion of counterintuitive occurrence in real life, but this could be highly dependent on the area of application (and introductory statistics courses are often tailored for majors in a particular subject, such as business, psychology, etc.). A more defensible perspective is that it should depend on the goal of instruction. If the goal is merely computation and basic skills, then the level should clearly be LOW, while if the goals emphasize metacognition and critical thinking, then the level should be HIGH. In the syllabus-driven model in section 4.7, there could usually be one or more major counterintuitive examples discussed or experienced in class per curriculum topic, some of which (e.g., Simpson's Paradox) should be considered more "core" than others.

When counterintuitive examples are used, should they PRECEDE or FOLLOW any intuitive examples from the same topic?

The syllabus-driven model makes clear that counterintuitive examples are best used to expand and explore concepts after they have already been introduced in a more intuitive, student-centered way, in a similar manner to the reference by Falk and Konold at the end of section 4.5. Because of their ability to motivate, however, such examples may also occasionally be used to introduce new topics as well (e.g., Gordon 1991). Gonick and Smith (1993) use de Méré's Paradox as a vehicle to introduce basic probability definitions and rules.

Should counterintuitive examples be telegraphed (either pre-announced or setup with a sequence of leading questions) when used, or simply presented with an open-ended question for students to explore?

While many books attempt no such distinction, others put special markings by problems that are "particularly thought-provoking or have no 'exact' answer" (Berenson and Levine 1989, pp. vii–viii). There is not only a lack of consistency here between textbooks, but even *within* textbooks. For example, Moore and McCabe (1993, pp. 195–197) give exercises (e.g., 2.77, 2.81) that are telegraphed as well as not (2.80). This variable, while there does not seem to be research on it, is probably not a crucial one in that students will quickly get the point that a specific example is counterintuitive before too long into an activity, and will certainly be less surprised to encounter another one later in the same semester.

Because real-life data sets do not come tagged with an announcement of counterintuitiveness, students should not have counterintuitive examples pre-announced as such, if a goal of instruction is to prepare students for real-life experiences (a goal advocated by the NCTM (1989), as mentioned in section 3.2).

How counterintuitive should the examples be?

This is problematic to measure, as discussed in section 2.4.2. Fischbein (1987, p. 78) states that "there is no systematic experimental data available concerning the intuitiveness of various logical rules. It seems that researchers have not been concerned so far with this question." Piaget (1985, p. 150) refers to "optimizing equilibration" as "a process that leads to better equilibrium rather than simply returning to more stable forms of a former equilibrium. It unites constructions and compensations in an indissociable way." However, Piaget gives no guidance as to how to achieve this or measure this in the real-life setting of a classroom environment.

Festinger (1957, p. 55) talks of looking "for some overall measure that will reflect the total magnitude of dissonance which exists. In the experiment which we conducted … the actual measure employed was the subject's rating of how confident he was that his decision was the correct one. This measure was

used on the assumption that the greater the dissonance … the lower would be the confidence expressed by the subject." Festinger's model of the relation (p. 130) between magnitude of dissonance and active seeking of new information suggests that there is indeed an optimal amount of dissonance (which interacts with whether or not new information is expected to increase dissonance) if the goal is to encourage students to seek new information. This seems related to staying within what L. S. Vygotsky referred to as a student's zone of proximal development. It seems clear that dissonance must be defined relative to what each individual student already knows or believes to be true, and so (even assuming you can always accurately measure what students know and believe in the first place) there can be no one-size-fits-all measure developed or applied.

If a goal is to create dissonance (under the assumption that this will ultimately lead to a lesson with greater impact), then group activities in which members are forced to commit to answers before and after the activity seem to be most effective. This was found in psychology research in general (e.g., Lewin 1952), and has been utilized in experiments in statistics education as well (e.g., Shaughnessy 1977). In practice, it will come down to each individual instructor's assessment of her students, goals and time available as to which counterintuitive examples should be included.

How should individual student differences be taken into account?

As mentioned in section 2.4, a group anchor may not also be an individual anchor for every individual. Wicklund and Brehm (1976, pp. 225–228) discuss the influence of factors due to individual differences such as self-esteem and introversion-extraversion. For example, Wicklund and Brehm's (p. 225) finding that "… in experiments by Glass (1964) and Gerard, Blevans and Malcolm (1964) … self-esteem was manipulated experimentally, and both of them found that dissonance was greater among high self esteem individuals" is certainly consistent with the findings of Dweck and Leggett (1988), who might suggest, for example, that a student whose goal orientation is learning would thrive on the challenge of a counterintuitive example, while a student with a performance

orientation might feel discouraged. Ultimately, however, the focus and form of the model emerged in such a way as not to emphasize individual differences.

4.7 SPICE: Structured Progression Involving Counterintuitive Examples

While some introductory statistics books such as Moore and McCabe (1993) are organized at least as much by conceptual theme as by tool, others follow a more traditional sequence. For example, Moore and McCabe (1993) introduce regression and time series before introducing probability models and estimation. The topics in the Structured Progression Involving Counterintuitive Examples (hereafter abbreviated as SPICE), are in the sequence that many introductory courses typically use, although some topics might be omitted or covered in less depth due to time constraints. Indeed, it would be extremely challenging for a one-semester course to cover the entire SPICE, and a current trend in education reform is to cover fewer topics, but in greater depth.

The general features of this model are that intuitive metaphors and activities are used to develop an initial conceptual framework for a topic. The counterintuitive examples are then used for expanding and refining the framework, encouraging higher-order thinking skills and appreciation of statistics as a field of real-life inquiry, as discussed in section 4.3.

Some topics feature a closer connection between their counterintuitive and intuitive examples than other topics do. This is because some of the counterintuitive examples involve a specific counterintuitive result that is counterintuitive in at least a qualitative sense (e.g., Simpson's Paradox, Central Limit Theorem), while other examples simply involve a tendency to consistently overestimate or underestimate a quantity in an otherwise straightforward situation (e.g., combinatorial growth, runs in a random process).

The level of complexity of the counterintuitive examples increases fairly naturally in a parallel to that of the corresponding topic. If one accepts Brewer's premise (1985, p. 253) that "statistical inference [as opposed to descriptive statistics] is subject to more misinterpretation because of its probabilistic character," then it is precisely the latter (and more difficult) part of the syllabus that has the greatest need for a device such as a counterintuitive example to

expose the misinterpretations. Indeed, the counterintuitive examples for the descriptive topics at the beginning of the syllabus often seem scarce or somewhat marginal in fundamental significance.

What follows is a brief listing of the SPICE, followed by a more in-depth prose discussion of each topic in the table. For each topic, the in-depth SPICE discussion begins with suggested intuitive activities before describing counterintuitive activities. The brief version of the SPICE, however, lists only the counterintuitive examples for several reasons: (1) it is the part that is more likely to be the "new" or "unusual" feature for an instructor; (2) it is the part more readily listed in a way that is both concise and clear; (3) the "intuitive" example(s)
for each topic are often of a similar form (i.e., students try to generate their own list first); (4) the counterintuitive examples often need more careful structure and scaffolding in their presentation.

The examples also vary considerably in how likely they are to be included in a Traditional course. Currently, the few examples that typically would be included in any significant depth are marked with an asterisk (*). Ironically, some of the starred examples may be actually be more counterintuitive to students than several of the nonstarred examples. This syllabus is not meant to be exhaustive in every detail, but merely to provide at least one example per major topic. For example, there is perfect correspondence between this study's syllabus and the typical syllabus listed in Garfield and Ahlgren (1988, p. 46), who present topics in three categories—descriptive statistics, probability theory, and inferential statistics—but later critique (p. 57) such an organization.

It should be also stressed that, while the model features a syllabus of examples and topics, the underlying motivations and context of the model as discussed throughout section 4 should be considered part of the model as well. As in the Chance Plus curriculum (see Konold 1991), the fundamental curriculum unit can be thought of as a lab involving an introduction to a new topic with questions and structured activities for individual and group work. The exact number of students per group, time allotted, and resources used, of course, must remain at the discretion of the instructor. There are also certainly variations

possible on the ordering of topics, as discussed by Field (1984) and implemented by Moore and McCabe (1993).

Structured Progression Involving

Counterintuitive Examples (SPICE)—brief version

<u>TOPIC</u>	<u>COUNTERINTUITIVE</u>
<u>EXAMPLE</u>	
Types of Numerical Data	Type can depend on context
Tabular/Graphical Displays	Histograms with unequal class intervals;
	Histograms vs. bar graphs;
	bar graphs vs. pictograms
Measures of	Averaging-the-averages misconception;
Central Tendency	average class size; Simpson's Paradox
Measures of Variability	None known at present; see detailed model
Independent Events	Disjoint □ independent*
Randomness	Longest run; number of runs & lead changes
Combinatorics	Factorial growth rates
Probability	Birthday problem*; deMéré's Paradox
Conditional Probability	Classification Paradox
Independent Random Variables	Dependence □ correlation_____
Distributions, Limit Theorems	Central Limit Theorem,* Law of
	Large Numbers
Estimation	Usefulness & having important properties
	need not imply each other
Regression	Regression fallacy
Correlation	Correlation □ causation*
	Zero correlation □ no relationship*
Hypothesis Testing	Classification Paradox
	Statistical significance □ practical sig.
	Practical significance □ statistical sig.
Experimental Design/ANOVA	Interaction effects; see detailed model
Sampling	For small n/N, precision depends on n,
	not n/N; Inspection Paradox

Structured Progression Involving Counterintuitive
Examples (SPICE)--detailed version

Types of Numerical Data

Students can be asked for examples of numerical data until they have given possible examples of nominal, ordinal, interval, and ratio data. They can then be asked to generate a list of ways how they are different and how they are similar and what can be inferred from them (e.g., it will be "intuitive" that there is no meaningful average of football jersey numbers, or that 80 degrees is not "twice as hot" as 40 degrees).

Students will have an excellent opportunity for critical thinking if they then consider the "numbered raffle tickets" and "number of cylinders in a car's engine" examples of Velleman and Wilkinson (1993, p. 69), which demonstrate that "[s]cale type … is not an attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we may have."

Tabular/Graphical Displays of Data

Students can first be given a set of quantitative data and asked to find several tabular and graphical ways to summarize it. The graphical methods they generate (with some guidance as needed) should correspond to some subset of the following: histogram, dotplot, frequency polygon, frequency curve, ogive, box-and-whisker plot. Students can also discuss the usefulness of such tabular methods as: frequency distribution, cumulative frequency distribution, relative frequency distribution, cumulative relative frequency distribution, and stem-and-leaf display. Finally, students can be asked what graphical and tabular methods are appropriate for qualitative data (bar chart and pie chart; frequency distribution and relative frequency distribution) and why others are not.

