THE EFFECT OF STUDENT-DRIVEN PROJECTS ON THE DEVELOPMENT OF STATISTICAL REASONING

by

Melissa M. Sovak

B.S. Mathematics, Carlow University, PA, 2003

M.S. Computational Mathematics, Duquesne University, PA, 2006

Submitted to the Graduate Faculty of

the Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

STATISTICS DEPARTMENT

This dissertation was presented

by

Melissa M. Sovak

It was defended on

July 16th 2010

and approved by

Leon J. Gleser, Ph.D., Professor, Department of Statistics

Satish Iyengar, Ph.D., Professor, Department of Statistics

Henry Block, Ph.D., Professor, Department of Statistics

Nancy Pfenning, Ph.D., Senior Lecturer, Department of Statistics

Clement Stone, Ph.D., Professor, Department of Psychology in Education

Dissertation Director: Leon J. Gleser, Ph.D., Professor, Department of Statistics

THE EFFECT OF STUDENT-DRIVEN PROJECTS ON THE DEVELOPMENT OF STATISTICAL REASONING

Melissa M. Sovak, PhD

University of Pittsburgh, 2010

Research has shown that even if students pass a standard introductory statistics course, they often still lack the ability to reason statistically. Many instructional techniques for enhancing the development of statistical reasoning have been discussed, although there is often little to no experimental evidence that they produce effective results in the classroom.

The purpose of this study was to produce quantitative data from a designed comparative study to investigate the effectiveness of a particular teaching technique in enhancing students' statistical reasoning abilities. The study compared students in a traditional lecture-based introductory statistics course with students in a similar introductory course that adds a semester-long project. The project was designed to target three main focus areas found in an introductory statistics course: (i) distributions, (ii) probability and (iii) inference. Seven sections of introductory statistics courses were used. One section at each level served as an experimental section and used a five part project in the course curriculum. All other sections followed a typical introductory curriculum for the specific course level.

All sections involved completed both a pre-test and a post-test. Both assessments were designed to measure reasoning ability targeted by the project in order to determine if using the project aids in the increased development of statistical reasoning.

Additional purposes of this research were to develop assessment questions that target students' reasoning abilities and to provide a template for a semester-long data analysis project for introductory courses.

Analysis of the data was completed using methods that included ANCOVA and contin-

gency tables to investigate the effect of the project on the development of students' statistical reasoning. A qualitative analysis is also presented to provide information on aspects of the project not covered by the quantitative analysis.

Analysis of the data indicated that project participants had higher learning gains overall when compared with the gains made by students not participating in the project. Results of the qualitative analysis also suggest that, in addition to providing larger learning gains, projects were also enjoyed by students. These results indicate that the use of projects are a valuable teaching technique for introductory statistics courses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation is dedicated to the memory of my grandfather, Edward HuWalt, who inspired my love of numbers at a young age.

Acknowledgements

I would like to thank all the members of my committee for the time and effort that they put forth to make this dissertation a success. In particular, I would like to thank my advisor, Dr. Gleser, for always being there to give great advice, constructive criticism and pep talks. I also extend my thanks to Dr. Pfenning, who spent a significant amount of time helping with the design of the study and the assessments used. To Dr. Iyengar, thank you for believing in me and always being encouraging. To Dr. Block, thank you for providing me with great suggestions and always being there to answer questions. I would also like to thank Dr. Stone, who provided insight into the educational research environment.

I also want to thank Laurel Chiappetta for inspiring the idea for this study through her own creative teaching methods, providing ideas for the project's development and for providing me with a course section for my reliability analysis. I would like to extend gratitude to all of the instructors and students who participated in this study. Without your help and participation, this study would not have happened. My appreciation also goes to Mary and Kim who always have the answers to my questions (or at least will find the answer for me!).

To all of my family, thanks for all your prayers and support. In particular, to my grandma, Catherine, thank you for all your prayers and for insisting since I was a child that I would one day make a good teacher. Finally to my parents, thank you for supporting me and encouraging me while I went through this journey.

1.0 INTRODUCTION

Statistics education has long sought to provide reasonings on why students seem to miss big ideas in statistics and fail to be able to apply their knowledge to real world problems. In recent years, there has also been an attempt to not only identify why students have trouble comprehending and using statistical knowledge, but also to establish learning techniques that better equip students to develop statistical reasoning and to be able to apply this knowledge outside of the statistics classroom. Research into the area of learning techniques, however, is sometimes largely based on anecdotal studies, where authors recount their own experience with certain techniques in their own classrooms without providing any statistical evidence to support the success of the technique.

One such technique suggested to improve student learning is the use of a project. This technique comes in a variety of forms and has been suggested by a number of researchers. However, again, no quantitative evidence has been provided to show the positive effects of the project on students' understanding of statistical material. Producing quantitative experimental results to strengthen the qualitative claims made by other researchers is the purpose of this study.

1.1 PURPOSE OF STUDY

The purpose of this study is to address whether or not a semester-long, student-driven data anaylsis project enhances students' reasoning in the areas of distribution, probability, and inference.

The following research questions will be studied:

1. Does a student-driven project with interspersed feedback from the instructor have a significant impact on students' conceptual understanding of distributions, probability, and inference?
2. What differences in the understanding of distributions, probability, and inference are there between students whose interaction with these topics occurs through a lecture, recitation, homework format and students whose interaction with these topics includes lecture, recitation, homework and student-led projects?
3. Are students who participate in the lecture, recitation, homework, plus project format better equipped to articulate statistical knowledge than students whose training does not involve a project?

1.2 CONTRIBUTION TO THE FIELD

There are three major contributions to the field of statistics education made by this study. First, it aims to provide quantitative data obtained from a designed comparative study on the effect of the use of projects in an elementary statistics course. While the use of projects has been extensively discussed by other authors, no quantitative data has been provided to verify that projects have a positive effect on students' learning. This study not only attempts to provide quantitative data to show the effect of projects, but also to provide information for instructors on what types of learning gains they can expect to see from students if they incorporate a project into their course.

Second, the study provides a template for a semester-long project in an elementary statistics course. The project protocol developed for this study seeks to incorporate all the major themes covered in an elementary statistics course. It also provides instructors with a grading scheme designed to reward students for developing statistical reasoning and connecting ideas in statistics as well as connecting statistical ideas to the larger context of the problem.

Finally, the study provides two assessment forms designed to assess statistical reasoning abilities. These forms focus on three key areas: displays and descriptives, classification of

variables and inferential test selection. Because the questions on the forms require students to interrelate knowledge about types of variables, relationships between variables and other statistical information, they go beyond basic statistical literacy questions to assess reasoning skills.

1.3 LIMITATIONS OF STUDY

Several limitations should be considered as they pertain to this study. The sample chosen for this study was considered to be the most generalizable group available to the researcher at this academic level. However, this sample of students may not be entirely generalizable to students at other institutions or with different academic interests.

Second, the sample size may not be of an appropriate size to produce results of sufficient accuracy concerning the effect size. It also may not be large enough to produce the classically accepted power for statistical anaylsis. Also, the percentages of students participating in each assessment and who completed both assessments varied. Table 1 shows the various percentages for participation for each course. As shown, the percentages for each section vary to a certain degree, with percentages for post-test participation consistently lower than pre-test participation. Also, percentages for students who completed both assessments is consistently lower than students completing just one form. Percentage of students participating is also listed in Chapter 4 when sample sizes differ from those shown in the table below. Also, Table 1 shows the time of each course section (day or night). At the STAT 1000 level, both sections were day sections. At the STAT 0200 level, the traditional and project sections were both night sections, while the Big Picture sections were all day sections. This difference should be noted since there may be differences between students who take day courses and students who take night courses. However, at each level, the project group and the traditional control group are directly comparable because they occur at the same time.

Finally, there may be variables that produce differences that were not measured in this study. Variables considered in this study included pre-test score, post-test score, difference of scores, year level of student, prior experience of student and instruction method. Other

variables may also influence the outcome, but were not considered.

Section	Course Level	Group	Time	Pre-test	Post-test	Both
1	STAT 1000	Experimental	Day	98.9%	82.0%	78.7%
2	STAT 0200	Experimental	Night	90.0%	57.5%	50.0%
3	STAT 0200	Control (Big Picture)	Day	95.0%	83.3%	75%
4	STAT 0200	Control (Big Picture)	Day	74.7%	68.1%	60.4%
5	STAT 0200	Control (Big Picture)	Day	100%	95.4%	89.5%
6	STAT 0200	Control	Night	74.7%	68.7%	56.6%
7	STAT 1000	Control	Day	79.5%	69.3%	53.4%

Table 1: Percentage of Students Participating for Each Course Section

1.4 ORGANIZATION OF THE DISSERTATION

This dissertation is organized into five chapters. Chapter one provides a motivation for the research, an overview of the study, its contributions to the field, and a discussion of the study. Chapter two reviews prior research into statistics education, including research about the three core reasoning areas of the project. It also focuses on pedagogical theory surrounding the use of projects in the classroom. Chapter three contains information about the methods used to analyze the data and why these methods were chosen. Chapter four presents the results of the analyzed data. Chapter five contains a discussion of the research findings, the implications of these findings and directions for future research studies.

2.0 LITERATURE REVIEW

2.1 STATISTICAL REASONING

In the last 20 years, the number of statistics courses taught on college campuses around the United States has risen dramatically. Many of these courses are designed to give students of varying backgrounds a basic understanding of statistical concepts. Despite a strong push to integrate statistics and probability into the K-12 classroom, these introductory courses are often a student's first formal experience with statistics. Therefore, there has been a significant amount of literature produced concerning statistics education at the college level.

Prior to advancements in technology such as the graphing calculator and the personal computer, introductory statistics courses were primarily focused on teaching students how to perform calculations. However, with advancements in technology, this is no longer necessary in statistics courses of today and so the focus has shifted from learning how to complete complex calculations to learning how to reason statistically. This shift in the classroom has caused a shift in the focus of statistical education research. Much of the research deals with the educational issues that arise when educators try to teach abstract concepts and students try to learn them [7].

Since the focus of research has shifted, it has become increasingly important to define what is meant by statistical reasoning and how it differs from statistical literacy and statistical thinking. Several studies attempt to define differences among the three; however, many authors still use these terms interchangeably. In recent years, there has been a considerable effort made to agree on definitions for these terms. The following definitions are the result of several papers written on the subject.

- Statistical literacy encompasses basic and important skills including being able to organize and display data appropriately and work with different data representations as well as understanding concepts, vocabulary and symbols [7]. It also includes the ability to interpret, evaluate and communicate statistical information and arguments [27].
- Statistical reasoning requires interpreting summaries or representations of data and involves connecting one concept to another or combining ideas. "Reasoning means understanding and being able to explain statistical processes and being able to fully interpret statistical results" [7] (pg. 7).
- Statistical thinking involves understanding the "big ideas" of statistics. It consists of an understanding of how and why statistical investigations are conducted. Thinking statistically involves keeping in mind a constant relation to the context of the problem and being able to interpret conclusions in non-statistical terms as well as being able to see the complete process with constant revisions for each component part [7, 15].

Despite these well-formed definitions, there are still differing opinions on how they relate to each other. Several authors employ the model described by Figure 1 where these three concepts overlap each other, but also have distinct parts. This model maintains that each domain has independent content from the other two, while there is also some overlap [22]. Other researchers feel that statistical reasoning and thinking are completely subsets of statistical literacy. This model is represented by Figure 2. This viewpoint suggests that reasoning and thinking are not independent from literacy and are, in fact, subgoals within the development of the all-encompassing goal of statistical literacy [22].

In either model, the concepts overlap in some way and this leads to differing use of the terms. In any case, recent studies focus their attention on the issues of developing statistical reasoning.

An alternative perspective on how to view the differences among literacy, thinking and reasoning was provided in 2002 by delMas [22]. He suggested that what sets the three domains apart are tasks rather than statistical content. His commentary provided the following categorization of tasks into the three domains, as seen in Table 2. This perspective provides a distinct view of each of the three domains without having significant overlap and thus alleviates problems separating each domain into distinct learning goals.

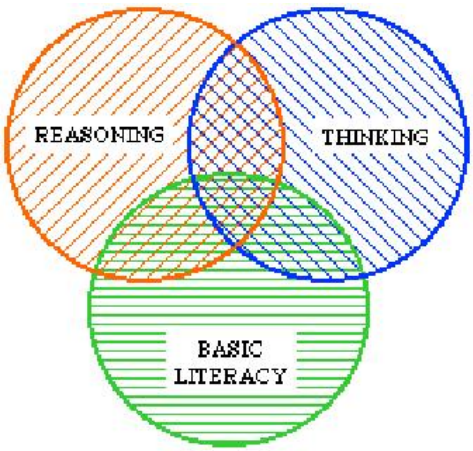


Figure 1: Overlap Model

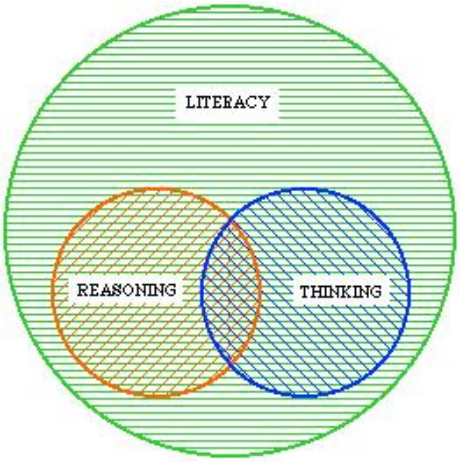


Figure 2: Subset Model

BASIC LITERACY	REASONING	THINKING
Identify	Why?	Apply
Describe	How?	Critique
Rephrase	Explain	Evaluate
Translate	(the process)	Generalize
Interpret		
Read		

Table 2: Categorization of tasks into the three domains

2.2 RESEARCH INTO STATISTICS EDUCATION

Early studies in statistics education concentrated on understanding how students reason in the areas of probability and statistics - specifically, how people make judgments under uncertainty. This work is best represented by Kahneman, Slovic, and Tversky's (1982) book, *Judgment under Uncertainty: Heuristics and Biases* [42]. Much of the work included in this book deals with behaviors exhibited while reasoning about statistical situations, and in many cases demonstrates specific errors made while reasoning about these situations. This work determined that people rely on a limited number of heuristic principles which aid in reducing the complex tasks required in statistical reasoning, but which can also lead to serious errors. Two of the heuristics that were identified as a result of this work were representativeness and availability. The representativeness heuristic is typically used when individuals are presented with probabilistic questions and states that people tend to make decisions about the probability of that event occurs based on how similar the event is to the distribution from which it is drawn [70]. For example, an individual would decide that the sequence of coin flips H H T H H H is not likely because it does not contain 50% heads. The availability heuristic states that people tend to assess the probability of an event by determining how easily they can provide examples of the event occurring [42]. For example, people may base their assessment of the probability of a middle-aged person having a heart

attack on how many middle-aged people they knew who had heart attacks. This research paid little attention to the realities of everyday problem solving and classroom learning; however it provided a foundation for constructs that could be applied in a more practical setting.

In the mid-to late-1980s researchers started to become interested in confirming and understanding how to correct the misconceptions found in Kahneman and Tversky's earlier work. Two experiments were conducted by Pollatsek, Konold, Well, and Lima (1984) [65] in which they used one particular type of assessment question that Kahneman and Tversky had used when identifying the representativeness heuristic. In the first experiment, students were asked to determine the average for a sample given the population mean and the first value from the sample, then asked what they expected the average for the remaining students in the sample to be. In the second experiment, interviewers suggested alternative solutions to the questions to the students in order to assess how strongly students held to their original answer. The alternative solutions also were suggested to students in order to probe further what kind of reasoning the students were using. In the study, Pollastek et. al. put forth their own model to classify the reasoning students were using, but, concluded that the evidence did not support their model, and that students' reasoning *was* consistent with the representativeness heuristic proposed by Kahneman and Tversky. One of the implications from this study was that "since students' actual heuristic, representativeness, is so different in form from the appropriate mechanistic belief, it may not be easy...to effect any lasting change in students' beliefs" [65] (pg. 401).

In 1986, Fong, Krantz and Nisbett [26] experimented to determine how well statistics students could handle questions presented to them outside of a classroom context. Specifically, Fong et. al. wanted to determine whether the level of statistical training affected the statistical "quality" of the responses to the questions. For this study, statistical quality was defined as how much formal statistical reasoning was used to answer the assessment questions. Fong et. al. selected students at random from a college-level introductory statistics course; half of the students were tested during the first week of the semester and the other half were tested during the last week of the semester. Students were contacted outside of the context of the course and asked questions in the framework of a sports issues survey.

The conclusion from the study was that whereas statistical training did have some effect on the statistical quality of the responses to some questions, it played no role in the statistical quality of the responses to half of the questions. The results of their work suggested that formal statistical training did not necessarily guarantee a higher level of statistical reasoning skills [54].

In 1989, Konold [44] tested his model of statistical reasoning, called the "outcome approach," against the representativeness heuristic. His results suggest that there were "additional misconceptions that ought to be taken into account" (p. 93). In the study, Konold observed that students used casual, informal experiences without applying any mathematical model to reason about probability questions. Konold used these results to demonstrate that, whereas Kahneman and Tversky's work was a start to understanding student misconceptions, there was more to be understood.

After these studies in the 1980s, research began to evolve to match the new reform movement in statistics instruction. By the 1990s, many textbooks and classroom curricula were focusing more on reasoning about data and less on memorizing and working with formulas. This brought about a new focus in research where the motivation was now to document students' strengths and weaknesses under the reformed curriculum and to evaluate specific instructional instruments [54].

In 1991, Garfield and delMas [32] used a three-stage model to develop and evaluate an instructional unit on probability. The first stage attempted to identify misconceptions; the second stage involved the development of an instructional approach based on a theory or model of learning; the third stage was assessment. During the study, students were given a pre-test on probability at the start of an introductory course, then given a post-test on probability at the end of the course. The study showed that, while students showed an overall increase in correct responses, they still maintained particular misconceptions after instruction.

Other studies from this time period focused on evaluating various new teaching techniques. Several studies by Cohen and various colleagues [16, 17, 18] tested whether or not students' conceptual understanding of statistics was affected by the use of a software package. While the results showed that students who used the software package exhibited greater

learning gains than students who did not, students who had used the software only scored 57% correct on average on the post test. This demonstrated that the students' ability to reason statistically left much room for improvement.

More recent research has focused on how students develop statistical reasoning skills. One specific hypothesis for these studies was that students develop better statistical reasoning if they are exposed to statistics through exploratory data analysis (EDA). EDA is an approach to teaching statistics that employs open-ended questions with an emphasis on interpreting graphical displays in order to develop students' ability to analyze and explore data [6]. Research has been done involving several teaching experiments which used EDA as the primary method of instruction. Ben-Zvi [6] studied two above-average seventh grade students in an EDA classroom. The development of the students' reasoning was tracked over two and a half months using video recordings, classroom observations, interviews and assessment of the students' notebooks. The results of the study showed that students develop their reasoning by working with the ideas related to the philosophy of an EDA environment, including collecting data, making hypotheses and formulating questions. A similar study was completed by Groth in 2003 [35]. This study sought to explore the defining characteristics of different patterns of high school students' statistical thinking while collecting, describing, organizing, representing, and analyzing data. The study was able to identify patterns of responses that matched with levels of sophistication previously described by other researchers.

Another focus of an EDA classroom has been to involve more technology. This is reflected in two of Biehler's 1997 studies [8]. In these studies, students in a data analysis course used a software tool throughout the course and also to complete a data analysis project at the end of the course. In this study, Biehler was able to identify difficulties that students encounter in a data analysis course more extensively.

Another study by Lovett [54] sought to use EDA to understand students' learning processes and improve their statistical reasoning through the use of new instructional tools. The study followed students through a lecture-based course with weekly lab activities completed using technology. The lab activities consisted of students using commercially available statistics packages to complete data analysis. The study provided insight concerning the design of computerized learning activities for students. Because the study's researchers were comprised

of cognitive psychologists, statistics instructors, educational researchers, and instructional technologists, the study was also able to offer insight from many different perspectives on how students learn effectively.

In summary, most of the work in the seventies, eighties and early nineties focused on reasoning in probability situations. The research from the 1970s focused on the theoretical aspects of misconceptions in statistical reasoning. The research from the 1980s focused primarily on testing the theories of how these misconceptions arise while documenting the abilities of statistics students. Finally, the research in the early 1990s focused on documenting students' difficulties in statistical reasoning. More recent research has focused primarily on the effects of EDA in the classroom.

2.3 REASONING ABOUT DESCRIPTIVES AND DISTRIBUTION

The concept of distribution is one of the first topics introduced in an introductory statistics course. Many researchers argue that without this concept, students cannot properly describe and summarize data [34, 59]. There are two main types of distributions, empirical and theoretical. "Empirical distributions are the foundation of students' work in an introductory statistics course" [34] (p. 167). Several researchers have studied students' conceptual understanding of distribution.

In 1995, Mokros and Russell [59] completed a study on elementary school students to determine how they develop the notion of average. They found that until students can think of a data set as a unit rather than a series of individual values, it cannot be appropriately described and summarized. They also found that until the data sets can be viewed as a whole, averages cannot make sense.

A similar view was shared by Bakker and Gravemeijer [4] who developed the structure seen in Table 3.

The structure can be read both upwards and downwards with the upward perspective representing the typical reasoning pattern for novices in statistics [4]. Students also need a downward perspective so that they can sensibly choose between measures. In their study,

distribution (conceptual entity)			
center	**spread**	**density**	**skewness**
mean, median, midrange, ...	range, standard deviation, interquartile range, ...	(relative) frequency, majority, quartiles	position majority of data
data (individual values)			

Table 3: Between data and distribution

Bakker and Gravemeijer hoped to answer the question of how students develop a notion of distribution by observing four different seventh grade classrooms in the Netherlands over the course of a year. Using a specifically designed series of interactive web applets, the researchers were able to track how students develop reasoning about distributions. The researchers demonstrated that by asking students to move back and forth between interpreting distributions and constructing distributions, students were able to represent many elements of distributions through graphs. They determined that using graphs was a very effective tool in developing students' conceptual understanding of distribution.

Several other studies have also advocated that graphical representations help students to reason about distributions. Some of this research has focused on difficulties that students encounter when using graphs to learn about data sets [10, 43, 77]. Bright and Friel [10] documented students' confusion between differences in bar charts and histograms, which lead them to try to describe shape, center and spread of bar graphs or to confuse the meaning of bars in a histogram. Wild and Pfannkuch [77] along with Konold and Pollatsek [43] found that students often use graphs as illustrations rather than as tools to learn about a data set.

More recent research has recommended that the instructional focus shift from drawing various kinds of graphs to learning graphing skills for the purposes of detecting patterns,

working with hypotheses and making sense of the data. In 2005, Pfannkuch [64] studied Year 11 students' assessment responses as they learned about comparing boxplots. The students were also introduced to back-to-back stem-and-leaf plots, but, as the study found, students tended to reason only with the boxplot representation. When using the stem-and-leaf plots, students tended to reason as if there were no underlying data. This study suggests that the educational practice of teaching multiple graphs may not help students learn to reason about data, but that focusing on particular graphs does aid students in their reasoning about data.

In a study done by delMas, Garfield and Ooms [19], college students were tested on their literacy and reasoning involving graphical representations. The researchers analyzed student performance using a series of multiple-choice items. They also found that college students confused bar graphs and histograms and attempted to estimate shape, center and spread when it was inappropriate. The researchers additionally identified other errors made in reading and interpreting these types of graphs. Based on these difficulties, the researchers questioned whether students should even have exposure to graphs other than dotplots and boxplots. However, after extensive conversations with colleagues, it was determined that the need for histograms in the curriculum was vitally important, especially for learning about large data sets.

In summary, after several studies, it has been shown that to clearly understand descriptive measures of a data set, and hence a data set overall, it is important to have a clear concept of distributions. Graphical representations of the data are a good tool to help students better understand the concept of distributions.

2.4 REASONING ABOUT PROBABILITY

Several researchers have documented and studied reasoning and difficulties in reasoning about probability. Early work in this area by Kahneman and Tverksy was discussed above. Their work presented heuristics that people use to make probabilistic judgments. These heuristics are based on a collection of intuitive assessments and often violate traditional probability rules used to correctly analyze problems. In 1983, Nisbett, Krantz, Jepson and

Kunda [62] made the argument that people can analyze a situation probabilistically with little formal training when both the sample space is easily recognizable and the role of chance is prominent in the problem. In contrast to this, however, Kahneman and Tversky showed that even people with considerable training in probabilistic models can be led to apply intuitive assessments even when they know situations call for probabilistic analysis.

In 1989, Konold [45] expanded on the heuristics described by Kahneman and Tversky's work. The purpose of his study was to show that errors in how people reason when uncertainty is present include both the application of intuitive assessment as well as analyses based on confusion of what is being asked. In his study, Konold proposed the "outcome approach". In this model, the goal in dealing with uncertainty is merely to predict the outcome of the next single trial. For example, given an irregularly shaped bone to roll and asked which side was most likely to land upright, most participants interpreted the question as asking for a prediction of the outcome of the next trial. These same participants tended to evaluate their answer as correct or incorrect based on only one trial. Konold concluded, on the basis of his study, that the heuristics presented by Kahneman and Tversky were not enough to capture the reasoning behind particular errors that are made when reasoning about probability, and that the outcome approach could provide further explanation on different sets of errors.

In 1993, Konold, Pollatsek, Well, Lohmeier and Lipson [46] argued that there was sense in both Kahneman and Tversky's theories on natural assessment as well as Nisbett, et. al.'s claims that reasoning can develop with little formal training. They asserted that a typical person has knowledge about many uncertain situations, but that the knowledge that they have is incomplete and not integrated. Because of this, a person may appear to reason correctly in one problem, but in another problem, this same person may reason in ways that violate probabilitistic theory. They concluded that painting a picture of how students reason about chance was not simple because students bring a variety of beliefs and perspectives into the classroom. They suggested that specific problems needed to be developed to discriminate between individuals who reason according to specific heuristics.

More recent research into reasoning about probability models has focused on specific teaching techniques or the development of specific problem types. For example, Rossman and Short [68] advocate that there is still a place for teaching conditional probability in

reformed statistics education and present a variety of realistic examples that can be used. They maintain that the introduction of conditional probability is an important concept to aid in clarification of statistical inference, and should not be omitted from introductory statistics courses. A similar study was completed by Warner, Pendergraft and Webb [75]. Since many students have no formal experience with probability concepts, the authors encouraged the use of Venn diagrams in the guise of pizzas to facilitate students' understanding of probability theory. They maintain that using this form that is more familiar to students allows them to discover advanced probability topics in a familiar setting. There are several studies of this nature concerning the best way to introduce probability concepts. Many researchers, however, do not focus specifically on reasoning about probability because these topics are sometimes discouraged in introductory statistics courses under the reformed curriculum.

2.5 STATISTICAL INFERENCE

Statistical inference is formally defined by the Collins English Dictionary [12] as "the theory, methods, and practice of forming judgments about the parameters of a population, usually on the basis of random sampling." Statistical inference should move beyond the data to draw conclusions about a wider universe while taking variation into account and remembering that conclusions are uncertain [60]. Much of the research on statistical reasoning has focused on students' difficulties in understanding and using statistical inference.

An early study completed by Oakes [63] provides a context for several more recent studies. In Oakes' study, 70 academic psychologists were surveyed with an instrument consisting of six true-false items associated with a problem scenario reporting statistically significant results. Each of the six statements were incorrect interpretations of statistical significance. Despite this, more than half of the 70 psychologists thought at least half of the statements were true and only three correctly marked all statements as being false. In the second half of the exercise, participants were asked to select from a list the interpretation of a p-value that they typically use. Only eight of the 70 respondents wrote a correct interpretation. The study provided evidence that many of the participants misinterpreted the meaning of statistical

significance.

A similar study was completed by Falk and Greenbaum [24] using items similar to those used by Oakes. This study focused on university students with two previous statistics courses and gathered the students' interpretations of the items. In one of the courses taken, students read a paper warning readers about common difficulties with interpreting statistically significant results. Despite reading this paper, only seven of the 53 respondents chose the correct option stating that all statements were false. Falk and Greenbaum concluded that merely citing or warning students of common misconceptions is insufficient to help them overturn them.

In 2002, Haller and Krauss [36] replicated Oakes' study across six German universities. The sample consisted of methodology instructors, scientists from non-methods courses and psychology students. Results from this study showed that 80% of instructors, 89.7% of scientists and 100% of students selected true for at least one misinterpretation of the p-value. Based on these results, Haller and Krauss concluded that students are not the only ones with misconceptions, but that these misconceptions are often shared by instructors as well.

A study by Vallecillos and Holmes [74] focused on students' understanding about "proving" the truth or falsity of statistical hypotheses. Over 400 students from varying backgrounds were surveyed with 20 true-false items. The results showed that nearly a third of responses reflected that students believed that significance tests definitively proved the truth or falsity of the null or alternative hypothesis.

Additional misconceptions were discovered in 1997 by Wilkerson and Olson [78]. They surveyed graduate students with a six-item questionnaire. The focus of the study was whether interpretations of significance tests reflected an understanding of the relationship between treatment effects, sample size, and Type-I and Type-II error. Results from this study suggested that the impact of sample size on treatment effects was not well understood.

Additional misunderstandings were highlighted in two studies done by Williams [79, 80]. In the first of these studies, Williams identified several sources of students' misunderstanding of p-values. In the latter study, Williams interviewed eighteen introductory statistics students to explore their conceptual understanding of significance using concept-mapping.

In the concept-mapping portion, students arranged concept labels on a page and were required to provide the links connecting the concepts. Procedural knowledge was also tested by providing data to students and asking them to calculate two statistical significance tests. Williams found that students had difficulty completing the concept-map. In follow-up interviews, students failed to be able to provide a correct description of the p-value and only one student gave a correct description of statistical significance. Williams attributed this to students' difficulties with statistical languages rather than misconceptions. One particular misconception that Williams did point out, however, was that the students in the study seemed to believe that the p-value was always low.

Additional difficulties in reasoning about inference noted by Batanero [5] included the belief that the significance level is the probability that the null hypothesis is true, the belief that the p-value is the probability that the result is due to chance and the belief that the formalism of rejection regions and levels of significance is a matter of arbitrary convention as opposed to justified mathematical theory. Batanero also suggested that introductory students misapply the Boolean logic of the converse in some of their statistical analyses by switching the hypothesis and conclusion of a conditional.

A study by Mittag and Thompson [58] focused on members of the American Educational Research Association and considered topics directly related to interpretation of p-values and statistical significance. This study found that many respondents interpreted non-significant findings as unimportant, confused p-values with Type I error and believed that p-values tested the probability of results occuring in the sample rather in than the population.

More recent research seeks to give suggestions for teaching methods to improve students' reasoning with respect to statistical inference. In 1999, delMas, Garfield and Chance [23] introduced a program called *Sampling Distributions* which allows students to change the shape of a theoretical population and run simulations. The hope was to improve the instruction of sampling and hence reasoning about inference. While the study showed a significant change from pretest to post-test, many students still displayed some serious misunderstandings of sampling distributions.

In 2004, Watson [76] and Chance, delMas, and Garfield [14] suggested that using good simulation tools and activities for teaching concepts such as sampling distribution and the

Central Limit Theorem could help students better understand statistical inference. However, Lipson [50] cautioned that simulation tools were not enough to properly introduce students to sampling because the software packages used were often distribution specific and so they have no specific role in illustrating links between concepts. Lipson, Kokonis and Francis [51] commented that the use of software tools often was not enough to improve students' understanding of sampling and hence of statistical reasoning.

2.6 USE OF REAL DATA AND PROJECTS

In 2005, the American Statistical Association published the GAISE (Guidelines for Assessment and Instruction in Statistics Education) report. One of the main recommendations of the GAISE report is to use real data in the classroom [1]. Many recommendations have been made to use real data in classroom examples to illustrate statistical concepts. Another way that real data is often incorporated into the classroom is through projects or experiments because it has been noted by several researchers that students tend to be more invested in understanding why a dataset behaves a certain way if it is their own data [38, 55].

Several authors have advocated the use of projects to present students with real data experience. One of the earliest authors to suggest the use of a project was Hunter [40], who suggested the use of a three week project assignment in an experimental design course. He urged the use of projects because they provided students with a deeper understanding of material. Nearly 15 years later, Hogg [38] advocated the use of projects for similar reasons, stating that projects give students experience in asking questions, formulating hypotheses, summarizing data and communicating findings.

In the early 1990s, several researchers [11, 37, 73, 81, 83] advocated the use of projects to teach statistical concepts. All of these authors believed that the use of projects provides students with much-needed hands-on experience with statistical concepts. In particular, Wolfe [82] stated that learning statistics was more meaningful when students collected data of interest to them and analyzed their own hypotheses. In 1992, Roberts [67] discussed his experiences with projects in introductory statistics courses for MBA students. He classified

student projects into three main types: sample surveys, predictive time series and process improvement and descibed his implementation of each type of project in the classroom. Mackisack [55] discussed the benefits of using a project for second year majors specializing in mathematics. She found the students' reactions to the project were positive and conjectured that these projects helped students connect p-values, etc. to the context of the data and make statistics more readily transferable to other situations.

In 1994, Fillebrown [25] provided details of her implementation of a project in an elementary statistics course. The semester-long project involved students choosing a topic of interest to them and examining associations between their variables. The final product consisted of a written report detailing the results of their project as well as displaying their data in graphical and tabular form. Fillebrown indicated that overall the projects were very good, however, she advised spending class time doing sample projects to assist students in navigating through their own project. She observed that the projects made teaching the course more enjoyable and that students felt that the projects helped tie the material together.

Other research from the early 1990s includes Garfield's [33] research on using practical projects as an assessment technique in high school statistics courses. She suggested that alternate forms of assessing statistical knowledge were needed to inform teachers about how well students communicate using statistical language and to indicate how well students understand statistics as an interrelated set of ideas. Garfield suggested using a practical project as an assessment method. The project she outlined consisted of two versions, the first where students collected a small dataset of interest to them, the second where students collected information about themselves for three to five weeks. Garfield found that the projects were useful in indicating students' understanding of statistical ideas and their ability to apply these ideas in analyzing data.

In 1995, Ledolter [49] and Garfield [28] both advocated the use of projects in the classroom. Again, Garfield discussed the use of projects to aid in students' understanding of concepts and also as an assessment technique. In his paper, Ledolter discusses the use of projects in a second-year statistical methods course for business students. He gives some examples of projects which are student-driven, and states that he has had a positive experience with using projects in the classroom. He asserts that it is important that students

collect their own data, so that they are involved in the statistical process from beginning to end. He also mentions that he has used similar projects in his introductory statistics course with similar, positive results. He concludes that projects are useful at any level. Also around this time, Sevin [69] presented some tips to help introductory statistics students with data analysis projects. He emphasized that clear and detailed instructions should be given to students whenever possible in order to assist them in completing a successful project.