Because of factors such as similarity in appearance, students often confuse bar graphs and histograms. A histogram needs no vertical scale because it is the

area, not the height, of a block that corresponds to the number of observations in the class interval (which need not have the same range as all other class intervals) covered by the bottom edge of that block. It is not uncommon for the rightmost class interval to be larger than others when the data is sparsely distributed beyond a certain value. Therefore, there can actually be more people in the rightmost income class interval than in the leftmost income class interval, even if the rightmost block is not as tall as the leftmost block (e.g., Freedman et al. 1991, p. 31, figure 2). As for bar graphs, this is an appropriate place in the course to have students dissect certain graphs that are deliberately misleading, as in Huff (1954) or Reichmann (1961). One example is bar graph data that are depicted in a pictogram such that only the icons' heights (as opposed to the heights and the areas) are proportional to the numbers they represent. Another example is quantitative data plotted in a line graph in which zero has been omitted (without acknowledging this omission with a broken or jagged line) from the vertical axis. In section 5.2, there is discussion about a group activity used by Shatz (1985) to illustrate concepts such as how the size of a class interval can affect the appearance of a frequency distribution. Berenson, Friel and Bright (1993) have documented the tendency of elementary teachers to fixate on one graphical feature at a time (e.g., range, horizontal scale, modes, most frequent frequency, concentration of data points, absence of data point, number of columns) to interpret statistical data, so an activity such as a counterintuitive example that promotes critical thinking may help disrupt this fixation.

Measures of Central Tendency

Students can often themselves come up with most of these measures (the first three of which receive the most attention in a typical introductory course) of center or location: mean, median, mode, midrange ($[X_{min} + X_{max}]/2$), midhinge ($[Q_1 + Q_3]/2$), trimmed mean. Students can then discuss them (as in the sample group activity in the appendix of Garfield 1993), discovering relationships such as a unimodal right-skewed distribution would tend to have these relative positions: mode < median < midhinge < mean < midrange. In section 5.2, there is

discussion about a group activity used by Shatz (1985) to illustrate concepts such as how the skewed data can affect measures of central tendency.

Counterintuitive examples for this topic include the "average class size" example discussed in section 3.3. This example is related to the Inspection Paradox which is discussed in the topic of sampling. Also, students can be exposed to situations involving weighted averages and eventually confront the fact that the average of a set of averages need not be (but can be under certain conditions) the same as the average of all the original individual numbers. Since categorical data has already been introduced, Simpson's Paradox (also thoroughly discussed in section 3.3) can be discussed as well. Following Bruner's recommended progression from concrete to iconic to abstract, students could be exposed to the representation in figure 3, then figure 2, then figure 1. While most of the focus of this topic has been on means (arguably the most important measure of center), it should be pointed out that there are also examples involving modes or medians (e.g., Romano and Siegel 1986, pp. 55–58).

Measures of Spread, Variability or Dispersion

With a structured sequence of questions, students can be led to "discover" and appreciate most of these measures: range, interquartile range, mean absolute deviation (from the mean), mean squared deviation, variance, standard deviation (an "average" deviation from the mean in the original units of measurement), coefficient of variation.   Intuitive connections can then be made between the standard deviation and the normal distribution in several ways. First, after collecting data that should be expected to be reasonably symmetric and unimodal, students can empirically discover the so-called empirical rule, that about 2/3 of the data will be within one standard deviation of the mean, etc. Also, students can visually see that for a normal distribution, one standard deviation away from the mean is exactly where the direction of curvature changes from convex to concave (this description is a viable alternative to using the term "inflection point" since it is not assumed that students in introductory statistics have had calculus).

There does not seem to have been any counterintuitive examples yet reported for the topic of variability at the level of mere descriptive statistics. There are, however, a number of possible examples for the general phenomenon of variability, once randomness and independence have been discussed. For example Hogarth (1987, p. 18) describes how "[t]he amount of variability is positively related to the degree of randomness exhibited by the phenomenon" and illustrates it (p. 19) with the following problem of Kahneman and Tversky: "Boys are a majority (65%) in program A, and a minority (45%) in program B. There is 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 program A or to program B?" Also, Shaughnessy (1992, p. 478) discusses how students deny the existence of variability in the real world and do not understand the law of large numbers.

Independent Versus Dependent Events

It is admittedly difficult to generate examples of independent events that are related to the real world and yet involve neither approximated independence nor imprecision in specifying the sample space. A reasonable example students might explore is "Whether it rains today in Austin" and "Whether it rains tomorrow in Austin" (dependent events), versus "Whether it rains today in Austin" and "Whether it rains a year from today in Austin" (independent events, for practical purposes).

Traditionally, the concepts of disjoint and independent events are introduced on the same day, namely when the addition and multiplication probability rules, respectively, are introduced. Student difficulty with these concepts may be aggravated by a Venn diagram of disjoint events, in which the physical separateness of the circles erroneously seems to suggest independence of the events. This is, in fact, only true when at least one of the events has a probability of zero. So, while "the independence condition $P(A \cap B) = P(A)P(B)$ will always be satisfied when $P(B) = 0$" (Romano and Siegel 1986, p. 5), disjoint events in general will not be independent.

Randomness

Students at the nonstatistical or naïve statistical level (see Shaughnessy, 1992, p. 485) are likely to have a classical, or equiprobable, model of probability that assigns equal probability to all points in the space. While this model ultimately needs to be seen as just one of several, it is adequate for establishing a concept of randomness. Randomness is also related to the "intuitive" representativeness heuristic, as discussed in section 4.1.

Students tend to underestimate the expected length of longest run in a sequence of coin tosses, as well as overestimate the number of runs in the sequence. Schilling (1990) describes an interesting classroom experiment in which the class is divided into two groups, one group instructed to record the sequence of 200 coin tosses and the other group instructed to write down a reasonable simulation. Using only students' tendency to underestimate the longest run length, Schilling (p. 197) is able to correctly classify which group a student was originally in with about 85% accuracy: "The fact that one can easily and in a matter of minutes separate the two groups quite well stimulates considerable student interest … while at the same time strikingly driving home the message that human beings make rather poor randomization devices." Schilling (1994) discusses this same topic at a less technical level with more emphasis on real-world applications. Most students are surprised that, for example, there is a 50-50 chance of a run (of heads or tails) of length 3 or more when a coin is flipped 5 times. Riehl (1994) demonstrates a simple way to compute the probability of r runs recursively, without needing knowledge of Markov chains to understand the calculation. The number-of-runs test for randomness (e.g., Berenson and Levine, 1989, pp. 513–518; Lapin 1987, pp. 536–542) can be used as an important example of nonparametric statistics, a topic not often included in most introductory courses. Length-of-runs can be applied to control charts (e.g., Cryer and Miller 1991, p. 292).

Also, as Feller (1968, pp. 78, 80) cites a misconception that "a so-called law of averages should ensure that in a long coin-tossing game each player will be

on the winning side for about half the time, and that the lead will pass not infrequently from one player to the other" when in fact "it is quite likely that in a long coin-tossing game one of the players remains practically the whole time on the winning side, the other on the losing side." It turns out that the fraction of time that, say, heads is in the lead is least likely to be 1/2 and most likely to be the extremes of 0 or 1. The so-called arcsine laws for random walks can be investigated by computer simulations.

Combinatorics

Certain combinatoric relationships such as "n choose k" equals "n choose n – k" are intuitive because students readily agree, if not come up with the observation themselves, that choosing k of n people is equivalent to rejecting n – k people. The factorial formula for permutations (without replacement, has a very intuitive basis as described in section 2.3.3, but how fast n! grows is not so intuitive. Students consistently underestimate the number of batting order lineups for a 9 person baseball team (it's 9! = 362,880). Students are quite shocked to hear that, for example, there are more possible seating arrangements for a class of 30 students (and 30 desks) than there are grains of sand that could fit inside the Earth.

Probability

Many probability rules, such as the inclusion rule, "P (A) ≤ P (B) if A is a subset of B," are found extremely intuitive by students (see Bar-Hillel and Neter 1993). Others, such as P (A or B) = P (A) + P (B) – P (A and B) are found very intuitive once the Venn diagram is drawn. Students also find intuition in a geometric approach to solving probability problems such as the area model in the *Standards* (NCTM 1989, p. 111). Tree diagrams are also very intuitive for modeling a chronological sequence of actions and random events. The classical (equiprobable) view of probability that naïve-statistical students initially have, according to delMas and Bart (1989, p. 40), "may be regarded as a special case of a more general frequentist orientation of probability." Situations involving natural

generalizations such as this fit the Traditional Position quite well. Counterintuitive examples in this topic include the birthday problem and déMére's Paradox, as discussed in section 3.4. Additional examples by Fischbein were given in section 3.1.

Conditional Probability

Once again, a Venn diagram can be used to make formulas such as $P(A|B) = P(A \text{ and } B) / P(B)$ intuitive. Shaughnessy (1992, pp. 473–475) discusses prob–lems that illustrate the "Falk phenomenon" (which involves a conditioning event occurring after the event that it conditions), difficulties in selecting a conditioning event, and confusion between conditional and its inverse, which is related to the Classification Paradox discussed in section 3.5 and the Taxi Problem (Shaughnessy 1992). Finally, there is the Monty Hall problem (see section 3.2).

Independent versus Dependent Variables

Students can easily provide their own examples. For comparison purposes, it is often helpful to have examples of each grounded in the same real world situation (as was done with the "rain" example for independent and dependent events). To use dice (a fair red die and a fair green die) as this common context, an example of dependent variables is as follows:

W = "whether or not the sum of a red die and a green die equals 3"
and
Z = "the value of the red die."

A trivial example of independent variables would be

W = "the value of the green die"
and
Z = "the value of the red die."

What is not initially intuitive, however, is that two variables can appear to have "overlapping" sample spaces, and yet be statistically independent, such as

W = "whether or not the sum of a red die and a green die equals 7"
and
Z = "the value of the red die."

Also, dependent variables need not be correlated. While this fact may not have a great deal of practical use or significance for students in an introductory course, it is a fact they are capable of verifying, and may help them be less quick to assume that, for example, correlation implies causation in a particular situation. See, for example, Romano and Siegel (1986, p. 63, example 4.11), in which X and Y are the sum and difference, respectively, of two coin tosses where 1 = heads and 0 = tails. X and Y are uncorrelated since Cov $(X,Y) = E(XY) - E(X) E(Y) = 0 - (1)(0) = 0$. But X and Y are not independent, because $P(X = 0$ and $Y = 0) = 0 \neq (1/4)(1/2) P(X = 0) P(Y = 0)$.

Distributions, Limit Theorems

The intuitiveness of the general shapes of the most commonly used distributions can be brought out, such as the symmetry of the normal curve or the asymmetry of a binomial distribution with $p \neq .5$. Weinberg (1981, p. 280) provides intuition for the shape of the F distribution. Specific distributions, especially discrete distributions, often have a physical balls and urn model that can serve as a concrete representation. For example, the probabilities of drawing six balls without replacement from balls numbered 1 to 50 (as in Lotto Texas) can be represented with the hypergeometric distribution. Also, students consistently find it intuitive that the expected value of the binomial distribution is the product of its two parameters, n and p.