In 1998, several researchers again reviewed their experiences with projects in statistics courses. Love [53] discussed the use of projects in a second statistics course for data analysis. In this course, students were asked to plan, gather and analyze four sets of data over the course of the semester. Love felt that the addition of the project made the course more enjoyable for both the instructor and the students. Around this time, Anderson-Cook [2] also discussed the use of projects, this time in a design of experiments course. The project required students to design their own experiment at the start of a course and then comment on their design. Again it was found the students' reaction to the use of projects was generally positive.

Also in 1998, Smith [71] discussed the use of projects in an introductory statistics course. In his introductory course, Smith introduced a semester long project that consisted of bi-weekly assignments that matched the coverage of the course material. After completing the semester, students were surveyed. Of the 30 students surveyed, 24 indicated that the project format was a "great idea" while the remaining 6 thought it was a "good idea". None of the students listed the format as a terrible idea or a bad idea. Smith also notes that examination scores seemed to have improved with the introduction of the project, citing that in pre-project the mean midterm exam score was 80.79 with a standard deviation of 16 and the mean final exam score was 80.27 with a standard deviation of 12.56, whereas the project course had a midterm mean exam score of 92.13 with a standard deviation of 6.96 and a final exam mean score of 88.12 with a standard deviation of 8.28. Smith took this as an indication that student projects helped students learn about statistics and also produced more effective communication skills.

Melton, Reed and Kasturiarachi [57] also discussed the use of projects in an introductory statistics course. They examined the use of two different real-life data projects in two

elementary statistics courses. In one of the courses, students worked with projects geared toward their fields of study, whereas in the second course students worked with a comprehensive project based on data from a local industry. The purpose of this research was two-fold: first to determine what student attitudes were towards the comprehensive project and the major-oriented project and second to determine if non-traditional, returning students, who are typically underprepared, performed well in the course. Course grades were examined and showed that non-traditional students scored very close to the average scores of traditional students, which had not been the case in previous courses. These results reinforced the idea that projects are valuable in increasing students' statistical understanding.

In the following year, Holcomb and Ruffer [39] wrote an article designed to provide resources to fulfill the growing desire for educators to use projects in introductory statistics courses. They suggested using one dataset in a semester-long project, proposing that using the same dataset throughout the semester allows students to discover connections among statistical techniques. They also suggested that the use of computer software and the requirement of presenting the project in written form leads to a greater understanding of statistical methods. The authors then describe their proposed project along with a grading rubric to assess students' work. Students were surveyed after participating in a course where this project was used to determine the benefit of the project; however, no assessments on student learning were made. Their survey indicated that 94% of students felt the project helped better their understanding of statistical results.

In 2002, several authors further discussed the use of projects in the classroom. Richardson [66] introduced "The World of Chance", a non-typical introductory statistics course, in which the main assessment item is a project consisting of a report and a poster. Richardson's research did not provide any specific assessment data concerning whether or not this type of introductory statistics course provided any benefit in developing students' reasoning, but was suggested to be a fresh approach to teaching statistical concepts. Binnie [9] addressed the use of projects in the classroom in his discussion of statistics education reform. He reinforced that projects where students collect their own data are important because they require students to use real data and become actively involved because they have chosen the subject, which forces enhanced learning. He also stated that he believed projects also require students to

make all the decisions about the analysis which helps in learning statistical concepts.

More recent research concerning projects in the classroom has focused on student attitudes toward statistics. In 2008, Carnell [13] studied two introductory statistics courses to determine if student attitudes toward statistics were improved in a course using a project versus a traditional course. The study used the Survey of Attitudes Toward Statistics which students completed on the first and last days of the semester. Students in the project-based course were found to not exhibit a more positive attitude toward statistics than students in the traditional course, which disagrees with several earlier studies. Carnell mentions the need for further studies on differentl structured projects.

In summary, several researchers have discussed the use of projects in statistics courses, some as an alternate assessment technique, others because they feel that it enhances students' development of statistical understanding. Several researchers have found that students' attitudes toward statistics have improved and that students generally enjoy completing projects in their course. However, no data has been collected to determine if projects improve students' statistical reasoning using a controlled study. All evidence provided by other authors has been primarily anecdotal or based on a one sample design.

2.7 ASSESSMENT

In statistics education, there has been some question as to how to assess student learning, particularly how to assess statistical reasoning and thinking. Garfield [28] has suggested that, in addition to traditional methods such as tests and quizzes, alternative methods of assessment are necessary to accurately measure students' conceptual understanding. Some researchers [56] have relied on interviews and open-ended questioning techniques to study students' statistical reasoning; others have used self-generated pretests and posttests [70].

In their article about the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) website [3], delMas, et. al. [20] hypothesize that the reason for the absence of research into the effect of statistics reform may be because of the lack of a standard assessment instrument. Several assessment instruments have been developed but fail to

provide a general assessment for statistical reasoning due to a lack of broad enough focus of material or audience. One of these assessment instruments is the Statistical Reasoning Assessment (SRA). The SRA is a multiple-choice test consisting of 20 items, developed by Konold and Garfield to evaluate the effectiveness of a new statistics curriculum in US high schools [44, 29, 30, 31]. Each item in the SRA describes a problem and offers several choices of response, most of which include a statement of reasoning in favor of a particular choice. Unfortunately, this assessment focuses heavily on probability and does not include items related to data collection and statistical inference [31].

Another assessment instrument is the Statistics Concepts Inventory SCI [72], which was developed to assess statistical understanding using closed-form items. The SCI, however, was written for a specific audience (engineering students) and may not have broad enough coverage to test the statistical understanding of college students in general [20]. In response to this, the ARTIST authors attempted to create an assessment instrument that would have broader coverage of the statistical content covered in an introductory statistics course as well as applying to a broader range of students.

The ARTIST website currently provides a number of assessment resources designed to evaluate students' statistical literacy, reasoning and thinking. One of these resources is the Comprehensive Assessment of Outcomes in Statistics (CAOS) test. The development of CAOS had a dual purpose. First, CAOS was intended to provide a reliable assessment that covered topics that any student completing an introductory statistics course would be expected to understand. Secondly, CAOS was intended to identify areas where students do or do not make significant gains in reasoning [20]. All items included in the CAOS test have a realistic context and are designed to require students to reason. There are no items included in the assessment that require students to compute or recall definitions. The CAOS test underwent several phases of testing and modification. The current CAOS test consists of 40 questions which produced a Cronbach's alpha coefficient of 0.77 [3].

The ARTIST website also provides 11 online unit tests designed to measure conceptual understanding in important areas of an introductory course [21]. The topics for the 11 unit tests are: Data Collection, Data Representation, Measures of Center, Measures of Spread, Normal Distribution, Probability, Bivariate Quantitative Data, Bivariate Categorical Data,

Sampling Distributions, Confidence Intervals and Tests of Significance. Each test consists of seven to twelve multiple-choice items designed to assess literacy and reasoning [21]. The website also provides a data bank of over 1100 items, searchable by topic and by learning outcome (literacy, reasoning, thinking).

3.0 METHODS

The goal of this study is to identify if learning gains are made in the area of statistical reasoning when students are exposed to a project and also to discover the size of these gains. While several researchers have advocated the use of projects in the classroom, no quantitative data obtained from controlled or comparative designed studies has been provided to substantiate the claims that projects significantly increase students' understanding of major ideas in statistics.

This chapter gives an overview of the design for this study. It includes a general discussion of the design along with a description of the participants and the plan of analysis used to determine the effect of the project.

3.1 PARTICIPANTS

This study involved students in seven separate sections of introductory level courses, five of which were at the STAT 0200 level and two of which were at the STAT 1000 level. The STAT 0200 level of introductory statistics teaches descriptive and inferential statistics, allowing students who complete the course to conduct their own analysis of standard one- and two-sample data sets, to follow statistical reasoning and to read statistical reports with understanding. The STAT 1000 level of introductory statistics is a more intensive introduction to statistical methods and is designed for students who want to complete data analysis and continue studies of statistical ideas beyond the scope of the introductory course. The emphasis in this course is placed on the statistical reasoning underlying the methods. Students could not be randomly placed into groups for the purpose of this study. Courses were

selected with instructor agreement to serve as either an experimental section or a control section. Students were not informed prior to registration that their course would be participating in a study nor were they informed of which section (experimental or control) their course would serve as. Students registering for a participating course had no way of knowing prior to registration.

All students registered in the courses were asked to participate in the study regardless of their level of experience with statistics. Students who complete both the pre-test and post-test were included in the study. Students who completed only one of the tests were excluded as were students who did not complete the entire post-test.

3.2 RESEARCH DESIGN

This study researched subjects in two types (STAT 0200, STAT 1000) of introductory statistics courses on the University of Pittsburgh's main campus. A third type of introductory course (STAT 1100) was used to provide a reliability analysis for the measuring instrument. The sections of both STAT 0200 and STAT 1000 were separated into: (i) a control group and (ii) an experimental group who will complete the project. In each course type, one section was designated as the experimental group and at least one other section was designated as a control group. For STAT 0200, there were four separate sections used as controls taught by two different instructors. The control sections for STAT 0200 used two different teaching methods, traditional and big picture. The big picture method is further discussed later in this section. The experimental group for the STAT 0200 level was taught by a third instructor. At the STAT 1000 level, there was one section for both control and experimental groups, each taught by a different instructor (these instructors differed from the instructors who taught at the STAT 0200 level). Sections of each course type were assigned to treatment or control groups based on instructor preference. A summary of this design can be found in Table 4.

Students in all levels and both groups completed a pre-test and a post-test designed to assess their reasoning abilities in specific areas where the researcher expected that students

with exposure to the project would make greater learning gains than students not exposed to the project. Because these questions asked students to interrelate ideas in a contextual setting, the questions assess reasoning abilities rather than basic literacy. The purpose of the pre-test was two-fold. First, it provided a background on the participating students to determine if there were significant differences between the experimental and control sections. Second the pre-test provided a baseline from which the researcher can determine the significance of learning gains made. There were two forms of the measuring instrument used in the study as a pre-test and post-test (Form A and Form B). Students completing one form as a pre-test completed the other form as a post-test. The development of these forms is further discussed in a later section. A summary that shows which forms were completed as a pre-test for each course section can be found in Table 4.

Section	Course Level	Instructor	Group	Pretest Form
1	STAT 1000	1	Experimental	A
2	STAT 0200	2	Experimental	B
3	STAT 0200	3	Control (Big Picture)	B
4	STAT 0200	3	Control (Big Picture)	A
5	STAT 0200	3	Control (Big Picture)	B
6	STAT 0200	4	Control	A
7	STAT 1000	5	Control	B

Table 4: Summary of Study Design

Students in the control group had a traditional lecture/recitation/homework format in their course. Lecture sections were permitted to use the Exploratory Data Analysis (EDA) format and real data analysis, where EDA is defined as a teaching approach using open-ended questions with an emphasis on interpreting graphical displays. Recitations were also permitted to use this format and group activities were also permitted to explore topic concepts. Homeworks were to be assigned from the textbook or from an outside source, but they were required to be structured in such a way that students were given the data to analyze and told what questions needed to be answered. Assignments or group work in which students

gathered their own data, generated their own hypotheses or probabilistic questions or guided their own analysis was not permitted in the control section.

In three sections of the STAT 0200 level, one instructor used the "Big Picture" approach to teach statistics. This approach was designed to guide students learning of specific techniques while keeping in mind a broader context. None of the techniques used in this approach violated the control group environment as defined above; however, the techniques differ from that of a traditional introductory statistics course. For this reason, this group will be considered as a third, separate instructional technique to be compared with traditional methods and courses containing the project.

Students in the experimental group had the lecture/recitation/homework/project format. This format followed a similar format to the control group with the exception of the project. That is, the students will have a similar lecture/recitation/homework experience with the added experience of the project, described in depth below.

Both experimental and control groups had similar exposure to concepts within their statistics courses. Because the project was designed as an outside assignment, similar to homeworks, lectures and recitations were similar between groups. Also in order to ensure that exposure was similar between groups, homework assignments for project groups were appropriately shortened in order to provide the experimental group with a similar amount of work outside of the classroom.

3.2.1 Project Description

The project consisted of five distinct assignments: an assignment on each of the three focus areas, an initial assignment reporting the data that will be used and also a fifth assignment that required students to summarize all of their findings in written form. Students worked in groups of 4-5 to complete the project. The assignments were designed to emphasize the "How" task from Table 2 by requiring students to determine the appropriate course of analysis for their data without specific instructions from the instructor.

Students were introduced to an overview of the project and groups were selected in Week 1 of the semester. By the end of week 2, students reported the topic of their research and

located necessary data to be used for the project. Students submitted a one page description of their topic, the data (with variables of interest) and the source of their data.

The first of the three topic assignments was submitted during Week 4 of the semester. This assignment dealt with descriptive statistics and required that students describe the subject matter that they have chosen using numerical and graphical displays.

The second assignment was submitted during Week 8. This assignment required students to identify both single and multiple events that could occur in the context of their topic and then use probability techniques from their lecture to assign probabilities to these events. Students were also asked to comment on the probability distributions events may follow.

The third assignment was submitted during week 12. This assignment required students to test three hypotheses about their topic. Students were required to formulate each hypothesis, and then find and use an appropriate method to test it.

The final overview assignment was submitted during Week 15. This assignment required that students tie together all of the methods used, questions asked and answered and conclusions drawn. Students were also asked to provide a reflection on the assignment as a whole.

To increase the likelihood that the division of labor in each group was fair, students were asked to submit a peer evaluation in which they evaluated their own work along with all of their group members' work for each portion of the assignment. Copies of each individual assignment along with all grading rubrics can be found in the appendix.

3.2.2 Assessment

As mentioned above, students involved in the study completed both a pre-test and a post-test. These tests were designed to measure students' reasoning ability in three specific areas: displays/descriptives, classification of variables, test selection. Two assessments were designed for the study. Some of the questions contained in the assessments were modified versions of questions provided in the ARTIST database. In some cases, the modifications were provided to change open-ended questions to closed-form questions in order to make the assessment more manageable for students to complete. In other cases, modifications were

provided to make the problem statement less vague. Still other questions in the assessments were generated by the researcher. These questions were designed to be similar to the ARTIST questions and to target the areas where the researcher expected learning gains would be made by students exposed to the project. Questions were designed to test reasoning abilities by emphasizing that students be able to correctly select methods for testing, describing and relating data. These tasks specifically correspond to the "How?" task listed under "Reasoning" in Table 2 because they ask students how to perform appropriate analysis for specific contexts and variables. Also, because students are asked to relate concepts within a contextual framework, these questions specifically target the reasoning abilities discussed in the definition of statistical reasoning in Chapter 2.

Students who completed Form A as a pre-test completed Form B as a post-test and vice versa. Pre-test forms were assigned randomly to each section involved in the study. The pre-test was given during week 1 of the course and the post-test was given during week 15 or week 16 of the course, depending on each instructor's preference.

Each assessment form was 16 questions in length. After the assessments were written, 8 similar questions were swapped between the two forms to ensure that the forms were as equivalent as possible. A reliability analysis was also performed using a third and separate type of introductory course, STAT 1100. Students in this course completed both forms of the assessment in two sequential lectures in order to judge the equivalence of the forms. Time between completion of each assessment was minimized in order to more accurately determine reliability. A copy of both forms can be found in Appendix A.

3.2.2.1 Display and Descriptive Questions The first eight assessment questions on each form focused on displays and descriptive statistics. Four of these questions asked students to identify the appropriate display for a particular variable. The other four questions asked students to identify the appropriate descriptive statistic(s) for the variable in question. The focus of each question along with its question number for each form is summarized in Table 5.

To answer these questions correctly, students needed to understand what types of displays and descriptive statistics are appropriate for what types of variables and also what displays

and descriptive statistics are appropriate in the presence of outliers and skewed data.

This focus area was selected by the researcher because it was believed that students who have worked with their own data via the project would have a better understanding of how outliers and skewed data affect descriptive statistics. This understanding, most likely cultivated by working with their own non-trivial datasets, should aid students in selecting appropriate descriptive statistics. Students working with their own data were also exposed to determining variable types in a real-world data setting. This experience was expected to aid students in classifying variables correctly and hence selecting appropriate displays based on this classification.

Question Focus	Form A Question	Form B Question
Single Categorical Display	1	8
Categorical → Quantitative Display	2	2
Single Quantitative Descriptive	3	3
Single Categorical Descriptive	4	7
Quantitative → Quantitative Descriptive	5	5
Quantitative → Quantitative Display	6	6
Single Quantitative Display	7	1
Single Quantitative with Outlier Descriptive	8	4

Table 5: Summary of Display/Descriptive Questions on Assessment Forms

3.2.2.2 Classification of Variables Questions 9 through 12 on each assessment focused on the classification of variables. Specifically, it was required that students be able to classify variables not only as categorical or quantitative but that students were also able determine which variable was the explanatory and which was the response. A summary of each question's focus along with it's question number for each form can be found in Table 6.

To answer these questions correctly, students not only needed to be able to appropriately classify variables, but to understand the relationship between the variables.

Classification of variables was selected as a focus area because, as mentioned above, it

was believed that students with exposure to the project would have experience in classifying variables. Also, because students were asked to formulate their own questions in the context of the projects, they would gain more experience in understanding and articulating relationships between variables.

Question Focus	Form A Question	Form B Question
Categorical → Categorical	12	11
Categorical → Quantitative	9	12
Quantitative → Categorical	10	9
Quantitative → Quantitative	11	10

Table 6: Summary of Classification Questions on Assessment Forms

3.2.2.3 Test Selection The final four questions on each assessment focused on test selection. These questions asked students to correctly identify the most appropriate statistical test given the variables and hypotheses involved. The focus of these questions along with their question numbers on each form can be found in Table 7.

To answer these questions correctly, students needed to be able to correctly identify the variable types involved, the relationship between these variables and how the question being asked translated into a hypothesis.

Test selection was chosen by the researcher as a focus area because it directly correlated to one of the project assignments. In the third topic assignment, students were required to form questions, translate these into hypotheses and then select the appropriate method to test their hypotheses. It was expected that students exposed to the project would more easily be able to identify appropriate tests based on the situations given because of their experience in guiding their own analyses.

Question Focus	Form A Question	Form B Question
Regression	13	14
χ^2	14	15
ANOVA	15	13
Paired t-test	16	16

Table 7: Summary of Test Selection on Assessment Forms

3.3 DATA ANALYSIS

This section describes the methods that might be used and the methods that were used to determine effects in a study of this type. The main hypothesis to be tested is that "project" has an effect on the difference between pre-test and post-test scores over and above what would be expected as a result of exposure to a traditional lecture-based statistics course. It was also the goal of this study to determine where and how large these effects may be.

3.3.1 Multilevel Models

Multilevel models are often used in educational research and are used for data that is clustered into hierarchically organized groups. The nature of this study fits well with the multilevel model format since we have students nested within teachers nested within course type. Because students are all affected by the same teacher, these students cannot be considered independent observations. Also, students and sections within a certain course level would be expected to have similarities due to their grouping into this particular course type. Even though both course types represent introductory statistics courses, different course types may be required by different majors. This may lead to a clustering of certain majors within a STAT 0200 course that may not occur in a STAT 1000 course. Multilevel models account for this correlation between observations within clusters unlike single level models, which assume independence of all observations.

Given the appropriate amount of data, the multilevel model for this data would consist of the following levels: student level, class level and course level. The following equations outline the appropriate model for this data. The level-one model can be seen in Equation (3.1). The equation contains an intercept, the three potential predictors: EXPERIENCE, METHOD and PRETESTSCORE along with an interaction term for PROJECT and EXPERIENCE. All slopes in the model, along with the constant, are random. These can be changed to fixed if it is determined that random slopes are not necessary.

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}EXPERIENCE + \beta_{2jk}METHOD \\ &+\beta_{3jk}PRETESTSCORE + \beta_{4jk}(EXPERIENCE * METHOD) + \epsilon_{ijk} \end{aligned} \tag{3.1}$$

The level-two model equations are shown in Equation (3.2). The left-hand side of each of these equations represents one of the terms from the level-one model (intercept or slope term). In the first equation, β_{00k} represents the mean intercept value from the level-one model in the k^{th} course level and β_{01k} represents the slope for the level-two predictor, TEACHER. In the next four equations, the first term on the right-hand side of the equation represents the mean value of the the level one slope for the k^{th} course level. Each equation includes the predictor, TEACHER. Again, all slopes in these equations are random, which can also be changed to fixed if it can be determined that higher level variance components are not significant.

$$\begin{aligned} \beta_{0jk} &= \beta_{00k} + \beta_{01k}TEACHER + u_{0jk} \\ \beta_{1jk} &= \beta_{10k} + \beta_{11k}TEACHER + u_{1jk} \\ \beta_{2jk} &= \beta_{20k} + \beta_{21k}TEACHER + u_{2jk} \\ \beta_{3jk} &= \beta_{30k} + \beta_{31k}TEACHER + u_{3jk} \\ \beta_{4jk} &= \beta_{40k} + \beta_{41k}TEACHER + u_{4jk} \end{aligned} \tag{3.2}$$

Finally, the level-three model equations are shown in (3.3). The left-hand side of each equation represents a term from the level-two models. On the right-hand side of the first equation, the first term represents the mean value for the intercept from levels one and two. In the latter equations, the first term represents the mean value for the appropriate slope from levels one and two. The second term in each equation incorporates the predictor,

LEVEL. The slopes and intercepts in each of these equations are fixed parameters because there will be no significant higher level variance components due to the fact that this is the highest level.

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + \gamma_{001}LEVEL + u_{00k} \\
\beta_{10k} &= \gamma_{100} + \gamma_{101}LEVEL + u_{10k} \\
\beta_{20k} &= \gamma_{200} + \gamma_{201}LEVEL + u_{20k} \\
\beta_{30k} &= \gamma_{300} + \gamma_{301}LEVEL + u_{30k} \\
\beta_{40k} &= \gamma_{400} + \gamma_{401}LEVEL + u_{40k} \\
\beta_{01k} &= \gamma_{010} + \gamma_{011}LEVEL + u_{01k} \\
\beta_{11k} &= \gamma_{110} + \gamma_{111}LEVEL + u_{11k} \\
\beta_{21k} &= \gamma_{210} + \gamma_{211}LEVEL + u_{21k} \\
\beta_{31k} &= \gamma_{310} + \gamma_{311}LEVEL + u_{31k} \\
\beta_{41k} &= \gamma_{410} + \gamma_{411}LEVEL + u_{41k}
\end{aligned} \tag{3.3}$$

These equations can be combined by substituting the values from the level-three equations into the level two-equations, and then substituting the new level-two equations into the level-one equations. This will create a mixed model with some effects being random and some fixed.

Due to the nature of the data collected for this study, the sample sizes at levels two and three were too small to estimate the multilevel model. Consequently, other methods had to be employed to analyze the dataset obtained during this research study.

3.3.2 Actual Quantitative Analysis

Because the two course types used in this study are considered to be different, each course level was analyzed separately. Also, due to differences in the design at each level, each level was analyzed using differing models and techniques that were appropriate for the data within each course level.

3.3.2.1 STAT 1000 Analysis At the STAT 1000 level, one section was an experimental group and one section was a control group. Because of this, it was impossible to separate

project and teacher effect. For this reason, the multilevel model was abandoned in favor of using an analysis of covariance model (ANCOVA). The ANCOVA model incorporates both factors and covariates that may influence the dependent variable. In this model, both the group means of the covariate as well as any linear relationship between the covariate and the dependent variable are taken into account [52]. This procedure also allows us to compare pairs of means to find differences.

For the STAT 1000 level of this study, the dependent variable was the difference between pre-test and post-test scores. The factor was "teaching method" (project or traditional) and the covariate was pre-test score. The main hypothesis of interest was whether there was a difference between the mean difference in scores from the project course versus the mean difference in scores from the non-project course:

$H_0 : \tau_{proj} = \tau_{noproj} = 0$
H_a : not both τ_{proj} and τ_{noproj} equal zero

Table 8: Hypotheses associated with the ANCOVA for STAT 1000

Comparisons were also made between the two means to determine the direction of any differences existing between the project and non-project groups. These comparisons were made by halving the p-value for the factor from the ANCOVA model and determining the sign of the difference in means. This gives a one-sided test for the equality of means. The hypotheses shown in Table 9 were tested.

$H_0 : \mu_{proj} \leq \mu_{noproj}$
$H_a : \mu_{proj} > \mu_{noproj}$

Table 9: Hypotheses associated with comparisons of group means for STAT 1000

Following these two analyses, subtest and individual question analyses were performed to determine in what focus areas the project group performed better than the non-project group. Differences between subtest pre-test scores and subtest post-test scores for each of the three focus areas were examined to determine if there was a particular area in which students

in the project group showed higher differences than students in the non-project group. This analysis was also performed using an ANCOVA model in which the subtest difference was the dependent variable, project (presence or absence) was the factor and subtest pre-test score was the covariate.

Finally, contingency tables were formed for each individual question, showing how correct and incorrect answers changed from pre-test to post-test. Conditional probabilities were computed for the off-diagonal entries. These conditional probabilities highlighted the percentage of students who went from incorrect on the pre-test to correct on the post-test and the percentage of students whose answers went from correct on the pre-test to incorrect on the post-test. These percentages were then compared between project and non-project courses to determine the questions on which the project group performed better.

3.3.2.2 STAT 0200 Analysis Analysis for the STAT 0200 level followed a similar course to the analysis for the STAT 1000 level. However, at the STAT 0200 level, there were five sections involved. One section acted as the experimental group while four sections were designated as control groups. Of the four control group sections, three were taught by one instructor using the "Big Picture" approach. The other section was taught by a separate instructor using traditional methods.

To analyze the data for this course type, two ANCOVA models similar to the model used for the STAT 1000 type were used. Because there were two different types of control groups, an analysis of the project group versus the traditional group only was provided to compare directly to the results found at the STAT 1000 course type. A separate ANCOVA model was then used to compare all three groups. The dependent variable in both ANCOVA models was the difference of pre-test and post-test scores. The factor and covariate used in the model were again "teaching method" and pre-test score, respectively. For the comparison between the project and traditional method, the factor had the same categories as in the STAT 1000 model; for the comparison of project with both control groups, the factor had three categories: project, big picture or traditional.

The first hypothesis tested at this level using the ANCOVA determined if differences existed between the project and traditional methods. The specific hypotheses tested can be

$H_0 : \tau_{proj} = \tau_{noproj} = 0$
H_a : not both τ_{proj} and τ_{noproj} equal 0
$H_0 : \tau_{proj} \leq \tau_{noproj}$
$H_a : \tau_{proj} \geq \tau_{noproj}$

Table 10: Hypothesis associated with the ANCOVA for STAT 0200 two group comparison

.

found in Table 10. Again, in the case of the two group analysis, the p-value for the factor from the ANCOVA model was halved and the sign of the difference was noted to provide a one-sided test for the equality of means. A second ANCOVA model was used to test the hypotheses found in Table 11 to determine if there were differences between the three groups.

$H_0 : \tau_{proj} = \tau_{bigpic} = \tau_{trad} = 0$
H_a : not all τ equal zero

Table 11: Hypotheses associated with the ANCOVA for STAT 0200 three group comparison

Following the ANCOVA for the three groups, an intersection-union test was used to determine if project scores were better than both big picture and traditional methods simultaneously. This procedure was introduced by Berger [41] in 1982 to test composite hypothesis of the form $H_0 : \bigcap H_{0j}$. The procedure requires that all H_{0j} be rejected in order to reject H_0. This test is also known as the min-test, a term introducted by Laska and Meisner [48], who showed that the composite hypothesis of the form $H_0 : \mu_0 \leq \mu_1$ or $\mu_0 \leq \mu_2$ is rejected if and only if the minimum t-statistic computed from the two individual t-tests has a p-value below the significance level α. This is equivalent to rejecting the composite null hypothesis if the maximum p-value from both tests is less than the significance level α. This procedure provides an α-level test and can be used even when test statistics are dependent. It is also a better confirmatory procedure for this type of comparatory analysis than classical multiple comparisons. The composite hypotheses tested using this procedure are shown in Table 12.

$H_0 : \mu_{proj} \leq \mu_{trad}$ or $\mu_{proj} \leq \mu_{bigpic}$
$H_a : \mu_{proj} > \mu_{trad}$ and $\mu_{proj} > \mu bigpic$

Table 12: Hypotheses associated with comparisons of group means for STAT 0200

Following these analyses, subtest and individual question analysis was also performed similar to the analysis done at the STAT 1000 level. Again, each analysis was performed first on the project and traditional data and then using all three groups.

3.3.2.3 Reliability Analysis Using a section of a different course type of introductory statistics, a reliability analysis was performed on the two assessments used in this study. To determine the reliability of the two forms of the assessment, A and B, several techniques were used. First, as is typical in parallel-forms reliability, the correlation between total scores on the two forms was found. In addition, a paired t test was also used to determine if there were differences between the mean scores on each form. A paired t test provides a comparison of the test and retest means along with the correlation, allowing us to see not only if the scores were strongly associated but also whether, on average, scores on the second form were the same as scores on the first form [61]. In this case, we look for non-significant values on the t test to indicate that there was not a difference between the two forms.

Individual question analysis was also performed to highlight questions where performance may have been different from form to form. As part of this individual question analysis, Kappa values were also generated for each question. Kappa values are used as a measure of agreement. The strength of Kappa values can be determined using Table 13, taken from Landis and Koch [47]. Negative values of Kappa suggest that the forms agree less than expected by chance.

Kappa	**Strength**
0	Poor
0.01 - 0.20	Slight
0.20-0.40	Fair
0.40-0.60	Moderate
0.60-0.80	Substantial
0.80-1.00	Almost perfect

Table 13: Interpretation of Strength of Kappa Values

4.0 RESULTS

The results of the study are reported in this chapter. They are organized into three sections. The first section focuses on the reliability analysis performed for the two forms of the assessment. It includes overall results along with a question-by-question analysis. The second section reports results from the STAT 1000 course level of the study, including an analysis of pretest scores, model results, subtest analysis results and individual question analysis. The third section provides the same information for the STAT 0200 course level.

4.1 RELIABILITY ANALYSIS

The reliability analysis was performed on a section of a separate introductory statistics course, STAT 1100. The two assessment forms were given in two consecutive lecture meetings. Only students who completed both Form A and Form B were used for the reliability analysis. Using only students who completed both forms resulted in having 46 observations for the analysis.

4.1.1 Correlation and Paired t Results

For the first part of the reliability analysis, means for both forms were calculated. These means are shown in Table 14. As can be seen in this table and in Figure 3, the means and distributions appear to have no significant differences.

The correlation between forms was found to be .591 with a p-value of 0, indicating that it is significantly different from 0. The results of the paired sample t-test to determine

Form	Mean	Standard Error	Median
Form A	9.22	.396	9.5
Form B	9.80	.413	10

Table 14: Mean scores on Assessment Forms for Reliability Group

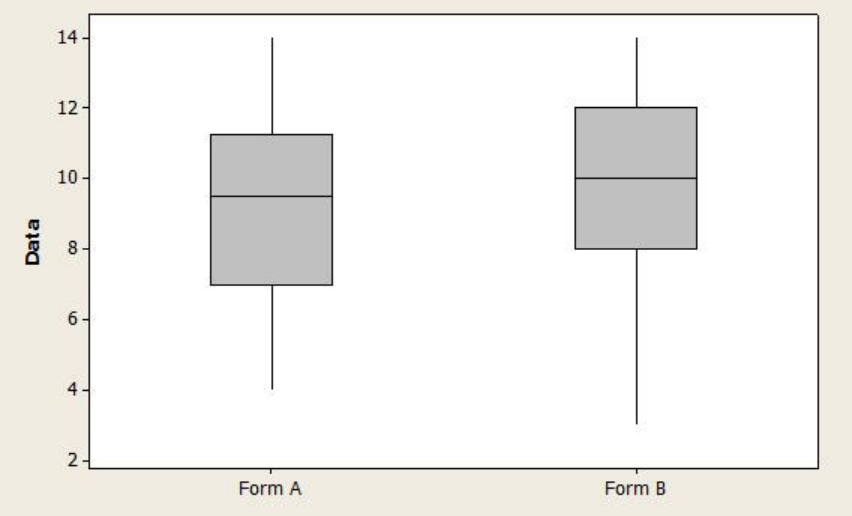


Figure 3: Side-by-Side Boxplots representing Scores on each Form for Reliability Group

if differences existed between the mean score for Form A (9.22) and the mean score for Form B (9.80) was not significant with t(45) of -1.604 and two-tailed p-value of .116. Also, Cronbach's α was calculated for the two forms to be .741. The longer CAOS test that was a source of model for some of the questions on Forms A and B produced a Cronbach's α coefficient of .77. So the reliability of these forms is comparable to that of the CAOS test.

Figure 4 shows a scatterplot of scores from the two forms. As can be seen in the the scatterplot, there are three observations that may be considered outliers. These observations are those where scores on the second form were significantly lower than scores on the first form (a difference of 5 or more). When these observations were removed, the sample size was reduced to 43. The correlation without these observations was .716 with a p-value of

0. A paired t-test was also performed without these observations and produced a p-value of .163 when comparing the mean of Form A (9.35) and Form B (9.79), indicating that there were no significant differences between the two forms.

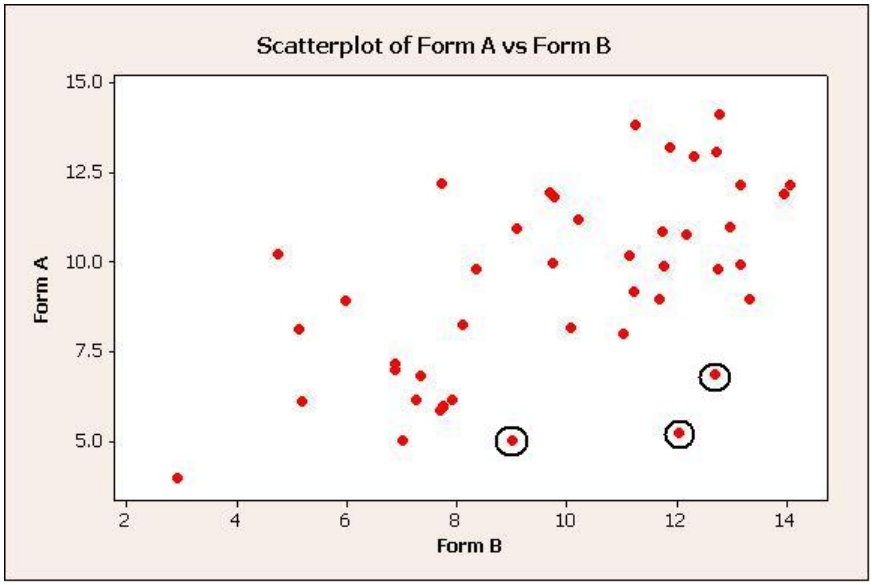


Figure 4: Scatterplot of Form A Score versus Form B Score

4.1.2 Subtest Analysis

Further testing was performed for each of the three subtests to determine if there were any significant differences between scores on these forms. Means from each form for each of the subtests are shown in Table 15. These means were calculated without excluding the outliers mentioned above.