While it is intuitive that the sampling distribution of the sample mean has less variance than the original population, it is not at all intuitively obvious why it is so nearly normal for modest values of n, regardless of the shape of the original population. The typical proofs, of course, require calculus, and so an introductory course can best demonstrate the Central Limit Theorem through simulation, either on computer or with in-class low-tech techniques (e.g., Johnson 1986). The Law

of Large Numbers also has both intuitive and counterintuitive aspects (as mentioned in section 3.1), and can be used to confront the erroneous Law of Small Numbers (or Law of Averages, or Gambler's Fallacy) that students often have. Indeed, Moore and McCabe (1993) do this.

Estimation

The target analogy of Moore and McCabe (1993, pp .293–294) effectively illustrates the ideas of bias and variability in a point estimator: "Think of the parameter as the bull's-eye on a target and the sample estimate as an arrow shot at the target." This analogy is also applied to interval estimation in Gonick and Smith (1993, p. 116): "Consider an archer-pollster shooting at a target. Suppose that she hits the 10 cm radius bull's-eye 95% of the time. … Sitting behind the target is a brave detective, who can't see the bull's-eye. The archer shoots a single arrow. Knowing the archer's skill level, the detective draws a circle with 10cm radius around the arrow. He now has 95% confidence that his circle includes the center of the bull's-eye!"

From a class taught from the Traditional Position, it is not always clear that some simple or commonly used estimators may lack important desirable properties. For example, students can readily see that the sample range underestimates the population range. (In general, most maximum-likelihood estimators are also biased estimators.)

It is also not immediately obvious that some estimators with desired properties permit impossible results. There are numerous such examples when calculus is permitted, such as DeGroot's (1975, p. 353) example of X denoting the number of failures that occur before the first success is obtained (the trials are independent Bernoulli random variables with common unknown success probability p). X has a geometric distribution with pmf $f(x) = pq^x$, $q = 1 - p$. The unique unbiased estimator $T(x)$ of p from the observation x is 1 if $x = 0$ and 0 if $x \geq 1$. This can be obtained by setting $E(T(x)) = \sum T(x)pq^x$ equal to p, and equating powers of q. But "if the first success is obtained on the second trial, i.e., if $x = 1$, then it is silly to estimate that the probability of success p is 0." A

simpler example is the method of moments estimator for the maximum of a discrete uniform distribution. In other words, when sampling with replacement is done from a population numbered 1 to N, the method of moments estimator for N is easily seen to be 1 less than twice the sample mean. This estimator could be less than the largest observation we have seen, and we know N is at least as big as that observation!

Regression

In parallel with the development of measures of spread, a structured sequence of questions can help lead students through most of the "stages" often used to motivate simple, linear least-squares regression in a textbook (e.g., Wonnacott and Wonnacott 1977, chapter 11): fitting a line by "eye," fitting a line to minimize the sum of errors, fitting a line that minimizes the sum of the absolute errors, and finally, fitting a line that minimizes the sum of the squares of the errors. The regression line itself can be thought of a smoothed version of the graph of averages of Y given X, and its slope can be thought of as average increase of Y per unit increase of X, etc.

Students are often unaware and puzzled by what Freedman (1991, p. 160, emphasis in original) calls the regression effect: "In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test and the top group will on average fall back. This is the *regression effect*. The *regression fallacy* consists of thinking that the regression effect must be due to something important, not just the spread around the line." This can be demonstrated with in-class activities. For example, Levin (1993) describes a method of demonstrating this effect to an introductory statistics class using two standard decks of playing cards. Lock and Moore (1991) relate an even simpler version by Colin Sacks with index cards numbered 1 to the number of students. Karylowski (1985) gives a classroom demonstration using dice but grounded in a real-life context. The regression effect is given "numerical intuition" by Moses (1986, p. 328): "$(y - y)/s_y = r \, (x - x)/s_x$ … the extremeness of a given x (by *extremeness* we mean distance from x, measured in

standard deviation units) is not repeated in the extremeness of y (the estimated average of y at the given x); rather, a discount factor equal to r is first applied."

Correlation

In addition to the merry-go-round horses metaphor in section 2.3.1, there are such intuition builders as the square of correlation interpreted as the proportion of variation in one variable explained by the other variable, or the correlation interpreted as the slope of the regression involving the variables after each is standardized (transformed into z-scores). Students knowing trigonometry (see section 2.3.2) can be shown the correlation between X and Y as the cosine of the angle between two vectors whose lengths are proportional to the standard deviations of X and Y, respectively, such that the near diagonal of the completed parallelogram has length proportional to the standard deviation of X + Y.

Moses (1986, p. 319) lists an intuition-refining example that is usually omitted in a traditional course:

The strength of the correlation between x and y will be reduced in a subpopulation consisting only of individuals with x values lying above a certain cutoff value A, or only of those with x values lying below a certain cutoff value B, or only of those lying between two limits, A and B. In each instance, the variability in x is reduced … . If for some reason the subpopulation is chosen to consist of individuals with x values lying outside an interval A, B, the variability in x is thereby increased and the correlation coefficient will be larger.

Weisberg (1985, p. 76) adds "that even in the unusual event of analyzing data drawn from a multivariate normal population, if sampling of the population is not random, the interpretation of summary statistics such as R^2 may be completely misleading … ."

Most introductory statistics courses especially those preparing students for courses in the social sciences, stress like a mantra that "correlation does not imply causation." Unfortunately, the fact that this distinction is blurred by the media so frequently makes it difficult for students to keep in mind that correlation between X and Y could mean any of at least five interpretations, such as: X causes Y, Y causes X, X and Y cause each other, X and Y are caused by a third variable Z, or the correlation is completely spurious. Students can enjoy critiquing examples of

the latter such as "a high positive linear relationship between the birthrates in counties and the density of storks in the counties (Haack 1979, p. 264)" or between high tuberculosis death rates and living in Arizona. Cryer and Miller (1991, pp. 151–152) have a useful discussion of three criteria that would have to be established before "X causes Y" could be claimed: consistency across contexts, responsiveness of Y to X, and model or mechanism.

Also, zero correlation need not imply no relationship, as in the example of $Y = 6X – X^2$, or in the case of certain linear situations that have outliers (the correlation coefficient can be extremely sensitive to an outlier in small data sets). A more subtle example (perhaps subtle to the point of being non-intuitive rather than counterintuitive to most students) involves ecological correlations, which "tend to overstate the strength of an association," as Freedman et al. (1991, p. 141) discuss.

Hypothesis Testing

Students have already been exposed to the basic sequence of the scientific method, and can readily follow the structure of metaphors such as the courtroom or leaves from a tree trunk, as discussed in section 2.3.1. Also, students can steadily connect a confidence interval with a two-tailed hypothesis test, as many books do (e.g., Anderson, Sweeney, and Williams 1990). The implementation of testing, however, invokes serious questions.

As Freedman (1991, p. 510) makes the general point, "Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory. … The language of testing makes it easy to bypass the model, and talk about 'statistically significant' results." As a result, students often fail to realize that [ibid.] "[t]he test will not check to see whether this model [specified explicitly or implicitly by the investigator] is relevant or plausible. The test will not measure the size of a difference or its importance. And it will not identify the cause of the difference." Indeed, students are not always exposed to the fact that with a large enough sample, any difference (whether it has practical significance or not) can be statistically significant. A common example is birth rates between the genders.

With a too-often unquestioning use of a 95% confidence standard, students are even less likely to have been exposed to the idea that a practically significant difference may not always result in a statistically significant result. Students can benefit from a classroom ESP-detection experiment (Lock and Moore 1991, p. 3) in which each student tries to guess suits of 25 cards and is then forced to interpret the few "significant results" that will likely occur in a class of 30 students. In general, Traditional instructors may play these implications down not to draw attention to the fact that statistical and real-world views can be different. This, however, is a good opportunity to stress the need for good experimental design so that a practically significant result *will* also have statistical significance. The necessity of understanding a precise meaning of a word ("significance") in a statistical setting has already been encountered in a course with words such as *independence* and *population*.

The Classification Paradox of section 3.5 was referred to in the topic of conditional probability, but certainly can also be discussed in a setting of hypothesis testing. Also, the frequentist interpretation of p-value = P (data | H_0 is true) is not equal to P (H_0 true), as students often would like to assume. While it is beyond the scope of the present study to pursue this further, the reader can find relevant discussion in Albert (1994).

Experimental Design

Given a simple real-world experiment to construct and conduct, students on their own will usually express the need for a plan that invokes the concepts of (in ascending order of intuitiveness) blocking, randomization and replication, which are widely considered the three most important principles of experimental design. Such an experiment can be simulated by real-world advertisements, such as some variation of the "Coke vs. Pepsi" experiment described by Garfield (1993). The use of this experiment and the previous use of two others were described by Solomon (1979). A useful metaphor for experimental design is found in audio communication in that the goal is to get a specified quantity of information (precision) at lower cost by increasing signal-to-noise ratio. Nature sends both a "signal" (information) and "static" (uncontrolled background noise).

Other authors encouraged students to identify experimental design concepts in situations such as the sport of basketball (Polyson and Blick 1985).

There do not seem to be any counterintuitive concepts yet identified for this topic, partially because the basic concepts are in fact fairly intuitive, and also because many of the details are sufficiently complex as to be non-intuitive to a novice. The main exception is interaction effects, which are often omitted from an introductory course although they can be discussed with a minimum of technical discussion. Schaefer (1976, p. 103) goes so far as to say that interaction is "vital as a counter to some very general self-inhibiting human behaviors. Most of us … 'think' in very few dimensions (sometimes only one), rather than the many that are usually necessary to give a fairly adequate account of an event."

Analysis of Variance (ANOVA)

Numerical intuition for the formulas can be given, as described in section 2.3.3. Further intuition can be cultivated by demonstrating its equivalence to regression in which the independent variables are dummy (indicator) variables. Also, Dillbeck (1983, p. 21) describes a very useful activity for single-factor ANOVA in which "the questions led students sequentially through a process of reasoning which arrived at the goal of the lesson and which, at each step, allowed the students to experience an 'aha' of figuring out a point." This latter type of intuition building is more important than the numerical intuition because, as Rubin (1994) states, "teaching concepts in anything is more difficult than teaching computations, and the computations do not really help."

There do not seem to be any counterintuitive concepts yet identified for this topic, partially because the basic concepts are in fact fairly intuitive, and other details are sufficiently complex as to be non-intuitive to a novice.

Sampling

Dietz (1993, p. 104) describes an activity she has used in which "students have 'invented' simple random sampling, systematic sampling, stratified sampling, and various combinations thereof." Dietz describes the activity as follows: "Students are presented with a dataset consisting of gender, SAT verbal

score, SAT mathematics score, and high school grade point average for 317 freshmen from North Carolina State University. The students, who have not yet studied sampling, work in groups of three or four to generate three possible methods for selecting a representative sample of 20 freshmen from the population of 317." A variation of this exercise could be constructed to motivate additional sampling schemes such as cluster sampling or sequential sampling. The "random rectangles" activity (Bentley et al. 1993) gives studetns a chance to compare their judgmental samples with random samples in estimating rectangles' areas. The fact that students' judgmental samples consistently overestimated the truth is valuable evidence for the value of random sampling.

The initial counterintuitiveness of required sample size was discussed in section 3.6. A further blow to the overemphasis on sample size occurs in Lapin (1987, pp. 68–71), who gives six reasons why a census can be less desirable than a sample: economy, timeliness, population size, inaccessibillity, accuracy, destructive observations. A "less can be more" connection can be made to regression in that a model may not always be improved by adding more predictor variables, especially if they are highly correlated with variables already in the model.

Another excellent counterintuitive example to use with sampling is the Inspection Paradox because it is a chance to revisit and generalize the "average class size" example encountered in the earlier topic of central tendency measures. It is also a chance to explore a deeper type of bias that can result from sampling from a fixed point, a situation that students have likely already considered at a surface level, such as bias from an opinion poll about abortion if one surveys people passing by a point near a Roman Catholic church. Stein and Dattero

(1985, p. 96) provide two concrete examples of sampling bias that illustrate the paradox:

When faced with the problem of estimating the average size of a family, students frequently suggest polling their classmates or standing on a corner and asking people who walk past. In using these methods, bias may be introduced since we are sampling people and not families. Or ask the students how to measure the

average speed of all the cars on a freeway. Measuring the speed of cars passing a fixed point will give biased results.

Because larger families and faster cars will be sampled more frequently in these situations, there will be positive bias in estimating the means.

The families example is further grounded with connections to the greater likelihood of picking an American that's from Texas than from Idaho, or picking a random point that lies on a long piece of yarn when many long and short pieces are lined up end to end. Smith and Gonick (1993) explore a slight variation on the car scenario—with a moving observer.

Finally, the related Inspector's Paradox (Reinhardt 1981) is a useful demonstration of the importance of random sampling as well as a more subtle example than students are usually given of how nonrandomness can be introduced. This paradox involves looking at the length of a string of coin tosses ending in heads, and comparing this length to the theoretical average of two tosses for a fair coin. One method (A) selects the string of tosses that includes the tenth toss and a second method (B) selects the string that follows the A string. "Roughly speaking, B's procedure is random because the coin has no memory, but since long strings have a greater opportunity to include the tenth toss [or any other specific toss, for that matter], than short ones, A's is not" (p. 106). Reinhardt concludes by relating an illustration that is easily carried out in a classroom (p. 107):

Ask the students in the class to give the number of boys and girls in their families. Tabulate the results separately for the boys and girls in the class and calculate the average number of boys per family and the average number of girls per family. … The claim is that families tend to 'run' to children of one sex as the data superficially indicate. The difficulty is that the families of the boys in the class are not a random sample of families, nor are the families of the girls.

CHAPTER 5

CONNECTIONS TO TEACHING AND LEARNING

5.1 Small-Group Cooperative Learning

Many of the activities detailed in the "intuitive" sections of the syllabus-driven model are well suited for small-group (i.e., 2–5 students per group) cooperative learning, a mode of learning that has been well documented in mathematics education, but only somewhat in statistics education specifically. Garfield (1993) gives the most complete overview of the latter, including definitions, motivations, guidelines, strategies, and resources. Although there have been hundreds of "research studies documenting the effectiveness of cooperative learning activities in classrooms," Garfield (1993) names only three studies that have examined the use of cooperative learning in college statistics courses. One is Dietz (1993), which was discussed in section 4.7, at the end of the SPICE. Also, "Shaughnessy (1977) found that the use of small groups appeared to help students overcome some misconceptions about probability and enhance student learning of statistics concepts. … Jones (1991) introduced cooperative learning activities in several sections of a statistics course and observed dramatic increase in attendance, class participation, office visits, and student attitudes." Studies not mentioned by Garfield include Borresen (1990), Cumming (1977, 1983, 1984), and Reynolds et al. (1991).

Under the heading "concerns about using small groups", Garfield (1993) gives instructors a caution that could be applied to counterintuitive examples in general: "Instructors may be discouraged by students who resist an activity that appears challenging and difficult, forces them to think, and does not allow them to be passive learners, because students are used to sitting in lectures where they are not required to talk, solve problems, or struggle with learning new material. Students may want the teacher to do more explaining, and telling them the right answers, rather than struggle with a problem themselves." The good news is that

the strategies and references Garfield then offers to overcome this may help an instructor using the SPICE.

A final point is that cooperative learning is arguably as appropriate for statistics education as for any other field in that real-life statistics is usually a multidisciplinary endeavor involving a team of programmers, consultants, high- and low-level statisticians, managers, subject-matter experts, etc.

5.2 Structured Controversies

The SPICE of section 4.7 offers a number of counterintuitive examples which may cause controversy, at least initially. The controversy is occurring as cognitive dissonance within each person, and is also being played out between students in the classroom as well. Structured controversies are a particular form of group activity, and may be thought of as either a special case of cooperative learning or a different type of group learning, depending on one's definitions. Either way, much of the logistics involved with cooperative learning groups apply to structured controversies.

Derry et al. (pp. 6–7, in press) offer a connection between statistics, controversy and learning opportunities:

[S]tatistics as a subject matter represents controversial knowledge (e.g., Nicholls & Nelson 1992), an idea that should influence how statistics is taught. For if statistics is controversial subject matter, then it should not be taught as a set of final-form, universally accepted concepts that can be handed down by authority and conveyed to students by professors and textbooks. At a minimum, statistics courses might inform students about statistical debates. We believe that students could gain even more by actually discovering and participating in such controversies. Participating in statistical controversy is what scientists do when they conduct research, for their work involves selecting statistical tools that are perceived to be appropriate for the problem at hand, using them as a basis for conceptually analyzing that problem, and, often, defending their choices to other scientists.

Indeed, statistics *is* controversy, ranging from specific conflicting polls and studies published by the media to current debates among statisticians on paradigms of inference (Bayesian, frequentist, etc.). As Moore (1992, p. 16) states, "What is

most striking is that the position that statisticians take on foundational issues has an immediate impact on their statistical practice. … Even elementary text cannot avoid taking sides, even if only implicitly, on questions such as the role of prior information in inference and whether statistical tests are decision procedures or merely assess the weight of evidence." Any method of instruction that exposes

students to this process of real-life statistics supports the spirit of current mathematics education reform.

There are a variety of instructional styles which are oriented towards utilizing the conflict between normative conceptions and students preconceptions, (i.e., students' primary and secondary intuitions). Borasi (1994, p. 171) states: "Research on conceptual change in science learning (e.g., Brown & Clement, 1989; Hewson, 1981; Strike & Posner, in press) has also suggested that students' specific misconceptions could be used as a way to generate conflicts that, in turn, may expose and challenge the students' limited theories about how the world operates. A similar principle informs 'conflict teaching,' a strategy developed for mathematics instruction by Bell and his colleagues (Bell, 1983, 1986; Swan, 1983)."

In an article discussing a related idea called academic controversy, Johnson and Johnson (1993, p. 43) use language that seems very consistent with that of the NCTM (1989):

If students are to become citizens capable of making reasoned judgments about the complex problems facing society, they must learn to use the higher-level reasoning and critical thinking processes involved in effective problem solving, especially problems for which different viewpoints can be plausibly developed. To do so, students must enter empathically into the arguments of both sides of the issue, ensure that the strongest possible case is made for each side, and arrive at a synthesis based on rational, probabilistic thought.

Johnson and Johnson (1992, p. 125) define controversy as "when one student's ideas, information, conclusions, theories, and opinions are incompatible with those of another, and the two seek to reach an agreement." Without describing the methodology (except for mentioning that one study was a meta-analysis), they present the results of their studies from 1979 through the present to discover the effects of structured controversy. They found that "compared with concurrence-seeking, debate, and individualistic efforts", controversy tends to yield: greater mastery and retention of subject matter, greater ability to generalize, higher quality decisions and solutions to complex problems, more frequent creative insights, and various other benefits. Johnson and Johnson (1993, p. 44)

cite research by Ames and Murray which concludes that "conflict qua conflict is not only cognitively motivating but that the resolution of the conflict is likely to be in the direction of correct performance." Classroom prerequisites are given in Johnson and Johnson (1988).

Roth (1990, p. 162) found conceptual conflict could be created in many ways, including laboratory demonstrations, experiments and student writing. In addition to this flexibility of instruction, Johnson, Johnson, and Smith (1991, 7:15) described these specific benefits from "Structured Controversies in Science":

To ensure that higher-level reasoning, critical thinking and metacognition take place, however, students need the intellectual challenge resulting from conflict among ideas and conclusions.. Controversy, compared with concurrence seeking, debate and individualistic efforts, results in higher achievement, higher-quality decisions and problem-solving, more creative thinking, more higher-level reasoning and critical thinking, greater perspective-taking accuracy, greater task involvement, more positive relationships among group members, and higher academic self-esteem. … Students make an initial judgment, present their conclusions to other group members, are challenged with opposing views, become uncertain about the correctness of their views, actively search for new information and understanding, incorporate others' perspectives and reasoning into their thinking, and reach a new set of conclusions. While this process sometimes occurs naturally within cooperative learning groups, it may be considerably enhanced when teachers structure academic controversies.

As delMas and Bart (1989, pp. 42–43) add:

Ross and Anderson argue that the incorporation of new experiences into current beliefs is usually a biased assimilation; confirmatory experiences are anticipated and readily accepted, while contradictory experiences are seen as exceptions and/or treated with skepticism. .In order to overcome these factors, Ross and Anderson suggest that effective discrediting experiences are those which require subjects to act upon their beliefs [this is consistent with the protocol of Shaughnessy 1977] and increase the dissonance between their expectations and the actual outcomes. … There are several reasons as to why this approach may be effective. First of all, when subjects encounter a situation which they believe can be assimilated into their existing schemas but which are not resolved when the schemas are applied, the contradictions prepare the subjects to restructure or accommodate their schemas. The argument is that a subject is more likely to accept and learn a new strategy if it resolves the contradiction. Second a contradictory situation helps to highlight conflicts between a subject's present strategy and the correct strategy. Such conflicts provide contrasts between the

two frameworks which can aid better recall of the new strategy. Finally, the contradictory situation helps a subject focus on key relationships among variables and to disregard the variables which may lead to misjudgments or misinterpretations of causal effects.