For the display and descriptive subtest, the correlation between forms was found to be .698 (0). The results of the paired sample t-test to determine if differences existed between the mean score for Form A and Form B produced a p-value of .606, indicating that significant differences between the forms did not exist.

The classification subtest produced a correlation statistic of .370 (.011). The paired t-test for equality of means produced a p-value of 1, indicating that, even though the correlation is low, there are not significant differences between the two forms. Finally, the test selection

subtest produced a correlation statistic of .202 (.179), indicating that the correlation is not significantly different from 0. The paired t-test of means produced a significant p-value of .008, indicating that there are significant differences between the two forms for this subtest. As seen in Table 15, the mean difference of scores between the forms is different by .5 points.

Display and Descriptive Subtest		
Group	Mean	Standard Error
Form A	5.46	.256
Form B	5.57	.279
Classification Subtest		
Group	Mean	Standard Error
Form A	2.67	.179
Form B	2.67	.176
Test Selection Subtest		
Group	Mean	Standard Error
Project	1.09	.131
Non-project	1.57	.141

Table 15: Results for Subtest Analysis for Reliability Analysis

4.1.3 Individual Question Analysis

Table 16 outlines the percentage of students whose score remained the same on each form, i.e., a correct response on Form A and a correct response on Form B, or an incorrect response on Form A and an incorrect response on Form B. For each question the percentage of participants for which the score did not change was at least 50%. For ten of the questions the percentage of unchanged scores was at least 70%. Also shown in Table 16 is the Kappa value for each question. Following the interpretation from Table 13, 3 of the Kappa values shown are considered poor strength, 5 are considered slight strength, 6 are considered fair strength, 1 is considered moderate strength and one is considered substantial strength. These

results conflict with the results obtained in the subtest analysis. In the subtest, analysis, the test selection subtest was shown to have significantly different means on each form, however, there are moderate and fair strength Kappa values within this subtest.

Question Focus	**Percentage Remaining the Same**	**Kappa**
Single Categorical Display	80.4%	.119
Categorical → Quantitative Display	50%	.172
Single Quantitative Decriptive	73.9%	-.007
Single Categorical Descriptive	73.9%	.330
Quantitative → Quantitative Descriptive	73.9%	.235
Quantitative → Quantitative Display	84.4%	.661
Single Quantitative Display	50%	.043
Single Quantitative with Outlier Descriptive	58.7%	.180
Categorical → Categorical	82.6%	.237
Categorical → Quantitative	63.0%	.200
Quantitative → Categorical	73.9%	.363
Quantitative → Quantitative	71.7%	.290
Regression	84.8%	.148
χ^2	67.4%	.271
ANOVA	71.7%	.424
Paired t-test	54.3%	-.1

Table 16: Form-to-Form Percentage of Unchanged Answers for Reliability Group

4.2 STAT 1000 RESULTS

The following sections present the results of the analysis performed on the STAT 1000 course level. At the STAT 1000 level there was one section that made up the experimental group and one section that made up the control group. Data used for the analysis consisted of scores of

participants who completed both the pre-test assessment and the post-test assessment. For the experimental group there were 70 participants for which matched scores were available and for the control group there were 47 participants with matched scores.

4.2.1 Pretest Results

Students participating in the study in either the experimental or control group completed a pre-test designed to measure their statistical reasoning in three specific areas of statistics. The pre-test was a sixteen question multiple choice assessment with questions concerning appropriate displays and descriptive statistics for variables, classification of variables and test selection for various situations. Scores on the pre-test indicated the number of questions that students correctly answered on the test as a whole. The mean and its standard error along with the median scores for each group can be seen in Table 17. The distribution of scores overall can be viewed in the side-by-side boxplots shown in Figure 5.

Group	Mean	Standard Error	Median
Project	6.81	.255	7
Non-project	9.15	.360	10

Table 17: Mean and Median Scores for STAT 1000 Pre-test

It is clear that the pre-test scores are lower for the project group than for the non-project group. A one-sided two-sample t test showed that these differences were significant, with a p-value of approximately 0. Further investigation of this phenomenon revealed that there were differing proportions of students in each class level (freshman, sophomore, junior and senior) in each group. A summary of the proportions of class level in each group can be found in Table 18.

As seen in Table 18, there are higher proportions of students at the junior and senior level in the project group, while there are higher proportions of students at the freshman and sophomore level in the non-project group. These proportions were shown to be significantly different when tested using a χ^2 test for equal distributions (.002).

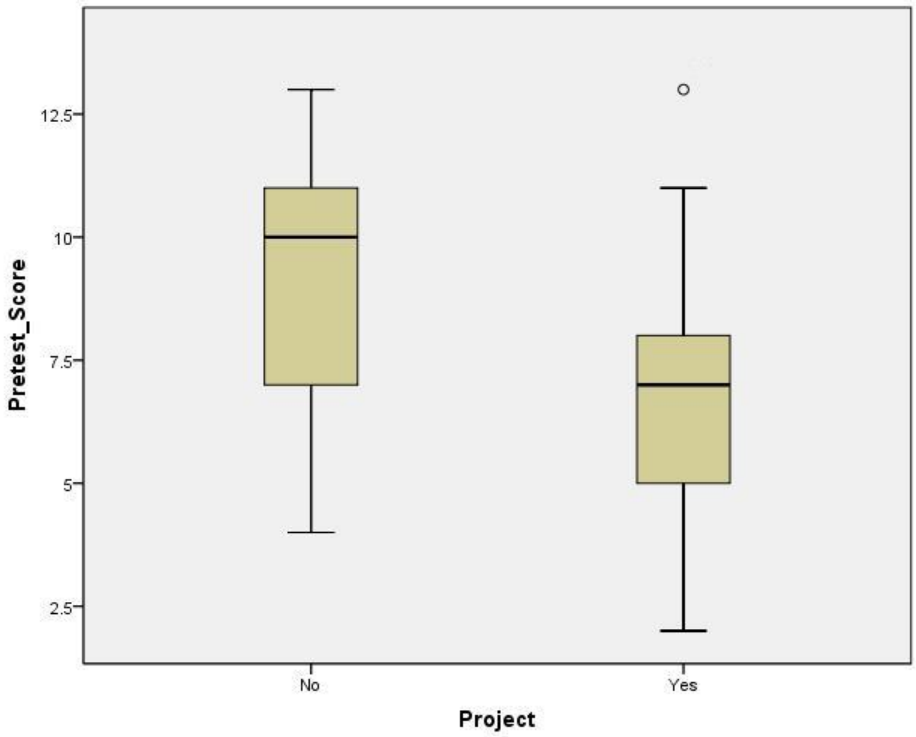


Figure 5: Side-by-Side Boxplots of Pre-test Scores for STAT 1000

		Class Level			
		Freshman	Sophomore	Junior	Senior
Project	No	27.7%	46.8%	17.0%	8.5%
	Yes	8.6%	32.9%	40.0%	18.6%

Table 18: Percentages of Class Level for Project and Non-project Groups

Pretest scores were also examined based on the experience of the student with statistics courses. On the pre-test, students were asked to select their prior experience with statistics courses from the following choices: no experience, high school statistics course or college statistics course. As can be seen in Figure 6, experience did not appear to affect the distribution of pre-test scores in either group. In this figure, A represents students with no

prior statistics experience, B represents students who have completed a high school statistics course and C represents students who have completed a college statistics course.

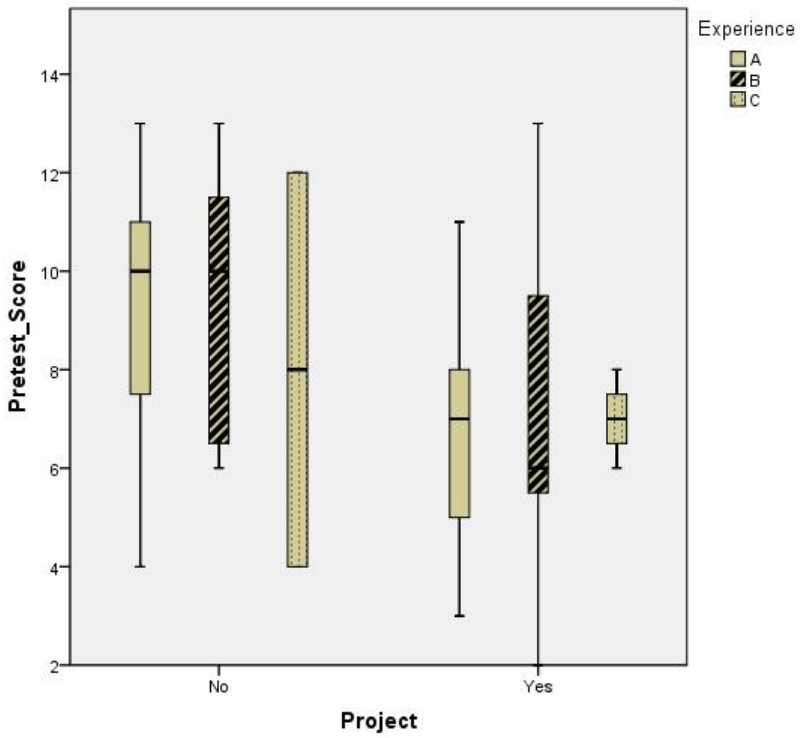


Figure 6: Side-by-Side Boxplots of Pre-test Scores Split by Experience Nested Within Project Groups

It is conjectured by the researcher that the different class level make-up of the groups could have led to the different pre-test performance. A higher number of upperclassmen suggetsions a lower number of students who have delayed taking the required introductory statistics course due to anxiety or difficulty with the subject. These students would then not perform as well as their younger counterparts. It is also likely that the span of time between the last similar subject for these students is larger than for a freshman or sophomore.

4.2.2 Model Analysis

The model used to analyze the data was an ANCOVA, which is represented in the following equation:

$$DIFF_{ij} = \mu_{\cdot} + \tau_i + \gamma PRETESTSCORE_{ij} + \epsilon_{ij} \tag{4.1}$$

where $\mu_{\cdot}$ is an overall mean, the τ_i are fixed treatment effects, γ is a regression coefficient for the relation between DIFF and PRETESTSCORE (assumed the same in both groups), and ϵ_{ij} are independent $Normal(0, \sigma^2)$ random errors.

The covariate used was the score on the pre-test and the factor was project (presence or absence). Before performing this analysis, a scatterplot of pre-test scores versus difference in pre-test and post-test scores was examined. It was determined that there were outliers existing at the pre-test scores of three and below. For this reason, observations with a pre-test score of three or below were excluded from the model analysis.

Also, observations with pre-test scores above twelve were excluded due to the problem caused by a restriction in range, since post-test scores could at most be 16 and thus post-test minus pre-test differences could be at most 4. Excluding these observations resulted in 65 observations (92.9% of students completing both assessments) from the project group and 38 observations from the non-project group (80.6% of students completing both assessments) to be used in the analysis.

Before running the model, it is necessary to determine if the slopes of the covariate for each group are parallel. As can be seen in Figure 7, the slopes for the covariate are roughly parallel. Testing for equality of slopes resulted in a p-value of .455, suggesting that the slopes are indeed parallel.

Also, year level (freshman, sophomore, junior, senior) was considered as a covariate in the model. However, year level was not significant.

It can also be seen in Figure 7 that the regression lines are parallel but distinct, with the estimated line for the project group above the estimated line for the non-project group, an indication that the project group produced, on average, higher differences for each pre-test score than the non-project group.

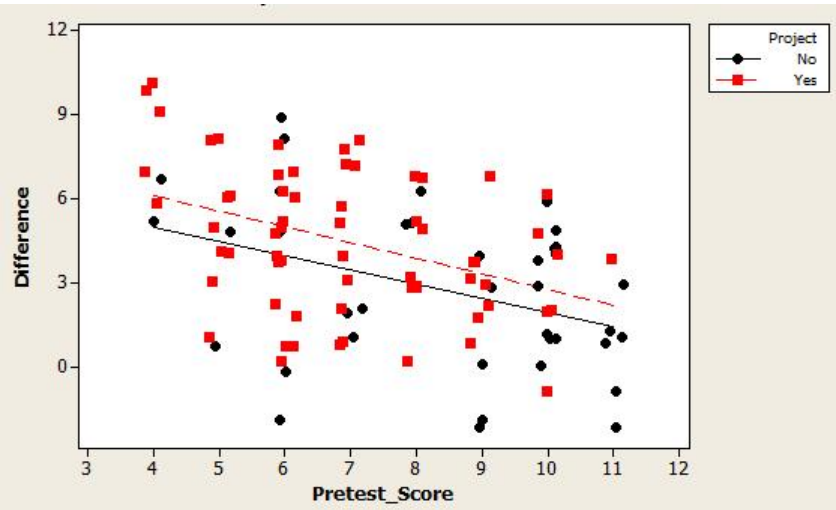


Figure 7: Scatterplot of Pre-test Scores vs Difference Scores for each Group at the STAT 1000 Level

Average and median differences for each group can be seen in Table 19. As can be seen in Figure 7, the difference between the unadjusted means is approximately 2.

Group	Mean	Standard Error	Median
Project	4.46	.317	4
Non-project	2.74	.468	3

Table 19: Mean and Median Scores for STAT 1000 Difference of Scores

The adjusted means from running the ANCOVA are shown in Table 20 and were calculated using a pre-test covariate mean of 7.5. The p-value for the teaching method factor of .083. The one-sided test comparing the project group's adjusted mean score (4.174) and the non-project group's adjusted mean score (3.229) indicated that these two means were significantly different with a p-value of approximately .0415. These results indicate that the project group had a significantly higher increase in their mean scores than did the non-project group.

Finally, it is noted that an analysis was performed without restricting the data to certain

Group	Mean	Standard Error
Project	4.174	.315
Non-project	3.229	.419

Table 20: Adjusted Means for STAT 1000 Difference of Scores

pre-test scores. This analysis provided similar results to the analysis performed above. In this case, the project group again had a higher adjusted mean than the non-project group.

4.2.3 Subtest Analysis

ANCOVA models were also used to analyze the three subtests that made up the overall assessment and to determine more specifically where differences between groups existed. The ANCOVA model used was similar to the ANCOVA model used to analyze the assessment overall. The response variable for these models was the difference in scores on the subtest only. The covariate was the pre-test score for the subtest only and the factor was teaching method (project or traditional). For consistency with the analysis of the overall assessment, only observations not already excluded based on pre-test score were used.

Adjusted means for each group and each subtest are shown in Table 21. This table also includes 95% confidence intervals for the difference of each set of means. The adjusted means for the displays and descriptives subtest were calculated using a pre-test subtest score of 4.25. The p-value from the ANCOVA for project when this subtest was examined was 0. Also, the one-sided test for equality of mean of subtest scores produced a p-value of approximately 0. This indicates that there are significant differences between means of the two groups for this particular subtest. For this subtest, the project group outperformed the non-project group.

For the classification subtest, the adjusted means were calculated using a pre-test subtest score of 2.50. The p-value from the ANCOVA model for the factor project was .011 and one-sided test for equality of mean subtest scores had a p-value of .0055. This indicates that there was a difference between the two means for the classification subtest. In this case, the

Display and Descriptive Subtest		
Group	Adjusted Mean	Standard Error
Project	2.651	.169
Non-project	1.150	.224
Classification Subtest		
Group	Adjusted Mean	Standard Error
Project	.625	.115
Non-project	1.114	.150
Test Selection Subtest		
Group	Adjusted Mean	Standard Error
Project	.789	.148
Non-project	1.150	.193

Table 21: Results for Subtest Analysis for STAT 1000

non-project group outperformed the project group.

The test selection subtest adjusted means were calculated using a pre-test subtest score of .75. The p-value from the ANCOVA model for project in this subtest was .142. This indicates that there are not significant differences between groups for the test selection subtest.

4.2.4 Individual Question Analysis

Data was tabulated for each individual question on the pre-test and post-test. Contingency tables were produced to show how the number of students who scored correctly or incorrectly on the pre-test later scored on the post-test for each question. An example of one of the contingency tables used can be found in Table 22. All participants with matched pre-test and post-test scores were used in this analysis since there was no problem of restriction on the range for this type of analysis. At the STAT 1000 level, the sample sizes for the individual question analysis were 70 for the project group and 47 for the non-project group.

		Post-test		
		Incorrect	Correct	Total
Pre-test	Incorrect			
	Correct			
	Total			

Table 22: Sample Contingency Table for Question-by-Question Analysis

Of particular interest from these tables were the "off-diagonal" entries. These entries show the number of students who either had an incorrect answer on the pre-test and a correct answer on the post-test or a correct answer on the pre-test and an incorrect answer on the post-test. The percentage of students who improved their scores from pre-test to post-test (went from incorrect on the pre-test to correct on the post-test) was calculated for both the project and non-project group. A summary of those percentages can be found in Table 23.

From Table 23, it can be seen that for eleven of the sixteen questions, the project group's improvement percentage was higher than the non-project group's improvement percentage. On six of the eight display and descriptive questions, the percentage for the improvement percentage for the project group was higher than the non-project group. This was also the case for three of the four classification questions and two of the four test selection questions.

Further testing was performed using the z-test for two proportions when appropriate and Fisher's exact test when sample sizes were too small for the Normal approximation. The p-values for questions where differences were significant are shown in Table 24. Significance was determined using a 5% comparison-wise signficance level. Using the Bonferroni correction, the significance level for each individual test was .003. In one of the three cases where significant differences were indicated (indicated in the table with a *), the project group's score was higher than the non-project group's score.

From Table 25 it can be seen that the project group had a lower percentage of students who answered questions incorrectly on the post-test after answering them correctly on the

Question Focus	Project Group	Non-project Group
Single Categorical Display	80.0%	35.0%
Categorical → Quantitative Display	98.0%	50.0%
Single Quantitative Descriptive	75.0%	85.7%
Single Categorical Descriptive	84.1%	78.9%
Quantitative → Quantitative Descriptive	70.8%	61.1%
Quantitative → Quantitative Display	78.4%	70.0%
Single Quantitative Display	77.1%	83.3%
Single Quantitative with Outlier Descriptive	72.5%	58.8%
Categorical → Categorical	100%	77.8%
Categorical → Quantitative	40.0%	81.0%
Quantitative → Categorical	84.6%	44.4%
Quantitative → Quantitative	100%	92.9%
Regression	27.5%	22.0%
χ^2	25.8%	56.8%
ANOVA	77.4%	75.0%
Paired t-test	25.5%	45.5%

Table 23: Summary of Percentages for Incorrect to Correct for STAT 1000

Question Focus	p-value
Categorical → Quantitative Display	.001*
Categorical → Quantitative	.003
χ^2	.001

Table 24: Incorrect to Correct Significant Differences and their p-values

Question Focus	Project Group	Non-project Group
Single Categorical Display	12.7%	40.7%
Categorical → Quantitative Display	5.0%	28.2%
Single Quantitative Descriptive	6.5%	7.5%
Single Categorical Descriptive	3.8%	17.9%
Quantitative → Quantitative Descriptive	4.3%	24.1%
Quantitative → Quantitative Display	12.1%	22.2%
Single Quantitative Display	13.6%	20.7%
Single Quantitative with Outlier Descriptive	36.8%	26.7%
Categorical → Categorical	9.3%	2.6%
Categorical → Quantitative	54.3%	0%
Quantitative → Categorical	12.9%	2.6%
Quantitative → Quantitative	15.4%	0%
Regression	0%	50.0%
χ^2	75.0%	0%
ANOVA	35.3%	39.1%
Paired t-test	69.6%	57.1%

Table 25: Summary of Percentages for Correct to Incorrect for STAT 1000

pre-test on nine of the sixteen questions. Seven of these questions were from the display and descriptive section of the test, and two were from the test selection portion of the test.

Testing was performed using a two-proportion z-test or Fisher's exact test to determine where differences between the groups were significant. The p-values for questions where differences were significant are shown in Table 26. Again, significance was determined using a 5% comparison-wise significance level, using the Bonferroni correction to adjust for multiple tests. There was one question for which a significance difference was indicated, as shown in Table 26. In this case, the project group's percentage was higher than the non-project group's percentage.

Question Focus	p-value
Categorical → Quantitative	0

Table 26: Correct to Incorrect with Significant Differences and their p-values

4.3 STAT 0200 RESULTS

4.3.1 Pretest Results

Students in all participating sections at the STAT 0200 level completed a pre-test assessment. Mean and median scores for each group at the STAT 0200 level are shown in Table 27. The distribution of pre-test scores for each group can be seen in Figure 8.

Group	Mean	Standard Error	Median
Project	5.65	.437	5
Big Picture	7.96	.163	8
Traditional	7.28	.324	7

Table 27: Mean and Median Scores for STAT 0200 Pre-test

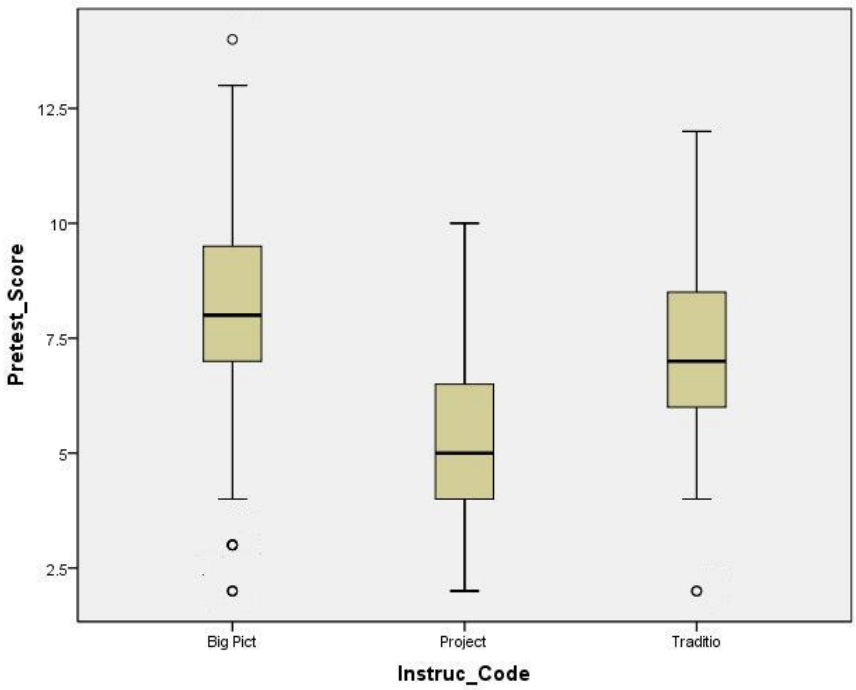


Figure 8: Side-by-Side Boxplots of Pre-test Scores for STAT 0200

As with the STAT 1000 level, pre-test scores for the project group were lower than pre-test scores for both other groups. An ANOVA and multiple pairwise comparisons using the Bonferroni correction were used to determine if significant differences existed between groups. The p-value for the ANOVA test for differences between means of each group was approximately 0, indicating significant differences between group means. Significant differences were shown to exist between the project and big picture groups and the project and traditional groups with comparison-wise p-values for the comparisons equaling 0 and .02, respectively. The comparisons showed no significant difference existing between the big picture and traditional groups.

Class year data for each group is shown in Table 28. The p-value from the χ^2 test for equality of distributions was .052. A further analysis showed significant differences between the project and traditional groups. There are fewer freshman in the project group, similarly to the STAT 1000 level and it is conjectured that this is the reason for the difference in

scores between these two groups. The percentages for the project group were roughly the same as the percentages for the big picture group. However, as was noted in Chapter 1, the project course was a night section course while the big picture groups were all day section courses. It is conjectured by the researcher that the time difference between the two courses is the reason for the differences in scores between these two groups.

Pre-test scores were examined based on the experience of the student. Again, scores did not significantly differ based on experience level for any group. Figure 9 shows the distribution of pre-test scores based on experience level nested within instructional method. In this figure, A represents no prior experience, B represents that the student has completed a statistics course at the high school level and C represents that the student has completed a statistics course at the college level.

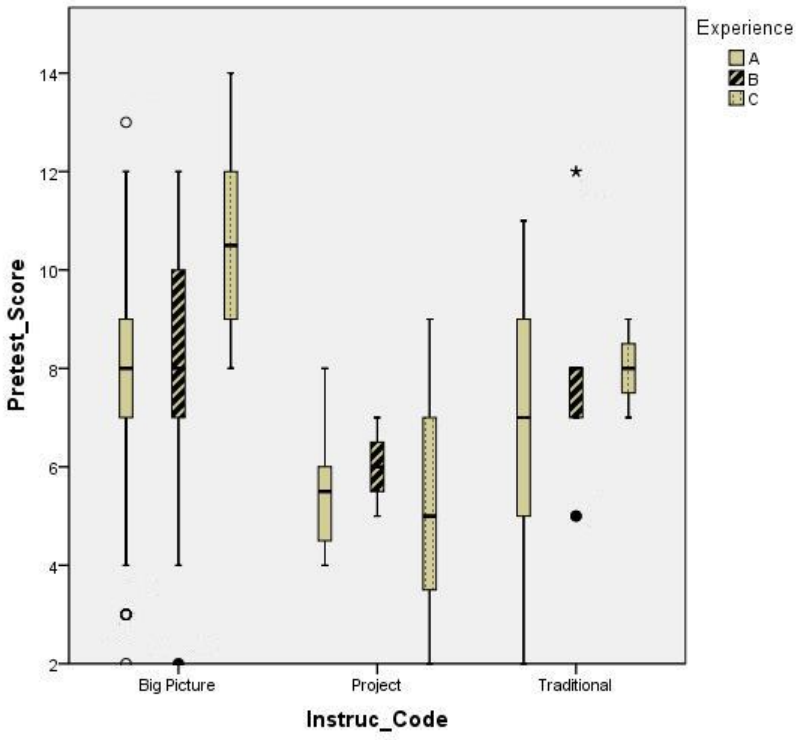


Figure 9: Side-by-Side Boxplots of Pre-test Scores Split by Experience Nested Within Instructional Method

		Class Level			
		Freshman	Sophomore	Junior	Senior
Method	Project	47.2%	36.1%	11.1%	5.6%
	Big Picture	48.7%	33.3%	12.0%	6.0%
	Traditional	58.1%	12.9%	16.1%	12.9%

Table 28: Percentages of Class Level for Project and Non-project Groups

4.3.2 Model Analysis

4.3.2.1 Project and Traditional Groups Only Analysis As with the STAT 1000 course type, an ANCOVA model was used to analyze the STAT 0200 data for the overall assessment to compare the project and traditional teaching methods. The ANCOVA model for this comparison was identical to the model used to compare groups for STAT 1000.

Prior to running the ANCOVA, a scatterplot was generated to determine if the lines relating to the post-test minus pre-test scores to pre-test scores for these two instructional methods were parallel. The range was restricted to pretest scores below eleven because of the problem of restriction of range and also because there was no data for the project group above this pretest score level. The sample size for the project group given this restriction was 20 (100% of students completing both assessments) and the sample size for the traditional group given the restriction was 44 (93.6% of students completing both assessments).

As can be seen in Figure 10, the slopes are not parallel to each other. However, the test for equality of slopes produced a p-value of .179, indicating that the difference between the two slopes is not significant. Therefore, an interaction term was not considered in the ANCOVA model.

Again, year level was considered as a covariate, but was shown not to be significant.

Table 29 shows the unadjusted means and medians for each group's difference in pre-test and post-test scores. These means show that the project group scored over four points better than the traditional group. Table 30 shows the adjusted means found from the ANCOVA

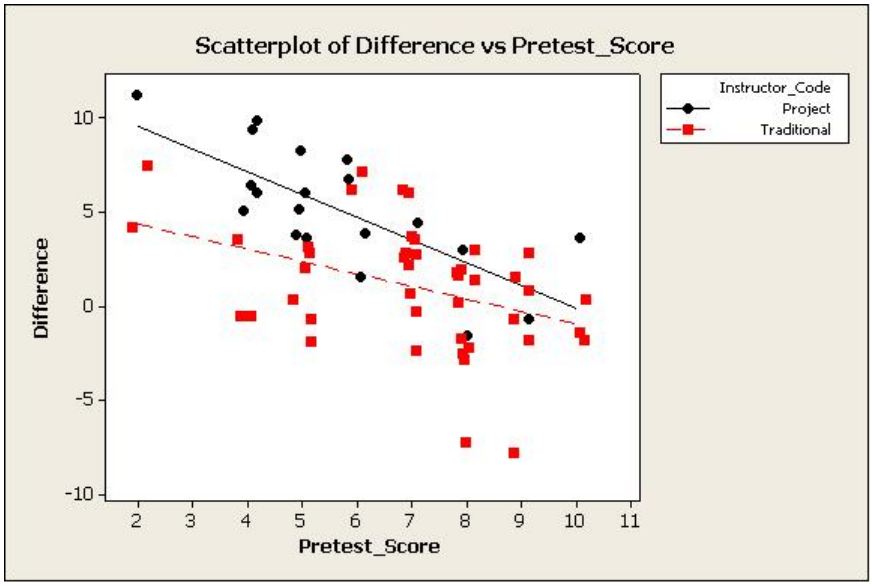


Figure 10: Scatterplot of Pre-test Scores vs Difference Scores for Project and Traditional Groups

model using a pre-test score of 6.56. As with the unadjusted means, the project group's adjusted mean is higher than the traditional group's mean.

Group	Mean	Standard Error	Median
Project	5.15	.734	5
Traditional	1.05	.499	1.5

Table 29: Mean and Median Scores for Project and Traditional STAT 0200 Groups

The ANCOVA model produced a p-value of .001 for the factor (teaching method). The p-value for the one-sided comparison of mean difference scores produced a p-value of .0005. This indicates that there were significant differences between the two groups, with the project group scoring better than the traditional group.

Again, an analysis was performed for the data with no restrictions on pre-test score. Again, results matched the restricted results discussed above, with a significant difference indicated and the project group scoring higher than the traditional group.

Group	Mean	Standard Error
Project	4.394	.671
Traditional	1.389	.444

Table 30: Adjusted Means for Project and Traditional STAT 0200 Groups

4.3.2.2 Project, Big Picture and Traditional Group Analysis As in the previous two-group analysis, an ANCOVA model was used to analyze the data for the overall assessment. In this case the factor was teaching method and had three levels: project, big picture and traditional.

Prior to running the ANCOVA, a scatterplot was generated to determine if the regression slopes for each instructional method were equal. As can be seen in the scatterplot shown in Figure 11, the slopes are not all equal to one another. Also, the estimated regression lines for project and big picture cross at a pre-test score of 8. This indicates that, particularly for lower pre-test scores, the project group's difference in scores is higher than both of the other two groups, while the big picture group's difference in scores is higher than the traditional method group's score. For pre-test scores of 8 and above, it appears that big picture and project methodologies perform similarly, with big picture performing slightly better and that both of these methods perform better than traditional methods. However, this is difficult to determine accurately due to a lack of data for the project group at high pre-test score levels. For this reason, the data will be split into separate ANCOVA analyses to determine if there are significant differences between groups for low pre-test scores and high pre-test scores. Low pre-test scores are defined as a score below 8, while high pre-test scores are considered to be 8 and above. The dataset involving the low pre-test scores had a sample size of 16 for the project group (80% of students completing both assessments), 75 for the big picture group (37.7% of students completing both assessments) and 24 for the traditional group (51% of students completing both assessments).

A scatterplot of pre-test scores below 8 versus differences is shown in Figure 12. As can be seen in the figure, the regresion slopes are not entirely equal, but each estimated regression

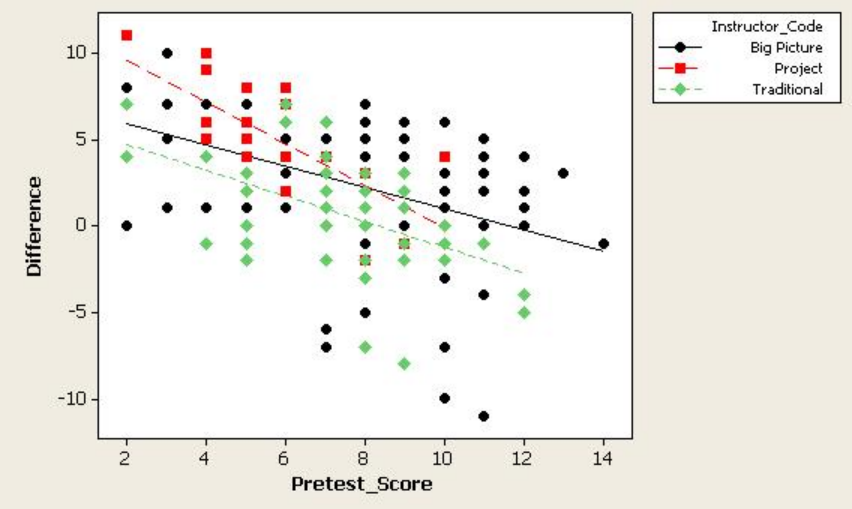


Figure 11: Scatterplot of Pre-test Scores vs Difference Scores for each Group at the STAT 0200 Level

line is distinct from one another with the project group having the highest estimated line and the traditional group having the lowest estimated line. While the slopes are not equal, a model including a linear interaction term was run and the p-value for the interaction term was .140 which was not significant. Hence, the interaction term was dropped from the model. Again, year level was also determined not to be significant, so it was not used in the model.

Table 31 shows the medians and unadjusted means for each group's difference in pre-test and post-test scores. These unadjusted means show that the project group had the highest mean difference score, scoring about 2.4 points on average above the big picture group and 3.7 points on average above the traditional group. Also, the big picture group scored approximately 1.3 points better than the traditional group. Adjusted means resulting from running the ANCOVA model are shown in Table 32. These means were calculated using a pre-test score of 5.53. As seen in the table, the project group's mean difference is still the highest of the three groups and the traditional group's mean difference is the lowest.