5.3 Constructivism

Using counterintuitive examples is consistent with a constructivist perspective. Constructivism has a strong relationship to primary and secondary intuitions (discussed in section 2.2) in that the instructor must recognize that the student will be actively involved in the construction of any secondary intuition that would replace the primary intuition he entered the classroom already having. Also, the connection between constructivism and the Gestalt organization of knowledge supported by Fischbein's origins classification of intuitions is discussed at the end of section 2.2. Constructivism, a theory and perspective in which learners actively construct their own knowledge, has been a major trend in mathematics education and has made an impact on statistics education as well, despite the lack of research that explicitly addresses this. Many statistics curriculum materials are constructivist in nature, such as the ELASTIC software of the Reasoning Under Uncertainty curriculum (Rosebery and Rubin 1988, p. 207), which employs "interactivity, visualization, and dynamic links to create a laboratory in which students can explore the underlying meaning of basic statistical concepts and processes." According to Stern-Dunyak (1993, p. 13): "The 'new' teaching, as described by [David S.] Moore, includes: learning is considered a construction of knowledge; teachers are guides and motivators; students work in groups on open-ended problems; students discuss problems, get feedback; teachers cover less material, but students learn more."

Such motivation was present in early experiments such as Shaughnessy (1977, p. 299, emphasis in original):

A small-group, problem-solving and model-building approach was undertaken in the experimental groups … perhaps the transition for students from preconceptions and misconceptions of probability to mathematizations of probabilistic laws can be facilitated if students are encouraged to experience

elementary probability and statistics as a *process* of describing observed experimental phenomena more and more accurately, rather than as a *system* of rules, axioms, and counting techniques that must be learned and applied to problems.

The bottom line is (Shaughnessy 1992, p. 472, emphasis in original): "Our students are not *tabulae rasae*, waiting for the normative theory of probability to descend from our lips. Students already have their own built-in heuristics, biases and beliefs about probability and statistics." This agrees with Mevarech (1983, p. 420): "Evidently, erroneous schemata are so deeply ingrained in a student's knowledge base that simply being exposed to another statistics course is not sufficient to overcome these errors."

Some educators (e.g., Melvin and Huff 1992) confront these primary intuitions by supplementing textbooks with lists of common errors. Some mathematics textbooks (e.g., Kolman et al. 1993) already incorporate warnings of standard errors (with explanations and examples explaining why they are wrong) into the text presentation. Some statistics textbooks (e.g., Moore and McCabe 1993, p. 327) incorporate this idea with certain topics, such as acknowledging both the hot hand theory and the law of averages in discussing the law of large numbers. Of course, having students read about common errors, while helpful, cannot be expected to be as effective as having students directly confront errors they have actually just made themselves.

Given how resistant erroneous schemata can be and given the limited success of the Traditional Position in overcoming many of these, it seems that a bolder method such as the SPICE may well be one of the most effective or efficient ways to provide the active process of feedback and correction needed to confront these erroneous schemata. Confrey (1990) goes so far as to say that a constructivist teacher is concerned more with teaching students how to develop their cognition than about mathematical structures. Certainly counterintuitive examples are very effective in developing cognition (and metacognition) for the reasons discussed in section 4.3.

An important question to consider is to what extent constructivism in statistics education is or should be expected to be any different than in

mathematics or science education as a whole. Most of the conceptual change and bridging analogies literature, for example, has been in science education. Konold (1993) does not expect "that the constructivist approach would play out any differently in application to statistics education than it has in mathematics education, and that what it is mostly about revolves around the importance of considering the nature and development of student thinking in the design of educational materials and approaches."

Nevertheless, it does seem natural that the extent to which mathematics and statistics have fundamental differences in foundations (e.g., Moore 1993) would be reflected in the constructivist educational approaches. Statistics involves constructing models, graphs and data collection instruments, and the fact that there is usually more than one defensible approach seems in the spirit of individual students each constructing his or her own knowledge. In fact, the fact that the SPICE points out the pitfalls of media graphs that have been misconstructed further emphasizes the dynamic, human process that statistics is.

The key point here is that constructivist philosophy underlies both cooperative learning and structured controversies in that students play an active role in constructing their own knowledge, are engaged in confronting their conceptions, and are involved in reflective thinking. Garfield (1993) suggests this connection when she states, "Small-group learning activities may be designed to encourage students to construct knowledge as they learn new material, transforming the classroom into a community of learners, actively working together to understand statistics. The role of the teacher changes accordingly from that of 'source of information' to 'facilitator of learning.' "

Unfortunately, it appears that while many researchers and reformers are embracing on the surface a perspective of constructivist cognition over a cookbook approach, few are fully implementing it, and fewer still have given serious attention to the natural role of counterintuitive examples within this perspective. The situation for mathematics and statistics instruction seems similar to the observation of Stofflett and Stoddart (1994, pp. 32–33):

The traditional didactic instruction that occurs in the majority of science content courses rarely challenges or improves student preconceptions about science content (Wittrock 1986). … This traditional approach to science teaching persists despite research on scientific cognition that has identified the process of conceptual change as a necessary prerequisite for the formation of scientifically validated theories (Hewson & Hewson 1988; Posner, Strike, Hewson, & Gertzog, 1982). Consequently, even those students who have passed college science content courses retain misconceptions about the content they were taught (Champagne, Gunstone & Klopfer 1985; McClosky 1983). This problem is particularly significant for elementary education majors who take only a few science content courses and then are expected to be able to teach a broad range of science content to children.

Teachers need time and training to learn how to (National Science Foundation 1992, p. 38) "probe for a misconception, ask questions to clarify students' beliefs, suggest events that contradict students' flawed beliefs, encourage non-demeaning discussion, guide students toward constructing valid scientific concepts, and finally reevaluate students' comprehension."

In summary, constructivism and counterintuitive examples share many things, including: acknowledgment of students' beliefs prior to instruction; active engagement of those prior beliefs in a way that leads to deeper understanding and empowerment; the role of metacognition; the role of teacher as a facilitator more than as a source of static knowledge, etc.; the amenability to exploration, group work, critical thinking; and difficulty in use by teachers who were taught in the traditional manner.

CHAPTER 6

CONCLUSIONS

6.1 Summary

The power of paradox has been explored throughout this study. Authors such as Eves (1990) and Rapaport (1967) discuss the important role counterintuitive results have had throughout the history of mathematics in the actual growth of the discipline. This study has further argued that counterintuitive examples, used appropriately, can be a powerful facilitating force in student learning. Some counterintuitive examples temporarily descend into chaos before then resolving to a deeper sense of order. Perhaps developing a spirit of the intuition of the counterintuitive (to modify Fischbein's phrase "intuition of the non-intuitive"), can help teachers and learners of statistics become more comfortable with the flow between order and chaos, as Davis and Hersh (1981, pp. 172–179) illustrate.

Definitions of intuition, intuitiveness, non-intuitiveness, and counter-intuitiveness were stated and explored. Connections in both directions were made between statistics and learning theory. This study has provided teachers of statistics with a comprehensive survey of the role intuition can play in their classroom, including conceptual, geometric, and numerical types of illumination. A surplus of terminology and difficulties in measurement were addressed as problematic issues. While there is no unification in sight to the former, given the multidisciplinary nature of the field, there are hopeful results for the latter that have been operationalized in related fields.

Four criteria for counterintuitive examples were set forth and several of the most important or popular examples were discussed, including Simpson's Paradox, average class size, deMéré's paradox, the birthday problem, the Classification Paradox and the required sample size for estimating a proportion.

The Traditional Position and Alternative Positions were discussed and critiqued, before being compared and synthesized in a new model named SPICE. After addressing model building concerns, the SPICE was presented around the structure of a syllabus, with counterintuitive examples matched to virtually every

major topic on an introductory college statistics course syllabus. While many courses now being taught currently use material that correspond to parts of the SPICE, no course is using all of it, and few courses using any of it are using it with the overall context of purpose and intention discussed in chapter 4.

Finally, chapter 5 presented connections and additional discussion concerning group work and constructivist learning, both of which naturally complement and support the SPICE.

6.2 Problematic Issues

Assessment instruments in statistics are still sorely lacking. This is an obstacle to assessing the results of a lesson or course that may include counterintuitive examples, as standard assessment instruments rarely take into account higher-order thinking or allow subjects to give reasons for their answers. It is quite possible to give the same answer (correct or incorrect) using completely incompatible heuristics. [e.g., Shaughnessy, 1992, p. 479: "This is reminiscent of the gambler's fallacy, though the children may … have been using Konold's outcome approach."] Garfield (1991) reviews five categories of research "relevant to the assessment of statistical understanding. These are: 1) students' attitudes and anxiety towards learning statistics, 2) students' computational skills in using probability and statistics, 3) students' misconceptions of probability and statistics, 4) conceptual frameworks for assessing statistical learning, and 5) methods of assessing mathematical learning and problem solving." The measurements of dissonance, deep-seatedness and difficulty from psychology and science education may have more to contribute to our attempts to assess intuitiveness of content or intuition of students (see section 2.4.).

The dynamics of intuition and intervention may be different at different levels (see, for example, Shaughnessy 1992, p. 485: non-statistical, naïve-statistical, emergent-statistical, pragmatic-statistical) of statistical maturity. This study limited its focus to introductory college courses, which would have only a part of this spectrum (certainly there should be no pragmatic-statistical students in such courses!), so this relationship should be investigated further. What is required to effect a transition from the first to the second level may be

qualitatively different from what is required to effect a transition from the second level to the third, for example.

While teacher support will be needed to use these new curricula and pedagogical techniques, it should become less and less necessary as teachers gradually incorporate group work into their instruction in general, and as teachers become more aware of their own misconceptions and how they overcame them. As mentioned in section 5.1, Garfield (1993) lists numerous resources for teachers wanting to become more confident and competent in the use of cooperative learning activities. Structured controversies in particular may be less familiar to teachers than cooperative learning groups in general. Also, if structured controversies are used with counterintuitive examples, they will require more care in setup, rather than more intuitive group work in which the instructor can spend more time waiting for students to discover the knowledge.

A specific kind of resistance that may be encountered from some instructors concerns the counterintuitive examples for which the power of statistics to deceive is explored. For example, section 4.1 cited Burrill's call not to emphasize this power. Care must be taken to ensure that such examples are explored with the utmost attention to ethics and responsibility as well as empowerment and self-defense.

Teachers' resistance to counterintuitive examples can be addressed by showing them supporting passages from the NCTM (1991, pp. 128, 134):

[Mathematics and mathematics education] should be designed deliberately to help teachers rethink their conceptions of what mathematics is, what a mathematics class is like, and how mathematics is learned. … Teachers need opportunities to revisit school mathematics topics in ways that will allow them to develop deeper understandings of the subtle ideas and relationships that are involved between and among concepts.

A more general dilemma that remains (Stofflett and Stoddart, p. 45) is

the expectation that teachers can learn to be constructivist teachers when they have not been constructivist learners. The fundamental assumption of constructivism is that learners construct understanding through personal experience. Teachers' own experiences as learners powerfully influence their instructional

beliefs and practice (Lortie, 1975). It should come as no surprise that teachers who learned science through didactic methods teach science didactically.