The ANCOVA model also produced a p-value of .001 for the teaching method factor, indicating that there were significant differences between teaching methods. The intersection-union test was used to determine if project scores were significantly higher than big picture

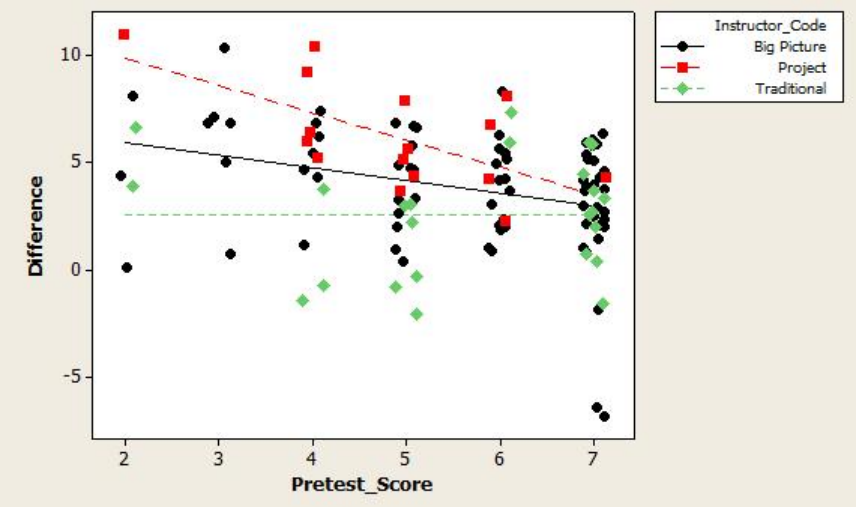


Figure 12: Scatterplot of Pre-test Scores below 8 vs Difference Scores for each Group at the STAT 0200 Level

Group	Mean	Standard Error	Median
Project	6.19	.621	6
Big Picture	3.77	.327	4
Traditional	2.54	.571	3

Table 31: Mean and Median Scores for STAT 0200 Difference of Scores for Low Pre-test Scores

Group	Mean	Standard Error
Project	5.853	.682
Big Picture	3.829	.311
Traditional	2.590	.549

Table 32: Adjusted Means for STAT 0200 Difference of Scores for Low Pre-test Scores

scores and traditional method scores. Two t-tests were performed comparing the adjusted means for each group. The t-test comparing project and big picture groups produced a t-statistic of 11.62 with a corresponding p-value of approximately 0 and the t-test comparing project and traditional groups produced a t-value of 15.99 again with a corresponding p-value of approximately 0. Following the procedure described by Laska and Meisner [48], the p-value for the first t-test was selected and compared to the 5% significance level. Using this method, the null hypothesis shown in Table 12 can be rejected at the 5% level.

It was also noted that there were possible outlying observations at the pre-test score of 7. In these 3 cases, the post-test scores were substantially lower than the pre-test scores, with post-test scores showing only one correct answer. Removing these observerations did not change the outcome of the ANCOVA analysis, which produced a factor p-value of 0. The intersection union test to compare the mean difference score for the traditional mean (2.590), the big picture mean (4.094) and the project mean (5.937) produced a p-value of 0, indicating that significant positive differences existed between the project group and both control groups.

Although comparison between big picture and traditional methods was not the main focus of this research, a separate one-side t-test comparing the adjusted means for each group was used to compare these methods at the low pre-test score level. This t-test resulted in a t-statistic of 10.53 and a corresponding p-value of approximately 0, indicating that there were significant differences between these groups with the big picture group performing better than the traditional group on average.

An analysis of unrestricted data was also performed to determine if results were different when restricted data was used. In this case, again, the teaching factor was indicated to be significant with the project group scoring the highest out of the three groups.

Analysis for high pre-test scores was limited to a comparison of big picture methodology and traditional methodology due to a small sample size for the project group. The ANCOVA that was run for the high pre-test data showed significant differences between the two instructional methods, with a p-value of approximately 0. The test for difference in the adjusted mean of the big picture group (1.304) and the adjusted mean of the traditional group (-1.249) showed a significant difference between the groups with a p-value of approximately

0. This shows that the big picture group performed better than the traditional group.

4.3.3 Subtest Analysis

4.3.3.1 Project and Traditional Groups Only Analysis A subtest analysis similar to the analysis performed on STAT 1000 data was also completed for the STAT 0200 data. As with the STAT 1000 data, the response variable for the ANCOVA model was the difference in scores for the subtest, the covariate was the pre-test score for the subtest and the factor was teaching method, with two factor levels: project and traditional. For consistency, only scores included in the overall assessment analysis were used for subtest analysis.

Adjusted means for each subtest can be found in Table 33. The adjusted means for the displays and descriptives subtest were calculated using a pre-test subtest score of 3.70. The p-value for the factor from the ANCOVA was .427, indicating that there were no significant differences between the two groups for this subtest. The adjusted means for the classification subtest were calculated using a pre-test subtest score of 2.30. The p-value for the factor from the ANCOVA was .107, again, indicating that there were no significant differences between the two groups for the classifications subtest. The adjusted means for the test selection subtest were calculated using a pre-test subtest score of .56. The p-value for the factor from the ANCOVA was approximately 0, indicating that there were significant differences between the two groups for this subtest. The one-sided test comparing the two means also produced a p-value of approximately 0, indicating that the project group performed better than the traditional group on this portion of the assessment.

4.3.3.2 Project, Big Picture and Traditional Group Analysis A subtest analysis was also performed for all three groups at the STAT 0200 level. As with other subtest analyses, the response variable for the ANCOVA models was difference in pre-test and post-test scores, the covariate was pre-test score and the factor was teaching method. In this case, the factor had three levels: project, big picture and traditional.

Intersection-union tests were also used to determine if the project group had significantly higher differences than both other groups. For consistency with the overall assessment

Display and Descriptive Subtest		
Group	Adjusted Mean	Standard Error
Project	1.131	.384
Traditional	.759	.258
Classification Subtest		
Group	Adjusted Mean	Standard Error
Project	.907	.284
Traditional	.338	.187
Test Selection Subtest		
Group	Adjusted Mean	Standard Error
Project	2.387	.212
Traditional	.279	.143

Table 33: Results for Subtest Analysis for Project and Traditional Groups

analysis performed in the previous section, scores were separated into high and low pre-test scores. Since the focus of this research is on project performance, only low pre-test scores were considered for the subtest anaylsis.

Adjusted means for each subtest can be found in Table 34. The adjusted means for the displays and descriptives subtest were calculated using a pre-test subscore of 2.88. The p-value from the ANCOVA for the factor teaching method was .390, indicating that there were not signficant differences between groups. An intersection-union test was again used to determine if the project group scored better than both the big picture and traditional groups. The hypothesis tested was similar to the hypothesis shown in Table 12 with the means applying only to the subtest score. The individual t-tests produced a t-statistic of -2.26 with corresponding p-value of .981 for the comparison of project and big picture and a t-statistic of 2.08 with a corresponding p-value of .024 for the comparison of project and traditional groups. Using the methodology for the min-test, the maximum p-value was .981, indicating that the project group did not score better than both the big picture group and

the traditional group. As seen by the individual tests and the adjusted means, the project group scored better than the traditional group but not better than the big picture group.

Display and Descriptive Subtest		
Group	Adjusted Mean	Standard Error
Project	1.910	.425
Big Picture	2.156	.196
Traditional	1.645	.347
Classification Subtest		
Group	Adjusted Mean	Standard Error
Project	.947	.261
Big Picture	.899	.115
Traditional	.600	.205
Test Selection Subtest		
Group	Adjusted Mean	Standard Error
Project	2.712	.224
Big Picture	.815	.103
Traditional	.473	.182

Table 34: Results for Subtest Analysis for STAT 0200

The adjusted means for the classification subtest were calculated using a pre-test subscore of 2.17. The p-value from the ANCOVA for the factor was .410, again indicating that signficant differences did not exist between teaching methods. The intersection-union test for the comparison of the project group with the big picture and traditional groups was not significant, again, indicating that the project group did not score signficantly better than both other groups. However, individual test results and adjusted means suggest that the project group did score significantly better than the traditional group but not significantly better than the big picture group.

Finally, the adjusted means for the test selection subtest were calculated using a pre-test subscore of .48. The p-value for the factor from the ANCOVA model, which was approxi-

mately 0, indicated significant differences between the groups. Individual t-tests for comparing the project group to each other group gave the following: the t-statistic for comparing project and big picture groups was 33.14 with corresponding p-value of approximately 0 and the t-statistic for comparing the project and traditional groups was 33.32 with a corresponding p-value of approximately 0. Using the min-test procedure, the null hypothesis can be rejected, indicating that the project group scored significantly better than the other two groups.

4.3.4 Individual Question Analysis

4.3.4.1 Project and Traditional Groups Only Analysis As with the STAT 1000 level, individual question analyses were performed to determine if there were significant differences in the "off-diagonal" entries of Table 22 between the project and traditional groups. Sample sizes for this portion of the analysis were not restricted, resulting in a sample size of 20 for the project group and 47 for the traditonal group.

Percentages of students who improved their score from pre-test to post-test by correctly answering the matched question on the post-test after answering it incorrectly on the pre-test along with percentages of students who answered a question correctly on the pre-test and then answered incorrectly on the post-test were calculated for each instructional method. The results of these calculations are shown in Table 35 and Table 37.

From Table 35, it can be seen that for thirteen of the sixteen questions, the project group's improvement percentage was higher than the traditional group's percentage. Of these thirteen questions, six were from the display and descriptives subtest, three were from the classification subtest and the remaining four were from the test selection subtest.

Analysis to determine where the percentages differed significantly was performed using Fisher's exact test due to small sample sizes. The results of these analyses can be found in Table 36. Significance was determined using a 5% comparison-wise error. Using the Bonferroni correction, the appropriate significance level for comparison with individual p-values is .003. There were four questions for which significant differences were indicated. In three of the four cases, the project group had the higher percentage. These are indicated in

Question Focus	Project Group	Traditional
Single Categorical Display	100%	14.3%
Categorical → Quantitative Display	0%	51.6%
Single Quantitative Descriptive	77.8%	64.3%
Single Categorical Descriptive	10%	45%
Quantitative → Quantitative Descriptive	85.7%	60%
Quantitative → Quantitative Display	77.8%	42.4%
Single Quantitative Display	50.0%	35%
Single Quantitative with Outlier Descriptive	82.4%	52.2%
Categorical → Categorical	100%	66.7%
Categorical → Quantitative	33.3%	40%
Quantitative → Categorical	83.3%	55.6%
Quantitative → Quantitative	76.9%	75%
Regression	78.9%	8.9%
χ^2	52.6%	4.7%
ANOVA	66.7%	33.3%
Paired t-test	93.3%	34.2%

Table 35: Summary of Percentages for Incorrect to Correct for STAT 0200 for Project and Traditional Groups

the Table 36 with a *.

Table 37 shows that the project group had a lower percentage of students who answered correctly on the pre-test then incorrectly on the post-test for ten of the sixteen questions. Of these ten questions, four were from the display and descriptives subtest, two were from the classification subtest and four were from the test selection subtest. Table 38 shows that there was only one question for which a signficant difference was indicated. For this question, the project group's percentage was higher than the traditional group's percentage.

Question Focus	p-value
Categorical → Quantitative Display	.002
Regression	0*
χ^2	0*
Paired t	0*

Table 36: Incorrect to Correct Significant Differences and their p-values for Project and Traditional Groups

4.3.4.2 Project, Big Picture and Traditional Group Analysis Individual question analyses were also performed to determine if there were significant differences in the "off-diagonal" entries of Table 22 between project and both non-project groups. In this case, all sections of controls were combined into the non-project. All participants with matched pre-test and post-tests were used for the individual question analysis. Sample sizes for this portion of the analysis were as follows: project group, 20, big picture group, 199, traditional group, 47.

Again, percentages of students who improved their score from pre-test to post-test by correctly answering the matched question on the post-test after answering it incorrectly on the pre-test along with percentages of students who answered a question correctly on the pre-test and then answered incorrectly on the post-test were calculated for each instructional method. The results of these calculations are shown in Table 39 and Table 41.

From Table 39, it can be seen that for ten of the sixteen questions, the project group's improvement percentage was higher than the non-project group's percentage. Of these ten questions, there were four questions from the eight display and descriptives questions for which the project group scored the highest and two from the four classification questions for which the project group scored the highest. The remaining four questions for which the project group scored highest made up the entire test selection portion of the assessments. For each of the questions where the project group's percentage was higher than the non-project group's percentage, it was determined that the project group had the highest percentage

Question Focus	Project Group	Traditional
Single Categorical Display	0%	35%
Categorical → Quantitative Display	100%	37.5%
Single Quantitative Descriptive	0%	18.2%
Single Categorical Descriptive	80.0%	28.6%
Quantitative → Quantitative Descriptive	38.5%	44.4%
Quantitative → Quantitative Display	0%	28.6%
Single Quantitative Display	75.0%	51.9%
Single Quantitative with Outlier Descriptive	66.7%	29.2%
Categorical → Categorical	20%	8.6%
Categorical → Quantitative	12.5%	40.9%
Quantitative → Categorical	25%	20.7%
Quantitative → Quantitative	0%	33.3%
Regression	0%	100%
χ^2	0%	50%
ANOVA	20.0%	64.3%
Paired t-test	20.0%	88.9%

Table 37: Summary of Percentages for Correct to Incorrect for STAT 0200 for Project and Traditional Groups

Question Focus	p-value
Categorical → Quantitative Display	.002

Table 38: Correct to Incorrect with Significant Differences and their p-values

of the three groups when compared separately. For all questions for which the non-project group's percentage was higher, it was determined that the big picture group's percentage

Question Focus	Project Group	Non-project Group
Single Categorical Display	100%	52.7%
Categorical → Quantitative Display	0%	69.4%
Single Quantitative Descriptive	77.8%	82.7%
Single Categorical Descriptive	10%	55.3%
Quantitative → Quantitative Descriptive	85.7%	64.4%
Quantitative → Quantitative Display	77.8%	63.0%
Single Quantitative Display	50.0%	50.7%
Single Quantitative with Outlier Descriptive	82.4%	47.8%
Categorical → Categorical	100%	76.7%
Categorical → Quantitative	33.3%	58.2%
Quantitative → Categorical	83.3%	72.4%
Quantitative → Quantitative	76.9%	85%
Regression	78.9%	24.8%
χ^2	52.6%	17.2%
ANOVA	66.7%	40.9%
Paired t-test	93.3%	53.9%

Table 39: Summary of Percentages for Incorrect to Correct for STAT 0200

was higher than both the project and traditional groups percentages.

Analyses to determine if percentages differed significantly between the groups was performed using a two proportion z-test or Fisher's exact test when sample sizes were small. Results of these analyses can be found in Table 40. Significance was determined using a 5% comparison-wise significance level. Again, a Bonferroni correction was used leading to a significance level for individual p-values of .003. In three of the four cases where significant differences were indicated, the project group's score was higher than the non-project group's score. These differences are indicated in the table with a *.

Question Focus	p-value
Categorical → Quantitative Display	0
Regression	0*
χ^2	.001*
Paired t-test	.002*

Table 40: Incorrect to Correct Significant Differences and their p-values

Table 41 shows that the project group had a lower percentage of students who answered correctly on the pre-test then incorrectly on the post-test for nine of the sixteen questions. Three of these questions were from the display and descriptive subtest, two were from the classification subtest and the remaining four were from the test selection subtest. In each case where the project percentage was lower, it was the lowest of the three groups when groups were compared separately. In the remaining seven cases where the non-project group was lower, the big picture group achieved the lowest percentage for five of the questions and the traditional group was the lowest in two cases.

Testing was again performed using either a two-proportion z-test or Fisher's exact test to determine where differences between the groups were significant. The p-values for questions where differences were significant are shown in Table 42. Again, significance was determined using a 5% comparison-wise significance level. Again, the Bonferroni correction was used to account for multiple tests. In the three questions for which significance differences were indicated, the project had a lower percentage of students who incorrectly answered a question on the post-test after correctly answering the similar question on the pre-test on one question. This question is indicated with * in the table.

Question Focus	Project Group	Non-project Group
Single Categorical Display	0%	34.0%
Categorical → Quantitative Display	100%	32.8%
Single Quantitative Descriptive	0%	10.9%
Single Categorical Descriptive	80.0%	28.5%
Quantitative → Quantitative Descriptive	38.5%	35.9%
Quantitative → Quantitative Display	0%	20.0%
Single Quantitative Display	75.0%	39.8%
Single Quantitative with Outlier Descriptive	66.7%	41.6%
Categorical → Categorical	20%	11.6 %
Categorical → Quantitative	12.5%	30.6 %
Quantitative → Categorical	25%	14.3%
Quantitative → Quantitative	0%	18.3%
Regression	0%	75.0%
χ^2	0%	64.0%
ANOVA	20.0%	56.0%
Paired t-test	20.0%	43.0%

Table 41: Summary of Percentages for Correct to Incorrect for STAT 0200

Question Focus	p-value
Single Categorical Display	.002*
Categorical → Quantitative Display	0
Single Categorical Descriptive	.002

Table 42: Correct to Incorrect with Significant Differences and their p-values

4.4 COMPARISON OF STAT 1000 AND STAT 0200

Since analyses of the data were performed separately for each introductory level course, it is of interest to compare these analyses to determine if project performance was similar for both levels. In both course levels, pre-test scores for the project group were significantly lower than the pre-test scores for the non-project groups. This anomaly in the data may be explained by the differing percentages of class levels in each section at the STAT 1000 level. At the STAT 0200 level, differing percentages of class levels within each section may account for the differences in pre-test scores between the traditional and project groups. Differences between project and big picture groups may be accounted for by the differences in class times, with the project group being a night course and the big picture group being a day course. With pre-test scores being significantly different between project and non-project groups, it is clear that a random or quasi-random assignment into groups was not attained by the study. This is to be expected, however, since the research could not randomly assign students into particular courses, but rather, students selected courses based on their schedules and personal preferences.

Analysis of overall differences in pre-test assessment scores revealed a similar pattern between both course levels. Project groups appeared to be significantly different from their non-project counterparts, particularly for lower pre-test scores. This trend occured in both analyses performed at the STAT 200 level and in the overall analysis performed at the STAT 1000 level. However, in the subtest analysis the two class levels did not perform similarly. At the STAT 1000 level, the project group scored significantly better than the non-project group on the displays and descriptives subtest but for the STAT 0200 level, the project group scored significantly better than its non-project counterparts on the test selection subtest.

In the individual question analyses, project sections had a higher percentage of improvement for eleven of the sixteen questions at the STAT 1000 level and for ten of the sixteen questions for the STAT 0200 level. For incorrect to correct, the project sections matched results for eleven of the sixteen questions. That is, both project groups had the highest improvement percentage or both project groups did not outscore their non-project counterparts. For the percentage of students who answered correctly on the pre-test then incorrectly

Question Focus	Project Groups Matched on Incorrect to Correct	Project Groups Matched on Correct to Incorrect
Single Categorical Display	Yes	Yes
Categorical → Quantitative Display	No	No
Single Quantitative Descriptive	Yes	Yes
Single Categorical Descriptive	No	No
Quantitative → Quantitative Descriptive	Yes	No
Quantitative → Quantitative Display	Yes	Yes
Single Quantitative Display	Yes	No
Single Quantitative with Outlier Descriptive	Yes	Yes
Categorical → Categorical	Yes	Yes
Categorical → Quantitative	Yes	No
Quantitative → Categorical	Yes	Yes
Quantitative → Quantitative	No	No
Regression	Yes	Yes
χ^2	No	No
ANOVA	Yes	Yes
Paired t-test	No	Yes

Table 43: How Project Groups Matched Question-by-Question

on the post-test, the STAT 1000 level project group had the lower percentage for nine of the sixteen questions and the STAT 0200 project group also had the lowest percentage for nine of the sixteen questions. For nine of these sixteen questions, both project groups matched, meaning that either both project groups had the lowest percentage when compared with non-project groups or both project groups did not have the lowest percentage when com-

pared with non-project groups. A summary of these results can be found in Table 43. The comparison shown in Table 43 was performed using all three groups at the STAT 0200 level. The STAT 1000 versus the two-group STAT 0200 comparisons do not entirely agree with the results shown in Table 43.

5.0 DISCUSSION

Introductory statistics courses are designed to provide students with an overall understanding of techniques and methods used for simple statistical analysis and to provide students with a basis for future statistical coursework or simple research endeavors in their field of study. Research into statistics education, however, has provided a glimpse at what students truly gain from their introductory statistics course, which is, unfortunately, significantly less than what most instructors would hope.

Prior research has shown that students exiting introductory statistics courses have little to no data sense, an inability to appropriately classify variables and hence, an inability to select appropriate statistical techniques to analyze these variables. One particular reasoning for why this phenomenon occurs frequently is due to the contrived nature of many textbook examples and exercises. These exercises often significantly reduce the complexity of problems in an effort to make the material understandable and manageable. While these efforts are well-intentioned, these textbook problems rarely capture the true nature of real-world data, leaving students ill-prepared for the challenges they will face in a typical data analysis setting.

It has also been observed in introductory courses that students develop a "chapter mentality" when approaching statistical analysis. That is, students tend to determine the validity of their analytical approaches to problems based on the chapter in which the problem appears, rather than based on the situation described and the nature of the variables associated with the problem. This phenomenom is often well-evidenced on final exams, where students are asked and expected to be able to determine appropriate methods to analyze situations and often cannot correctly identify these methods.

The purpose of this study was to determine if requiring students to gather and guide an analysis of real-world data would allow them to develop more sophisticated statistical

reasoning skills than students who were not exposed to the project. In this chapter, the results of the study are explained, implications of these findings are discussed and future research directions are offered.

5.1 EXPLANATION OF RESULTS

5.1.1 Quantitative Results

The results at both the STAT 1000 and STAT 0200 level introductory courses show that projects better equip students to develop statistical reasoning, particularly for students having low pre-test scores. These results confirm prior opinions that the use of projects as an assessment and learning tool enhanced students' understanding of statistical concepts.

It is clear from the data provided in this study that students' increase from pre-test to post-test was significantly different overall for students participating in the project. These results were consistent across course types and instructors, providing strong evidence that the project produces effective classroom results. It is also evident that students participating in the project were better equipped to appropriately select statistical approaches in the presence of data that was closer to what one might see in the real-world (i.e. data containing outliers or skewedness). While subtest analysis reveals different areas where each project group excelled, overall differences in scores were higher by about two points on average for both STAT 1000 and STAT 0200.

Individual question analysis reveals that students involved in the project group at both introductory course levels had a better improvement percentage (incorrect to correct) on a majority of the questions when compared with students not completing a project. It also revealed that students in the project group also had a better retention percentage (correct to correct) than students in the non-project groups. This indicates, as suggested by prior research, that students completing projects not only better develop statistical understanding, but also that projects aid in reinforcing statistical topics explored in lectures and other out-of-class assignments.

Results of this study also provide a first comparison between project methodology and big picture methodology, a methodology which reminds students of how specific processes fit into the larger statistical background. These results indicate that, particularly for low pre-test scores, the project method provides greater learning gains than big picture methods. However, big picture's methodology has also been shown to perform significantly better than traditional methods. While the comparison between big picture and project methodologies is new to the literature, the results for the comparison between big picture and traditional methods reaffirm the big picture creator's work showing that this method aids student learning of the subject.

The quantitative results of this study show that projects do increase students' ability to reason statistically. These results are the first quantitative results from a designed comparative study to show the extent of learning gains made when a project is introduced to an introductory statistics course. These results reinforce claims made by other researchers about the ability of projects to better equip students with reasoning skills.

There are some caveats to this study that should be investigated further. First, the reliability analysis showed that reliability for particular questions was substantially lower than what is traditionally acceptable. After these questions are modified or removed from the the assessments, a further reliability analysis should be conducted to determine if performance of the two forms is more comparable. Also, as noted in Chapter 4, project groups outscored their non-project counterparts in separate subtest areas. These differences should be further investigated in order to determine where project groups consistently score differently than non-project groups. Finally, the data showed that both project groups scored significantly lower than their non-project counterparts on the pre-test assessment. Although pre-test and post-test differences were consistently higher for both project groups than for non-project groups, post-test scores were essentially equivalent with the exception of the STAT 0200 level, where traditional post-test scores were approximately two points lower on average than post-test scores for big picture and project groups.

"We liked this project because we got to choose a topic that we enjoyed and learned statistics about it at the same time."
"This project was beneficial and interesting because we were able to complete a larger analysis on a real-life topic..."
"...was much more entertaining and engaging than completing homework problems in a textbook..."
"The project was fun rather than a chore."
"This project inspired a creative approach to the practice of statistics."
"As a whole, we really enjoyed this project..."

Table 44: Comments Regarding Student Attitude Toward Projects

5.1.2 Qualitative Results

Prior research has also indicated that the use of projects in statistics courses makes statistics more approachable and enjoyable to students. This research corroborates this claim as well. Students who participate in the project course were asked periodically to provide feedback on project assignments and, in particular, how these assignments affected their learning.

Overall, students' reactions to the project were positive, as can be seen in Table 44 outlining a sample of comments regarding attitudes to the project. The only negative comments received related to clarifying assignment write-ups that may have been too vague for students to understand what was being asked of them. However, even these comments were not numerous.

Students also felt that the project greatly enhanced their understanding of statistics. Some of the comments given on the project by students are shown in Table 45. No students indicated that the project failed to aid in their learning of the material, nor did students indicate that the projects were overwhelming.

As a whole, the project component of the course was well-received by students, as was indicated in previous research. Many students completed more than the required data anal-

"Rather than blindly plugging numbers into equations, this project helped us grasp the meaning behind the math."
"It [the project] was a very effective way to learn the material and allowed us to investigate something actually of interest."
"This project ... allowed us to experience the practical application of statistics. It also made much of the course material less abstract..."
"Overall, seeing a project through from data collection to statistical analysis has been an invaluable learning experience..."
"Applying what we learned in class to this project helped us to see how we might be able to use statistics in the future for other classes or our careers."

Table 45: Comments Regarding Student Thoughts How Projects Aided in Learning

ysis on each assignment, simply because the data was interesting to them. Students also commented throughout project assignments that techniques may not be entirely appropriate due to violations of assumptions or simplicity of the approach to the complex data. Statements of this nature indicate that students are connecting ideas and viewing problems from a larger, global perspective rather than focusing entirely on completing a specified technique.

Using data that was of interest to students also provided a crucial benefit. Students who are supplied datasets often complain that they have no idea what questions to ask or how to connect variables. However, since the topic was selected (and assumedly of interest) to the students, they could easily generate questions and determine which variables were most appropriate to use in answering their questions.

While not required to do so, students began generating questions and connecting variables early into their project's run. It was often the case that, rather than provide separate graphical and numerical displays of each variable in their dataset, groups provided displays linking two or more variables together in order to best begin analysis for specific questions they had about the data. These results not only show students' interest in the data, but also evidenced that fact that they are connecting their analyses back to the global situation.

Prior research also indicates that students seem to lack a global understanding of problems that they work with in their introductory statistics courses. Students often fail to remember contextual details which may alter or change statistical methods being used and they fail to relate results obtained from these procedures back to the problem's context. Often, in introductory courses and even beyond, instructors find that students do not answer a question by describing the results of their analysis in relation to the problem context, but rather that students merely list a p-value and state "accept" or "reject". Because students participating in project sections were required to structure project assignments in written form, they gained exposure to technical writing. The written nature of the assignments required that students provide background on the project topic, complete statistical analyses and include results appropriately and then discuss these results in the context of their topic. This formatting required that students be able to relate their results back to the context of their topic, furthering students ability to connect ideas and think globally about a problem. Because these students were accustomed to relating results to context, they should be better prepared to articulate statistical knowledge overall.

In summary, the project component was a popular addition to the introductory statistics course sections in which it was used. Students conveyed that the project benefited them by providing them with interesting data to reinforce concepts explored in lectures and homework assignments. Because the project required that students gather real data, the assignments they worked through required them to examine the data carefully and determine solutions that would work both with the type of data they had as well as to answer their questions. Qualitatively, the project seemed to be a worthwhile learning tool for students.

5.2 IMPLICATIONS OF RESULTS

The results of this study provide several implications for the field of statistics education. First, it provides the field with a framework for projects in an introductory statistics course. The project used for the purpose of this research was designed to incorporate major aspects included in most introductory statistics courses. It was also designed to be a semester-long

project, incorporating assignments periodically throughout a sixteen week semester so that students have reinforcement for topics throughout the semester.

This project was designed not only for the purposes of this research but also as a template for instructors. The template can be used with or without modification for any standard introductory statistics course.

This study also provides assessment material designed to analyze areas where students completing the project should excel. The assessments are designed for straightforward grading using multiple choice questions only. However, these questions can also be easily modified into open-ended questions in order to further test students' reasoning and thinking abilities. Since there are two assessment forms, they can be used in a pre-test/post-test manner or they can be used as a single assessment with multiple forms.

A final implication of this study is the quantitative evidence it provides to corroborate qualitative claims made by other researchers as to the effectiveness of projects in the classroom. While projects have been suggested by many statistics educators, this study is the first to provide quantitative evidence documenting the effect that projects have on increasing reasoning abilities. It also provides information on the extent to which reasoning abilities are increased when compared with other more traditional teaching methods.

5.3 CONCLUSIONS AND FUTURE DIRECTIONS

The results of this study are significant for several reasons. This study provides the first quantitative designed comparative study linking the use of projects in introductory statistics courses to increases in reasoning abilities. In the study design, instructors selected the method that they would use for instruction. No instructors were assigned to use a particular method by the researcher. Also, there were no instructors who taught different sections using different methodologies, therefore there are no instructors for whom we can compare directly results without a teacher effect. Future research should expand upon this study in both size and design. In particular, this research should focus on comparing project and traditional methodologies when both of these practices are used by all of the instructors. Future research

may also include an analysis on change in attitude toward statistics and data analysis after completing the project. Also, further research may be performed to determine if project scores aid in predicting post-test scores.

This study also provides short assessment tools for analyzing the statistical reasoning abilities of students. A reliability analysis was performed on the parallel forms. Future studies may be needed to verify the reliability of the forms or to expand or modify the forms to further test statistical reasoning by adding additional assessment material or modifying from multiple-choice questions to open-ended questions. The modification to open-ended questions may allow for the development of questions that better assess contextual inference.

Finally, the study provides a template for introductory statistics course projects. This template was designed to provide all the tools necessary to include a project component in an introductory level course. The project was also designed to cover three major aspects of these courses: displaying and describing the data, probability and inference. Future studies should refine the project template by appropriately modifying project assignments to more effectively focus on developing reasoning skills. These studies should also compare reasoning abilities of students completing modified project designs to reasoning abilities of students completing the current project design and students not completing any projects.

In summary, this study has provided data concerning the differences in learning gains made by students completing a project versus students who do not complete a project in an introductory statistics course, along with other important contributions. This study is a firstsstep in developing and identifying instructional techniques that provide significantly different classroom results. It is the hope of the researcher that this study will encourage instructors to experiment with different teaching techniques in addition to traditional methodologies.

APPENDIX A

ASSESSMENT MATERIALS

A.1 ASSESSMENTS

On the following pages are a copies of the assessments that students completed. Students completed either Form A for a pre-test and Form B as a post-test or Form B as a pre-test and Form A as a post-test. Course sections were randomly assigned to complete a specific form for a pre-test. Note that the documents are exactly as they appeared to students, including pagination and formatting. Form A is listed first and Form B is listed following Form A.

1. A class survey asked students to indicate if they are MAC or PC users. Of the following graphs, which is most appropriate to display their results?

(a) Pie chart

(b) Histogram

(c) Scatterplot

(d) None of the above

2. The dean of a college would like to know if IQ scores differ for students on academic probation versus students who are not on academic probation. The data he collected is listed below. Of the following graphs, which is the most appropriate to display this data?

Probation GPAs	Non-Probation GPAs
3.2	3.4
2.0	3.3
2.5	3.2
3.0	3.5
2.8	3.0

(a) Scatterplot

(b) Histogram

(c) Two bar charts

(d) Side-by-side boxplots

3. A class survey asks students to indicate how long it takes them to travel from campus to home. Of the following, which is the most appropriate summary for the data?

(a) Mean and standard deviation

(b) Mode and range

(c) Proportions

(d) Correlation

CONTINUED

4. A class survey gathered the following data. Of the following, which is the most appropriate summary for the data?

Education Level of Parent	Number of Responses
Less than high school	4
High school graduate	12
Some college	7
Associate degree	3
Bachelor's degree	8
Advanced degree	2

(a) Mean and standard deviation

(b) Mode and range

(c) Proportions

(d) Correlation

Items 5 and 6 refer to the following situation: The admissions office at a small college would like to know if scores on an entrance exam help predict a student's performance in college (measured by their GPA after one semester). The results for a small sample are listed in the table below.

Student	Exam Score	GPA
1	76	3.0
2	88	3.2
3	65	2.5
4	74	2.6
5	56	2.7
6	92	3.4

5. What is an appropriate summary for the data to answer the admissions office's question?

(a) Mean and standard deviation

(b) Median and interquartile range

(c) Proportions

(d) Correlation

6. Of the following, which is the most appropriate way to display the results?

(a) Histogram

(b) Scatterplot

(c) Stem-and-leaf plot

(d) Side-by-side boxplots

CONTINUED

Items 7 and 8 refer to the following situation: A state police officer records the speed of several cars on a particular stretch of highway. The following is the dataset obtained, where Car Number indicates which car was observed and Speed indicates the speed of the car as recorded by a radar gun.

Car Number	Speed
1	60
2	62
3	56
4	60
5	66
6	59
7	61
8	119

7. Of the following, which is the most appropriate way to display the results?

(a) Boxplot

(b) Scatterplot

(c) Pie chart

(d) Side-by-side boxplots

8. What is an appropriate summary for this dataset?

(a) Mean and standard deviation

(b) Median and interquartile range

(c) Proportions

(d) Correlation

Items 9 through 12 refer to the following situation:

In studies of employment discrimination, several variables are often relevant: an employee's age, sex, race, years of experience, salary, whether promoted, and whether laid off. For each question, select the appropriate classification of explanatory and response variables.

9. Are men paid more than women?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

CONTINUED

10. Can an employee's age help us predict whether or not he or she will be laid off?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

11. For every additional year of experience, about how much higher is a worker's salary?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

12. Are whites more likely than blacks to be promoted?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

Items 13 through 16 refer to the following situation: For each research situation, decide what statistical procedure would most likely be used to answer the research question posed. Assume all assumptions have been met for using the procedure.