This in turn brings to mind a point by Schoenfeld (1987, p. 27) that some misconceptions arise from the general classroom experience with formal mathematics. Perhaps the greatest challenge is to train people to teach in a way that at first will feel unfamiliar if not outright counterintuitive!

6.3 Additional Directions for Research

The considerable body of literature on misconceptions dominates most of the literature on the teaching and learning of statistics (see Shaughnessy 1992), and should remain a major area of interest for some time. There is room for further connections to be made between counterintuitive examples and misconceptions. These connections can be on the general level, such as the idea that counterintuitive examples correspond to a subset of misconceptions that students might have. These connections can also be specific, such as the average-the-averages misconception and Simpson's Paradox, as discussed in section 3.3. Also, research on utilizing student errors (Borasi 1994) needs to be further examined for connections to this study.

A controlled experiment involving the effect of structured controversies with counterintuitive examples could be very helpful. In covering the topic "relations in categorical data" (e.g., Moore and McCabe 1993, section 2.5), one group could work through a real-life Simpson's Paradox situation, while a control group works routine two-way table problems from the end of the chapter. Then both groups take the same one-problem post-test involving making a correct interpretation from a three-way table of data. For an example for the topic "conditional probability," one group would work through a real-life situation with the counterintuitive result of the "Classification Paradox," while the control group would work a routine "tree" problem such as p. 358, exercise 4.81 in Moore and McCabe (1993).

Individual differences were not examined in this study (see section 4.6), but it may be valuable to do so. Also, there may be group differences relevant to this study. For example, Fidelman (1992) relates paradox resolution to

hemispheric dominance, which he in turn discusses in the context of gender differences.

It needs to be investigated how to present statistics so that students are not turned off by what they might perceive to be a needlessly cautious or legalistic flavor of many of the statements. For example, students should not merely be told the assumptions of a hypothesis test without also being given a feeling for how robust they are (e.g., the F test for variances is much more sensitive to nonnormality than is the t-test for means). Also, when covering the "A does not imply B" statements in the SPICE, students and instructors should be aware of how common and also how potentially important the counterexamples truly are. If students suspect that something is virtually always true, they may just consider it true, even though they may dutifully recite mantras such as "correlation does not imply causation" on exams. If instructors truly want to impart a cautiousness in statistical reasoning, they must choose their points selectively and have concrete examples (perhaps counterintuitive examples) ready to back them up.

Just as not all outliers are equally influential, it is assumed that not all counterintuitive examples will be equally effective in achieving the benefits and goals of instruction listed in section 4.3. It is hoped that future research will help determine which counterintuitive examples should be given highest priority for inclusion in the introductory course.

Qualitative research methods, which have only recently begun to play more than a limited role in mathematics education (see Eisenhart 1988), could play an invaluable role in providing a deeper understanding of the process activated by a group of students grappling with a counterintuitive example in a real-life context. As Garfield and Ahlgren (1988, p. 55) point out, "It is possible that a great deal of research, in focusing on the correctness of *answers*, has missed the subjects' perception of what the question was—and so misestimated the subjects' reasoning."

Surveys should be conducted to determine what examples students consider counterintuitive, and these results should be compared to teachers' judgments and to a priori considerations (see section 3.1). Also, teachers should be surveyed to determine what examples they most regularly use and why. It is

also recommended that an analysis be conducted of the similarities among, and differences between, this study and styles of teaching (e.g., structured controversy, academic controversy, conflict teaching, errors as a springboard for inquiry) mentioned in this study. In general, there is room for more synthesis of the strands of relevant research by educators of psychology, statistics, mathematics, and science.

Although most traffic between statistics and mathematics has been in one direction—from mathematics (e.g., Moore 1993), and although mathematics and science education have had a huge head start on statistics education as organized fields, this research could be among the first instances of statistics education contributing to "other branches of mathematics." While the SPICE is a strong model for the assumptions made and issues addressed, it is not claimed to be the only possible way of synthesizing the Traditional and Alternative Positions. It is fully expected that there will be considerable scholarly discussion concerning what features of the model to incorporate into the curriculum.

BIBLIOGRAPHY

Albert, J. (1994). Teaching statistical inference using Bayes, unpublished manuscript.

Allen, J. L., Walker, L. D., Schroeder, D. A., & Johnson, D. E. (1987). Attributions and attribution-behavior relations: The effect of level of cognitive development. Journal of Personality and Social Psychology, 52, 1099–1109.

American Statistical Association (1994). G. Burrill (Ed.), <u>Teaching Statistics: Guidelines for Elementary to High School</u>. Palo Alto, CA: Dale Seymour.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (1990). <u>Statistics for Business and Economics</u> (4th ed.). St. Paul, MN: West.

Ash, C. (1993). <u>The Probability Tutoring Book: An Intuitive Course for Engineers and Scientists (and Everyone Else!)</u>. Piscataway, NJ: IEEE Press.

Ball, D. L. (1988). Unlearning to teach mathematics. Issue Paper 88-1. National Center for Research on Teacher Education. ERIC Document Reproduction Service No. ED 302 382.

Barbeau, E. (1993). Fallacies, flaws and flimflam. <u>College Mathematics Journal</u>, <u>24</u>(2), 149–154.

Barbin, E. (1994). The meanings of mathematical proof: On relations between history and mathematical education. In J. M. Anthony (ed.), <u>In Eves' Circles</u> (pp. 41–42). Washington, DC: Mathematical Association of America.

Bar-Hillel, M. & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. <u>Journal of Personality and Social Psychology</u>, <u>65</u>(6), 1119–1131.

Beins, B. (1985). Teaching the relevance of statistics through consumer-oriented research. <u>Teaching of Psychology</u>, <u>12</u>(3), 168-169.

Bell, A. (1986). Diagnostic teaching: Two developing conflict-discussion lessons. <u>Mathematics Teaching</u>. No. 116, 25–29.

Bentley, D., Lock, R., Moore, T., Parker, M., & Witmer, J. (1993, January). <u>Teaching the Introductory Statistics Course</u>. Minicourse presented at the annual meeting of the Mathematical Association of America.

Berenson, M.L. & Levine, D.M. (1989). <u>Basic Business Statistics</u> (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Berenson, S.B., Friel, S., & Bright, G. (1993, April). Elementary teachers' fixations on graphical features to interpret statistical data. Paper presented at the annual meeting of the American Educational Research Association.

Bergman, D. A., & Pantell, R. H. (1986). The impact of reading a clinical study on treatment decisions of physicians and residents. Journal of Medical Education, 61, 380–386.

Berresford, G. (1980). The uniformity assumption in the birthday problem. Mathematics Magazine, 53(5), 286–288.

Bickel, P.J., Hammel, E.A., & O'Connell, J.W. (1975). Sex bias in graduate admissions: Data from Berkeley. Science, 187, 398–404.

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. Journal for Research in Mathematics Education, 25(2), 166–208.

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

Brewer, J.K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? Journal of Educational Statistics, 10(3), 252–268.

Brown, D.E. (1993). Refocusing core intuitions: A concretizing rule for analogy in conceptual change. Journal of Research in Science Teaching, 30(10), 1273–1290.

Brown, D.E. (1992). Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. Journal for Research in Science Teaching, 29(1), 17–34.

Burrill, G. (1994, March 16). Personal correspondence.

Burrill, G. (1990). Implementing the Standards: Statistics and probability. Mathematics Teacher, 83(2), 113–118.

Carkenord, D. M. & Bullington, J. (1993). Bringing cognitive dissonance to the classroom. Teaching of Psychology, 20(1), 41–43.

Chesterton, G.K. (1959). Orthodoxy. Garden City, NY: Doubleday Image.

Chu, D. & Chu, J. (1992). A "simple" probability problem. Mathematics Teacher, 85(3), 191–195.

Cleary, R. J. (1992). Mathematics education issues in the teaching of statistics,

unpublished manuscript.

Clement, J., Brown, D.E. & Zietsman, A. (1989). Not all preconceptions are misconceptions: finding "anchoring concepts" for grounding instruction on students' intuitions. International Journal of Science Education, 11(5), 554–565.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. Journal of Research in Science Teaching, 30(10), 1241–1257.

Cochran, W. G. (1977). Sampling Techniques (3rd ed.). New York: Wiley.

Cohen, J. (1994, May 9). She hopes for a lotto legal luck. National Law Journal.

Cohen, J. E. (1986). An uncertainty principle in demography and the unisex issue. American Statistician, 40(1), 32–39.

Confrey, J. (1990). What constructivism implies for teaching. In R.B. Davis, C. A. Maher, & N. Noddings (Eds.), Constructivist Views on the Teaching and Learning of Mathematics (pp. 107–122). Reston, VA: National Council of Teachers of Mathematics.

Cryer, J. D. & Miller, R. B. (1991). Statistics for Business: Data Analysis and Modelling. Boston: PWS-KENT.

Cumming, G. (1977). A group approach to the introductory statistics course in psychology. Australian Psychologist, 12, 293–302.

Cumming, G. (1983). The introductory statistics course: Mixed student groups preferred to streamed. Teaching of Psychology, 10, 34–37.

Cumming, G. (1984). Self-selection of student groups in a group Keller course. Psychology, 11(3), 181–182.

Davis, P.J. & Hersh, R. (1981). The Mathematical Experience. Boston: Houghton Mifflin.

DeGroot, M. H. (1975). Probability and Statistics. Reading, MA: Addison-Wesley.

DelMas, R. C. & Bart, W. M. (1989). The role of an evaluation exercise in the

resolution of misconceptions of probability. Focus on Learning Problems in Mathematics, 11(3), 39–54.

Derry, S., Levin, J.R. & Schauble, L. (in press). Stimulating statistical thinking through situated simulations. Teaching of Psychology.

Devore, J. & Peck, R. (1990). Introductory Statistics. St. Paul, MN: West.

Dewey, J. (1933). How we think. Boston: D.C. Heath.

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

Dillbeck, M.C. (1983). Teaching statistics in terms of the knower. Teaching of Psychology, 10, 18–20.

Duit, R. (1991). On the role of analogies and metaphors in learning science. Science Education, 75(6), 649–672.

Durrett, R. (1989). Counterexamples in probability [book review]. American Scientist, 77(4), 405.

Dweck, C. S. & Leggett, E.L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, 95(2), 256–273.

Eisenhart, M. A. (1988). The ethnographic research tradition and mathematics education research. Journal for Research in Mathematics Education, 19(2), 99–114.

Elliot, D. (1993, August 28). Professor seeks to even odds of lottery. Austin American-Statesman, B1, B3.

Evans, G. F. (1986). Getting through statistics with the help of metaphors. Journal of Education for Business, 62(1), 28–30.

Eves, H. (1990). An Introduction to the History of Mathematics (6th ed.). Philadelphia: Saunders College Publishing.

Ewing, A. (1941). Reason and intuition. Proceedings of the British Academy, 27, pp. 67–107.

Falk, R. & Bar-Hillel, M. (1980). Magic possibilities of the weighted average. Mathematics Magazine, 53(2), 106–107.