13. Do students' IQ scores help to predict how well they will perform on a test of science achievement?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

CONTINUED

14. Is ethnicity related to political party affiliation (Republican, Democrat, Other)?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

15. Are typical blood pressure readings the same for groups of patients who have been assigned to take one of four possible medications?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

16. In sets of boy-girl twins, do the boys differ from their sisters in reading achievement?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

---

1. An instructor records the IQ scores for each of her students. Of the following graphs, which is most appropriate to display the instructor's data?

(a) Pie chart

(b) Histogram

(c) Scatterplot

(d) None of the above

2. Researchers compared the ages of actors and actresses at the time they won Oscars. The results for recent winners from each category are listed in the table below. We want to use the data to decide which group—men or women—tends to have older Oscar winners. What type of graph should be used?

Women	21 24 26 26 26 27 28 30 30 31 31 33 33 34 34 34 34 35 35 35 37 37 38 39 41 41 41 42 44 49 50 60 61 61 74 80
Men	31 32 32 32 33 35 36 37 37 38 39 39 40 40 41 42 42 43 43 44 45 45 46 47 48 48 51 53 55 56 56 60 60 61 62 76

(a) Histogram

(b) Two pie charts

(c) Side-by-side boxplots

(d) Scatterplot

3. A particular customer service center records the length of all calls made to the center for one month. Of the following, which is the most appropriate summary for the data?

(a) Correlation

(b) Proportions

(c) Mode and range

(d) Mean and standard deviation

CONTINUED

4. A random sample of salaries at a certain company produced the following data. Given this data, what is the most appropriate summary?

Employee 1	50,000
Employee 2	54,000
Employee 3	49,000
Employee 4	52,000
Employee 5	50,000
Employee 6	104,000

(a) Mode and range

(b) Median and interquartile range

(c) Correlation

(d) Mean and standard deviation

Items 5 and 6 refer to the following situation: Researchers would like to determine if the amount of money spent on advertising directly affects sales. In the table below is data from 10 companies indicating each company's spending and revenue (both in millions).

Company	Advertising Exp.	Revenue
Company 1	1.2	5.40
Company 2	1.1	5.54
Company 3	1.6	5.32
Company 4	1.5	5.49
Company 5	1.76	6.0
Company 6	1.86	5.87
Company 7	2.00	6.12
Company 8	2.03	6.08
Company 9	1.93	5.78
Company 10	1.99	5.99

5. What is an appropriate summary for the data to answer the reseachers's question?

(a) Mean and standard deviation

(b) Median and interquartile range

(c) Proportions

(d) Correlation

6. Of the following, which is the most appropriate way to display the results?

(a) Histogram

(b) Scatterplot

(c) Stem-and-leaf plot

(d) Side-by-side boxplots

CONTINUED

Items 7 and 8 refer to the following situation: Suppose we obtained the data set below, where Subject indicates the subject's ID number and Hair Color is coded as follows: 1=Blonde, 2=Brunette, 3=Red.

Subject	Hair Color
1	3
2	2
3	2
4	1
5	2
6	3
7	2
8	1

7. What is an appropriate numerical summary for this type of data?

(a) Mean and standard deviation

(b) Median and interquartile range

(c) Proportions

(d) Correlation

8. Of the following, which is the most appropriate to display the results?

(a) Histogram

(b) Pie chart

(c) Stem-and-leaf plot

(d) Scatterplot

Items 9 through 12 refer to the following situation:

In studies of employment discrimination, several variables are often relevant: an employee's age, sex, race, years of experience, salary, whether promoted, and whether laid off. For each question, select the appropriate classification of explanatory and response variables.

9. Can an employee's age help us predict whether or not he or she will be promoted?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

CONTINUED

10. Is age a predictor of salary?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

11. Are men more likely to be promoted than women?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

12. Are whites paid more than blacks?

(a) Categorical explanatory and categorical response

(b) Categorical explanatory and quantitative response

(c) Quantitative explanatory and categorical response

(d) Quantitative explanatory and quantitative response

Items 13 through 16 refer to the following situation: For each research situation, decide what statistical procedure would most likely be used to answer the research question posed. Assume all assumptions have been met for using the procedure.

13. Is the amount of time spent on cell phones the same for Americans, Canadians and Europeans?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

CONTINUED

14. Does knowing a college student's SAT score tell us anything about his or her first year college grade point average?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

15. Does support for a school bond issue (For or Against) differ by neighborhood in the city?

(a) Test one mean against a hypothesized constant.

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

16. Is there a difference between city gas mileage and highway gas mileage for minicompact cars?

(a) Test one mean against a hypothesized constant

(b) Test the difference between two means (independent samples).

(c) Test the difference in means between two paired or dependent samples.

(d) Test for a difference in more than two means (one way ANOVA).

(e) Test if the slope in the regression equation is 0.

(f) Use a chi-squared test of association.

APPENDIX B

PROJECT ASSIGNMENTS

This appendix contains project assignments and grading rubrics.

Assignment 1
Describing the Data

Assignment Overview: Describe the data that you've chosen using appropriate numerical and graphical displays.

What you should include:

- A minimum of three different graphical displays of your data with interpretation
- Appropriate numerical displays of all variables included in your dataset with interpretation
- A description of how the data was obtained by you and/or by the source where you found the data (if data was not collected directly by your group)

Formatting: Your submission should be written with the following sections: Introduction, Results and Discussion. Your introduction should include basic background information to familiarize the reader with your chosen topic as well as information on how you obtained the data and/or how the data was obtained by the source that you used to acquire it (if data was not collected directly by your group). The Results section should include your data displays, appropriately labeled. The Discussion section should include your interpretations of the displays with clear references to your figures from the Results section and comments on the the way the data was obtained.

Grade Sheet for Assignment 1

NAME:

Section	Topic	Points Assigned
Introduction		
	Provide basic background information	
	Indicate how data was obtained from source	
	Indicate how source obtained data	
Results		
	Graph 1	
	Graph 2	
	Graph 3	
	Numerical Displays	
Discussion		
	Interpretation of graph 1	
	Interpretation of graph 2	
	Interpretation of graph 3	
	Interpretation of numerical displays	
	Comments on data acquisition	
	TOTAL	

Grading Rubric for Assignment 1 Total Possible Points for Assignment 1 = 30

Introduction: Total Points Possible = 6

Provide Basic Background Information:

Score	Description
0	No background information provided
1	Limited or average background provided; key information left out
2	Good overview provided; all key information included

Indicate how data was obtained from source:

Score	Description
0	No information provided
1	Minimal information provided; source not clearly indicated
2	Specific information provided; source clearly referenced

Indicate how source obtained data:

Score	Description
0	No information provided
1	Minimal information provided
2	Specific information provided; clear references to collection method(s)

Note: If the source does not provide information on how the data was collected or if students collected their own data, students should indicate this clearly and the score should be 2.

Results: Total Points Possible = 10

Graphical Displays:

Score	Description
0	No graph
1	Graph submitted but inappropriate for data type
2	Appropriate graph submitted

Numerical Displays:

Score	Description
0	No numerical displayed provided
1	Numerical displays provided but most (more than 70%) are inappropriate for the data type and/or many variables do not have displays provided
2	Numerical displays provided but some (between 30% to 70%) are inappropriate for the data type and/or some variables do not have displays provided
3	Numerical displays provided but few (less than 30%) are inappropriate for the data type and/or few variables do not have displayed provided
4	All variables have appropriate displays provided

Discussion: Total Points Possible = 14
Interpretation of graphs:

Score	Discussion
0	No interpretation given
1	Completely inappropriate interpretation given
2	Some aspects of interpretation are correct but some aspects are incorrect or left out
3	Appropriate interpretation using correct statistical language

Interpretation of Numerical Displays:

Score	Description
0	No interpretation given
1	Completely inappropriate interpretation given
2	Some aspects of interpretation are correct but some aspects are incorrect or left out
3	Appropriate interpretation using correct statistical language

Comments on data acquisition:

Score	Description
0	No comments given
1	Comments are vague and do not use appropriate statistical language to discuss
2	Comments are insightful and include appropriate use of statistical language

Assignment 2
Probability and Random Events

Assignment Overview: Identify events that may occur with reference to your topic and discuss the probabilities associated with them. Use the probabilities you assigned to discuss independence of these events. Discuss what probability distribution one of your variables might follow. Using this distribution, estimate a characteristic (e.g. estimate the expected number of HR hit by some baseball player).

What you should include:

- Identification of at least 2 single events that may occur (Note: A single event is defined as a single outcome of a sample space.)
- Identification of at least 2 multiple events that may occur (Note: A multiple event is defined as an event made up of more than one single event.)
- Probability assignments for all events identified (using the classical approach, relative frequency approach or subjective approach)
- Discussion of independence (or lack of) for two of your identified events
- Discussion of probability distribution that one of your variables may follow
- Estimation of some characteristic of your topic based on the probability distribution you identified.

Formatting: Your submission should be written with the following sections: Introduction, Results and Discussion. The introduction section should provide for the reader a brief reminder of the data involved with your topic. The results section should include the identification of the 2 single and 2 multiple events that you have selected, along with the probabilities you assigned to each and finally the estimated value for the characteristic of your topic. The Discussion section should include a discussion of what method you used to assign probabilities and a discussion on whether or not two of your events are independent. It should also include a discussion of the probability distribution that you chose as the distribution that your selected variable follows and why you chose it. Finally, you should discuss relevant details of the characteristic you chose to estimate and its estimator.

Grade Sheet for Assignment 2

NAME:

Section	Topic	Points Assigned
Introduction		
	Overview	
Results		
	Identification of single event	
	Identification of single event	
	Identification of multiple event	
	Identification of multiple event	
	Probability assignment - First	
	Probability assignment - Second	
	Probability assignment - Third	
	Probability assignment - Fourth	
Discussion		
	Probability assignment method	
	Discussion of independence of events	
	Probability distribution discussion	
	Estimator	
	Total	

Grading Rubric for Assignment 2

Total Possible Points = 32

Introduction: Total Possible Points = 3

Brief overview provided:

Score	Description
0	No overview provided
1	Average overview provided
2	Good, concise overview provided

Results: Total Possible Points = 16

Identification of Single Event:

Score	Description
0	No single event identified
1	Event identified but not a valid, single event
2	Appropriate single event identified

Identification of Multiple Event:

Score	Description
0	No multiple event identified
1	Event identified but not a valid, multiple event
2	Appropriate multiple event identified

Probability Assignments:

Score	Description
0	No probabilities assigned
1	Invalid probabilities assigned
2	Valid probabilities assigned

Discussion: Total Possible Points = 13

Discussion of probability assignment method used:

Score	Description
0	No discussion given
1	Invalid method or inappropriate method
2	Valid and appropropriate method given

Discussion of Independence of Events:

Score	Description
0	No discussion given
1	Incorrect conclusion drawn
2	Correct conclusion drawn but inappropriate reasoning given
3	Correct conclusion drawn with appropriate reasoning without correct usage of statistical language
4	Correct conclusion drawn with appropriate reasoning with correct usage of statistical language

Discussion and Identification of probability distribution:

Score	Description
0	No probability distribution identified and no discussion
1	Inappropriate distribution given for event chosen
2	Appropriate distribution selected for event but inappropriate reasoning given
3	Appropriate distribution selected for event and appropriate reasoning given without correct usage of statistical language
4	Appropriate distribution selected for event and appropriate reasoning given with correct usage of statistical language

Estimator:

Score	Description
0	No estimator and no discussion given
1	Inappropriate estimator given with or without discussion
2	Correct estimator without discussion
3	Correct estimator and discussion

Assignment 3
Statistical Inference

Assignment Overview: Test hypotheses about your topic using appropriate statistical methods and discuss the results.

What you should include:

- 3 different hypotheses about your topic
- Application of appropriate statistical techniques to test your hypotheses
- Discussion of results obtained from the tests

Formatting: You submission should be in written form with the following sections: Introduction, Results and Discussion. Your Introduction section should provide a brief reminder for the reader of the data involved with your topic and the questions (hypotheses) you will be testing, written in the context of your topic. Your Results section should formalize these hypotheses (specifically state the null and alternative hypotheses) and should include results from the the tests you performed. The Discussion section should include a discussion of what tests were used and why as well as a formal discussion of the results in the context of your topic.

Grade Sheet for Assignment 3

NAME:

Section	Topic	Points Assigned
Introduction		
	Overview	
	Hypothesis 1	
	Hypothesis 2	
	Hypothesis 3	
Results		
	Formal Hypothesis 1	
	Formal Hypothesis 2	
	Formal Hypothesis 3	
	Test Results 1	
	Test Results 2	
	Test Results 3	
Discussion		
	Test Discussion 1	
	Test Discussion 2	
	Test Discussion 3	
	Results in Context 1	
	Results in Context 2	
	Results in Context 3	
	TOTAL	

Grading Rubric for Assignment 3

Total Possible Points = 47

Introduction: Total Possible Points = 11

Brief overview provided:

Score	Description
0	No overview provided
1	Average overview provided
2	Good, concise overview provided

Hypotheses in context:

Score	Description
0	No hypothesis given
1	Inappropriate hypothesis given
2	Appropriate hypothesis, no contextual wording
3	Appropriate hypothesis given with context

Results: Total Possible Points = 18

Formal hypothesis:

Score	Description
0	No formal hypothesis provided
1	Inappropriately constructed null and alternative hypotheses
2	One hypothesis inappropriately constructed
3	Both hypotheses formed correctly

Test Results:

Score	Description
0	No results provided
1	Inappropriate test performed
2	Appropriate test performed with incorrect results
3	Appropriate test performed with correct results

Discussion: Total Possible Points = 18

Discussion about tests:

Score	Description
0	No discussion given
1	Wrong test selected because of incorrect reasoning about the data
2	Correct test selected but incorrect reasoning about the data used to choose it
3	Correct test selected and correct reasoning used to select it

Discussion of results in context:

Score	Description
0	No discussion of results given
1	Results stated with no conclusions drawn
2	Results stated, conclusions drawn but with no context
3	Results stated, conclusions drawn within the context of the topic

Final Assignment

Assignment Overview: Collate information from all assignments into one cohesive document and provide reflections on the project.

What you should include:

- Introduction
- Results
- Discussion
- Conclusion and Reflection

Formatting: Your submission should be in written form and should include the above sections. The Introduction section should describe your project, including a background on your topic and reasons why you chose the topic. The Results section should include all results from all three previous assignments.. The Discussion section should include analysis and interpretation of any and all results included in the previous section. The interpretations should be in the context of your topic. The Summary section should include conclusions about your topic, suggestions for future research and your group's personal reflections on the projects.

Grade Sheet for Final Assignment

NAME:

Section	Topic	Points Possible
Introduction		
	Background provided	
	Reason for topic choice	
Results		
	Results provided	
Discussion		
	Discussion provided	
Summary		
	Conclusions drawn	
	Future research	
	Reflection	
	TOTAL	

Grading Rubric for Final Assignment

Total Possible Points = 16

Introduction: Total Points = 4

Background provided:

0	No background provided
1	Minimal background provided
2	Good overview provided

Reasons for choosing topic:

0	No reasons given
1	Minimal reasons given; little discussion
2	Good reasons given; good discussion

Results: Total Points Possible = 4

Results:

0	No results given
1	Few results given (less than 30% of all results)
2	Some results given (between 30% and 70% given)
3	Most results given (more than 70% of all results)
4	All results given

Discussion: Total Points Possible = 3

Discussion:

0	No discussion given
1	Poor discussion given; results not explained in context and results not explained correctly
2	Average discussion given; results explained correctly but not given in context
3	Good discussion given; results explained correctly and given in context

Summary: Total Possible Points = 5

Conclusions

0	No conclusions given
1	Minimal conclusions given
2	Good conclusions given

Future research:

0	No future research suggestions given
1	Future research suggestions given but not relevant
2	Relevant future research suggestions given

Reflections:

0	No reflections given
1	Reflections given

Peer Evaluation Form

Name:
Assignment:

Group Member Name	% of Work	Comments	Score
Me			

How to use this sheet:

The chart given asks you to rate each group member for the amount of work they have done. The total should add up to 100%. Give comments and examples of each member's achievements and then give a score based on the scale below. THIS SHEET IS CONFIDENTIAL SO SPEAK HONESTLY!

Scoring Guidelines:

6= Student was the group leader. S/he came up with the majority of the ideas and assigned tasks and did more than his/her share of the work

5= This student was one of the group leaders, paid attention, and did more than his/her share of the work.

4= This person was a significant contributory to the group's efforts and did his/her share of the work.

3= This person did most of his/her share of the work and contributed to the overall product

2= This person was generally unproductive and didn't contribute their fair share to the group's effort but still gave some assistance to the group's efforts.

1= This person did not contribute at all.

0= This person did not contribute to the group and negatively affected other people in the group with his/her behavior.

Note that if someone gets a 5 or 6, they have made up work for another group member(s) and this(these) members should receive below a 3.

APPENDIX C

SAMPLE PROJECTS

This appendix contains sample projects written by students involved in this study. Special thanks goes to all of the students who provided electronic copies and permission to reprint their work in this dissertation.

C.1 SAMPLE ASSIGNMENTS: ASSIGNMENT 1

C.1.1 Is Talbot Worth His Salary?

Stat 1000 Project
February 3, 2010

Introduction

Max Talbot defied all expectations in his performance during the final game of the Stanley Cup finals when he scored both goals to defeat the Red Wings 2-1. This type of performance is invaluable but was Max Talbot worth his salary based on his performance during the 2008-2009 regular season? Max Talbot was drafted 234th by the Penguins in 2002 and was offered $700,000 to play as a center last season. To try to evaluate his value as a player we compared his stats over 75 games with other centers that played at least half the season or 42 games. Stats that were analyzed included total goals, total assists, ice time, and total penalty minutes. All the data used except for salary information was found on the official NHL website. Data is recorded by NHL officials throughout the season on a game by game basis. The National Hockey League employs an official scorekeeper for every game. Salary data was collected by the National Hockey League Players' Association (NHLPA), which keeps track of player contracts for use in negotiations with the league. The salary data was found using three different databases: USA Today's salaries database, a fan page for the Chicago Blackhawks, and the NHLPA website.

Results

Figure 1

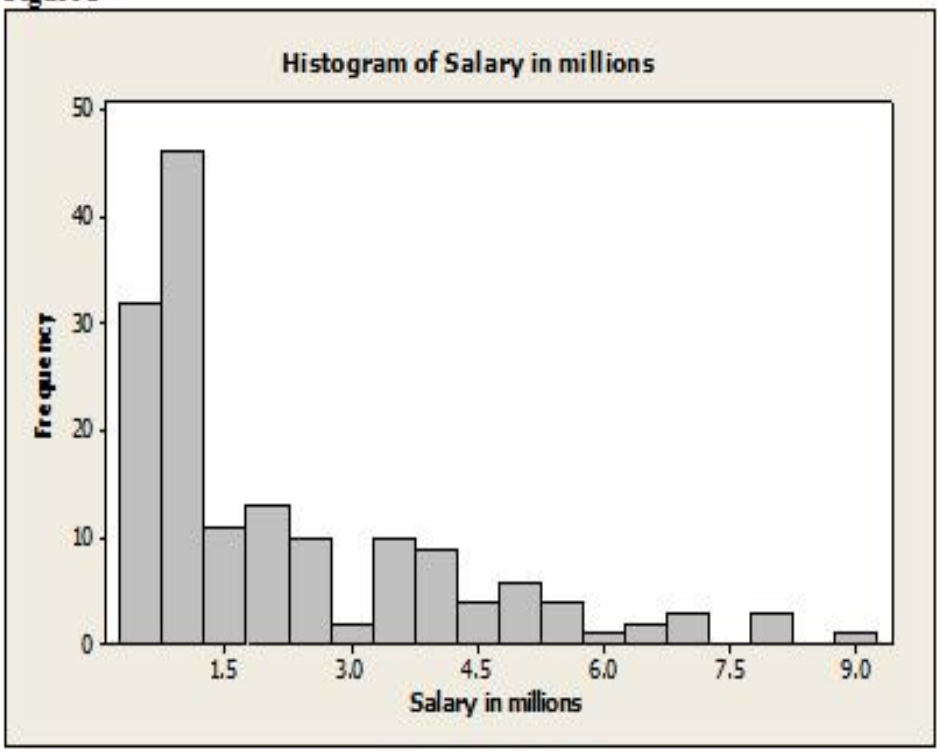


Figure 2

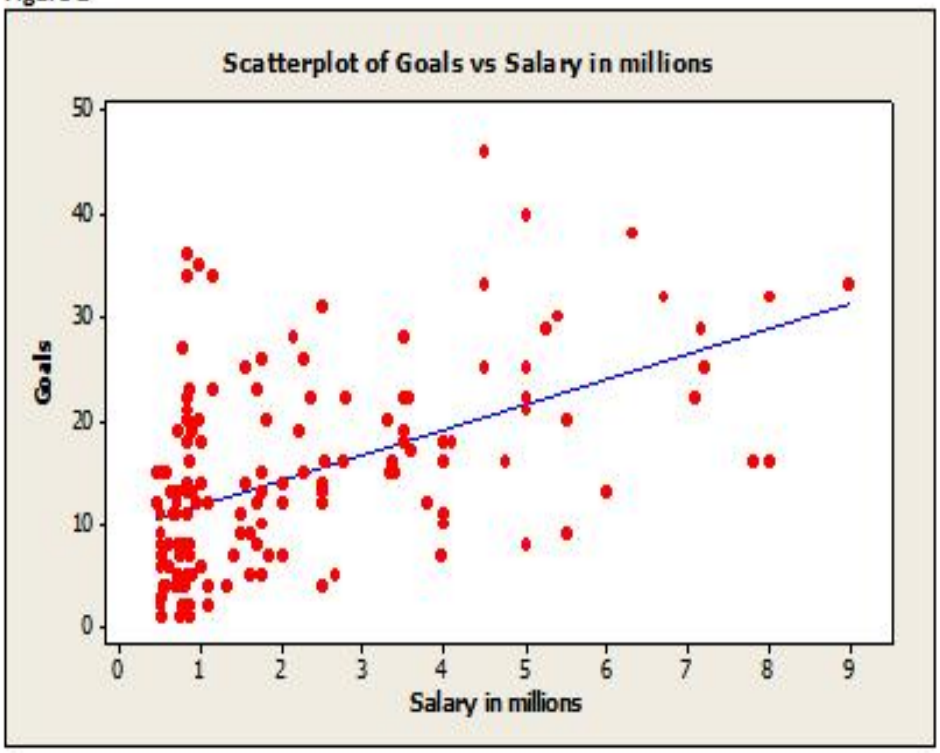


Figure 3

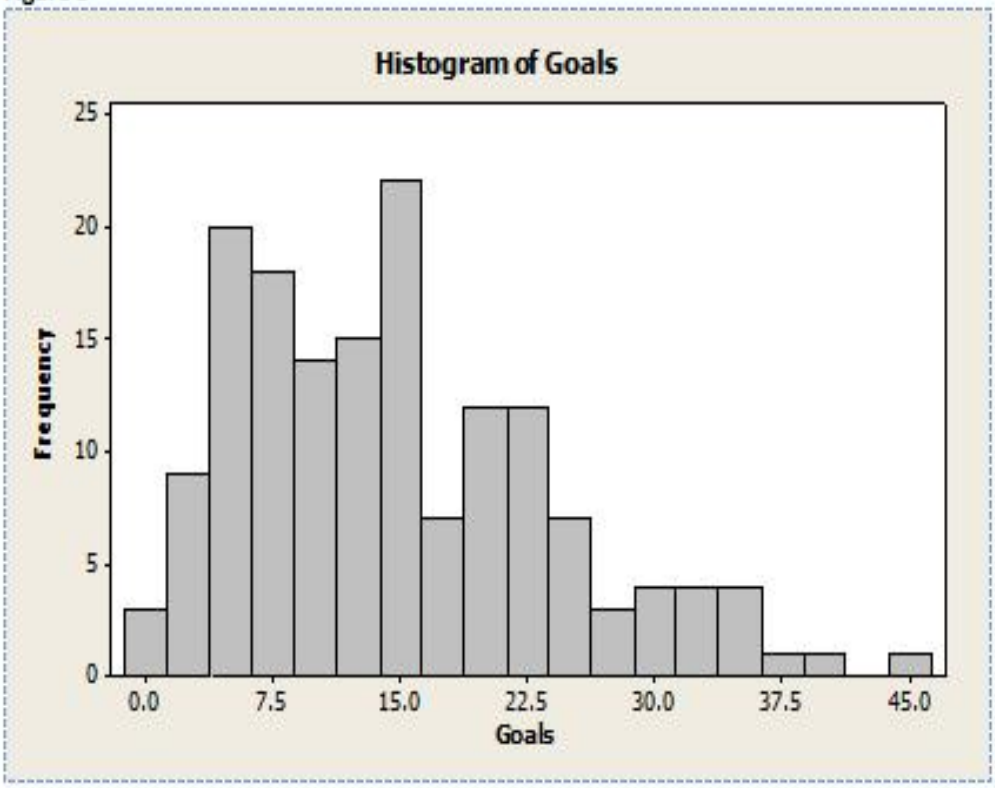


Figure 4

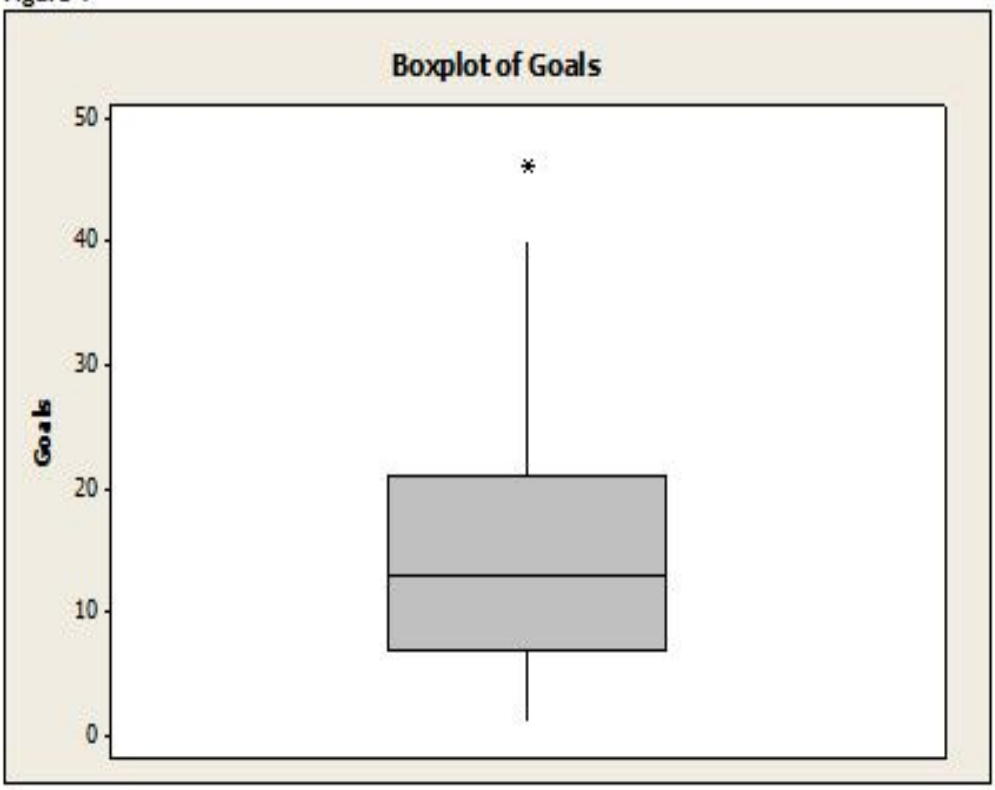


Figure 5

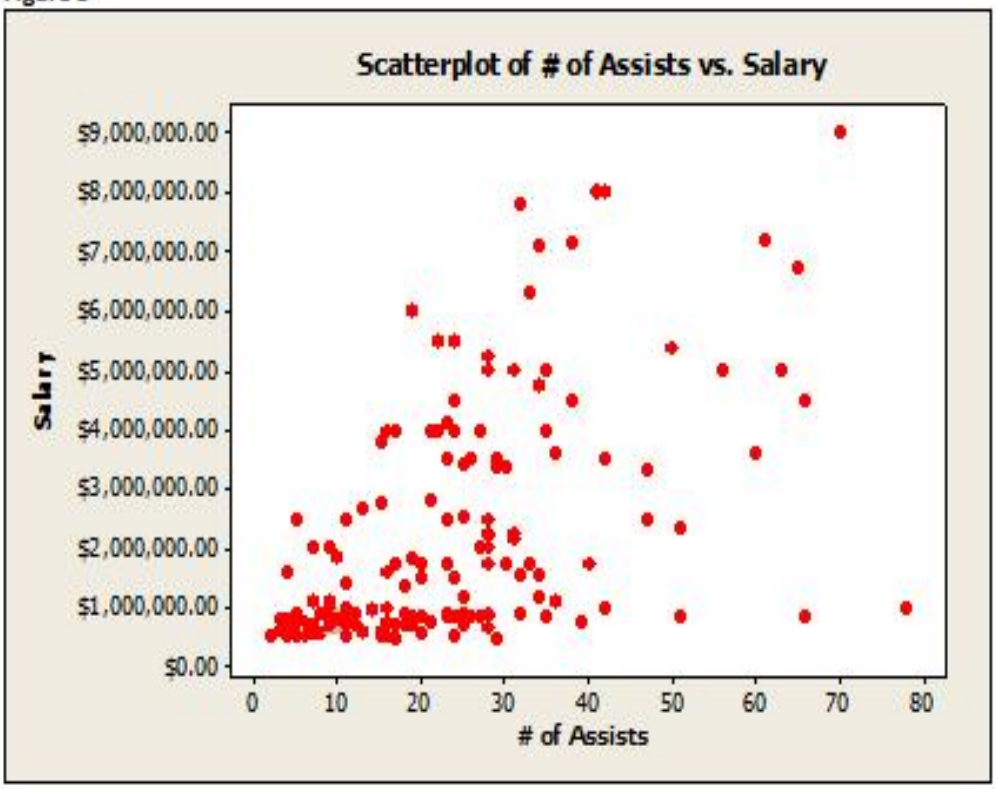


Figure 6

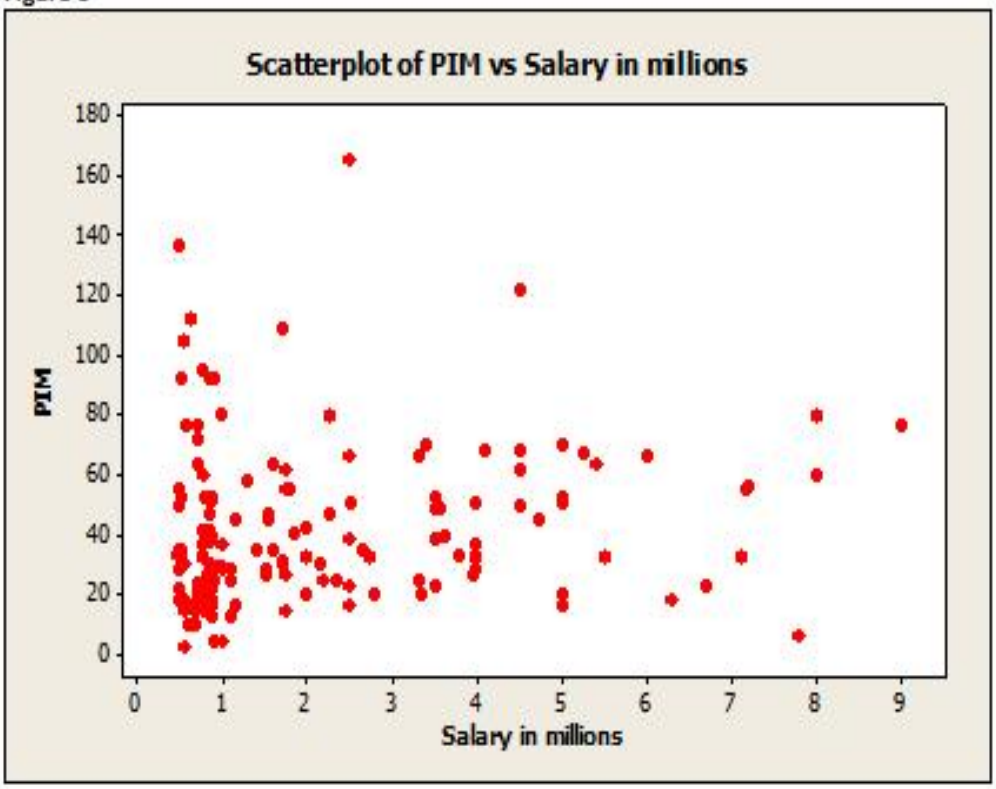


Figure 7

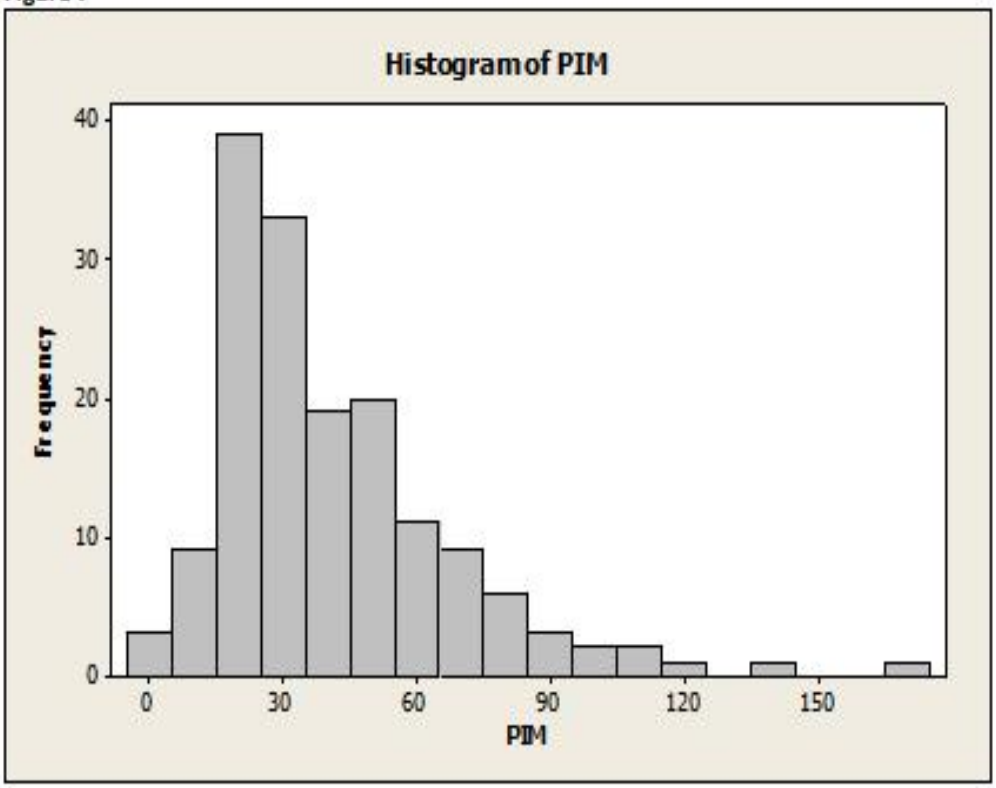


Figure 8

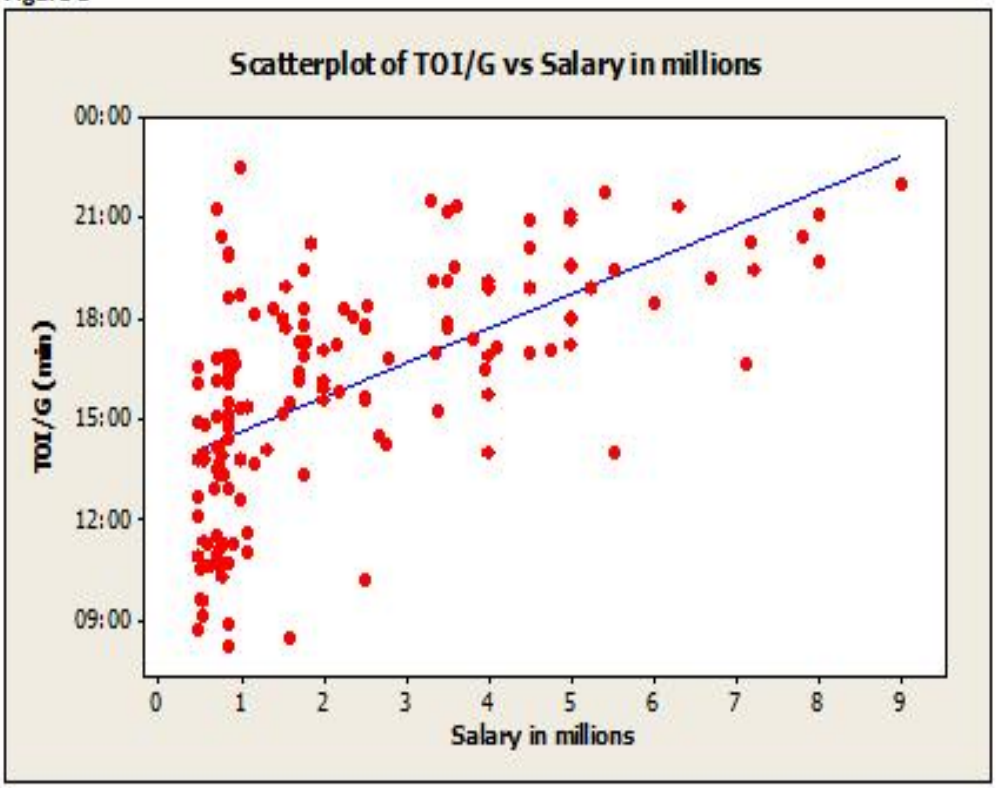
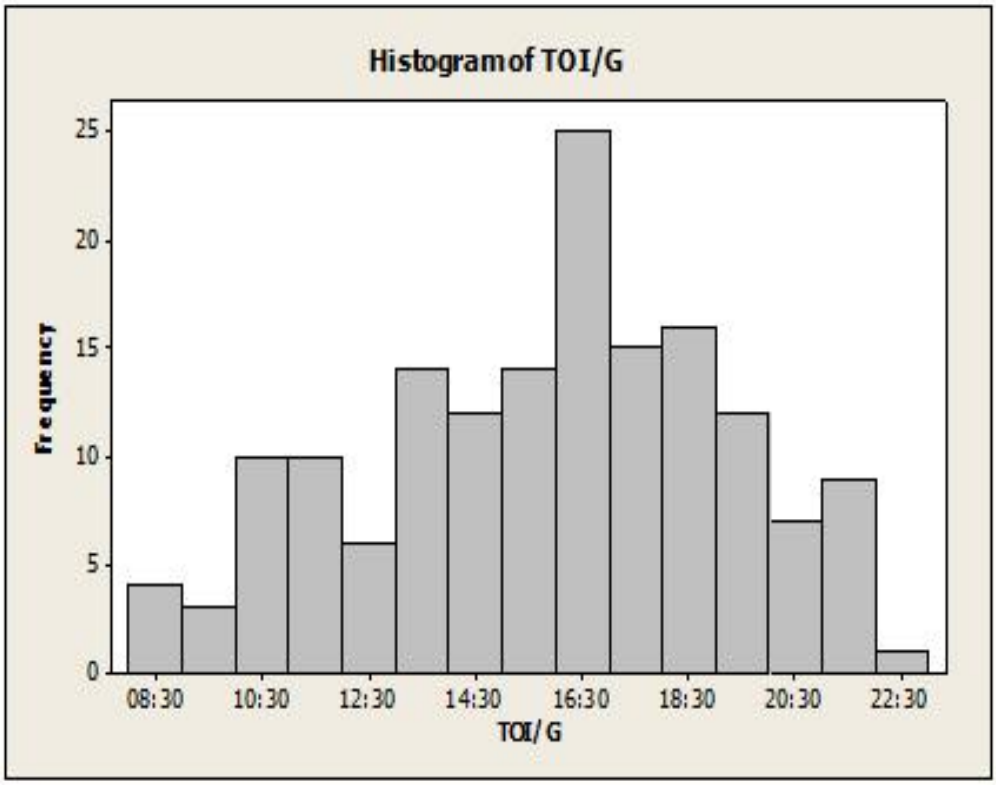


Figure 9



Numerical Display 1

Salary

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3
Salary in millions	157	0	2.193	1.952	0.475	0.770	1.315	3.450

Variable	Maximum
Salary in millions	9.000

Numerical Display 2

Goals

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Goals	157	0	14.777	9.329	1.000	7.000	13.000	21.000	46.000

Numerical Display 3

Assists

Mean # of assists=24 Standard Deviation=15 Mode=11 Minimum=2 Maximum=78

Numerical Display 4

Penalty Minutes

Mean: 40.91 Standard Deviation: 26.27 Minimum: 2.00 Q1: 22.00 Median: 33.00
Q3: 52.00 Maximum: 165.00

Numerical Display 5

Average Time on Ice per Game

Maximum	Minimum	Mean	Median
22:31	8:13	15:64	16:20

Discussion

The information on assists, goals, penalty minutes, and ice time were taken from nhl.com. The information on goals came from three sites: USA Today's salaries database, a fan page for the Chicago Blackhawks, and the NHLPA website. The numerical displays were obtained from minitab.

Figures 1

Several things are readily visible from the data on salaries and on the corresponding graphs. First and foremost, there is a dramatic rightward skew to the data, as seen on the histogram of player salaries. The center of the graph, the median, is 1.315 million dollars per year, which, as can be seen on the histogram, roughly corresponds to the mode. However, the highest salary is $9 million. The first half of the data only has a range of $840,000 (1.315 million-the minimum, .475 million) , yet the second half has a range of $9 million – $1.315 million = $7.69 million. This suggests that there is a large cluster of players making a small salary, then a few making moderate ones, and a few making very large ones. These conclusions are borne out by the histogram, showing a dramatic tailing-off after about $3 million.

Figure 2-4

So, it is fairly clear that some people score an exceptional number of goals, and some people make an exceptional amount of money. But are they necessarily the same people? And, if so, is salary a good predictor of performance on the ice?

From the looks of the scattergram, sort of. When salary is plotted against goals, though there is a linear upward trend, that trend is pretty weak. Further, around $1 million we see a strong vertical line of individuals paid the same salary, yet scoring anything from 2 to 35 goals. However, overall the pattern shows a correlation of 0.508, which is moderately strong. The regression equation for that scatterplot is Goals = 9.45 + 2.43 Salary in millions. For Maxime Talbot, if we plug in his salary as .7, we predict that he should have scored 11.151 goals. He actually scored 12. So, the regression analysis suggests that, so far as can be reasonably predicted, Talbot outperformed his salary. However, this does face the caveat that the linear correlation between salary and goals scored is far from perfect.

Numerical Display 1

The inter-quartile range of the salary data is $2.68 million, meaning suspected outliers lie above $7.47 million. By this calculation, there are 2, possibly 3 suspected outliers in salary.

Maxime Talbot ends up having a rather low salary based on these figures. His earnings of $700,000 places him below the 1st quartile, so he is in the bottom 25% of earners.

Numerical Display 2

What we note here is that there are must be a few suspected outliers, as the IQR is 14. Therefore, anything above 35 goals might be exceptionally good performance. The histogram of

goals suggests a slight right skew, again suggesting that a few people score lots of goals, and most people score a moderate number of goals.

Talbot, with 12 goals, falls between the 1st quartile and the mean, which is a higher position than he fell when his salary was considered.

Figure 5

This scatterplot graphically shows the # of assists obtained by center hockey players in the National Hockey league that played 42 or more games for the 2008-2009 season and the salary for each player. Max Talbot, our player of interest, is represented by a gray triangle on the scatterplot. Based on the scatterplot, Max Talbot makes a satisfactory salary in relation to other players that made the same or similar number of assists for the 2008-2009 season. However, the graph leads us to believe that number of assists is not the best predictor for a hockey players' salary. Although the graph does show a positive linear association between the two variables, the relationship is weak according to the r value obtained from the Pearson Correlation (r=0.533).

Numerical Display 3

Numerical Display discussion: The Mean number of assists for center hockey players that played a total of 42 or more games for the 2008-2009 season is 24 with a standard deviation of 15. The minimum number of assists was 2 and the maximum was 78. The mode number of assists was 11. Max Talbot obtained 10 assists for the 2008-2009 season which shows that his statistic falls close to what a majority of players obtained for the 2008-2009. Again, this shows that Max Talbot deserves the salary he is earning in relationship to the other players that scored the same or similar number of assists for the 2008-2009 season.

Figure 6

This is a scatterplot that shows the relationship between salary and number of penalty minutes (PIM) in the 2008-2009 season of centers who played over half of the regular season games. There is not much of a relationship between these two viable. The data is fairly even across the plot with most of the data falling between 10 and 80 minutes and is not influenced by salary. Max Talbot falls into this range with 63 total penalty minutes.

Figure 7

This is a histogram of the total penalty minutes of centers who played over half of regular season games in the 2008-2009 season. Unlike the previous penalty minutes of PIM this shows a trend, in this case a right skewed trend with the majority of player's total penalty time within the range of 10 minutes to 60 minutes so Max Talbot's penalty minutes are just outside this range. This is a much smaller range than predicted from the scatterplot showing that this is the preferred graphical tool for that information.

Numerical Display 4

This numerical display describes the penalty minute data from centers that played over half of the regular season games during the 2008-2009 regular season. The average number of penalty minutes was 40.91 minutes with a standard deviation of 26.27. Max Talbot had a total of 63 penalty minutes so he falls within one standard deviation from the mean. This leaves him somewhere in the center of the data since the lowest amount of penalty minutes was 2 minutes and the highest being 165 minutes.

Figure 8

This graph compares the ice time per game and salaries of all centers that played 42 or more games during the 2008-2009 season. On this graph Max Talbot's point is (.70000, 14:08). He falls almost directly on the regression line suggesting that, in this category, Talbot is right at the expected value for his salary. The scatter plot of ice time per game versus salary has a positive linear relationship. Unfortunately, this relationship is not very strong. The Pearson correlation of salary in millions and TOI/G is 0.592. Even when we do not consider the regression line, Talbot's point falls in the middle of that salary line.

Figure 9

This histogram displays the time on ice per game for each player data. The histogram is not dramatically skewed in either direction and it is relatively symmetrical. There is a major peak at 16:30 that shows the most frequent section of the time on ice data. The data for time on ice per game ranges from 8 minutes to 23 minutes. Talbot's value has a frequency of about 12.

Numerical Display 5

The average value for ice time per game is 16:04 when 158 different centers are considered. Talbot's average ice time per game is just below that at 14:08. The difference between Talbot's value and the average is 1.56. The lowest ice time value is 8:13. The highest value is 22:31. The fact that Talbot's salary falls close to the mean value illustrates the fact that he earned his salary during the 08-09 NHL regular season.

C.1.2 Presidential Approval Ratings

STATS 1000 Sovak
February 3, 2010

Assignment 1: The Data

Introduction

Our group chose to analyze data concerning President Obama's approval ratings. This data is interesting and relevant and, importantly, readily available. We chose to focus in on four different subdivisions of approval ratings- gender, race, age, and political party to better understand approval trends in the American population. We got our data from Gallup.com, a highly regarded polling service. They survey about 3400 people per week, however they do not publish their response percentages.

Data Table Attached

Results/Discussion

GENDER

Figure 1

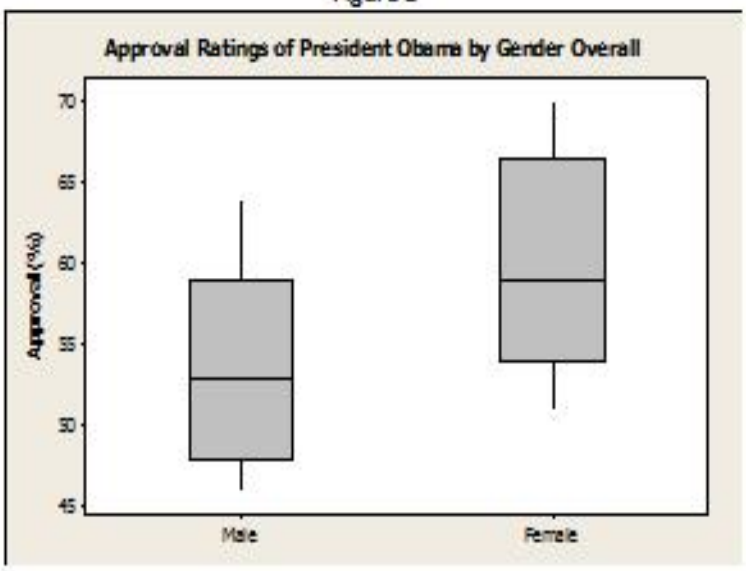
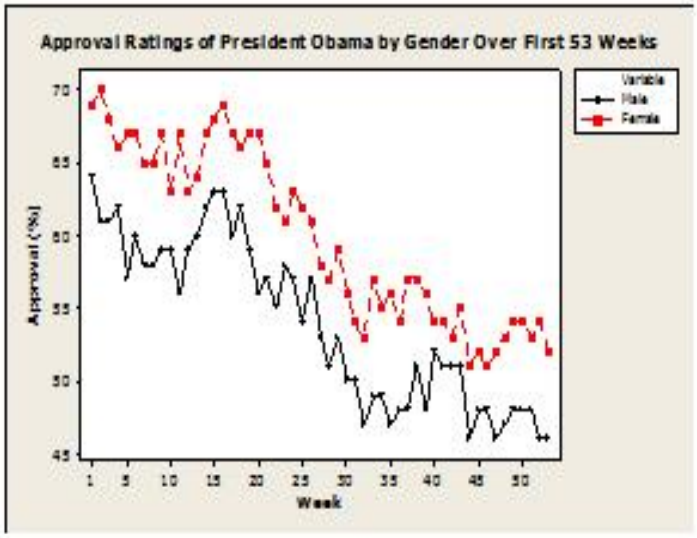


Figure 2



Overall approval ratings for President Obama have fallen over the past 53 weeks. He began with a high rating from both men and women, but over the first year, the approval ratings dropped from both groups (Figure 2). Women have a higher rating overall of Obama (Figure 1). At present, about 6% more women than men approve of Obama. Data was copied from Gallup into Excel and then into MiniTab- graphs were created from said data to show overall approval by gender and approval by gender over time.

By looking at raw numerical data, it becomes clear that women have a higher overall approval rating, with a mean of 59.943%, than men, who had a mean of 53.906%. The women also had a higher maximum (70.000%) and minimum (51.000%) approval than men (max of 64.000%, min of 51.000%).

RACE

GRAPH FOR RACE- FIGURE 3 Attached

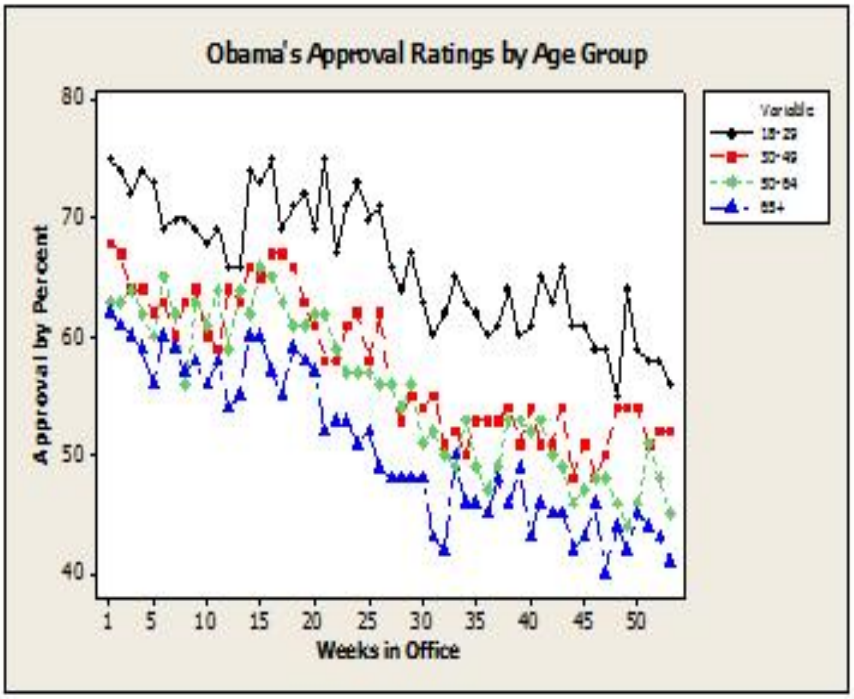
By closely examining the time series plot (figure 3), one is able to see that the approval ratings for most races presented in the study have declined. This plot depicts the approval ratings from oldest (January 25, 2009) to most recent approvals (January 18, 2010). The four categories of races consist of White, Non-White, Black, and Hispanic. Among the races the sharpest decline from the beginning of the study, January 2009, was White with a decline from 63% to 39%. Hispanics also experienced a mild decline in ratings but have been the most inconsistent group surveyed in that their percentages fluctuated significantly from week to week. The approvals ratings for Non-White Americans has also declined but their approval ratings were a bit more consistent on a week to week basis in comparison to Hispanics. The Black Race had the highest approval ratings from the beginning of the survey and almost identical ratings at the end of the survey (74 % in 2009 and 71% in 2010). There was not much fluctuation between the weeks when examining Black approval ratings. Overall, Black Americans had the highest and most consistent approval ratings. Non-White Americans began with the second highest approval ratings followed by Hispanic Americans. White Americans had the sharpest and most consistent decline in approval ratings from January 2009 to January 2010.

Under further evaluation, the values for the median and mean were calculated to be almost identical when separating the different races and comparing data. This allows the reader to know that

there is a symmetric distribution. The individual means are White – 49.5, Non-White – 77.1, Black – 92.1, and Hispanic – 73. The overall mean for the different races is 74.5 approval ratings. When configuring the standard deviation of the entire time series plot, the value is 17.6. This collective data and summary allows the reader to better understand the Race approval ratings in numerical terms.

AGE

Figure 4



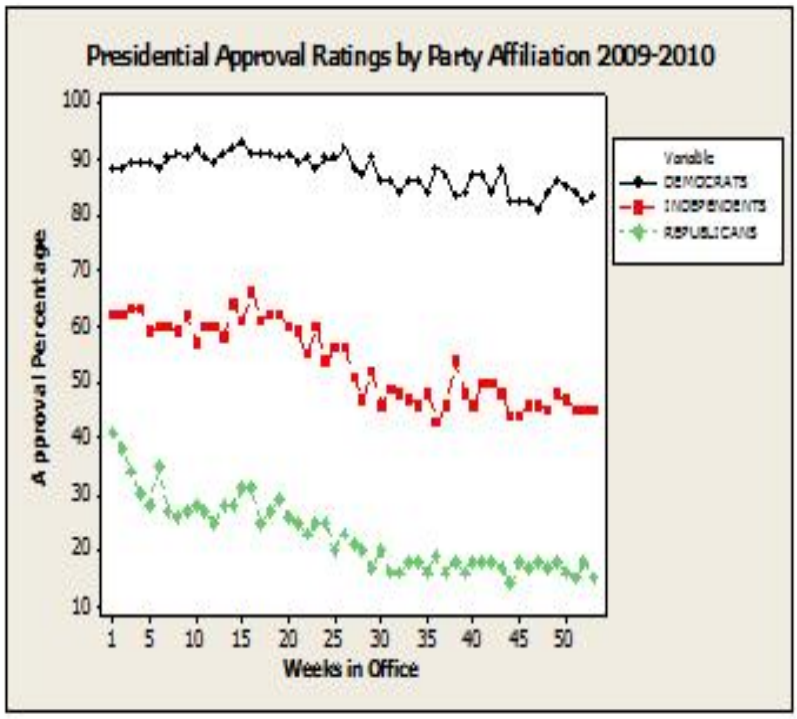
The results of this graph (figure 4) show that over the past year, President Barak Obama's approval ratings have had an overall decreased. Looking at each individual age group, we can see that the younger population has consistently given higher ratings than all other age groups. The oldest age group has, for the most part, always given the lowest ratings. There are two times when the second to oldest age group gave weaker ratings than the oldest. The two middle age groups have gone back and forth in who gave higher or lower scores. Their lines are the closest together, therefore being the most similar.

Variable	StDev
18-29	5.548
20-49	5.869
50-64	6.556
65+	6.547

Age Group	Mean	Q1	Median	Q3	Maximum	IQR
18-29	66.151	61.000	66.000	71.000	75.000	10.000
30-49	57.528	52.500	56.000	63.000	68.000	10.500
50-64	55.453	49.000	56.000	62.000	66.000	13.000
65+	50.698	45.000	49.000	57.000	62.000	12.000

Figure 5

POLITICAL PARTY



The graph (figure 5) of presidential approval by party affiliation shows a significant difference between approval ratings. Democrats report a significantly higher approval rating percentage than Independents, which is significantly higher than the approval rating of Republicans. Additionally, the Democrats approval rating is a relatively consistent value, while Independents showed a slowly decreasing approval percentage, and the Republicans' approval rapidly decreased in the first weeks, and then steadily decreased at a slower rate after that, showing a greater decrease than the Independents.

Numerical analysis reveals that the approval rating for the first year of Barack Obama's presidency varies widely by party. The average approval rating of Democrats is 87.5%, while Independents' average approval rating is 53.5%, and Republicans' average approval rating is 22.6%. This varies from current approval ratings, which are slightly lower in all categories, with Democrats' approval at 83%, Independents' at 45% and Republicans' at 15%. The year has significant differences in minimums and maximums as well. The maximum Democrat maximum approval rating over the past year is 93% and minimum approval is 81%. The maximum Independents' approval over the past year is 66% and the minimum Independent approval is 43%. Republicans' maximum approval over the past year is 41%, and the minimum is 14%.

Data Reliability

Gallup.com, a reliable source of polling information, collected our data. Their methods and response rates are not published, but the sample size averages over 3000 people/week. Polling is done by telephone, with Primary Sampling Units selected, and then stratified by population size and geography. Additionally, the strata are broken into clusters of households, which are contacted multiple times. If the house refuses, a simple substitution is used. Questions were posed by trained interviewers in an Approve, Disapprove, Indifferent answer style with data on age, race, gender, political party also collected. We chose to analyze the data in by demographic because we feel it gives far more insight into trends in the American attitude towards Barack Obama than simply analyzing overall approval ratings, as there are distinct differences in approval depending on which categorical variable analyzed. Gallup calculates their results to a 95% confidence rating, far higher than the confidence of any data we could collect ourselves.

Source:
"Gallup World Poll Methodology," from Gallup.com
http://www.gallup.com/consulting/worldpoll/108079/methodological-design.aspx#2

C.1.3 Analysis of National Gas Prices

As a group we have analyzed the average gas prices of eight regions within the United States over the period of one year. Additionally we have analyzed the average gas prices per quarter year over the span of ten years. We chose to use scatterplot graphs to represent the majority of this data in order to show the relationships between gas prices and the month of the year. We also used a bar graph and a table that represents our data numerically.

We used government data[i] for gas prices from the 2009 year to determine average prices per month of the 8 United States regions. We chose to compile 8 graphs, corresponding to each region, to observe how prices have changed over the each month in the year. We took the average price per month from each region and created a scatterplot in order to visually observe to increase and decrease in prices over the past year. We expected to see variation over the different regions, but instead we found that the regions all followed a very similar trend. From the months of January to June, the prices in each region all increased dramatically. This is generally to be expected due to higher gas prices at the beginning of the summer months. However, following the month of June each region exhibited the variation we had originally predicted. Gas prices increased in one month then decreased in the next, or vice versa. While we can say that there was an obvious trend for the first half of the year, each region's average price per month varied considerably for the remainder of the year.

We also constructed a bar graph which represented the overall average gas price for each region for the year. This graph led us to the conclusion that the Gulf Coast had the smallest average gas prices of the 8 regions while the West Coast has the highest average gas prices. We can speculate that this 10% increase is most likely attributed to distribution costs.

Our final two graphs were also representative of the average gas prices in the United States for a year. One of them is only representative of the year 2009, while the other is representative of all quarterly averages from the year 1999-2009. As observed in the earlier scatterplots of the 8 regions, and based on those averages, we expected a similar trend in which prices would rise during the first 6 months of the year, and fluctuate during the last six. Our expectations were correct, but we needed to compare it to the previous years to conclude whether or not this trend occurred regularly or if it was an anomaly. To do this, we examined the seasonal cyclical trends in gas prices over the past 10 years in the United States. In order to do this, we split up the twelve months into quarterly groups of January-March, April-June, July-September, and October-December. We then gathered the data and calculated the quarterly averages for each year. Before plotting the data, we made predictions of what the final graph would show us. We predicted that the gas prices would either remain relatively similar for a few years, then increase rapidly, followed by a decline. We based this information on the struggling economy over the past couple of years. After plotting the data, our graph showed that gas prices were relatively stable, occasionally jumping for a year followed by a decline, up until 2004. After this time, the average price per year increased to a much higher number than in the earlier years. Overall, we confirmed a general trend of increasing gas prices throughout any given year as evidenced by all years, excluding the outlier years of 2006 and 2008. In these two years, Q3-Q4 experiences a sharp drop likely due to the speculative oil market and shifting demands.

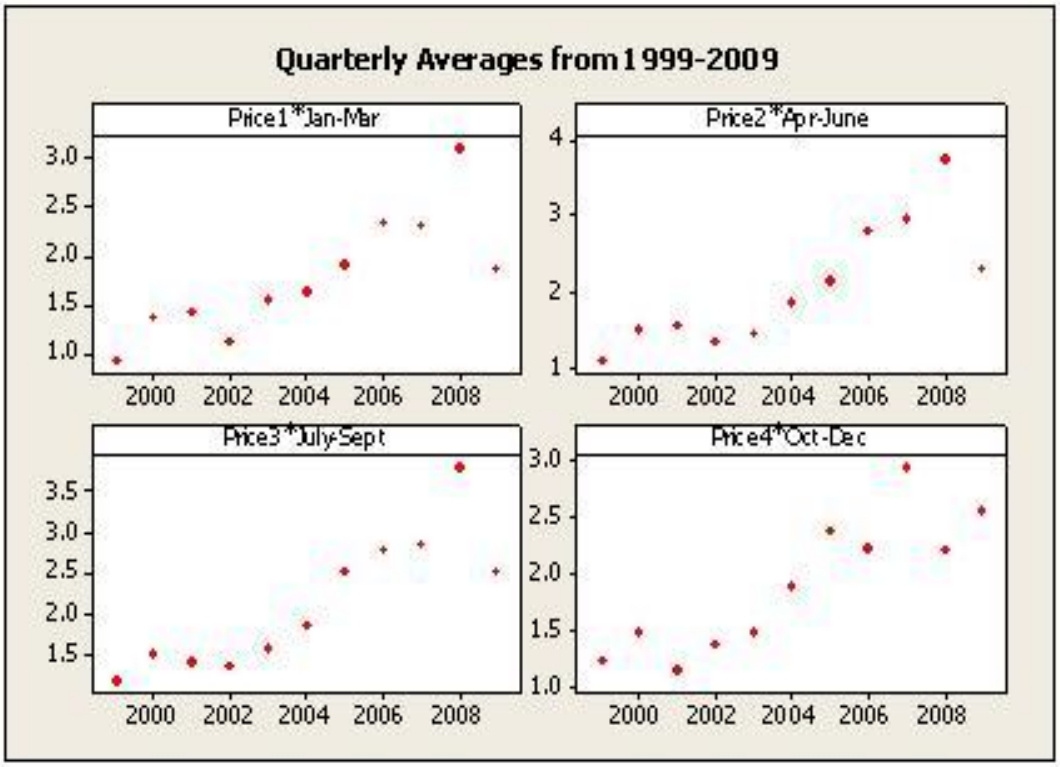
Our final data representation is a numerical display of the data from the 8 regions, and the United State average price in 2009. This representation shows us that the averages for each of this is relatively close to $2.30, which the exception of the Rocky Mountain average which is

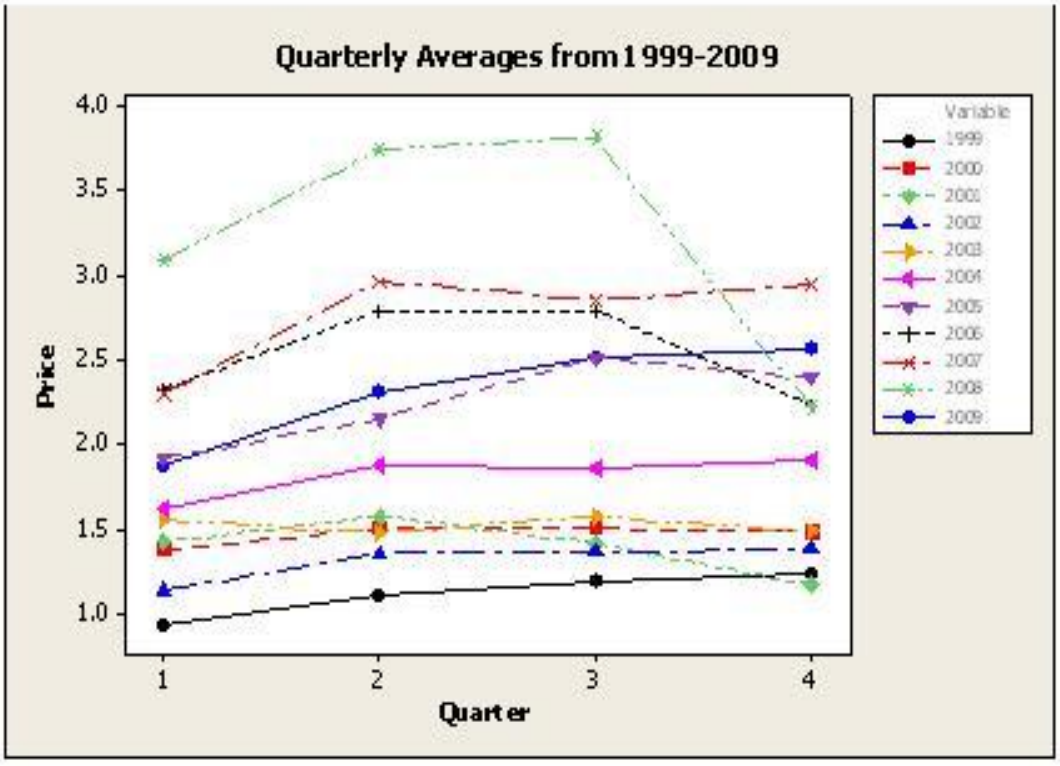
significantly higher ($2.70). Although the majority of these averages are close to being the same, their medians are a range of values from 2.31 to 2.74, none of which are equal to the mean. This information supports our graphical evidence that no symmetry exists. Additionally, each of these averages has a standard deviation of approximately .3, except for the "Average price 2009" which has a standard deviation of .0969.

With the aid of graphical and numeric representations of our collected data, we were able to predict and observe trends of a ten year period of gas price changes, as well as how prices changed within 8 United States regions over the past year.