Falk, R. & Konold, C. (1992). The psychology of learning probability. In F. Gordon & S. Gordon (Eds.), Statistics for the Twenty-First Century (pp. 151–164). Washington, DC: Mathematical Association of America.

Falletta, N. (1990). The Paradoxicon. New York: Wiley.

Feffer, M. (1988). Radical Constructionism. New York: New York University.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications: Vol. 1 (3rd ed.). New York: Wiley.

Festinger, L. (1957). A Theory of Cognitive Dissonance. Stanford, CA: Stanford University Press.

Festinger, L. & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology, 58, 203–210.

Fidelman, U. (1992). The brain, feminism and mathematics. Journal of Structural Learning, 11(2), 155–166.

Field, C. (1984). A "reverse-order" elementary statistics course. American Statistician, 38(2), 117–119.

Fischbein, E. (1975). The Intuitive Sources of Probabilistic Thinking in Children. Dordrecht, Holland: D. Reidel.

Fischbein, E. (1987). Intuition in Science and Mathematics. Dordrecht, Holland: D. Reidel.

Foster, D. & George, E. (1994, April 26). Games Bayesians play. Colloquium talk at University of Texas at Austin, Department of MSIS.

Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). Statistics (2nd ed.). New York: W. W. Norton.

Friedlander, R., & Wagon, S. (1993). Double Simpson's Paradox. Mathematics Magazine, 66(4), 268.

Gardner, M. (1976). On the fabric of inductive logic, and some probability paradoxes. Scientific American, 234(3), 119–124.

Garfield, J. (1991). Newsletter of the International Study Group for Research on Learning Probability and Statistics, 4(2).

Garfield, J. (1993). Teaching statistics using small-group cooperative learning. Journal of Statistics Education, 1(1).

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

Gelbaum, B. R. & Olmsted, J. M. H. (1964). Counterexamples in Analysis. San Francisco: Holden-Day.

Gigerenzer, G. & Murray, D. J. (1987). Cognition as Intuitive Statistics. Hillsdale, NJ: Lawrence Erlbaum.

Girosi, F. (1994, May 10). Posting to Edstat-L electronic forum.

Goertzel, B. (1993). The structure of intelligence. New York: Springer-Verlag.

Gonick, L. & Smith, W. (1993). The Cartoon Guide to Statistics. New York: Harper Collins.

Gordon, M. (1991). Counterintuitive instances encourage mathematical thinking. Mathematics Teacher, 84(7), 511–515.

Green, D. R. (1983). A survey of probability concepts in 3000 pupils aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), Proceedings of the First International Conference on Teaching Statistics (pp. 766–783). Sheffield, U.K.: Teaching Statistics Institute.

Greeno, J.G. (1991). Number sense as situated knowing in a conceptual domain. Journal for Research in Mathematics Education, 22(3), 170–218.

Haack, D. G. (1979). Statistical Literacy: A Guide to Interpretation. North Scituate, MA: Duxbury.

Hahn, H. (1956). The crisis in intuition. In J. R. Newman(Ed.), The World of Mathematics, pp.1957-1976. New York: Simon and Schuster.

Hansen, W. (1994). Using graphical misrepresentation to stimulate student

interest. Mathematics Teacher, 87(3), 202–205.

Hemenway, D. (1982). Why your classes are larger than "average." Mathematics Magazine, 55(3), 162–164.

Hogarth, R. (1987). Judgment and Choice (2nd ed.). New York: John Wiley.

Hudak, M.A. & Anderson, D.E. (1990). Formal operations and learning style predict success in statistics and computer science courses. Teaching of Psychology, 17(4), 231–234.

Huff, D. (1954). How to Lie with Statistics. New York: W.W. Norton.

Johnson, D. E. (1989).   An intuitive approach to teaching analysis of variance. Teaching of Psychology, 16(2), 67–68.

Johnson, D. E. (1986). Demonstrating the Central Limit Theorem. Teaching of Psychology, 13(3), 155–156.

Johnson, D. W., Johnson, R. T. & Smith, K. A. (1991). Active Learning: Cooperation in the College Classroom. Edina, MN: Interaction.

Johnson, D. W. & Johnson, R.T. (1992). Encouraging thinking through constructive controversy. In N. Davidson & T. Worsham (Eds.), Enhancing Thinking Through Cooperative Learning (pp. 120–137). New York: Teachers College Press.

Johnson, D. W. & Johnson, R.T. (1988). Critical thinking through structured controversy. Educational Leadership, 45(8), 58–64.

Johnson, D. W. & Johnson, R.T. (1993). Creative and critical thinking through academic controversy. American Behavioral Scientist, 37(1), 40–53.

Johnson, R. E. & Herr, D. G. (1993, August). Direction: The invisible player. Proceedings of the Section on Statistical Education [annual meeting of the American Statistical Association], 48–52.

Jones, L.V. (1991). Using cooperative learning to teach statistics.   (Research Report No. 91-2). Chapel Hill: University of North Carolina, L.L. Thurstone Psychometric Laboratory.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430–454.

Karylowski, J. (1985). Regression toward the mean effect: No statistical background required. Teaching of Psychology, 12(4), 229–230.

Kernighan, M. D. (1994, May 21–June 10). [letter to the editor] Chance News.

Kohn, A. (1992). Defying intuition: Demonstrating the importance of the empirical technique. Teaching of Psychology, 19(4), 217–219.

Kolman, B., Levitan, M.L., & Shapiro, A. (1993). College Algebra (3rd ed.). Orlando, FL: Saunders College Publishing.

Konold, C. (1991). Chance Plus: A computer-based curriculum for probability and statistics. [Annual review (year 2). NSF Grant # MDR-8954626.] (Paper No. 251). Amherst, MA: University of Massachusetts, Scientific Reasoning Research Institute.

Konold, C. (1993). Personal communication to the author.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. Journal for Research in Mathematics Education, 24(5), 392-414.

Konold, C. (1994). Teaching probability through modeling real problems. Mathematics Teacher, 87(4), 232–235.

Kuhn, T. (1970). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Kunoff, S. & Pines, S. (1986). Teaching elementary probability through its history. College Mathematics Journal, 17(3), 210–219.

Lapin, L. L. (1987). Statistics for Modern Business Decisions (4th ed.). Austin: Harcourt Brace Jovanovich.

Lee, A. S. (1989). The pattern forming mode of teaching and learning statistics. International Journal of Mathematical Education in Science and Technology, 20(2), 321–328.

Lembke, L. O. & Reys, B. J. (1994). The development of, and interaction between, intuitive and school-taught ideas about percent. Journal for Research in Mathematics Education, 25(3), 237–259.

Levin, J. R. (1993). An improved modification of a regression-toward-the-mean demonstration. American Statistician, 47(1), 24–26.

Lewin, K. (1952). Group decision and social change. In G. Swanson, T. Newcomb
& E. Hartley (Eds.), Readings in social psychology (pp. 459-473). New York: Henry Holt.

Lock, P.F. & Lock, R.H. (1993, January). Parallels between calculus reform and statistics education reform. Paper presented at the meeting of the Mathematical Association of America, San Antonio, TX.

Lock, R. H. & Moore, T. L. (1991). Low-Tech Ideas for Teaching Statistics. (Technical Report No. 91-008). Claremont, CA: Pomona College, Statistics in Liberal Arts Workshop.

Lord, N. (1990). From vectors to reversal paradoxes. Mathematical Gazette, 74(467), 55–58.

Marzano, R.J., Brandt, R.S., Hughes, C.S., Jones, B., Presseisen, B.Z., Rankin, S.C. & Suhor, C. (1988). Dimensions of Thinking: A Framework for Curriculum and Instruction. Alexandria, VA: Association for Supervision and Curriculum Development.

Melvin, K. B. & Huff, K.R. (1992). Standard errors of statistics students. Teaching of Psychology, 19(3), 177–178.

Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. Educational Studies in Mathematics, 14, 415-429.

Mitchem, J. (1989). Paradoxes in averages. Mathematics Teacher, 82(4), 250–253.

Moore, D. S. & McCabe, G. P. (1993). Introduction to the Practice of Statistics (2nd ed.).   New York: W.H. Freeman.

Moore, D.S. (1988). Should mathematicians teach statistics? College Mathematics Journal, 19(1), 2–35.

Moore, D. S. (1993). Teaching statistics as a respectable subject. In F. Gordon & S. Gordon (Eds.), Statistics for the Twenty-First Century (pp. 14–25). Washington, DC: Mathematical Association of America.

Morgan, J. L. & Ginther, J. L. (1994). The magic of mathematics.

Mathematics Teacher, 87(3), 150–153.

Morris, C. N. (1988). Personal communication.

Moses, L.E. (1986). Think and Explain with Statistics. Reading, MA:
Addison-Wesley.

National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation
Standards for School Mathematics. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1991). Professional Standards for
Teaching Mathematics. Reston, VA: NCTM.

National Science Foundation. (1992). The science of learning math and science.
Mosaic, 23(2), 37-43.

Nesbitt, R.E., Krantz, D.H., Jepson, C., & Kunda, Z. (1983). The use of statistical
heuristics in everyday inductive reasoning. Psychological Review,
90(4), 339–363.

Ohlsson, S., Ernst, A.M., & Rees, E. (1992). The cognitive complexity of learning
and doing arithmetic. Journal for Research in Mathematics Education,
23(5), 441–467.

Osberg, T. M. (1993). Psychology is not just common sense: An introductory
psychology demonstration. Teaching of Psychology, 20(2), 110–111.

Paik, M. (1985). A graphic representation of a three-way contingency table:
Simpson's paradox and correlation. American Statistician, 39(1), 53–54.

Paulos, J. A. (1994). Counting on dyscalculia. Discover, 15(3), 30-36.

Pearson, E. S. (1962). Some thoughts on statistical inference. Annals of
Mathematical Statistics, 33, 394-403.

Perkins, D.N. & Simmons, R. (1988). Patterns of misunderstanding: An
integrative model for science, math and programming. Review of
Educational Research, 58(3), 303–326.

Piaget, J. (1985). The Equilibration of Cognitive Structures. Chicago: The
University of Chicago Press. (Original work published 1975).

Piaget, J. & Inhelder, B. (1975). The Origin of the Idea of Chance in Children. New York: Norton. (Original work published 1951).

Pollatsek, A., Lima, S., & Weil, A. D. (1981). Concept or computation: Students' understanding of the mean. Educational Studies in Mathematics, 12, 191–204.

Polyson, J.A. & Blick, K. A. (1985). Basketball game as psychology experiment. Teaching of Psychology, 12(1), 52–53.

Pomeranz, J.B. (1984). The dice-problem: Then and now. College Mathematics Journal, 15(3), 229–237.

Pringle, D. (1994). Who's the DNA fingerprinting pointing at? New Scientist, 141(1910), 51–52.

Rapoport, A. (1967). Escape from paradox. Scientific American, 217(1), 50-56.