---

[1] Data obtained from the website:
http://www.eia.doe.gov/oil_gas/petroleum/data_publications/wrgp/mogas_history.html


Quarterly Averages from 1999-2009
Price1*Jan-Mar
Price2*Apr-June
Price3*July-Sept
Price4*Oct-Dec
2000
2002
2004
2006
2008


Quarterly Averages from 1999-2009
Price
Quarter
Variable
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009

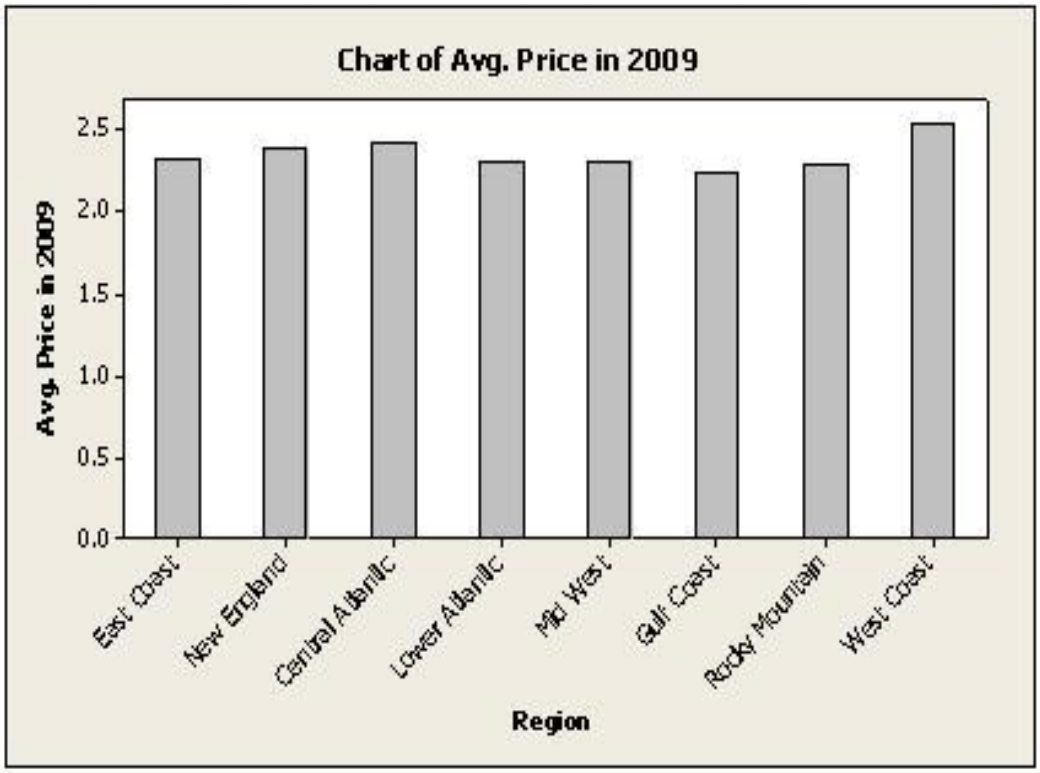
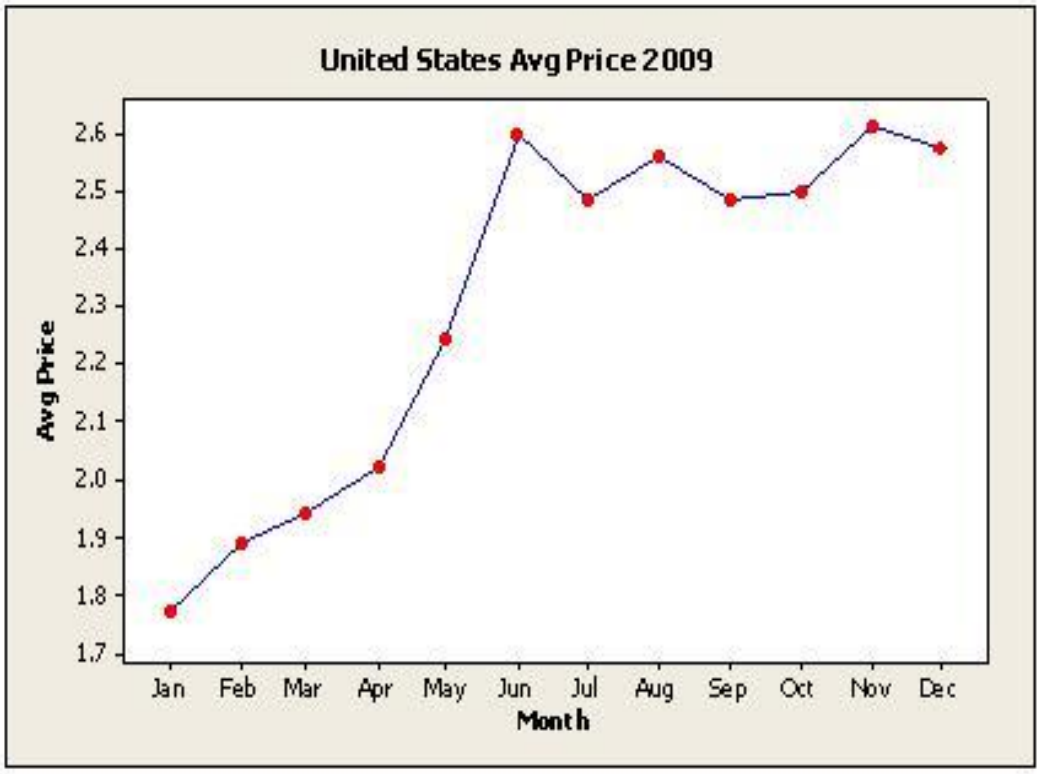
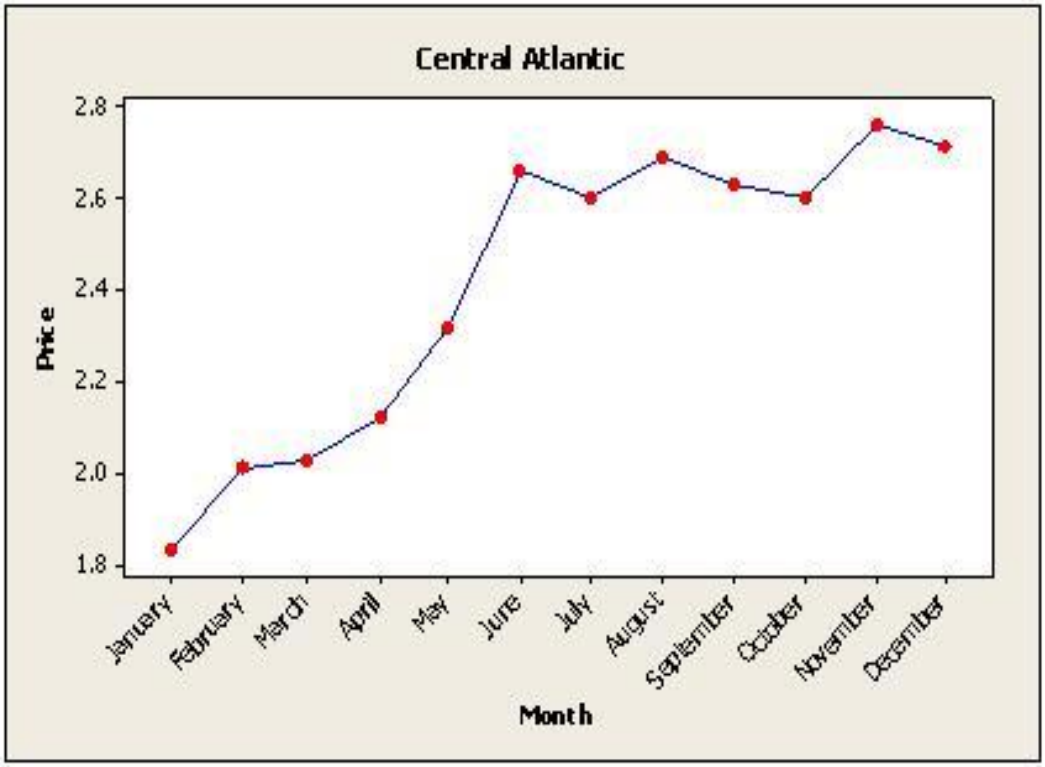
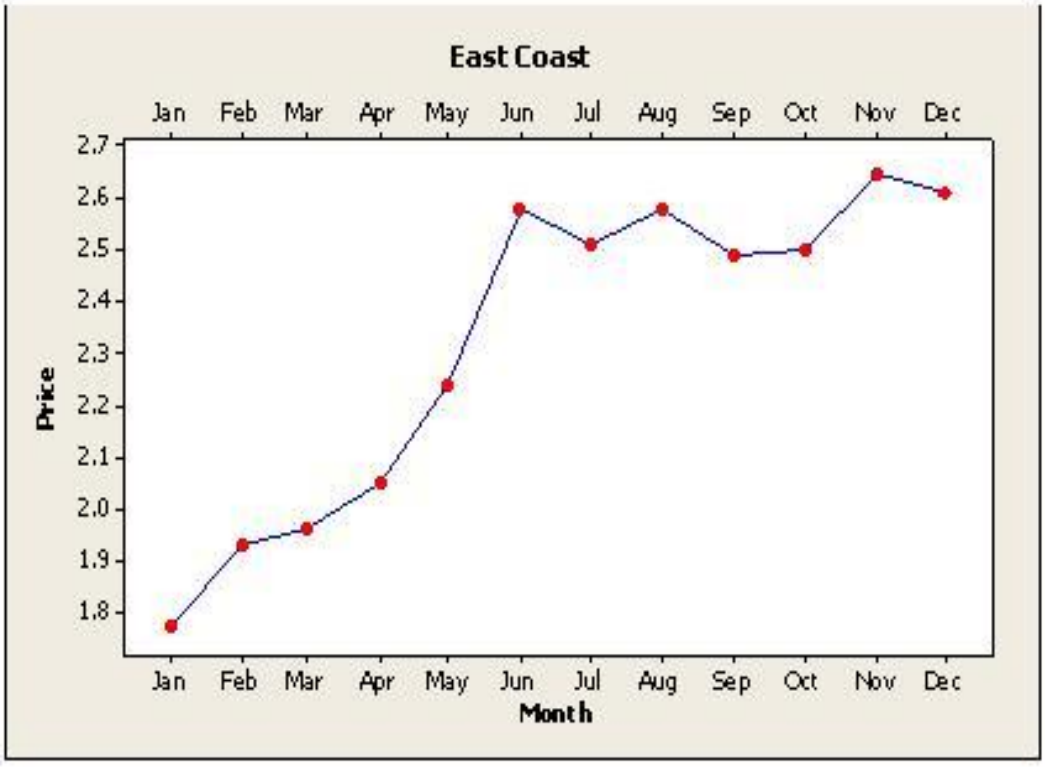
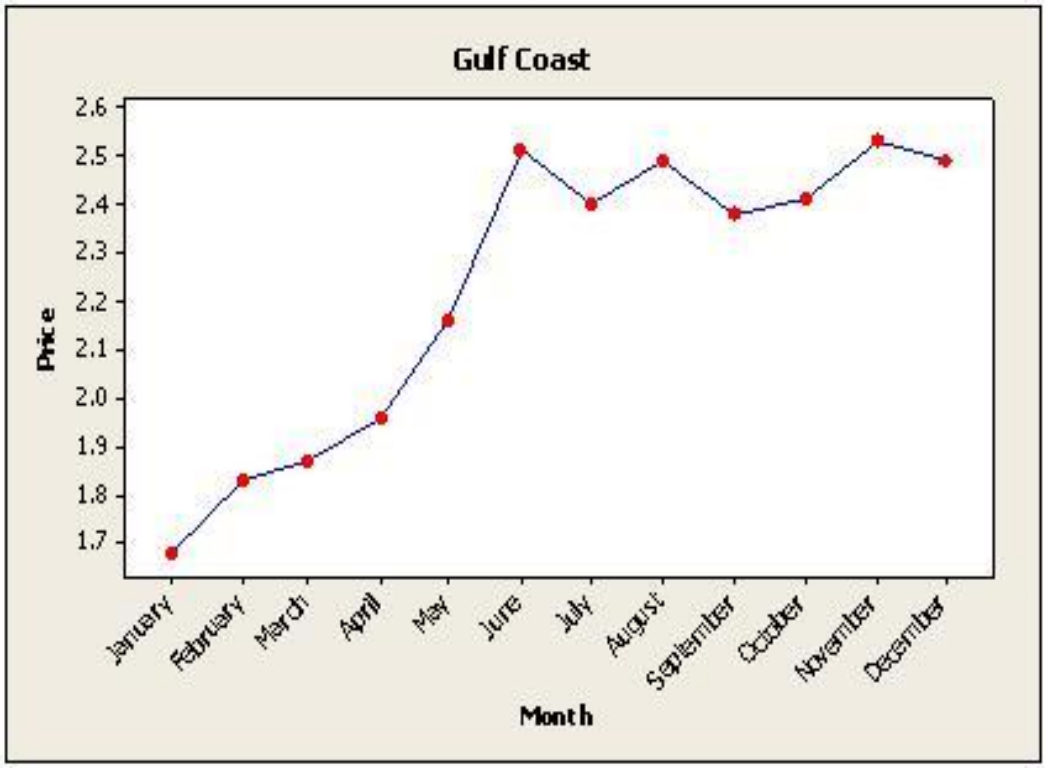


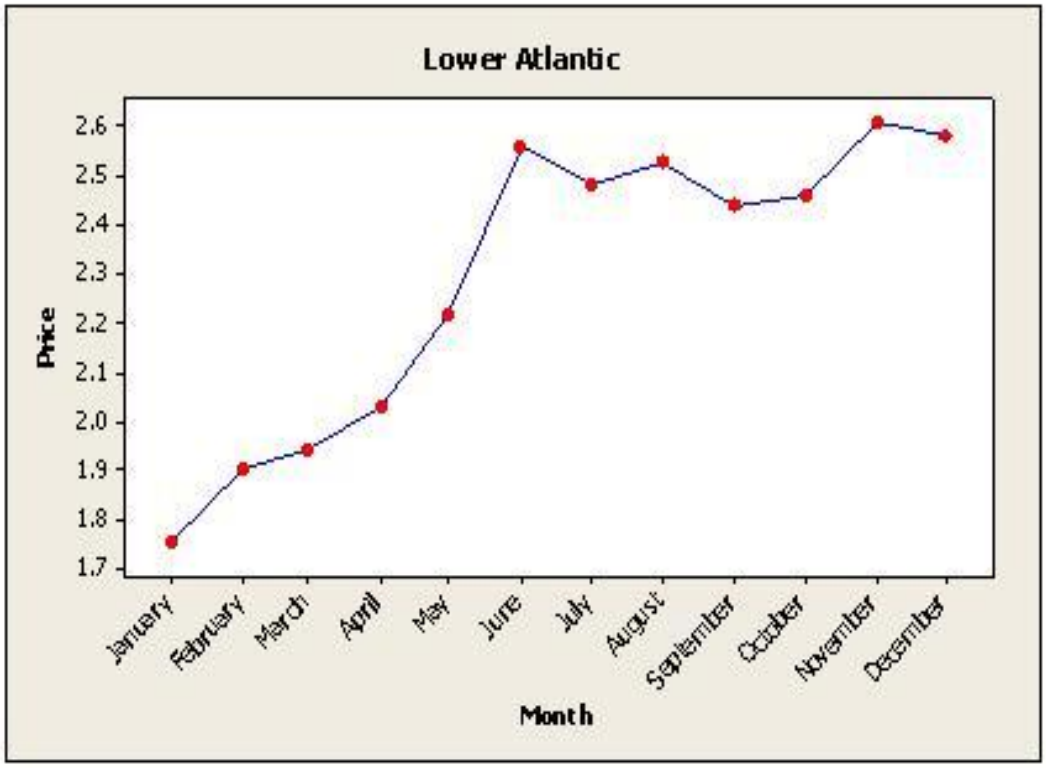
Chart of Avg. Price in 2009
Avg. Price in 2009
2.5
2.0
1.5
1.0
0.5
0.0
East Coast
New England
Central Atlantic
Lower Atlantic
Mid West
Gulf Coast
Rocky Mountain
West Coast
Region

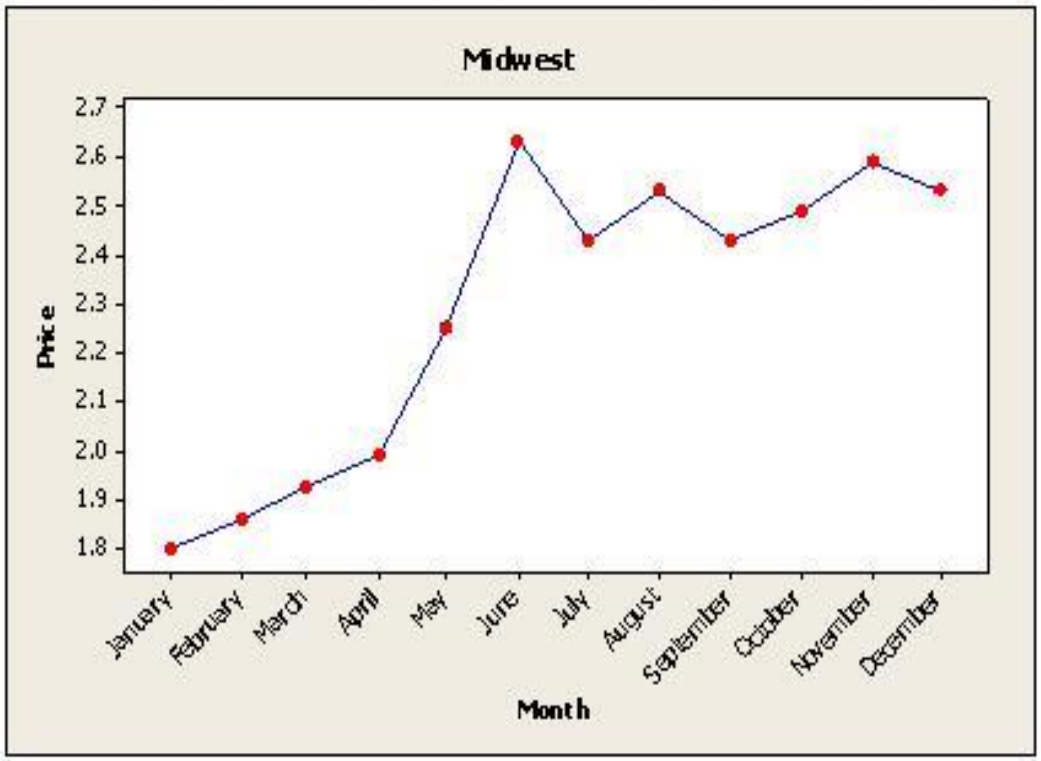

United States Avg Price 2009
Avg Price
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

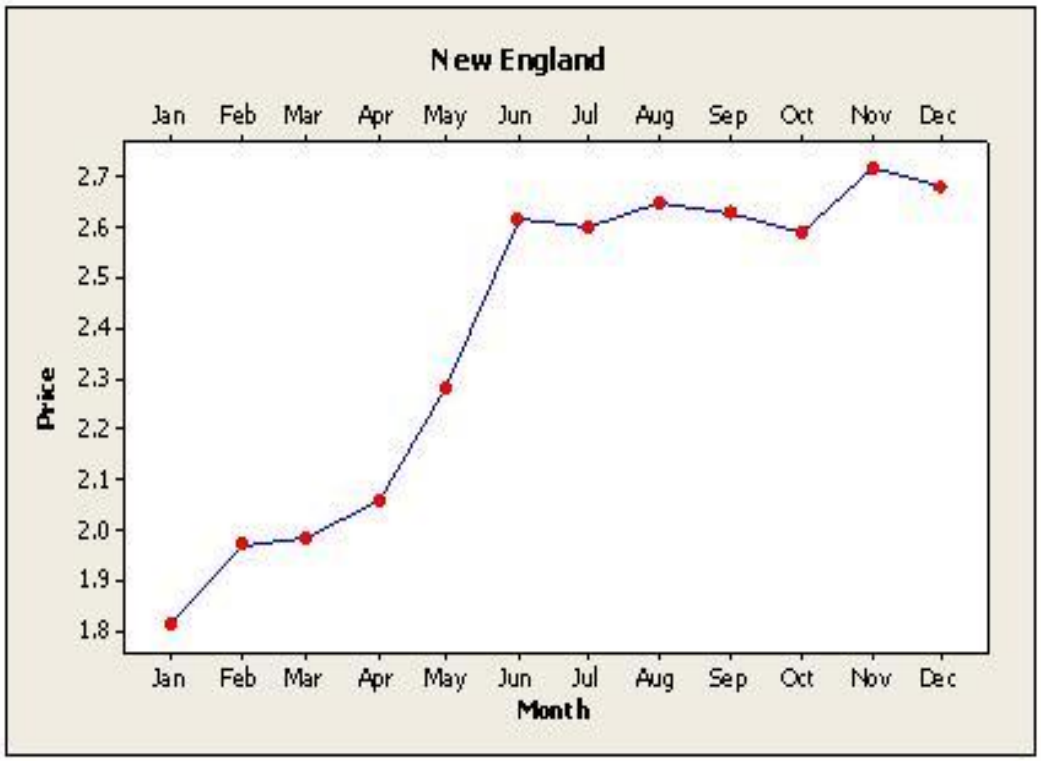

Central Atlantic
Price
2.8
2.6
2.4
2.2
2.0
1.8
January
February
March
April
May
June
July
August
September
October
November
December
Month

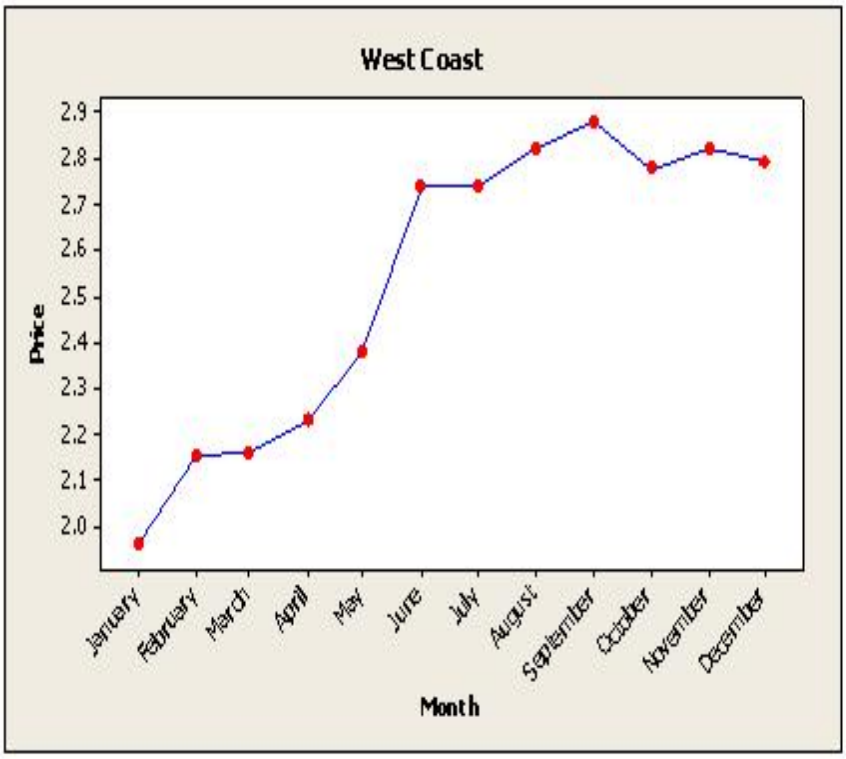

East Coast
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Price
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
Month


Gulf Coast
Price
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
January
February
March
April
May
June
July
August
September
October
November
December
Month


Lower Atlantic
Price
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
January
February
March
April
May
June
July
August
September
October
November
December
Month


Midwest
Price
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
January
February
March
April
May
June
July
August
September
October
November
December
Month


New England
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Price
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
Month


West Coast
Price
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
January
February
March
April
May
June
July
August
September
October
November
December
Month

C.2 SAMPLE ASSIGNMENTS: ASSIGNMENT 2

C.2.1 Salary vs. Performance Analysis for Pittsburgh Penguins

Introduction:

As mentioned previously, the main purpose of this study is to determine if there is any correlation between a professional hockey player's annual salary and their performance throughout the hockey season. A specific team, The Pittsburgh Penguins, was chosen for three reasons: there would be a wide range of salaries to study; the season length for every player is roughly the same; and the locality of the team. The data used for this portion is still the same as the data used in the previous portion of this study. As a reminder, players who played less than ten games during the 2008-2009 hockey season were omitted from the study as their data would appear as outliers.

Results:

Each player considered in the data set played a certain number of games. A player might play a low number of games, a medium number of games, or a high number of games. The events of a player having less than 20 games played, having from 20 to 60 games played, or having greater than 60 games played were used to examine the likelihood of a player playing many or few games. The probability of a player having either less than 20 games played or greater than 60 games played was 0.76.

The event of having a salary greater than 1 million dollars was picked as one to be analyzed. It was found that the probability of a randomly chosen player on the Pittsburgh Penguins having a salary of greater than a million dollars is 0.64 or 64%.

Having a shot on goal was picked as an event that may occur. However, to take a few lurking variables (missed games due to injury, trades, etc.) into account, the probability of having a shot on goal per game was used instead. This was found by dividing the number of total shots on goal by the number of games played during the season. The probability of a randomly chosen player on the Pittsburgh Penguins had at least one shot on goal per game was found to be 0.80. In other words, there is a 80% chance that a randomly chosen player on the Pittsburgh Penguins had at least one shot on goal per game.

Another event that was picked was assists. To take lurking variables into account, the event was again was converted to a per game event, assists per game. This was calculated by dividing the total number of assists awarded during the season and dividing by the total number of games played during the said season. The probability of a randomly chosen player on the Pittsburgh Penguins having at least one assist per game was 0.64.

Goals, or points scored, are perhaps the most important variable in any sport. Once again to take lurking variables into account, the event was converted to a per game event, goals per game. For this project, the value of 0.25 was chosen to be meaningful (one goal every four games) due to the fact that no player had a goals per game value of greater than 0.5. The probability of goals per game being greater than one goal every four games can be written as such: $P(\text{Goals per Game} \geq 0.25)=0.28$. This indicates that there is a 28% chance that any given player will score a goal once every four games.

Discussion:

Other variables might depend on the number of games a player played. For this reason, the probability of a player playing less than 20 or greater than 60 games was calculated using the relative frequency approach. 19 players had a number of games played that placed them in the less 20 category or the greater than 60 category, leaving 6 players in the category of games played from 20 to 60. The probability of a player having less than 20 or greater than 60 games played was thus 19/25, or 0.76, so that P(Games Played < 20 or Games Played > 60) = 0.76. This result indicates that most players played either many games or few games.

The probability of having a random player on the Pittsburgh Penguins have a salary of over a million dollars was found by using the relative frequency method. The number of players who had a salary of over a million dollars were counted up (16 players) and was divided by the total number of players on the team who played more than ten games during the 2008-2009 hockey season (25). The probability was found to be 0.64. Mathematically this can be expressed as P($>1000000)=0.64

Salary, for the purposes of our project is independent of the other variables we looked at. Although players get paid based on merit and skill, their salaries for the 2008 season were based on performances from seasons prior. Ultimately, their salary for the 2008 season we looked at has no bearing on how they did that season. The players with larger salaries may be expected to contribute more, but there is no way to say that they are statistically dependent on one another in terms of their stats for the 2008 season (but their pay in the future may be dependent on their performance in 2008).

The independence of the event of a player having a salary greater than $1,000,000 dollars and the event of a player having least 10 goals during the season was considered. A salary of over $1,000,000 was chosen because $1,000,000 is generally recognized as a socially significant amount of money. A number of goals of at least 10 was chosen as an indication of a relatively low standard for a player's success. The probability of a player having a salary greater than $1,000,000 dollars was P(Salary > 1,000,000) = 0.64. The probability of a player having at least 10 goals during the season was P(Total Goals >= 10) = 0.44. The two events are independent if the product of their probabilities is equal to the probability of the two events occurring together. The probability of the events occurring together is P(Salary > 1,000,000 and Total Goals >= 10) = 0.28. The product of the two events' probabilities is P(Salary > 1,000,000) * P(Total Goals >= 10) = 0.64 * 0.44 = 0.2816. As these two values are approximately equal, it is likely that the events are independent

We can use the probability that a player will earn more that one million dollars to estimate this probability for all the other teams in the NHL. This is because the NHL has a salary cap, meaning there is a limit to how much a team can spend. Unlike in Major League Baseball, no team can afford to many high wage players. This causes teams to have a similar distribution of salaries. Some players will be on the lower end in terms of pay (under a million), some will make a little more (around 1 to 4 million), and then most teams have a few star players who make the most money.

The probability of having a shot on goal per game was found by using the relative frequency approach. The sample space was defined as the players on the Pittsburgh Penguins who played more

than 10 games during the 2008-2009 season. The number of players who had one or more shot on goal per game were counted up (20 players) and divided by the total number of players in the sample space (25 players). This number came out to be 0.80 and is the probability that a randomly selected player on the Pittsburgh Penguins who played more than 10 games during the 2008-2009 hockey season had at least one shot on goal per game. This can also be expressed mathematically as P(SOG/G≥1)=0.80.

The count of players with at least one shot on goal per game could potentially follow a binomial distribution. For a binomial distribution to be applicable, four conditions must be met. First, there must be a fixed number of observations. For this study, the number of observations is the number of players being considered, which is fixed at 25. Second, each observation must be independent. The observations for this study could be dependent if there were only a limited number of shooting opportunities available to the team as a whole. In that case, the more shots one player took, the less shots other players could take. However, considering that a player could make a shot anytime they had control of the puck, the total number of shots is not necessarily limited in this manner and the shots on goal per game for each player may be considered independent. Third, for a binomial distribution to apply, each observation must fall into one of two categories, indicating success or failure. In this study, the categories were that a player had at least one shot on goal per game or that a player did not have at least one shot on goal per game. A player having at least one shot on goal per game indicated success. Fourth, the probability of success must be the same for each observation. Each player included in the study played in a certain number of games. In each game an individual player played, that player had an opportunity to make a shot on goal anytime they had control of the puck. With the way a hockey puck moves from player to player on the ice, each player would have an equal chance of gaining control of the puck and so have an equal chance of making a shot. Additionally, though different players played different numbers of games, because this variable considers the average per game, players do not have a different chance of success due to playing fewer games. Because these four conditions are met, a binomial distribution is likely a good estimate for the variable counting the number of players with at least one shot on goal per game.

The number of observations is the number of players in the sample, n = 25. The probability, p, of success of a single observation is estimated by the sample proportion $\hat{p}$ = 0.80. The distribution is then approximately B(25, 0.8). The mean of this distribution is μ_x = n*p = 25 * 0.8 = 20 and its standard deviation is $\sigma_x = \sqrt{n\,p\,(1-p)} = 2$.

The probability of having an assist per game was also found by using the relative frequency approach. The sample space was, as before, defined as the players on the Pittsburgh Penguins who played more than 10 games during the 2008-2009 season. When the assists per game was calculated, it was found that no player had 1.0 assists per game or greater. Therefore, it was then specified that having 0.2 assists per game or more would be significant and would count as if the player had at least one assist per game. The number of players who averaged at least 0.2 assists per game were counted (16 players) and divided by the total number of players in the sample space (25 players) as was done with the shots on goal per game. Using this information, 0.64 (64%) was found to be the probability that a randomly selected player on the Pittsburgh Penguins who played more than 10 games during the

2008-2009 hockey season had at least 0.2 assists per game. This can also be expressed mathematically as P(Assists/G≥.2)=0.64.

The variable of goals per game was a difficult variable to assess, mostly due to its dependence on other variables. For instance, shots on goal per game, assists per game, salary, and even games played, should obviously affect the number of goals scored per game. The correlation values between variables like this were relatively high (goals vs. shots: 0.888, goals vs. salary: 0.478, etc…). As far as probabilities go, in order to determine this variable's dependence (or lack thereof), we need to examine the other variables that may potentially affect the number of goals per game. In these examples, let: A= goals per game, B=shots on goal per game, C=assists per game, D=salary, and E=games played. The most important was the probability of goals per game and salary, P(A and D), as the point of this study is to determine the correlation between performance and pay. To do this, we find each player who had both greater than .25 goals per game and a salary of greater than $1,000,000. Because 7 of 25 people meet the criteria for A and 5 out of those 7 meet criteria for D, the percentage of D that meet A's criteria (P[A and D]) is 19.88% (or 0.1988). Now, we need to find P(D|A), which can be found by using P(D|A)= P(A and D)/P(A) [0.1988/0.64= 0.311]. A probability is said to be independent when P(D|A)=P(D). We can clearly see that 0.311≠0.64, thus these two variables are not independent (causality cannot, of course, be determined though). The same calculations were made with variables B, C, and E yielding similar results suggesting that in each case it is not independent of A [P(B|A)= 0.35, P(C|A)= 0.4375, and P(E|A)=0.2615]. Given these calculations, the variable of goals per game can be said to be dependent on each and every one of these variables discussed above.

C.2.2 Comparison of Goalies in the NHL

Introduction

"In hockey, goaltending is 75 percent of the game—unless it's bad goaltending. Then it's 100 percent of the game, because you're going to lose."

-Gene Ubriaco, Former NHL Forward

The objective of this project was to juxtapose the statistics of NHL goalies who played between 1988 and 2009. We selected 25 goalies who played a minimum of 500 games during that span. In part two of our project, we analyzed some specific events, single and multiple, in the careers of each goalie. In particular, we looked at which trophies they were awarded at the end of the season. We utilized ESPN.com and NHL.com to amass our data; both websites boast countless statisticians and sports writers who meticulously gathered our statistics.

Our project has two main sections. The "results" section identifies the relevant goaltender events we selected. Also, it includes the probability assignments of each event and the estimated value of a characteristic of that event. The "discussion" section includes an analysis on how we found our probabilities, an evaluation of event independence, a discussion of our selected distribution (the Bernoulli), and an extrapolation on details of our estimated characteristic and chosen estimator.

Results

An event refers to an outcome or set of outcomes from a chance phenomenon. A *single event* denotes one outcome from the sample space of the phenomenon. Numerous outcomes from a sample space constitute a *multiple event*. The following lists portray the pertinent events we chose for our goaltenders:

Single Events:

1) Winning the Stanley Cup
2) Winning the Vezina Trophy
3) Winning the Jennings Trophy

Multiple Events:

1) Winning the Stanley Cup and the Vezina Trophy
2) Winning the Vezina Trophy and the Jennings Trophy
3) Winning the Stanley Cup, Vezina Trophy, and Jennings Trophy

Utilizing the observed occurrences of events, we were able to assign a *probability* of each event occurring. The probability refers to the proportion of occasions the event transpires in an extensive survey of repetitions. Our survey consists of 25 goalies* whose award collections are listed below:

Single Events:

1) Winning the Stanley Cup: (1) (2) (4) (5) (6) (8) (10) (12) (24) (25)
2) Winning the Vezina Trophy: (1) (2) (5) (11) (21) (24)
3) Winning the Jennings Trophy: (1) (2) (4) (5) (6) (24) (25)

Multiple Events:

1) Winning the Stanley Cup and the Vezina Trophy: (1) (2) (5) (24)
2) Winning the Vezina Trophy and the Jennings Trophy: (1) (2) (5) (24)
3) Winning the Stanley Cup, Vezina Trophy, and Jennings Trophy: (1) (2) (5) (24)

*We numbered each goalie to ameliorate our analysis. The numbered list, including each goalie's name, can be found in Appendix A.

Results (cont'd)

From the previous event occurrences, we were able to assign the following probabilities:

Single Events:

1) Winning the Stanley Cup: 10/25 = 0.40
2) Winning the Vezina Trophy: 6/25 = 0.24
3) Winning the Jennings Trophy: 7/25 = 0.28

Multiple Events:

1) Winning the Stanley Cup and the Vezina Trophy: 4/25 = 0.16
2) Winning the Vezina Trophy and the Jennings Trophy: 4/25 = 0.16
3) Winning the Stanley Cup, Vezina Trophy, and Jennings Trophy: 4/25 = 0.16

Since the whole of our events followed a specific *distribution* (Bernoulli), we were able to estimate a characteristic of their distributions; i.e. we determined the mean and standard deviation. For our example, our event of interest is multiple event #2:

Mean (p): 4/25 = .16

Standard Deviation ($p(1-p)$) : .16*(1-.16) = .16*.84 = .1344

Discussion

I. Probability Assignments

For all of our events (1-6), the relative frequency approach was used to find the probabilities. Our probabilities were based on observed data rather than on prior knowledge or a person's opinion of the likelihood of an event. The total number of goalies is 25. Therefore, the probabilities were found by dividing the number of goalies that satisfied a certain event and divided that number by 25. For example, 7 goalies in our dataset have won the Jennings Trophy. Therefore, the probability for the Jennings event is 7/25 = 0.28.