Read, K.L.Q. & Riley, I.S. (1983). Statistics problems with simple numbers. American Statistician, 37(3), 229-231.

Reichmann, W.J. (1961). Use and Abuse of Statistics. Harmondsworth, Middlesex, England: Penguin Books Ltd.

Riehl, G. (1994, spring). More computer generated thinking. Teaching Statistics, 9–11.

Reinhardt, H. E. (1981). Some statistical paradoxes. In A. P. Shulte(Ed.), Teaching Statistics and Probability, pp. 100–108. Reston, VA: National Council of Teachers of Mathematics.

Resnick, L.B. (1986). The development of mathematical intuition. In M. Perlmutter (Ed.), Perspectives on Intellectual Development (pp. 159–194). Hillsdale, NJ: Lawrence Erlbaum.

Reynolds, S.B., Patterson, M.E., Skaggs, L.P., & Danserau, D.F. (1991). Knowledge hypermaps and cooperative learning. Computers and Education, 16(2), 167–173.

Ringo, M. (1993, August 29). Cable News Network Headline News.

Romano, J. P. & Siegel, A.F. (1986). Counterexamples in Probability and

Statistics. Belmont, CA: Wadsworth.

Rosebery, A. S. & Rubin, A. (1989). Reasoning under uncertainty: Developing statistical reasoning. Journal of Mathematical Behavior, 8, 205–219.

Roth, K. J. (1990). Conceptual understanding in science. In B.F. Jones and L. Idol (Eds.), Dimensions of Thinking and Cognitive Instruction (pp. 138–175). Hillsdale, NJ: Lawrence Erlbaum.

Rubin, H. (1994, May 5). Posting to Edstat-L electronic bulletin board.

Samuels, M. L. (1993). Simpson's paradox and related phenomena. Journal of the American Statistical Association, 88(421), 81–88.

Saville, D.J. & Wood, G.R. (1986). A method for teaching statistics using n-dimensional geometry. American Statistician, 40(3), 205–214.

Schaefer, V. H. (1976). Teaching the concept of interaction and sensitizing students to its implications. Teaching of Psychology, 3(3), 103–114.

Schey, H.M. (1993). The relationship between the magnitudes of SSR(X2) and SSR(X2|X1): A geometric description. American Statistician, 47(1), 26-30.

Schilling, M.F. (1994, spring). Long run predictions. Math Horizons, 10–12.

Schilling, M.F. (1990). The longest run of heads. College Mathematics Journal, 21(3), 196–207.

Schoenfeld, A. H. (1987). Cognitive Science and Mathematics Education. Hillsdale, NJ: Lawrence Erlbaum.

Scholz, R. W. (1991). Psychological research in probabilistic understanding. In R. Kapadia & M. Borovcnik (Eds.), Chance Encounters: Probability in Education (pp. 213–249). The Netherlands: Kluwer.

Shuster, E.F. (1993). Accuracy of proportion estimators: A simple rule. Metrika, 40, 325-332.

Shatz, M.A. (1985). The Greyhound strike: Using a labor dispute to teach descriptive statistics. Teaching of Psychology, 12(2), 85-86.

Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment

with a small-group, activity-based, model building approach to introductory probability at the college level. Educational Studies in Mathematics, 8, 295–316.

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.

Shimp, C. P. & Hightower, F. A. (1990). Intuitive statistical inference: How pigeons categorize binomial samples. Animal Learning Behavior, 18(4), 401–409.

Siegel, M. & Carey, R. F. (1989). Critical Thinking: A Semiotic Perspective. Bloomington, IN: ERIC Clearinghouse on Reading and Communication Skills.

Skagestaad, P. (1981). The Road of Inquiry. New York: Columbia University.

Slonim, M.J. (1960). Sampling in a Nutshell. New York: Simon & Schuster.

Smith, P.T. (1987). Levels of understanding and psychology students' acquisition of statistics. In J.A. Sloboda & D. Rogers (Eds.), Cognitive Processes in Mathematics (pp. 157–168). Oxford: Clarendon Press.

Smith, W. & Gonick, L. (1993, August). The mean, the median, and the mundane: Exploring everyday images for statistical insight. Proceedings of the Section on Statistical Education [annual meeting of American Statistical Assocation], 175–179.

Solomon, P. R. (1979). Science and television commercials: Adding relevance to the research methodology course. Teaching of Psychology, 6(1), 26–30.

Steen, L. A. & Seebach, J. A., Jr. (1978). Counterexamples in Topology (2nd ed.). New York: Springer-Verlag.

Stein, W.E. & Dattero, R. (1985). Sampling bias and the inspection paradox. Mathematics Magazine, 58(2), 96–97.

Stern-Dunyak, A. (1993). Culture clash: Change and technological evolution coming to the statistics classroom. Amstat News, No. 202, 13–14.

Stofflett, R.T. & Stoddart, T. (1994). The ability to understand and use conceptual change pedagogy as a function of prior content learning experience. <u>Journal of Research in Science Teaching</u>, <u>31</u>(1), 31–51.

Stoyanov, J. M. (1987). <u>Counterexamples in Probability</u>. New York: Wiley.

Tan, A. (1986). A geometric interpretation of Simpson's Paradox. <u>College Mathematics Journal</u>, <u>17</u>(4), 340–341.

Tanur, J.M., Mosteller, F., Kruskal, W.H., Lehmann, E.L., Link, R.F., Pieters, R.S., & Rising, G.R. (1989). <u>Statistics: A Guide to the Unknown</u> (3rd ed.). Belmont, CA: Wadsworth.

Tennyson, R. D. (1990). Cognitive learning theory linked to instructional theory. <u>Journal of Structural Learning</u>, <u>10</u>(3), 249–258.

Thomas, G. & O'Quigley, J. (1993). A geometric interpretation of partial correlation using spherical triangles. <u>American Statistician,</u> <u>47</u>(1), 30–32.

Thomma, S. (1994, April 22). House approves most expensive crime bill ever. <u>Austin American-Statesman</u>, A5.

Ulep, S. A. (1991). Strategies preservice secondary mathematics teachers use in solving problems involving uncertainty (Doctoral dissertation, University of Texas at Austin, 1990). <u>Dissertation Abstracts International,</u> <u>52</u>, 105A.

Velleman, P.F. & Wilkson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. <u>American Statistician</u>, <u>47</u>(1), 65–72.

Vos Savant, M. (1990, September 9). Ask Marilyn. <u>Parade</u>.

Vos Savant, M. (1993, March 28). Ask Marilyn. <u>Parade</u>.

Wagner, C. H. (1981). Paradoxes in sampling. In A. P. Shulte(Ed.), <u>Teaching Statistics and Probability</u>, pp.162-167. Reston, VA: The National Council of Teachers of Mathematics.

Watts, D. G. (1991). Why is introductory statistics difficult to learn? And what can we do to make it easier? <u>American Statistician,</u> <u>48</u>(4), 290–291.

Weaver, K. A. (1992). Elaborating selected statistics concepts with common experience. <u>Teaching of Psychology</u>, <u>19</u>(3), 178–179.

Weaver, W. (1963). <u>Lady Luck</u>. Garden City, NY: Doubleday Anchor.

Weinberg, G. H., Schumaker, J. A., & Oltman, D. (1981). <u>Statistics: An Intuitive Approach</u> (4th ed.). Monterey, CA: Brooks/Cole.

Weisberg, S. (1985). <u>Applied Linear Regression</u> (2nd ed.). New York: Wiley.

Well, A. D., Pollatsek, A. & Boyce, S. (1990). Understanding the effects of sample size on the mean. <u>Organizational Behavior and Human Decision Processes</u>, 47, 289–312.

Westcott, M. R. (1968). <u>Toward a Contemporary Psychology of Intuition</u>. New York: Holt, Reinhart, Winston.

Wicklund, R. A. & Brehm, J.W. (1976). <u>Perspectives on Cognitive Dissonance</u>. Hillsdale, NJ: Lawrence Erlbaum.

Wonnacott, T.H. & Wonnacott, R.J. (1977). <u>Introductory Statistics for Business and Economics</u> (2nd ed.). New York: Wiley.

VITA

Lawrence Mark Lesser was born in Princeton, New Jersey on June 24, 1964. His parents are Linda Farfel Lesser and Herbert Arthur Lesser. After completing his work at Bellaire Senior High School, Bellaire, Texas, in 1982, he entered Rice University in Houston, Texas. While at Rice, he was active in contest and recreational mathematics and also gained valuable experience tutoring individuals and groups for the Rice University Mathematics Department, Houston Scholastic Services, and for an outreach project. He received the degree of Bachelor of Arts in Mathematics and Mathematical Sciences from Rice University in May, 1986.

In September, 1986, he entered the Graduate School of the University of Texas at Austin, where he began work as a teaching assistant. Over the next four years, he worked at the University in several other capacities, including consultant, assistant instructor, course coordinator and mentor/tutor. He also assisted Dr. Carl Morris, Center for Statistical Sciences director, in his work applying maximum likelihood estimation and hierarchical modeling to psycho-metric models of test equating. In August, 1989, he completed a Master's Report,

Reliability and the Construction and Interpretation of Tests, and received the degree of Master of Science in Statistics from the Mathematics Department.

After receiving the M.S. degree, he was employed as a Lecturer for the Department of Management Science and Information Systems and then as a statistician and research associate for the redistricting project of the Texas Legislative Council's Research Division. During his two years at the TLC, he took additional courses in statistics, completing all courses required for a Ph.D. from the Department of Management Science and Information Systems. He also began teaching as an adjunct instructor for St. Edward's University, teaching a variety of mathematics and statistic courses, from service to upper-division level. He later taught at Southwestern University (Georgetown, Texas) and for the Texas Union Informal Classes (at the University of Texas at Austin). His teaching consistently received above-average evaluations from both his students and his colleagues.

Since formally entering the Mathematics Education program in 1992, he has given well-received presentations to educators at conferences on the local, state and national levels (including the 1994 joint meetings of the American Mathematical Society and Mathematical Association of America), as well as research colloquia at universities in six states. He has also had several letters-to-the-editor published concerning topics such as statistical misuse. In the summer of 1993, he was awarded the 1993–1994 Jewel Popham Raschke Endowed Presidential Scholarship and also received extensive media coverage (CNN Headline News) for a course he created on the psychology and probability underlying the Texas Lottery. In the fall of 1993, he was invited to present a paper at the fourth International Conference on Teaching Statistics, to be held July 24–30, 1994 in Morocco. In the spring of 1994, he began serving as a referee for *Mathematics Teacher* and *Journal of Statistics Education* and was invited to join the Pi Lambda Theta and Delta Pi Kappa honor societies in education. In August 1994, he will begin working as an Assistant Professor at the University of Northern Colorado.

Permanent Address: Department of Mathematical Sciences,

University of Northern Colorado, Greeley, CO 80639.

This dissertation was typed by the author, with expert formatting assistance from Mr. Lindsey Eck of Corner Oak Publications.