II. Independence

Two events A and B are independent if knowing that one occurs does not change the probability that the other occurs. From this definition, none of our events are independent. Essentially, the more skilled a goalie is, the better chance that goalie has at winning trophies. Therefore, better goalies would tend to not only win more trophies but win them more often.

Looking at the equation P(A|B)=P(B), we see that this is not true for any of our events. If we know whether a goalie is skilled enough to win one trophy, it gives us some indication as to their chances of winning another trophy, which violates the definition of independence listed above.

Another proof that none of our events are independent are by looking at the mathematical equation for independent variables: P(A and B) = P(A) * P(B). As an example, let's look at single events 2 and 3: P(2) = 0.24 and P(3) = 0.28. So P(2 and 3) should be (0.24*0.28) = 0.0672. However, multiple event 2 is the occurrence of single events 2 and 3 happening together, and the probability is 0.16, not 0.0672. This also shows that they are not independent.

III. Probability Distribution

A random variable, the variables our events include, is a variable whose value is a numerical outcome of a random phenomenon. The probability distribution describes the range of possible values that a random variable can have and the probability that the value of the variable is within any division of that range.

The event we will look at is multiple event #2: goalies who have won the Vezina Trophy and the Jennings Trophy. As was discussed in the probability assignment section above, we can see that this and all of our variables follow a Bernoulli distribution. This is true because there are only two possible outcomes (Yes or No), and the sample proportion represents the probability of success (Yes).

IV. Estimator

Since our events followed a Bernoulli distribution, we utilized specific formulas to estimate their mean and standard deviation. In selecting the mean, we used the probability (p) of successes derived from our observations. To find the standard deviation, we applied the formula $p*(1-p)$.

C.2.3 Pittsburgh Panther Win Record at The Pete

Project Assignment 2: Probabilty and Random Events

Introduction:

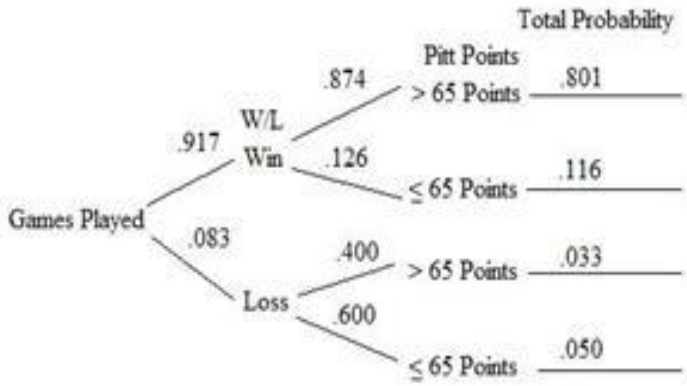
As mentioned previously, the Pittsburgh Panthers have an outstanding at home record with only ten losses at the Petersen Events Center at the close of the 2008-2009 season . The purpose of this project is to investigate which factors contribute into a Panther's victory at home such as offensive and defensive statistics. Additionally, the effect of lurking variables will be assessed for each factor.

Results:

Single Event #1: Q_{sn}: {W}
Probability of a Pitt Win
P(W)=111/121= .917

Multiple Event #1: Q_{sn}: {W & Pitt Points >65 }
Probability of a Pitt Win and Points Scored Greater than 65
P(W and Pitt Pts > 65)= 97/121 =.801

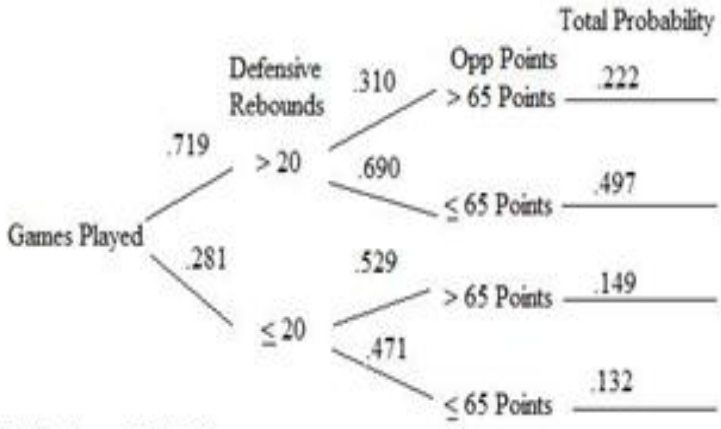
Independence Test: P(Pitt Pts > 65 | W) = P(Pitt Pts > 65) ?
97/111 ≠ 101/121
.874 ≠ .834
Events are not independent



Single Event #2: Q_{S2}: {DR>20}
Probabilty of Defenisve Rebounds Greater than 20
P(DR > 20) = 87/121 = .719

Multiple Event #2: Q_{M2}: {DR>20 & Opp Points >65}
Probabilty of Defensive Rebounds Greater than 20
and Opponents Points Greater than 65
P(DR > 20 and Opp Pts > 65) = 27/121 = .222

Independence Test: P(Opp Pts > 65 | DR > 20) = P(Opp Pts > 65) ?
27/87 ≠ 45 / 121
.310 ≠ .372
Events are not independent



Probability Distribution of a Pitt Win:
Binomial: B(121, .917)
Mean= np = 121(.917) = 111
Standard Deviation= sqrt(n*p*(1-p))=sqrt(121*.917*(1-.917)) = 3.03

Discussion:

The first single event was the probability of a Pittsburgh win and the multiple event was the probability Pittsburgh win and Pittsburgh scoring more than 65 points. The second single event was the probability of more than 20 defensive rebounds per game and the multiple event was the probability of more than 20 defensive rebounds and the opponents scoring more than 65 points. Probabilities were assigned through the relative frequency approach with the recorded data of the Pittsburgh Panthers from 2001-2002 to 2008-2009.

In both cases, events were not independent. This is logical because in the case of our first multiple event, it is expected that if the Panthers score more than 65 points, they are more likely to win. In the second event, if the Panthers obtain over 20 defensive rebounds, the opponent team is less likely to score over 65 points.

The probability distribution that fits the Panthers win-loss results the best is a binomial distribution, which models a count distribution of Pitt victories. In order for a probability distribution to be binomial,

each of the observations has to fall into one of two categories. In the data, the two categories are either a Pittsburgh win or loss, which is why we chose a binomial distribution. Additionally, there are a fixed number of observations, *n*, in the 2001-2002 to 2008-2009 seasons of 121 games. Each observation is assumed to be independent, as the outcome of one game should have no effect on the outcome of another. The probability of each success is also assumed to be the same.

As stated above, one of the characteristics that was estimated is the probability of a Pittsburgh win. In a binomial distribution, the mean is np and the standard deviation is the sqrt(np(1-p)). The mean was calculated by multiplying (121)(.917), which equaled 111, or the number of games Pitt has won. The standard deviation was solved through calculating the sqrt(121*.917*(1-.917)). This equaled a standard deviation of 3.03. In these calculations, the probability of success *p* was assumed to be equal to $\hat{p}$, which is the sample probability. This is because the true value of *p* was not known. In this case though, $\hat{p}$ serves as a great estimator for *p*.

C.3 SAMPLE ASSIGNMENTS: ASSIGNMENT 3

C.3.1 Presidential Approval Ratings

STAT 1000 Sovak

Presidential Approval Ratings Part III

Our project consists of analyzing the approval ratings of President Barak Obama. In previous sections of our project we have noted the significant changes these approvals have taken over the past 53 weeks of Obama's term in office. In this part (III) of the project, we have made three separate hypotheses about our topic. We have analyzed approval rating on the basis of Political Party, Age, and Gender. Our Null Hypotheses (H_0) is that there is *no* relationship between political party and age group with approval ratings and that men and women approval equally. Our alternate hypotheses (H_a) state the opposite, that there *is* a relationship between political party and age group with approval ratings and that men and women approve unequally. To determine the probabilities we are using 1 specific week- Jan 18-24, 2010. This was a sample of 3,638 individuals. Again our three categories of statistical study ar, political party, age, and gender.

Test I: Political Party & Approval Ratings

H_0: There is no relationship between political party and approval of the president
H_a: There is a relationship between political party and approval of the president

Two Way Table:

	Political Party			
Observed	Democrats	Republicans	Independents	Total
Approval	1093	180	506	1779
Disapproval/No Opinion	224	1017	618	1859
Total	1317	1197	1124	3638

X^2 Test:
Expected count = (row total x column total)/n
Expected (Sample) = (1317 x 1779)/3638 = 644.02, (1317 x 1859)/3638 = 672.98
$X^2 = \Sigma$ (observed—expected)^2/expected
Sample: (1093—644.02)^2/644.02 + (224—672.98)^2/672.98 + ...
X^2 = 1168.632 (from software)
Degrees of Freedom = (c-1)(r-1) = 2
P value < 0.000001

Test II: Age & Approval Ratings

Hypotheses

H_0: There is no association between age group and approval

H_a: There is an association between age group and approval

Table of Observed Counts

	Age Group				
	18-29	30-49	50-64	65+	Total
Approval	420	668	403	291	1782
Disapproval/No Opinion	330	616	492	418	1856
Total	750	1284	895	709	3,638

Table of Expected Counts

	Age Group				
	18-29	30-49	50-64	65+	Total
Approval	367.37	621.26	433.04	343.05	1782
Disapproval/No Opinion	377.95	647.06	451.02	357.29	1856
Total	750	1284	895	709	3683

Calculating X^2

$$\frac{(420-367.37)^2}{367.67} + \frac{(668-621.26)^2}{621.26} + \frac{(403-433.04)^2}{433.04} + \frac{(291-343.05)^2}{343.05} + \frac{(330-377.95)^2}{377.95} +$$

$$\frac{(616-647.06)^2}{647.06} + \frac{(492-451.02)^2}{451.02} + \frac{(418-357.29)^2}{357.29} =$$

$= 7.531 + 3.516 + 2.084 + 7.897 + 6.083 + 1.491 + 3.723 + 10.316 = 42.641$

Degrees of Freedom

Df=(2-1)(4-1)=1*3=3

P-Value (From Minitab) P value < 0.000001

Test III: Gender & Approval Ratings

Hypotheses

$H_0 = p1 = p2$ (Males and Females approve of Obama equally/no difference)

$H_a = p1 \neq p2$ (Males and Females do not approve equally of Obama)

p1 represents *females'* approval of Obama

p2 represents *males'* approval of Obama

Comparing 2 Proportions

Population	Sample Size (n)	Successes (X)	X/n (p-hat)
1) Female	1783	928	928/1783= 0.52
2) Male	1855	855	855/1855=0.46

$$Z = \frac{\text{p-hat1} - \text{p-hat2}}{\sqrt{\text{p-hat}\,(1 - \text{p-hat})\,(1/n1 + 1/n2)}}$$

$$\text{P-hat} = \frac{x1 + x2}{n1 + n2}$$

P-hat = 0.49

Z = 3.53

P- Value = 0.0004

Conclusion

Overall, our calculations showed if there were correlations between approval ratings and political party, age groups, and gender. When considering political party, we utilized a two-way table to display the information gathered from our data. We then performed a chi-squared test and calculated the p-value. The p-value for this category calculated to < 0.000001. This extremely small P value indicates that the survey is statistically significant at the 0.01 level. Therefore, we can reject the null hypothesis, and conclude that there is some relationship between political party and approval of the president. This is reasonable, because approval of the president is not a random event, but the result of an individual's political beliefs, among other variables. As for our Test II, we again utilized a chi-squared test. This was useful because of the number of age groups represented in our data. Calculating the p-value (< 0.000001) we were able to reject the null hypotheses at the 0.01 significance level and conclude that there is an association between the age group of an individual and their approval ratings. The results from Test II are reasonable because certain groups are more politically involved than others. The Age Group 30-49 had significantly higher numbers than that of 65+. For Test III we examined gender differences. Because this category had two divisions (male and female) we decided to utilize the proportions test. After calculating the Z statistic (3.53) we then found the p-value (0.0004). Although the p-value was extremely small, it was very significant at the .05 level. It showed that men and women *do* differ in their approval of Obama. This conclusion is also reasonable because men and women play different social roles in society and therefore would have different approvals and disapprovals on a number of issues.

C.3.2 Pittsburgh Panther Win Record at The Pete

Project Assignment 3: Statistical Inference

Introduction:

Over the past seven seasons, the University of Pittsburgh's Men's Basketball team has had a very impressive home record. Throughout the project, many different factors that may have helped the Panther's secure a victory have been investigated. Three hypotheses regarding these factors will be examined:

1. Mean points scored per season is constant.
2. Mean points scored with a win equals mean points scored with a loss.
3. Mean free throw percentage with a win equals mean free throw percentage with a loss.

All of the data that has been gathered in previous sections of this project will either support or reject these hypotheses.

Results:

1. Mean points scored per seasons is constant

$$H_0: \mu_{Pts\,02-03} = \mu_{Pts\,02-03} = \mu_{Pts\,03-04} = \mu_{Pts\,04-05} = \mu_{Pts\,05-06} = \mu_{Pts\,06-07} = \mu_{Pts\,07-08} = \mu_{Pts\,08-09}$$

H_a: The mean points scored per game does in fact vary between seasons.

One-way ANOVA: Pitt Points Scored Versus Season					
Source	DF	SS	MS	F	P
Year	6	2834	472	4.57	0.000
Error	114	11794	103	---	---
Total	120	14628	---	---	---

The probability level is less than α=.01 or .05, so one can reject the null. The mean number of points scored per season is NOT constant.

2. Mean points scored with a win = Mean points scored with a loss

$$H_0: \mu_{Pts\ win} = \mu_{Pts\ loss}$$

$$H_a: \mu_{Pts\ win} > \mu_{Pts\ loss}$$

Two-Sample T-Test

Pitt Points Scored			
Group	N	$\bar{x}$	S
Win	111	77.51	10.48
Loss	10	63.1	8.40

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(77.51 - 63.1) - 0}{\sqrt{\frac{10.48^2}{111} + \frac{8.40^2}{10}}} = 5.08$$

$$P(T > 5.08) = 0.000$$

The probability level is less than α=.01 or .05, so one can reject the null. The mean points score for a win is GREATER than the mean points scored for a loss.

3. Mean free throw % with a win = mean free throw % with a loss

$$H_0: \mu_{FT\%\ win} = \mu_{FT\%\ loss}$$

$$H_a: \mu_{FT\%\ win} > \mu_{FT\%\ loss}$$

Two-Sample T-Test

Pitt Free Throw Percentage			
Group	N	$\bar{x}$	S
Win	111	.6623	.1096
Loss	10	.5827	.1026

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(.6623 - .5827) - 0}{\sqrt{\frac{.1096^2}{111} + \frac{.1026^2}{10}}} = 2.34$$

$$P(T > 2.34) = 0.042$$

The probability level is less than α=.05, so one can reject the null at a 95% confidence level. The mean free throw percentage for a win is GREATER than the mean free throw percentage for a loss.

Discussion

In generating hypotheses for different variables, one either had to compare the mean values of a win and a loss, or compare mean values for every season. The first hypothesis that was formulated was that the average number of points scored in a season was significantly different each year. Because this was testing several mean variables, a one-way ANOVA test was used. To find the probability level, a table was used where the factor was the seasons for which our data was measured. The sum of squares and the mean square was found in order to calculate the F-statistic, 4.75. Once this was found, the probability level could be found using a table of F distribution critical values, and the number obtained was 0.0000. This makes the p-value significant on any interval, thus making it possible to reject the null hypothesis that each season had the same average number of points scored per game. Therefore, it could be concluded that the mean number of points scored per season was not constant.

The second and third hypotheses under investigation regarding the mean points scored with a win equals the mean points scored with a loss and mean free throw percentage with a win equals mean free throw percentage with a loss are perfect examples for utilizing a two-sample t statistic. In these situations the means of the population are unknown and are compared using the t statistic distribution. The purpose of this test is to examine the null hypothesis $H_0: \mu_1 = \mu_2$ where $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2, n_1, n_2$ are all known.

In both situations the t values 5.08 and 2.34 yielded probability level less than α=.05 allowing the null hypothesis to be rejected at the 95% confidence level. For the second hypothesis this means that the mean points scored with a win was greater than mean points scored with a loss, and for the third hypothesis the mean free throw percentage with a win was greater than the mean free throw percentage with a loss.

C.3.3 Comparison of Goalies in the NHL

Introduction

**"Because the demands on a goalie are mostly mental,
it means that for a goalie, the biggest enemy is himself.
Not a puck, not an opponent, not a quirk of size or style…
The successful goalie understands these neuroses, accepts them, and puts them under control.**

-Ken Dryden, Former NHL Goaltender

Our project involved juxtaposing data of NHL goalies who played 500 games between 1988 and 2009. We selected 25 goalies from this time period. The third part of our project required creating and analyzing hypotheses, or testable statements, associated with our data. The ultimate goal entailed proving that goaltenders who played over 500 games (G500) were statistically superior hockey players than the average goaltender.

The basic assumption states that if a goaltender remains in the NHL for 500 games, he must possess greater than average skill to do so. Furthermore, we assumed three pieces of information regarding average or lackluster goaltenders: 1) they win one in every two games; 2) their goals against average (GAA) is 2.75; 3) their save percentage (SV%) is .850. Accordingly, our hypotheses examined each of these assumptions.

Our first hypothesis tested the claim that the mean career win count for G500 exceeded 250 games. Secondly, we examined the claim that the mean career GAA for G500 fell below 2.75. Finally, we analyzed the claim that the mean SV% for G500 surpassed .850. Our predetermined conclusion: the G500 should meet these criteria.

Results

Hypotheses are statements evaluated to determine some conclusion about information. Typically, we use two hypotheses, a *null* (Ho) and an *alternative* (Ha), when analyzing our data. The null refers to the statement we desire to disprove, and the alternative is the statement we want to buttress. In this case, our null reflects the typical goaltender while the alternative represents the G500.

Hypotheses 1: *The mean career win count for goalies who play over 500 games is greater than 250 games*

Ho: Ã = 250
Ha: Ã > 250

Hypothesis 2: *The mean career GAA for goalies who play over 500 games is less than 2.75*

Ho: Ã = 2.75
Ha: Ã < 2.75

Hypothesis 3: *The mean SV% for goalies who play over 500 games is greater than .850*

Ho: Ã = .850
Ha: Ã > .850

Choosing an Appropriate Statistical Technique:

A *significance test* refers to a procedure which helps us ascertain the validity of our hypotheses. We use different significance tests depending on the data we possess about our topic. Since we are using a sample of the population and lack information like the population mean and standard deviation, we must use a *t-distribution*. In this particular significance test, we know sample statistics such as the mean and standard deviation. The t-distribution utilizes the formula

$$t = \frac{x\text{bar} - \mu}{\frac{s}{\sqrt{(n)}}}$$

where *"xbar"* is the mean of the SRS, and μ is the mean value you are testing. *S* represents the standard deviation of the SRS, and *n* is the sample size of the SRS. For the purposes of this project, we assume the population to be distributed normally.

Results (cont'd)

Testing Our Hypotheses:

In order to make a conclusion from a significance test, we must derive a *p-value*. Essentially, a p-value is a probability that helps us determine information about our hypotheses. We test a p-value against a *significance level* in order to make a conclusion about our hypotheses; our significance level will be a=.05. If our p-value from the significance test exceeds .05, *we do not have enough evidence to reject the null hypothesis*. In other words, we cannot prove our alternative hypothesis.

In the following tests, we utilized MINITAB software to determine the sample mean, standard deviation, and p-value.

Test 1:

xbar = 303.64; *S* = 105.73

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{(n)}}} = \frac{303.64 - 250}{\frac{105.73}{\sqrt{(25)}}} = 2.54$$

Ha: $\mu > 250 \rightarrow P(T>2.54) = .009$

$P < a \rightarrow$ We have enough evidence to reject the null and claim that G500 will win more 250 games.

Test 2:

xbar = 2.7512; *S* = .305

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{(n)}}} = \frac{2.7512 - 2.75}{\frac{.305}{\sqrt{(25)}}} = .0197$$

Ha: $\mu < 2.75 \rightarrow P(<.0197) = .508$

$P > a \rightarrow$ We do not have enough evidence to the reject the null and claim that G500 will have a GAA greater than 2.75.

Results (cont'd)

Test 3:

xbar = .905 and *S* = .0089

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{(n)}}} = \frac{.905 - .850}{\frac{.0089}{\sqrt{(25)}}} = 30.89$$

Ha: $\mu > .850 \rightarrow P(T > 30.89) = .000$

$P < a \rightarrow$ We have enough evidence to reject the null and claim that G500 will get a SV% greater than .850.

Discussion

I. Test Used:

Since we had an SRS, we could not determine a population standard deviation. However, we could determine a sample mean and standard deviation. Thus, it was appropriate to use the t-distribution test.

II. Results in Context:

Our hypotheses provided us with many novel findings about the dynamics of goaltending in the NHL. Our first hypothesis helped us verify that goaltenders who play more than 500 games (about 8 seasons) win more often than the "average" goaltender. Therefore, we can conclude that longevity in the NHL correlates with winning. Secondly, the last two hypotheses unveiled unique information about goaltenders puck-stopping abilities. Since GAA did not decrease below 2.75, but SV% surpassed .850, we established that goalies must be facing numerous shots; i.e. they maintained an above average SV% while still allowing in over 2 goals per game.

BIBLIOGRAPHY

[1] M. Aliaga, G. Cobb, C. Cuff, J. Garfield, R. Gould, R. Lock, T. Moore, A. Rossman, B. Stephenson, J. Utts, et al. Gaise college report. American Statistical Association, 2005.

[2] C.M. Anderson-Cook. Designing a first experiment: A project for design of experiment courses. *The American Statistician*, 52(4):338–342, 1998.

[3] Assessment resource tools for improving statistical thinking (artist). https://app.gen.umn.edu/artist/, retrieved on October 26, 2009.

[4] Arthur Bakker and Koeno P.E. Gravemeijer. Learning to reason about distribution. In J. Garfield and D. Ben-Zvi, editors, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pages 147–168. Kluwer Academic, Boston, 2004.

[5] C. Batanero. Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, 2(1):75–97, 2000.

[6] Dani Ben-Zvi. Reasoning about data analysis. In Dani Ben-Zvi and Joan Garfield, editors, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pages 121–145. Kluwer Academic, BOSTON, 2004.

[7] Dani Ben-Zvi and Joan Garfield. Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In Dani Ben-Zvi and Joan Garfield, editors, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pages 3–15. Kluwer Academic, Boston, 2004.

[8] R. Biehler. Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalization from results of two exploratory studies. In G. Burrill and J. Garfield, editors, *Research on the role of technology in teaching and learning statistics: 1996 Proceedings of the 1996 IASE Round Table Conference*, pages 169–190. International Statistical Institute, 1997.

[9] Neil Binner. Using projects to encourage statistical thinking. In *Proceedings of the sixth international conference on teaching statistics*, 2002.

[10] G.W. Bright and S.N. Friel. Graphical representations: Helping students interpret data. In S.P. Lajoie, editor, *Refletions on Statistics: Agendas for learning, teaching, and assessment in K-12*, pages 63–88. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

[11] G.R. Bryce. Data driven experiences in an introductory statistics course for engineers using student collected data. In *1992 Proceedings of the Section on Statistical Education*, pages 155–160, 2003.

[12] J. Butterfield and S. Anderson. *Collins English Dictionary*. HarperCollins, 2003.

[13] Lisa J. Carnell. The effect of a student-designed data collection project on attitudes towards statistics. *Journal of Statistics Education*, 16(1), 2008.

[14] Beth Chance, Robert delMas, and Joan Garfield. Reasoning about sampling distributions. In Dani Ben-Zvi and Joan Garfield, editors, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pages 295–323. Kluwer Academic, Boston, 2004.

[15] B.L. Chance. Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3):1–18, 2002.

[16] S. Cohen and R.A. Chechile. Overview of constats and the constats assessment. In J.B. Garfield and Burrill, editors, *Research on the role of technology in teaching and learning statistics*, pages 101–110. International Statistical Institute, Voorburg, The Netherlands, 1997.

[17] S. Cohen, G. Smith, R.A. Chechile, G. Burns, and F. Tsai. Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics*, 21(1):35–54, 1996.

[18] S. Cohen, F. Tsai, and R.A. Chechile. A model for assessing student interaction with educational software. *Behavior Research Methods, Instruments, and Computers*, 27(2):251–256, 1995.

[19] R.C. delMas, J. Garfield, and A. Ooms. Using assessment to study the development of students' reasoning about sampling distribution. In K. Makar, editor, *Reasoning about Distribution: A collection of research studies. Proceedings of the Seventh International Conference on Teaching Statistics: Working cooperatively in statistics education*, 2005.

[20] Robert delMas, Joan Garfield, Ann Ooms, and Beth Chance. Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2):28–58, 2007.

[21] Robert delMas, Ann Ooms, Joan Garfield, and Beth Chance. Assessing students' statistical reasoning. In *Proceedings of the Seventh International Conference on Teaching Statistics*, 2006.

[22] Robert C. delMas. Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, 10(3), 2002.

[23] Robert C. delMas, Joan Garfield, and Beth L. Chance. A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, 7(3), 1999.

[24] R. Falk and C.W. Greenbaum. Significance tests die hard: The amazing persistence of a a probabilistic misconception. *Theory and Psychology*, 5:75–98, 1995.

[25] Sandra Fillebrown. Using projects in an elementary statistics course for non-science majors. *Journal of Statistics Education*, 2(2), 1994.

[26] Geoffrey T. Fong, David H. Krantz, and Richard E. Nisbett. The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18(3):253–292, 1986.

[27] Iddo Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1):1–51, 2002.

[28] Joan Garfield. How students learn statistics. *International Statistical Review*, 63(1):25–34, 1995.

[29] Joan Garfield. Assessing student learning in the context of evaluating a chance course. *Communication in statistics: Theory and methods*, 25(11):2863–2873, 1996.

[30] Joan Garfield. Challenges in assessing statistical reasoning. In *AERA Annual Meeting presentation*, 1998.

[31] Joan Garfield. Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1):22–38, 2003.

[32] Joan Garfield and Robert delMas. Students' conceptions of probability. In D. Vere-Jones, S. Carlyle, and B.P. Dawkins, editors, *Proceedings of the Third International Conference on Teaching Statistics: Vol 1*, pages 340–349. International Statistical Institute, Voorburg, The Netherlands, 1991.

[33] Joan B. Garfield. An authentic assessment of students' statistical knowledge. In Norman L. Webb and Arthur F. Coxford, editors, *Assessment in the Mathematics Classroom*, pages 187–196. National Council of Teachers of Mathematics, 1993.

[34] Joan B. Garfield and Dani Ben-Zvi. *Developing Students' Statistical Reasoning*. Springer, 2008.

[35] Randall Groth. *Development of a High School Statistical Thinking Framework*. Dissertation, Illinois State University, 2003.

[36] H. Haller and Krauss S. Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1):1–20, 2002.

[37] K.T. Halvorsen and T.L. Moore. Motivating, monitoring and evaluating student projects. In *Proceedings of the Section on Statistical Education*, pages 20–25, 1991.

[38] Robert V. Hogg. Statistical education: Improvements are badly needed. *The American Statistician*, 45(4):342–343, 1991.

[39] John P. Holcome and Rochelle L. Ruffer. Using a term-long project sequence in introductory statistics. *The American Statistician*, 54(1):49–53, 2000.

[40] W.G. Hunter. Some ideas about teaching design of experiments, with 25 examples of experiments conducted by students. *The American Statistician*, 31(1):12–17, 1977.

[41] Multiparameter hypothesis testing and acceptance sampling. Berger, roger. *Technometrics*, 24:295–300, 1982.

[42] Daniel Kahneman, Paul Slovic, and Amos Tversky, editors. *Judgment under Uncertainty: Heuristics and biases*. Cambridge University Press, New York, 1982.

[43] C. Konold and A. Pollatsek. Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33(4):259–289, 2002.

[44] Clifford Konold. Informal conceptions of probability. *Cognition and Instruction*, 6(1):59–98, 1989.

[45] Clifford Konold. Informal conceptions of probability. *Cognition and Instruction*, 6(1):59–98, 1989.

[46] Clifford Konold, Alexander Pollatsek, Arnold Well, Jill Lohmeier, and Abigail Lipson. Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education*, 24(5):392–414, 2003.

[47] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–74, 1977.

[48] Eugene M. Laska and Morris J. Meisner. Testing whether an identified treatment is best. *Biometrics*, 45(4):1139–1151, 1989.

[49] Johannes Ledolter. Projects in introductory statistics courses. *The American Statistician*, 49(4):364–367, 1995.

[50] K. Lipson. The role of computer based technology in developing understanding of the concept of sampling distribution. In *Proceedings of the Sixth International Conference on Teaching Statistics*, 2002.

[51] K. Lipson, S. Kokonis, and G. Francis. Investigation of students experiences with a web-based computer simulation. In *Proceedings of the 2003 IASE Satellite Conference on Statistics Education and the Internet*, 2003.

[52] R.G. Lomax. *Statistical concepts: A second course for education and the behavioral sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition, 2007.

[53] Thomas E. Love. A project-drive second course. *Journal of Statistics Education*, 6(1), 1998.

[54] Marsha Lovett. A collaborative convergence on studying reasoning processes: A case study in statistics. In Sharon M. Carver and David Klahr, editors, *Cognition and Instruction: Twenty-Five Years of Progress*, pages 347–384. Lawrence Arlbaum Associates, Mahwah, New Jersey, 2001.

[55] Margaret Mackisack. What is the use of experiments conducted by statistics students? *Journal of Statistics Education*, 2(1), 1994.

[56] K. Makar and J. Confrey. "variation-talks": Articulating meaning in statistics. *Statistics Education Research Journal*, 4(1):27–54, 2005.

[57] Austin Melton, Beverly M. Reed, and A. Kasturiarachi. A project-based elementary statistics course. In *Proceedings of The Delta '99 Symposium on Undergraduate Mathematics*, pages 142–147, 1999.

[58] K.C. Mittag and B. Thompson. A national survey of aera members' perceptions of statistical signficance tests and other statistical issues. *Educational Researcher*, 29(4):14–20, 2000.

[59] Jan Mokros and Susan Jo Russell. Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26(1):20–30, 1995.

[60] D.S. Moore. *The basic practice of statistics*. W.H. Freeman, New York, 3rd edition, 2004.

[61] George A. Morgan. *SPSS for introductory statistics: use and interpretation*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition, 2007.

[62] R.E. Nisbett, D.H. Krantz, C. Jepson, and Z. Kunda. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90:339–363, 1983.

[63] M. Oakes. *Statistical inference: A commentary for the social and behavioral sciences*. Wiley, New York, 1986.

[64] M. Pfannkuch. Informal inferential reasoning: A case study. In K. Makar, editor, *Reasoning about Distribution: A collection of research studies. Proceedings of the Seventh International Conference on Teaching Statistics: Working cooperatively in statistics education*, 2005.

[65] Alexander Pollatsek, Clifford E. Konold, Arnold D. Well, and Susan D. Lima. Beliefs underlying random sampling. *Memory and Cognition*, 12(4):395–401, 1984.

[66] Alice Richardson. Experimental research in a statistical concepts course. In *Proceedings of the sixth international confernece on teaching statistics*, 2002.

[67] H.V. Roberts. Student-conducted projects in introductory statistics courses. In F. Gordon and S. Gordon, editors, *Statistics for the 21st Century*, pages 109–121. Mathematical Associate of America, Washington, DC, 1992.

[68] Allan J. Rossman and Thomas H. Short. Conditional probability and education reform: Are they compatible? *Journal of Statistics Education*, 3(2), 1995.

[69] A. Sevin. Some tips for helping students in introductory statistics classes carry out successful data analysis projects. In *ASA Proceedings of the Section of Statistical Education*, pages 159–164, 1995.

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

[71] Gary Smith. Learning statistics by doing statistics. *Journal of Statistics Education*, 6(3), 1998.

[72] A. Stone, K. Allen, T.R. Rhoads, T.J. Murphy, R.L. Shehab, and C. Saha. The statistics concept inventory: A pilot study. In *Frontiers in Education Conference*, 2003.

[73] DL Sylvester and RW Mee. Student projects: An important element in the beginning statistics course. In *American Statistical Association: 1992 Proceedings of the Section on Statistical Education*, pages 137–141, 1992.

[74] JA Vallecillos and P. Holmes. Students' understanding of the logic of hypothesis testing. In *Proceedings of the Fourth International Conference on Teaching Statistics*, Marrakech, Morocco, 1994. International Statistical Institute.

[75] Bradley A. Warner, David Pendergraft, and Timothy Webb. That was venn, this is now. *Journal of Statistics Education*, 6(1), 1998.

[76] Jane M. Watson. Developing reasoning about samples. In Dani Ben-Zvi and Joan Garfield, editors, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pages 277–294. Kluwer Academic, Boston, 2004.

[77] C.J. Wild and M. Pfannkuch. Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3):223–265, 1999.

[78] M. Wilkerson and J.R. Olson. Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology*, 131(6):627–631, 1997.

[79] A.M. Williams. Students' understanding of hypothesis testing: The case of the significant concept. In F. Biddulph and K. Carr, editors, *People in mathematics education: Proceedings of the twentieth annual conference of the mathematics education research group of Australasia*, pages 585–591. MERGA, Rotirua, New Zealand, 1997.

[80] A.M. Williams. Novice students' conceptual knowledge of statistical hypothesis testing. In J.M. Truran and K.M. Truran, editors, *Making the difference: Proceedings of the twenty-second annual conference of the mathematics education research group of Australasia*, pages 554–560. MERGA, 1999.

[81] J. Witmer. Using a class project to teach statistics. In *1991 Proceedings of the Section on Statistical Education*, pages 26–28, 1992.

[82] Christopher R. Wolfe. Quantitative reasoning across a college curriculum. *College Teaching*, 41(1):3–9, 1993.

[83] D.A. Zahn. Student projects in a large lecture introductory business statistics course. In *1992 Proceedings of the Section on Statistical Education*, pages 147–154, 1993.